Uniwersytet Warszawski
Wydział Nauk Ekonomicznych
ul. Długa 44/50, 00-241 Warszawa
Studia Podyplomowe
Metody Statystyczne w Biznesie

Mariusz Gromada

REDUKCJA WYMIARU PRÓBY METODĄ ANALIZY SKŁADOWYCH GŁÓWNYCH

Praca wykonana pod kierunkiem:
Dr Piotra Wójcika

WARSZAWA, WRZESIEŃ 2009 r.

Oświadczenie kierującego pracą

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki określone dla prac dyplomowych.

.....	.....
Data	Dr Piotr Wójcik

Oświadczenie autora pracy

Mając świadomość odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem świadectwa studiów podyplomowych lub tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersja elektroniczną.

.....	.....
Data	Mariusz Gromada

Streszczenie

Praca przedstawia metodę analizy składowych głównych, jako jedną z metod redukcji wymiaru próby. Zmniejszenie liczby zmiennych objaśniających pozwala często uprościć model, ułatwić interpretację wyników, zachowując jednocześnie niezbędny do dalszej analizy poziom informacji. W rozdziale 1 wprowadzamy podstawy teoretyczne metody w oparciu o algebrę macierzową opisaną w dodatku A. Część praktyczna pracy to redukcja wymiaru zbioru danych opisujących różne typy szkła w zależność od jego składu chemicznego oraz właściwości fizycznych. Wynik w kolejnym kroku stanowi podstawę do przeprowadzenia analizy dyskryminacji metodą lasów losowych, co pozwala zmierzyć pogorszenie jakości klasyfikacji, w konsekwencji ocenić poziom utraty informacji w zbiorze o zredukowanym wymiarze.

Wstęp

Postęp informatyzacji życia codziennego dokonał się za sprawą ogromnego wzrostu możliwości komputerów, w szczególności ich mocy obliczeniowej, jak też zdolności do przechowywania olbrzymich ilości danych. Liczba *baz danych*[1] rośnie w niewiarygodnym tempie. Tak szerokie wykorzystanie systemów gromadzących dane wynika bezpośrednio z relatywnie prostej konstrukcji oraz akceptowalnych kosztów. W efekcie przechowuje się miliony (a nawet miliardy) rekordów[2] opisanych wektorem obserwacji o bardzo dużym wymiarze (setki / tysiące cech). Jednym z celów gromadzenia danych jest *odkrywanie ukrytych w nich zależności*[3]. Nowoczesnych metod analizy takich danych dostarcza współczesna *statystyka wielowymiarowa*, jednak radzenie sobie z dużymi zbiorami danych wymaga konstrukcji specjalnych technik.

Jest mało prawdopodobne, aby w bazach danych zawierających miliony rekordów opisanych tysiącami zmiennych, nie wystąpiły *zależności* pomiędzy tymi wszystkimi zmiennymi. Analityk danych powinien w szczególności unikać *współliniowości zmiennych objaśniających* (sytuacje kiedy istnieje *korelacja* pomiędzy kilkoma zmiennymi objaśniającymi). Współliniowość może prowadzić do *niestabilności* w przestrzeni rozwiązań. Według Bellmana [16] wraz ze wzrostem liczby zmiennych objaśniających wykładniczo rośnie rozmiar próby niezbędnej do dopasowania funkcji od nich zależnej. Użycie zbyt wielu zmiennych niepotrzebnie komplikuje również interpretację wynikowego modelu, naruszając *zasadę zwięzłości*, mówiącą, że należy dążyć do takiej liczby zmiennych, które można łatwo zinterpretować. Wykorzystanie zbyt wielu zmiennych prowadzić może również do *nadmiernego dopasowania* modelu do danych trenujących, co uniemożliwia dokonanie dalszych uogólnień. Należy przy tym zwrócić uwagę, że kilka zmiennych objaśniających może w naturalny sposób utworzyć jeden spójny komponent, opisujący ten sam aspekt danych. Przykładowo *stan konta*, *wartość domu*, *wartość portfela akcji*, może przedstawiać *majątek* danej osoby. Innymi słowy *przestrzenie wielowymiarowe* rzadko istnieją. U podstaw metod *redukcji wymiaru* leżą powyższe intuicje. Celem redukcji wymiaru jest: z*mniejszenie liczby zmiennych objaśniających poprzez utworzenie nowych zmiennych*, *uzyskanie niezależności nowych zmiennych opisujących*, *stworzenie możliwości interpretacji nowych zmiennych opisujących.* Podstawowe metody redukcji wymiaru to: *Analiza Składowych Głównych*, *Analiza Czynnikowa*[4], *Skalowanie Wielowymiarowe* [5].

[1] Baza danych - zbiór danych zapisanych w ściśle określony sposób w strukturach odpowiadających założonemu modelowi danych. W potocznym ujęciu obejmuje dane oraz program komputerowy wyspecjalizowany do gromadzenia i przetwarzania tych danych [7].

[2] Rekord - pojedynczy wpis do bazy danych.

[3] Eksploracja danych (ang. data mining) - to jeden z etapów procesu odkrywania wiedzy z baz danych (ang. Knowledge Discovery in Databases, KDD) [5].

[4] Analiza czynnikowa (ang. Factor analysis) - to metoda statystyczna, służąca odnajdywaniu struktur w zbiorze zmiennych losowych.

[5] Skalowanie wielowymiarowe (ang. Multidimensional scaling - MDS) – technika statystyczna, mająca na celu wykrycie zmiennych ukrytych, które choć nie obserwowane bezpośrednio, wyjaśniają podobieństwa i różnice pomiędzy badanymi obiektami [3].

Wszystkie powyższe metody wykorzystują fakt, że z reguły wielowymiarowe obserwacje (powiedzmy z przestrzeni $\mathbb{R}^p$) nie są równomiernie „rozrzucone“ wzdłuż wszystkich kierunków, lecz koncentrują się w pewnych *podprzestrzeniach* przestrzeni wyjściowej. Kierunki, wzdłuż których znajduje się większość obserwacji, nie muszą pokrywać się z kierunkami osi wyjściowego układu współrzędnych. Powstaje więc zadanie odszukania tych podprzestrzeni, *zrzutowania* na nie obserwacji, w konsekwencji uzyskania redukcji wymiaru próby. Jest to szczególnie użyteczne podczas *wizualizacji* danych (ich graficznego przedstawienia), gdzie bardzo istotne są metody odpowiedniego rzutowania na określone płaszczyzny lub powierzchnie.

Przywołując literaturę przedmiotu należy wskazać następujące pozycje: Koronacki, Ćwik „Statystyczne systemy uczące się” [13], Krzyśko, Wołyński, Górecki „Systemy uczące się“ [14], Larose „Metody i modele eksploracji danych” [16], Larose „Odkrywanie wiedzy z danych“ [15].

W niniejszej pracy przedstawiamy metodę *składowych głównych*. Podajemy jej podstawy teoretyczne. W części praktycznej przeprowadzamy redukcję wymiaru danych opisujących różne typy szkła w zależność od jego składu chemicznego oraz właściwości fizycznych (zawartość różnych pierwiastków chemicznych). Skuteczność metody znajduje potwierdzenie w wynikach, gdzie znacząca redukcja wymiaru nie spowodowała dużej utraty zmienności oraz znaczącego powiększenie błędu klasyfikacji w dalej przeprowadzonej *analizie dyskryminacyjnej*[6]. Wszelkie obliczenia przeprowadzone zostały przy użyciu statystycznego pakietu `R-cran`[7] w wersji `2.9.2` [17].

[6] Analiza dyskryminacyjna (ang. discriminant analysis) – zespół metod wielowymiarowej analizy danych, które czynią z tej techniki niezwykle efektywne narzędzie do zagadnień klasyfikacyjnych i technik eksploracja danych. Zalicza się do metod klasyfikacji pod nadzorem. W metodzie wyróżnia się dwa główne etapy: etap uczenia / budowy modelu, etap klasyfikacji / wykorzystania modelu [2].

[7] GNU R - język programowania i środowisko do obliczeń statystycznych i wizualizacji wyników. Jest to projekt GNU podobny do języka i środowiska S stworzonego w Bell Laboratories (dawniejsze AT&T, obecnie Lucent Technologies) przez Johna Chambersa i jego współpracowników. R może być traktowane jako implementacja języka i całego środowiska S [6].

Spis treści

Streszczenie 2

Wstęp 2

1. Analiza składowych głównych 6
1.1. Wprowadzenie 6
1.2. Podstawy teoretyczne metody 7
1.2.1. Formalna definicja modelu 7
1.2.2. Kryteria dla liczby składowych - metody pomijania 9
1.2.3. Interpretacja składowych 11

2. Analiza danych 13
2.1. Charakterystyka zbioru danych 13
2.2. Wstępna analiza 14
2.3. Cel badania 16
2.4. Algorytm postępowania 17
2.5. Zastosowanie metody składowych głównych 18
2.5.1. Liczba składowych 18
2.5.2. Rzutowanie danych 19
2.5.3. Utrata informacji w zredukowanym zbiorze 22
2.6. Opis wyników - wnioski 23

3. Podsumowanie 24

A. Algebra liniowa 25
A.1. Macierze i wyznaczniki 25
A.2. Wymiar przestrzeni liniowej 29
A.3. Wartości własne i wektory własne 31

B. Statystyka matematyczna 33
B.1. Miary położenia, rozproszenia, współzależności 33

Rozdział 1

Analiza składowych głównych

1.1. Wprowadzenie

Pierwszą postać analizy składowych głównych[1] zaproponował na początku XX wieku Karl Pearson[2]. Rozwinięcie metody podał Harold Hotelling[3] w pracy z roku 1934 [13]. Hotelling stosując metodę do analizy wyników testów szkolnych wykazał, że istnieją dwie główne składowe wyników: zdolności słowne oraz zdolności matematyczne .

Analiza składowych głównych jest techniką eksploracji danych bez nadzoru. Celem analizy składowych głównych nie jest predykcja wartości zmiennej objaśnianej, nie stosuje się podziału na zmienne objaśniające i objaśniane (wszystkie zmienne traktuje się równorzędnie - każda obserwacja składa się jedynie z wektora cech). Badacz dąży do redukcji wymiaru (redukcji liczby zmiennych). Aby taka redukcja była możliwa, między zmiennymi oryginalnymi muszą zachodzić statystyczne zależności. Istotą metody jest wytłumaczenie zmienności wielowymiarowych danych z wykorzystaniem nowych nieskorelowanych zmiennych, będących liniowymi kombinacjami (def. A.2.3) zmiennych obserwowanych. Nowe zmienne (główne składowe) porządkuje się ze względu na znaczenie: pierwsza tłumaczy największą część wariancji, każda kolejna (i-ta) największą część wariancji pod warunkiem braku korelacji z każdą poprzednią $(1, 2, \ldots, i-1)$. Kolejny istotny cel metody to nadanie interpretacji składowym poprzez ich powiązania z wyjściowymi zmiennymi.

Metoda ta jest obecnie często stosowana. Przekształcenie licznego wektora obserwacji w kilka niezależnych składowych głównych stanowi na ogół wstęp do zastosowania kolejnych metod wielowymiarowej statystycznej analizy danych, głównie analizy skupień oraz analizy dyskryminacji [13] [14] [15] [11].

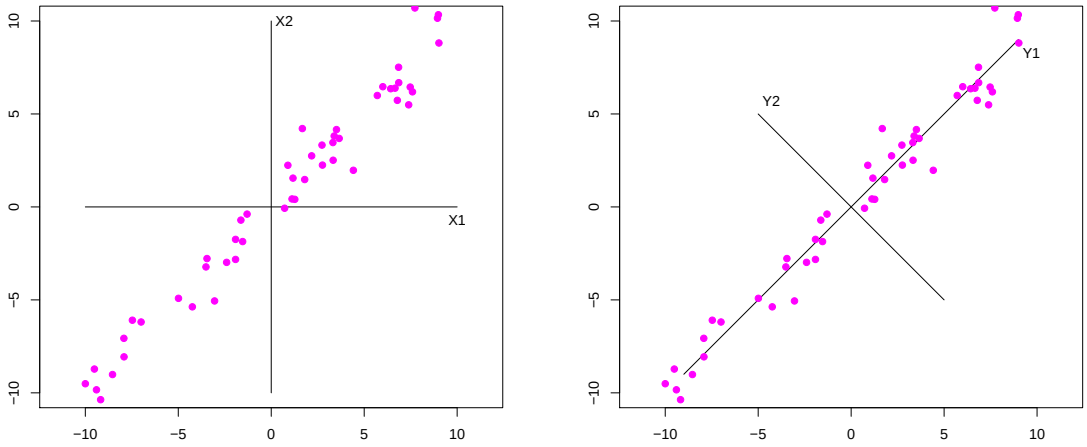
W dalszym tekście wykorzystujemy pojęcia algebry macierzowej, której podstawy (wraz z podstawowymi terminami statystyki matematycznej) zostały przestawione w dodatkach A, B (w szczególności wprowadzono pojęcie *wymiaru przestrzeni liniowej* (def. A.2.5)).

[1] Analiza składowych głównych (ang. Principal Component Analysis, PCA)
[2] Karl Pearson (1857 - 1936) - angielski matematyk, prekursor statystyki matematycznej [4].
[3] Harold Hotelling (1895 - 1973) - statystyk i ekonomista amerykański [1].

1.2. Podstawy teoretyczne metody

Metoda głównych składowych to transformacja p-wymiarowego układu współrzędnych w układ q-wymiarowy ($q \leq p$) (bardziej formalnie przestrzeni p-wymiarowej w przestrzeń q-wymiarową - pojęcie wymiaru przestrzeni liniowej przedstawiono w A.2.5). W przypadku, gdy $q = p$ zachodzi jedynie obrót oryginalnej przestrzeni (brak redukcji wymiaru). Zwykle w praktyce redukcja jest osiągana ($q < p$), poszukujemy takiej liczby głównych składowych, aby wyjaśnić zmienność wyjściowych zmiennych z niewielką utratą informacji.

Na rysunku 1.1 przedstawiliśmy przykładową transformację dwuwymiarowego układu współrzędnych. Wykres pierwszy (współrzędne X_1, X_2) wskazuje na duża zmienność w zakresie każdej ze zmiennych. Prosta transformacja układu (obrót do współrzędnych (Y_1, Y_2)) pozwala stwierdzić, że w rzeczywistości zmienność obserwowanych danych zachodzi głównie w jednym z dwóch wymiarów (tutaj Y_1). Rezygnacja z wymiaru Y_2 nie prowadzi do znacznej utraty informacji (oryginalne zmienne są silnie zależne).



Rysunek 1.1: Transformacja układu współrzędnych (X_1, X_2) na nowy układ (Y_1, Y_2)

1.2.1. Formalna definicja modelu

Załóżmy, że dana jest próba losowa pochodząca z populacji o ciągłym rozkładzie w przestrzeni $\mathbb{R}^p$ z wektorem wartości oczekiwanych $\mathbf{m}$ oraz macierzą kowariancji $\mathbf{\Sigma}$ (def. B.1.4) .

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p \end{bmatrix}$$

Przez $\mathbf{x}$ oznaczamy wektor losowy pochodzący z tej populacji. Niech $\mathbf{a} \in \mathbb{R}^n$ będzie dowolnym wektorem jednostkowym (def. A.2.7) . Rozpatrzmy standaryzowaną kombinację liniową (def. A.2.8) $\mathbf{a}^T\mathbf{x}$ wektora losowego $\mathbf{x}$. Celem jest określenie kolejnych głównych składowych wektora $\mathbf{x}$.

Definicja 1.2.1 Zmienną losową

$$\gamma_{(1)}^T(\mathbf{x}-\mathbf{m}) \tag{1.1}$$

gdzie $\gamma_{(1)} \in \mathbb{R}^p$ spełnia warunek

$$\mathrm{Var}\Big(\gamma_{(1)}^T\mathbf{x}\Big) = \max_{\mathbf{a}\in\mathbb{R}^p,\ ||\mathbf{a}||=1} \Big\{\mathrm{Var}(\mathbf{a}^T\mathbf{x})\Big\} \tag{1.2}$$

nazywamy *pierwszą składową główną* wektora $\mathbf{x}$ [13]. W (1.2) pisząc $||\mathbf{a}||$ rozumiemy normę wektora $\mathbf{a}$ (def. A.2.7) .

Poszukujemy standaryzowanej kombinacji liniowej maksymalizującej zmienność mierzoną wariancją. W interpretacji geometrycznej poszukujemy kierunku $\gamma_{(1)} \in \mathbb{R}^p$, by rzut ortogonalny wektora losowego $\mathbf{x}$ na ten kierunek maksymalizował zmienność. Bez wpływu na wariancję zmiennej losowej $\gamma_{(1)}^T\mathbf{x}$ przeprowadza się „centrowanie“ poprzez odjęcie stałej $\gamma_{(1)}^T\mathbf{m}$.

Definicja 1.2.2 $k+1$ *składową główną* wektora $\mathbf{x}$ nazywamy zmienną losową

$$\gamma_{(k+1)}^T(\mathbf{x}-\mathbf{m}) \tag{1.3}$$

gdzie $\gamma_{(k+1)} \in \mathbb{R}^p$ spełnia warunek

$$\mathrm{Var}\Big(\gamma_{(k+1)}^T\mathbf{x}\Big) = \max_{\mathbf{a}\in\mathbb{R}^p,\ ||\mathbf{a}||=1} \Big\{\mathrm{Var}(\mathbf{a}^T\mathbf{x})\Big\} \tag{1.4}$$

pod warunkiem, że zmienna ta jest nieskorelowana z wcześniejszymi składowymi głównymi: pierwszą, drugą, ..., k-tą. Wektory $\gamma_{(1)}, \gamma_{(2)}, \ldots, \gamma_{(p)}$ nazywamy *wektorami ładunków* [13].

Kolejne wektory ładunków wyznaczają kolejne kierunki największej zmienności wektora losowego $\mathbf{x}$.

Twierdzenie 1.2.1 Niech $\mathbf{x} \in \mathbb{R}^p$ będzie wektorem losowym o wektorze wartości oczekiwanych $\mathbf{m}$ i macierzy kowariancji $\mathbf{\Sigma}$ i niech wartości własne tej macierzy (def. A.3.4) $\lambda_1, \lambda_2, \ldots, \lambda_p$, spełniają warunek

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$$

Wektor ładunku $\gamma_{(k)}$ odpowiadający k-tej składowej głównej wektora $\mathbf{x}$ równy jest jednostkowemu wektorowi własnemu (def. A.3.2) odpowiadającemu wartości własnej λ_k [13].

Dowód powyższego twierdzenia można znaleźć w [13] (strony 235 - 237). Wspomniany dowód wskazuje również, że wariancja k-tej składowej równa jest k-tej wartości własnej λ_k macierzy kowariancji $\mathbf{\Sigma}$ [13] [14] [16].

$$\mathrm{Var}\big(\gamma_{(k)}^T(\mathbf{x}-\mathbf{m})\big) = \lambda_k \quad \text{dla } k = 1, 2 \ldots, p \tag{1.5}$$

Całkowita wariancja to suma wartości własnych macierzy $\mathbf{\Sigma}$, która dana jest śladem (def. A.1.8) macierz $\mathbf{\Sigma}$.

$$\sum_{k=1}^{p} \lambda_k = \mathrm{tr}(\mathbf{\Sigma}) \tag{1.6}$$

Definicja 1.2.3 Miarą wyjaśnionej zmienności wektora losowego $\mathbf{x}$ przez k jego pierwszy składowych głównych nazywamy wskaźnik

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} 100\% \tag{1.7}$$

Macierz $\mathbf{\Sigma}$ można przedstawić w postaci *rozkładu spektralnego*, co prowadzi do wniosku, że pierwszych k składowych wyjaśnia również nie wprost kowariancje [14] [13]

Wszystkie p składowe tworzą wektor $\mathbf{y}$ postaci

$$\mathbf{y} = \mathbf{\Gamma}^T(\mathbf{x} - \mathbf{m}) \tag{1.8}$$

Wektor $\mathbf{y}$ powstaje więc z wektora $\mathbf{x}$ poprzez jego przesunięcie (scentrowanie) o wektor $\mathbf{m}$, a następnie obrót przez macierz ortogonalną $\mathbf{\Gamma}^T$. Składowe główne są więc względem siebie ortogonalne (prostopadłe). Jako kombinacje liniowe składowe główne reprezentują hiperpłaszczyzny, na które są rzutowane ortogonalnie obserwowane punkty. Składowe główne nie są niezmiennicze względem zmiany skali zmiennych [14] [13].

1.2.2. Kryteria dla liczby składowych - metody pomijania

Stwierdzenie możliwości redukcji wymiaru danych wymaga posiadania kryterium. Ile składowych wybrać? Które uznać za nieistotne i w konsekwencji pominąć? Do osiągnięcia powyższego celu badacz może wykorzystać kilka rozsądnych metod.

Kryterium wykresu osypiskowego

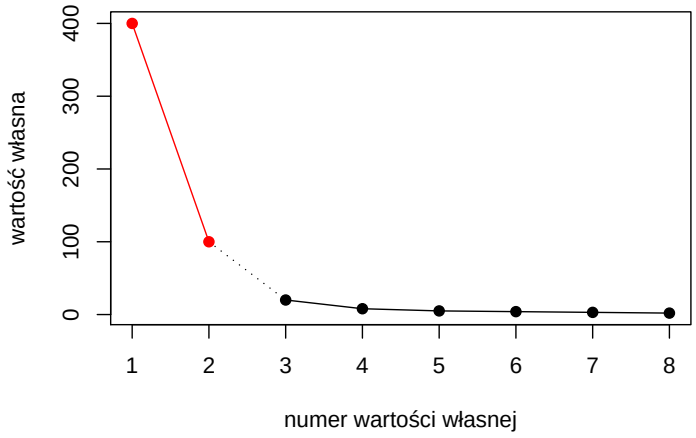
Na osi odciętych zaznaczamy numer wartości własnej (wartości własne uprzednio uporządkowane w kolejności nierosnącej), na osi rzędnych nanosimy wielkość wartości własnej. Tak powstałe punkty łączymy liniami. Otrzymany wykres nazywamy *wykresem piargowym* (lub *wykresem osypiska*). Przykładowy wykres osypiska zaprezentowany jest na rysunku 1.2. Wiele problemów analizowanych metodą składowych głównych będzie przedstawiało właśnie taki typ wykresu.

Zwróćmy uwagę, że pierwsze dwie wartości własne istotnie różnią się od pozostałych, wykazujących trend liniowy o niewielkim spadku. W takim przypadku pomijamy wszystkie składowe główne poza dwiema pierwszymi. Niektóre przypadki mogą wymagać pozostawienia również składowej odpowiadającej wartości własnej rozpoczynającej płaski trend liniowy.

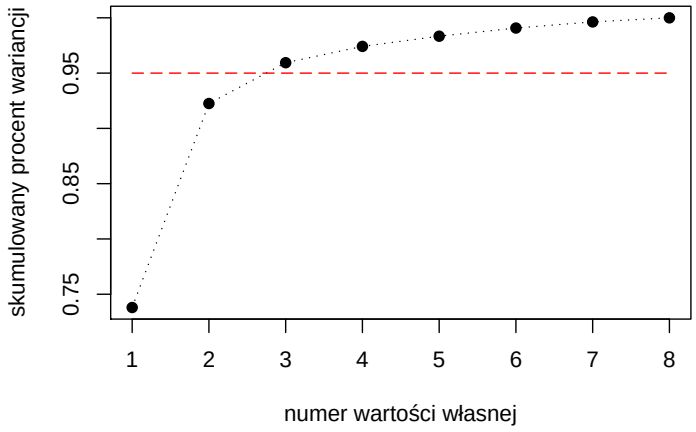
Kryterium części wariancji

W tym przypadku analityk określa, jaka część zmienności powinna być wyjaśniona przez składowe główne. W kolejnym kroku stosowany jest wskaźnik 1.7 pozwalający dobrać minimalne k, przy którym wartość wskaźnika spełnia postawiony warunek. Rysunek 1.3 przedstawia przykładowy wykres skumulowanego procentu wyjaśnionej wariancji w zależności od liczby składowych głównych. Badacz wymagając wyjaśnienia zmienności na poziomie 95% wybiera 2 pierwsze składowe.

Nietrudno zauważyć podobieństwo interpretacyjne metody opartej na kryterium części wariancji do analizy wartości współczynnika determinacji dla zadania regresji [12]. Przy regresji



Rysunek 1.2: Przykładowy wykres osypiska dla problemu z ośmioma składowymi głównymi.



Rysunek 1.3: Przykładowy wykres skumulowanej wariancji wyjaśnionej w zależności od liczby składowych głównych.

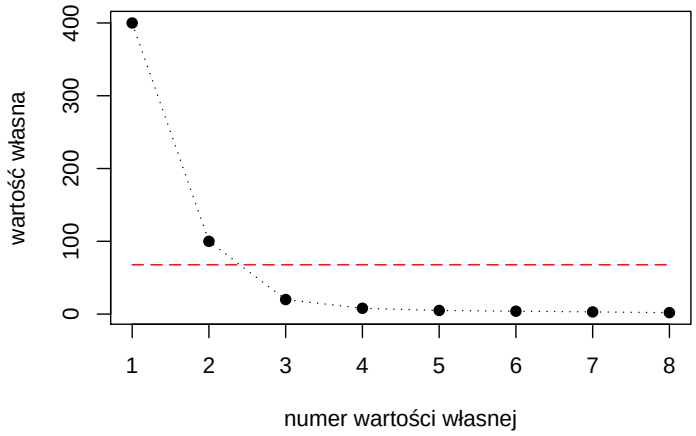
liniowej pytamy jaka wartość współczynnika determinacji jest odpowiednia? Analogiczne pytanie jest również zasadne w przypadku analizy składowych głównych. Odpowiedź w dużej mierze zależy od dziedziny, w której prowadzone są badania.

Kryterium średniej wartości własnej

Zgodnie z (1.6) całkowita wariancja składowych głównych dana jest śladem macierzy kowariancji $\mathbf{\Sigma}$, który równy jest sumie wszystkich wartości własnych macierzy $\mathbf{\Sigma}$. Wyznaczając średnią wartość własną

$$\bar{\lambda} = \frac{1}{p}\sum_{k=1}^{p}\lambda_k = \frac{1}{p}\mathrm{tr}(\mathbf{\Sigma})$$

otrzymujemy średnią wariancję składowych głównych, która jest zarazem średnią wariancją zmiennych pierwotnych. W kryterium średniej wartości własnej pomijamy wszystkie składowe, których wartości własne są mniejsze od średniej wartości własnej. Zastosowanie kryterium średniej wartości własnej do problemu przedstawionego na rysunku 1.4 nakazuje pozostawić jedynie dwie pierwsze główne składowe.



Rysunek 1.4: Przykładowy wykres osypiska z naniesioną średnią wartością własną.

Kryterium wartości własnej

W praktyce często w miejsce macierzy kowariancji stosuje się jej „unormowany" odpowiednik w postaci macierzy korelacji (def. B.1.5) .

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

Macierz korelacji na diagonali posiada same jedynki. W tak rozważnym problemie całkowita zmienność to liczba zmiennych pierwotnych (p) oraz zmienność średnia wynosi 1. Stosując kryterium średniej wartości własnej odrzucamy te składowe, których wartości własne są mniejsze od 1.

1.2.3. Interpretacja składowych

Interpretacja składowych polega najczęściej na analizie *wkładu* zmiennych pierwotnych w budowę składowej głównej. Odbywa się to poprzez porównanie modułów współczynników

stojących przy danej zmiennej pierwotnej. Maksymalna wartość modułu współczynnika stojącego przy zmiennej pierwotnej wskazuje na maksymalny wkład tej zmiennej w budowę składowej głównej. Jeżeli składową główną przedstawioną jako kombinacja liniowa zmiennych pierwotnych

$$a_1x_1 + a_2x_2 + \ldots + a_px_p$$

zapiszemy w postaci

$$a_{i_1}x_{i_1} + a_{i_2}x_{i_2} + \ldots + a_{i_p}x_{i_p}$$

gdzie

$$|a_{i_1}| \geq |a_{i_2}| \geq \ldots \geq |a_{i_p}|$$

dla $i_j = 1, 2, \ldots p$, oraz $i_j \neq i_k$ przy $j \neq k$

to wkład w budowę składowej głównej kolejno od największego do najmniejszego wnoszą zmienne $x_{i_1}, x_{i_2}, \ldots, x_{i_p}$.

Do interpretacji składowych głównych niektórzy autorzy zalecają wykorzystanie wartości współczynnika korelacji pomiędzy zmienną pierwotną a składową główną, jednak zgodnie z [14] (str. 339) nie jest to obecnie rekomendowane podejście ze względu na utratę wielowymiarowej informacji odnośnie łącznego wkładu zmiennych w budowę składowej głównej.

Rozdział 2

Analiza danych

Do analizy wybrano dane opisujące różne typy szkła, pochodzące z badania przeprowadzonego przez Vina Spiehler, Ph.D., DABFT z Diagnostic Products Corporation. Motywem badania typu szkła okazały się śledztwa kryminalne prowadzone w USA (USA Forensic Science Service), gdzie szkło pozostawione na miejscu zbrodni mogło zostać wykorzystane jako dowód, jeżeli zostało poprawnie sklasyfikowane. Dane jak też opis zbioru znajduj się na stronie `UCI Machine Learning Repository` pod adresem:

`http://archive.ics.uci.edu/ml/datasets/Glass+Identification` [8]

Analiza danych w pełni przeprowadzono w pakiecie statystycznym `R cran`. Wykorzystane moduły:

1. `stats` - The R Stats Package.
2. `randomForest` - Breiman and Cutler's random forests for classification and regression.

2.1. Charakterystyka zbioru danych

Nazwa zbioru: `Glass Identification`
Liczba obserwacji w zbiorze: 214
Liczba atrybutów: 9 (+ 1 określający klasę - typ szkła)

Wyróżniono 6 typów szkła:

1. `building windows float processed` (szkło okienne - szyba),
2. `building windows non float processed` (szkło okienne - inne niż szyba),
3. `vehicle windows float processed` (szkło okienne samochodowe - szyba),
4. `vehicle windows non floaat processed` (szkło okienne samochodowe - inne niż szyba),
5. `containers` (pojemniki),
6. `tableware` (zastawa stołowa, szkło użytku kuchennego),
7. `headlamps` (reflektor samochodowy).

Szkło opisano 9 atrybutami:

1. `RI (Refractive Index)` - współczynnik załamania światła,
2. `Na (sód)` - zawartość procentowa sodu,
3. `Mg (magnez)` - zawartość procentowa magnezu,
4. `Al (glin)` - zawartość procentowa glinu,
5. `Si (silikon)` - zawartość procentowa silikonu,
6. `K (potas)` - zawartość procentowa potasu,
7. `Ca (wapń)` - zawartość procentowa wapnia,
8. `Ba (bar)` - zawartość procentowa baru,
9. `Fe (żelazo)` - zawartość procentowa żelaza.

Jednostki dla atrybutów (2 - 9) - `procent wagi danego pierwiastka w odpowiadającym tlenku`.

2.2. Wstępna analiza

W zbiorze znajduje się 214 obserwacji dla 6 typów szkła (brak klasy o numerze 4 „szkło okienne samochodowe – inne niż szyba“). Rozkład klas

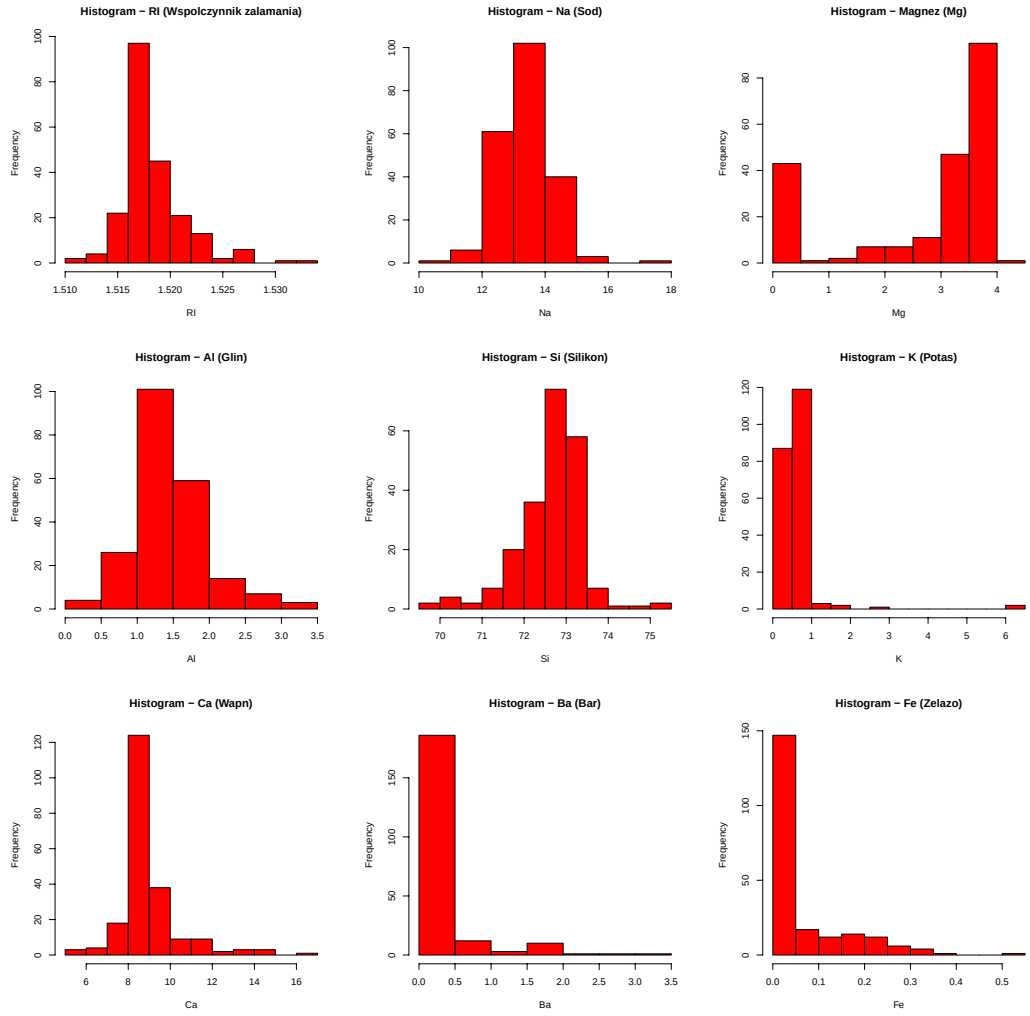
1. `32,7%` (`70` obserwacji),
2. `35,5%` (`76` obserwacji),
3. `7,9%` (`17` obserwacji),
4. `0%` (`0` obserwacji),
5. `6,1%` (`13` obserwacji),
6. `4,2%` (`9` obserwacji),
7. `13,6%` (`29` obserwacji).

Podstawowe statystyki (*wartość minimalna*, *wartość maksymalna*, *mediana*, *średnia*, *1 oraz 3 kwartyl*) dla poszczególnych zmiennych objaśniających zawarte są w tabeli 2.1.

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
Min.	1.511	10.730	0.000	0.290	69.810	0.000	5.430	0.000	0.000
1st Qu.	1.517	12.910	2.115	1.190	72.280	0.123	8.240	0.000	0.000
Median	1.518	13.300	3.480	1.360	72.790	0.555	8.600	0.000	0.000
Mean	1.518	13.410	2.685	1.445	72.650	0.497	8.957	0.175	0.057
3rd Qu.	1.519	13.820	3.600	1.630	73.090	0.610	9.172	0.000	0.100
Max.	1.534	17.380	4.490	3.500	75.410	6.210	16.190	3.150	0.510

Tablica 2.1: Podstawowe statystyki dla zmiennych występujących w zbiorze danych glass.

Rozkłady (*histogramy*) poszczególnych zmiennych objaśniających przedstawione są na rysunku 2.1. Rysunek 2.5 przedstawia wykresy *pudełkowe*.

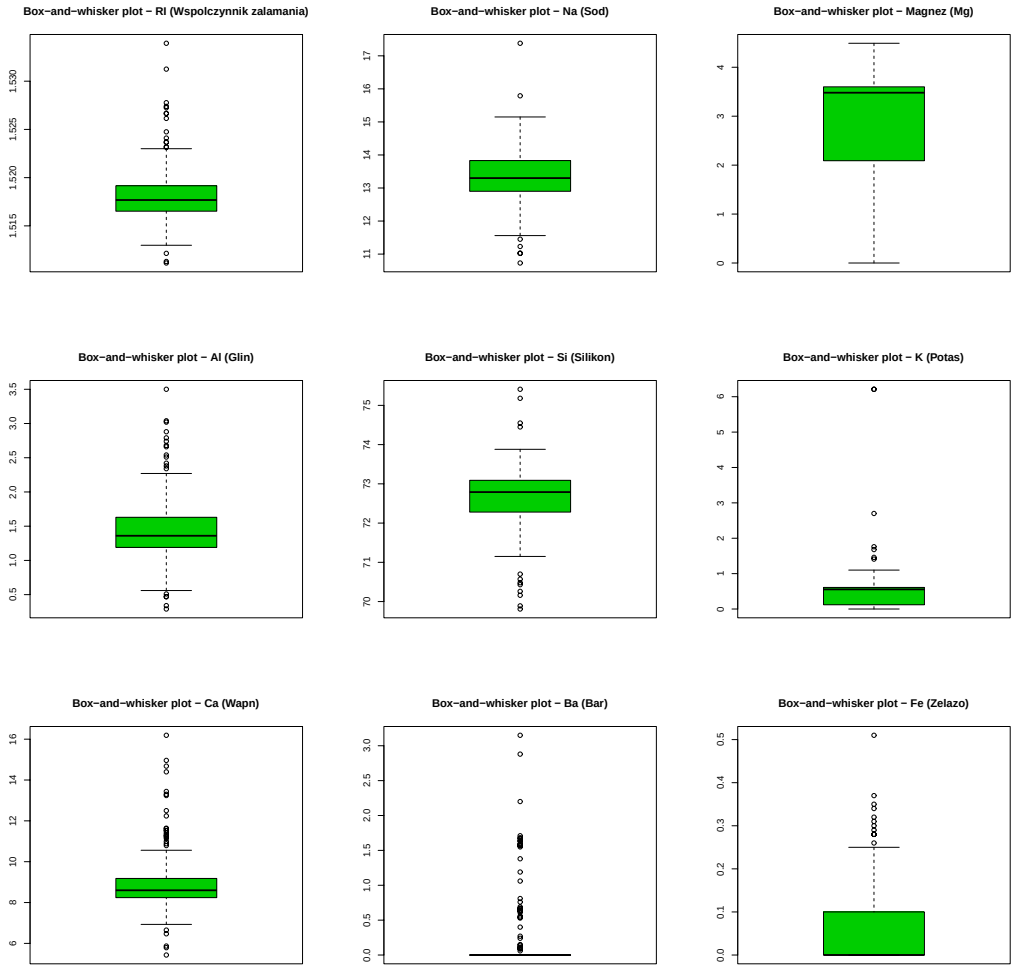


Rysunek 2.1: Histogramy poszczególnych zmiennych ze zbioru danych glass.

Wszystkie rozkłady są *jednomodalne* (oprócz zmiennej `Mg` opisującej procentową zawartość magnezu w badanym szkle). Wykresy pudełkowe pozwalają zaobserwować *skośności*. Dodatkowo dla większości zmiennych występują *obserwacje odstające* ze względu na badaną cechę (wynikające z różnic pomiędzy typami analizowanego szkła).

Macierz korelacji zaprezentowana w tabeli 2.2 wskazuje na dodatnią korelacje (`0.81`) pomiędzy zawartością wapnia (`Ca`) oraz współczynnikiem załamania światła (`RI`).

Brak innych ewidentnych korelacji, co potwierdza również zbiór wykresów rozproszenia dla poszczególnych par zmiennych objaśniających.



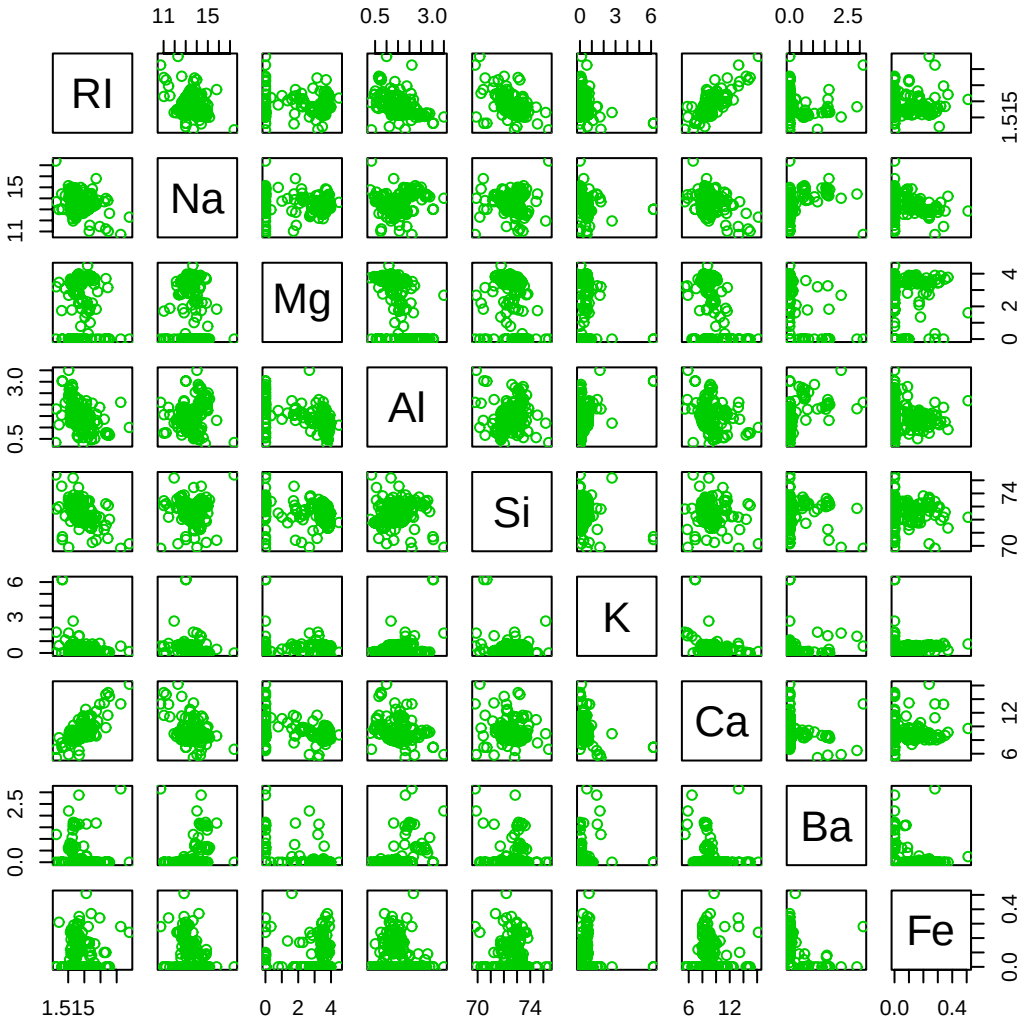
Rysunek 2.2: Wykresy pudełkowe poszczególnych zmiennych ze zbioru danych glass.

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
RI	1.00	-0.19	-0.12	-0.41	-0.54	-0.29	**0.81**	0.00	0.14
Na	-0.19	1.00	-0.27	0.16	-0.07	-0.27	-0.28	0.33	-0.24
Mg	-0.12	-0.27	1.00	-0.48	-0.17	0.01	-0.44	-0.49	0.08
Al	-0.41	0.16	-0.48	1.00	-0.01	0.33	-0.26	0.48	-0.07
S	-0.54	-0.07	-0.17	-0.01	1.00	-0.19	-0.21	-0.10	-0.09
K	-0.29	-0.27	0.01	0.33	-0.19	1.00	-0.32	-0.04	-0.01
Ca	**0.81**	-0.28	-0.44	-0.26	-0.21	-0.32	1.00	-0.11	0.12
Ba	0.00	0.33	-0.49	0.48	-0.10	-0.04	-0.11	1.00	-0.06
Fe	0.14	-0.24	0.08	-0.07	-0.09	-0.01	0.12	-0.06	1.00

Tablica 2.2: Macierz korelacji dla poszczególnych zmiennych ze zbioru danych glass.

2.3. Cel badania

Celem badania jest r*edukcja wymiaru danych glass* przy wykorzystaniu analizy składowych głównych. Redukcja wymiaru powinna zostać tak przeprowadzona, aby utrata informacji



Rysunek 2.3: Wykresy rozproszenia dla par poszczególnych zmiennych ze zbioru danych glass.

(mierzona wyjaśnioną wariancją) nie przekraczała około 5% (nowe zmienne powinny wyjaśnić około 95% zmienności modelu). Jako dodatkowy miernik utraty informacji przedstawimy porównanie średnich błędów klasyfikacji dla *lasów losowych* Breimana[1] przed i po redukcji wymiaru.

2.4. Algorytm postępowania

1. Wyznaczenie macierzy kowariancji.
2. Wyznaczenie wartości własnych macierzy kowariancji.
3. Wyznaczenie wektorów własnych (składowych).

[1]

4. Wybór składowych głównych (różne kryteria - dyskusja).
5. Rotacja danych (redukcja wymiaru - powstaje nowy zbiór danych).
6. Wybór identyfikatorów obserwacji do zbioru uczącego oraz do zbioru testowego.
7. Wybór zbioru uczącego i testowego na podstawie identyfikatorów z punktu 6 ze zbioru oryginalnego (bez redukcji) - las losowy - pomiar błędu klasyfikacji na zbiorze testowym.
8. Wybór zbioru uczącego i testowego na podstawie identyfikatorów z punktu 6 z nowego zbioru (po redukcji) - las losowy - pomiar błędu klasyfikacji na zbiorze testowym.
9. 20 krotne powtórzenie punktów 6 - 8.
10. Wyznaczenie średnich błędów klasyfikacji dla lasu losowego na zbiorze przed redukcją oraz po redukcji.
11. Porównanie średnich błędów klasyfikacji jako dodatkowo miara utraty informacji przy redukcji wymiaru metodą składowych głównych.

Dlaczego lasy losowe?

Metoda klasyfikacji oparta na lasach losowych należy do tych wykorzystujących *rodziny klasyfikatorów* (inne przykładowe metody wykorzystujące rodziny klasyfikatorów to *bagging*, *boosting*). W lesie losowym budujemy kilka drzew[10], następnie decyzja o wyborze klasy podejmowana jest na podstawie *reguły głosowania* (ostateczny wybór pada na najczęściej wskazywaną klasę). W ten sposób otrzymuje się *wzmocnienie siły klasyfikacji*, co wiąże się z lepszym odzwierciedleniem wewnętrznych zależności w zbiorze danych. Nasz wybór padł na lasy losowe również z innego powodu. Drzewa wchodzące w skład lasu losowego budowane są w pewien specjalny sposób, różniący je od tradycyjnych drzew klasyfikacyjnych. W momencie *wzrostu* (budowy kolejnych węzłów), decyzje o wyborze atrybutów użytych do podziałów w tych węzłach, podejmowane są na podstawie l*isty wylosowanych atrybutów* z ogólnie dostępnych. W ten sposób każdy z atrybutów ma szanse być *wykorzystanym* (lub *pominiętym*), nawet te z niewielką (lub bardzo dużą) miarą siły klasyfikacyjnej. Metoda głosowania dokonuje pewnego rodzaju *uśrednienia ostatecznego wyniku*. Po więcej informacji na temat lasów losowych zainteresowanego czytelnika odsyłamy do literatury [10].

2.5. Zastosowanie metody składowych głównych

2.5.1. Liczba składowych

Postępując zgodnie z algorytmem wskazanym powyżej rozpoczynamy od wyznaczenia macierzy kowariancji dla 9 zmiennych objaśniających - tabela 2.3.
W kolejnym kroku analizujemy wartości własne macierzy. Z tabeli 2.4 wynika, że 4 pierwsze składowe wyjaśniają `94,92%` ogólnej wariancji, co jest bardzo bliskie progu określonemu w celu badania.

Kryterium średniej wartości własnej (około `0.7`) wskazuje na wybór 2 pierwszych składowych. Zauważmy jednak, że wartości własne odpowiadające 2 i 3 składowej są jedynie niewiele mniejsze od średniej wartości własnej (są bardzo bliskie `0.7`). Istotne różnice występują poczynając od 5 składowej. 3 i 4 składowa wyjaśniają razem ponad `20%` zmienności w danych,

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
RI	0.000009	-0.000476	-0.000536	-0.000618	-0.001275	-0.000574	0.003502	-0.000001	0.000042
Na	-0.000476	0.666841	-0.322422	0.063926	-0.044154	-0.141713	-0.320106	0.132611	-0.019204
Mg	-0.000536	-0.322422	2.080540	-0.346967	-0.185375	0.005076	-0.910916	-0.353047	0.011674
Al	-0.000618	0.063926	-0.346967	0.249270	-0.002136	0.106138	-0.184450	0.119010	-0.003620
S	-0.001275	-0.044154	-0.185375	-0.002136	0.599921	-0.097662	-0.230085	-0.039340	-0.007109
K	-0.000574	-0.141713	0.005076	0.106138	-0.097662	0.425354	-0.295006	-0.013820	-0.000491
Ca	0.003502	-0.320106	-0.910916	-0.184450	-0.230085	-0.295006	2.025366	-0.079848	0.017329
Ba	-0.000001	0.132611	-0.353047	0.119010	-0.039340	-0.013820	-0.079848	0.247227	-0.002844
Fe	0.000042	-0.019204	0.011674	-0.003620	-0.007109	-0.000491	0.017329	-0.002844	0.009494

Tablica 2.3: Macierz kowariancji dla danych glass.

	wartość własna	% wariancji	skumulowany % wariancji
λ_1	3.002009	47.62%	47.62%
λ_2	1.659173	26.32%	73.94%
λ_3	0.679577	10.78%	84.72%
λ_4	0.643166	10.20%	94.92%
λ_5	0.208457	3.31%	98.23%
λ_6	0.101165	1.60%	99.83%
λ_7	0.008999	0.14%	99.98%
λ_8	0.001477	0.02%	100.00%
λ_9	0.000001	0.00%	100.00%

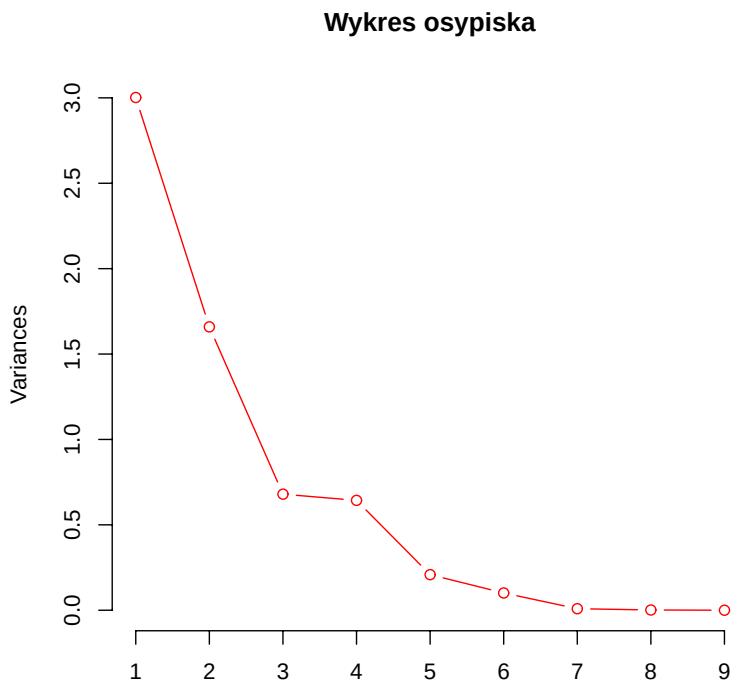
średnia wartość własna
0.700447

Tablica 2.4: Wartości własne macierzy kowariancji oraz procenty wyjaśnionej wariancji dla danych glass.

co pozwala przyjąć, że na podstawie kryterium średniej wartości własnej również wybieramy 4 pierwsze składowe.
Wykres osypiska zamieszczony na rysunku 2.5.1 także przedstawia stabilny trend, bliski liniowemu, rozpoczynający się wraz z wartością własną odpowiadającą 5 składowej głównej. Powyższa argumentacja wraz z określonymi celami badania pozwala jednoznacznie zdecydować o redukcji 9 wymiarowej przestrzeni do jej 4 wymiarowego odpowiednika.

2.5.2. Rzutowanie danych

Wyznaczając wektory własne macierzy kowariancji odpowiadające kolejnym wartościom własnym otrzymujemy współczynniki kombinacji liniowych tworzących nowe składowe główne. W tabeli 2.5 przedstawiamy 9 składowych głównych (`PC1 - PC9`).
Przypomnijmy, iż zdecydowaliśmy o redukcji do 4 pierwszych wymiarów, zatem wybieramy pierwsze 4 składowe główne. Przy analizie wkładu poszczególnych zmiennych oryginalnych w budowę składowych głównych, pomijamy te zmienne, dla których wartości bezwzględne współczynników im odpowiadających są istotnie mniejsze od `0.4`. Dodatkowo szarym kolorem zaznaczono współczynniki większe od `0.3`, ale mniejsze od `0.4` - tabela 2.6.



Rysunek 2.4: Wykres osypiska dla wartości własnych macierzy kowariancji danych glass.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
RI	-0.0009	0.0015	-0.0014	0.0003	0.0007	0.0018	0.0003	0.0041	1.0000
Na	-0.0172	-0.3988	-0.6549	-0.3466	-0.3984	-0.0156	0.0377	0.3622	-0.0014
Mg	0.7235	0.5431	-0.1312	-0.0987	0.0768	-0.0478	0.0750	0.3753	-0.0018
Al	-0.0463	-0.2588	0.0557	0.2709	0.3135	-0.7804	0.0748	0.3761	0.0000
S	0.0077	-0.1941	0.6920	-0.5701	-0.1033	0.0603	0.0587	0.3751	-0.0002
K	0.0784	-0.1038	0.2186	0.6777	-0.5080	0.2652	0.0604	0.3794	-0.0014
Ca	-0.6797	0.6167	-0.0788	-0.0539	-0.0657	-0.0289	0.0738	0.3714	-0.0031
Ba	-0.0764	-0.2235	-0.1339	0.0971	0.6807	0.5601	0.0789	0.3656	-0.0030
Fe	-0.0009	0.0168	0.0072	0.0111	0.0267	-0.0009	-0.9842	0.1739	-0.0004

Tablica 2.5: Wektory własne odpowiadające poszczególnym wartościom własnym macierzy kowariancji dla danych glass.

Profile składowych głównych

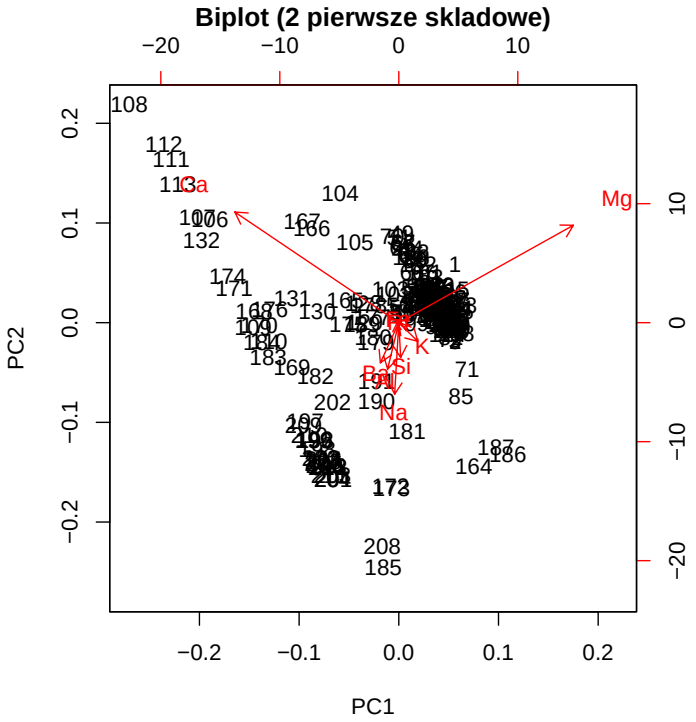
`PC1` pierwsza składowa główna jest skonstruowana głównie ze zmiennych określających zawartość wapnia (`Ca`) oraz magnezu (`Mg`). Obie zmienne wnoszą podobny wkład w składową. Przeciwne znaki współczynników wskazują na ujemną korelację, co świadczy o powiązaniu wysokich zawartości magnezu z niskimi poziomami wapnia (i odpowiednio - niskich zawartości magnezu z wysokimi poziomami wapnia). Zgodnie z tabelą 2.4 pierwsza składowa wyjaśnia około `47.6%` zmienności w danych.

`PC2` druga składowa główna również skonstruowana jest głównie ze zmiennych określają-

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
RI									1.0000
Na		-0.3988	-0.6549	-0.3466	-0.3984			0.3622	
Mg	0.7235	0.5431						0.3753	
Al					0.3135	-0.7804		0.3761	
S			0.6920	-0.5701				0.3751	
K				0.6777	-0.5080			0.3794	
Ca	-0.6797	0.6167						0.3714	
Ba					0.6807	0.5601		0.3656	
Fe							-0.9842		

Tablica 2.6: Profile składowych głównych dla danych Glass. Ukryto współczynniki istotnie mniejsze od 0.4, współczynniki większe od 0.3 ale mniejsze od 0.4 zaznaczono szarym kolorem.

cych zawartość wapnia (`Ca`) oraz magnezu (`Mg`). Podobnie jak w przypadku pierwszej składowej - obie zmienne wnoszą podobny wkład w składową. Tym razem znaki współczynników są dodatnie (dodatnia korelacja), co świadczy o relacji opisującej nieco inną część zmienności w danych. Dodatkowo w skład drugiej składowej z ujemnym znakiem wchodzi zmienna określająca zawartość sodu (`Na`). Składowa wiąże wysokie zawartości magnezu z wysokimi poziomami wapnia oraz niskimi poziomami sodu) (i odpowiednio - niskie zawartości magnezu z niskimi poziomami wapnia oraz wysokimi poziomami sodu). Zgodnie z tabelą 2.4 druga składowa wyjaśnia około `26.3%` zmienności w danych.



Rysunek 2.5: Wykres typu biplot dla analizy składowych głównych danych glass.

PC3 w skład trzeciej składowej wchodzą głównie zmienne określających zawartość sodu (Na) oraz silikonu (Si). Obie te zmienne wnoszą podobny wkład w składową. Znaki współczynników są ujemne (ujemna korelacja). Składowa wiąże niskie zawartości sodu z wysokimi poziomami silikonu (i odpowiednio - wysokie zawartości sodu z niskimi poziomami silikonu). Zgodnie z tabelą 2.4 trzecia składowa wyjaśnia około 10.8% zmienności w danych.

PC4 w skład czwartej składowej wchodzą głównie zmienne określających zawartość silikonu (Si), potasu (K) oraz sodu (Na). Największy wkład wnosi potas, następnie silikon i sód. Znaki współczynników świadczą o ujemnym powiązaniu łącznej zawartości sodu i silikonu z poziomem potasu. Składowa wiąże niskie zawartości silikonu i sodu z wysokimi poziomami potasu (i odpowiednio - wysokie zawartości silikonu i sodu odpowiadają niskim poziomom potasu). Zgodnie z tabelą 2.4 składowa wyjaśnia około 10.2% zmienności w danych.

Z obrazu konstrukcji wybranych składowych oraz wykresu typu biplot (rysunek 2.5) wyłania się duże znaczenie zmiennych przedstawiających zawartość magnezu (Mg), wapnia (Ca) i sodu (Na).

2.5.3. Utrata informacji w zredukowanym zbiorze

Podstawową miarą utraty informacji w zredukowanym zbiorze jest procent ogółu wariancji wyjaśnionej wybranymi składowymi głównymi. W naszym przypadku 9 oryginalnych zmiennych objaśniających zredukowaliśmy do 4 nowych składowych głównych zachowując blisko 95% ogólnej wyjściowej zmienności. Kolejnym krokiem jest wykonanie analizy dyskryminacyjnej w celu porównania możliwości jakości klasyfikacji na zbiorze przed redukcją i po redukcji. Jak wyjaśniono powyżej wybrany klasyfikatora to lasy losowe Breimana.

W celu zapewnienia maksymalnej zgodności zbiory uczące i testowe dla danych przed i po redukcji zostały zbudowane z tych samych obserwacji (te same indeksy obserwacji). Parametry klasyfikacji również ustawiono w sposób identyczny:

1. minimalny rozmiar węzła - 10,
2. maksymalna głębokość drzewa - 2,
3. liczba drzew w lesie losowym - 1000.

Algorytm

Liczba obserwacji 214. Współczynnik podziału zbiór uczący / zbiór testowy = 2/3 (143 rekordy w zbiorze uczącym / 71 rekordów w zbiorze testowy.

1. Wylosuj bez zwracania 143 indeksy (od 1 do 214) do zbioru testowego (train), pozostałe 71 indeksów zapisz do zbioru testowego (test)
2. Zbuduj zbiory: $\texttt{train}_{\texttt{or}}$ - zbiór uczący (obserwacje oryginalne), $\texttt{test}_{\texttt{or}}$ - zbiór testowey (obserwacje oryginalne), $\texttt{train}_{\texttt{red}}$ - zbiór uczący (obserwacje po redukcji), $\texttt{test}_{\texttt{red}}$ - zbiór testowy (obserwacje po redukcji).
3. Zbuduj klasyfikatory metoda lasów losowych: $\texttt{train-RF}_{\texttt{or}}\texttt{(train}_{\texttt{or}}\texttt{)}$ - las losowy na podstawie zbioru uczącego zawierającego obserwacje oryginalne, $\texttt{train-RF}_{\texttt{red}}\texttt{(train}_{\texttt{red}}\texttt{)}$ - las losowy na podstawie zbioru uczącego zawierającego obserwacje po redukcji.

Poprawność klasyfikacji (las losowy)

iteracja	9 atrybutów (oryginalny zbiór)	4 składowe główne
1	75.9%	66.7%
2	75.9%	70.4%
3	75.9%	70.4%
4	70.4%	66.7%
5	70.4%	64.8%
6	79.6%	72.2%
7	79.6%	68.5%
8	70.4%	75.9%
9	66.7%	61.1%
10	75.9%	68.5%
11	70.4%	61.1%
12	85.2%	79.6%
13	75.9%	72.2%
14	70.4%	64.8%
15	79.6%	63.0%
16	81.5%	75.9%
17	83.3%	74.1%
18	81.5%	72.2%
19	70.4%	59.3%
20	74.1%	63.0%
średnia	75.6%	68.5%

Tablica 2.7: Wynik 20 krotnej analizy dyskryminacji metodą lasów losowych dla danych glass przed redukcją i po redukcji.

4. Sprawdź poprawność klasyfikacji: $\texttt{test-RF}_{\texttt{or}}(\texttt{test}_{\texttt{or}})$, $\texttt{test-RF}_{\texttt{red}}(\texttt{test}_{\texttt{red}})$
5. Zapamiętaj poprawność klasyfikacji.
6. Powtórz 20-krotnie kroki `(1 - 5)`.
7. Wyznacz średnie poprawności klasyfikacji dla zbiorów przed i po redukcji.

Realizację kroków `1 - 7` przedstawiono w tabeli 2.7.

2.6. Opis wyników - wnioski

Ucząc i testując klasyfikator przy wykorzystaniu obserwacji ze zbioru oryginalnego osiągnęliśmy średnio `75.6%` poprawnych klasyfikacji, natomiast przy zastosowaniu zbioru zredukowanego otrzymaliśmy wynik na poziomie `68.5%`. Różnica rzędu `7.1%` wskazuje, że redukując zbiór metodą składowych głównych niepoprawnie sklasyfikowano średnio więcej o `7%` obserwacji ze zbioru testowego (więcej o około `5` obserwacji). Zatem `5%` utrata wariancji przełożyła się na około `7%` zwiększenie błędu klasyfikacji, co jest wynikiem akceptowalnym, dającym się uzasadnić w sposób bezpośredni właśnie zmniejszeniem wymiaru próby.

Rozdział 3

Podsumowanie

Przedstawiając zagadnienia związane z redukcją wymiaru, wskazaliśmy przyczyny, dla których taka procedura wydaje się być celowa. Są nimi najczęściej rozmiary współczesnych baz danych. Zmniejszenie liczby zmiennych objaśniających pozwala często uprościć model, ułatwić interpretacje wyników, zachowując jednocześnie niezbędny do dalszej analizy poziom informacji. Należy pamiętać, że takie metody są zazwyczaj wstępem do kolejnych analiz, często do analizy dyskryminacyjnej, bądź też analizy skupień. W części praktycznej pracy przedstawiliśmy przypadek 9 wymiarowego zbioru danych. Przy pomocy metody składowych głównych zredukowaliśmy jego wymiar do 4 nowych zmiennych, tracąc 5% wyjaśnionej ogólnej zmienności (wariancji) oraz zwiększając średnio o 7% błąd klasyfikacji w wykonanej dalej analizie dyskryminacyjnej.

Dodatek A

Algebra liniowa

Wszelkie definicje oraz twierdzenia zawarte w dodatku A wprowadzono na podstawie pozycji [20]. Czytelnika bardziej zainteresowanego odsyłamy właśnie do tej literatury.

A.1. Macierze i wyznaczniki

Definicja A.1.1 Niech $\mathbb{P}$ będzie pierścieniem przemiennym z jedynką . *Macierzą typu* $m \times n$ nad $\mathbb{P}$ (lub $m \times n$ - macierzą nad $\mathbb{P}$) nazywamy funkcję

$$\mathbf{A} : \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\} \to \mathbb{P}$$

tj. funkcję, która każdej parze (i, j), gdzie $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$, przyporządkuje element $\mathbf{A}(i, j)$ pierścienia $\mathbb{P}$. Najczęściej zamiast $\mathbf{A}(i, j)$ pisze się a_{ij} i przedstawia się macierz $\mathbf{A}$ w postaci

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Wartości a_{ij} nazywa się wyrazami macierzy $\mathbf{A}$. Często pisze się również $\mathbf{A} = [a_{ij}]_m^n$. Symbolem $\mathbb{P}_m^n$ oznaczać będziemy zbiór wszystkich macierzy typu $m \times n$ nad pierścieniem $\mathbb{P}$. (cytat [20]).

Definicja A.1.2 Niech $\mathbf{A} \in \mathbb{P}_m^n$. Jeśli $\alpha = (i_1, i_2, \ldots, i_p)$ jest ciągiem elementów zbioru $\{1, 2, \ldots, m\}$ to symbolem $\mathbf{A}_\alpha$ oznaczamy macierz typu $p \times n$ nad $\mathbb{P}$ zdefiniowaną przez

$$\mathbf{A}_\alpha(k, j) = \mathbf{A}(i_k, j)$$

dla $k = (1, 2, \ldots, p)$, $j = (1, 2, \ldots, n)$. Analogicznie, jeśli $\beta = (j_1, j_2, \ldots, j_r)$ jest ciągiem elementów zbioru $\{1, 2, \ldots, n\}$, to symbolem $\mathbf{A}^\beta$ oznaczamy macierz typu $m \times r$ nad P daną wzorem

$$\mathbf{A}^\beta(i, l) = \mathbf{A}(i, j_l)$$

.

Łatwo zauważyć, że $(\mathbf{A}_\alpha)^\beta = (\mathbf{A}^\beta)_\alpha$ jest macierzą typu $p \times r$ nad $\mathbb{P}$. Macierz taką oznacza się krótko $\mathbf{A}_\alpha^\beta$

Definicja A.1.3 Jeżeli α i β są ciągami rosnącymi to macierz $\mathbf{A}_\alpha^\beta$ nazywamy *podmacierzą* macierzy $\mathbf{A}$.

Szczególnymi przykładami podmacierzy są macierze typu $1 \times n$ zwane *wierszami* macierzy $\mathbf{A}$ oraz macierze typu $m \times 1$ zwane *kolumnami* macierzy $\mathbf{A}$.

$$\mathbf{A}_{(i)} = [a_{i1}, a_{i2}, \ldots, a_{in}] \quad i - \text{ty wiersz}$$

$$\mathbf{A}^{(j)} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix} \quad j - \text{ta kolumna}$$

Definicja A.1.4 Jeżeli $\mathbf{A}, \mathbf{B} \in \mathbb{P}_m^n$ i niech $c \in \mathbb{P}$. *Sumą macierzy* $\mathbf{A}$ i $\mathbf{B}$ nazywamy $m \times n$ macierz $\mathbf{A} + \mathbf{B}$ nad $\mathbb{P}$ daną wzorem

$$(\mathbf{A} + \mathbf{B})(i, j) = \mathbf{A}(i, j) + \mathbf{B}(i, j)$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ b_{21} & \cdots & b_{2n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & \cdots & a_{2n} + b_{2n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$$

dla $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. Iloczynem macierzy $\mathbf{A}$ przez element c pierścienia $\mathbb{P}$ nazywamy $m \times n$ macierz $c\mathbf{A}$ nad $\mathbb{P}$ daną wzorem

$$(c\mathbf{A})(i, j) = c\mathbf{A}(i, j)$$

$$c\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} ca_{11} & \cdots & ca_{1n} \\ ca_{21} & \cdots & ca_{2n} \\ \vdots & \ddots & \vdots \\ ca_{m1} & \cdots & ca_{mn} \end{bmatrix}$$

dla $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$.

Definicja A.1.5 Niech $\mathbf{A} \in \mathbb{P}_m^n$, $\mathbf{B} \in \mathbb{P}_n^p$. *Iloczynem macierzy* $\mathbf{A}$ i $\mathbf{B}$ nazywamy taką $m \times p$ macierz $\mathbf{AB}$ nad $\mathbb{P}$, że

$$\mathbf{AB}(i, j) = \sum_{k=1}^{n} \mathbf{A}(i, k)\mathbf{B}(k, j)$$

$$\begin{array}{c} \\ \mathbf{A} = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ a_{i1} & a_{i2} & \cdots & a_{in} \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \end{array} \quad \begin{array}{c} \mathbf{B} \\ \| \\ \begin{bmatrix} \cdots & b_{1j} & \cdots \\ \cdots & b_{2j} & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & b_{nj} & \cdots \end{bmatrix} \\ \\ \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \sum_{k=1}^{n} a_{ik} b_{kj} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \\ \| \\ \mathbf{AB} \end{array} \tag{A.1}$$

dla $i = 1, 2, \ldots m$, $j = 1, 2, \ldots, p$.

Definicja A.1.6 *Macierzą zerową* nazywamy macierz, której wszystkie wyrazy są równe 0.

$$\mathbf{0}_m^n = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

Definicja A.1.7 Macierz typu $n \times n$ nazywamy *macierzą kwadratową*. Mówimy, że wyrazy $a_{11}, a_{22}, \ldots, a_{nn}$ leżą na *głównej przekątnej* macierzy kwadratowej. Macierz kwadratową $\mathbf{A}$ spełniającą warunek $\mathbf{A}(i,j) = 0$ dla $i \neq j$ nazywamy *macierzą diagonalną*.

$$\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \tag{A.2}$$

Definicja A.1.8 *Śladem macierzy* kwadratowej $\mathbf{A} \in \mathbb{P}_n^n$ nazywamy sumę elementów głównej przekątnej macierzy $\mathbf{A}$ i piszemy

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

Definicja A.1.9 Macierz diagonalną typu $n \times n$ nazywamy *macierzą jednostkową* i oznaczamy $\mathbf{I}_n$ lub $\mathbf{I}$ jeżeli jej główna przekątna składa się z samych jedynek.

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{A.3}$$

Twierdzenie A.1.1 Jeżeli $\mathbf{A}, \mathbf{A}' \in \mathbb{P}_m^n$, $\mathbf{B}, \mathbf{B}' \in \mathbb{P}_n^p$, $\mathbf{C} \in \mathbb{P}_p^r$, $a \in \mathbb{P}$, to zachodzą poniższe tożsamości

1° $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$

2° $(\mathbf{A} + \mathbf{A}')\mathbf{B} = \mathbf{AB} + \mathbf{A}'\mathbf{B}$ i $\mathbf{A}(\mathbf{B} + \mathbf{B}') = \mathbf{AB} + \mathbf{AB}'$

3° $(a\mathbf{A})\mathbf{B} = \mathbf{A}(a\mathbf{B}) = a(\mathbf{AB})$

4° $\mathbf{I}_m\mathbf{A} = \mathbf{A} = \mathbf{A}\mathbf{I}_n$

Definicja A.1.10 *Transpozycją macierzy* $\mathbf{A} \in \mathbb{P}_m^n$ (lub *macierzą transponową* względem macierzy $\mathbf{A}$) nazywamy taką macierz $\mathbf{A}^T \in \mathbb{P}_n^m$, że

$$\mathbf{A}^T(i,j) = A(j,i)$$

dla $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Macierz $\mathbf{A}$ spełniającą warunek $\mathbf{A} = \mathbf{A}^T$ nazywamy *macierzą symetryczną*.

Twierdzenie A.1.2 Własności transpozycji macierzy

1° $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$ dla $\mathbf{A}, \mathbf{B} \in \mathbb{P}_m^n$,

2° $(c\mathbf{A})^T = c\mathbf{A}^T$, dla $\mathbf{A} \in \mathbb{P}_m^n$, $c \in \mathbb{P}$,

3° $(\mathbf{A}^T)^T = \mathbf{A}$,

4° $\mathbf{I}^T = \mathbf{I}$,

5° $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ dla $\mathbf{A} \in \mathbb{P}_m^n$, $\mathbf{B} \in \mathbb{P}_n^p$.

Definicja A.1.11 Macierz kwadratową $\mathbf{A} \in \mathbb{P}_n^n$ nazywamy *macierzą odwracalną* jeżeli istnieje taka macierz $\mathbf{A}^{-1} \in \mathbb{P}_n^n$, że

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$$

Macierz $\mathbf{A}^{-1}$ nazywamy *macierzą odwrotną* do macierzy $\mathbf{A}$.

Twierdzenie A.1.3 Zachodzą poniższe tożsamości

1° macierz jednostkowa $\mathbf{I}$ jest macierzą odwracalną oraz $\mathbf{I}^{-1} = \mathbf{I}$,

2° jeżeli $\mathbf{A}$ jest macierzą odwracalną to $\mathbf{A}^{-1}$ jest macierzą odwracalną i $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$,

3° jeśli $\mathbf{A}$ i $\mathbf{B}$ są macierzami odwracalnymi typu $n \times n$ to $\mathbf{AB}$ jest macierzą odwracalną oraz $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$,

3° jeśli $\mathbf{A}$ jest macierzą odwracalną to $\mathbf{A}^T$ jest macierzą odwracalną i $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$

Definicja A.1.12 *Wyznacznikiem* macierzy kwadratowej $\mathbf{A} \in \mathbb{P}_n^n$ nazywamy element pierścienia $\mathbb{P}$ oznaczony $\det \mathbf{A}$ lub $|\mathbf{A}|$ dany zależnością rekurencyjną

1. wyznacznik macierzy typu 1×1 $\det[a] = a$,
2. jeżeli
$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$
to
$$\det \mathbf{A} = \sum_{i=1}^{n} a_{ij}^d \tag{A.4}$$
gdzie
$$a_{ij}^d = (-1)^{i+j} \det \mathbf{A}_{i,j} \tag{A.5}$$
oraz przez $\mathbf{A}_{i,j}$ rozumie się macierz typu $n-1 \times n-1$ otrzymaną z macierzy $\mathbf{A}$ poprzez usunięcie (skreślenie) i-tego wiersza oraz j-tej kolumny. Element a_{ij}^d nazywamy *dopełnieniem algebraicznym* wyrazu a_{ij} macierzy $\mathbf{A}$. Macierz
$$\mathbf{A}^D = [a_{ij}^d]^T$$
nazywamy *dopełnieniem* macierzy $\mathbf{A}$.

Twierdzenie A.1.4 Własności wyznacznika macierzy

1° $\det \mathbf{A} = \det \mathbf{A}^T$ dla dowolnej kwadratowej macierzy $\mathbf{A}$,

2° $\det(\mathbf{AB}) = \det\mathbf{A}\det\mathbf{B}$ dla dowolnych kwadratowych $n \times n$ macierzy $\mathbf{A}, \mathbf{B}$,

3° $\mathbf{A}\mathbf{A}^D = \mathbf{A}^D\mathbf{A} = (\det\mathbf{A})\mathbf{I}_n$ dla dowolnej kwadratowej macierzy $\mathbf{A} \in \mathbb{P}_n^n$,

4° jeśli $\mathbf{A} \in \mathbb{P}_n^n$ jest macierzą odwracalną to $\det\mathbf{A}$ jest odwracalnym elementem pierścienia $\mathbb{P}$ i $\det(\mathbf{A}^{-1}) = (\det\mathbf{A})^{-1}$,

5° jeśli $\det\mathbf{A}$ jest odwracalnym elementem pierścienia $\mathbb{P}$ to $\mathbf{A}$ jest macierzą odwracalną i $\mathbf{A}^{-1} = (\det\mathbf{A})^{-1}\mathbf{A}^D$,

6° jeśli $\mathbf{A}, \mathbf{B} \in \mathbb{P}_n^n$ oraz $\mathbf{AB} = \mathbf{I}_n$ lub $\mathbf{BA} = \mathbf{I}_n$, to macierz $\mathbf{A}$ jest odwracalna i $\mathbf{B} = \mathbf{A}^{-1}$,

7° jeśli $\mathbb{P}$ jest ciałem i $\mathbf{A} \in \mathbb{P}_n^n$ to $\mathbf{A}$ jest macierzą odwracalną wtedy i tylko wtedy, gdy $\det\mathbf{A} \neq 0$.

A.2. Wymiar przestrzeni liniowej

*Przestrzeń liniowa*a (lub *przestrzeń wektorowa*) to w matematyce zbiór obiektów (nazywanych wektorami) podlegających działaniom skalowania i dodawania Formalnie jest to zatem zbiór z określonymi dwoma działaniami: (+) dodawaniem wektorów, (·) mnożeniem wektora przez *skalar* (element pewnego ciała). Naturalnymi przykładami przestrzeni liniowych są *przestrzenie euklidesowe* (2, 3, i więcej *wymiarowe*, gdzie wymiar utożsamiany jest z liczbą współrzędnych wektora). Wektory w tych przestrzeniach można interpretować jako wektory geometryczne, poprzez intuicję *kierunku*, *zwrotu* oraz *długości*. Takie wektory mogą być sumowane przy użyciu reguły równoległoboku oraz mnożone przez dowolne liczby rzeczywiste (skalary z ciała $\mathbb{R}$ liczb rzeczywistych).

Definicja A.2.1 Trójkę $((V, +), \mathbb{K}, \cdot)$ nazywamy przestrzenią liniową (wektorową) nad ciałem $\mathbb{K}$ z określonym działaniem dodawania wektorów $+ : V \times V \to V$ oraz działaniem mnożenia wektora przez skalar $\cdot : \mathbb{K} \times V \to V$, jeśli spełnione są poniższe warunki:

(L1) Dodawanie wektorów jest *łączne*.

$$\forall\ \mathbf{u}, \mathbf{v}, \mathbf{w} \in V \quad \mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w} \tag{A.6}$$

(L2) Dodawanie wektorów jest *przemienne*.

$$\forall\ \mathbf{v}, \mathbf{w} \in V \quad \mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v} \tag{A.7}$$

(L3) Istnieje *element neutralny* (nazywany *wektorem zerowym*) ze względu na dodawanie wektorów.

$$\exists\ \mathbf{0} \in V \quad \forall\ \mathbf{v} \in V \quad \mathbf{v} + \mathbf{0} = \mathbf{v} \tag{A.8}$$

(L4) Istnieją *elementy odwrotne* (nazywane *wektorami przeciwnymi*) ze względu na dodawanie wektorów.

$$\forall\ \mathbf{v} \in V \quad \exists\ -\mathbf{v} \in V \quad -\mathbf{v} + \mathbf{v} = \mathbf{0} \tag{A.9}$$

(L5) Dodawanie jest rozdzielne względem mnożenia przez skalar.

$$\forall\ a \in \mathbb{K} \quad \forall\ \mathbf{v}, \mathbf{w} \in V \quad a(\mathbf{v} + \mathbf{w}) = a\mathbf{v} + a\mathbf{w} \tag{A.10}$$

(L6) Dodawanie skalarów jest rozdzielne względem mnożenia przez wektor.

$$\forall\ a, b \in \mathbb{K} \quad \forall\ \mathbf{v} \in V \quad (a+b)\mathbf{v} = a\mathbf{v} + b\mathbf{v} \tag{A.11}$$

(L7) Mnożenie przez skalar jest zgodne z mnożeniem skalarnym.

$$\forall\ a, b \in \mathbb{K} \quad \forall\ \mathbf{v} \in V \quad a(b\mathbf{v}) = (a \cdot b)\mathbf{v} \tag{A.12}$$

(L8) Istnieje element neutralny ze względu na mnożenie przez skalar.

$$\exists\ \mathbf{1} \in \mathbb{K} \quad \forall\ \mathbf{v} \in V \quad \mathbf{1}\mathbf{v} = \mathbf{v} \tag{A.13}$$

Definicja A.2.2 *Układem wektorów* nazywamy każdy ciąg skończony $\mathcal{A} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ wektorów z przestrzeni liniowej V nad ciałem $\mathbb{K}$. Każde wyrażenie (wektor) postaci

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \ldots a_n\mathbf{v}_n$$

gdzie $a_1, a_2, \ldots, a_n \in \mathbb{K}$ nazywamy *kombinacją liniową* układu wektorów $\mathcal{A}$. Zbiór wszystkich kombinacji liniowych układu wektorów $\mathcal{A}$ nazywamy *podprzestrzenią generowaną* przez układ wektorów $\mathcal{A}$ i oznaczamy $\mathcal{L}(\mathcal{A})$.

Definicja A.2.3 Mówimy, że układ wektorów $\mathcal{A} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ jest l*iniowo niezależny*, jeśli z równości

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \ldots a_n\mathbf{v}_n = \mathbf{0} \tag{A.14}$$

wynika, że $a_1 = a_2 = \ldots = a_n = 0$. Jeśli układ $\mathcal{A}$ nie jest liniowo niezależny to mówimy, że układ $\mathcal{A}$ jest *liniowo zależny*.

Liniową niezależność wektorów należy intuicyjnie utożsamiać z różnicą kierunku. W przypadku przestrzeni euklidesowej układ wektorów jest liniowo niezależny jeżeli każdy z wektorów posiada inny kierunek.

Definicja A.2.4 Niech V będzie przestrzenią wektorową nad ciałem $\mathbb{K}$. Układ $\mathcal{B} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ wektorów z przestrzenie V nazywamy bazą przestrzeni V jeżeli

B1 jest liniowo niezależny

B2 $\mathcal{L}(\mathcal{B}) = V$ - układ $\mathcal{B}$ generuje całą przestrzeń V.

Z definicji (def. A.2.4) wynika, że dla dowolnego wektora $\mathbf{v}$ istnieje taka kombinacja liniowa układu wektorów z bazy $\mathcal{B}$ przestrzeni V, że

$$\mathbf{v} = x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \ldots x_n\mathbf{v}_n$$

Skalary $x_1, x_2, \ldots, x_n \in \mathbb{K}$ nazywamy *współrzędnymi* wektora $\mathbf{v}$ w bazie $\mathcal{B}$ przestrzenie wektorowej V. Baza przestrzeni V wektorowej jest zarazem maksymalnym liniowo niezależnym układem wektorów z przestrzeni V oraz minimalnym układem generującym całą przestrzeń V. Szczególnym przykładem bazy jest baza *kanoniczna* oznaczana najczęściej $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$.

$$\mathbf{E} = \left(\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}^T, \quad \cdots, \quad \mathbf{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}^T \right) \tag{A.15}$$

Definicja A.2.5 Przestrzeń wektorową V posiadającą *skończoną* bazę $\mathcal{B} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ nazywamy *przestrzenią skończenie wymiarową*. Liczbę n elementów bazy $\mathcal{B}$ nazywamy *wymiarem przestrzeni* V oraz oznaczamy $\dim V = n$.

Powyższa definicja odpowiada intuicji pojęcia wymiaru jako liczby współrzędnych niezbędnych do opisania punktu / wektora / elementu.

Definicja A.2.6 Niech będzie dana n-wymiarowa przestrzeń euklidesowa $\mathbb{R}^n$ liczb rzeczywistych. Dla dowolnych wektorów kolumnowych (traktowanych jako kolumna macierzy) $\mathbf{x}^T = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, $\mathbf{y}^T = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ definiujemy działanie

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y} = \sum_{i=1n} x_i y_i$$

i nazywamy *iloczynem skalarnym*.

Iloczyn skalarny przyporządkowuje dwóm dowolnym wektorom liczbę rzeczywistą (skalar). Zauważmy, że tak zdefiniowany iloczyn skalarny reprezentuje kombinację liniową.

Definicja A.2.7 *Normą* (*długością*) wektora $\mathbf{x}^T \in \mathbb{R}^n$ nazywamy skalar

$$||\mathbf{x}|| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\mathbf{x}^T\mathbf{x}}$$

Jeżeli $||\mathbf{x}|| = 1$ to mówimy, że $\mathbf{x}$ jest wektorem *jednostkowym*.

Definicja A.2.8 Iloczyn skalarny $\mathbf{a}\cdot\mathbf{x}$ wektorów $\mathbf{a}^T = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$, $\mathbf{x}^T = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ nazywamy standaryzowaną kombinacją liniową jeżeli wektor $\mathbf{a}$ jest jednostkowy ($||\mathbf{a}|| = 1$).

Dla uproszczenia zapisów zamiast $\mathbf{x}^t \in \mathbb{R}^n$ pisać będziemy krótko $\mathbf{x} \in \mathbb{R}^n$.

A.3. Wartości własne i wektory własne

Definicja A.3.1 Niech V i V' będą przestrzeniami liniowymi nad tym samy ciałem $\mathbb{K}$. Odwzorowanie $F : V \to V'$ nazywa się *odwzorowaniem liniowym* (lub *operatorem liniowym*), jeśli jest *addytywne* i *jednorodne*

(A) $F(\mathbf{u} + \mathbf{v}) = F(\mathbf{u}) + F(\mathbf{v})$

(J) $F(a\mathbf{v}) = aF(\mathbf{v})$

dla dowolnych $\mathbf{u}, \mathbf{v} \in V$, $a \in \mathbb{K}$. Często zamiast $F(\mathbf{v})$ pisze się krotko $F\mathbf{v}$.

Przykład A.3.1 Naturalnymi przykładami operatorów liniowych w przestrzeniach euklidesowych są odwzorowania postaci $y = a\mathbf{x}$.

Przykład A.3.2 Niech $\mathbf{A} \in \mathbb{K}^n_m$ będzie macierzą typu $m \times n$. Odwzorowanie

$$L_{\mathbf{A}} : \mathbb{K}^1_n \to \mathbb{K}^1_m \tag{A.16}$$

gdzie $L_{\mathbf{A}}(\mathbf{X}) = \mathbf{AX}$ dla $\mathbf{X} \in \mathbb{K}^1_n$, jest operatorem liniowym.

Definicja A.3.2 *Wartością własną* operatora liniowego $F : V \to V$ w przestrzeni liniowej V nad ciałem $\mathbb{K}$ nazywamy taki skalar $\lambda \in \mathbb{K}$, że istnieje pewien niezerowy wektor $\mathbf{v} \in V$ spełniający warunek:

$$F\mathbf{v} = \lambda\mathbf{v} \tag{A.17}$$

Dla danej wartości własnej λ operatora liniowego F wektory $\mathbf{v} \in V$ spełniające warunek (A.17) nazywamy *wektorami własnymi* odpowiadającymi wartości własnej λ.

Zbiór wektorów własnych odpowiadających wartości własnej λ operatora liniowego $F : V \to V$ jest podprzestrzenią przestrzeni liniowej V nazywaną *podprzestrzenią własną*, a jej wymiar nazywany jest *wielokrotnością wartości własnej* λ.

Jeżeli przestrzeń liniowa V nad ciałem $\mathbb{K}$ jest skończenie wymiarowa to każdy operator liniowy $F : V \to V$ można przestawić jako macierz w zależności od ustalonej bazy $\mathcal{B}$ przestrzeni V.

Definicja A.3.3 Niech V będzie przestrzenią liniową nad ciałem $\mathbb{K}$ z bazą $\mathcal{B} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$. *Macierzą odwzorowania liniowego* $F : V \to V$ nazywamy taką macierz $\mathbf{A} \in \mathbb{K}_n^n$, której j-ta kolumna składa się ze współrzędnych wektora $F\mathbf{v}_j$ w bazie $\mathcal{B}$. Dokładniej, jeśli

$$\begin{aligned} F\mathbf{v}_1 &= a_{11}\mathbf{v}_1 + a_{21}\mathbf{v}_2 + \ldots a_{n1}\mathbf{v}_n \\ F\mathbf{v}_2 &= a_{12}\mathbf{v}_1 + a_{22}\mathbf{v}_2 + \ldots a_{n2}\mathbf{v}_n \\ &\vdots \\ F\mathbf{v}_n &= a_{1n}\mathbf{v}_1 + a_{2n}\mathbf{v}_2 + \ldots a_{nn}\mathbf{v}_n \end{aligned}$$

to macierz $\mathbf{A}$ odwzorowania F przyjmuje postać

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Definicja A.3.4 *Wartością własną macierzy* $\mathbf{A} \in \mathbb{K}_n^n$ nazywamy każdy pierwiastek *wielomianu charakterystycznego* $w_{\mathbf{A}}(\lambda)$ macierzy $\mathbf{A}$.

$$w_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) \tag{A.18}$$

Widmem (lub *spektrum*) macierzy $\mathbf{A}$ nazywamy zbiór jej wszystkich wartości własnych i oznaczamy $\sigma(\mathbf{A})$

$$\sigma(\mathbf{A}) = \{\lambda \in \mathbb{K} \ : \ \det(\mathbf{A} - \lambda\mathbf{I}) = 0\} \tag{A.19}$$

Wartość własna macierzy jest w istocie wartością własną operatora liniowego podanego w przykładzie A.3.2.

Dodatek B

Statystyka matematyczna

Poniższe definicje oraz własności, podane w oparciu o [9] [12], służą uzupełnieniu właściwej części pracy. Są to jedynie kluczowe pojęcia, bardziej zainteresowanego czytelnika odsyłamy do wyżej wskazanej lektury.

B.1. Miary położenia, rozproszenia, współzależności

Niech $\mathbf{E}X$ oznacza wartość oczekiwaną zmiennej losowej X ([9], str. 72-79)

Definicja B.1.1 *Wariancją* zmiennej losowej X nazywamy wielkość

$$\mathrm{Var}(X) = \mathbf{E}\big[(X - \mathbf{E}X)^2\big]$$

Wartość oczekiwana to miara położenia zmiennej losowej. Wariancja stanowi miarą rozproszenia (rozrzutu) zmiennej losowej.

Definicja B.1.2 *Kowariancją* zmiennych losowych X i Y nazywamy wielkość

$$\mathrm{Cov}(X, Y) = \mathbf{E}\big[(X - \mathbf{E}X)(Y - \mathbf{E}Y)\big]$$

Kowariancja zmiennych losowych mierzy siłę zależności liniowej pomiędzy zmiennymi losowymi.

Definicja B.1.3 *Współczynnikiem korelacji* między zmiennymi losowymi X i Y nazywamy wielkość

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Współczynnik korelacji jest rodzajem unormowanej miary siły zależności liniowej między zmiennymi.

Definicja B.1.4 Macierz $\mathbf{\Sigma}$, która dla wektora losowego $(X_1, X_2, \ldots, X_n)$ przyjmuje postać

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n \end{bmatrix} \tag{B.1}$$

gdzie

$$\sigma_i = \mathrm{Var}(X_i)$$

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$$

nazywamy *macierzą kowariancji* wektora losowego $(X_1, X_2, \ldots, X_n)$.

Macierz kowariancji jest uogólnieniem pojęcia wariancji na przypadek wielowymiarowy. Macierz kowariancji jest macierzą symetryczną, czyli macierzą o rzeczywistych wartościach własnych A.3.4.

Definicja B.1.5 Macierz ρ, która dla wektora losowego $(X_1, X_2, \ldots, X_n)$ przyjmuje postać

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{bmatrix} \tag{B.2}$$

gdzie

$$\rho_{ij} \quad - \text{współczynnik korelacji zmiennych} X_i, X_j$$

nazywamy *macierzą korelacji* wektora losowego $(X_1, X_2, \ldots, X_n)$.

W praktyce wykorzystuje się próbkowe odpowiedniki wyżej podanych miar. Niech $x_1, x_2, \ldots, x_n$ będzie próbą o liczności n.

Definicja B.1.6 *Wartością średnią* w próbie $x_1, x_2, \ldots, x_n$ nazywamy średnią arytmetyczną

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Definicja B.1.7 *Wariancją w próbie* $x_1, x_2, \ldots, x_n$ nazywamy wielkość

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Pierwiastek wariancji nazywany jest *odchyleniem standardowym w próbie*

$$s = \sqrt{s^2}$$

Dzielenie przez $n-1$ związane jest z usunięciem obciążenia estymatora, które występuje w przypadku dzielenia przez n.

Definicja B.1.8 *Próbkowym współczynnikiem korelacji* nazywamy wielkość

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

gdzie s_x oznacza wariancję z próby $x_1, x_2, \ldots, x_n$, s_y oznacza wariancję z próby $y_1, y_2, \ldots, y_n$.

Bibliografia

[1] Harold Hotelling [online]. wikipedia : wolna encyklopedia, 2008-11-27 01:34Z. [dostęp: 2009-04-19 18:10Z]. Dostępny w Internecie: `http://pl.wikipedia.org/wiki/Fullpagename?oldid=14973526`.

[2] Analiza dyskryminacyjna [online]. wikipedia : wolna encyklopedia, 2009-03-26 08:24Z. [dostęp: 2009-09-16 17:39Z]. Dostępny w Internecie: http://pl.wikipedia.org/wiki/Fullpagename?oldid=16473839.

[3] Skalowanie wielowymiarowe [online]. wikipedia : wolna encyklopedia, 2009-04-3 14:34Z. [dostęp: 2009-09-16 17:42Z]. Dostępny w Internecie: http://pl.wikipedia.org/wiki/Fullpagename?oldid=16567865.

[4] Karl Pearson [online]. wikipedia : wolna encyklopedia, 2009-04-4 21:55Z. [dostęp: 2009-04-19 17:58Z]. Dostępny w Internecie: `http://pl.wikipedia.org/wiki/Fullpagename?oldid=16584170`.

[5] Eksploracja danych [online]. wikipedia : wolna encyklopedia, 2009-07-22 11:56Z. [dostęp: 2009-09-16 17:30Z]. Dostępny w Internecie: http://pl.wikipedia.org/wiki/Fullpagename?oldid=17870884.

[6] R (język programowania) [online]. wikipedia : wolna encyklopedia, 2009-08-17 18:33Z. [dostęp: 2009-09-16 17:48Z]. Dostępny w Internecie: http://pl.wikipedia.org/wiki/Fullpagename?oldid=18162704.

[7] Baza danych [online]. wikipedia : wolna encyklopedia, 2009-09-11 11:13Z. [dostęp: 2009-09-16 17:23Z]. Dostępny w Internecie: http://pl.wikipedia.org/wiki/Fullpagename?oldid=18443530.

[8] A. Asuncion, D.J. Newman. UCI machine learning repository, 2007.

[9] Patrick Billingsley. *Prawdopodobieństwo i miara*. Państwowe Wydawnictwo Naukowe, Warszawa, wydanie 1, 1987.

[10] Leo Breiman, Jerome H. Friedman, Richard A. Olshen. *Classification and Regression Trees*. Chapman & Hall, New York, NY, wydanie 1, 1984.

[11] Paweł Cichosz. *Systemu uczące się*. Wydawnictwa Naukowo-Techniczne, Warszawa, wydanie 1, 2000.

[12] Jacek Koronacki, Jan Mielniczuk. *Statystyka dla studentów kierunków technicznych i przyrodniczych*. Wydawnictwo Naukowo-techniczne, Warszawa, wydanie 2, 2004.

[13] Jacek Koronacki, Jan Ćwik. *Statystyczne systemy uczące się*. Wydawnictwo Naukowo-techniczne, Warszawa, wydanie 1, 2005.

[14] Mirosław Krzyśko, Waldemar Wołyński, Tomasz Górecki, Michał Skorzybut. *Systemy uczące się*. Wydawnictwo Naukowo-techniczne, Warszawa, wydanie 1, 2008.

[15] Daniel T. Larose. *Odkrywanie wiedzy z danych*. Wydawnictwo Naukowe PWN, Warszawa, wydanie 1, 2006.

[16] Daniel T. Larose. *Metody i modele eksploracji danych*. Wydawnictwo Naukowe PWN, Warszawa, wydanie 1, 2008.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[18] Dariusz Szymański. Podstawy statystyki 2. Metody Statystyczne w Biznesie - studia podyplomowe, materiały do wykładów, rok akademicki 2008/2009.

[19] Piotr Wójcik. Statystyka w badaniach rynkowych. Metody Statystyczne w Biznesie - studia podyplomowe, materiały do wykładów, rok akademicki 2008/2009.

[20] Tadeusz Świrszcz. *Algebra liniowa z geometrią analityczną*. Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, wydanie 2, 1996.