

Analiza Głównych Składowych

Principal Component Analysis (PCA)

Paweł Gliwny

Uniwersytet Łódzki
Wydział Fizyki i Informatyki Stosowanej

Eksploracja Danych

Problem wyjściowy

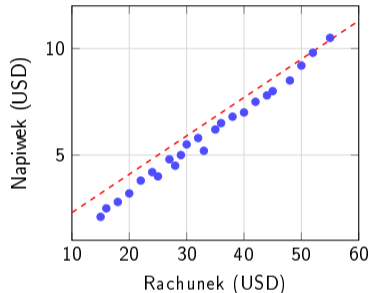
Obserwacja:

- Zmienne często **współzienne** (kowariancja)
- Część zmienności jednej zmiennej jest powielana przez inną
- Informacja jest **redundantna**

Przykład:

Rachunek w restauracji \leftrightarrow Napiwek

Gdy rachunek rośnie, napiwek też rośnie!



Korelacja dodatnia:

- Wzrost \leftrightarrow Waga osoby
- Powierzchnia mieszkania \leftrightarrow Cena
- Temperatura \leftrightarrow Sprzedaż lodów
- Wykształcenie \leftrightarrow Zarobki

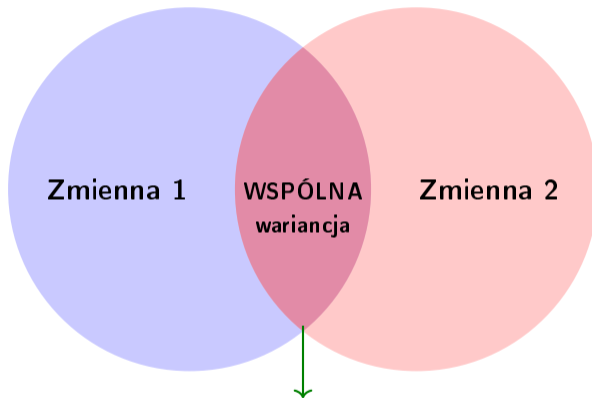
Korelacja ujemna:

- Przebieg auta \leftrightarrow Cena auta
- Temperatura \leftrightarrow Sprzedaż kurtek
- Wiek sprzętu \leftrightarrow Wydajność

Wniosek

Skoro zmienne „mówią to samo” \rightarrow można je skompresować!

Idea redundancji informacji



PCA eliminuje redundancję
i tworzy **niezależne składowe**

Czym jest PCA?

Definicja

PCA (Principal Component Analysis) – technika redukcji wymiarowości, która przekształca zbiór skorelowanych zmiennych w zbiór **nieskorelowanych** zmiennych zwanych **głównymi składowymi**.

Kluczowe cechy:

- Łączy wiele zmiennych w mniejszy zestaw **głównych składowych**
- Główne składowe = **ważone kombinacje liniowe** oryginalnych zmiennych
- Zachowuje jak największą część **wariancji** danych
- Działa **tylko dla zmiennych numerycznych**

Cel

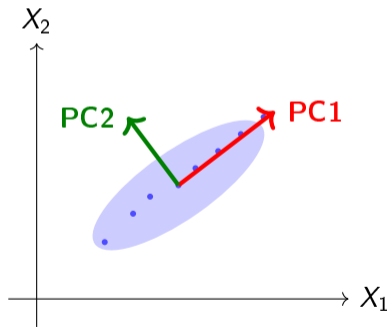
Redukcja wymiarowości przy minimalnej utracie informacji

PCA: intuicja geometryczna

Główne składowe to kombinacje liniowe zmiennych oryginalnych:

$$Z_1 = w_{1,1} \cdot X_1 + w_{1,2} \cdot X_2, \quad Z_2 = w_{2,1} \cdot X_1 + w_{2,2} \cdot X_2$$

Wagi $w_{i,j}$ = **ładunki (loadings)** – określają wkład zmiennych w składowe.



PC1 – kierunek max. wariancji; **PC2** – ortogonalny, wyjaśnia resztę.

Dane: $X \in \mathbb{R}^{n \times p}$ (n próbek, p cech), wycentrowane (odjęta średnia).

Macierz kowariancji:

$$C = \frac{1}{n-1} X^T X = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Rozkład na wartości własne: $Cv_i = \lambda_i v_i$

- λ_i – wartość własna = wariancja wzdłuż i -tej składowej
- v_i – wektor własny = kierunek i -tej głównej składowej

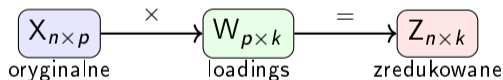
PCA: transformacja danych

Macierz transformacji (loadings) – k wybranych wektorów własnych:

$$W = [v_1 \ v_2 \ \cdots \ v_k] \in \mathbb{R}^{p \times k}$$

Projekcja danych (scores):

$$Z = XW \in \mathbb{R}^{n \times k}$$



Redukcja wymiarowości: p cech $\rightarrow k$ składowych (gdzie $k \ll p$).

Wyjaśniona wariancja

Procent wyjaśnionej wariancji przez i -tą składową:

$$\text{Var}_i = \frac{\lambda_i}{\sum_{j=1}^p \lambda_j} \times 100\%$$

Skumulowana wariancja (pierwsze k składowych):

$$\text{Var}_{1:k} = \frac{\sum_{i=1}^k \lambda_i}{\sum_{j=1}^p \lambda_j} \times 100\%$$

Przykład (zbiór IRIS)

	PC1	PC2	PC3	PC4
Wariancja (%)	72.96	22.85	3.67	0.52
Skumulowana (%)	72.96	95.81	99.48	100

⇒ PC1 + PC2 wyjaśniają **95.8%** wariancji!

Dlaczego standaryzacja jest kluczowa?

Problem

PCA jest wrażliwe na **skalę** zmiennych!
Zmienne o większych wartościach dominują analizę.

Przykład (zbiór Wine):

Cecha	Zakres wartości	Wariancja
Alkohol	11.0 – 14.8	0.66
Prolina	278 – 1680	98609

Bez standaryzacji: **Prolina zdominuje** całą analizę!

Standaryzacja (z-score)

Wzór standaryzacji

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

gdzie:

- μ_j – średnia j -tej cechy
- σ_j – odchylenie standardowe j -tej cechy

Po standaryzacji:

- Każda cecha ma $\mu = 0$ i $\sigma = 1$
- Wszystkie cechy mają **równy wpływ** na PCA
- Macierz kowariancji C = macierz korelacji R

W Pythonie

```
from sklearn.preprocessing import StandardScaler  
X_std = StandardScaler().fit_transform(X)
```

Algorytm PCA – krok po kroku

- 1 **Standaryzacja** danych (opcjonalnie, ale zalecane)

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

- 2 Obliczenie **macierzy kowariancji**

$$C = \frac{1}{n-1} \bar{X}^T \bar{X}$$

- 3 **Rozkład według wartości osobliwych** macierzy kowariancji

$$Cv_i = \lambda_i v_i$$

- 4 **Sortowanie** wektorów własnych wg malejących wartości własnych

- 5 **Wybór** k pierwszych składowych

- 6 **Projekcja** danych na nową przestrzeń

$$Z = \bar{X}W$$

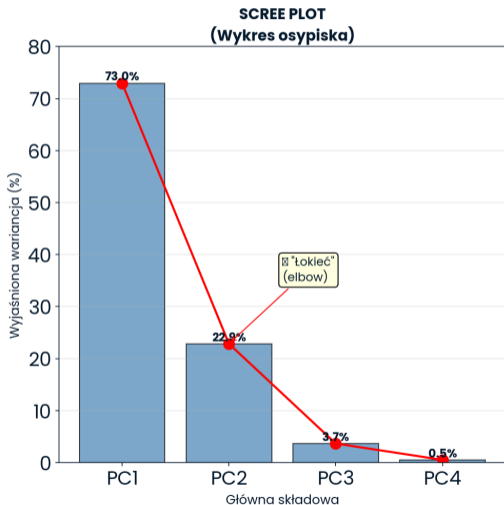
① Próg wyjaśnionej wariancji

- Wybieramy k takie, że $\text{Var}_{1:k} \geq 80\%$ lub 95%
- Obiektywne kryterium

② Metoda łokcia (Scree Plot)

- Szukamy punktu załamania na wykresie
- Subiektywna, ale intuicyjna

Scree Plot (wykres osypiska)



Jak czytać?

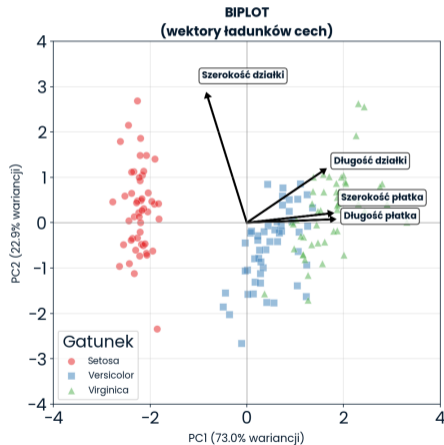
- Oś Y: wartość własna lub % wariancji
- Oś X: numer składowej
- Szukamy łokcia – miejsca załamania

Reguła:

- Zatrzymujemy składowe **przed** łokciem
- Reszta to szum

Nazwa: scree = osypisko skalne

Biplot – interpretacja ładunków



Elementy biplot:

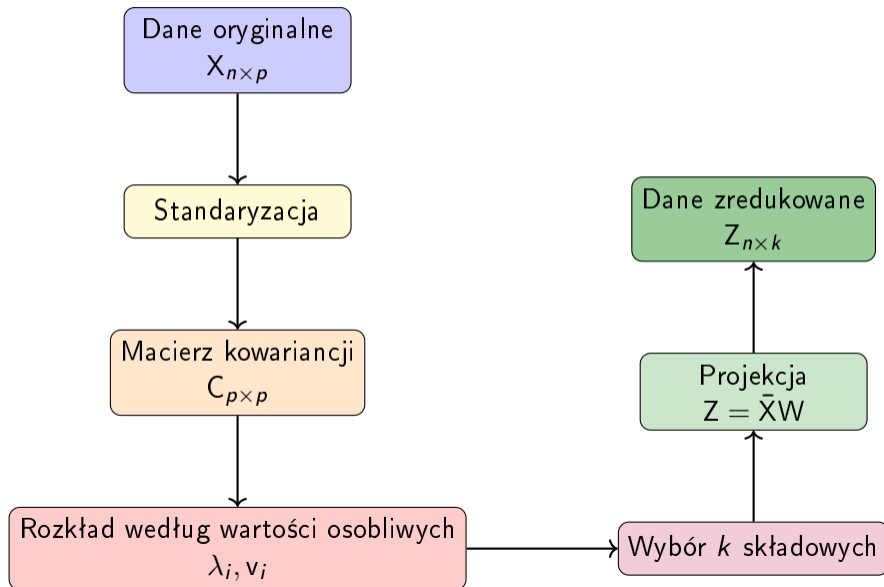
- Punkty – próbki w przestrzeni PC
- Strzałki – wektory ładunków cech

Interpretacja strzałek:

- Długość – siła wpływu cechy
- Kierunek – korelacja z PC
- Kąt między strzałkami – korelacja między cechami

Strzałki w tym samym kierunku \Rightarrow cechy skorelowane

Podsumowanie: schemat przepływu PCA



Kiedy stosować PCA?

Zalety:

- Redukcja wymiarowości
- Usunięcie korelacji między zmiennymi
- Wizualizacja danych wielowymiarowych
- Przyspieszenie uczenia modeli ML
- Redukcja szumu

Ograniczenia:

- Tylko zmienne numeryczne
- Zakłada liniowe zależności
- Wrażliwość na outliers
- Utrata interpretowalności
- Wymaga standaryzacji

Typowe zastosowania

- Eksploracyjna analiza danych (EDA)
- Preprocessing dla ML (redukcja wymiarów)
- Kompresja obrazów
- Analiza danych genetycznych, finansowych, sensorowych

Dziękuję za uwagę!

Materiały:

Przykład + zadania dostępne na Teams