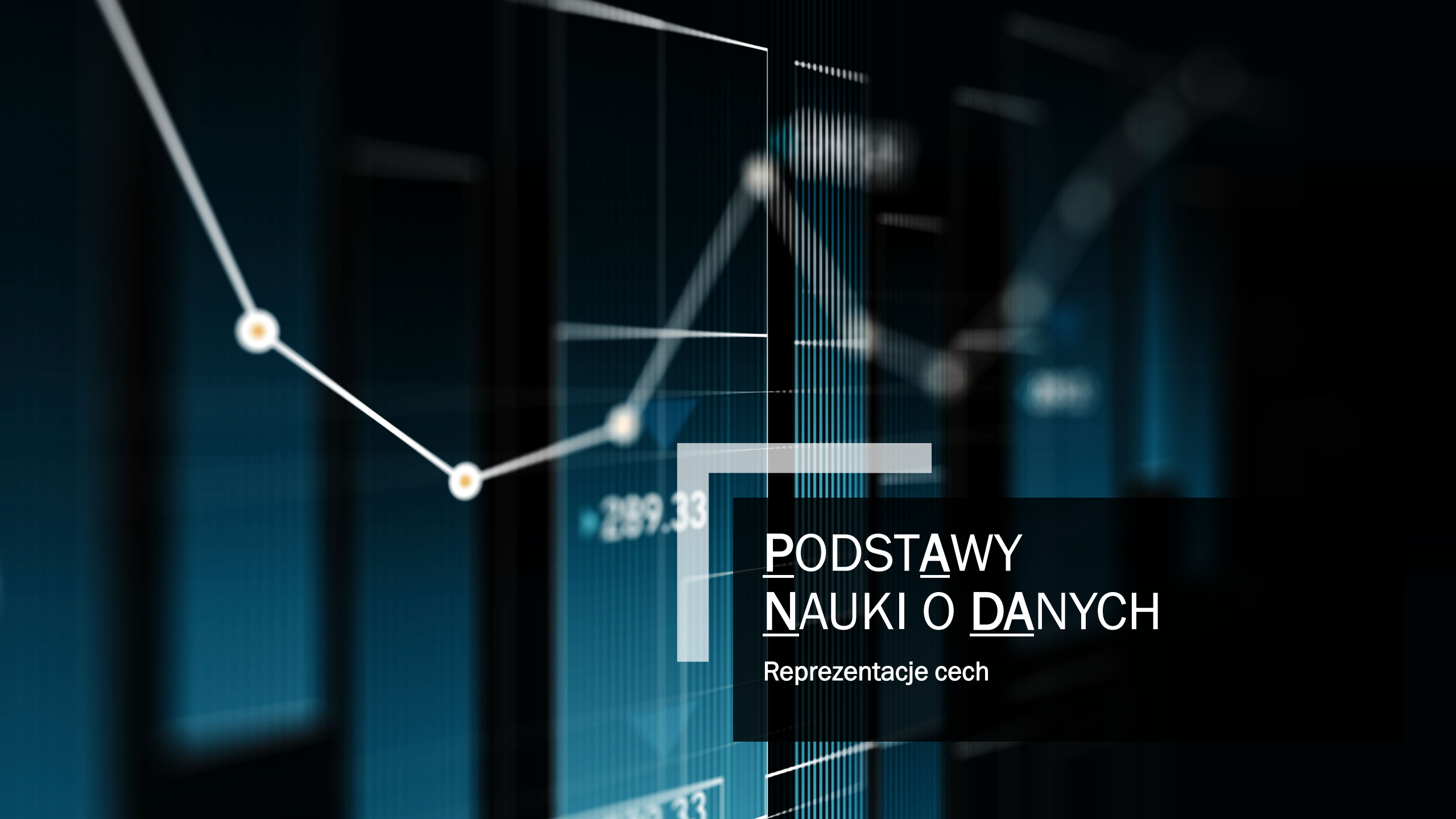
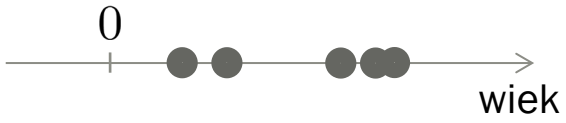


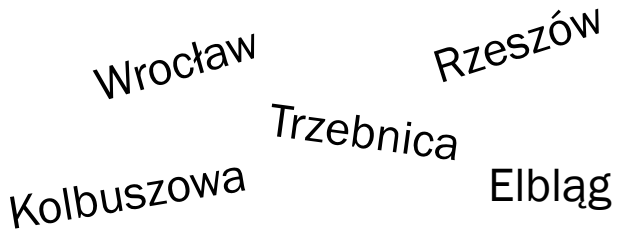
PODSTAWY NAUKI O DANYCH

Reprezentacje cech

PROSTE LICZBOWE REPREZENTACJE

Skale liczbowe





Absolutna: jest 0 i jednostka, np. ilość cukru, liczba subskrybentów
różnica i iloraz mają sens

Ilorazowa: sens ma stosunek wartości, np. temp. w Kelwinach, inflacja

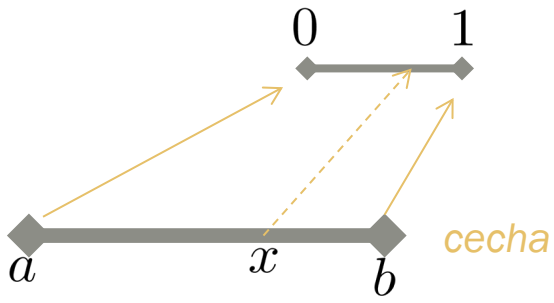
Interwałowa / przedziałowa: 0 jest umowne, np. temp. w C, data

Porządkowa: kolejność ma znaczenie, np. miesiąc, wykształcenie

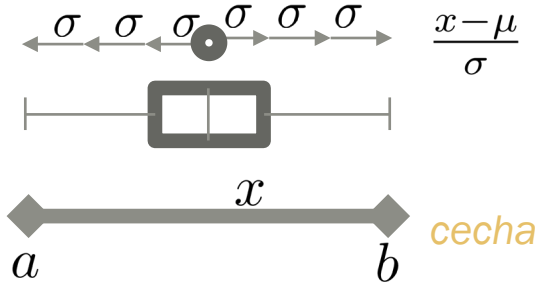
Nominalna: nie da się jednoznacznie uporządkować

Przekształcenia ciągłe $\Re \rightarrow \Re$

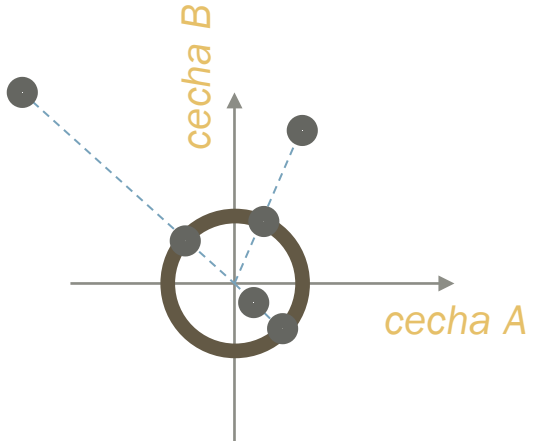
Skalowanie



Standaryzacja

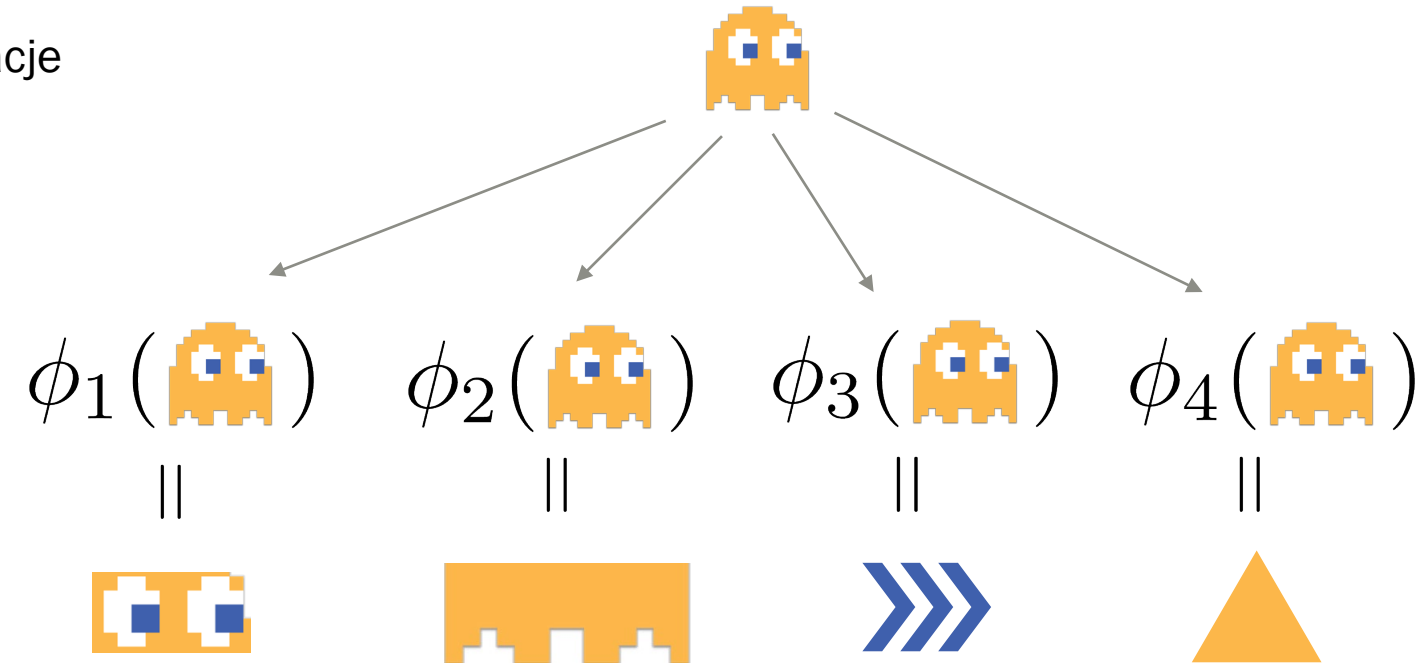


Normalizacja



Tworzenie cech pochodnych

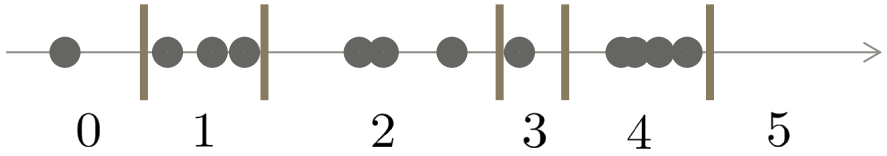
Transformacje



Przykład: cechy wielomianowe, funkcje bazowe

Tworzenie cech pochodnych

Dyskretyzacja



Tworzenie cech pochodnych

Kodowanie cech porządkowych

```
stan pacjenta = {
  'zdrowy': 0,
  'stabilny': 0.5,
  'niestabilny': 1,
  'chory': 4,
  'poważnie chory': 4.5,
  'zagrożenie życia': 8,
}
```

```
glucose level = {
  'Low': 0,
  'Medium': 1,
  'High': 2,
}
```

Tworzenie cech pochodnych

Kodowanie cech kategorialnych / nadawanie etykiet

```
[lekarz, inżynier systemów, dyrektor, inżynier systemów]
```

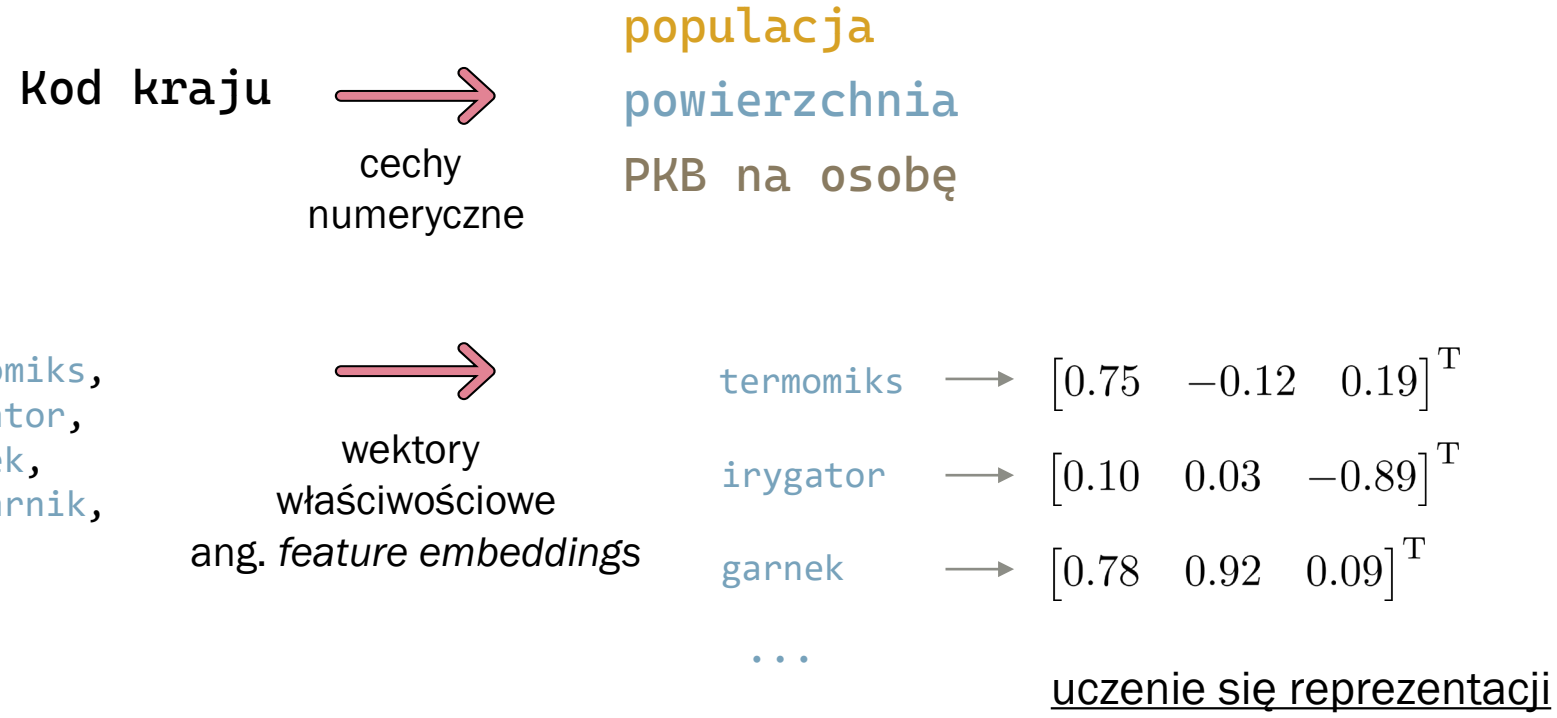
nr	lekarz	inżynier systemów	dyrektor
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

kodowanie gorącojedynkowe (ang. *one-hot encoding*)

Tworzenie cech pochodnych

Kodowanie cech kategorialnych / nadawanie etykiet

Problem kodowania gorącojedynkowego przy wielu kategoriach



EKSTRAKCJA CECH

Dane
po standaryzacji

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_N^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_N^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(K)} & x_2^{(K)} & \dots & x_N^{(K)} \end{bmatrix}$$

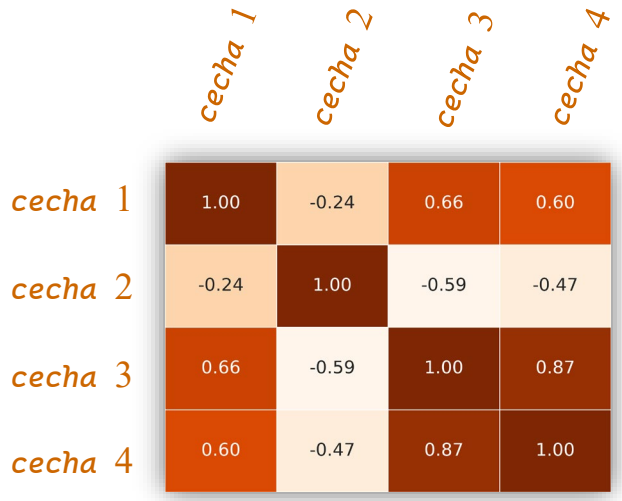
cecha 1
cecha 2
...
cecha K

obserwacja 1 | obserwacja 2 | ... | obserwacja N

Macierz kowariancji

$$\mathrm{Cov}\left[\mathbf{x}\right] = \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_K^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}} \equiv \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}}$$



$\mathrm{Cov}\left[\mathbf{x}\right]$

Pytanie problemowe: *Czy można tak obrócić osie układu współrzędnych, aby **nowe cechy** wyglądały na zdekorelowane?*

$$\mathbf{z}_i \longleftarrow \underset{\text{bazy}}{\text{zmiana}} \longleftarrow \mathbf{x}_i \qquad \text{dla} \quad i = 1, 2, \ldots, N$$

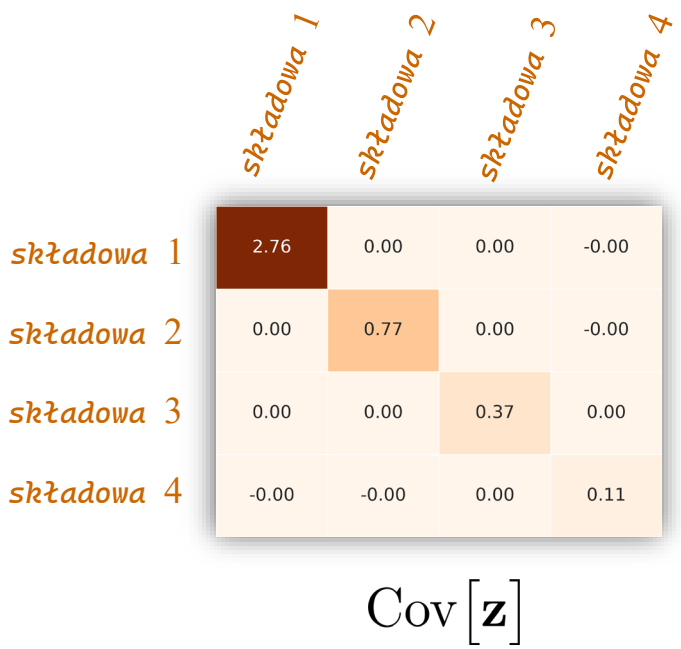
$$\mathbf{Z} \longleftarrow \underset{\text{bazy}}{\text{zmiana}} \longleftarrow \mathbf{X}$$

$$\mathbf{Z} = \mathbf{V}^{\mathrm{T}}\mathbf{X}$$

$\mathbf{V}$ – macierz obrotu

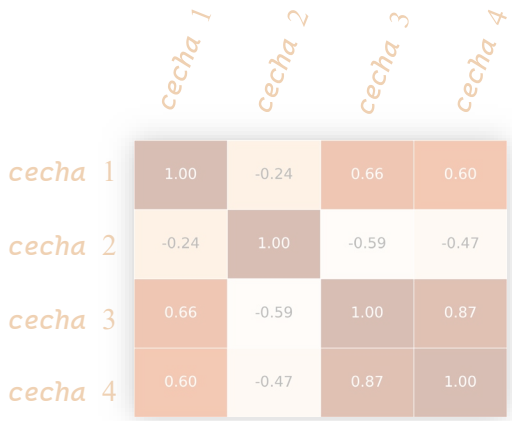
$$\mathbf{V}^{\mathrm{T}} = \mathbf{V}^{-1}$$

składowa główna = nowa cecha (w obróconej bazie)



$\mathrm{Cov}\left[\mathbf{z}\right]$

zmiana bazy ⟵



$\mathrm{Cov}\left[\mathbf{x}\right]$

Rozwiązanie: 1) *Wyznaczyć macierz kowariancji danych po obrocie w zależności od* $\mathbf{V}$

2) *Znaleźć taką macierz obrotu* $\mathbf{V}$*, dla której macierz kowariancji danych jest diagonalna*

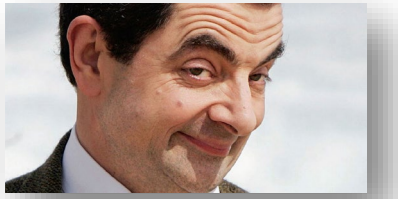
$$\mathrm{Cov}[\mathbf{z}] = \frac{1}{N}\mathbf{Z}\mathbf{Z}^{\mathrm{T}} = \frac{1}{N}\mathbf{V}^{\mathrm{T}}\mathbf{X}\left(\mathbf{V}^{\mathrm{T}}\mathbf{X}\right)^{\mathrm{T}} = \frac{1}{N}\mathbf{V}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{V} = \mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{V}$$

Nowe cechy (składowe główne) mają być **zdekorelowane:**

$$\mathbf{D} \equiv \mathrm{Cov}[\mathbf{z}] = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_K \end{bmatrix}$$

$$\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{V} = \mathbf{D}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{V} = \mathbf{V}\mathbf{D}$$

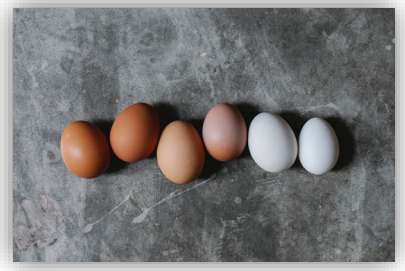


$\mathbf{V}$ – wektory własne macierzy kowariancji $\boldsymbol{\Sigma}_{\mathbf{x}}$

$\mathrm{diag}(\mathbf{D})$ – wartości własne macierzy kowariancji $\boldsymbol{\Sigma}_{\mathbf{x}}$

Analiza składowych głównych

Principal Component Analysis



$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_K \end{bmatrix}$$

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_S = \lambda_{\min}$$

– ciach! –

wariancja rzutowanych danych

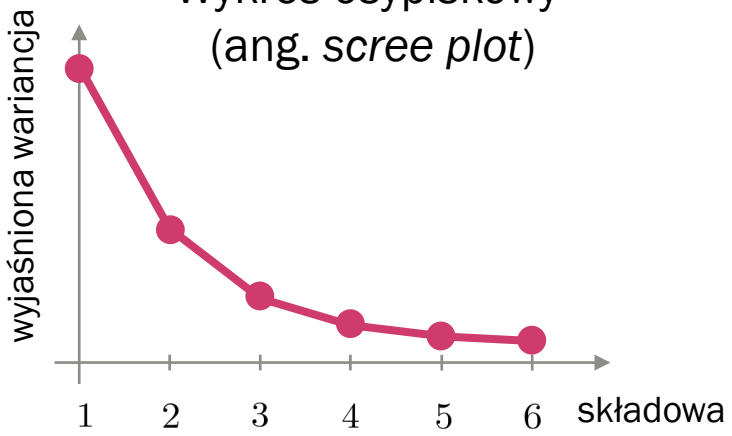
błąd aproksymacji danych

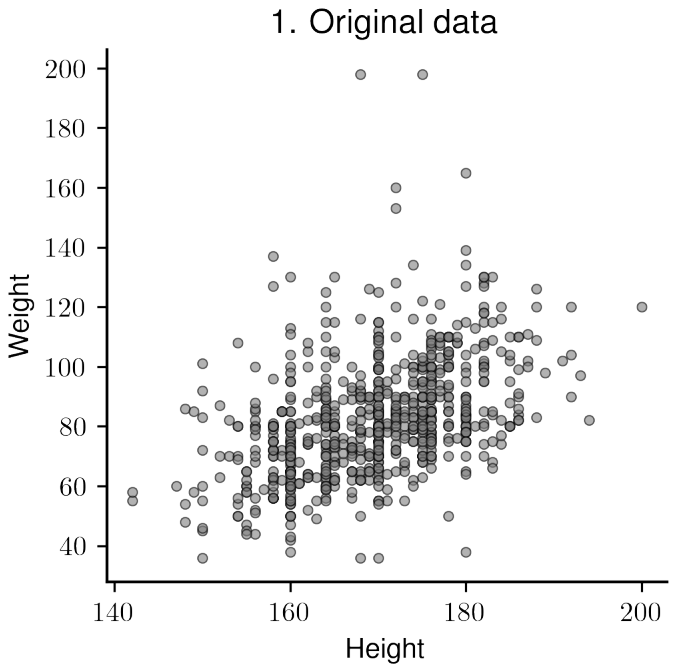
$$\frac{\lambda_1 + \cdots + \lambda_D + \lambda_{D+1} + \cdots + \lambda_K}{\lambda_1 + \lambda_2 + \cdots + \lambda_K}$$

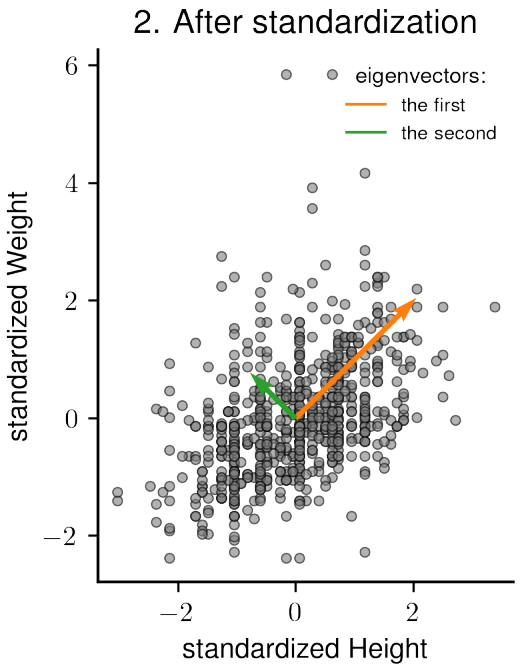
$$\frac{\lambda_1 + \cdots + \lambda_D}{\lambda_1 + \lambda_2 + \cdots + \lambda_K}$$

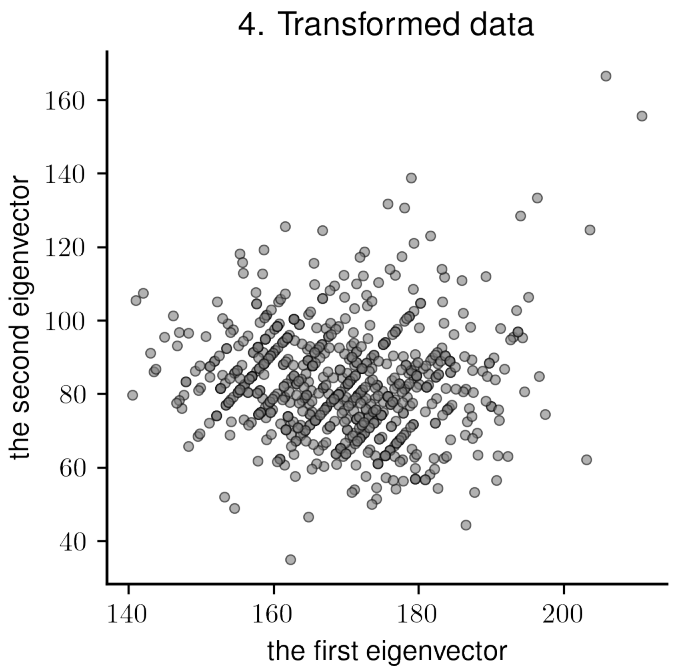
Wyjaśniona wariancja

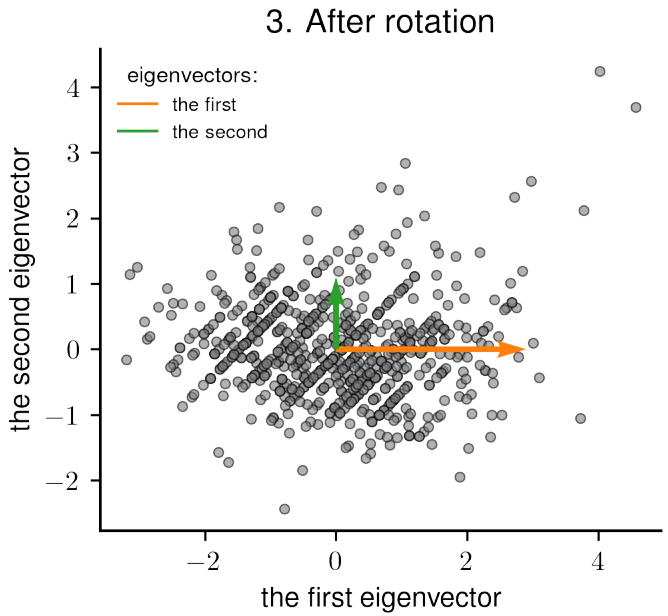
Wykres osypiskowy
(ang. *scree plot*)

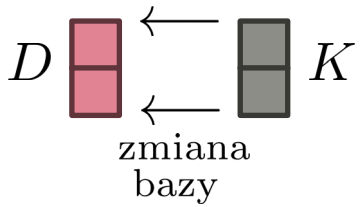


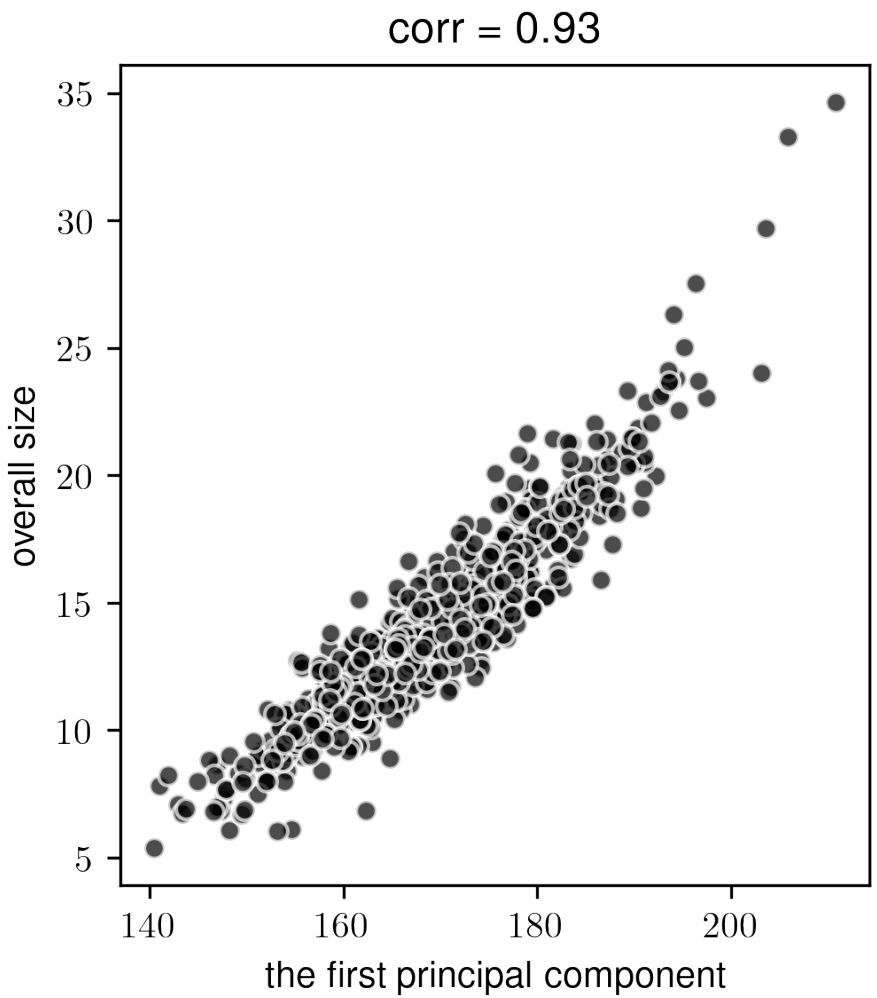

1. Original data
Weight
Height
200
180
160
140
120
100
80
60
40
140
160
180
200

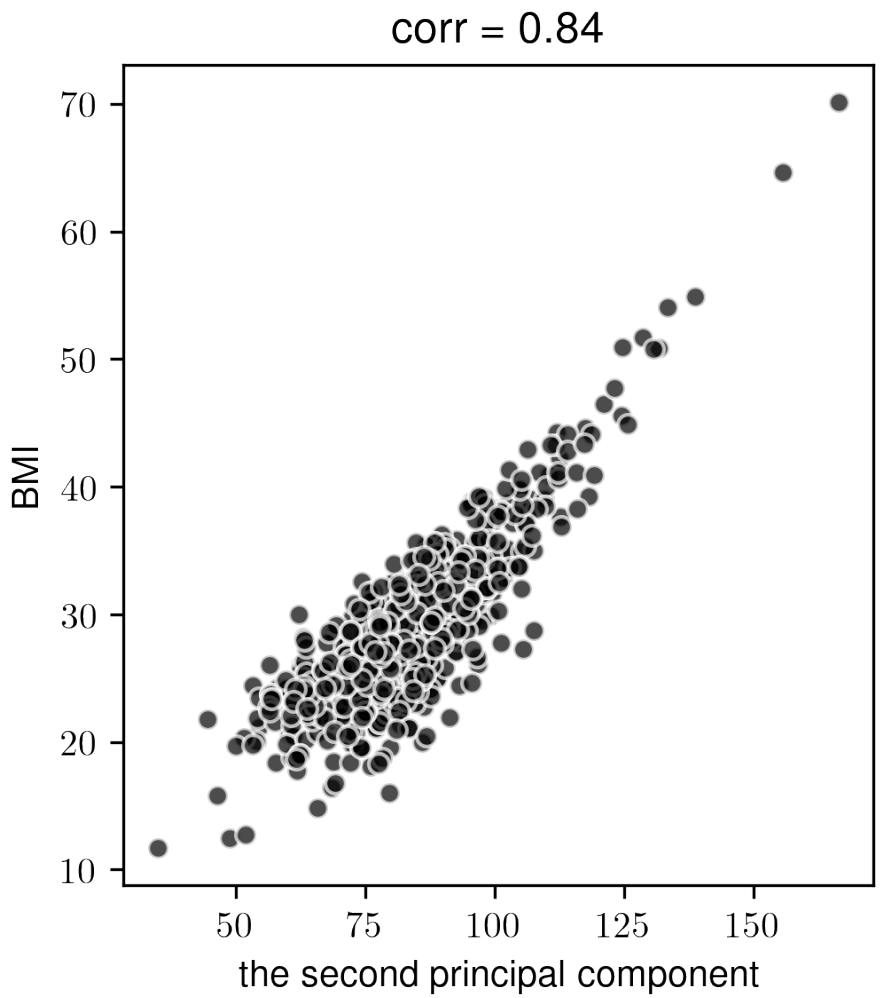

2. After standardization
eigenvectors:
the first
the second
standardized Weight
standardized Height
6
4
2
0
−2
−2
0
2

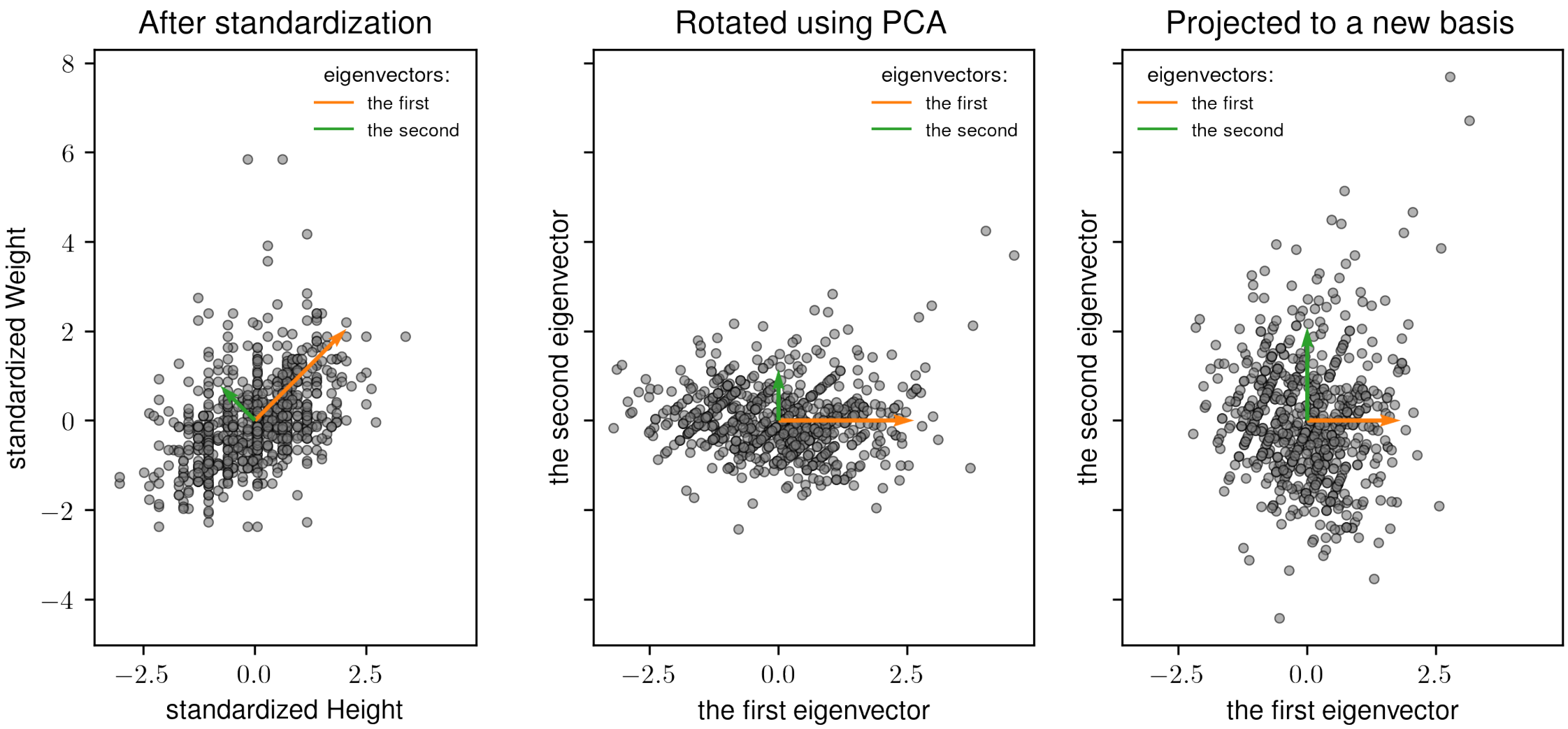

4. Transformed data
the second eigenvector
the first eigenvector
160
140
120
100
80
60
40
140
160
180
200


3. After rotation
eigenvectors:
the first
the second
the second eigenvector
the first eigenvector
4
3
2
1
0
−1
−2
−2
0
2
4

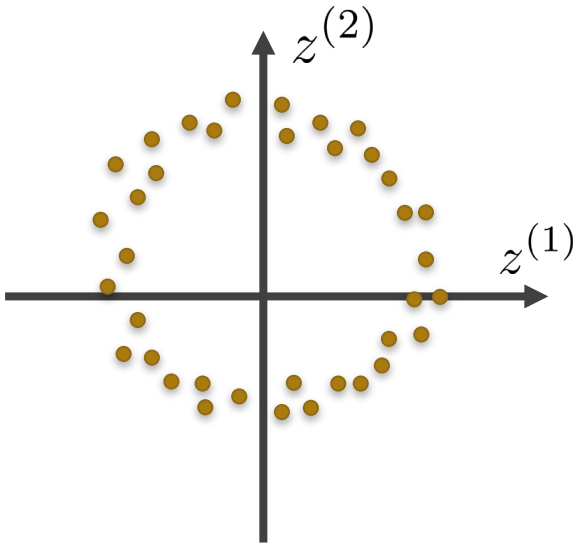

D
K
zmiana
bazy

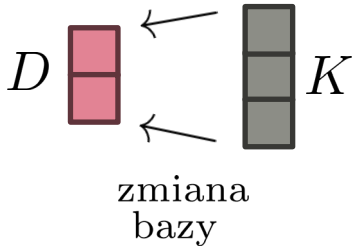

corr = 0.93
overall size
the first principal component
35
30
25
20
15
10
5
140
160
180
200

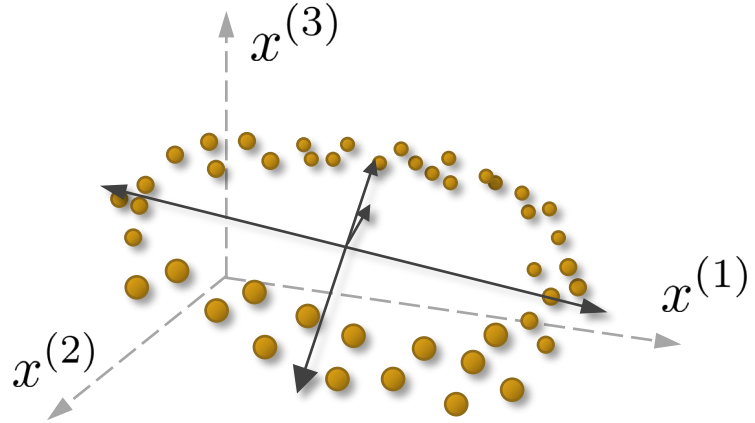

corr = 0.84
BMI
the second principal component
70
60
50
40
30
20
10
50
75
100
125
150


After standardization
Rotated using PCA
Projected to a new basis
eigenvectors:
the first
the second
standardized Weight
standardized Height
the second eigenvector
the first eigenvector
8
6
4
2
0
−2
−4
−2.5
0.0
2.5

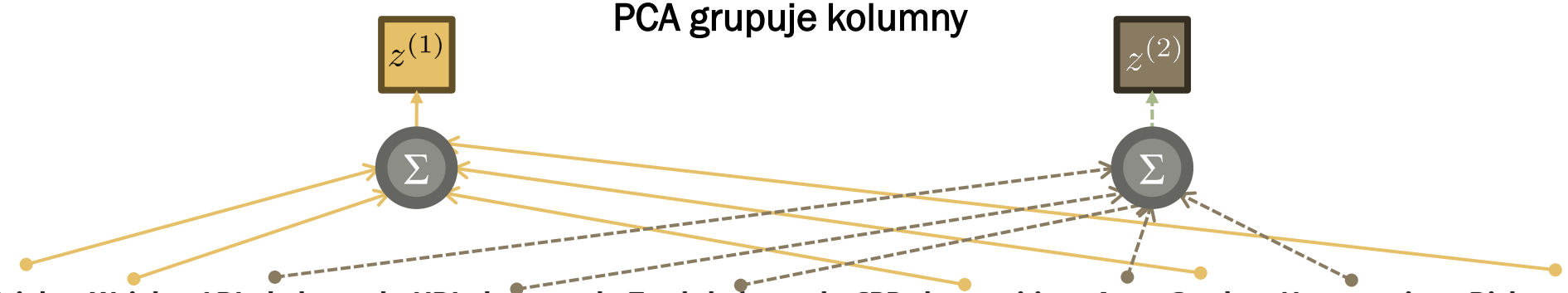
REDUKCJA WYMIAROWOŚCI


$z^{(2)}$
$z^{(1)}$


D
K
zmiana
bazy


$x^{(3)}$
$x^{(1)}$
$x^{(2)}$

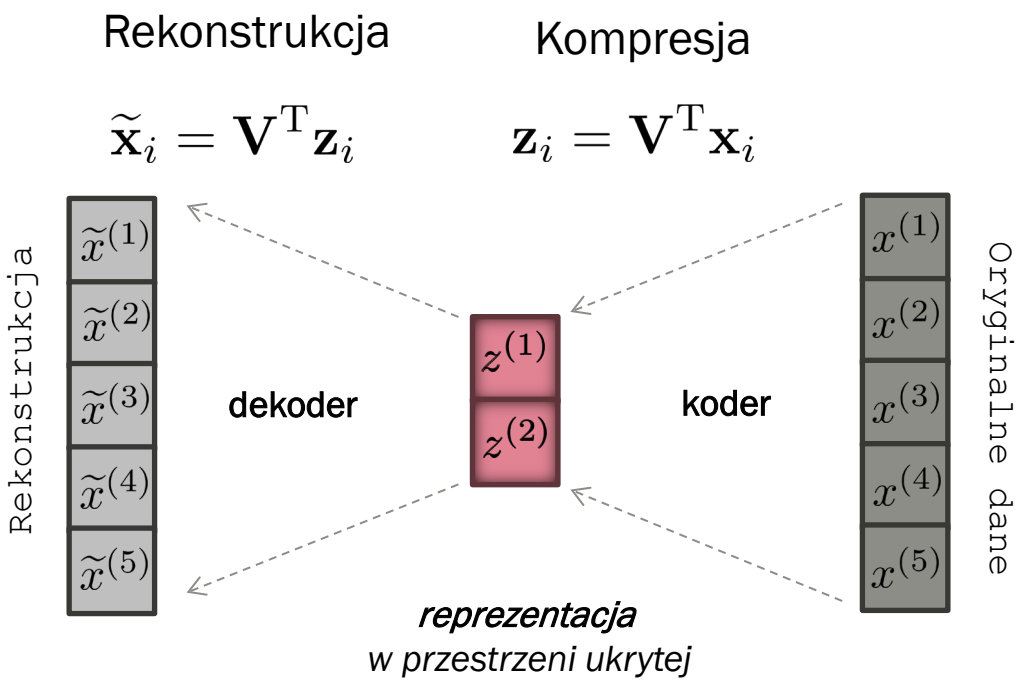
PCA grupuje kolumny



	Height	Weight	LDL cholesterol	HDL cholesterol	Total cholesterol	CRP ultrasensitive	Age	Gender	Hypertension	Diabetes mellitus	
140	182	85	76	47	138	2.14	86	Male	Yes	Yes	
89	177	110	85	45	150	8.04	72	Male	Yes	Yes	
448	164	110	120	37	172	10.98	77	Female	Yes	No	
453	156	70	105	28	162	8.88	89	Female	Yes	Yes	
6	174	100	53	25	96	21.44	81	Male	No	No	
84	185	80	68	36	119	3.77	68	Male	Yes	Yes	
561	170	83	84	34	136	16.42	71	Male	Yes	Yes	
659	170	70	41	33	85	1.17	80	Male	Yes	No	
289	176	85	83	58	155	9.53	53	Male	No	No	
372	160	64	82	53	162	0.92	85	Female	Yes	No	
632	176	107	84	48	150	19.30	56	Male	No	Yes	
132	161	68	25	54	109	1.28	83	Female	Yes	Yes	
379	169	126	8	30	69	6.49	67	Male	Yes	Yes	

Grupowanie grupuje dane

Schemat koder – dekoder



PCA minimalizuje błąd rekonstrukcji

$$\sum_i ||\mathbf{x}_i - \widetilde{\mathbf{x}}_i||_2^2$$

Karhunen-Loève Expansion

- gdy są dostępne etykiety C
- równoważne PCA z podziałem na $\boldsymbol{\Sigma}(C)$ dla każdej klasy C oddzielnie

WEKTORY WŁASNE

Przykład 1

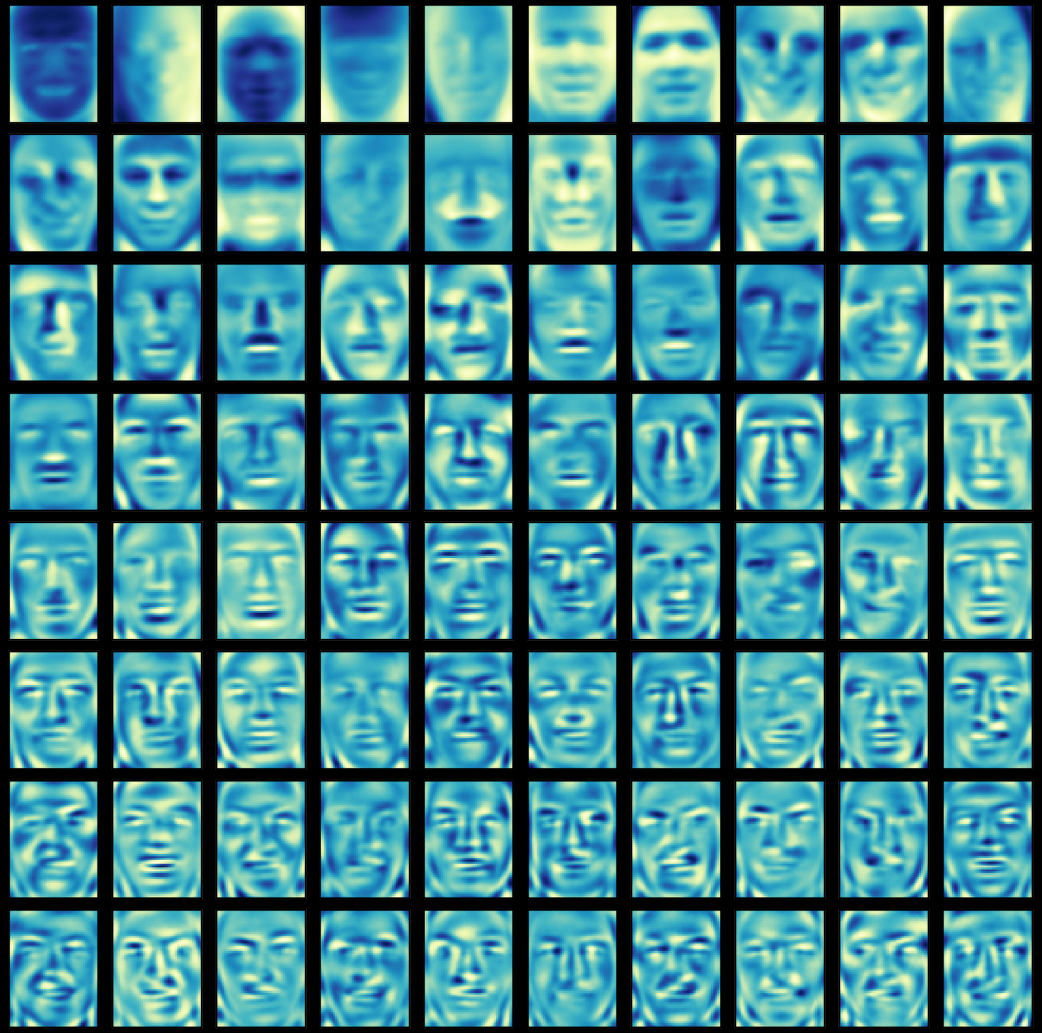
- https://psycnet.apa.org/record/1981-32589-001
- 1634 studentów z Los Angeles
- Każda substancja oceniona w skali 1-5

zażywanie	punktacja
nigdy	1
tylko raz	2
kilka razy	3
wiele razy	4
regularnie	5

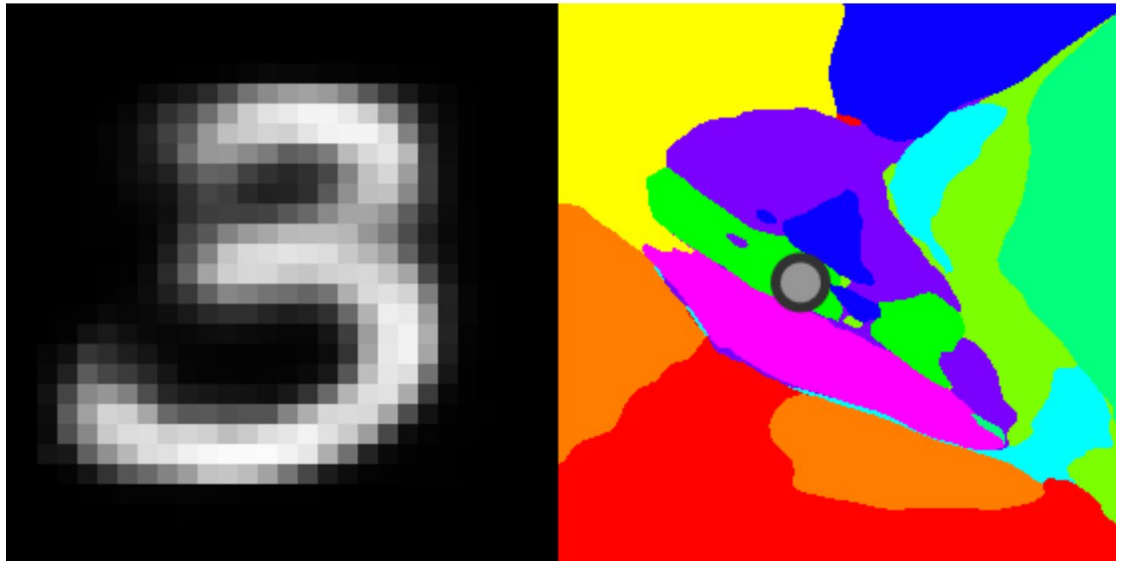
substancja	v_1	v_2
Papierosy	0.28	0.28
Piwo	0.29	0.40
Wino	0.26	0.39
Wódka	0.32	0.32
Kokaina	0.20	-0.29
Środki uspokajające	0.29	-0.26
Leki odurzające	0.18	-0.19
Heroina i opiaty	0.20	-0.32
Marihuana	0.34	-0.16
Haszysz	0.33	-0.05
Kleje itp.	0.28	-0.17
Halucynogeny	0.25	-0.33
Amfetamina	0.33	-0.23

62 x 47 = 2917

62 x 47 = 2917



„eigen-twarze”



Liniowa analiza dyskryminacyjna

Linear Discriminant Analysis

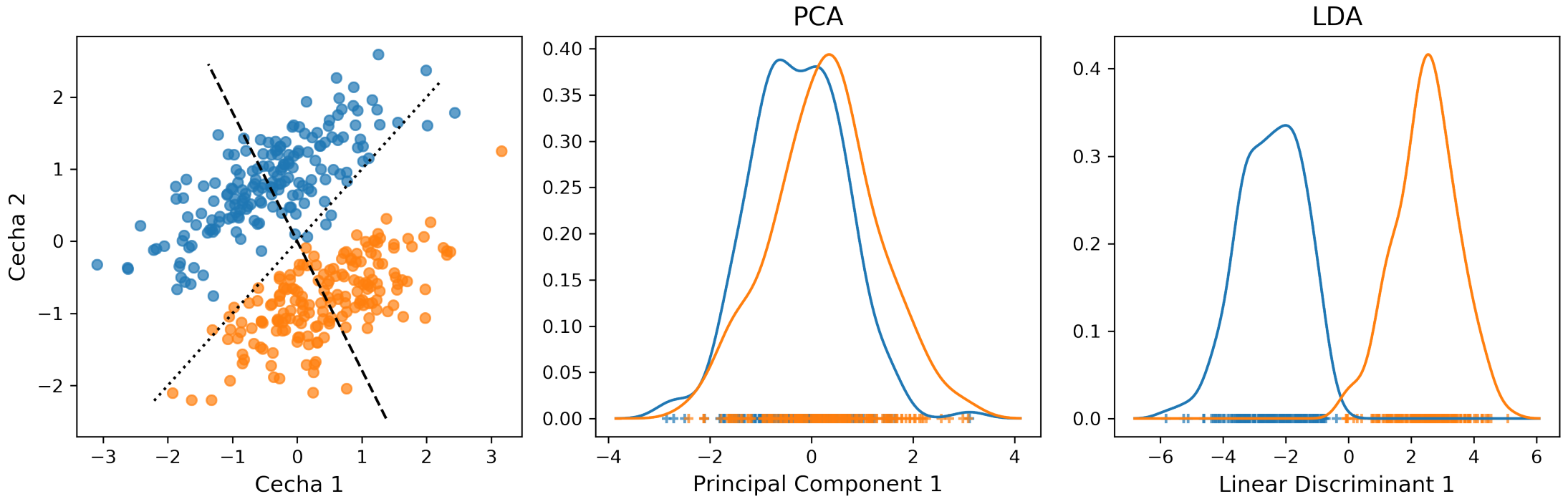
PCA

- nienadzorowana

LDA

- nadzorowana

Szukamy takiej podprzestrzeni cech, która **najlepiej separuje klasy**




The End