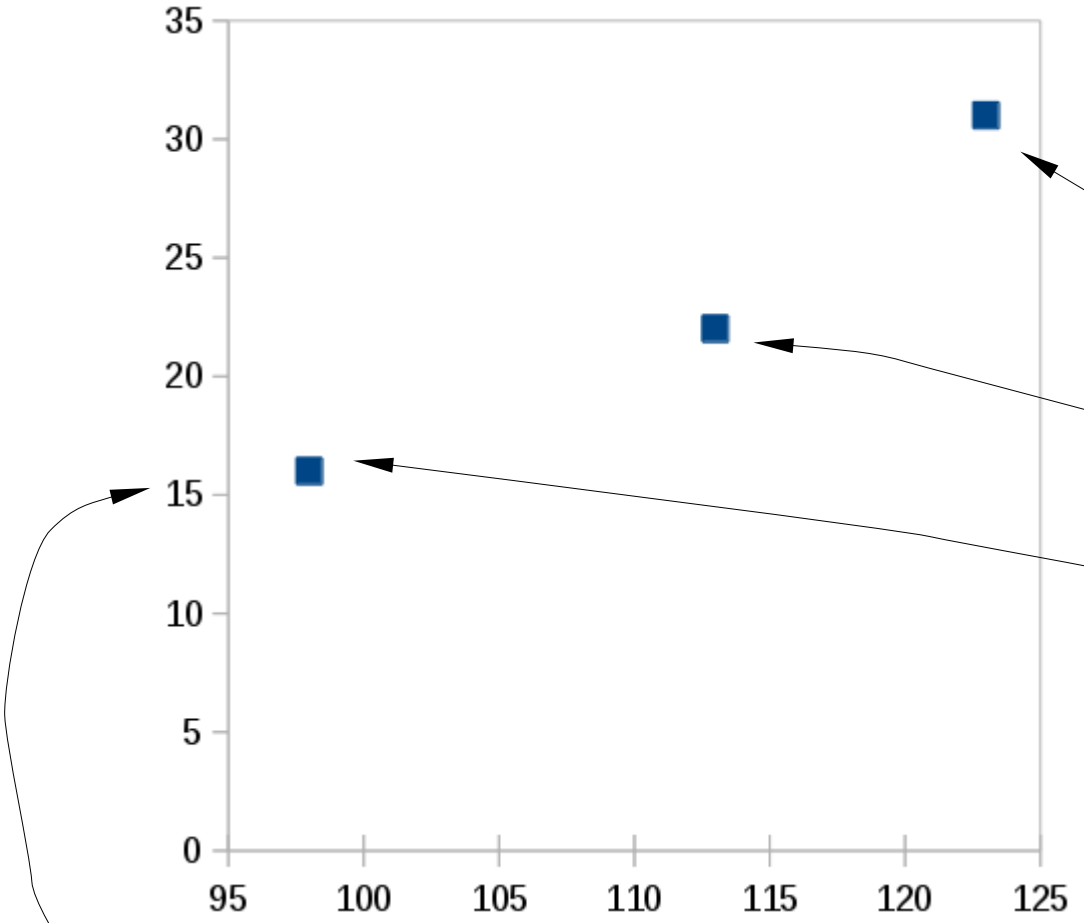
Analiza głównych składowych

Statystyka biomedyczna

Piotr M. Szczypiński

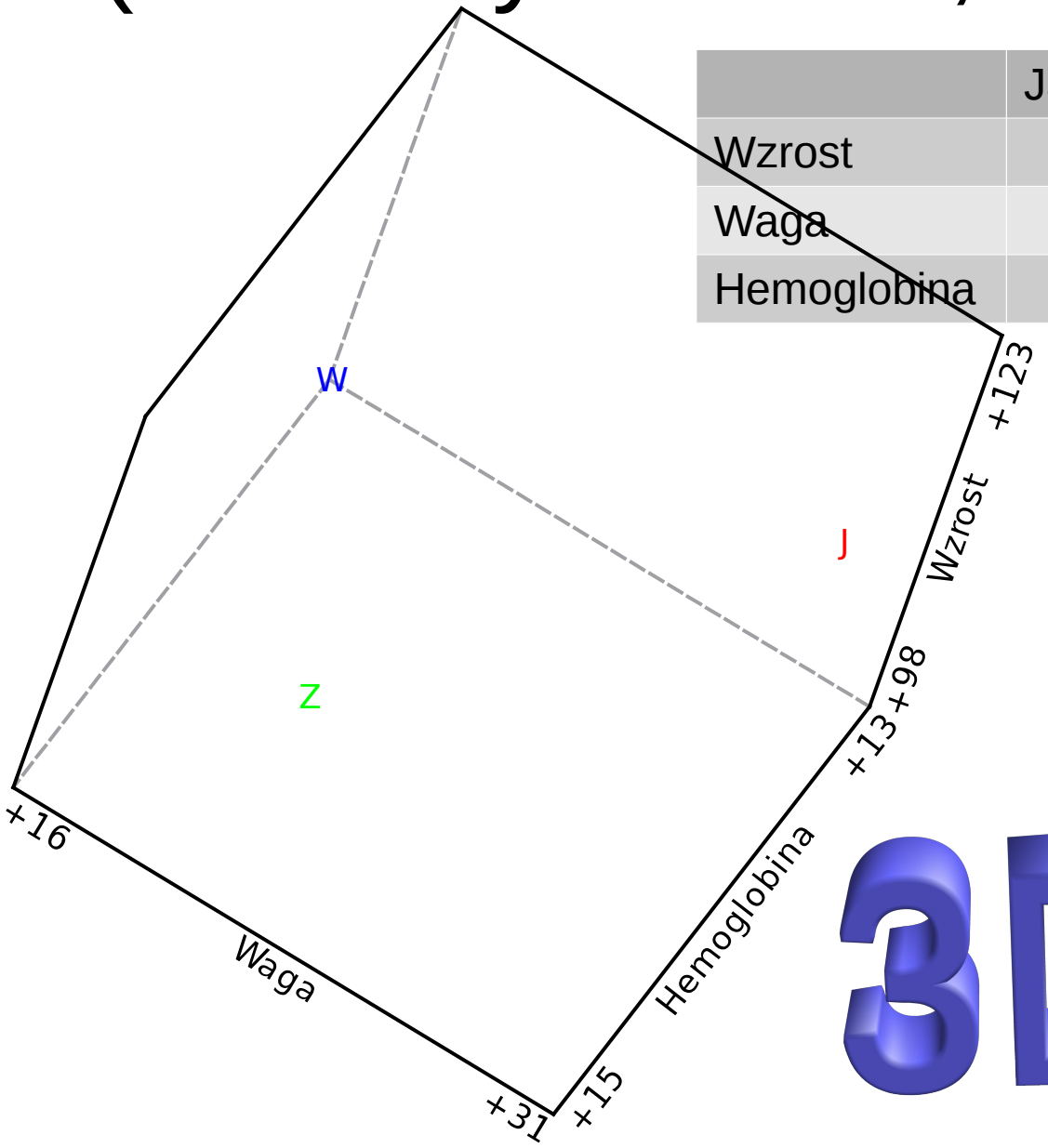
Analiza wielowymiarowa (wieloczynnikowa, multiwariacyjna)



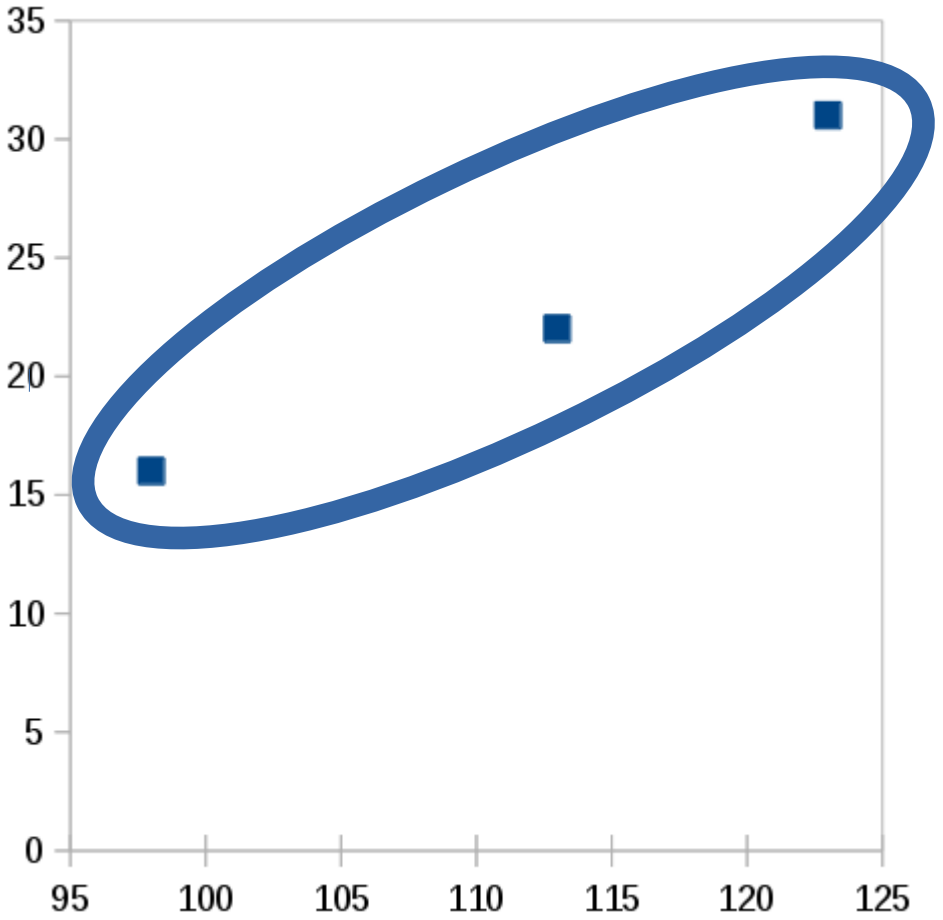
	Jaś	Zuzia	Wojtek
Wzrost	98	113	123
Waga	16	22	31

Analiza wielowymiarowa (wieloczynnikowa, multiwariacyjna)

	Jaś	Zuzia	Wojtek
Wzrost	123	113	98
Waga	31	22	16
Hemoglobina	14	16	13

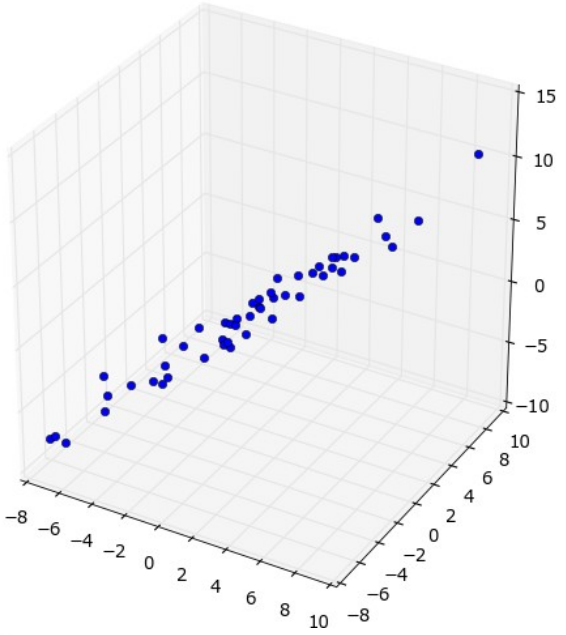


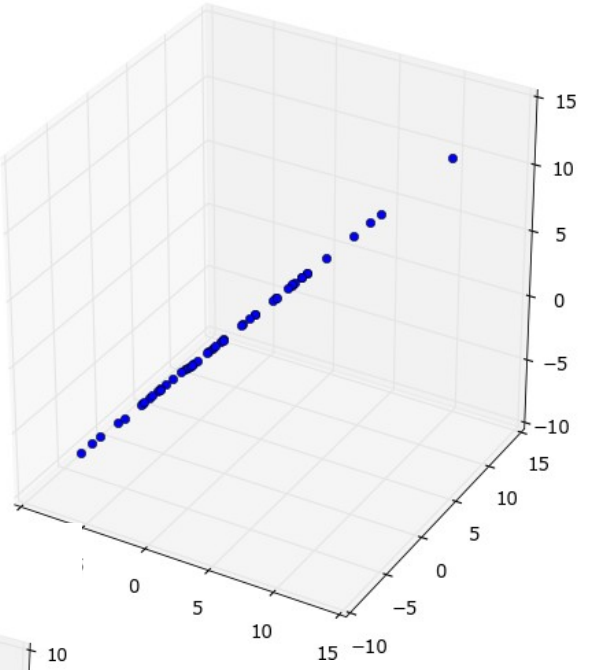
Czy zmienne są zależne?

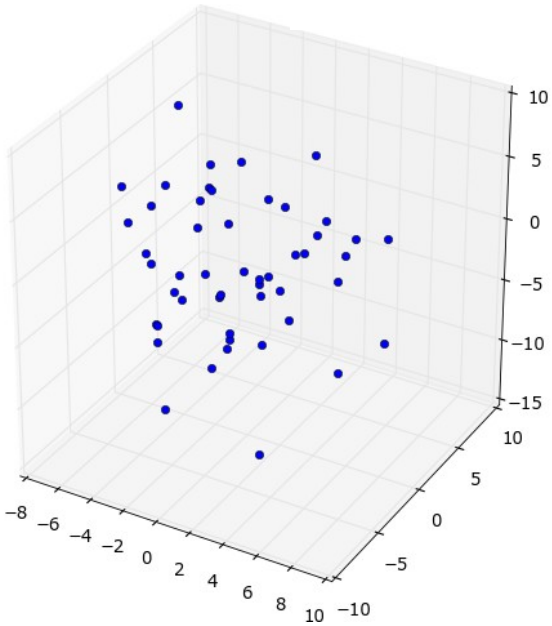


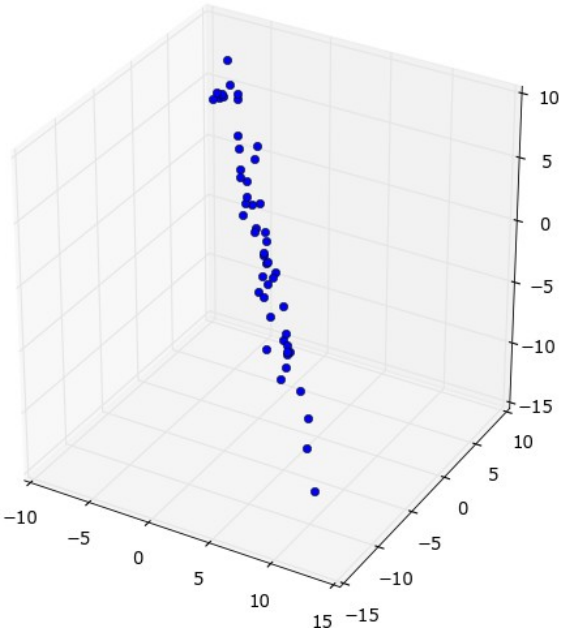
	Jaś	Zuzia	Wojtek
Wzrost	98	113	123
Waga	16	22	31

Czy zmienne są zależne?

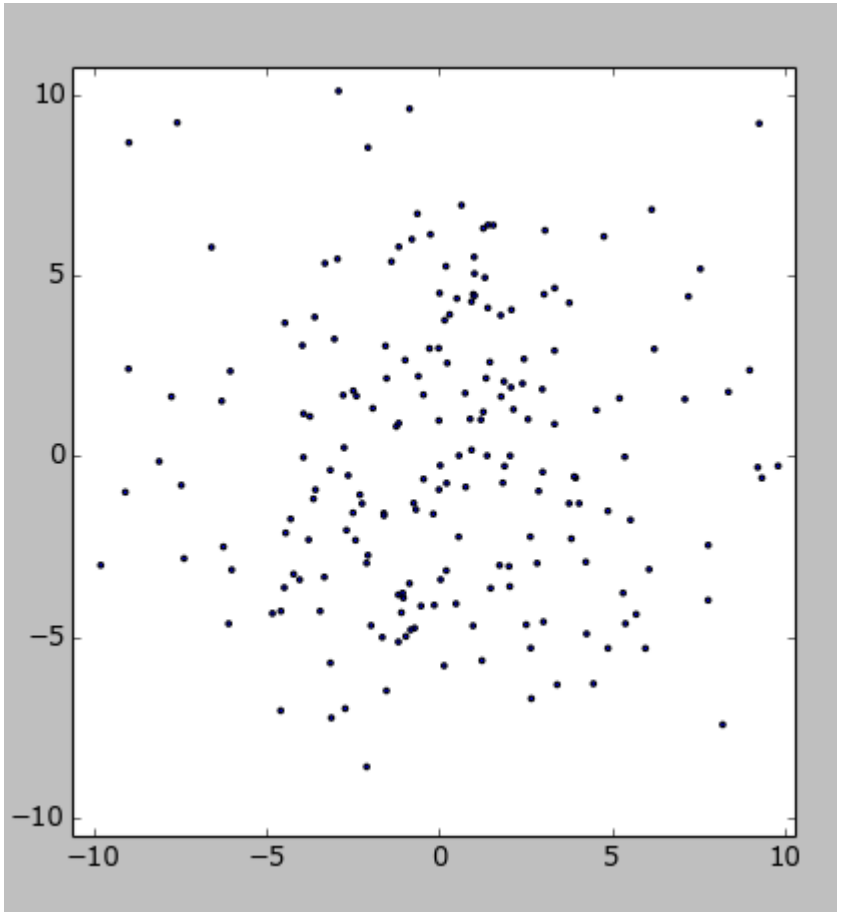


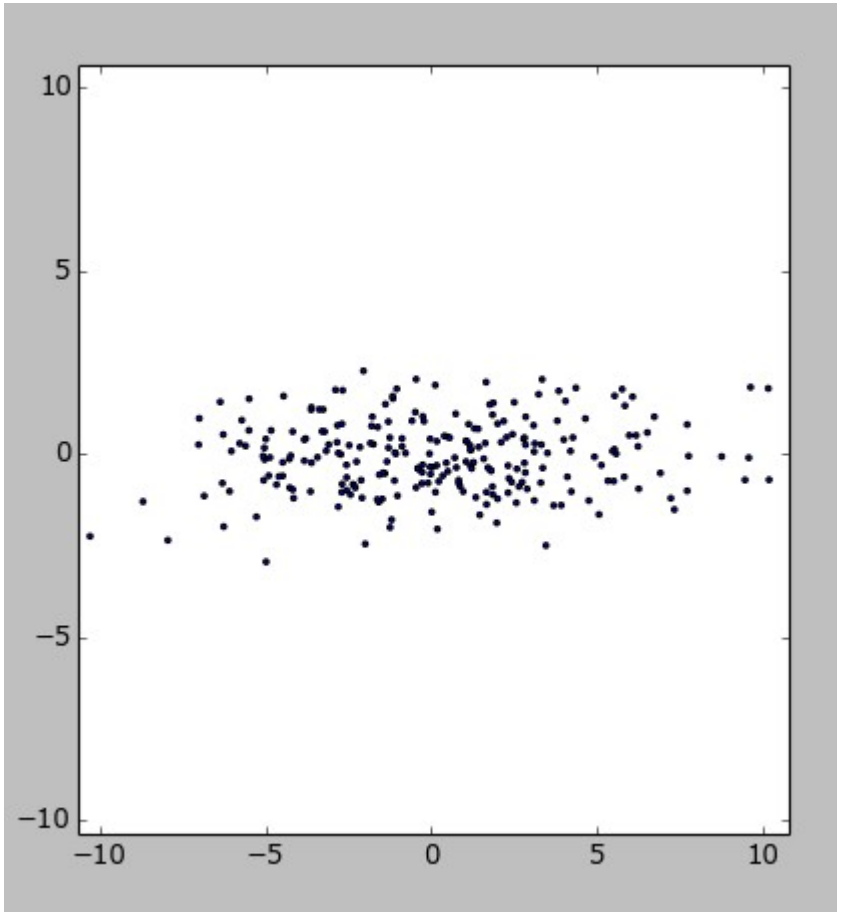




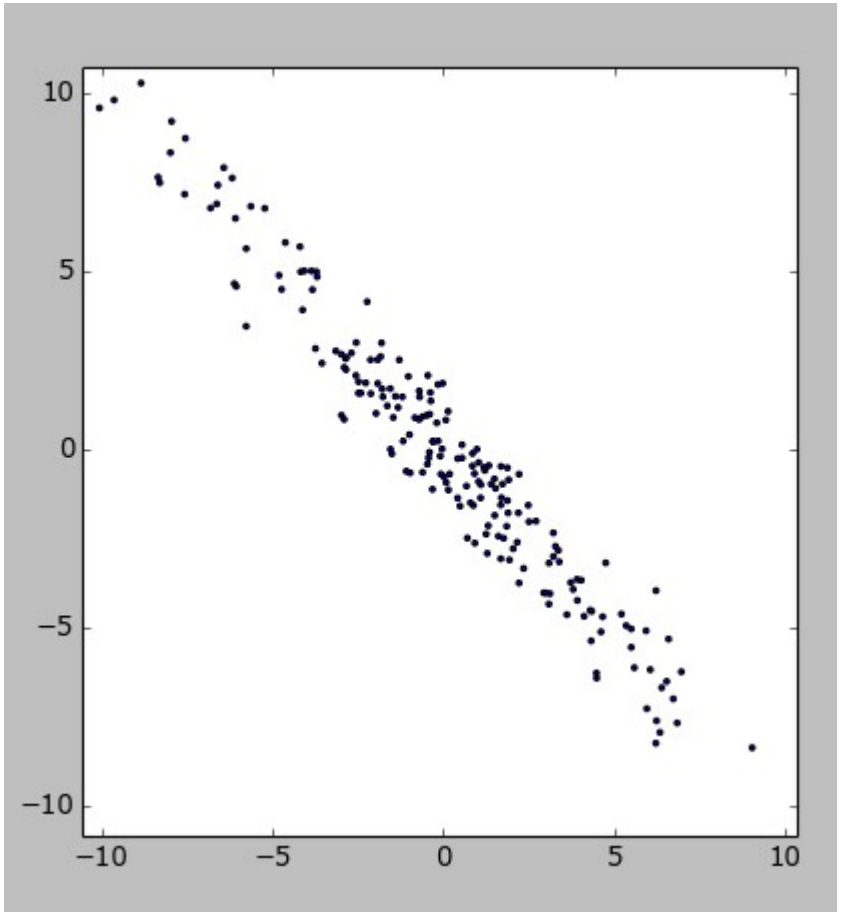


Ile jest różnych wariancji?





Macierz kowariancji



Macierz kowariancji jest uogólnieniem pojęcia wariancji na przypadek wielowymiarowy. Macierz taka dla wektora losowego (X_1, X_2, … X_n) ma postać:

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

Własności macierzy:

- jest symetryczna
- wyznacznik jest nieujemny
- Jeśli X_i, X_j są niezależne to $\sigma_{ij} = 0$
- $\Sigma = \mathrm{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}) - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$

Macierz kowariancji

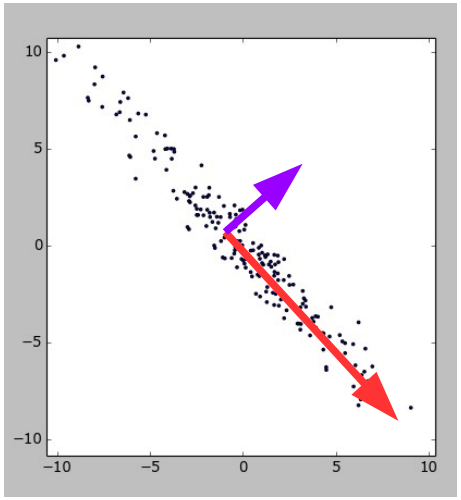
$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

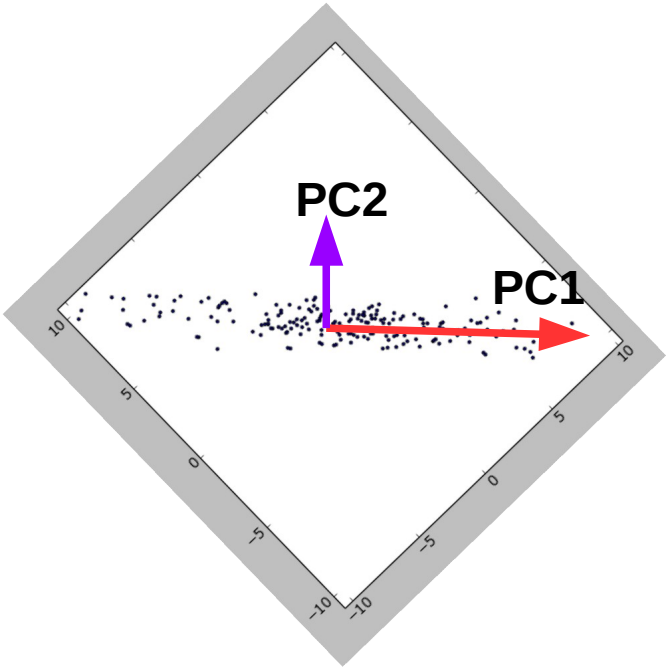
$$\mu_i = \mathrm{E}(X_i)$$

https://en.wikipedia.org/wiki/Covariance_matrix

Analiza głównych składowych

Principal Component Analysis (PCA)



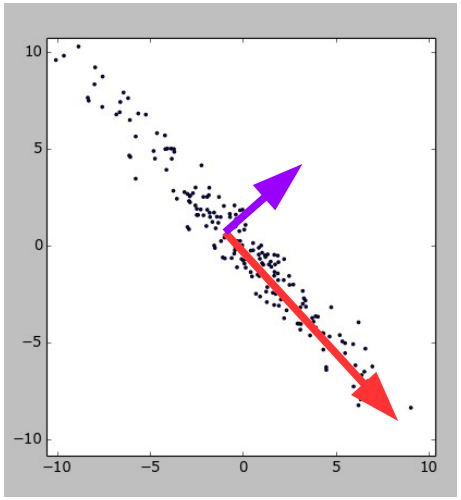


Celem PCA jest obrót przestrzeni *n*-wymiarowej w taki sposób, aby w nowej przestrzeni, zmienne były niezależne.

Kierunki nowej przestrzeni nazywane są **składowymi głównymi**.

Kolejność składowych głównych po przeprowadzeniu PCA jest od zmiennej o największej wariancji do zmiennej o najmniejszej wariancji

Analiza głównych składowych
Algorytm



1) Obliczenie macierzy kowariancji **Σ**
2) Rozkład macierzy symetrycznej **Σ** według wartości własnych (lub według wartości osobliwych):

$$\mathbf{\Sigma} = \mathbf{V}^{-1}\mathbf{D}\,\mathbf{V}$$

D macierz diagonalna wartości własnych
V macierz o kolumnach wektorów własnych

3) Uszeregowanie wartości własnych nierosnąco, uszeregowanie wektorów własnych zgodnie z kolejnością wartosci własnych
4) Transformacja liniowa przestrzeni za pomocą macierzy wektorów własnych **V**

Przykład zastosowania (5D)

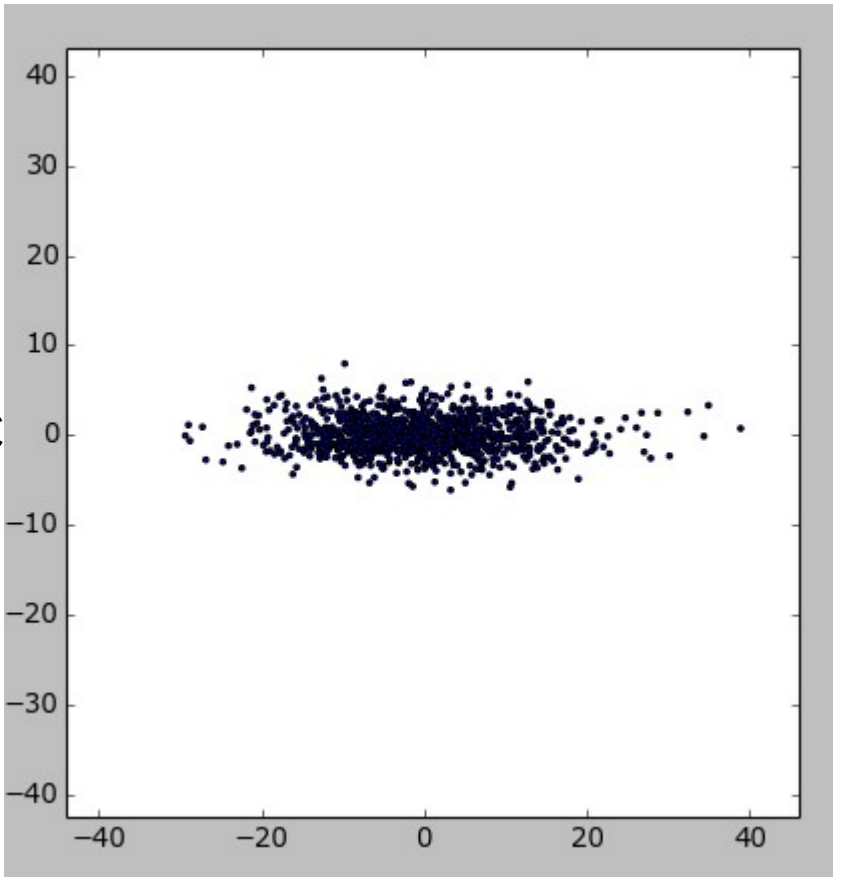
Mając zbiór danych dotyczących grupy osób charakteryzowanych przez 5 zmiennych (np. wzrost, waga, wiek, dochód, powierzchnia mieszkania) można przypuszczać, że zmienne "wzrost" i "waga" będą ze sobą dodatnio skorelowane (zależne). Po to żeby uzyskać większą przejrzystość danych lub uniknąć powielania się danych warto zastąpić dwie zmienne jedną (składową), którą można nazwać na przykład "wielkość". Podobnie skorelowane będą ze sobą zmienne "dochód" i "powierzchnia mieszkania", które można zastąpić czynnikiem "zamożność".

https://pl.wikipedia.org/wiki/Analiza_głównych_składowych

Przykład PCA w Pythonie

```
from numpy import *
from matplotlib.pyplot import *
from scipy.stats import *

standardowy = norm(loc=0.0,
scale=1.0)

# Liczba próbek
N=1000
# Zadane odchylenia standardowe
StdB = 10
StdC = 2

B = StdB * standardowy.rvs(N)
C = StdC * standardowy.rvs(N)

E = cov(B,C)
print("Macierz kowariancji B,C:")
print E
```

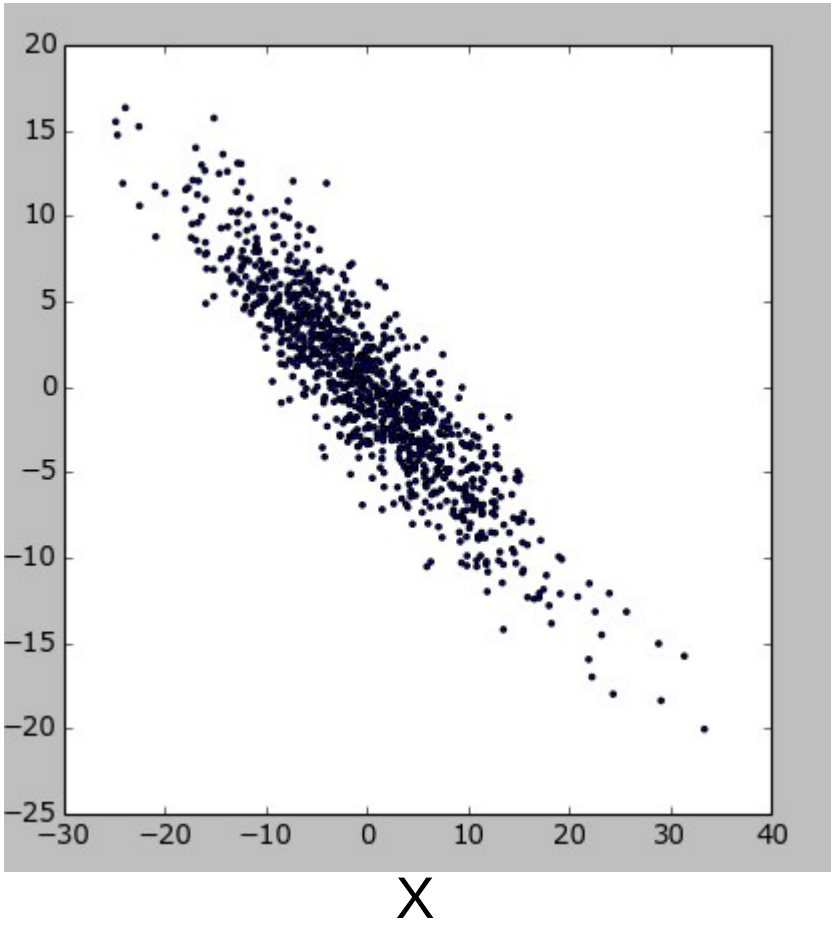
scatter(B, C, 5)



102.94	-0.51
-0.51	4.05

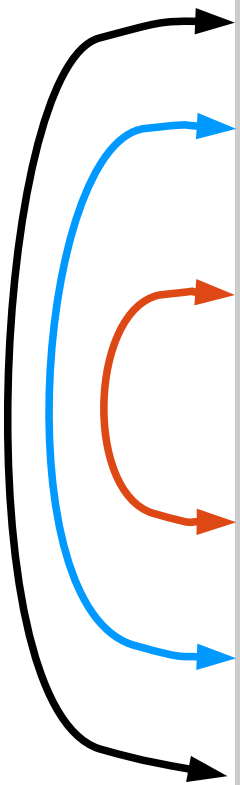
```
# Zadany kąt obrotu
A=0.56
CosA = cos(A)
SinA = sin(A)

print("Kąt obrotu zadany:")
print 180*A/3.14
print("Macierz obrotu zadana:")
print CosA, SinA
print -SinA, CosA

X = CosA*B+SinA*C
Y = -SinA*B+CosA*C
E = cov(X,Y)
print("Macierz kowariancji X,Y:")
print E
```

scatter(X, Y, 5)



74.57	-44.73
-44.73	32.42

```
eigvects, eigvals, V = linalg.svd(E, full_matrices=False)
print("Macierz wektorów własnych:")
print eigvects

print("Wartości własne:")
print eigvals
print("Kąt obrotu z PCA:")
print 180*math.atan2(eigvects[0][1], eigvects[0][0])/3.14
```



Kąt obrotu zadany:
32.101911
Macierz kowariancji B,C:
[[102.94466534 -0.51923671]
 [-0.51923671 4.05551266]]
Macierz obrotu zadana:
0.847255 0.531186
-0.531186 0.847255

Macierz wektorów własnych:
[[-0.84445445 0.53562737]
 [0.53562737 0.84445445]]
Wartości własne:
[102.94739162 4.05278638]
Kąt obrotu z PCA:
-32.213380