SPRAWOZDANIE

Zajęcia: Matematyka Konkretna
Prowadzący: prof. dr hab. Vasyl Martsenyuk

Zadanie 4
Temat: Analiza głównych składowych
Wariant 13

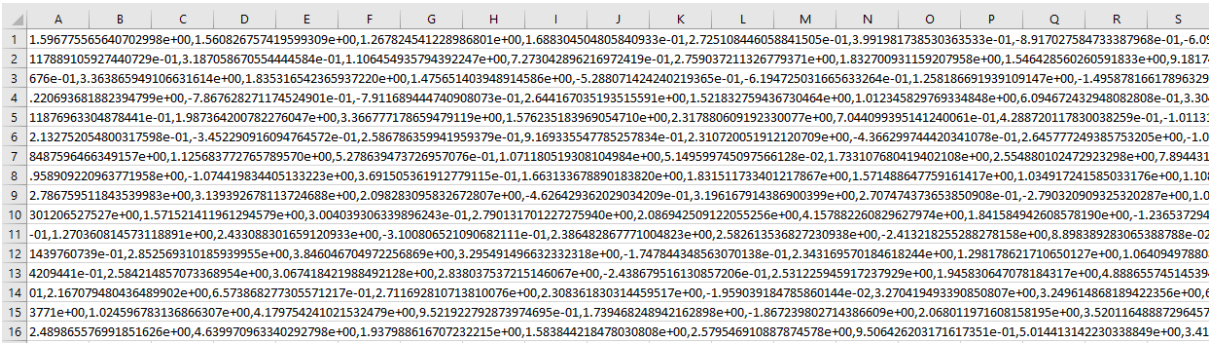
Łukasz Pindel
Informatyka II stopień,
stacjonarne,
2 semestr,
Gr. 1B

1. Polecenie:

Zadaniem do zrealizowania jest obliczenie środka, osi głównych oraz kąta obrotu danych dwuwymiarowych z pliku csv zgodnie z wariantem zadania.

2. Wprowadzane dane:

Wariant 13 – plik csv z wartościami



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	1.596775565640702998e+00,1.560826757419599309e+00,1.267824541228986801e+00,1.688304504805840933e-01,2.725108446058841505e-01,3.991981738530363533e-01,-8.917027584733387968e-01,-6.05																		
2	117889105927440729e-01,3.187058670554444584e-01,1.106454935794392247e+00,7.273042896216972419e-01,2.759037211326779371e+00,1.832700931159207958e+00,1.546428560260591833e+00,9.1817																		
3	676e-01,3.363865949106631614e+00,1.835316542365937220e+00,1.475651403948914586e+00,-5.288071424240219365e-01,-6.194725031665633264e-01,1.258186691939109147e+00,-1.495878166178963291																		
4	.220693681882394799e+00,-7.867628271174524901e-01,-7.911689444740908073e-01,2.644167035193515591e+00,1.521832759436730464e+00,1.012345829769334848e+00,6.094672432948082808e-01,3.30																		
5	1187696330487844 1e-01,1.987364200782276047e+00,3.366777178659479119e+00,1.576235183969054710e+00,2.317880609192330077e+00,7.044099395141240061e-01,4.288720117830038259e-01,-1.01131																		
6	2.132752054800317598e-01,-3.452290916094764572e-01,2.586786359941959379e-01,9.169335547785257834e-01,2.310720051912120709e+00,-4.366299744420341078e-01,2.645777249385753205e+00,-1.0																		
7	8487596466349157e+00,1.125683772765789570e+00,5.278639473726957076e-01,1.071180519308104984e+00,5.149599745097566128e-02,1.733107680419402108e+00,2.554880102472923298e+00,7.894431																		
8	.9589092209637719 58e+00,-1.074419834405133223e+00,3.691505361912779115e-01,1.663133678890183820e+00,1.831511733401217867e+00,1.571488647759161417e+00,1.034917241585033176e+00,1.108																		
9	2.786759511843539983e+00,3.139392678113724688e+00,2.098283095832672807e+00,-4.626429362029034209e-01,3.196167914386900399e+00,2.707474373653850908e-01,-2.790320909325320287e+00,1.0																		
10	3012065275 27e+00,1.571521411961294579e+00,3.004039306339896243e-01,2.790131701227275940e+00,2.086942509122055256e+00,4.157882260829627974e+00,1.841584942608578190e+00,-1.236537294																		
11	-01,1.270360814573118891e+00,2.433088301659120933e+00,-3.100806521090682111e-01,2.386482867771004823e+00,2.582613536827230938e+00,-2.413218255288278158e+00,8.898389283065388788e-02																		
12	1439760739e-01,2.852569310185939955e+00,3.846046704972256869e+00,3.295491496632332318e+00,-1.747844348563070138e-01,2.343169570184618244e+00,1.298178621710650127e+00,1.064094978 80																		
13	4209441e-01,2.584214857073368954e+00,3.067418421988492128e+00,2.838037537215146067e+00,-2.438679516130857206e-01,2.531225945917237929e+00,1.945830647078184317e+00,4.88865574514539																		
14	01,2.167079480436489902e+00,6.573868277305571217e-01,2.711692810713810076e+00,2.308361830314459517e+00,-1.959039184785860144e-02,3.270419493390850807e+00,3.249614868189422356e+00,6																		
15	3771e+00,1.024596783136866307e+00,4.179754241021532479e+00,9.521922792873974695e-01,1.739468248942162898e+00,-1.867239802714386609e+00,2.068011971608158195e+00,3.520116488872964 57																		
16	2.489865576991851626e+00,4.639970963340292798e+00,1.937988616707232215e+00,1.583844218478030808e+00,2.579546910887874578e+00,9.506426203171617351e-01,5.014413142230338849e+00,3.41																		

Rysunek 1: Zawartość pliku csv

3. Wykorzystane komendy:

Wczytywanie i przygotowanie danych:

data = pd.read_csv('13.csv', sep=',', header=None)
data = data.to_numpy()

Dane z pliku "13.csv" są wczytywane za pomocą funkcji **read_csv()** z biblioteki Pandas. Parametr sep=',' określa separator kolumn w pliku, a header=None oznacza, że plik nie zawiera wiersza nagłówkowego. Następnie dane są konwertowane na tablicę NumPy za pomocą funkcji **to_numpy()**, aby były gotowe do dalszej analizy.

Generowanie chmury punktów:

X = R @ np.diag(sig) @ data + np.diag(xC) @ np.ones((2,nPoints))

Na podstawie wczytanych danych tworzona jest chmura punktów X. Dane są transformowane przy użyciu macierzy rotacji R, wektora odchylenia standardowego sig i wektora średnich xC.

Obliczanie średniej i macierzy kowariancji:

```
Xavg = np.mean(X,axis=1)
B = X - np.tile(Xavg,(nPoints,1)).T
```

W tej części obliczana jest średnia wartość dla każdej współrzędnej punktów w chmurze danych X, a następnie tworzona jest macierz kowariancji B poprzez odjęcie od chmury danych X macierzy średnich Xavg.

Dekompozycja SVD:

```
U, S, VT = np.linalg.svd(B/np.sqrt(nPoints),full_matrices=0)
```

Dekompozycja SVD macierzy kowariancji B. Wynikiem są macierz lewych wektorów singularnych U, wektor wartości singularnych S oraz macierz prawych wektorów singularnych VT.

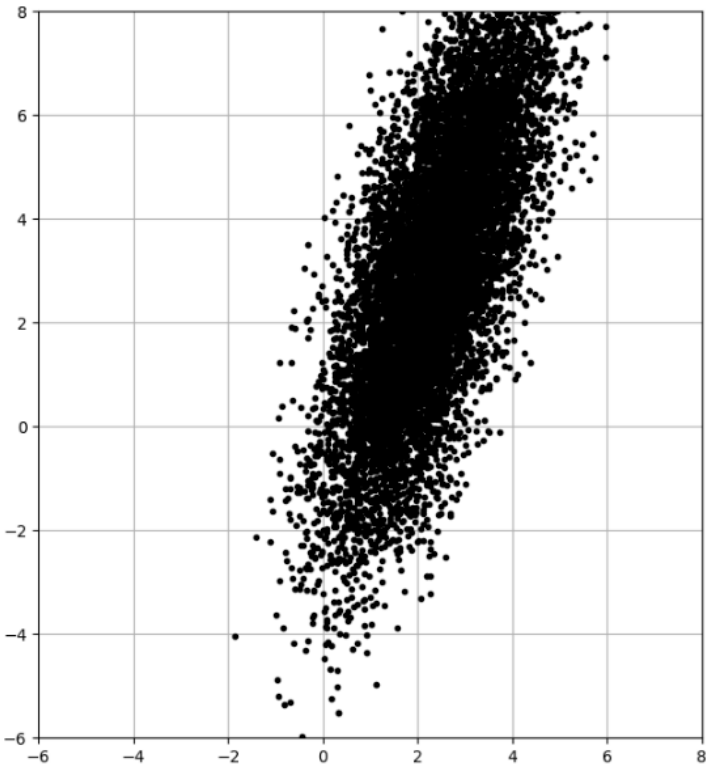
Rysowanie wykresów:

```
ax2.plot(X[0,:],X[1,:], '.', color='k')
ax2.plot(Xavg[0] + Xstd[0,:], Xavg[1] + Xstd[1,:],'-',color='r',linewidth=3)
```

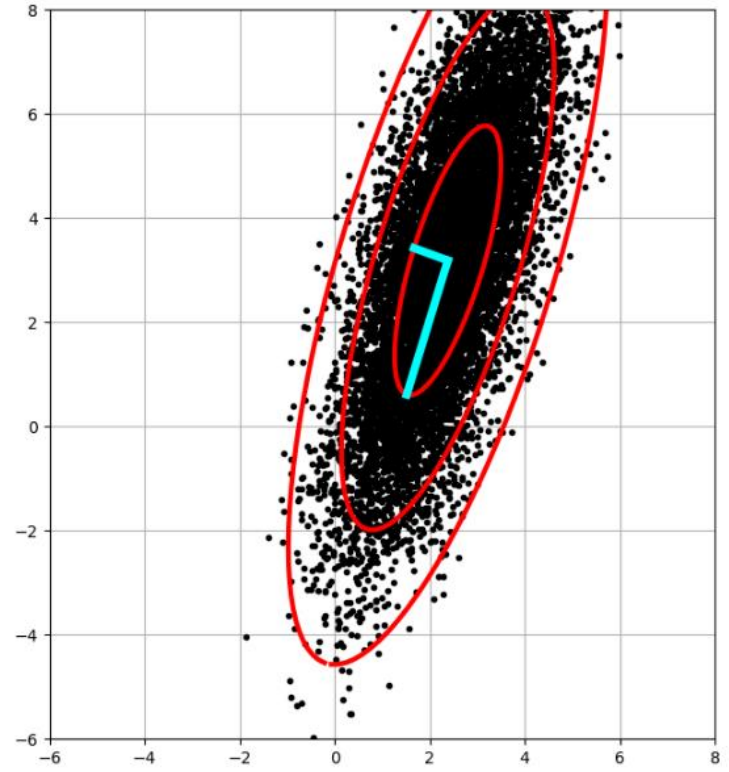
Na pierwszym wykresie przedstawiona jest chmura punktów X, gdzie każdy punkt reprezentuje jedną obserwację z danych. Na drugim wykresie ta sama chmura punktów zostaje ponownie przedstawiona w celu nałożenia na nią analizy PCA. Dodatkowo na drugim wykresie rysowane są elipsy reprezentujące 1-, 2- i 3-krotne odchylenia standardowe od środka danych oraz linie reprezentując

Link do repozytorium:
https://github.com/denniak/MK/tree/main/MK_4

4. Wynik działania:



Rysunek 2: Przedstawienie obserwacji w postaci chmury punktów



Rysunek 3: Przedstawienie obserwacji wraz z nałożoną analizą PCA

5. Wnioski:

Na podstawie otrzymanego wyniku można stwierdzić, że przeprowadzona analiza PCA umożliwiła identyfikację środka danych oraz głównych osi zmienności. Średnia wartość dla każdej współrzędnej punktów wskazuje na ich centrum, natomiast osie główne, otrzymane dzięki dekompozycji SVD, reprezentują kierunki maksymalnej zmienności w danych dwuwymiarowych. Dodatkowo, obecność elips na wykresie drugim wskazuje na rozkład punktów wokół środka danych oraz ilustruje ich zmienność w różnych kierunkach.