Sprawozdanie – Zastosowanie metody składowych głównych (PCA)

Jakub Zięba

Cel pracy

Celem pracy jest przedstawienie metody składowych głównych i opisanie jej założeń, następnie zastosowanie metody na odpowiednio dużym zbiorze danych z uprzednią weryfikacją ww. założeń, interpretacja uzyskanych wyników oraz porównanie ich z wynikami uzyskanymi przy zaniedbaniu części z nich.

Opis zbioru danych i założeń

Opis datasetu

Wybrany zbiór danych składa się z informacji o słuchotkach, ślimakach morskich, wyglądem przypominających małże czy ostrygi. Uwzględnione cechy zwierząt to:

- Sex – płeć zwierzęcia, męska lub żeńska;
- Length – pomiar skorupy w najdłuższym jej miejscu, długość wyrażona w mm;
- Diameter – pomiar skorupy w miejscu prostopadłym do miejsca wymienionego wyżej, tj. ze zmiennej Length. Długość wyrażona w mm.
- Height – wysokość zwierzęcia, mierzona od miejsca styku z podłożem do najwyższego miejsca skorupy, wyrażona w mm;
- Whole Weight – całkowita waga ślimaka, łącznie skorupy oraz jej wnętrzności, wyrażona w gramach;
- Shucked Weight – waga „mięsa” ślimaka, wyrażona w mm;
- Viscera Weight – waga wnętrzności po ich wykrwawieniu, wyrażona w mm;
- Shell Weight – waga muszli po wysuszeniu, wyrażona w mm;
- Rings – liczba pierścieni na muszli, liczba całkowita. Pierścieni można użyć, podobnie jak słojów w przypadku drzew, do określania wieku zwierząt. Można to osiągnąć dodając 1.5 do liczby pierścieni.

Opis założeń

Przy stosowaniu metody PCA należy spełnić szereg założeń, które pozwolą na poprawne jej wykonanie i w efekcie otrzymanie poprawnych wyników:

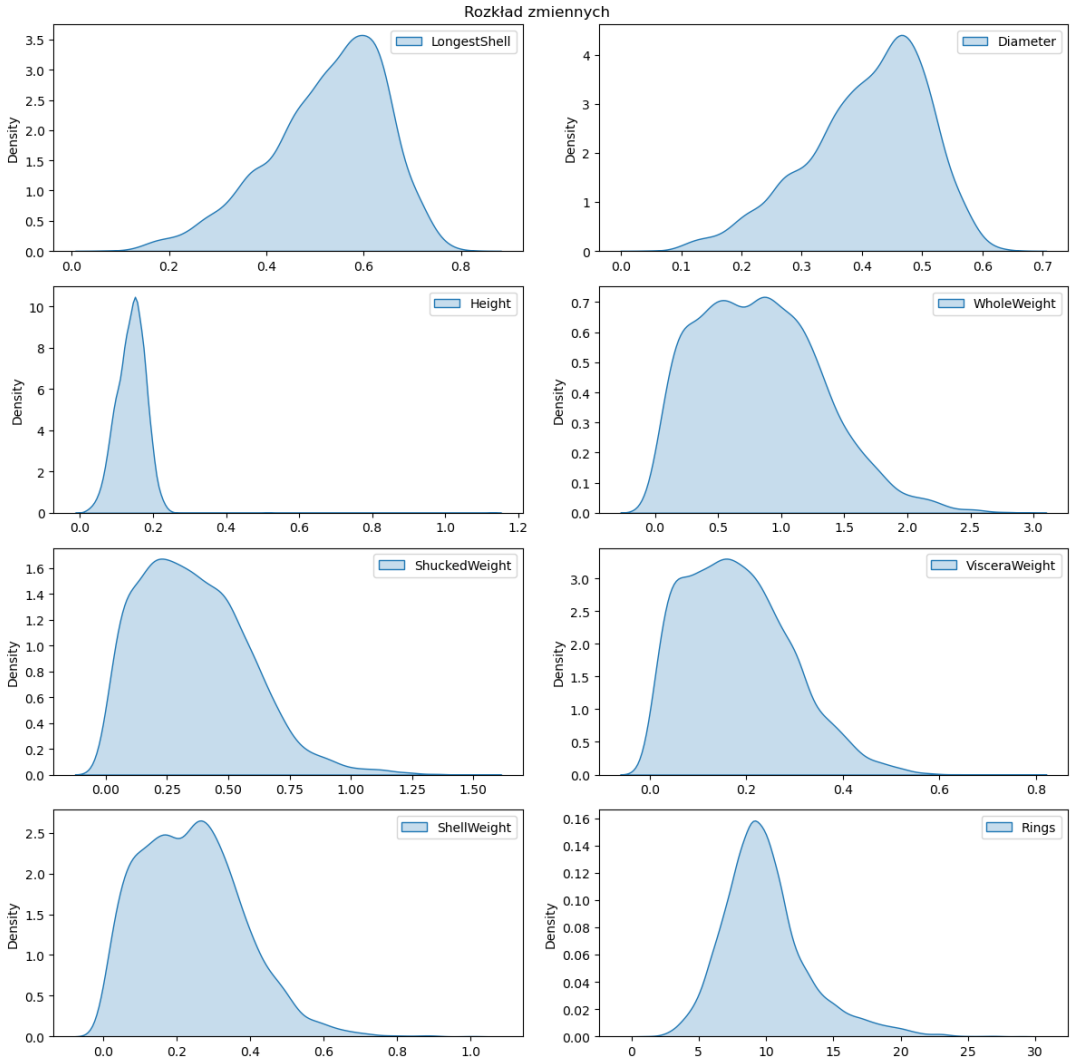
- Należy pozbyć się outlierów, ponieważ będą zakłócać przeprowadzanie badania, które może w efekcie jedynie podkreślić to że występują outlierzy;
- Dane należy zestandaryzować. Jest to wymagane szczególnie w przypadku danych mocno się od siebie różniących jeśli chodzi o rząd wielkości. W naszym przykładzie można spojrzeć na zmienne Diameter oraz Rings. Średnia wartość tej pierwszej to 0.4 mm, największa to 0.65 mm, natomiast dla zmiennej Rings już najniższą wartością jest 1, średnia to prawie 10 a maksimum 29. Takich zmiennych nie można badać przed ich standaryzacją, gołym okiem widać jak bardzo od siebie odstają.
- Poszczególne zmienne powinny mieć również rozkłady przypominające normalne.

Analiza i obróbka danych

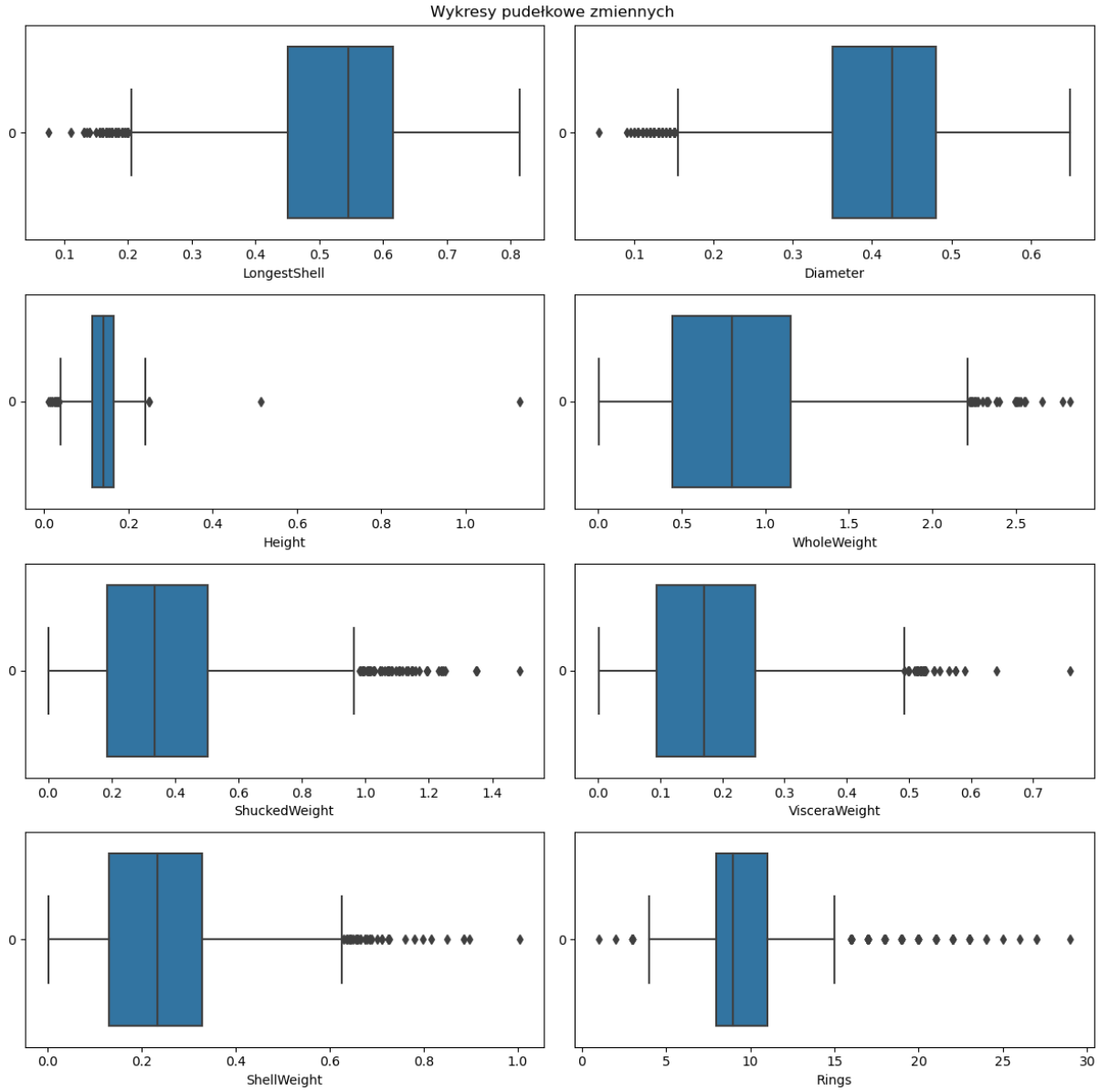
Przed przystąpieniem do obróbki i stosowania PCA, najpierw zajmiemy się podstawową analizą zbioru danych.

	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Rings
mean	0.5241	0.4079	0.1396	0.8290	0.3595	0.1807	0.2388	9.9351
std	0.1201	0.0992	0.0417	0.4903	0.2220	0.1096	0.1392	3.2242
min	0.0750	0.0550	0.0100	0.0020	0.0010	0.0005	0.0015	1.0000
25%	0.4500	0.3500	0.1150	0.4422	0.1862	0.0935	0.1300	8.0000
50%	0.5450	0.4250	0.1400	0.8000	0.3360	0.1710	0.2340	9.0000
75%	0.6150	0.4800	0.1650	1.1535	0.5020	0.2530	0.3288	11.0000
max	0.8150	0.6500	1.1300	2.8255	1.4880	0.7600	1.0050	29.0000

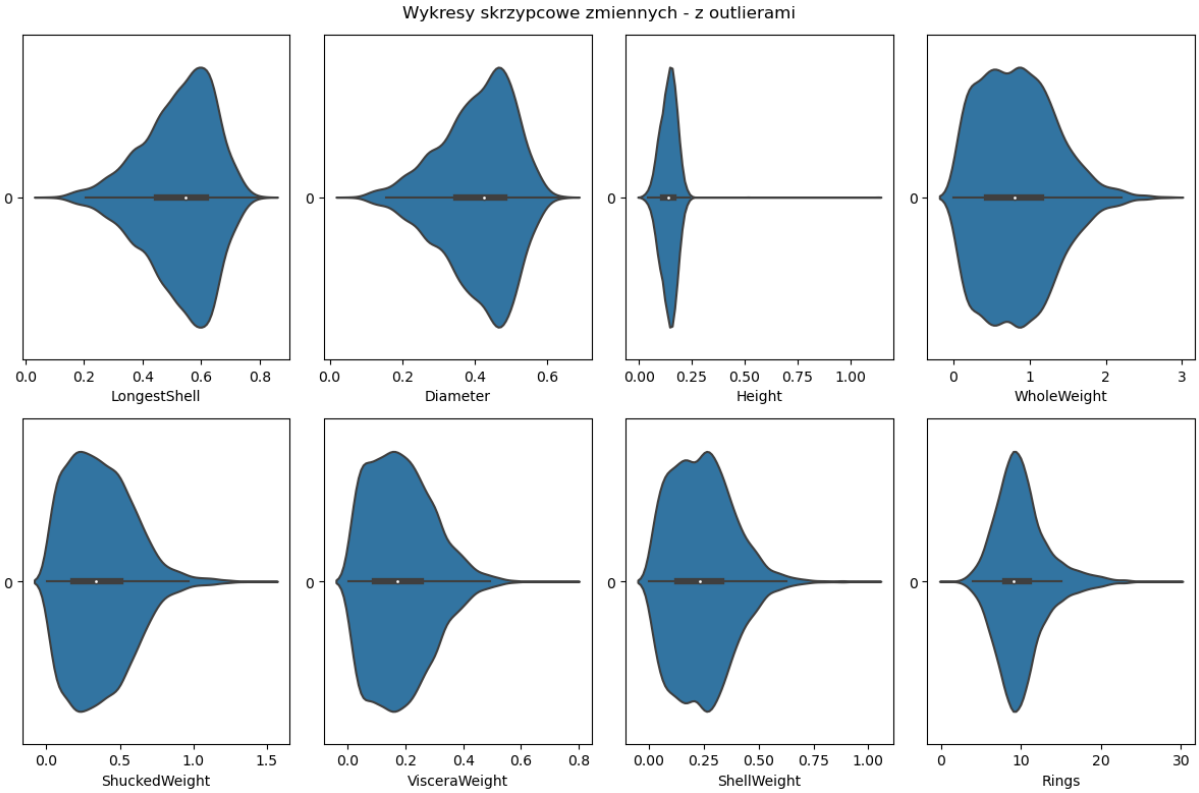
Analizując statystyki związane z wymienionymi wyżej cechami, możemy zobaczyć że dane zawierają informacje o bardzo różnorodnych osobnikach. Najmłodszy z nich miał niespełna 2.5 roku, najstarszy natomiast ponad 30 lat. Zbadano tutaj zarówno osobniki maleńkie, o wadze tak małej jak 0.002 grama, jak i nieporównywalnie większe względem tych poprzednich, ważące niemal 3kg.

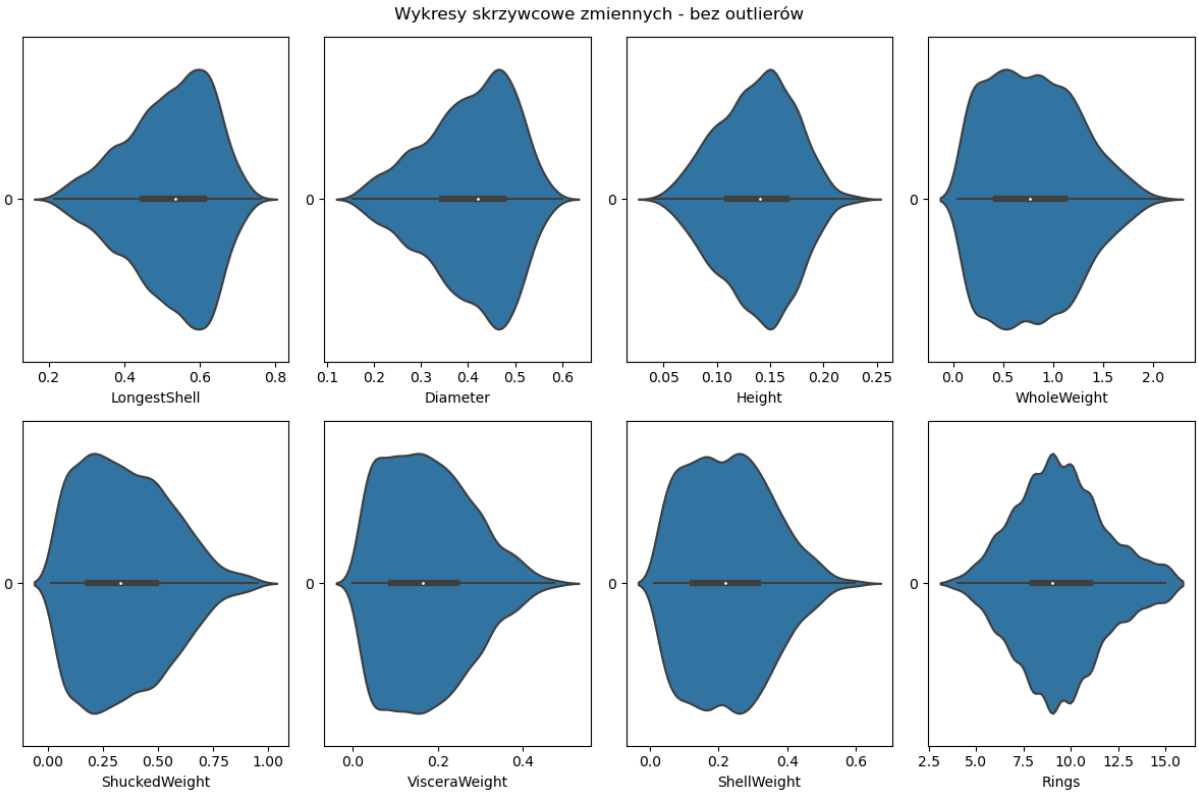


Rozkłady poszczególnych zmiennych wskazują na występowanie wartości odstających, których należy się pozbyć.



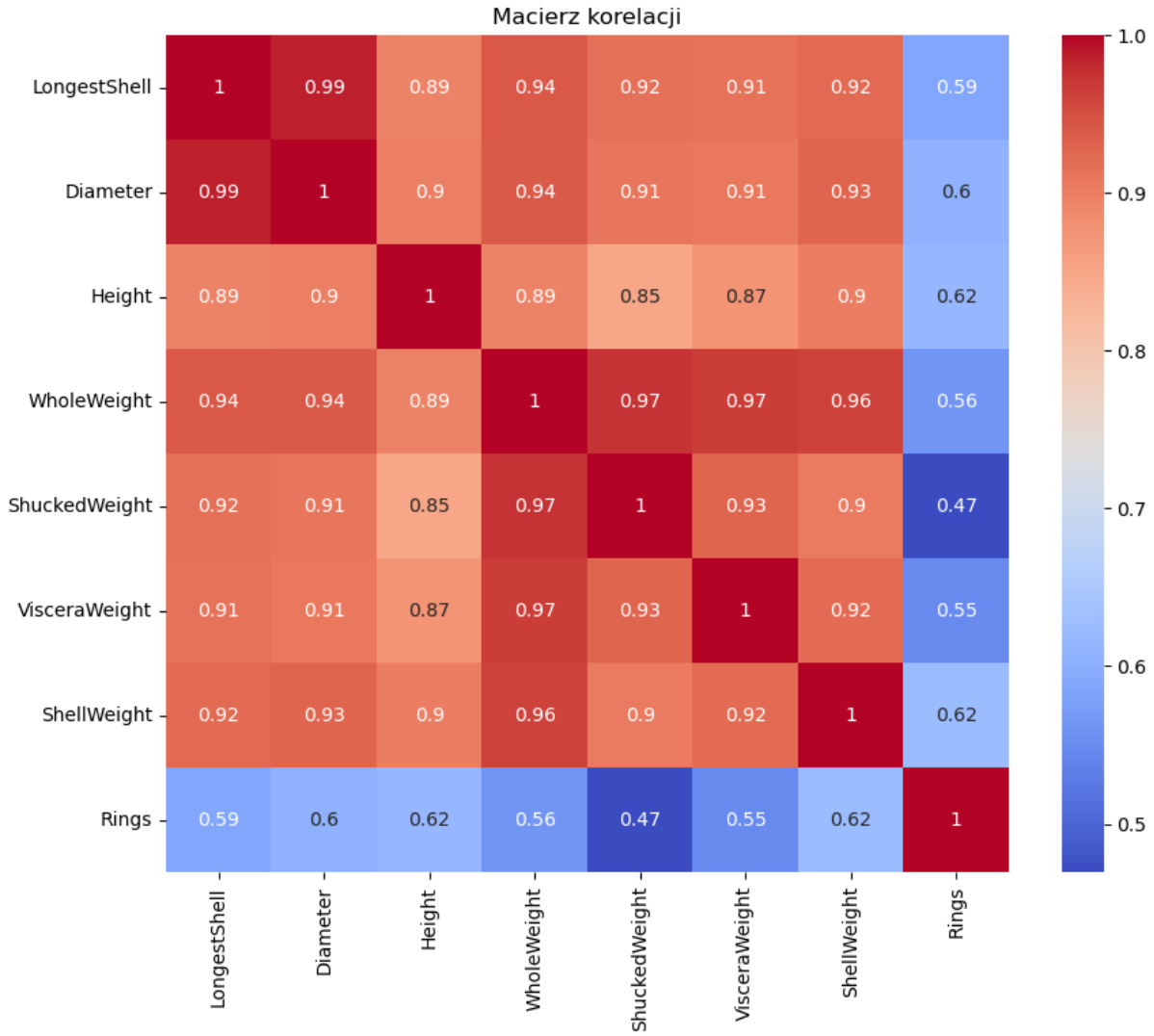
Wykresy pudełkowe potwierdzają to przypuszczenie, wartości odstających jest całkiem sporo. Aby się ich pozbyć można m.in. użyć metody wykorzystującej rozstęp ćwiartkowy.





Po wyrzuceniu outlierów można porównać zbiór danych na wykresach skrzypcowych – widać znaczną poprawę jeśli chodzi o wartości odstające i długość „ogonów” wykresów. Wykresy te pokazują również, że większość rozkładów zmiennych przypomina w pewnym stopniu rozkłady normalne, choć na pewno żaden z nich nie jest dokładnie taki.

Kolejnym elementem który należy poddać analizie są współczynniki korelacji między zmiennymi.



Współczynniki korelacji wydają się być wystarczająco wysokie, jednak sprawdzimy to jeszcze z pomocą formalnego testu – testu Bartletta.

```
#sprawdzenie testem Bartletta czy dane są odpowiednio powiązane
#H0: dane są niepowiązane
#H1: dane są powiązane
bartlett_ = bartlett(data['LongestShell'], data['Diameter'], data['Height'], data['WholeWeight'], data['ShuckedWeight'],data['VisceraWeight'], data['ShellWeight'], data['Rings'])
stat, p_value = bartlett(data['LongestShell'], data['Diameter'], data['Height'], data['WholeWeight'], data['ShuckedWeight'],data['VisceraWeight'], data['ShellWeight'], data['Rings'])
if p_value > 0.05:
    print('Dane nie są dostatecznie powiązane')
else:
    print('Dane są dostatecznie powiązane')
✓ 0.0s
Dane są dostatecznie powiązane
```

Według testu Bartletta dane są dostatecznie powiązane. Pozostało więc jedynie poddać dane standaryzacji i możemy przejść do stosowania metody PCA.

Opis i zastosowanie metody PCA

Analiza głównych składowych (PCA) to technika statystyczna wykorzystywana do analizy zbioru danych, który składa się z wielu zmiennych zawierających wiele obserwacji. Wizualizuje się ten zbiór danych jako chmurę punktów w wielowymiarowej przestrzeni. Głównym celem PCA jest obrót układu współrzędnych w taki sposób, aby w przypadku pierwszej współrzędnej wariancja była jak największa i malała dla kolejnych współrzędnych. To przekształcenie prowadzi do powstania nowych zmiennych, nazywanych składnikami głównymi, które najlepiej wyjaśniają zmienność w danych. Cechą

odróżniającą PCA od technik skalowania wielomiarowego jest właśnie to badanie zmienności, natomiast techniki skalowania badają odległości między obiektami.

PCA znajduje zastosowanie w różnych dziedzinach. Jednym z jego głównych zastosowań jest redukcja wymiarowości danych, co oznacza, że można usunąć mniej ważne składniki, zachowując jednocześnie dużą część informacji. Można także próbować interpretować te składniki w kontekście danych, co pozwala lepiej zrozumieć charakter danych, choć może być to trudne, zwłaszcza w przypadku dużej liczby zmiennych.

Z matematycznego punktu widzenia metoda polega na wyznaczeniu macierzy kowariancji, dla której następnie obliczane są wartości własne, a na ich podstawie wektory własne. Po ich wyznaczeniu dokonuje się na nie projekcji początkowego zbioru danych, co w efekcie daje nam zbiór przekształcony. W językach programowania przeznaczonych do analizy danych możemy łatwo znaleźć gotowe funkcje odpowiadające za zastosowanie tej metody.

Interpretacja wyników

W wyniku zastosowania funkcji sklearn.decomposition.PCA, służącej właśnie do zastosowania metody składowych głównych, otrzymujemy następujące wartości i wektory własne:

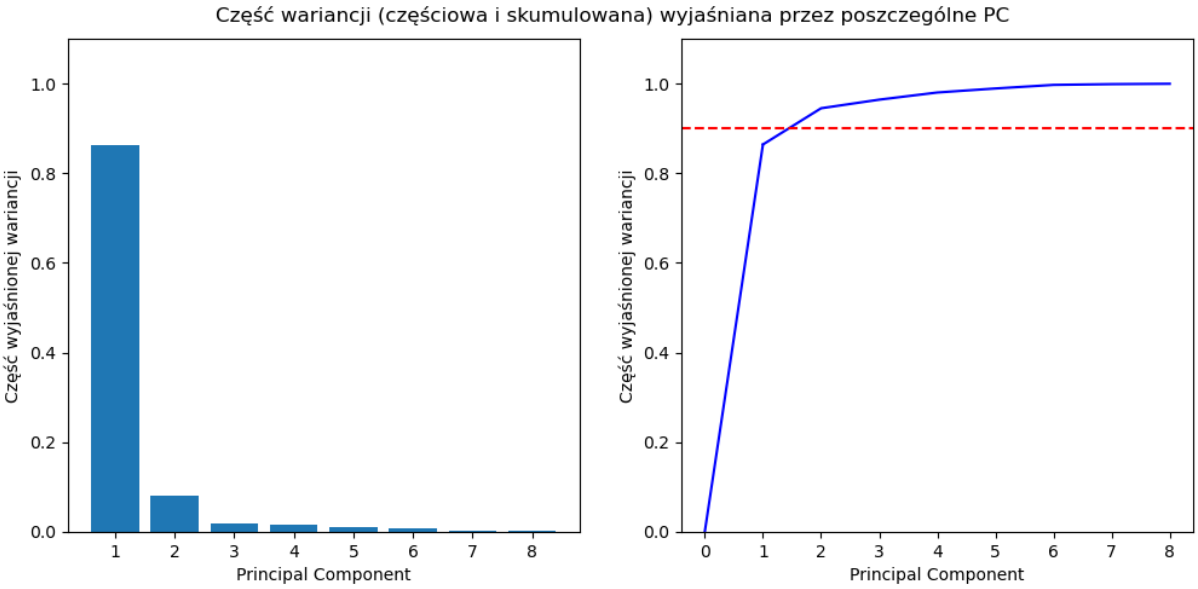
	STD
PC1	6.916791
PC2	0.647605
PC3	0.153803
PC4	0.127972
PC5	0.063873
PC6	0.072046
PC7	0.013351
PC8	0.006675

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
LongestShell	0.369300	-0.068230	0.228764	-0.544685	-0.141110	0.039477	0.698885	0.005979
Diameter	0.369785	-0.037922	0.262421	-0.526873	-0.088261	-0.047823	-0.710801	0.006788
Height	0.356319	0.044044	0.711507	0.547315	0.085361	0.240435	0.013286	0.005289
WholeWeight	0.373713	-0.146432	-0.281444	0.123552	0.140963	-0.044156	0.000329	0.850059
ShuckedWeight	0.360003	-0.274032	-0.360661	-0.018789	0.565598	0.435729	-0.020621	-0.393304
VisceraWeight	0.364631	-0.145845	-0.339282	0.271901	-0.768730	0.151322	-0.045981	-0.201923
ShellWeight	0.368933	-0.005240	-0.066093	0.189066	0.178013	-0.840785	0.059404	-0.285676
Rings	0.248158	0.934951	-0.209264	-0.029585	0.051037	0.129303	0.008714	-0.014777

Wartości te są trudne do zinterpretowania, co jest zdecydowaną wadą metody. Trzeba je interpretować na tle pozostałych wartości dla danej składowej. Możliwe jest też wyznaczenie dodatkowej macierzy, tzw. macierzy ładunków czynnikowych (loading factors), które przedstawiają korelację między początkowymi danymi a wyznaczonymi komponentami:

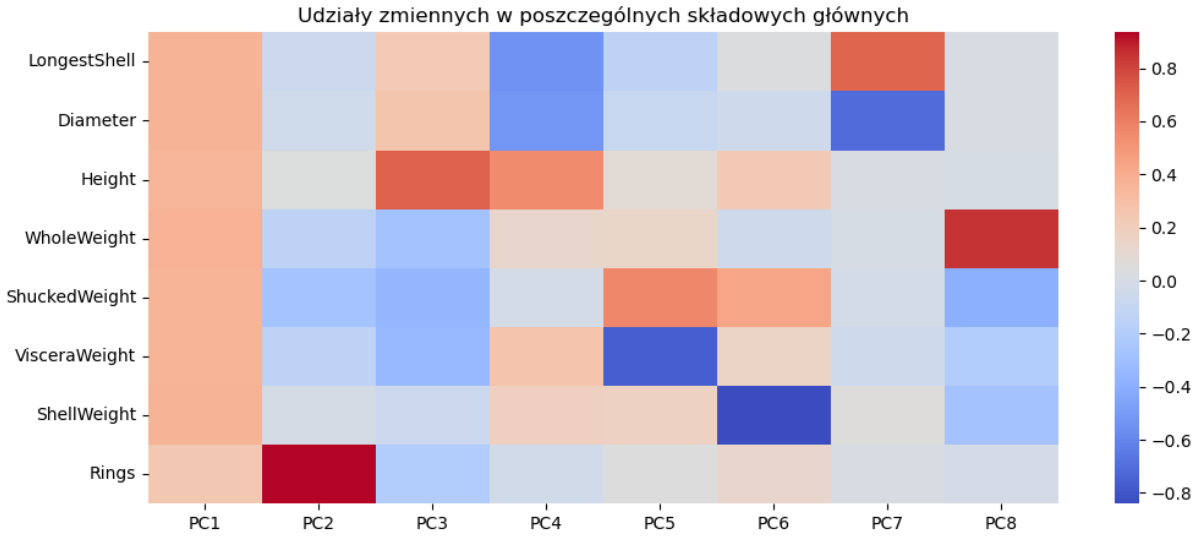
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
LongestShell	0.369300	-0.068230	0.228764	0.544685	-0.039477	0.141110	0.698885	-0.005979
Diameter	0.369785	-0.037922	0.262421	0.526873	0.047823	0.088261	-0.710801	-0.006788
Height	0.356319	0.044044	0.711507	-0.547315	-0.240435	-0.085361	0.013286	-0.005289
WholeWeight	0.373713	-0.146432	-0.281444	-0.123552	0.044156	-0.140963	0.000329	-0.850059
ShuckedWeight	0.360003	-0.274032	-0.360661	0.018789	-0.435729	-0.565598	-0.020621	0.393304
VisceraWeight	0.364631	-0.145845	-0.339282	-0.271901	-0.151322	0.768730	-0.045981	0.201923
ShellWeight	0.368933	-0.005240	-0.066093	-0.189066	0.840785	-0.178013	0.059404	0.285676
Rings	0.248158	0.934951	-0.209264	0.029585	-0.129303	-0.051037	0.008714	0.014777

W wielu miejscach widoczne jest podobieństwo między wektorami własnymi a loadingsami, nie są to jednak macierze jednakowe.
Analizując wyniki metody, w pierwszej kolejności chcielibyśmy sprawdzić ile komponentów tak naprawdę powinniśmy wziąć pod uwagę. Może nam do tego posłużyć wykres przedstawiający jaką część wariancji wyjaśniają poszczególne zmienne.

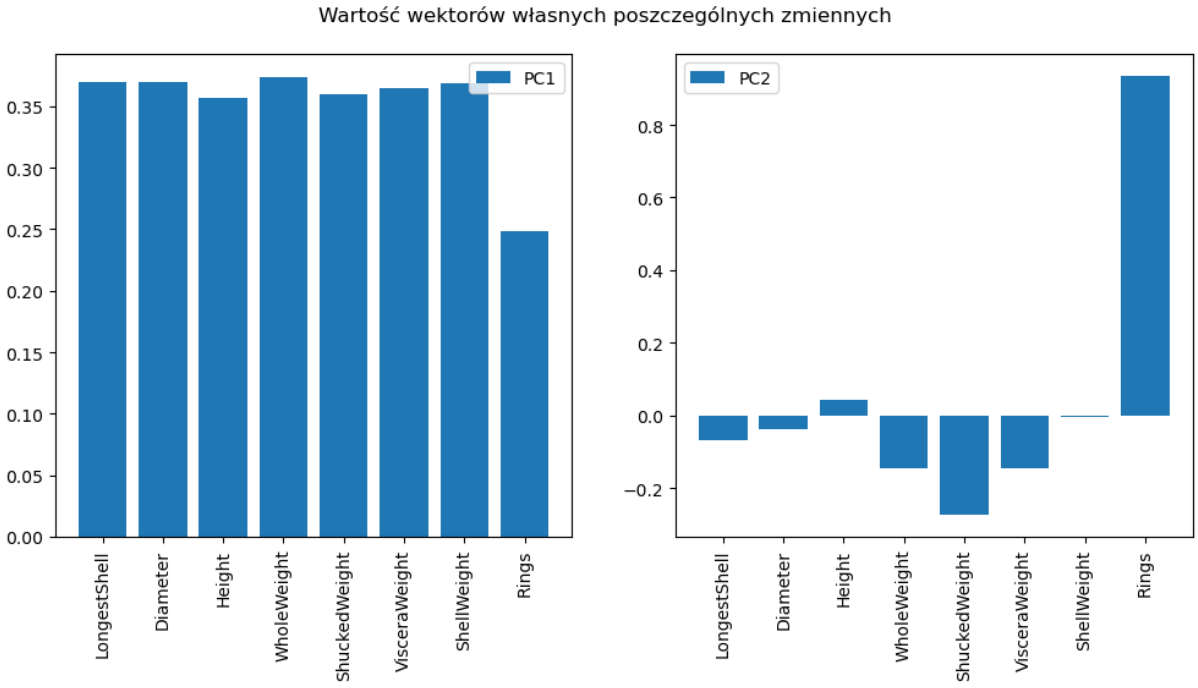


Pamiętając opis metody z poprzedniej części bardzo wysoki udział pierwszej zmiennej w wyjaśnianiu wariancji nie jest niczym dziwnym, a kolejne zmienne mają coraz mniejszy udział, co również nie jest zaskoczeniem. Wyboru komponentów można dokonać analizując skumulowane udziały w wyjaśnianej wariancji. Wybrane składowe powinny wyjaśniać około 80%-90% wariancji zbioru. W naszym przypadku odpowiada to dwóm pierwszym komponentom.

Zanim całkowicie skupimy się na pierwszych dwóch składowych, możemy sprawdzić jeszcze jak poszczególne zmienne wpływają na każdą ze składowych.

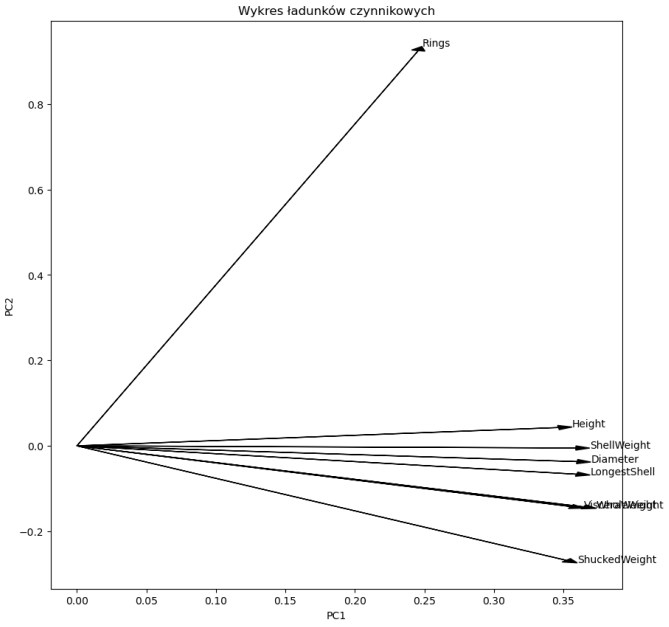


Jak widać nasza pierwsza składowa jest bardzo równomiernie wyjaśniana przez wszystkie ze zmiennych. Ciężko dopatrzeć się podobnego rozkładu w przypadku kolejnych komponentów. Szczególnie dla drugiej ze składowych bardzo duży nacisk pada na zmienną Rings, natomiast pozostałe zmienne wywierają znikomy wpływ.

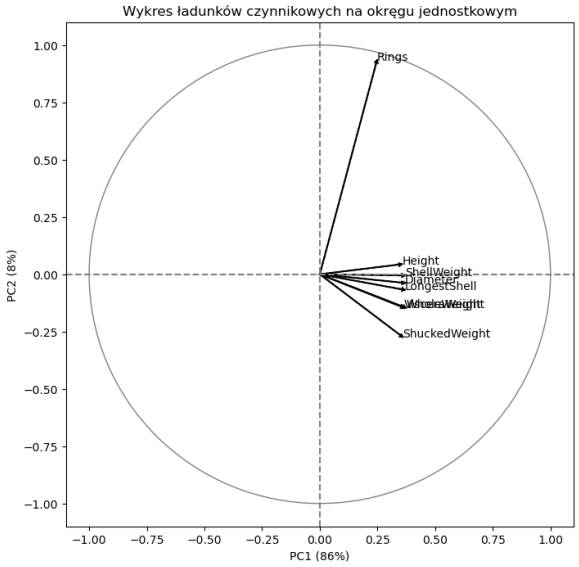


Na powyższym wykresie widać ten rozkład jeszcze lepiej. Po lewej pierwsza składowa, w przypadku której jedynie Rings odstają nieznacznie, natomiast pozostałe zmienne są na bardzo podobnym poziomie. Znaczy to, że poza pierścieniami na muszli, wartości zmiennych zmieniają się w bardzo podobnym tempie i kierunku. Ten komponent może być interpretowany jako wyznacznik tego, że badane osobniki podczas wzrostu rozwijają się równomiernie.

W przypadku drugiej składowej interpretacja jest dużo trudniejsza, ponieważ jedynie dwie zmienne wykazują tutaj wartości dodatnie, natomiast cała reszta wędruje poniżej osi. Może być ona wyznacznikiem tego, że liczba pierścieni, a więc wiek osobnika, niekoniecznie musi być związana z jego wzrostem, przynajmniej w przypadku zmiennych innych niż wysokość, co zdaje się być sensownym wnioskiem. Zwierzęta mogą różnić się wymiarami mimo podobnego wieku, natomiast jeśli chodzi o wysokość to musi ona rosnąć wraz z narastaniem kolejnych warstw tworzących pierścienie.



W przypadku wybrania zbioru składającego się z dwóch komponentów można przedstawić wyliczone wartości na wykresie, gdzie osie odpowiadają wartościom loadingów dla poszczególnych komponentów. Może on pełnić podobną funkcję do wykresów przedstawionych wyżej, natomiast w zależności od wyników może być łatwiejszy lub trudniejszy do zinterpretowania. Jego przewagą natomiast jest przedstawienie obu komponentów w jednej płaszczyźnie.



Wykres można również jeszcze bardziej doszlifować, wpisując go niejako w okrąg jednostkowy. Bardzo dobrze oddaje to skalę tego, jak mocno wartości loadingów odbiegają od siebie lub jak mocno są zbliżone. Na tym wykresie widać również bardzo dobitnie, że dla PC1 nie występują w zasadzie żadne wartości ujemne a dla PC2 nie są one bardzo wysokie.

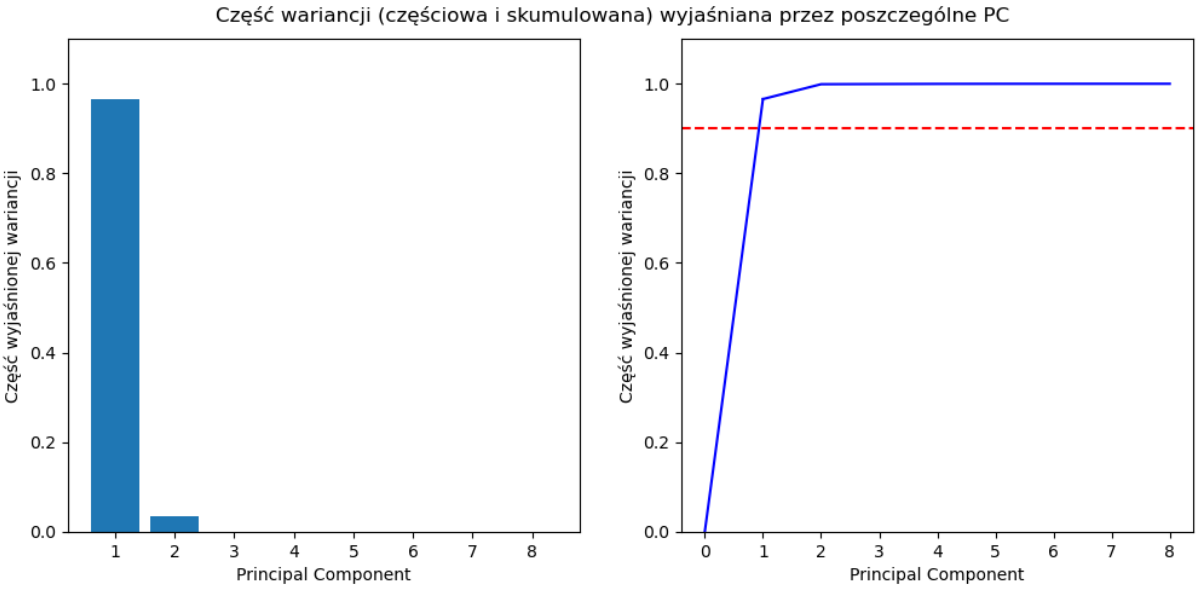
Porównanie z metodą stosowaną na danych niestandaryzowanych

Jednym z ważnych założeń PCA była praca na danych poddanych standaryzacji. Z tego powodu postanowiłem porównać wyniki uzyskane przez zastosowanie metody na danych zestandaryzowanych z wynikami uzyskanymi z danych przed standaryzacją.

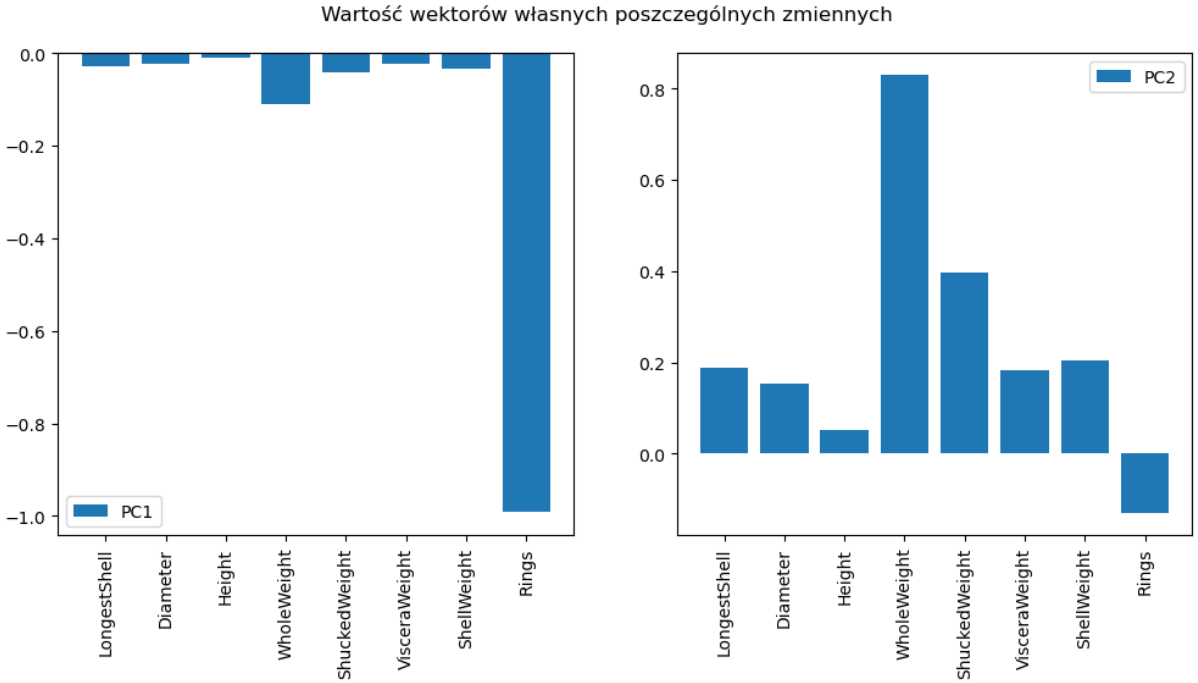
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
LongestShell	-0.028719	0.187924	-0.502539	0.543996	-0.118160	0.101306	-0.599412	-0.178503
Diameter	-0.024424	0.152306	-0.431975	0.416325	-0.013421	0.058702	0.779103	0.075161
Height	-0.009491	0.052852	-0.122534	0.024151	0.005974	-0.155391	-0.172574	0.963096
WholeWeight	-0.109977	0.829836	0.032185	-0.341335	-0.044472	0.422587	-0.001340	0.034250
ShuckedWeight	-0.042462	0.397857	0.603752	0.533375	0.207433	-0.383928	0.005165	-0.021119
VisceraWeight	-0.024382	0.182264	-0.072623	-0.167626	-0.773987	-0.567883	0.051330	-0.092905
ShellWeight	-0.033553	0.203982	-0.418144	-0.317881	0.584602	-0.560065	-0.035480	-0.157101
Rings	-0.991398	-0.130183	0.012881	0.003654	-0.001005	-0.000407	-0.000312	-0.001194

	STD
PC1	5.522339
PC2	0.189758
PC3	0.002157
PC4	0.001913
PC5	0.000822
PC6	0.000373
PC7	0.000133
PC8	0.000186

Pierwsze różnice widać już na poziomie własności i wektorów własnych. Nie są one jednak zbyt dobrze interpretowalne.



Kolejną różnicą jest całkowity brak istotnego wpływu komponentów dalszych niż dwa pierwsze. W poprzednim przypadku każda składowa w jakimś stopniu wyjaśniała wariancję, tutaj nie ma to miejsca.



Jeśli chodzi o wkład zmiennych w składowe to jest on skrajnie różny od tego, który widzieliśmy wcześniej. Pierwsza składowa o największej wariancji została zdominowana przez zmienną Rings. Nie jest to zadziwiający wynik biorąc pod uwagę, że wartości w tej zmiennej były w większości przypadków przynajmniej rząd wielkości większe. W przypadku drugiego komponentu najwyższą korelację widać między zmiennymi określającymi dwie różne wagi. Można to interpretować jako wskazanie na wpływ jednej z nich na drugą, jednak ciężko jest odpowiedzieć na pytanie czy wykazana zależność jest wystarczająco podobna do tej rzeczywistej. Ponadto ciężko nie zwrócić uwagi na znikome podkreślenie wpływu innych wymiarów zwierzęcia.
Na tym przykładzie można zauważyć, jak ważne jest spełnienie założenia o standaryzacji danych. Jeśli

nie chcemy aby nasze badanie zostało całkowicie zdominowane przez jedną ze zmiennych powinniśmy zawsze zwracać na nie uwagę.

Podsumowanie

Podsumowując uzyskane wyniki i wnioski, możemy z pewnością stwierdzić, że metoda składowych głównych może być bardzo przydatna. Mimo konieczności spełnienia pewnych założeń jest ona dość uniwersalna, wykorzystuje się ją również poza czystą analizą danych, np. przy przetwarzaniu sygnałów czy rozpoznawaniu obrazów. Działa ona również bardzo szybko ze względu na wykorzystanie macierzy i operacji na nich, a nie pożerających zasoby komputera algorytmów. Należy pamiętać jednak o konieczności spełnienia ww. założeń, w innym wypadku otrzymane wyniki nie będą poprawne. Ponadto metoda ta jest trudna w interpretacji a otrzymane wyniki mogą być niejednoznaczne.