Bartłomiej Wiśniewski

Badanie wpływu metody określania położenia punktu na wyniki

1. Informacja techniczna

Obliczenia były wykonywane przy użyciu Jupyter Notebook'a (Python 3.9) oraz biblioteki numpy w wersji 1.20.1. Maszyna wykonująca obliczenia wyposażona jest w procesor AMD Ryzen 7 3700U i działa pod kontrolą systemu operacyjnego Manjaro Linux 21.1.6 na jądrze w wersji 5.12.19-1. Wszystkie niecałkowite dane numeryczne zapisywane były w zmiennych typu numpy.float64 (podwójna precyzja).

2. Przebieg eksperymentu numerycznego

Eksperyment numeryczny przebiegał w kilku etapach:

1. Ustalenie tolerancji ϵ oznaczającej jak bardzo wynik może się różnić od oczekiwanego, aby został uznany za jemu równoważny. Obliczenia zostały przeprowadzone dla każdego z $\epsilon \in \{1e\text{-}4, 1e\text{-}8, 1e\text{-}12, 1e\text{-}14\}$
2. Wygenerowanie zbiorów danych:
   a) **dataset 1a:** 10^5 punktów o współrzędnych z przedziału [-1000, 1000]
   b) **dataset 1b:** 10^5 losowych punktów o współrzędnych z przedziału [-10^{14}, 10^{14}],
   c) **dataset 1c:** 1000 losowych punktów leżących na okręgu o środku (0,0) i promieniu R=100,
   d) **dataset 1d:** 1000 losowych punktów o współrzędnych z przedziału [-1000, 1000] leżących na prostej wyznaczonej przez wektor $AB = (a, b)$, gdzie $a = [-1.0, 0.0]$, $b = [1.0, 0.1]$.

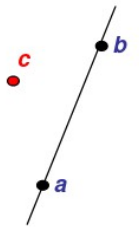
W podpunktach a) i b) punkty współrzędne były generowane z użyciem następującego kodu $numpy.float64(random.uniform(\langle minimalna\ wartość\ współrzędnej \rangle, \langle maksymalna\ wartość\ współrzędnej \rangle))$.

W podpunkcie c) generowany był parametr $t \in [0; 1)$ przy pomocy $numpy.random.uniform(1, 0)$, a następnie obliczano współrzędne punktów ze wzorów:

$$x = \cos(2\pi t)$$
$$y = \sin(2\pi t)$$

W podpunkcie d) generowana była współrzędna x (w ten sam sposób co w podpunkcie a)) i na jej podstawie obliczana była współrzędna y według wzoru $y=(\frac{y_B-y_A}{x_B-x_A})*(x-x_A)+y_A$. Punkt był dodawany do zbioru jeżeli $x, y \in [-1000, 1000]$

3. Zaklasyfikowanie każdego z punktów w każdym ze zbiorów do jednej z kategorii: LEFT, COLLINEAR, RIGHT w przypadkach kiedy kolejno: punkt leży na lewo od wektora AB, leży wystarczająco blisko prostej wyznaczonej przez ten wektor, aby uznać go za współliniowy i RIGHT, gdy leży na prawo od tego wektora. Klasyfikacja odbywała się czterema metodami wykorzystującymi wyznaczniki macierzy:



$$det(a,b,c)=\begin{vmatrix} a_x & a_y & 1 \\ b_x & b_y & 1 \\ c_x & c_y & 1 \end{vmatrix} \quad (1)$$

$$det(a,b,c)=\begin{vmatrix} a_x-c_x & a_y-c_y \\ b_x-c_x & b_y-c_y \end{vmatrix} \quad (2)$$

a) **Metoda A:** wzór (1) i własna implementacja liczenia wyznacznika

b) **Metoda B:** wzór (2) i własna implementacja liczenia wyznacznika

c) **Metoda C:** wzór (1) i biblioteczna implementacja liczenia wyznacznika (numpy.linalg.det)

d) **Metoda D:** wzór (2) i biblioteczna implementacja liczenia wyznacznika (numpy.linalg.det)

Jeżeli obliczony wyznacznik spełniał nierówność $|det(a,b,c)|<\epsilon$ to punkt był klasyfikowany jako COLLINEAR. W przeciwnym wypadku jeżeli $det(a,b,c)<0$ to punkt był klasyfikowany jako LEFT, a jeżeli $det(a,b,c)>0$ to jako RIGHT.

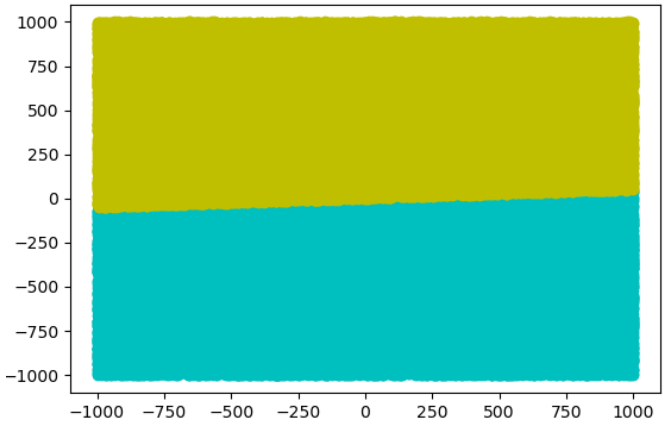
4. Porównanie parami wyników klasyfikacji każdą z tych metod.
5. Zmierzenie czasu klasyfikacji.

3. Wyniki

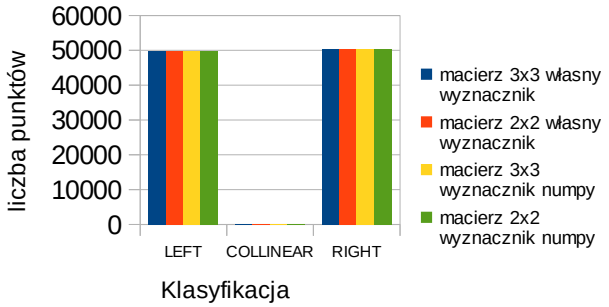
1. ϵ = 1e-4

a) klasyfikacja zbioru punktów dataset 1a

Metoda klasyfikacji punktów	LEFT	COLLINEAR	RIGHT
macierz 3x3 własny wyznacznik	49767	0	50233
macierz 2x2 własny wyznacznik	49767	0	50233
macierz 3x3 wyznacznik numpy	49767	0	50233
macierz 2x2 wyznacznik numpy	49767	0	50233



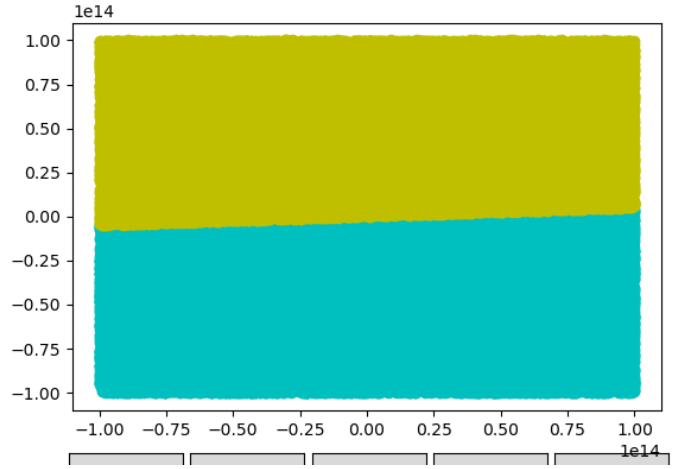
Rysunek 1: *Wizualizacja klasyfikacji punktów z dataset 1a*



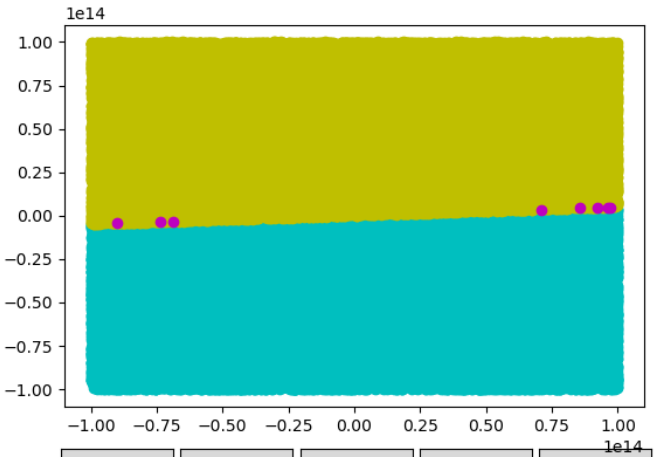
Wykres 1: *Klasyfikacja punktów z dataset 1a dla* $\epsilon = 1e\text{-}4$

b) klasyfikacja zbioru punktów dataset 1b

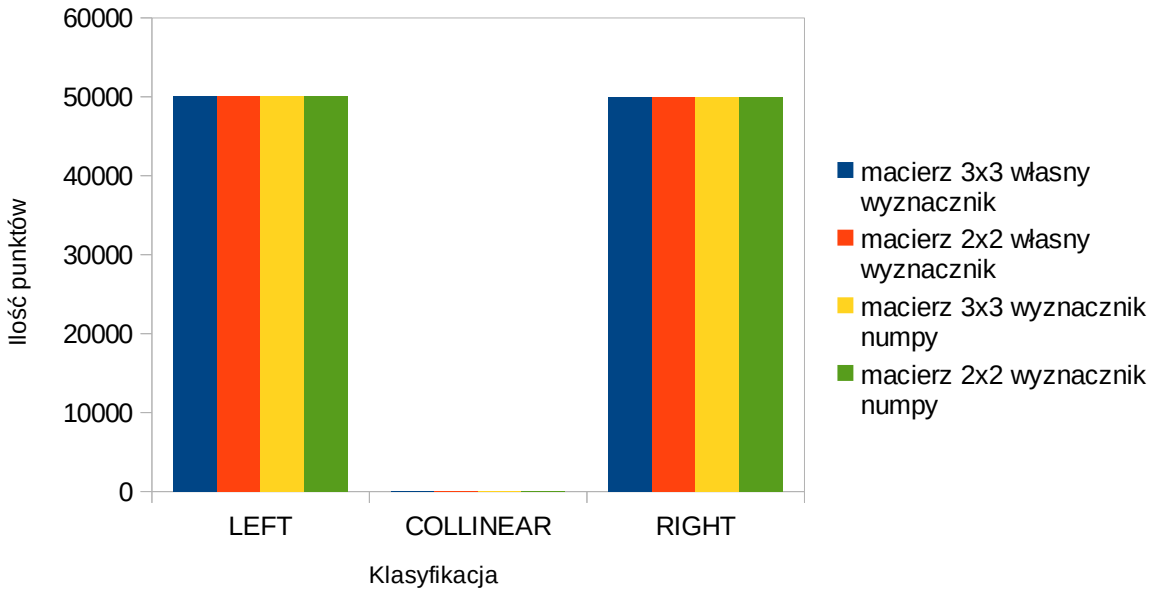
Metoda klasyfikacji punktów	LEFT	COLLINEAR	RIGHT
macierz 3x3 własny wyznacznik	50013	0	49987
macierz 2x2 własny wyznacznik	50007	8	49985
macierz 3x3 wyznacznik numpy	50013	0	49987
macierz 2x2 wyznacznik numpy	50008	10	49982



Rysunek 2: *Wizualizacja podziału punktów z dataset 1b metodami A i C*



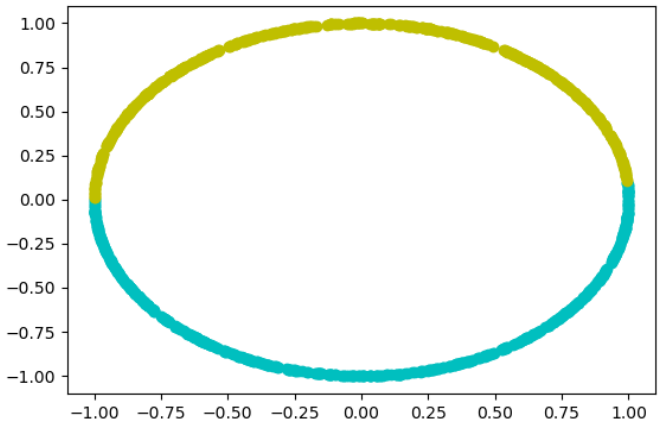
Rysunek 3: *Wizualizacja klasyfikacji punktów z dataset 1b metodą B*

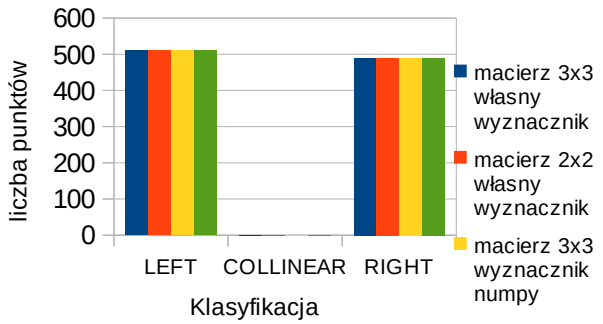


Wykres 2: *Klasyfikacja punktów z dataset 1b dla* $\epsilon = 1e\text{-}4$

c) klasyfikacja zbioru punktów dataset 1c

Metoda klasyfikacji punktów	LEFT	COLLINEAR	RIGHT
macierz 3x3 własny wyznacznik	510	0	490
macierz 2x2 własny wyznacznik	510	0	490
macierz 3x3 wyznacznik numpy	510	0	490
macierz 2x2 wyznacznik numpy	510	0	490



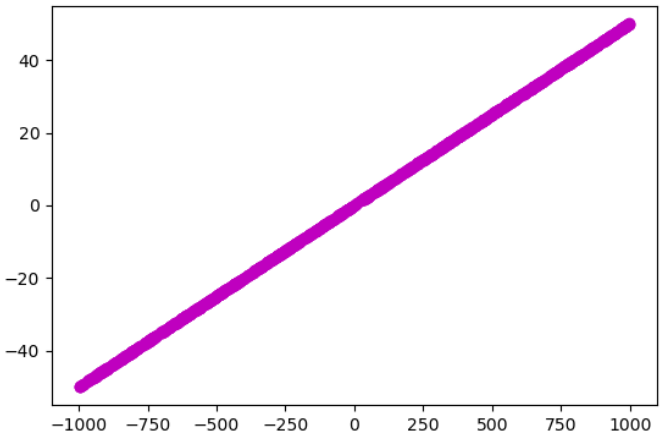


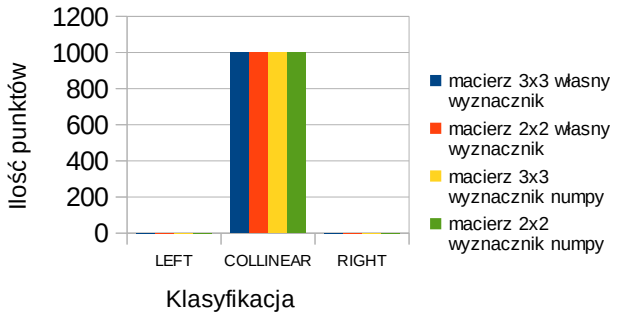
Wykres 3: *Klasyfikacja punktów z dataset 1c dla* $\epsilon = 1e\text{-}4$

d) klasyfikacja zbioru punktów dataset 1d

Metoda klasyfikacji punktów	LEFT	COLLINEAR	RIGHT
macierz 3x3 własny wyznacznik	0	1000	0
macierz 2x2 własny wyznacznik	0	1000	0
macierz 3x3 wyznacznik numpy	0	1000	0

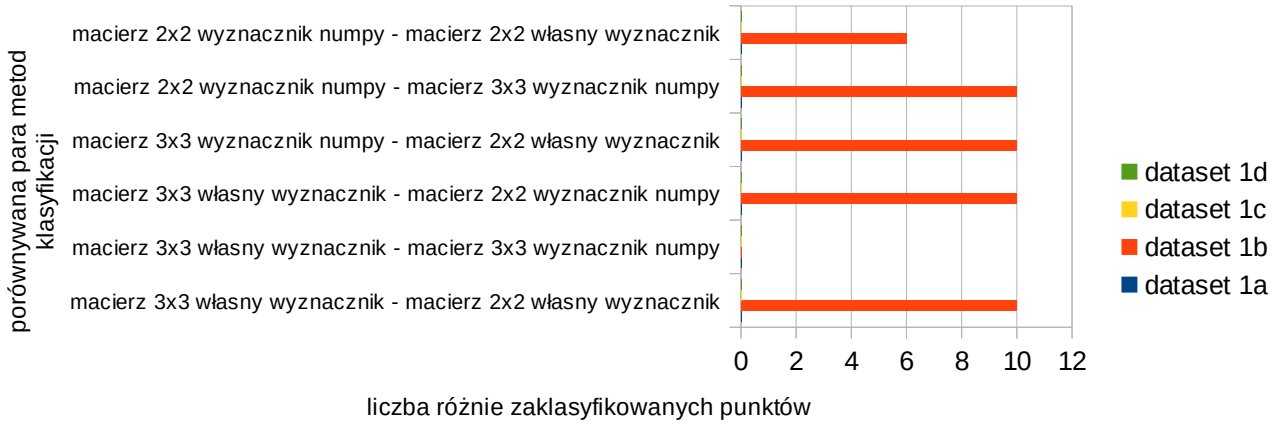
macierz 2x2 wyznacznik numpy	0	1000	0

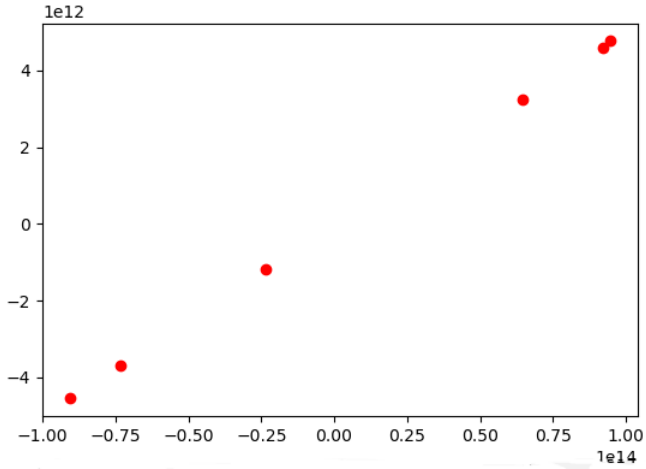




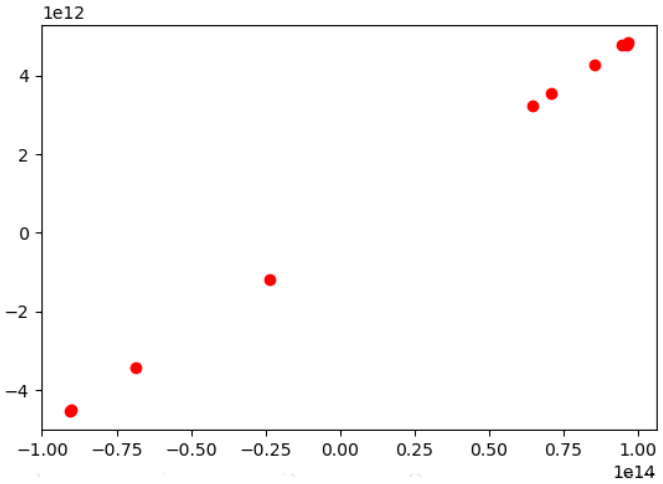
Wykres 4: *Klasyfikacja punktów z dataset 1d dla* $\epsilon = 1e\text{-}4$

e) porównanie klasyfikacji przez różne metody





Rysunek 4: *Punkty które zostały różnie zaklasyfikowane przez metody B i D dla dataset 1b*

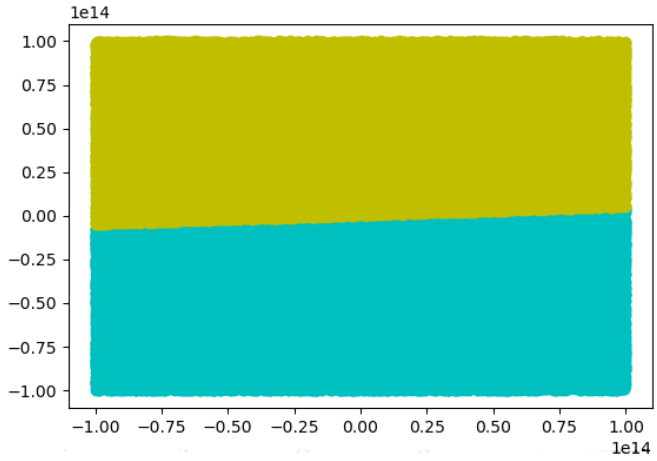


Rysunek 5: *Punkty które zostały różnie zaklasyfikowane przez metody A i D dla dataset 1b*

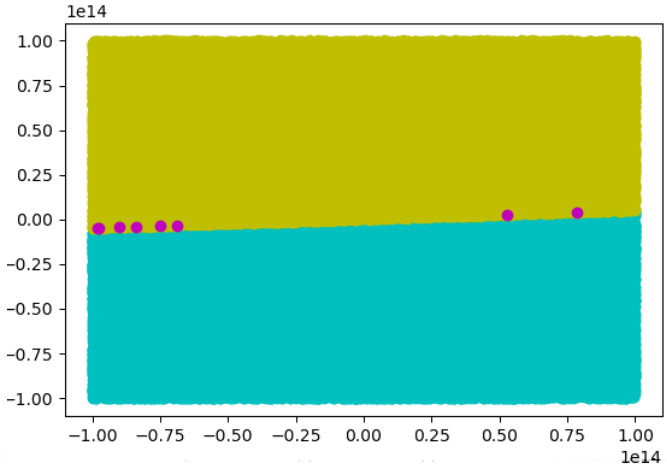
Różnice klasyfikacji przez pozostałe pary metod klasyfikacji wyglądają podobnie.

2. ϵ = 1e-8
   a) klasyfikacja zbioru punktów dataset 1a
      Punkty zostały zaklasyfikowane identycznie jak w przypadku ϵ = 1e-4 .
   b) klasyfikacja zbioru punktów dataset 1b

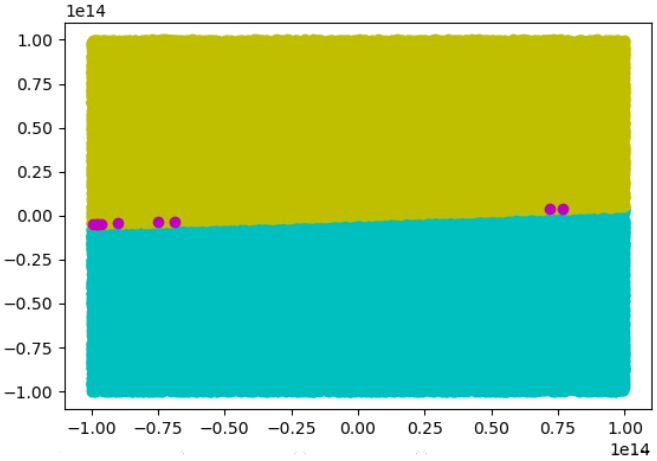
Metoda klasyfikacji punktów	LEFT	COLLINEAR	RIGHT
macierz 3x3 własny wyznacznik	50013	0	49987
macierz 2x2 własny wyznacznik	50010	8	49982
macierz 3x3 wyznacznik numpy	50013	0	49987
macierz 2x2 wyznacznik numpy	50010	9	49981



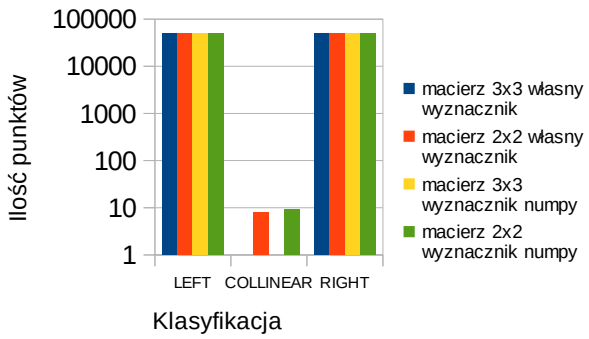
Rysunek 6: *Wizualizacja podziału punktów z dataset 1b metodą A*



Rysunek 7: *Wizualizacja podziału punktów z dataset 1b metodą B*



Rysunek 8: *Wizualizacja podziału punktów z dataset 1b metodą D*



Wykres 5: *Klasyfikacja punktów z dataset 1b.*
Uwaga: *skala logarytmiczna*

c) klasyfikacja zbioru punktów dataset 1c
Punkty zostały zaklasyfikowane identycznie jak w przypadku $\epsilon = 1e\text{-}4$.

d) klasyfikacja zbioru punktów dataset 1d
Punkty zostały zaklasyfikowane identycznie jak w przypadku $\epsilon = 1e\text{-}4$.

e) porównanie klasyfikacji przez różne metody

3. $\epsilon = 1e\text{-}12$ (zbiory punktów zostały wygenerowane na nowo w identyczny sposób)

a) klasyfikacja zbioru punktów dataset 1a

Metoda klasyfikacji punktów	LEFT	COLLINEAR	RIGHT
macierz 3x3 własny wyznacznik	49816	0	50184
macierz 2x2 własny wyznacznik	49816	0	50184
macierz 3x3 wyznacznik numpy	49816	0	50184
macierz 2x2 wyznacznik numpy	49816	0	50184

b) klasyfikacja zbioru punktów dataset 1b

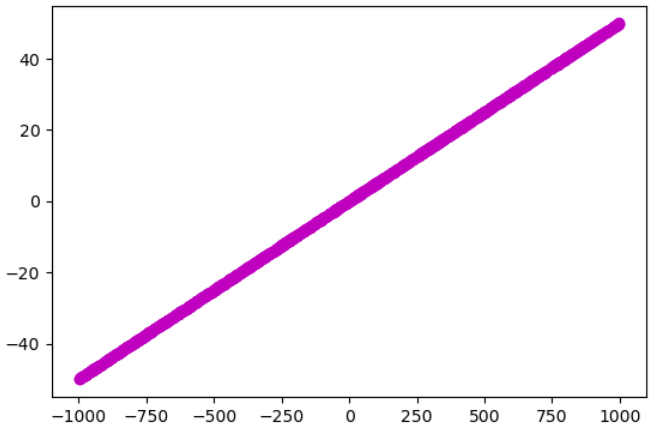
Metoda klasyfikacji punktów	LEFT	COLLINEAR	RIGHT
macierz 3x3 własny wyznacznik	50104	0	49896
macierz 2x2 własny wyznacznik	50101	7	49892
macierz 3x3 wyznacznik numpy	50104	0	49896
macierz 2x2 wyznacznik numpy	50100	8	49892

c) klasyfikacja zbioru punktów dataset 1c

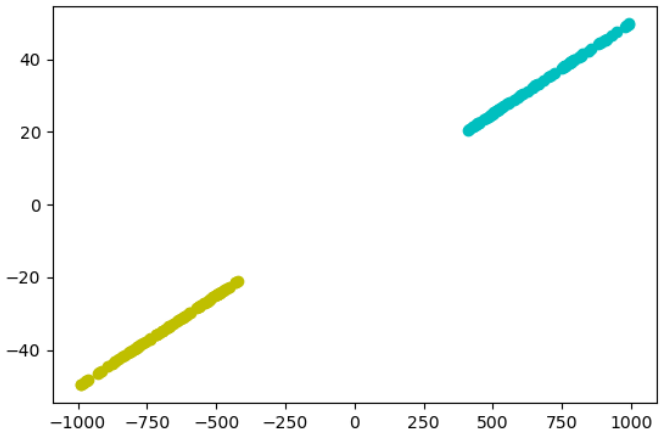
Metoda klasyfikacji punktów	LEFT	COLLINEAR	RIGHT
macierz 3x3 własny wyznacznik	524	0	476
macierz 2x2 własny wyznacznik	524	0	476
macierz 3x3 wyznacznik numpy	524	0	476
macierz 2x2 wyznacznik numpy	524	0	476

d) klasyfikacja zbioru punktów dataset 1d

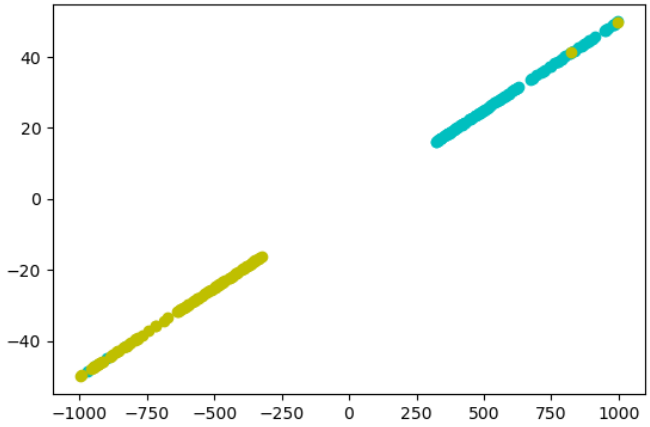
Metoda klasyfikacji punktów	LEFT	COLLINEAR	RIGHT
macierz 3x3 własny wyznacznik	0	1000	0
macierz 2x2 własny wyznacznik	79	839	82
macierz 3x3 wyznacznik numpy	0	1000	0
macierz 2x2 wyznacznik numpy	121	781	98



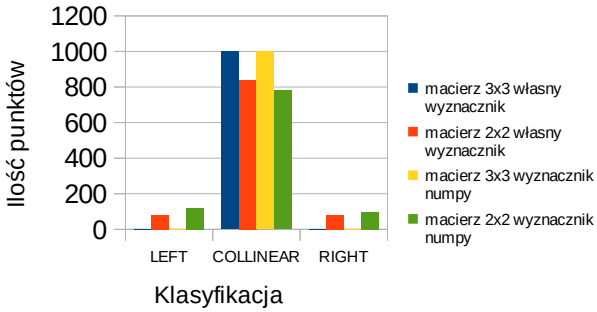
Rysunek 9: *Wizualizacja podziału punktów z dataset 1d metodami A i C*



Rysunek 10: *Wizualizacja podziału punktów z dataset 1d metodą B (bez punktów COLLINEAR)*

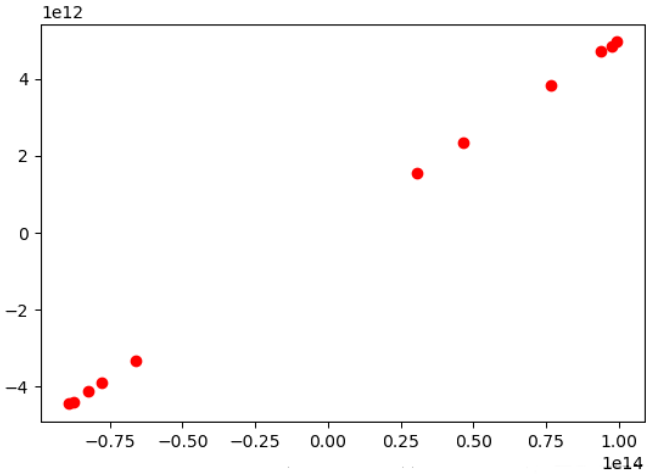


Rysunek 11: *Wizualizacja podziału punktów z dataset 1d metodą D (bez punktów COLLINEAR)*

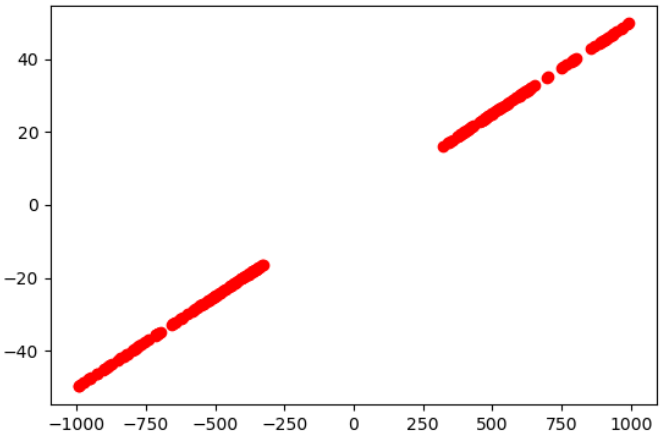


Wykres 6: *Klasyfikacja punktów z dataset 1d*

e) porównanie klasyfikacji przez różne metody

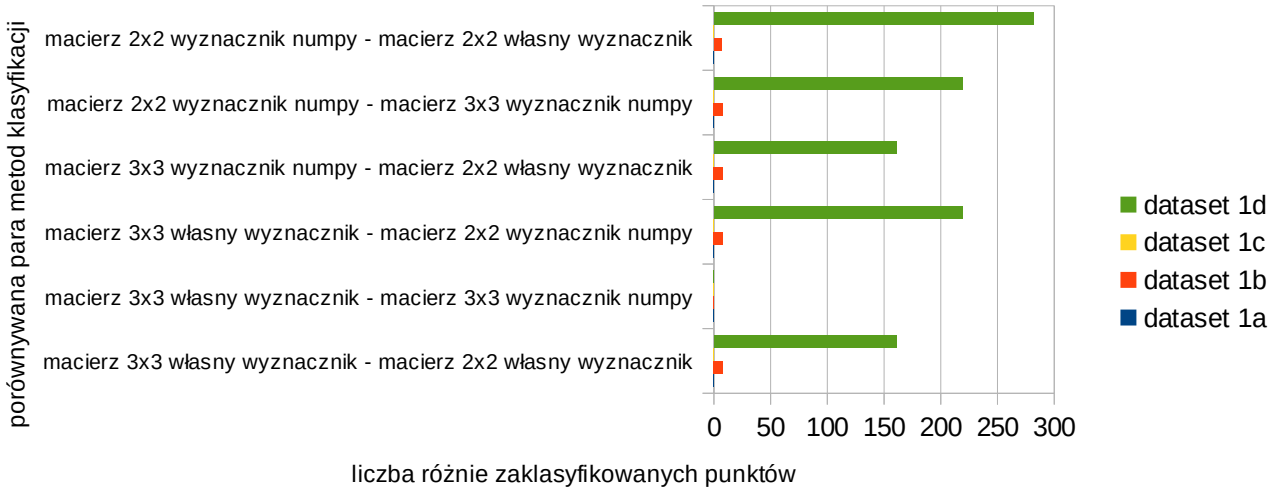


Rysunek 12: *Punkty które zostały różnie zaklasyfikowane przez metody A i D dla dataset 1b*



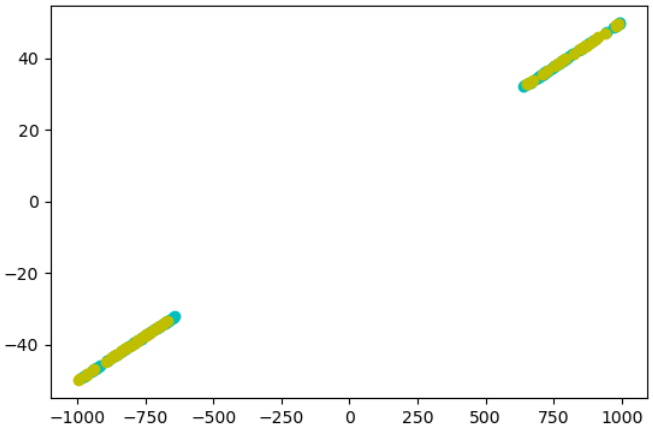
Rysunek 13: *Punkty które zostały różnie zaklasyfikowane przez metody B i C dla dataset 1d*

Różnice w klasyfikacji punktów dla pozostałych par metod klasyfikacji (za wyjątkiem A i C) wyglądają podobnie

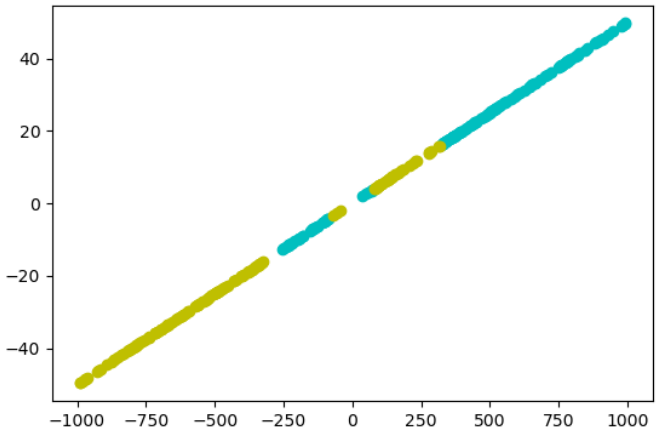


4. ϵ = 1e-14
   a) klasyfikacja zbioru punktów dataset 1a
      Punkty zostały zaklasyfikowane identycznie jak w przypadku ϵ = 1e-12 .
   b) klasyfikacja zbioru punktów dataset 1b
      Punkty zostały zaklasyfikowane identycznie jak w przypadku ϵ = 1e-12 .
   c) klasyfikacja zbioru punktów dataset 1c
      Punkty zostały zaklasyfikowane identycznie jak w przypadku ϵ = 1e-12 .
   d) klasyfikacja zbioru punktów dataset 1d

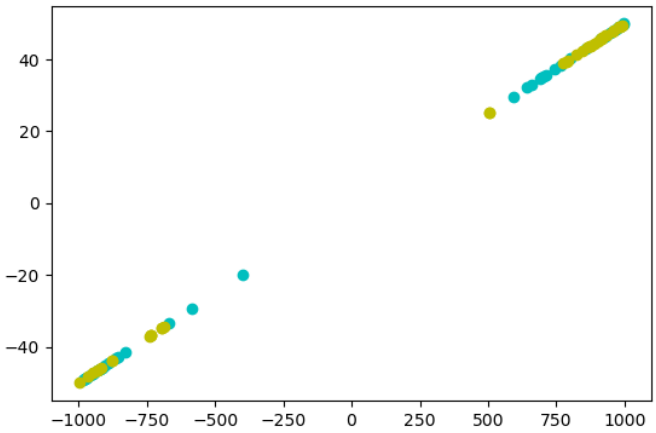
Metoda klasyfikacji punktów	LEFT	COLLINEAR	RIGHT
macierz 3x3 własny wyznacznik	105	827	68
macierz 2x2 własny wyznacznik	145	712	143
macierz 3x3 wyznacznik numpy	41	921	38
macierz 2x2 wyznacznik numpy	172	688	140



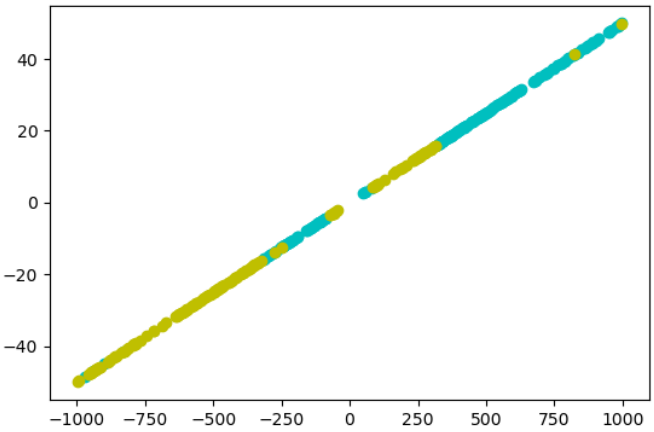
Rysunek 14: *Wizualizacja podziału punktów z dataset 1d metodą A (bez punktów COLLINEAR)*



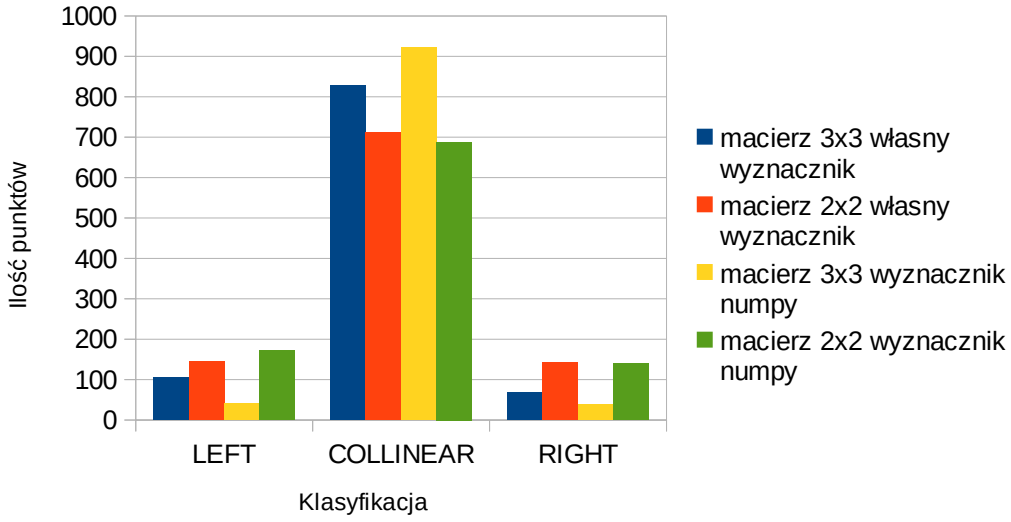
Rysunek 15: *Wizualizacja podziału punktów z dataset 1d metodą B (bez punktów COLLINEAR)*



Rysunek 16: *Wizualizacja podziału punktów z dataset 1d metodą C (bez punktów COLLINEAR)*

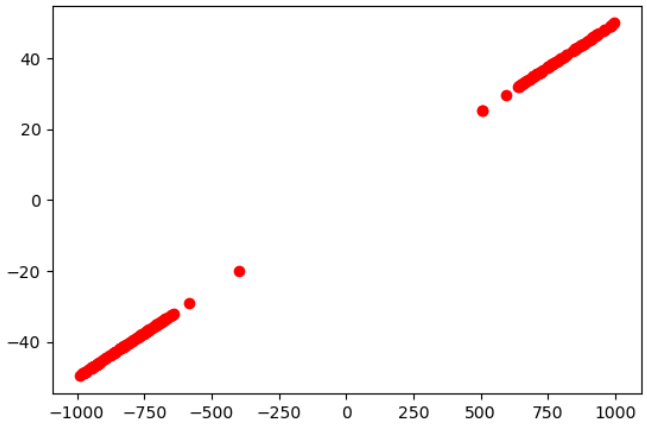


Rysunek 17: *Wizualizacja podziału punktów z dataset 1d metodą D (bez punktów COLLINEAR)*

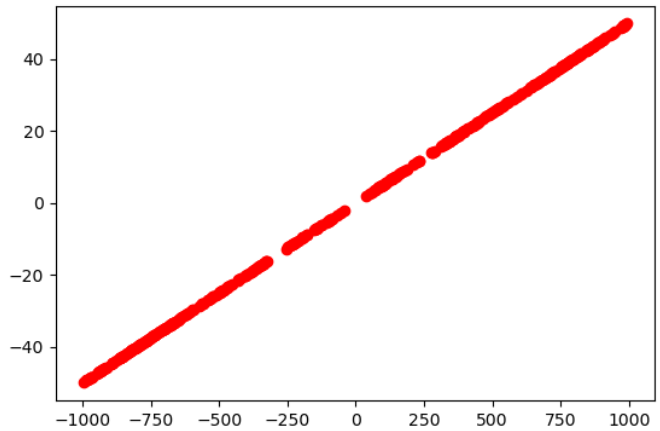


Wykres 7: *Klasyfikacja punktów z dataset 1d dla* $\epsilon = 1e\text{-}14$

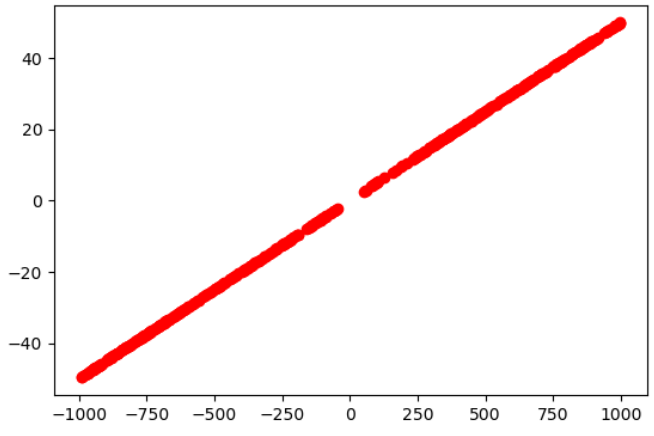
e) porównanie klasyfikacji przez różne metody



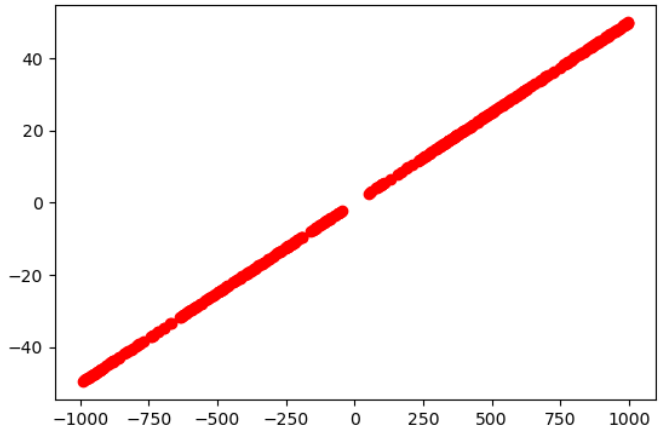
Rysunek 18: *Punkty które zostały różnie zaklasyfikowane przez metody A i C dla dataset 1d*



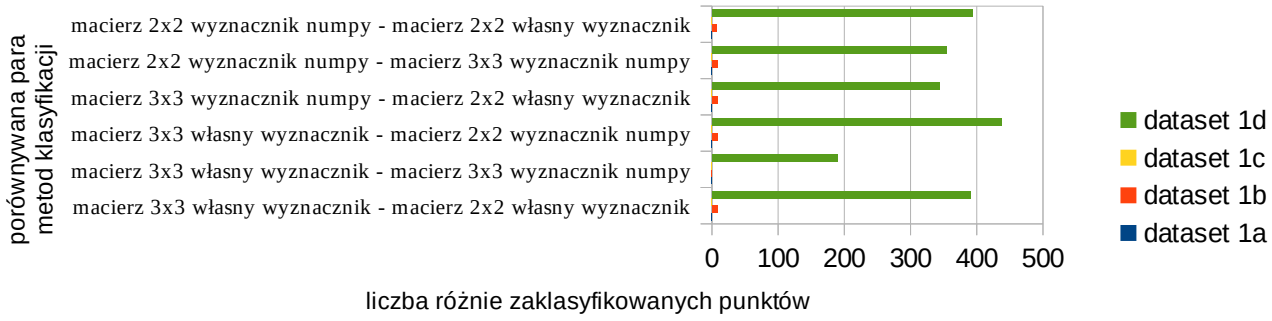
Rysunek 19: *Punkty które zostały różnie zaklasyfikowane przez metody A i B dla dataset 1d*



Rysunek 20: *Punkty które zostały różnie zaklasyfikowane przez metody A i D dla dataset 1d*



Rysunek 21: *Punkty które zostały różnie zaklasyfikowane przez metody C i D dla dataset 1d*



5. Czasy klasyfikacji
   Zmierzono ile czasu każda z metod klasyfikacji potrzebuje na klasyfikację każdego z punktów w danym zbiorze danych. Pomiary powtórzono 10 razy i obliczono średnią z tych pomiarów.

Metoda klasyfikacji punktów	Średnia czasu klasyfikacji (ns)
macierz 3x3 własny wyznacznik	21156315.5
macierz 2x2 własny wyznacznik	6911365.1
macierz 3x3 wyznacznik numpy	12199764.7
macierz 2x2 wyznacznik numpy	10435817.2

4. Wnioski

Pierwszą nasuwającą się obserwacją jest to, że im mniejsza jest wartość tolerancji ϵ tym więcej punktów zostaje różnie zakwalifikowanych przez różne (matematycznie równoważne) metody. Wniosek jest taki, że należy bardzo rozważnie dobierać parametr ϵ, ponieważ zbyt mała wartość może powodować zbyt silne uzależnienie wyniku od wybranej metody obliczeń, a zbyt duża wartość może spowodować skrajnie nieprecyzyjne wyniki.

Ciekawą obserwacją jest też zgodność metod A i C. Może to sugerować, podobną implementację algorytmów liczenia wyznacznika z macierzy.

Jednak najciekawszą obserwacją jest całkowity brak spójności wyników w przypadku $\epsilon = 1\text{e-}14$ dla zbioru punktów dataset 1d. Mimo, że wszystkie punkty znajdują się na prostej służącej nam do klasyfikacji (pod względem algebraicznym) to średnio ponad 200 punktów zostało zaklasyfikowane inaczej niż COLLINEAR, a porównanie klasyfikacji różnymi metodami pokazuje, że metody klasyfikacji nie są zgodne w przypadku średnio 35% wszystkich punktów. Jeżeli wziąć pod uwagę skończoną precyzję obliczeń podczas wyliczania drugiej współrzędnej danego punktu to można dojść do wniosku, że to właśnie te obliczenia wprowadzają błąd wystarczająco duży, aby punkt później został zakwalifikowany jako niewspółliniowy.

Dataset 1d pozwala zauważyć, że przy skrajnie małych wartościach ϵ metody klasyfikacji wykorzystujące macierz rozmiaru 3x3 wydają się znacznie bardziej tolerancyjne na błędy obliczeniowe niż te wykorzystujące macierz rozmiaru 2x2. Z kolei przy większych wartościach ϵ dla dataset 1b metody wykorzystujące macierz 2x2 są ze sobą zgodne i klasyfikują kilka punktów jako COLLINEAR podczas gdy inne metody tego nie robią. Sporne punkty znajdują się względnie daleko od punktu $(0,0)$. Wektor służący do klasyfikacji jest bardzo krótki i znajduje się blisko punktu $(0,0)$. Pierwszym krokiem metod wykorzystujących macierz 3x3 jest wykonanie odejmowania odpowiednich współrzędnych dwóch punktów. Tutaj odejmowane są liczby bardzo duże i małe (dalekie od zera i bliskie 0). Duże liczby mają małą precyzję po przecinku, a małe liczby wprost przeciwnie. Wynikiem odejmowania jest wciąż duża liczba, tak więc precyzja z jaką zapisana była mała liczba zostaje utracona. Przez to

reszta obliczeń wykonywana jest z dużym błędem. To może tłumaczyć dlaczego metody te są zgodne ze sobą, ale nie z pozostałymi. Wniosek jest taki, że w przypadku danych o dużym rozrzucie lepiej stosować metody A i C, a w przypadku danych o małym rozrzucie metody B i D.

Metoda B ma najniższy średni czas klasyfikacji co w połączeniu z jej dokładnością dla danych o małym rozrzucie prowadzi do wniosku, że w większości przypadków jest najlepszym wyborem z rozważanych tutaj metod.