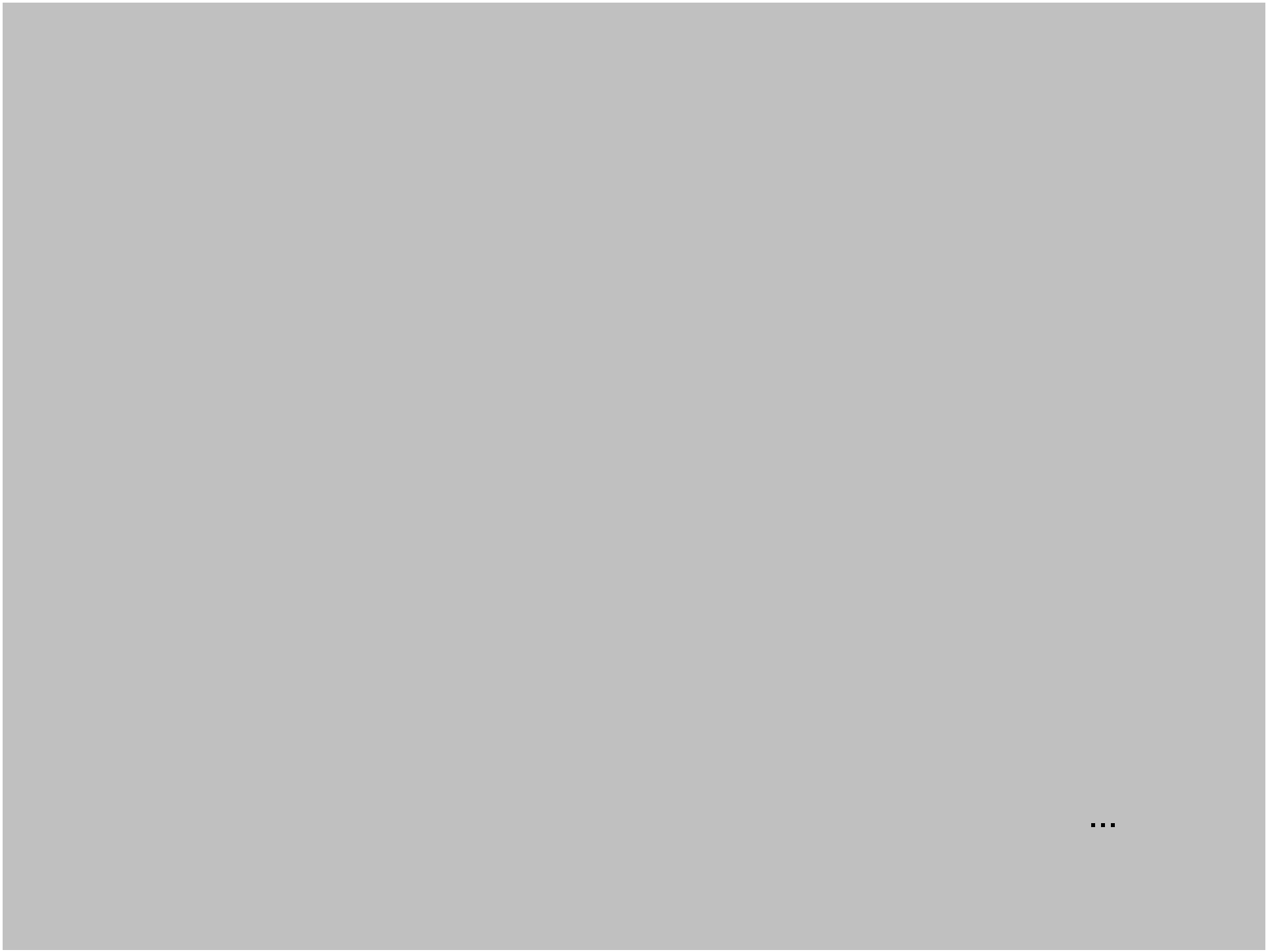


Robert Susmaga
Instytut Informatyki
ul. Piotrowo 2
Poznań

kontakt mail'owy
Robert.Susmaga@CS.PUT.Poznan.PL
kontakt osobisty
Centrum Wykładowe, „blok informatyki", pok. 7

Wyłączenie odpowiedzialności

Prezentowane materiały, będące dodatkiem pomocniczym do wykładów, z konieczności fragmentarycznym i niedopracowanym, należy wykorzystywać z pełną świadomością faktu, że mogą nie być pozbawione przypadkowych błędów, braków, wypaczeń i przeinaczeń :-)

Autor

...

Wizualizacja danych wielowymiarowych

- Prezentowana metoda
  - skalowanie wielowymiarowe (ang. multidimensional scaling, MDS)
  - dokonuje wizualizacji wielowymiarowych konfiguracji obiektów

Wizualizacja danych wielowymiarowych

- Przykład przewodni: dane o 15 modelach samochodów osobowych (tzw. obiekty, przykłady, obserwacje) opisanych w kategoriach 4 parametrów (tzw. zmienne, atrybuty)
  - zużycie paliwa
  - przyspieszenie
  - cena
  - wyposażenie
- Założenie: wszystkie parametry są traktowane jako równie ważne

Wizualizacja danych wielowymiarowych

- Opisy wybranych 3 samochodów:
  - Mitsubishi Carisma: zuż.=7,2; przysp.=12,0; cena=60,60; wyposaż.=9
  - Opel Astra II: zuż.=7,0; przysp.=12,0; cena=56,95; wyposaż.=8
  - Saab 9 3S: zuż.=9,7; przysp.=11,0; cena=85,35; wyposaż.=8
- Przykładowa analiza

  (nieskomplikowana /wobec niewielkiej liczby opisywanych obiektów/)
  - wniosek 1: „Carisma” bardzo zbliżony do „Astra II”
  - wniosek 2: „9 3S” mocno różny zarówno od „Carisma” jak i od „Astra II”

Wizualizacja danych wielowymiarowych

- Opisy wszystkich 15 samochodów (tabela informacyjna)

Marka	Model	Zużycie	Przyspieszenie	Cena	Wyposażenie
Alfa Romeo	156	8,1	9,3	71,14	9
Audi	A4	7,9	11,9	93,35	10
BMW	316I	7,5	12,3	81,79	8
Daewoo	Lanos	8,4	12,2	34,90	3
Honda	Civic	6,7	10,8	48,90	7
Hyunday	Accent	6,4	11,7	35,30	2
Lada	Samara	7,2	13,0	24,90	2
Mitsubishi	Carisma	7,2	12,0	60,60	9
Opel	Astra II	7,0	12,0	56,95	8
Peugeot	206 XR	6,6	13,2	38,36	4
Renault	Megane	7,0	9,9	50,05	9
Saab	9 3S	9,7	11,0	85,35	8
Seat	Cordoba	8,3	10,9	44,99	8
Toyota	Corrola	7,7	10,2	50,36	4
Volkswagen	Golf IV	8,3	9,9	61,62	6

Wizualizacja danych wielowymiarowych

- Przykładowa analiza

  (zdecydowanie trudniejsza ze względu na dużą liczbę obiektów)

  – oczywiście „Carisma” bardzo zbliżona do „Astra II”, ale:
    - czy to jedyna para obiektów o mocno zbliżonych opisach?
      – jeżeli istnieją inne takie pary, to jakie?
    - które obiekty są najbardziej zbliżone?

  – oczywiście „9 3S” mocno różny od „Carisma” i „Astra II”, ale:
    - czy to jedyna para obiektów o mocno zróżnicowanych opisach?
      – jeżeli istnieją inne takie pary, to jakie?
    - które obiekty są najbardziej zróżnicowane?

Wizualizacja danych wielowymiarowych

- Podejście bezpośrednie: utworzenie wykresu rozrzutu
  - niemożliwe ze względu na liczbę zmiennych (przekraczającą 3)

Wizualizacja danych wielowymiarowych

- Potencjalne podejście pośrednie: wprowadzenie miary odległości (najlepiej na znormalizowanych danych) i zobrazowanie tych odległości
  - dla każdej pary obiektów (bez względu na liczbę parametrów je opisujacych) odległość pomiędzy nimi wyraża się pojedynczą wartością skalarną

- Opisy wszystkich 15 samochodów (tabela informacyjna)

Marka	Model	Zużycie	Przyspieszenie	Cena	Wyposażenie
Alfa Romeo	156	8,1	9,3	71,14	9
Audi	A4	7,9	11,9	93,35	10
BMW	316I	7,5	12,3	81,79	8
Daewoo	Lanos	8,4	12,2	34,90	3
Honda	Civic	6,7	10,8	48,90	7
Hyunday	Accent	6,4	11,7	35,30	2
Lada	Samara	7,2	13,0	24,90	2
Mitsubishi	Carisma	7,2	12,0	60,60	9
Opel	Astra II	7,0	12,0	56,95	8
Peugeot	206 XR	6,6	13,2	38,36	4
Renault	Megane	7,0	9,9	50,05	9
Saab	9 3S	9,7	11,0	85,35	8
Seat	Cordoba	8,3	10,9	44,99	8
Toyota	Corrola	7,7	10,2	50,36	4
Volkswagen	Golf IV	8,3	9,9	61,62	6

- Znormalizowane opisy 15 samochodów (tabela informacyjna)

Marka	Model	Zużycie	Przyspieszenie	Cena	Wyposażenie
Alfa Romeo	156	0,57	-1,74	0,76	0,93
Audi	A4	0,34	0,46	1,87	1,31
BMW	316I	-0,11	0,80	1,29	0,56
Daewoo	Lanos	0,92	0,72	-1,05	-1,34
Honda	Civic	-1,03	-0,47	-0,35	0,18
Hyunday	Accent	-1,37	0,29	-1,03	-1,34
Lada	Samara	-0,46	1,40	-1,55	-1,72
Mitsubishi	Carisma	-0,46	0,55	0,23	0,93
Opel	Astra II	-0,69	0,55	0,05	0,56
Peugeot	206 XR	-1,14	1,57	-0,88	-0,96
Renault	Megane	-0,69	-1,23	-0,29	0,93
Saab	9 3S	2,40	-0,30	1,47	0,56
Seat	Cordoba	0,80	-0,39	-0,55	0,56
Toyota	Corrola	0,11	-0,98	-0,28	-0,96
Volkswagen	Golf IV	0,80	-1,23	0,29	-0,20

Wizualizacja danych wielowymiarowych

- Odległość Euklidesowa pomiędzy samochodami
  $\mathbf{a}^T = [z_a, p_a, c_a, w_a]$ i $\mathbf{b}^T = [z_b, p_b, c_b, w_b]$

$$d(\mathbf{a},\mathbf{b}) = ||\mathbf{a} - \mathbf{b}|| = ((z_a - z_b)^2 + (p_a - p_b)^2 + (c_a - c_b)^2 + (w_a - w_b)^2)^{1/2}$$

gdzie

  - z_a i z_b: znormalizowane zużycie paliwa pojazdów a i b
  - p_a i p_b: znormalizowane przyspieszenie pojazdów a i b
  - c_a i c_b: znormalizowana cena pojazdów a i b
  - w_a i w_b: znormalizowane wyposażenie pojazdów a i b
- Obliczanie odległości jest dokonywane na obiektach (czyli wierszach macierzy danych)

Wizualizacja danych wielowymiarowych

- Odległości pomiędzy (wybranymi wcześniej) trzema pojazdami:
  - d(„Carisma”, „Astra II”) = 0,48
  - d(„9 3S”, „Carisma”) = 3,25
  - d(„9 3S”, „Astra II”) = 3,51
- Te same odległości w innej formie (macierz odległości):

Carisma	0,00	0,48	3,25
Astra II	0,48	0,00	3,51
9 3S	3,25	3,51	0,00
	Carisma	Astra II	9 3S

Wizualizacja danych wielowymiarowych

- Odległości pomiędzy wszystkimi pojazdami (macierz odległości)

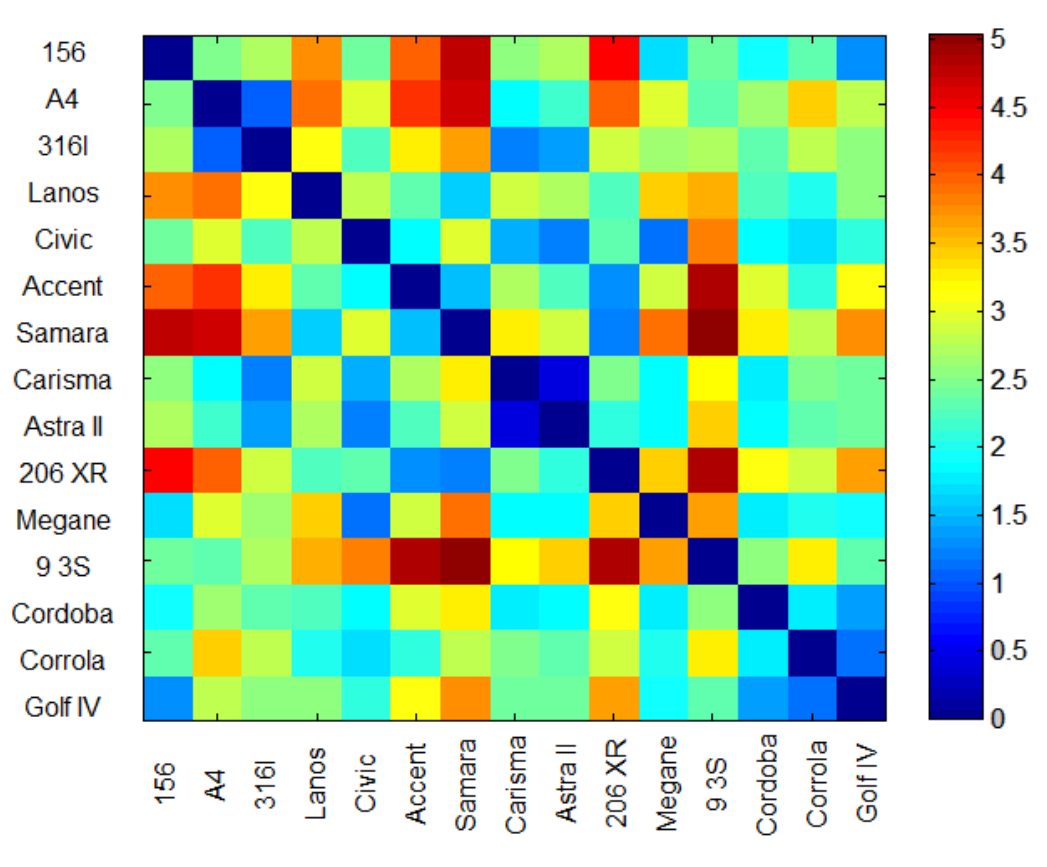
156	0,00	2,51	2,72	3,82	2,45	4,04	4,83	2,57	2,74	4,49	1,72	2,47	1,94	2,34	1,35
A4	2,51	0,00	1,11	4,00	3,00	4,29	4,73	1,86	2,23	4,02	2,96	2,36	2,71	3,45	2,81
316l	2,72	1,11	0,00	3,19	2,30	3,29	3,70	1,20	1,39	2,94	2,67	2,76	2,38	2,83	2,57
Lanos	3,82	4,00	3,19	0,00	2,83	2,33	1,66	2,95	2,72	2,27	3,48	3,63	2,25	2,07	2,63
Civic	2,45	3,00	2,30	2,83	0,00	1,86	2,97	1,51	1,21	2,40	1,13	3,91	1,88	1,69	2,12
Accent	4,04	4,29	3,29	2,33	1,86	0,00	1,57	2,77	2,30	1,36	2,92	4,95	3,00	2,13	3,18
Samara	4,83	4,73	3,70	1,66	2,97	1,57	0,00	3,31	2,92	1,24	3,95	5,04	3,31	2,86	3,77
Carisma	2,57	1,86	1,20	2,95	1,51	2,77	3,31	0,00	0,48	2,52	1,87	3,25	1,79	2,55	2,46
Astra II	2,74	2,23	1,39	2,72	1,21	2,30	2,92	0,48	0,00	2,10	1,86	3,51	1,86	2,32	2,45
206 XR	4,49	4,02	2,94	2,27	2,40	1,36	1,24	2,52	2,10	0,00	3,46	4,89	3,16	2,91	3,68
Megane	1,72	2,96	2,67	3,48	1,13	2,92	3,95	1,87	1,86	3,46	0,00	3,70	1,77	2,07	1,96
9 3S	2,47	2,36	2,76	3,63	3,91	4,95	5,04	3,25	3,51	4,89	3,70	0,00	2,58	3,33	2,33
Cordoba	1,94	2,71	2,38	2,25	1,88	3,00	3,31	1,79	1,86	3,16	1,77	2,58	0,00	1,79	1,41
Corrola	2,34	3,45	2,83	2,07	1,69	2,13	2,86	2,55	2,32	2,91	2,07	3,33	1,79	0,00	1,19
Golf IV	1,35	2,81	2,57	2,63	2,12	3,18	3,77	2,46	2,45	3,68	1,96	2,33	1,41	1,19	0,00
	156	A4	316l	Lanos	Civic	Accent	Samara	Carisma	Astra II	206 XR	Megane	9 3S	Cordoba	Corrola	Golf IV

Wizualizacja danych wielowymiarowych

- Obrazowanie odległości w postaci obrazu macierzy
  - idea
    - każdy element macierzy jest traktowany jako pewien piksel obrazu
    - wartości elementów macierzy są obrazowane jako kolory
  - jedynym problemem jest dobranie odwzorowania wartości na kolory
- Zobrazowanie w ten sposób macierzy odległości umożliwia łatwe dostrzeganie par obiektów charakteryzujących się małymi/dużymi odległościami

Wizualizacja danych wielowymiarowych

- Wizualizacja macierzy odległości (obraz macierzy)

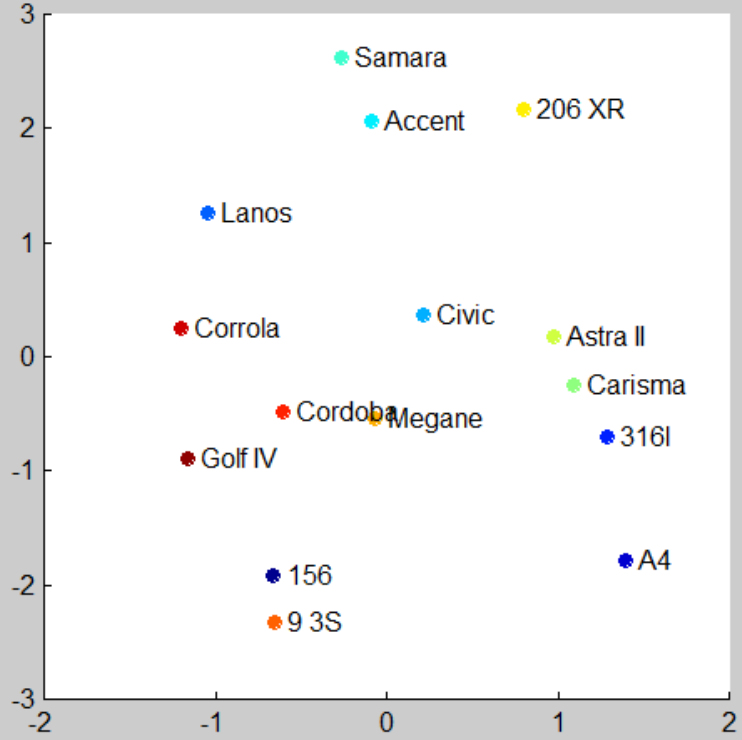


Wizualizacja danych wielowymiarowych

- Obrazowanie odległości w postaci mapy MDS
  - idea
    - każdy obiekt jest traktowany jako pewien punkt na płaszczyźnie
    - położenia punktów są tak dobrane, aby odległości między nimi były jak najbardziej zbliżone do odległości pomiędzy obiektami
  - zasadniczym problemem jest odpowiednie dobranie położeń punktów (problem ten jest rozwiązywany przez metodę MDS)
- Zobrazowanie w ten sposób obiektów umożliwia łatwe śledzenie względnych odległości pomiędzy wszystkimi parami obiektów

Wizualizacja danych wielowymiarowych

- Wizualizacja mapy MDS (wykres rozrzutu)



Wizualizacja danych wielowymiarowych

- Współrzędne MDS (tabela informacyjna)

Marka	Model	X	Y
Alfa Romeo	156	-1,74	0,57
Audi	A4	0,46	0,34
BMW	316I	0,80	-0,11
Daewoo	Lanos	0,72	0,92
Honda	Civic	-0,47	-1,03
Hyunday	Accent	0,29	-1,37
Lada	Samara	1,40	-0,46
Mitsubishi	Carisma	0,55	-0,46
Opel	Astra II	0,55	-0,69
Peugeot	206 XR	1,57	-1,14
Renault	Megane	-1,23	-0,69
Saab	9 3S	-0,30	2,40
Seat	Cordoba	-0,39	0,80
Toyota	Corrola	-0,98	0,11
Volkswagen	Golf IV	-1,23	0,80

Wizualizacja danych wielowymiarowych

- Zmienne generowane w metodzie MDS
  - są sztuczne
  - nie posiadają jednostek
  - nie są z założenia w żaden szczególny sposób związane z oryginalnymi zmiennymi
    - w praktyce mogą być jednak z nimi mniej lub bardziej skorelowane

…

Dygresja

- Co to jest simpleks?
  - simpleks n-wymiarowy:
    figura składająca się z n+1 równoodległych od siebie wzajemnie wierzchołków o długości boku równej 1
    - dla n=2: simpleksem jest odcinek
    - dla n=3: simpleksem jest trójkąt równoboczny
    - dla n=4: simpleksem jest czworościan foremny
    - dla n=5: simpleksem jest hiperczworościan foremny
    - …

…

Nierówność Cauchy–Bunyakovskii–Schwarz (CBS)

- Twierdzenie Cauchy'go–Bunyakovskii'ego–Schwarz'a:
  - dla dowolnych wektorów o zgodnych rozmiarach zachodzi:

$$|\mathbf{x}^T\mathbf{y}| \leq ||\mathbf{x}||||\mathbf{y}||$$

    gdzie $||\mathbf{u}||$ jest zdefiniowane jako $(\mathbf{u}^T\mathbf{u})^{1/2}$
    - $||\mathbf{u}||$ jest więc normą Minkowskiego przy p=2: $||\mathbf{u}|| \equiv ||\mathbf{u}||_2$
  - twierdzenie jest w oczywisty sposób spełnione dla $\mathbf{x} = \mathbf{0}$ lub $\mathbf{y} = \mathbf{0}$, dalej zakłada się więc, że $\mathbf{x} \neq \mathbf{0}$ i $\mathbf{y} \neq \mathbf{0}$
    - z czego wynika, że $||\mathbf{x}|| \neq 0$ i $||\mathbf{y}|| \neq 0$
  - zależność $|\mathbf{x}^T\mathbf{y}| \leq ||\mathbf{x}||||\mathbf{y}||$ implikuje oczywiście zależność $\mathbf{x}^T\mathbf{y} \leq ||\mathbf{x}||||\mathbf{y}||$
  - uzasadnienie

    (pominięte)

Nierówność Cauchy–Bunyakovskii–Schwarz (CBS)

- Dla dowolnego wektora **y** i niezerowego wektora **x** równość $|\mathbf{x}^T\mathbf{y}| = ||\mathbf{x}||||\mathbf{y}||$ zachodzi wtedy i tylko wtedy, gdy istnieje skalar a taki, że $\mathbf{y} = a\mathbf{x}$
  - uwagi
  - twierdzenie stanowi równościową wersję nierówności CBS dla niezerowych wektorów
  - uzasadnienie

    (pominięte)

Nierówność trójkąta

- Twierdzenie o nierówności trójkąta:
  - dla dowolnych wektorów o zgodnych rozmiarach zachodzi:

    $$\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\|+\|\mathbf{y}\|$$

    gdzie $\|\mathbf{u}\|$ jest zdefiniowane jako $(\mathbf{u}^T\mathbf{u})^{1/2}$
    - $\|\mathbf{u}\|$ jest więc normą Minkowskiego przy p=2: $\|\mathbf{u}\| \equiv \|\mathbf{u}\|_2$
  - twierdzenie jest spełnione dla $\mathbf{x}=\mathbf{0}$ lub $\mathbf{y}=\mathbf{0}$, dalej zakłada się więc, że $\mathbf{x} \neq \mathbf{0}$ i $\mathbf{y} \neq \mathbf{0}$
    - z czego wynika, że $\|\mathbf{x}\| \neq 0$ i $\|\mathbf{y}\| \neq 0$
  - uzasadnienie

    (pominięte)

…

Odległość Euklidesowa

- Miara odległości $\delta(\mathbf{x},\mathbf{y})$ między wektorami **x** i **y**
  - musi spełniać tzw. aksjomaty odległości

    $\delta(\mathbf{x},\mathbf{y}) \geq 0$, przy czym $\delta(\mathbf{x},\mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$

    $\delta(\mathbf{x},\mathbf{y}) = \delta(\mathbf{y},\mathbf{x})$

    $\delta(\mathbf{x},\mathbf{y}) + \delta(\mathbf{y},\mathbf{z}) \geq \delta(\mathbf{x},\mathbf{z})$ (nierówność trójkąta dla odległości)

Odległość Euklidesowa

- Odległość Euklidesowa między wektorami **x** i **y**

$$\delta_E(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}||$$

  - inaczej: $\delta_E(\mathbf{x},\mathbf{y}) = ||\mathbf{z}||$, gdzie $\mathbf{z} = \mathbf{x}-\mathbf{y}$
    - odległość Euklidesowa między wektorami **x** i **y** jest normą Euklidesową wektora będącego różnicą wektorów **x** i **y**
  - ponieważ $||\mathbf{z}|| = (\mathbf{z}^T\mathbf{z})^{1/2}$, więc $\delta_E(\mathbf{x},\mathbf{y}) = ((\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}))^{1/2}$

Odległość Euklidesowa

- Dla wektorów $\mathbf{a}^T = [a_1, a_2, \ldots, a_n]$ i $\mathbf{b}^T = [b_1, b_2, \ldots, b_n]$ mamy:

$$\delta_E(\mathbf{a},\mathbf{b}) = ||\mathbf{a} - \mathbf{b}|| = (|a_1 - b_1|^2 + |a_2 - b_2|^2 + \ldots + |a_n - b_n|^2)^{1/2}$$

Odległość Euklidesowa

- Przykład 1
  - zadanie: obliczyć odległość Euklidesową pomiędzy punktami, których współrzędne przedstawiają wektory $\mathbf{x} = [4, 2]^T$ oraz $\mathbf{y} = [7, -2]^T$
  - rozwiązanie klasyczne (na podstawie tw. Pitagorasa):

    $\delta_E(\mathbf{x},\mathbf{y}) = ((4-7)^2 + (2-(-2))^2)^{1/2} =$

    $= ((-3)^2 + (4)^2)^{1/2} = (9 + 16)^{1/2} = 25^{1/2} = 5$
  - rozwiązanie prezentowane (na podstawie normy wektora):

    $\delta_E(\mathbf{x},\mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| = || [4, 2]^T - [7, -2]^T || = || [-3, 4]^T || =$

    $= ([-3, 4]^T[-3, 4])^{1/2} = ((-3)\cdot(-3) + 4\cdot 4)^{1/2} =$

    $= ((-3)^2 + (4)^2)^{1/2} = (9 + 16)^{1/2} = 25^{1/2} = 5$

Odległość Euklidesowa

- Przykład 2
  - zadanie: jaka jest długość przekątnej kwadratu o boku długości 1?
  - rozwiązanie:
    - umieszczamy kwadrat w układzie współrzędnych, tak aby
      - pewien jego wierzchołek miał współrzędne $[0, 0]^T$
      - przeciwległy do niego wierzchołek miał współrzędne $[1, 1]^T$
    - obliczamy odległość pomiędzy wektorami $[0, 0]^T$ i $[1, 1]^T$
    - odpowiedź: $2^{1/2}$
  - (uwaga: kwadrat jest szczególnym przypadkiem hipersześcianu wielowymiarowego, a konkretnie hipersześcianem dwuwymiarowym)

Odległość Euklidesowa

- Quiz
  - zadanie: jaka jest długość przekątnej sześcianu o boku długości 1?
  - odpowiedź: $3^{1/2}$
  - (uwaga: sześcian jest szczególnym przypadkiem hipersześcianu wielowymiarowego, a konkretnie hipersześcianem trójwymiarowym)
  - zadanie: jaka jest długość przekątnej hipersześcianu czterowymiarowego o boku długości 1?
  - odpowiedź: $4^{1/2} = 2$
  - zadanie: przekątna ilowymiarowego hipersześcianu ma długość 3?
  - odpowiedź: 9

Odległość Euklidesowa

- Właściwości normy wektorowej:

  $||\mathbf{x}|| \geq 0$, przy czym $||\mathbf{x}|| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$

  $||a\mathbf{x}|| = |a| \cdot ||\mathbf{x}||$

  $||\mathbf{x}+\mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ (nierówność trójkąta dla normy)

Odległość Euklidesowa

- Właściwości odległości Euklidesowej między wektorami **x** i **y**
  - odległość Euklidesowa spełnia oczywiście aksjomaty odległości, a więc

    $||\mathbf{x}-\mathbf{y}|| \geq 0$, przy czym $||\mathbf{x}-\mathbf{y}|| = 0 \Leftrightarrow \mathbf{x}-\mathbf{y} = \mathbf{0} \Leftrightarrow \mathbf{x} = \mathbf{y}$
    - nieujemność i zerowość spełniana przez normę wektora

    $||\mathbf{x}-\mathbf{y}|| = ||\mathbf{y}-\mathbf{x}||$
    - symetryczność spełniana przez różnicę wektorów

    $||\mathbf{x}-\mathbf{y}|| + ||\mathbf{y}-\mathbf{z}|| \geq ||\mathbf{x}-\mathbf{z}||$, przy czym $||\mathbf{x}-\mathbf{y}|| = 0 \Leftrightarrow \mathbf{x}-\mathbf{y} = \mathbf{0} \Leftrightarrow \mathbf{x} = \mathbf{y}$
    - nierówność trójkąta spełniana przez normę wektora

Odległość Euklidesowa

- Spełnialność nierówności trójkąta przez odległość Euklidesową między wektorami **x** i **y**
  - odległość Euklidesowa $\delta_E(\mathbf{x},\mathbf{y})$ między wektorami **x** i **y** jest zdefiniowana jako $\delta_E(\mathbf{x},\mathbf{y}) = ||\mathbf{x}–\mathbf{y}||$
  - niech $\mathbf{a} = \mathbf{x} – \mathbf{y}$ oraz $\mathbf{b} = \mathbf{y} – \mathbf{z}$
  - wtedy $\mathbf{a} + \mathbf{b} = \mathbf{x} – \mathbf{y} + \mathbf{y} – \mathbf{z} = \mathbf{x} – \mathbf{z}$
  - ale zachodzi twierdzenie CBS i wynikające z niego $||\mathbf{a} + \mathbf{b}|| \leq ||\mathbf{a}|| + ||\mathbf{b}||$, zatem $||\mathbf{x} – \mathbf{z}|| \leq ||\mathbf{x} – \mathbf{y}|| + ||\mathbf{y} – \mathbf{z}||$
  - a więc ostatecznie zachodzi $\delta_E(\mathbf{x},\mathbf{z}) \leq \delta_E(\mathbf{x},\mathbf{y}) + \delta_E(\mathbf{y},\mathbf{z})$

Odległość Euklidesowa

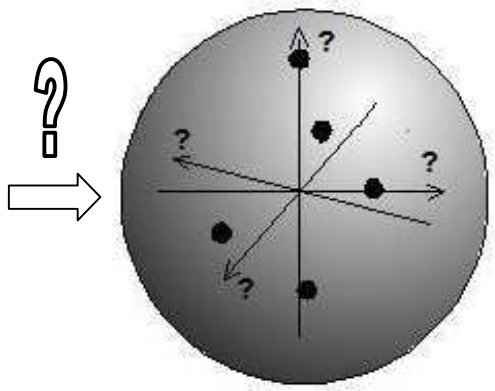
- Dany jest zbiór obiektów/wektorów $\mathbf{x}_1, \ldots, \mathbf{x}_m$
- Macierz odległości Euklidesowych $\mathbf{D} = [d_{ij}]$, gdzie $d_{ij} = \delta_E(\mathbf{x}_i, \mathbf{x}_j)$ spełnia następujące właściwości
  - $d_{ii} = 0$ (zerowa przekątna)
  - $d_{ij} = d_{ji}$ (symetria)
  - $d_{ik} \leq d_{ij} + d_{jk}$ (nierówność trójkąta)

…

Wizualizacja danych wielowymiarowych

- Problem wizualizowania danych wielowymiarowych

	X	Y	Z	…	…
a	12	1.5	56	…	…
b	21	1.8	76	…	…
c	44	1.9	71	…	…
d	18	1.0	59	…	…
e	22	1.3	64	…	…



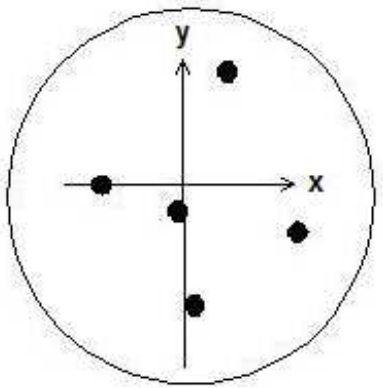
Wizualizacja danych wielowymiarowych

- Metoda MDS: generowanie współrzędnych z macierzy odległości

	a	b	c	d	e
a	0.00	1.31	0.23	3.19	5.42
b	2.10	0.00	1.26	5.26	3.14
c	4.14	2.98	0.00	1.12	2.65
d	3.55	1.07	4.21	0.00	3.30
e	1.91	0.15	0.36	3.14	0.00

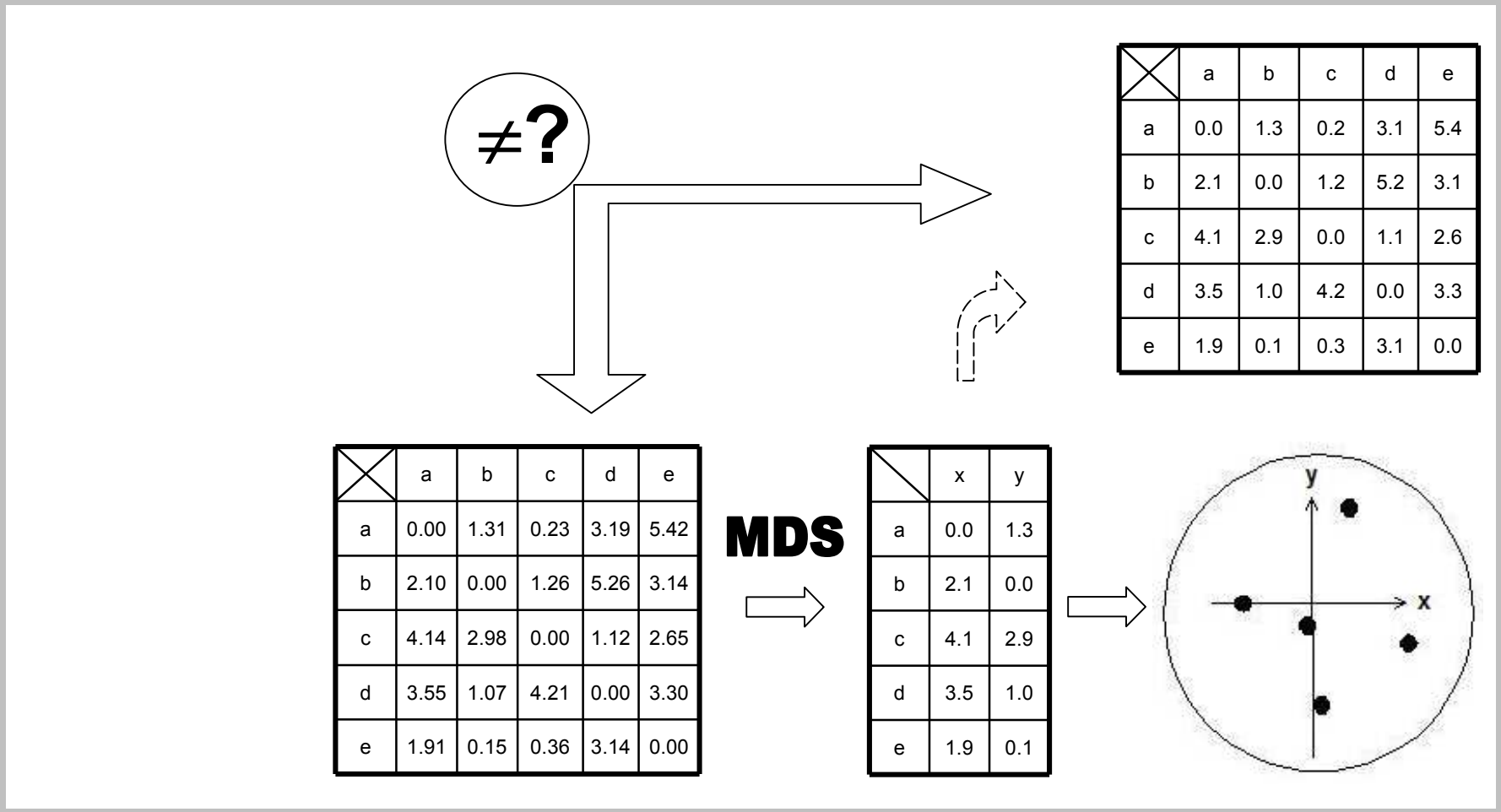
MDS

	x	y
a	0.0	1.3
b	2.1	0.0
c	4.1	2.9
d	3.5	1.0
e	1.9	0.1



Wizualizacja danych wielowymiarowych

- Nieodłączny problem MDS: niedokładność odwzorowania



	a	b	c	d	e
a	0.0	1.3	0.2	3.1	5.4
b	2.1	0.0	1.2	5.2	3.1
c	4.1	2.9	0.0	1.1	2.6
d	3.5	1.0	4.2	0.0	3.3
e	1.9	0.1	0.3	3.1	0.0

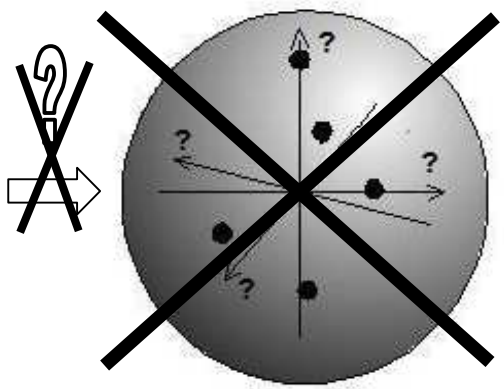
	a	b	c	d	e
a	0.00	1.31	0.23	3.19	5.42
b	2.10	0.00	1.26	5.26	3.14
c	4.14	2.98	0.00	1.12	2.65
d	3.55	1.07	4.21	0.00	3.30
e	1.91	0.15	0.36	3.14	0.00

	x	y
a	0.0	1.3
b	2.1	0.0
c	4.1	2.9
d	3.5	1.0
e	1.9	0.1

Wizualizacja danych wielowymiarowych

- Zastosowanie MDS: wizualizowanie danych wielowymiarowych

	X	Y	Z	…	…
a	12	1.5	56	…	…
b	21	1.8	76	…	…
c	44	1.9	71	…	…
d	18	1.0	59	…	…
e	22	1.3	64	…	…



Wizualizacja danych wielowymiarowych

- Zastosowanie MDS: wizualizowanie danych wielowymiarowych

	X	Y	Z	…	…
a	12	1.5	56	…	…
b	21	1.8	76	…	…
c	44	1.9	71	…	…
d	18	1.0	59	…	…
e	22	1.3	64	…	…

	a	b	c	d	e
a	0.00	1.31	0.23	3.19	5.42
b	2.10	0.00	1.26	5.26	3.14
c	4.14	2.98	0.00	1.12	2.65
d	3.55	1.07	4.21	0.00	3.30
e	1.91	0.15	0.36	3.14	0.00

Wizualizacja danych wielowymiarowych

- Zastosowanie MDS: wizualizowanie danych wielowymiarowych

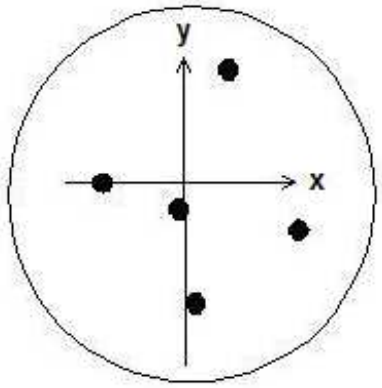
	X	Y	Z	…	…
a	12	1.5	56	…	…
b	21	1.8	76	…	…
c	44	1.9	71	…	…
d	18	1.0	59	…	…
e	22	1.3	64	…	…

	a	b	c	d	e
a	0.0	1.3	0.2	3.1	5.4
b	2.1	0.0	1.2	5.2	3.1
c	4.1	2.9	0.0	1.1	2.6
d	3.5	1.0	4.2	0.0	3.3
e	1.9	0.1	0.3	3.1	0.0

	a	b	c	d	e
a	0.00	1.31	0.23	3.19	5.42
b	2.10	0.00	1.26	5.26	3.14
c	4.14	2.98	0.00	1.12	2.65
d	3.55	1.07	4.21	0.00	3.30
e	1.91	0.15	0.36	3.14	0.00

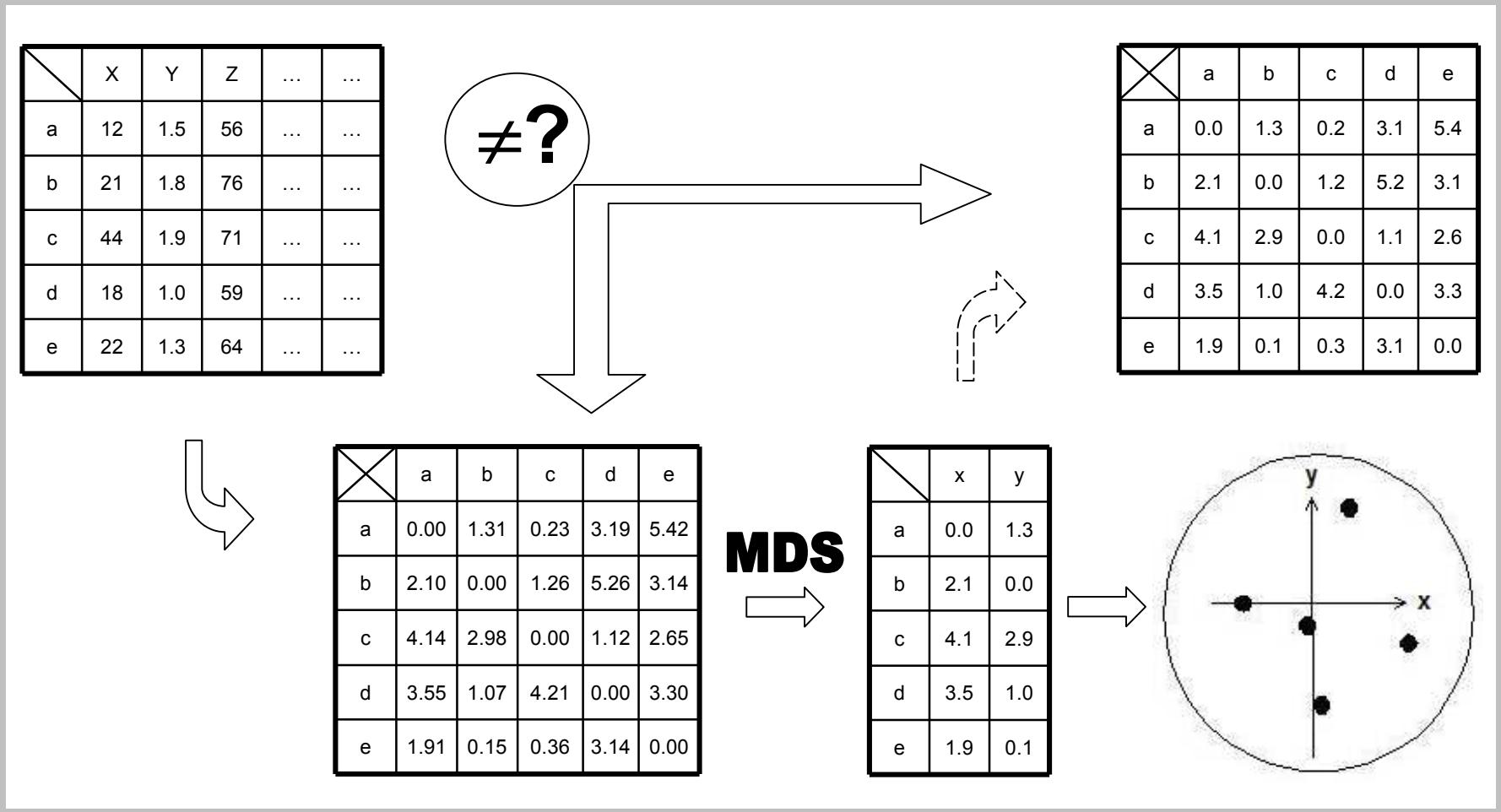
MDS

	x	y
a	0.0	1.3
b	2.1	0.0
c	4.1	2.9
d	3.5	1.0
e	1.9	0.1



Wizualizacja danych wielowymiarowych

- Zastosowanie MDS: wizualizowanie danych wielowymiarowych



	X	Y	Z	…	…
a	12	1.5	56	…	…
b	21	1.8	76	…	…
c	44	1.9	71	…	…
d	18	1.0	59	…	…
e	22	1.3	64	…	…

	a	b	c	d	e
a	0.0	1.3	0.2	3.1	5.4
b	2.1	0.0	1.2	5.2	3.1
c	4.1	2.9	0.0	1.1	2.6
d	3.5	1.0	4.2	0.0	3.3
e	1.9	0.1	0.3	3.1	0.0

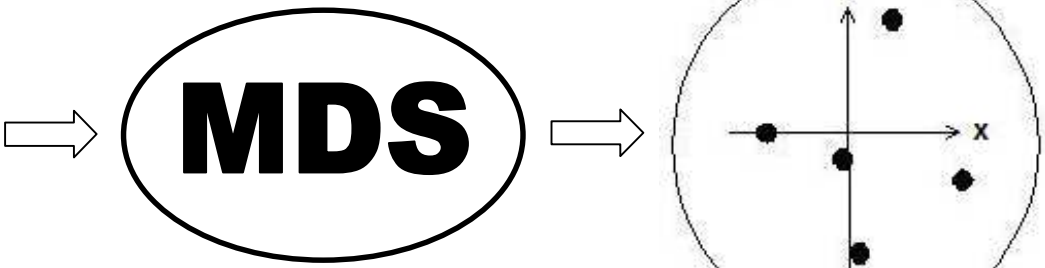
	a	b	c	d	e
a	0.00	1.31	0.23	3.19	5.42
b	2.10	0.00	1.26	5.26	3.14
c	4.14	2.98	0.00	1.12	2.65
d	3.55	1.07	4.21	0.00	3.30
e	1.91	0.15	0.36	3.14	0.00

	x	y
a	0.0	1.3
b	2.1	0.0
c	4.1	2.9
d	3.5	1.0
e	1.9	0.1

Wizualizacja danych wielowymiarowych

- Zastosowanie MDS: wizualizowanie danych wielowymiarowych

	X	Y	Z	…	…
a	12	1.5	56	…	…
b	21	1.8	76	…	…
c	44	1.9	71	…	…
d	18	1.0	59	…	…
e	22	1.3	64	…	…



Wizualizacja danych wielowymiarowych

- Bardziej złożone podejścia do wizualizacji danych wielowymiarowych

  („bardziej złożone”, tzn. wykraczające poza wizualizację wielowymiarowych opisów obiektów i wizualizację wielowymiarowych konfiguracji obiektów)

  – jednoczesna wizualizacja wielowymiarowych opisów obiektów oraz wielowymiarowych konfiguracji obiektów
    - obrazuje wzajemne położenia obiektów
    - obiekty są reprezentowane w postaci punktów
    - ułatwia dostrzeganie systematycznych różnic pomiędzy zbiorami obiektów

…

Wizualizacja danych wielowymiarowych

- Niech $\mathbf{A}^{(p)} = [(a_{ij})^p]$, gdzie
  - $\mathbf{A} = [a_{ij}]$ jest dowolną macierzą
  - p jest dowolnym skalarem

  oznacza macierz wartości a_{ij} podniesionych do potęgi p
  - operacja poelementowa
    - nie mylić z potęgowaniem macierzy!
- Przykład

A

1	2
3	4

$\mathbf{A}^2 = \mathbf{AA}$

7	10
15	22

$\mathbf{A}^{(2)}$

1	4
9	16

- Dla jakich **A** zachodzi $\mathbf{A}^{(p)} = \mathbf{A}^p$?
- Dla jakich diagonalnych **L** i jakich p zachodzi $\mathbf{L}^{(p)}\mathbf{L}^{(p)} = \mathbf{L}$?

Wizualizacja danych wielowymiarowych

- Klasyczne rozwiązanie problemu MDS: procedura
  - dane:
    - macierz odległości **D** o rozmiarach mxm
    - skalar n (liczba wymiarów mapy) spełniający $n \leq m$
      - zwyczajowo n = 2 lub n = 3
  - obliczamy $\mathbf{B} = -0.5 \cdot \mathbf{J}\mathbf{D}^{(2)}\mathbf{J}$,
    - gdzie $\mathbf{J} = \mathbf{I} - \mathbf{E}/m$ (macierz skalujaca)
  - znajdujemy rozkład EVD macierzy **B** postaci $\mathbf{B} = \mathbf{K}\mathbf{L}\mathbf{K}^T$
  - obliczamy $\mathbf{X} = \mathbf{K}\mathbf{L}^{(0.5)}$
  - wynik:
    - n kolumn macierzy **X** odpowiadających największym wartościom własnym macierzy **B**

Wizualizacja danych wielowymiarowych

- Klasyczne rozwiązanie problemu MDS: analiza
  - powstająca macierz **B** jest macierzą pewnych iloczynów skalarnych (odpowiednik macierzy kowariancji)
  - zastąpienie macierzy odległości macierzą iloczynów skalarnych pozwala na stosowanie nowych operacji: w szczególności odkrycie macierzy $\mathbf{X}_{mxn}$ takiej, że $\mathbf{XX}^T = \mathbf{B}$ rozwiązuje problem
    - w praktyce równość powyższa może zachodzić w przybliżeniu
  - macierz **X** odkrywa się stosując rozkład EVD macierzy **B**
    - jeżeli $\mathbf{X} = \mathbf{KL}^{(0.5)}$, to

      $\mathbf{XX}^T = \mathbf{KL}^{(0.5)}(\mathbf{KL}^{(0.5)})^T = \mathbf{KL}^{(0.5)}(\mathbf{L}^{(0.5)})^T\mathbf{K}^T = \mathbf{KL}^{(0.5)}\mathbf{L}^{(0.5)}\mathbf{K}^T =$

      $= \mathbf{KL}^{(1)}\mathbf{K}^T = \mathbf{KLK}^T = \mathbf{B}$

Wizualizacja danych wielowymiarowych

- Klasyczne rozwiązanie problemu MDS: analiza
  - jeżeli **D** jest macierzą odległości euklidesowych, to macierz **B** jest macierzą nieujemnie określoną, a więc wartości własne macierzy **B** są (rzeczywistymi) wartościami nieujemnymi
    - stosowanie tej procedury możliwe jest także do takich macierzy **D**, które nie spełniają wszystkich warunków nakładanych na macierz odległości euklidesowych
    - w praktyce wystarcza, aby n największych co do wartości bezwzględnej wartości własnych było (rzeczywistymi) wartościami dodatnimi

Wizualizacja danych wielowymiarowych

- Klasyczne rozwiązanie problemu MDS: uwagi efektywnościowe
  - wynikowa macierz $\mathbf{X} = \mathbf{K}\mathbf{L}^{(0.5)}$ posiada m kolumn; jej obliczanie jest dla n < m oczywiście nadmiarowe (szczególnie gdy n << m), ponieważ powstaje m–n niepotrzebnych kolumn, w praktyce wystarczy obliczyć

    $\mathbf{X} = \mathbf{K}_{1:m,1:n}\mathbf{L}^{(0.5)}{}_{1:n,1:n}$
    - macierz **X** posiada wtedy n kolumn

…

Dygresja

- Norma Frobeniusa $||.||_F$ macierzy $\mathbf{X} = [x_{ij}]$

  $||\mathbf{X}||_F = (\sum(x_{ij})^2)^{1/2}$

  (pierwiastek z sumy kwadratów wszystkich elementów)

...

Wizualizacja danych wielowymiarowych

- Zadanie optymalizacji w MDS (wersja podstawowa)
  - dane: macierz odległości **D** o rozmiarach mxm
  - niech:
    - **X** będzie macierzą zmiennych
      - **X** jest macierzą o rozmiarach mxd, gdzie d jest liczbą wymiarów konfiguracji wynikowej (dla konfiguracji na płaszczyźnie d = 2)
    - **D**(**X**) = pdist(**X**) będzie macierzą odległości Euklidesowych pomiędzy wektorami wierszowymi macierzy **X**
      - **D**(**X**) jest macierzą o rozmiarach mxm
    - $s(\mathbf{X}) = ||\mathbf{D}(\mathbf{X}) - \mathbf{D}||_F$ będzie wartością macierzowej normy Frobeniusa z różnicy pomiędzy macierzą **D**(**X**) a macierzą **D**
      - s(**X**) jest skalarem (funkcją skalarną od argumentu macierzowego)
  - zadanie: znaleźć minimum funkcji s(**X**)

Wizualizacja danych wielowymiarowych

- Cechy zadania optymalizacji w MDS (wersja podstawowa)
  - funkcja celu: s(**X**)
    - funkcja skalarna od argumentu macierzowego **X** o wymiarach mxd
    - liczba zmiennych skalarnych w problemie (liczba wymiarów przestrzeni, w której dokonywana jest optymalizacja): m*d
  - ograniczenia: brak

Wizualizacja danych wielowymiarowych

- Jakość rozwiązań generowanych w metodzie MDS
  - dokładne, tzn. takie, dla których s(**X**) = 0 (mogą nie istnieć)
  - przybliżone, tzn. takie, dla których s(**X**) > 0 (zawsze istnieją)
    - użyteczne/dobre, tzn. takie, dla których s(**X**) jest małe
    - nieużyteczne/słabe, tzn. takie, dla których s(**X**) jest duże

Wizualizacja danych wielowymiarowych

- Związane z danymi przyczyny niedokładności rozwiązań MDS
  - zbyt wysoka wielowymiarowość przestrzeni oryginalnej
  - niespełnialność aksjomatów miary odległości oryginalnej

Wizualizacja danych wielowymiarowych

- Niespełnialność aksjomatów odległości jako przyczyna niedokładności rozwiązań MDS
  - przykład (rozwiązanie przybliżone)
    - niech $M_1, \ldots, M_m$ będzie zbiorem miejscowości i niech $\mathbf{D} = [d_{ij}]$ będzie macierzą czasów przejazdu, tzn. macierzą której element d_{ij} jest czasem przejazdu z mniejscowości M_i do miejscowości M_j
    - w niektórych przypadkach może zachodzić $d_{ij} \neq d_{ji}$, a więc **D** może nie być macierzą symetryczną

Wizualizacja danych wielowymiarowych

- Niespełnialność aksjomatów odległości jako przyczyna, c.d.
  - niech $s(\mathbf{X}) = ||\mathbf{D}(\mathbf{X}) - \mathbf{D}||_F$ będzie minimalizowaną funkcją celu
  - wyrażenie $||\mathbf{D}(\mathbf{X}) - \mathbf{D}||_F$ oznacza normę macierzową (Frobeniusa), a więc zachodzi $||\mathbf{D}(\mathbf{X}) - \mathbf{D}||_F \geq 0$ oraz

    $$||\mathbf{D}(\mathbf{X}) - \mathbf{D}||_F = 0 \Leftrightarrow \mathbf{D}(\mathbf{X}) - \mathbf{D} = \mathbf{O} \Leftrightarrow \mathbf{D}(\mathbf{X}) = \mathbf{D}$$

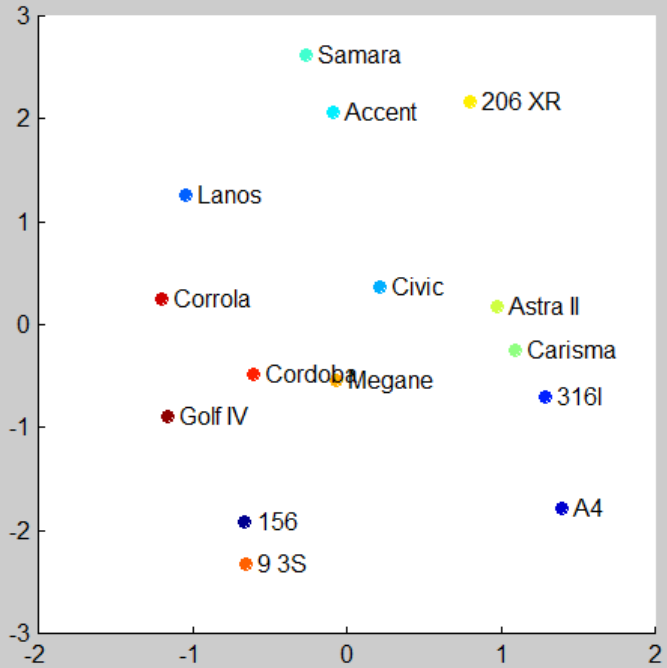
  - jednocześnie, ponieważ $\mathbf{D}(\mathbf{X}) = [d_{ij}]$ jest macierzą odległości Euklidesowych, więc spełnia warunki
    - $d_{ii} = 0$ (zerowa przekątna)
    - $d_{ij} = d_{ji}$ (symetria)
    - $d_{ik} \leq d_{ij} + d_{jk}$ (nierówność trójkąta)

Wizualizacja danych wielowymiarowych

- Niespełnialność aksjomatów odległości jako przyczyna, c.d.
  - jeżeli macierz **D** nie spełnia któregoś (lub wielu, potencjalnie wszystkich) spośród powyższych warunków, to zawsze będzie istniała jakaś różnica pomiędzy macierzą **D**(**X**) a macierzą **D**
  - czyli dla dowolnego rozwiązania **X** będzie zachodziło $\mathbf{D}(\mathbf{X}) \neq \mathbf{D}$, a więc $\mathbf{D}(\mathbf{X}) - \mathbf{D} \neq \mathbf{O}$, a więc $||\mathbf{D}(\mathbf{X}) - \mathbf{D}||_F \neq 0$, a to oznacza, że $||\mathbf{D}(\mathbf{X}) - \mathbf{D}||_F > 0$ (każde rozwiązanie **X** jest przybliżone)

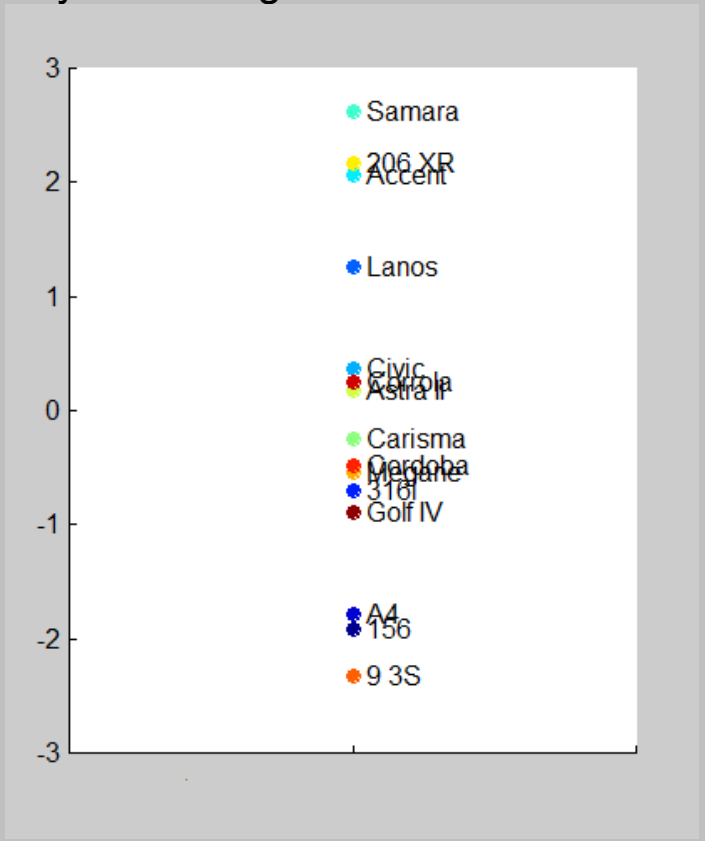
Wizualizacja danych wielowymiarowych

- Zbyt wysoka wielowymiarowość jako przyczyna niedokładności rozwiązań MDS
  - przykład skalowania z czterech wymiarów do dwóch
    - założenie: wszystkie odległości Euklidesowe



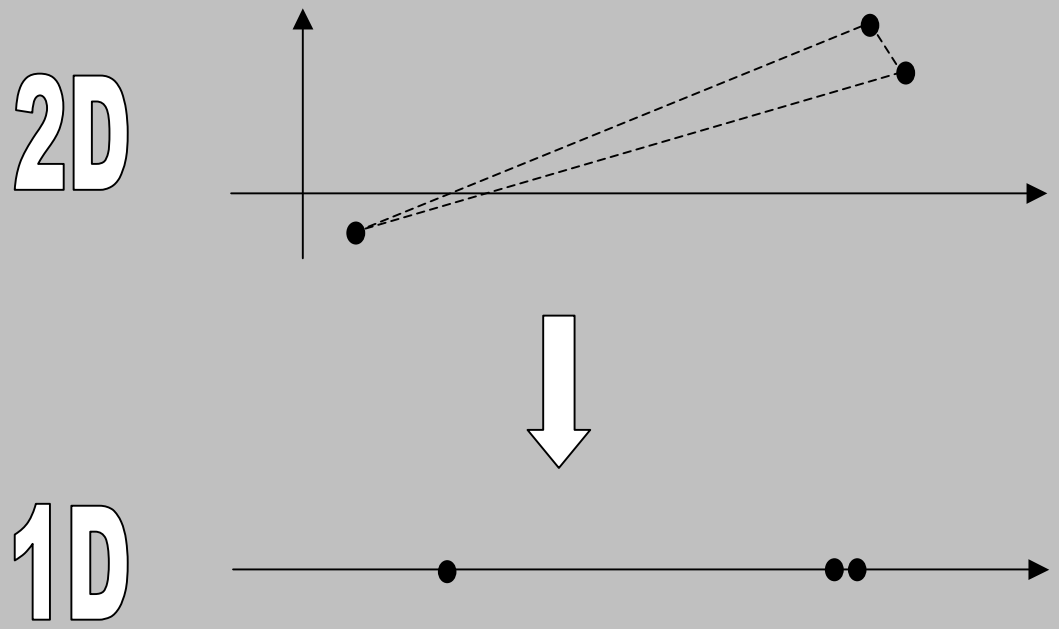
Wizualizacja danych wielowymiarowych

- Zbyt wysoka wielowymiarowość jako przyczyna niedokładności rozwiązań MDS
  - przykład skalowania z czterech wymiarów do jednego
    - założenie: wszystkie odległości Euklidesowe



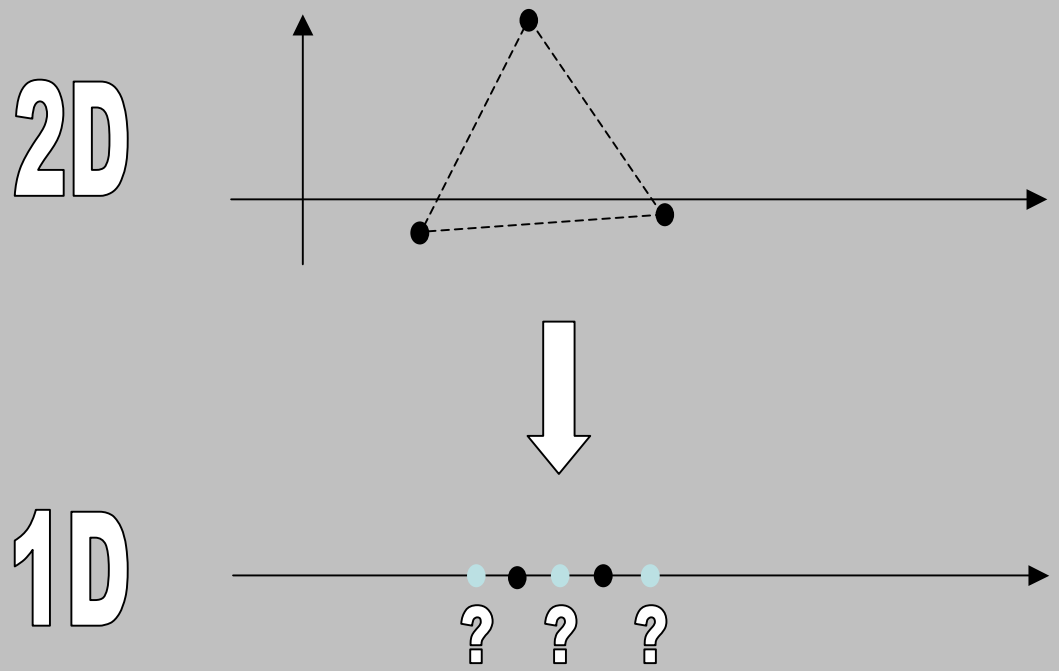
Wizualizacja danych wielowymiarowych

- Zbyt wysoka wielowymiarowość jako przyczyna niedokładności rozwiązań MDS
  - przykład skalowania z dwóch wymiarów do jednego
    - założenie: wszystkie odległości Euklidesowe
  - wynikowe rozwiązanie przybliżone ale użyteczne



Wizualizacja danych wielowymiarowych

- Zbyt wysoka wielowymiarowość jako przyczyna niedokładności rozwiązań MDS
  - przykład skalowania z dwóch wymiarów do jednego
    - założenie: wszystkie odległości Euklidesowe
  - wynikowe rozwiązanie przybliżone ale bezużyteczne



Wizualizacja danych wielowymiarowych

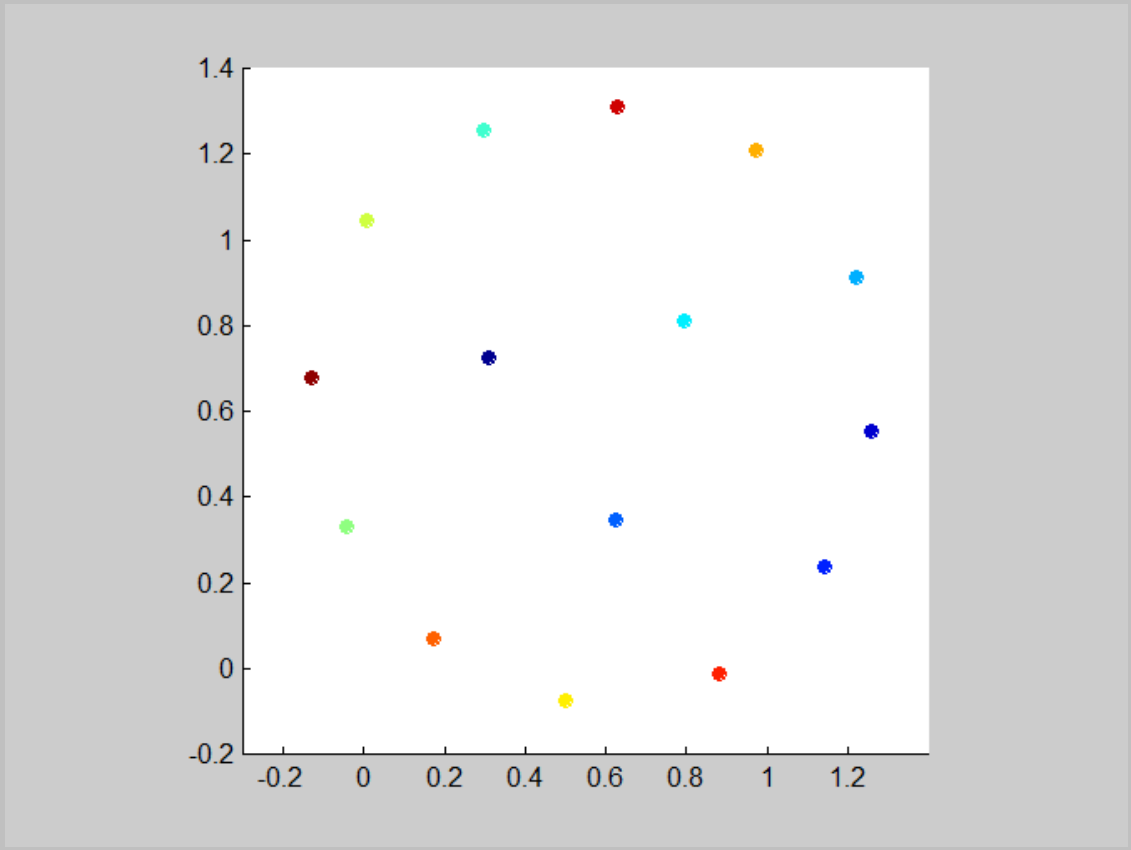
- Zbyt wysoka wielowymiarowość jako przyczyna niedokładności rozwiązań MDS
  - regularnej figury z przestrzeni r-wymiarowej nie można przedstawić w przestrzeni s-wymiarowej (gdy r > s) ze względu na brak odpowiedniej ilości miejsca
    - przykłady regularnych kształtów: hipersimpleksy, hipersześciany, hipersfery
  - problem dotyczy w praktyce wszystkich figur (mniej lub bardziej regularnych)
    - dokładność odwzorowania spada wraz ze wzrostem różnicy pomiędzy r i s
    - dla r = s (brak różnicy) dokładność ta powinna być zawsze pełna (otrzymujemy rozwiązanie dokładne)

Wizualizacja danych wielowymiarowych

- Rozwiązania dla macierzy odległości postaci $\mathbf{D}_{mxm} = \mathbf{1} \cdot \mathbf{1}^T - \mathbf{I}$
  - szczególnym przypadkiem regularnej figury n-wymiarowej jest simpleks n-wymiarowy (simpleks n-wymiarowy składa się z n+1 równoodległych od siebie wzajemnie wierzchołków) o długości boku równej 1
    - dla n=2: simpleksem jest odcinek
    - dla n=3: simpleksem jest trójkąt równoboczny
    - dla n=4: simpleksem jest czworościan foremny
    - dla n=5: simpleksem jest hiperczworościan foremny
    - …
  - dla każdego n simpleks n-wymiarowy charakteryzuje się tym, że macierz odległości pomiędzy jego wierzchołkami jest postaci $\mathbf{1} \cdot \mathbf{1}^T - \mathbf{I}$, czyli zawiera
    - zera na przekątnej
    - jedynki poza przekątną
  - rozwiązania proponowane przez metodę MDS dla tego rodzaju macierzy odległości są bardzo charakterystycznymi figurami

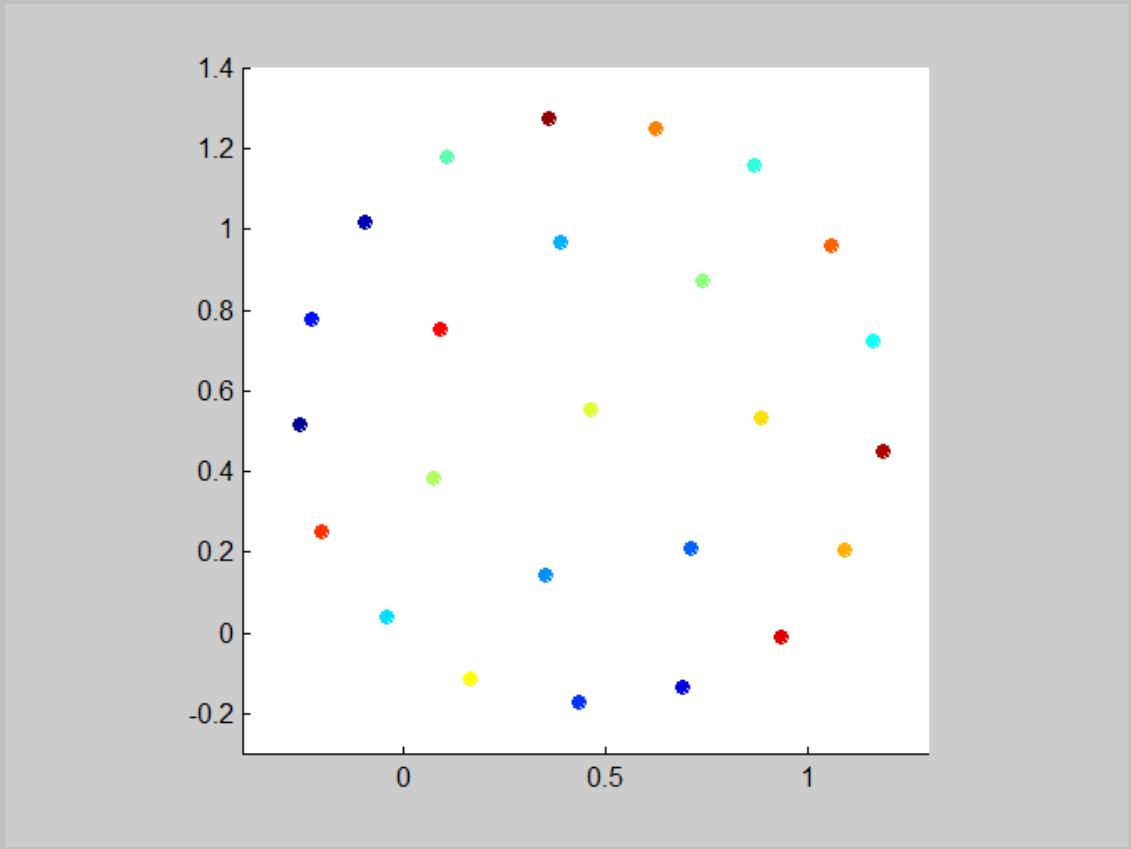
Wizualizacja danych wielowymiarowych

- Rozwiązania dla macierzy odległości postaci $\mathbf{D}_{mxm} = \mathbf{1} \cdot \mathbf{1}^T - \mathbf{I}$
  - m = 15



Wizualizacja danych wielowymiarowych

- Rozwiązania dla macierzy odległości postaci $\mathbf{D}_{mxm} = \mathbf{1} \cdot \mathbf{1}^T - \mathbf{I}$
  - m = 25



Wizualizacja danych wielowymiarowych

- Kiedy rozwiązanie problemu MDS jest dokładne w sytuacji, gdy znamy oryginalną konfigurację punktów (a nie tylko macierz odległości pomiędzy tymi punktami)?

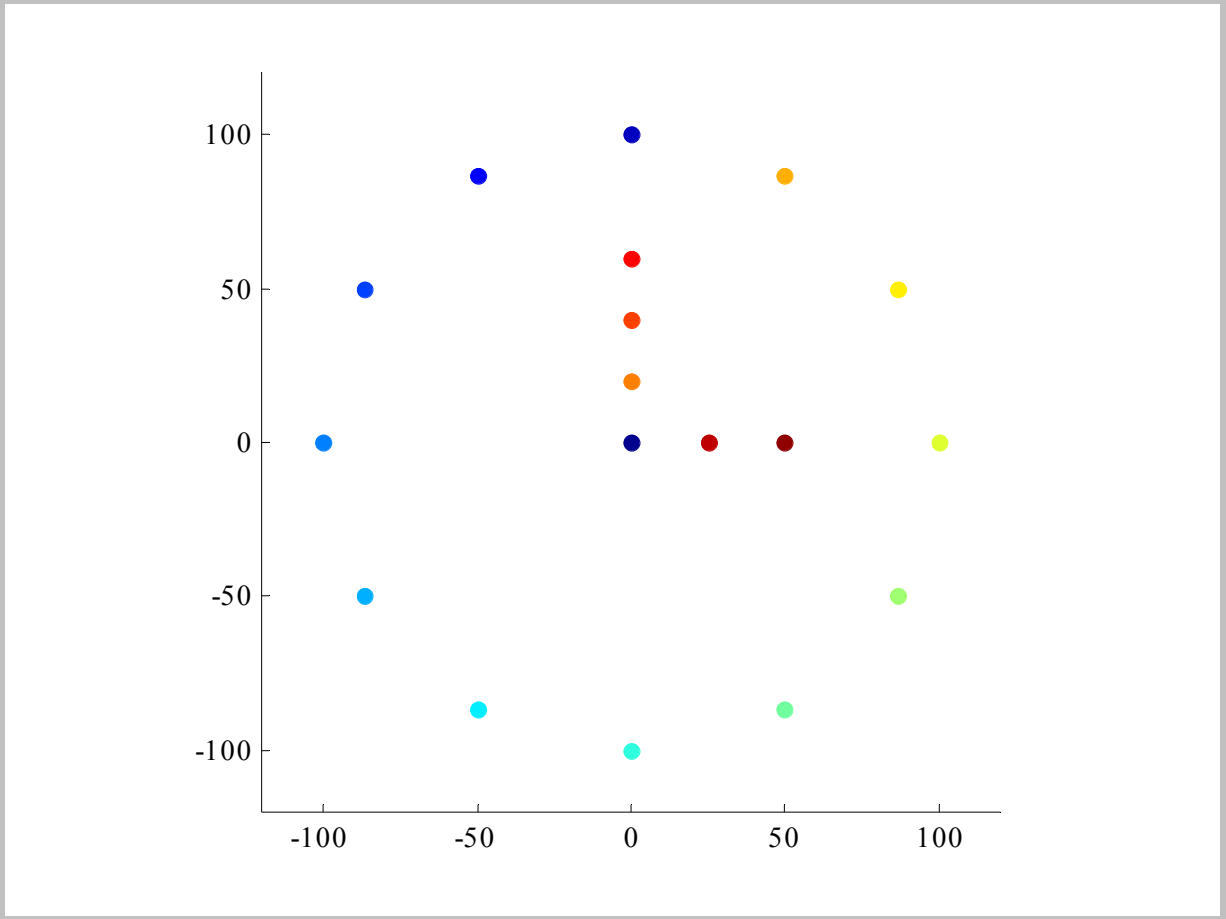
Wizualizacja danych wielowymiarowych

- Figura „clock”: współrzędne oryginalnych punktów

x¤	y¤
0,00¤	0,00¤
0,00¤	100,00¤
-50,00¤	86,60¤
-86,60¤	50,00¤
-100,00¤	0,00¤
-86,60¤	-50,00¤
-50,00¤	-86,60¤
0,00¤	-100,00¤
50,00¤	-86,60¤
86,60¤	-50,00¤
100,00¤	0,00¤
86,60¤	50,00¤
50,00¤	86,60¤
0,00¤	20,00¤
0,00¤	40,00¤
0,00¤	60,00¤
25,00¤	0,00¤
50,00¤	0,00¤

Wizualizacja danych wielowymiarowych

- Figura „clock”: wizualizacja (wykres rozrzutu) oryginalnych punktów



Wizualizacja danych wielowymiarowych

- Figura „clock”: macierz odległości oryginalnych punktów

0,00¤	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	20,00¤	40,00¤	60,00¤	25,00¤	50,00¤
100,00	0,00¤	51,76¤	100,00	141,42	173,21	193,19	200,00	193,19	173,21	141,42	100,00	51,76¤	80,00¤	60,00¤	40,00¤	103,08	111,80
100,00	51,76¤	0,00¤	51,76¤	100,00	141,42	173,21	193,19	200,00	193,19	173,21	141,42	100,00	83,28¤	68,35¤	56,64¤	114,56	132,29
100,00	100,00	51,76¤	0,00¤	51,76¤	100,00	141,42	173,21	193,19	200,00	193,19	173,21	141,42	91,65¤	87,18¤	87,18¤	122,29	145,47
100,00	141,42	100,00	51,76¤	0,00¤	51,76¤	100,00	141,42	173,21	193,19	200,00	193,19	173,21	101,98	107,70	116,62	125,00	150,00
100,00	173,21	141,42	100,00	51,76¤	0,00¤	51,76¤	100,00	141,42	173,21	193,19	200,00	193,19	111,36	124,90	140,00	122,29	145,47
100,00	193,19	173,21	141,42	100,00	51,76¤	0,00¤	51,76¤	100,00	141,42	173,21	193,19	200,00	117,75	136,12	154,89	114,56	132,29
100,00	200,00	193,19	173,21	141,42	100,00	51,76¤	0,00¤	51,76¤	100,00	141,42	173,21	193,19	120,00	140,00	160,00	103,08	111,80
100,00	193,19	200,00	193,19	173,21	141,42	100,00	51,76¤	0,00¤	51,76¤	100,00	141,42	173,21	117,75	136,12	154,89	90,14¤	86,60¤
100,00	173,21	193,19	200,00	193,19	173,21	141,42	100,00	51,76¤	0,00¤	51,76¤	100,00	141,42	111,36	124,90	140,00	79,34¤	61,97¤
100,00	141,42	173,21	193,19	200,00	193,19	173,21	141,42	100,00	51,76¤	0,00¤	51,76¤	100,00	101,98	107,70	116,62	75,00¤	50,00¤
100,00	100,00	141,42	173,21	193,19	200,00	193,19	173,21	141,42	100,00	51,76¤	0,00¤	51,76¤	91,65¤	87,18¤	87,18¤	79,34¤	61,97¤
100,00	51,76¤	100,00	141,42	173,21	193,19	200,00	193,19	173,21	141,42	100,00	51,76¤	0,00¤	83,28¤	68,35¤	56,64¤	90,14¤	86,60¤
20,00¤	80,00¤	83,28¤	91,65¤	101,98	111,36	117,75	120,00	117,75	111,36	101,98	91,65¤	83,28¤	0,00¤	20,00¤	40,00¤	32,02¤	53,85¤
40,00¤	60,00¤	68,35¤	87,18¤	107,70	124,90	136,12	140,00	136,12	124,90	107,70	87,18¤	68,35¤	20,00¤	0,00¤	20,00¤	47,17¤	64,03¤
60,00¤	40,00¤	56,64¤	87,18¤	116,62	140,00	154,89	160,00	154,89	140,00	116,62	87,18¤	56,64¤	40,00¤	20,00¤	0,00¤	65,00¤	78,10¤
25,00¤	103,08	114,56	122,29	125,00	122,29	114,56	103,08	90,14¤	79,34¤	75,00¤	79,34¤	90,14¤	32,02¤	47,17¤	65,00¤	0,00¤	25,00¤
50,00¤	111,80	132,29	145,47	150,00	145,47	132,29	111,80	86,60¤	61,97¤	50,00¤	61,97¤	86,60¤	53,85¤	64,03¤	78,10¤	25,00¤	0,00¤

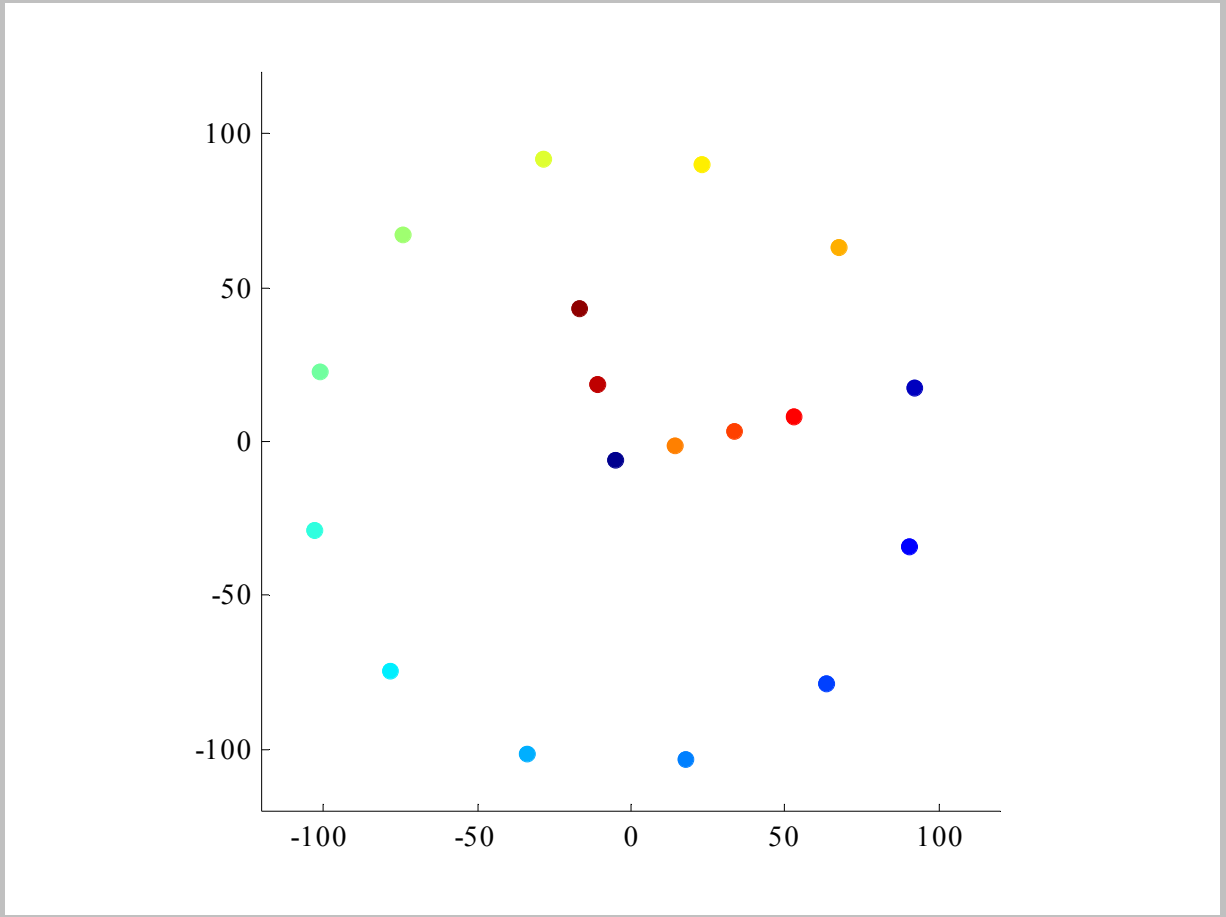
Wizualizacja danych wielowymiarowych

- Figura „clock”: współrzędne nowych punktów

u¤	v¤
-5,52¤	-5,59¤
91,77¤	17,50¤
90,29¤	-34,24¤
63,13¤	-78,31¤
17,57¤	-102,89¤
-34,17¤	-101,40¤
-78,24¤	-74,24¤
-102,82¤	-28,69¤
-101,33¤	23,05¤
-74,18¤	67,12¤
-28,62¤	91,70¤
23,12¤	90,22¤
67,19¤	63,06¤
13,94¤	-0,97¤
33,39¤	3,65¤
52,85¤	8,26¤
-11,30¤	18,73¤
-17,07¤	43,05¤

Wizualizacja danych wielowymiarowych

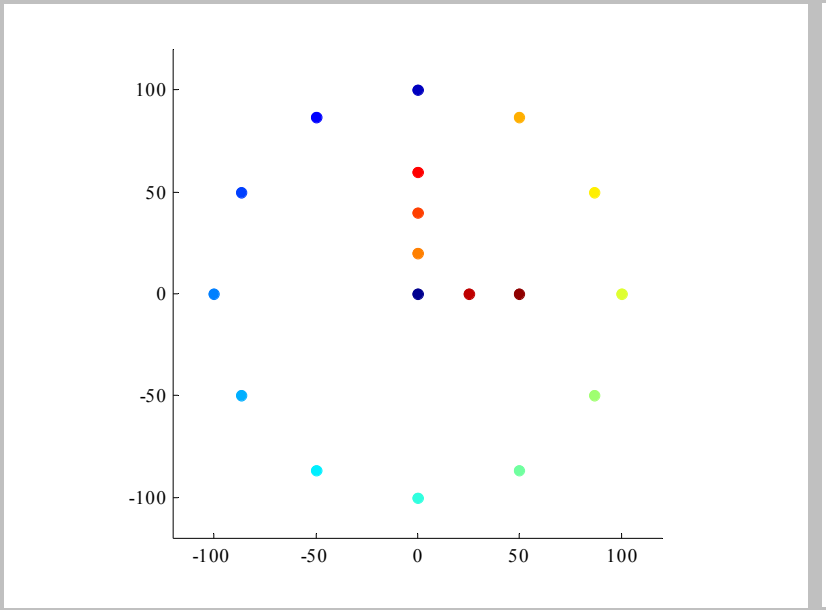
- Figura „clock”: wizualizacja (wykres rozrzutu) nowych punktów



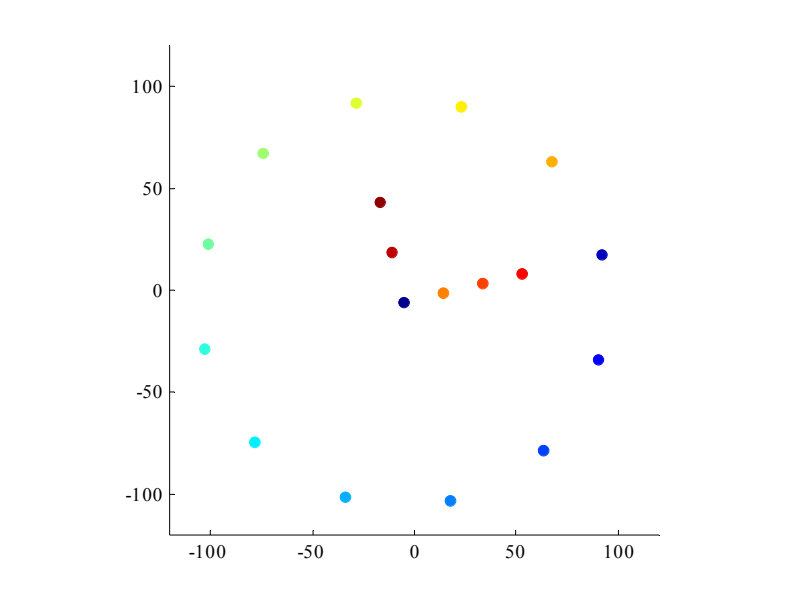
Wizualizacja danych wielowymiarowych

- Figura „clock”: porównanie konfiguracji

oryginał



wynik MDS



Wizualizacja danych wielowymiarowych

- W sytuacji, gdy znamy oryginalną konfigurację punktów (a nie tylko macierz odległości pomiędzy tymi punktami), rozwiązanie problemu MDS uznajemy za dokładne, gdy konfigurację nowych punktów można doprowadzić do konfiguracji oryginalnych punktów za pomocą złożenia następujących przekształceń
  - izometrii
    - obroty
    - symetrie osiowe
    - przesunięcia
  - (równomiernego) przeskalowania (wszystkich odległości)
- Przekształcenie będące złożeniem powyższych przekształceń jest nazywane przekształceniem dopuszczalnym (ang. feasible)

Wizualizacja danych wielowymiarowych

- Różne aspekty MDS
  - funkcja celu (w rozmaitych konkretnych postaciach) nosi ogólną nazwę stresu (ang. stress)
  - aby unikać pierwiastków przy obliczaniu odległości, stosuje się podnoszenie do kwadratu odległości oryginalnych
  - MDS dostarcza nietrywialnych zadań optymalizacyjnych o znanych rozwiązaniach (dokładnych/przybliżonych)
  - przedstawione wersje metody noszą ogólnie nazwę metrycznych (ang. metric), alternatywę stanowią tzw. wersje porządkowe (ang. ordinal)
  - porównywaniem rozwiązań znalezionych przez MDS zajmuje się tzw. analiza Prokrustesa

...

Wizualizacja danych wielowymiarowych

- Przykład obliczeniowy
  - oryginalna konfiguracja punktów nieznana
  - (czyli: typowe MDS)

Wizualizacja danych wielowymiarowych

- Dane są:
  - macierz odległości docelowych 3x3: **D =**

0	3	4
3	0	5
4	5	0

    - trzy obiekty
  - docelowy wymiar mapy: 2

Wizualizacja danych wielowymiarowych

- Tworzone są:
  - macierz zmiennych: **X** = $\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix}$
    - trzy punkty
      - każdy o dwóch współrzędnych
  - macierz odległości aktualnych **A** = pdist(**X**) = $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$,

    gdzie $a_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$

    - odległości pomiędzy punktami o współrzędnych postaci $\mathbf{x}_i = [x_{i1}, x_{i2}]^T$
  - norma $||\mathbf{A} - \mathbf{D}|| = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}\left(a_{ij} - d_{ij}\right)^2} = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2} - d_{ij}\right)^2}$
    - funkcja celu

Wizualizacja danych wielowymiarowych

- Powstały problem programowania
  - funkcja celu: $s(\mathbf{X}) = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2}-d_{ij}\right)^2}$
    - liczba zmiennych: 6
    - liczba ograniczeń: 0

  (charakter problemu: nieliniowy bez ograniczeń)

Wizualizacja danych wielowymiarowych

- Operacje upraszczające problem (eliminacja pierwiastkowania)
  - funkcja oryginalna:

$$s(\mathbf{X}) = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2}-d_{ij}\right)^2}$$

  - bez pierwiastka „zewnętrznego”

$$s'(\mathbf{X}) = \sum_{i=1}^{3}\sum_{j=1}^{3}\left(\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2}-d_{ij}\right)^2$$

  - bez pierwiastków „wewnętrznych”

$$s''(\mathbf{X}) = \sum_{i=1}^{3}\sum_{j=1}^{3}\left((x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2-d_{ij}^2\right)^2$$

Wizualizacja danych wielowymiarowych

- Funkcja celu (pełna wersja) w przypadku ogólnym
  - 5 obiektów wejściowych, 2 wymiary wyjściowe

$$s(\mathbf{X}) = \sqrt{\sum_{i=1}^{5}\sum_{j=1}^{5}\left(\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2}-d_{ij}\right)^2}$$

  - m obiektów wejściowych, 2 wymiary wyjściowe

$$s(\mathbf{X}) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2}-d_{ij}\right)^2}$$

Wizualizacja danych wielowymiarowych

- Funkcja celu (pełna wersja) w przypadku ogólnym
  - m obiektów wejściowych, 3 wymiary wyjściowe

$$s(\mathbf{X}) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+(x_{i3}-x_{j3})^2}-d_{ij}\right)^2}$$

  - m obiektów wejściowych, n wymiarów wyjściowych

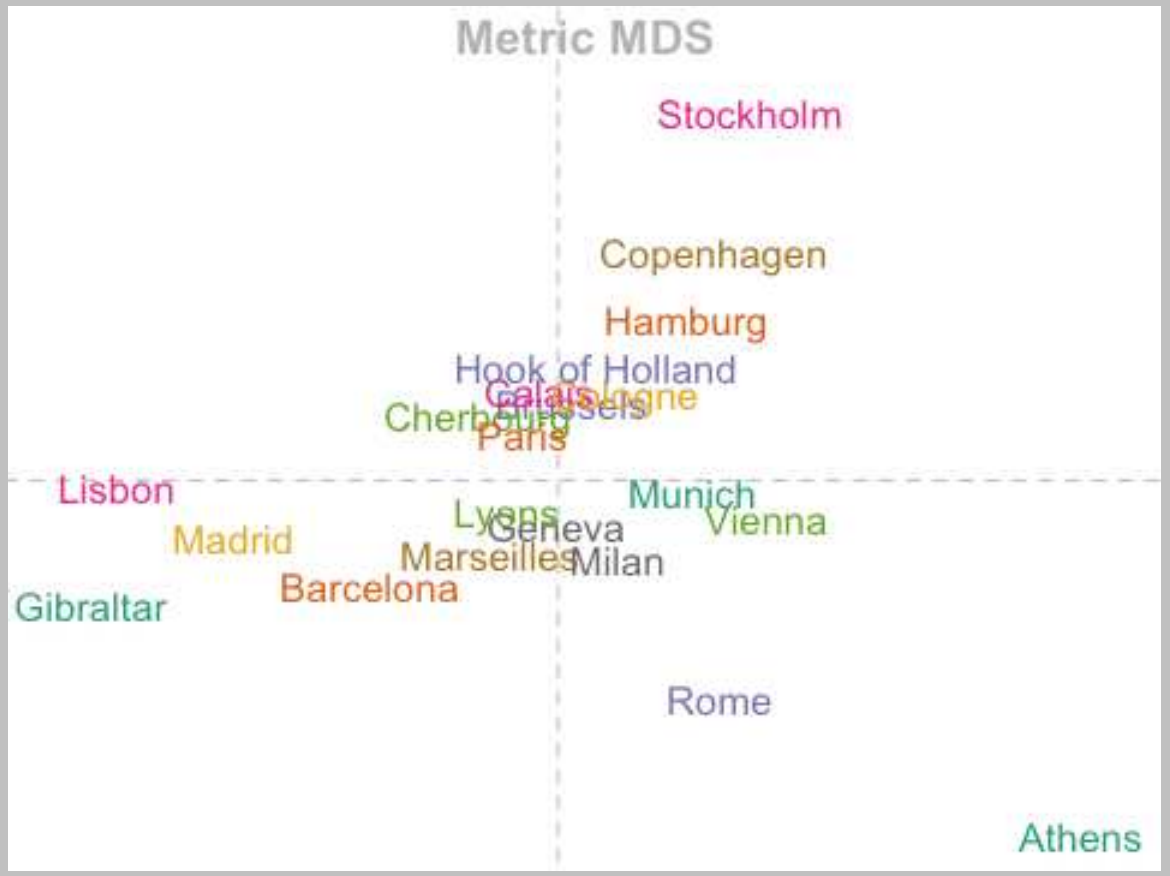
$$s(\mathbf{X}) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2}-d_{ij}\right)^2}$$

…

Przykładowe zastosowania

- Mapa wybranych miast w Europie (odległości drogowe)

Przykładowe zastosowania

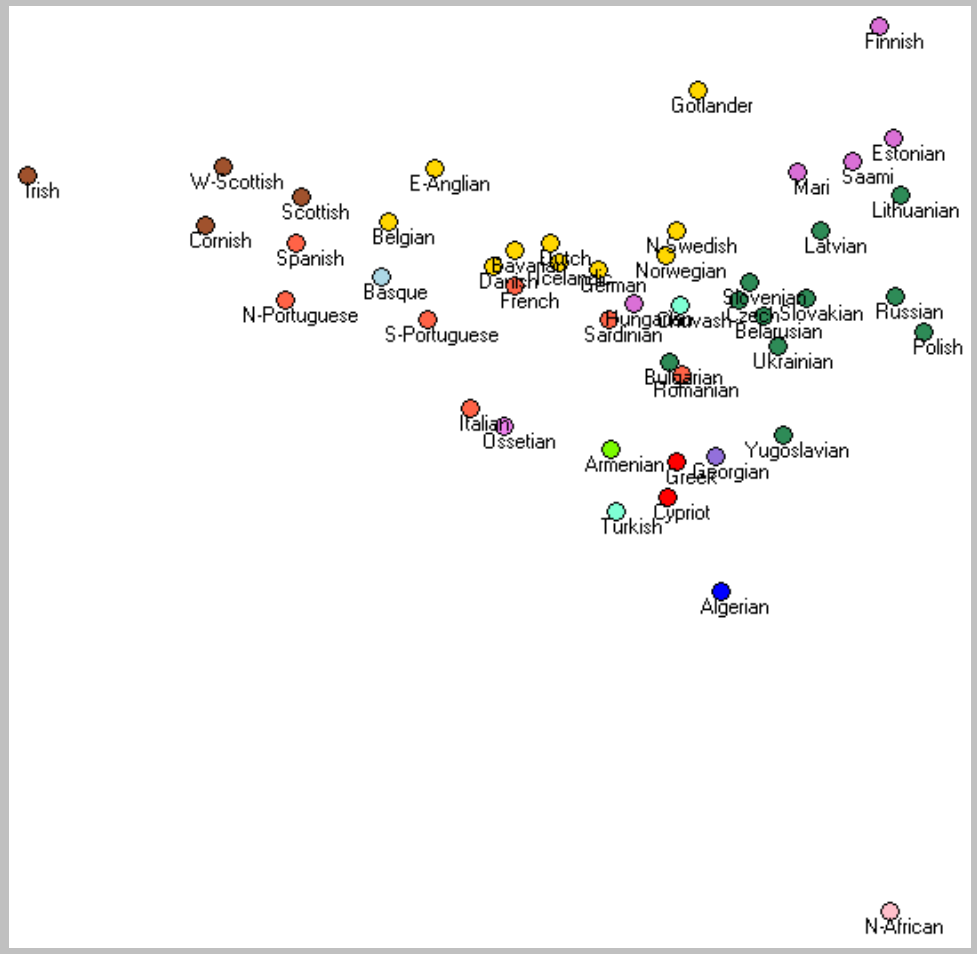


- Mapa odległości drogowych pomiędzy wybranymi miastami w Europie

Przykładowe zastosowania

- Mapa różnic genetycznych u Europejczyków (+części Afryki Płn)
  - kolor: grupa językowa

Przykładowe zastosowania

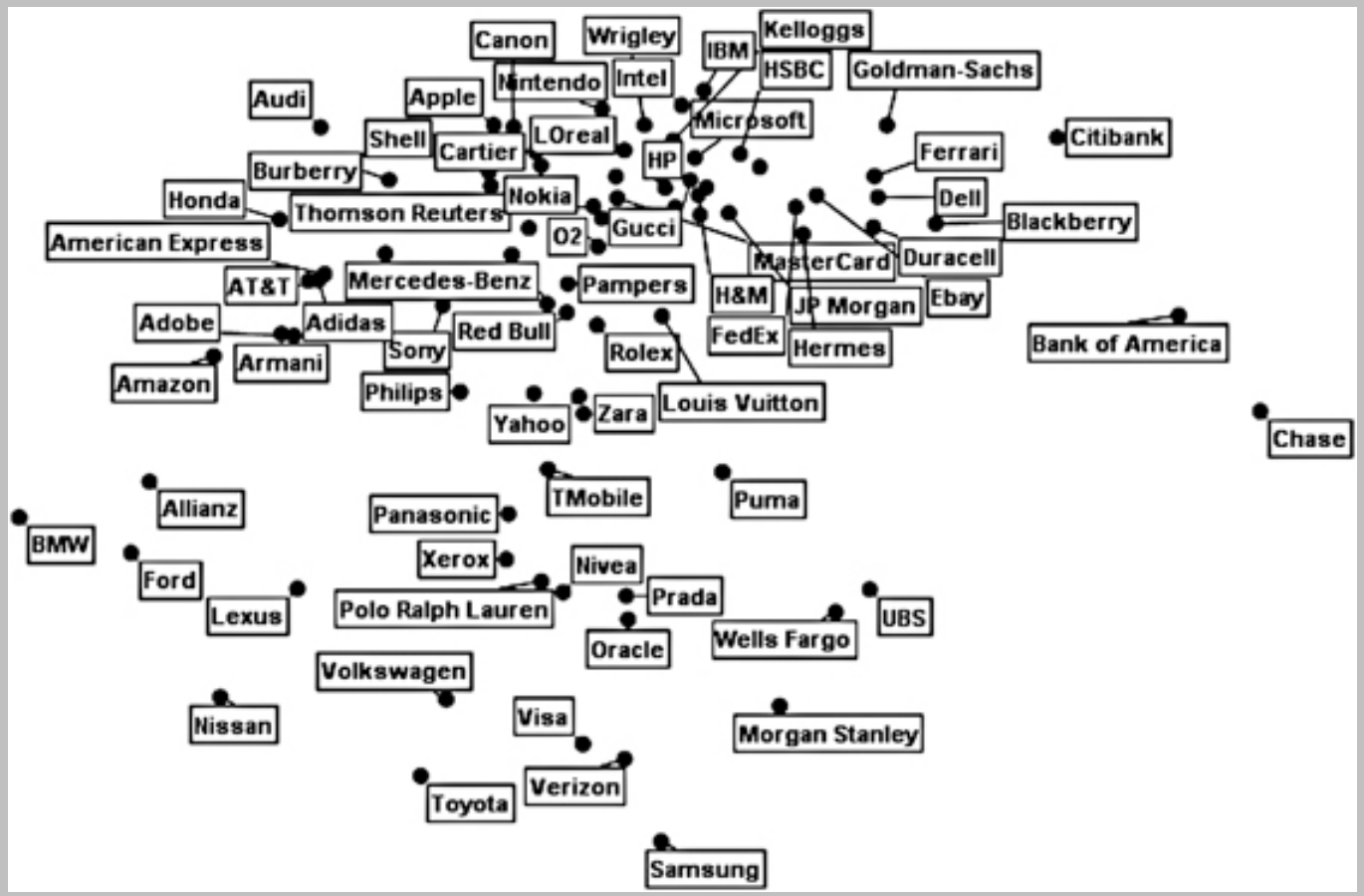


- http://andrewgelman.com/wp-content/uploads/2007/06/from_geo1.png

Przykładowe zastosowania

- Mapa postrzegania marek produktów

Przykładowe zastosowania

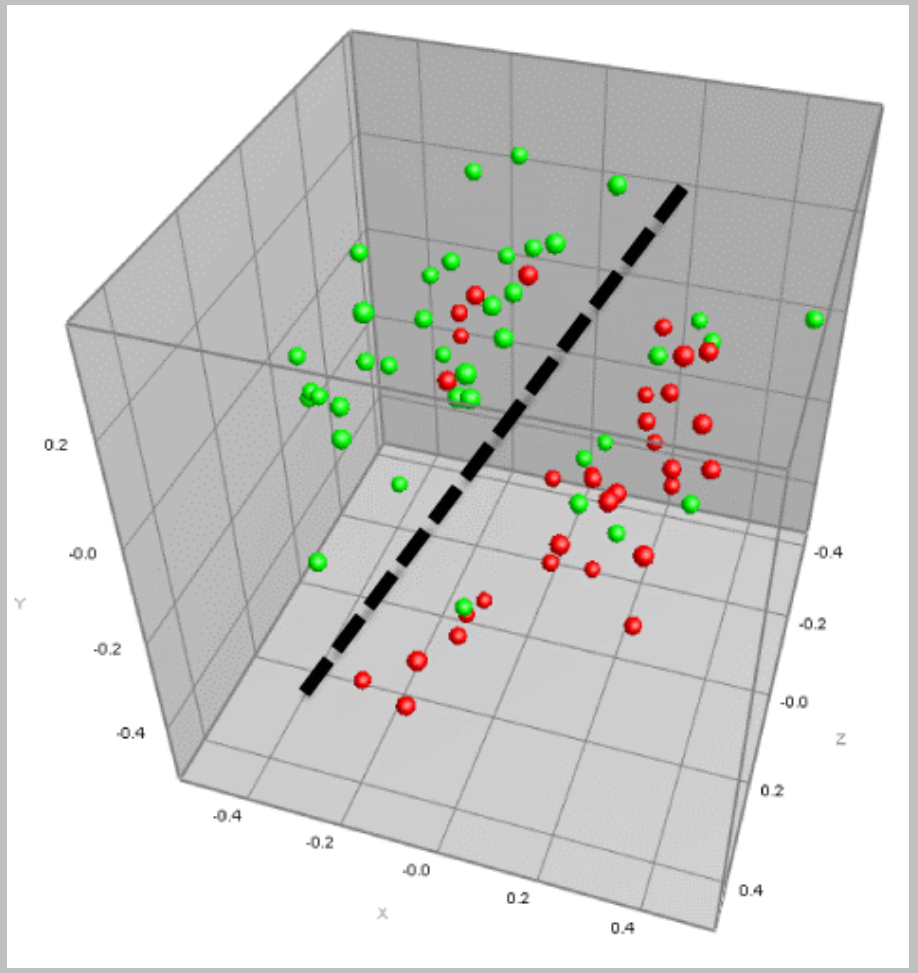


- http://www.palgrave-journals.com/bm/journal/v19/n6/images/bm201156f2.jpg

Przykładowe zastosowania

- Mapa wybranych patogenów człowieka

Przykładowe zastosowania



- http://www.omicsonline.org/evaluation-of-a-flow-through-depuration-system-to-eliminate%20the-human-pathogen-vibrio-vulnificus-from-oysters-2155-9546.1000103.php?aid=1517