# Skalowanie wielowymiarowe

Maciej Nasiński, Paweł Strawiński

Zajęcia 14
26 stycznia 2023

- Skalowanie wielowymiarowe (ang. *MultiDimension Scaling*) jest to technika statystyczna wykorzystywana do redukcji wymiarowości danych oraz ich wizualizacji
- Jest to zestaw technik porządkowania danych (ang. *ordination techniques*) stosowanych w wizualizacji informacji, w szczególności do wyświetlania informacji zawartych w macierzy odległości.
- Miary niepodobieństwa między obserwacjami w przestrzeni wielowymiarowej są reprezentowane w przestrzeni niskowymiarowej (zwykle w dwuwymiarowej), w taki sposób, że miary odległości w przestrzeni niskowymiarowej są zbliżone do miar odległości w przestrzeni wielowymiarowej.

- W praktyce nie oblicza się miar (nie)podobieństwa między obiektami a definiuje poprzez zbiór cech obiektów (zmiennych)
- Najczęściej wykorzystywana jest odległość L2 (euklidesowa) lub L1 (miejska)
- Zazwyczaj liczba obserwacji przekracza liczbę zmiennych, ale nie jest to warunek konieczny dla przeprowadzenia skalowania wielowymiarowego

# Klasyczne skalowanie wielowymiarowe

- Jest także określane jako *Principal Coordinates Analysis (PCoA)*
- Jeśli odległości są odległościami euklidesowymi, klasyczny MDS daje łatwe rozwiązanie algebraiczne
- Na podstawie macierzy niepodobieństwa obliczana jest wartość funkcji kryterium nazywanej *strain*

$$Strain(x_1, \ldots, x_n) = \left( \frac{\sum_{ij} (b_{ij} - x_i' x_j)^2}{\sum_{ij} b_{ij}^2} \right)^{1/2}$$

- Elementy $b_{ij}$ są wyliczane wg algorytmu

# Algorytm

- Klasyczne skalowanie wielowymiarowe zakłada odległość Euklidesową L2.
- Klasyczne skalowanie wielowymiarowe wykorzystuje fakt, że macierz $X$ można uzyskać poprzez przekształcenie macierzy $B = XX'$ na macierz wartości własnych
- Macierz $B$ jest obliczana z macierzy podobieństwa poprzez podwójne centrowanie
  1. Wyznacz macierz odległości $D = [d_{ij}^2]$
  2. Zastosuj podwójne centrowanie $B = -\frac{1}{2}CDC$, gdzie $C = I - \frac{1}{n}Jn$, gdzie $n$ to liczba obserwacji, $J$ oznacza macierz 1.
  3. Wyznacz $m$ jako największą wartość własną macierzy $B$
  4. Nowe $X = E_m \Lambda_m^{1/2}$. $E_m$ to macierz $m$ wektorów własnych, $\Lambda_m$ to diagonalna macierz $m$ wartości własnych macierzy $B$

# Metryczne skalowanie wielowymiarowe

- Jest uogólnieniem procedury optymalizacji na różne funkcje strat i macierze wejściowe o znanych odległościach z wagami
- Wykorzystywana jest funkcją kryterium straty nazywana stres
- Zazwyczaj minimalizowana jest się z wykorzystaniem procedury *stress majorization*.
- Metryczny MDS minimalizuje funkcję kryterium stress, która jest resztową sumą kwadratów

$$Stress(x_1, \ldots, x_n) = \sqrt{\sum_{i \neq j = 1, \ldots, N} (d(i,j) - ||x_i - x_j||)^2}$$