# missMDA

Pakiet do metod wielowymiarowych z brakami danych

Weronika Dyszkiewicz
Krzysztof Kłopotowski
Antoni Chudy

# Co ten pakiet robi?

- imputowanie brakujących wartości w zbiorach danych ciągłych przy użyciu modelu PCA, zbiorach danych kategorycznych przy użyciu MCA, danych mieszanych przy użyciu FAMD,  grupie zmiennych ciągłych lub kategorycznych lub które mogą być tablicą kontyngencji MFA
- generowanie wielu imputowanych zbiorów dla danych ciągłych przy użyciu modelu PCA i dla danych kategorycznych przy użyciu MCA
- wizualizacja imputacji wielokrotnej w PCA i MCA

# Czemu missMDA, a nie coś innego?

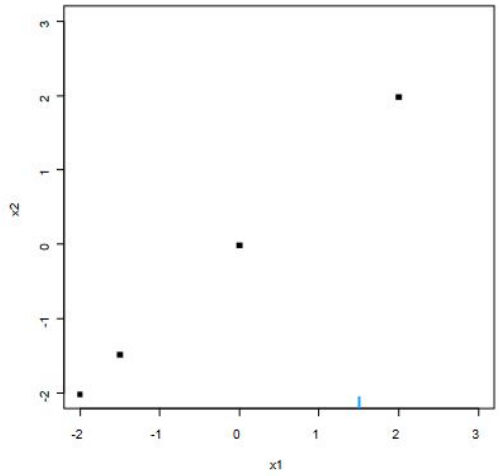
- większość metod głównych składowych dostępnych w pakietach mogą być stosowane tylko do kompletnych zbiorów danych
- graficzna reprezentacja, mimo braków danych
- działanie na różnych typach zmiennych (przypisywanie zmiennych nawet kategorycznych)
- pozwala wykonać standaryzowane PCA z brakami danych
- skalowanie jako część analizy, a nie pre-processingu
- graficzne przedstawienie przedziałów ufności
- dostarcza rozwiązań dla wyboru liczby wymiarów bazowych z niekompletnych zbiorów danych

# PCA (Principal component analysis)

- najlepiej najpierw sprawdzić czy między zmiennymi występuje tendencja liniowa
- metoda dla ciągłych zbiorów
- brakujące wartości są przewidywane przy użyciu iteracyjnego algorytmu PCA dla zdefiniowanej wcześniej liczby wymiarów
- na koniec algorytmu otrzymujemy zarówno parametry PCA z niekompletnego zbioru jak i uzupełniony zbiór
- gdy mamy nadmierne dopasowanie (np. gdy za dużo braków/dane są nieznaczące) stosujemy metody regularyzacji

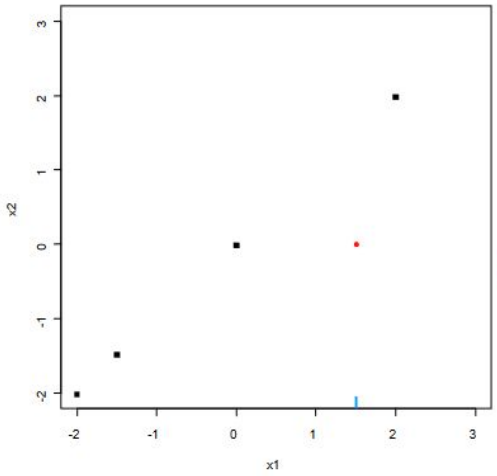
# Iterative PCA algorithm

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | NA |
| 2.0 | 1.98 |




| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | NA |
| 2.0 | 1.98 |

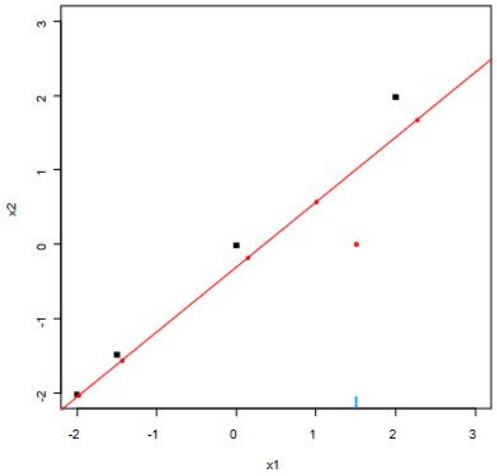
| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.00 |
| 2.0 | 1.98 |




| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | NA |
| 2.0 | 1.98 |

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.00 |
| 2.0 | 1.98 |

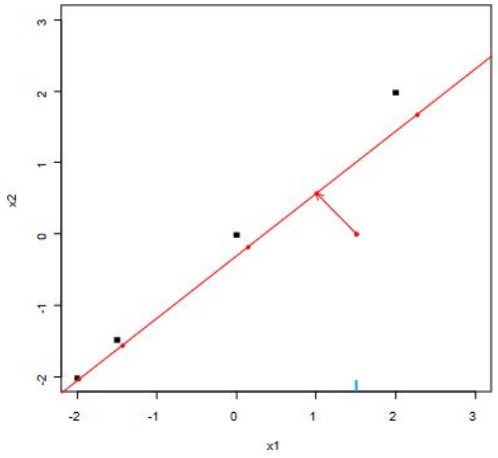
| ^x1 | ^x2 |
|---|---|
| -1.98 | -2.04 |
| -1.44 | -1.56 |
| 0.15 | -0.18 |
| 1.00 | 0.57 |
| 2.27 | 1.67 |



| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | NA |
| 2.0 | 1.98 |

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.00 |
| 2.0 | 1.98 |

| ^ x1 | ^ x2 |
|---|---|
| -1.98 | -2.04 |
| -1.44 | -1.56 |
| 0.15 | -0.18 |
| 1.00 | 0.57 |
| 2.27 | 1.67 |

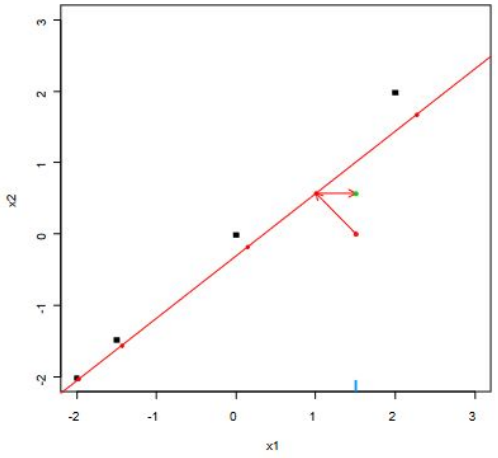



| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | NA |
| 2.0 | 1.98 |

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.00 |
| 2.0 | 1.98 |

| ^ x1 | ^ x2 |
|---|---|
| -1.98 | -2.04 |
| -1.44 | -1.56 |
| 0.15 | -0.18 |
| 1.00 | 0.57 |
| 2.27 | 1.67 |

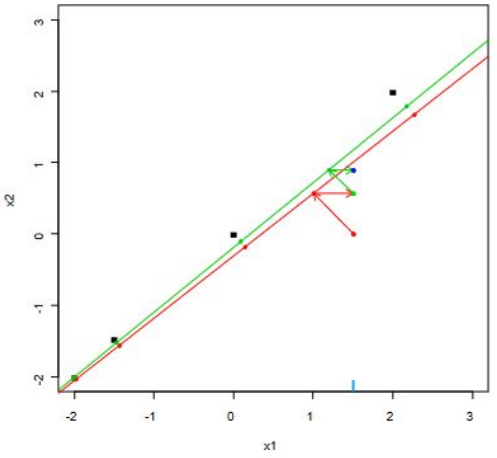
| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.57 |
| 2.0 | 1.98 |




| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | NA |
| 2.0 | 1.98 |

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.57 |
| 2.0 | 1.98 |

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.57 |
| 2.0 | 1.98 |

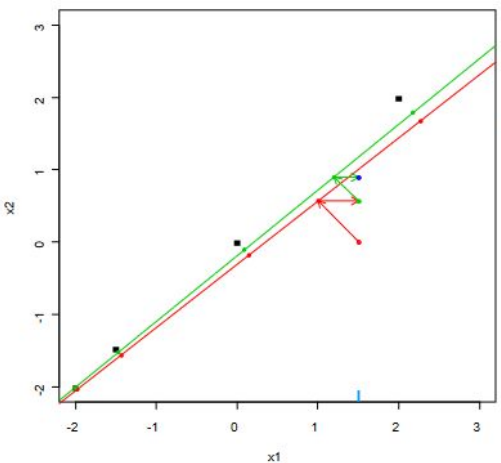



| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | NA |
| 2.0 | 1.98 |

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.57 |
| 2.0 | 1.98 |

| ^ x1 | ^ x2 |
|---|---|
| -2.00 | -2.01 |
| -1.47 | -1.52 |
| 0.09 | -0.11 |
| 1.20 | 0.90 |
| 2.18 | 1.78 |

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.90 |
| 2.0 | 1.98 |

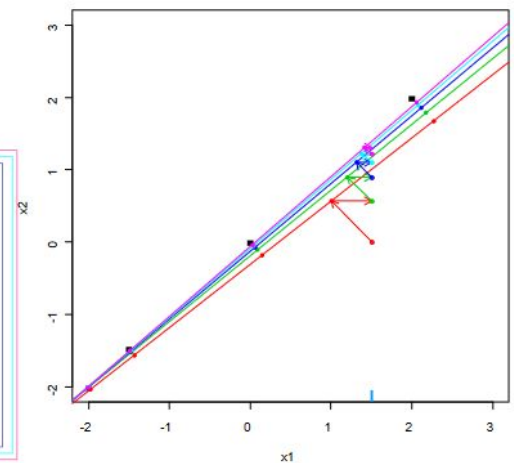



| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | NA |
| 2.0 | 1.98 |

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.00 |
| 2.0 | 1.98 |

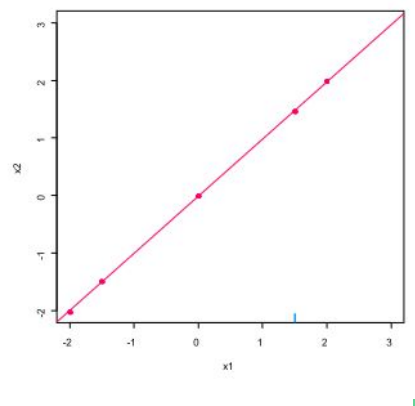
| ^ x1 | ^ x2 |
|---|---|
| -1.98 | -2.04 |
| -1.44 | -1.56 |
| 0.15 | -0.18 |
| 1.00 | 0.57 |
| 2.27 | 1.67 |

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 0.57 |
| 2.0 | 1.98 |

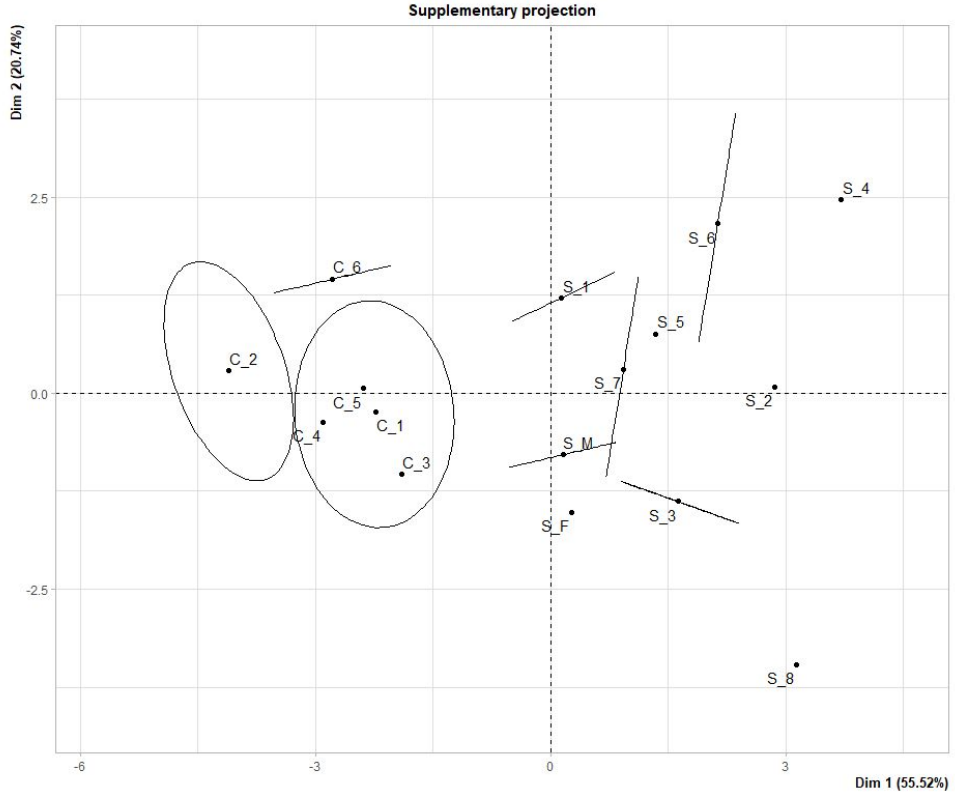



| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | NA |
| 2.0 | 1.98 |

| x1 | x2 |
|---|---|
| -2.0 | -2.01 |
| -1.5 | -1.48 |
| 0.0 | -0.01 |
| 1.5 | 1.46 |
| 2.0 | 1.98 |



# Obszary ufności



# MCA (Multiple correspondence analysis), a.k.a. kategoryczne PCA

- brakami danych zajmujemy się tak samo jak w PCA z kilkoma różnicami, takimi jak wyważenie zmiennych
- na początku uzupełniamy brakujące wartości np. proporcjonalnie do wielkości kategorii (analog do imputowania średnią wartością dla ciągłych zmiennych)
- potem wykonujemy MCA ze zfittowanymi wartościami i tak jak w PCA, kolejne iteracje coraz lepiej aproksymują model i estymują braki
- oczywiście można regularyzować

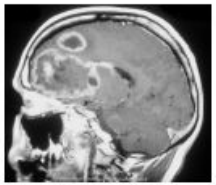
# FAMD (Factorial analysis for mixed data)

- w teorii można skategoryzować zmienne ciągłe i robić MCA, ale wtedy tracimy sporo informacji :c
- lepiej jest traktować zmienne kategoryczne jako ciągłe, przeskalować je, zestandaryzować zmienne ciągłe i zrobić PCA, ważąc odpowiednio wpływ obu typów zmiennych

# MFA (Multiple factor analysis)

- rozszerzamy przypadek jednej tabeli danych, zmienne są określane przez grupy cech, dane pochodzą z różnych źródeł
- chcemy badać podobieństwa między wierszami z wielowymiarowego punktu widzenia (porównywać informacje dostarczoną przez każdą z grup)
- duże ryzyko wystąpienia braków danych
- wyniki w postaci grafów interpretujemy podobnie jak dla PCA

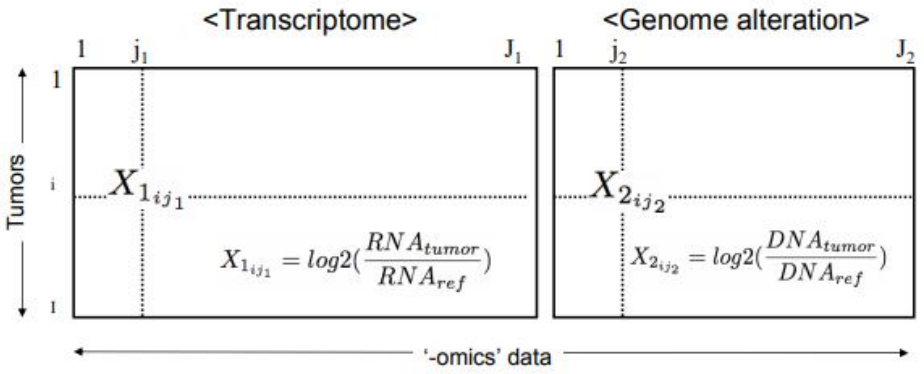
**Gliomas:** Brain tumors, WHO classification

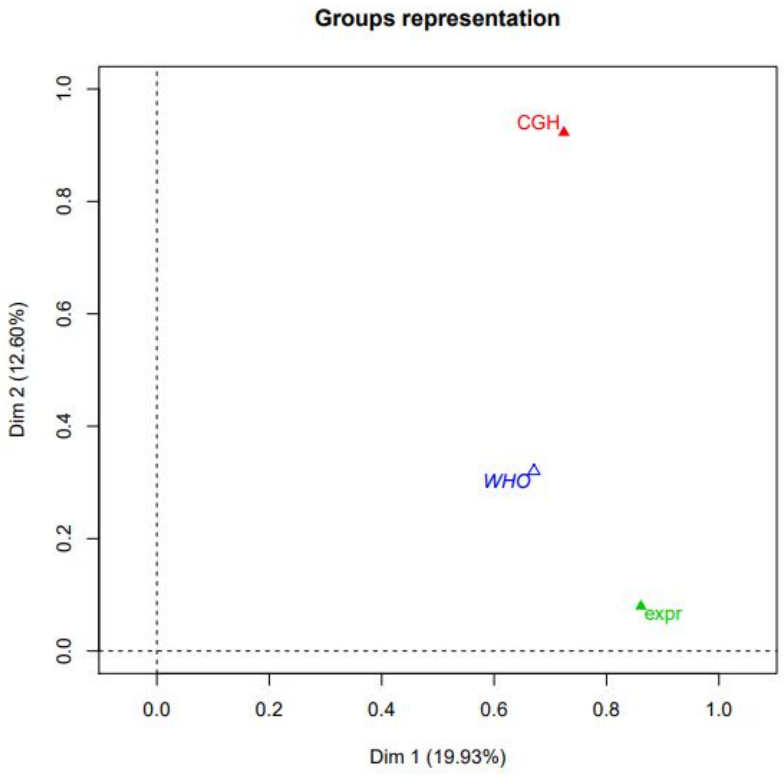


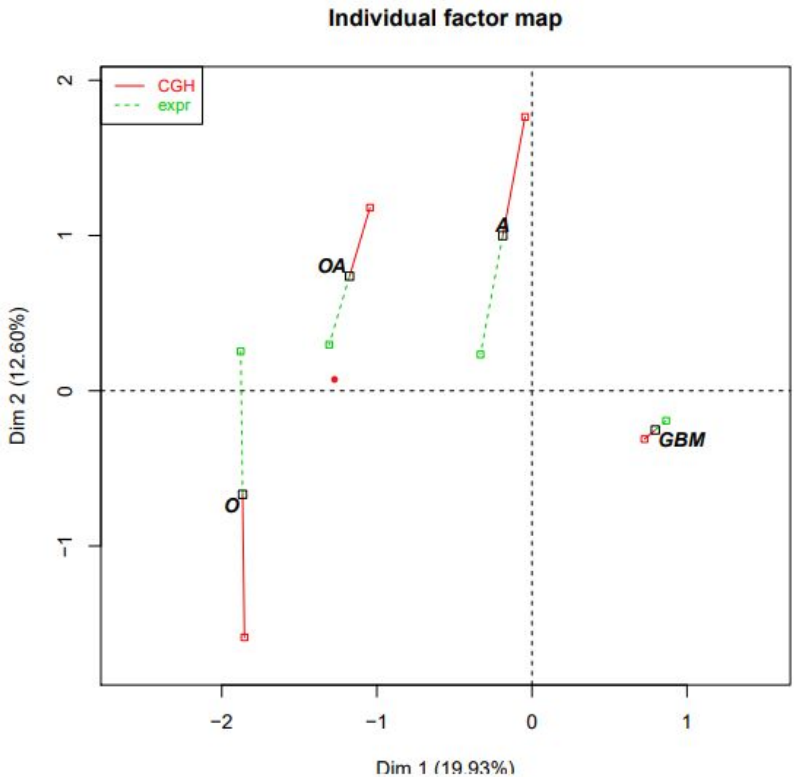
- astrocytoma (A) ..... x5
- oligodendroglioma (O) ..... x8
- oligo-astrocytoma (OA) ..... x6
- glioblastoma (GBM) ..... x24

43 tumor samples

(Bredel *et al.*,2005)




Groups representation
CGH
WHO
expr
Dim 1 (19.93%)
Dim 2 (12.60%)
0.0
0.2
0.4
0.6
0.8
1.0


Individual factor map
CGH
expr
A
OA
O
GBM
Dim 1 (19.93%)
Dim 2 (12.60%)
−2
−1
0
1
2

# Single/multiple imputation

- głównym celem przedstawianych metod było wykonanie Principial Component Methods pomimo braków danych
- dobra jakość dopasowania oraz alternatywa dla innych metod

# Bibliografia


- Josse, J., & Husson, F. (2016). missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. Journal of Statistical Software, 70(1), 1–31. https://doi.org/10.18637/jss.v070.i01
- https://cran.r-project.org/web/packages/missMDA/missMDA.pdf