Wrocław University of Science and Technology

"Exploratory data analysis of a clinical study group - revealing patient subgroups"

Bogumil M. Konopka

Department of Biomedical Engineering

Faculty of Fundamental Problems of Technology

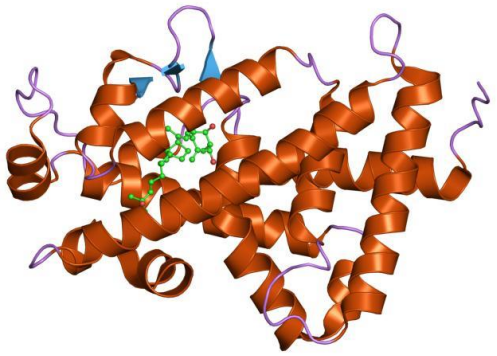
Wroclaw University of Science and Technology, POLAND

Assessing the influence of Vitamin D Receptor gene polymorphisms („changes” in DNA)

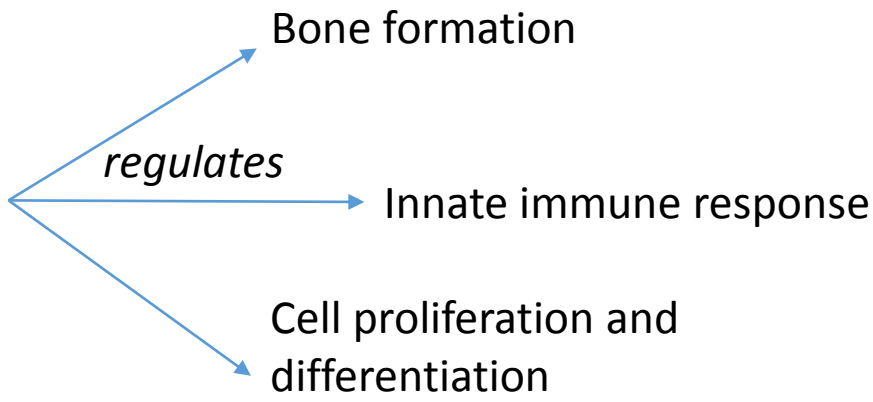
(Lukasz Laczmanski Ph.D., D. Sc. , Wroclaw Medical University)

UMW

Vitamin D Receptor



PDB_ID: 2hb8



- Our research hypothesis:
  - VDR polymorphisms influence sex hormone blood levels
- The plan:
  - **Explore gathered clinical data**
  - Build regression models relating VDR polymorphisms and blood levels of sex hormones

Dataset & questions

- Main questions:
  - Are there any outliers in the dataset?
  - What subgroups make up for the dataset?
  - What are the characteristics of particular subgroups?
  - What are the biological reasons that underlie such dataset structure?

515 samples

277 male ♂ 238 female ♀

44 attributes

Dataset & questions

- Main questions:
  - Are there any **outliers** in the dataset?
  - What **subgroups** make up for the dataset?
  - What are the **characteristics** of particular subgroups?
  - What are the **biological reasons** that underlie such dataset structure?

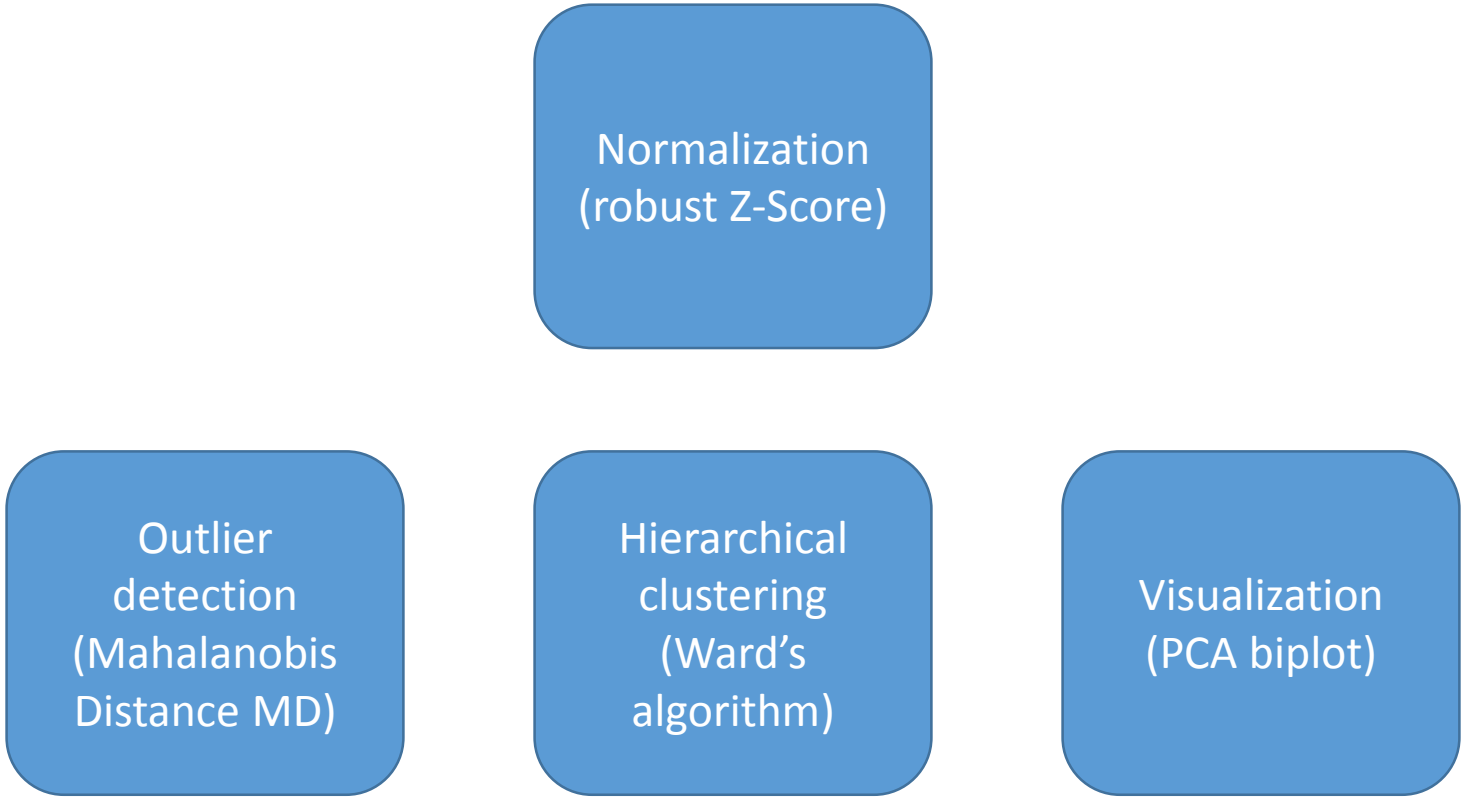
515 samples

277 male ♂ 238 female ♀

44 attributes

23 numerical	21 nominal
AGE, BMI, GLUCOSE, …	COUNTRY. REGION, YEAR.SEASON OBESITY,…

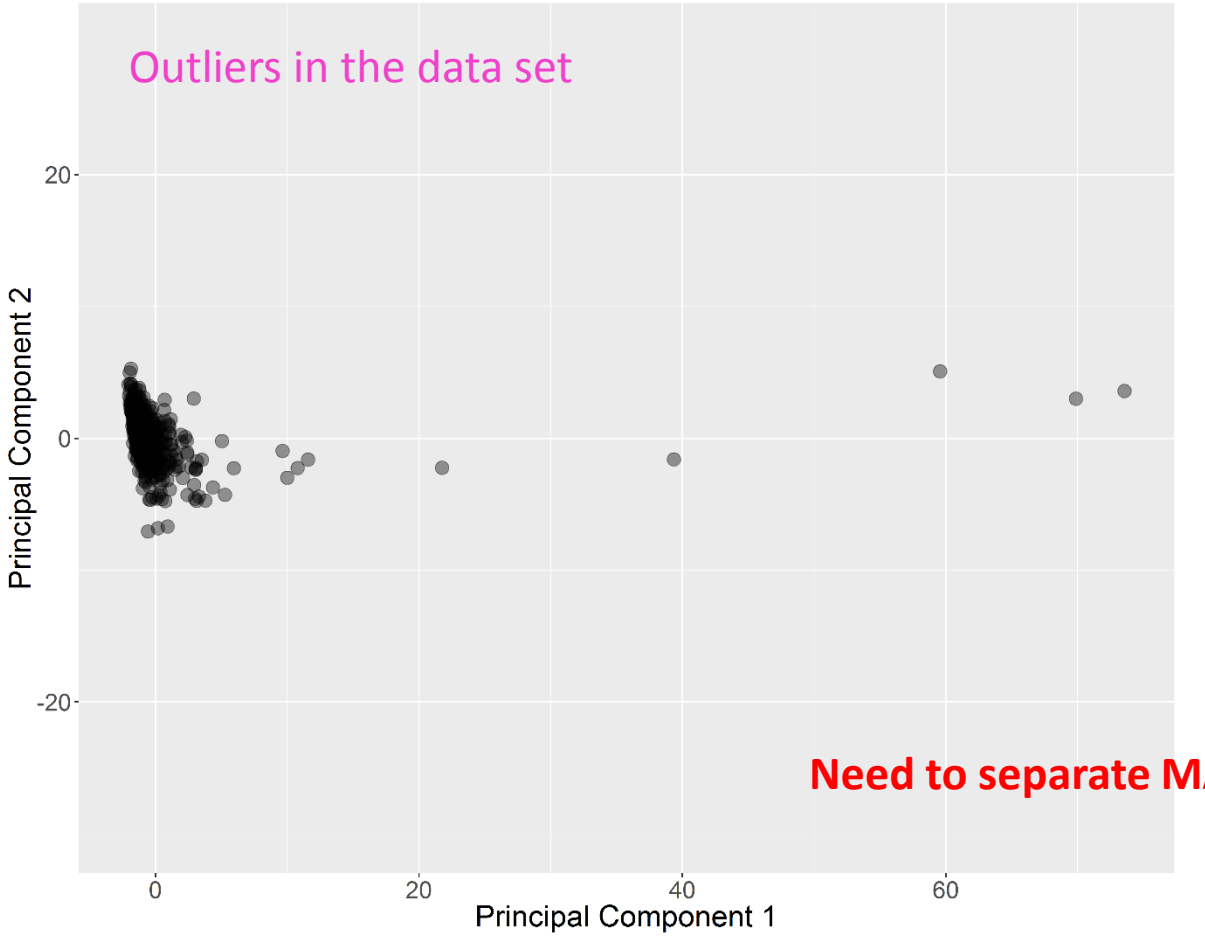
Overview of data processing procedure

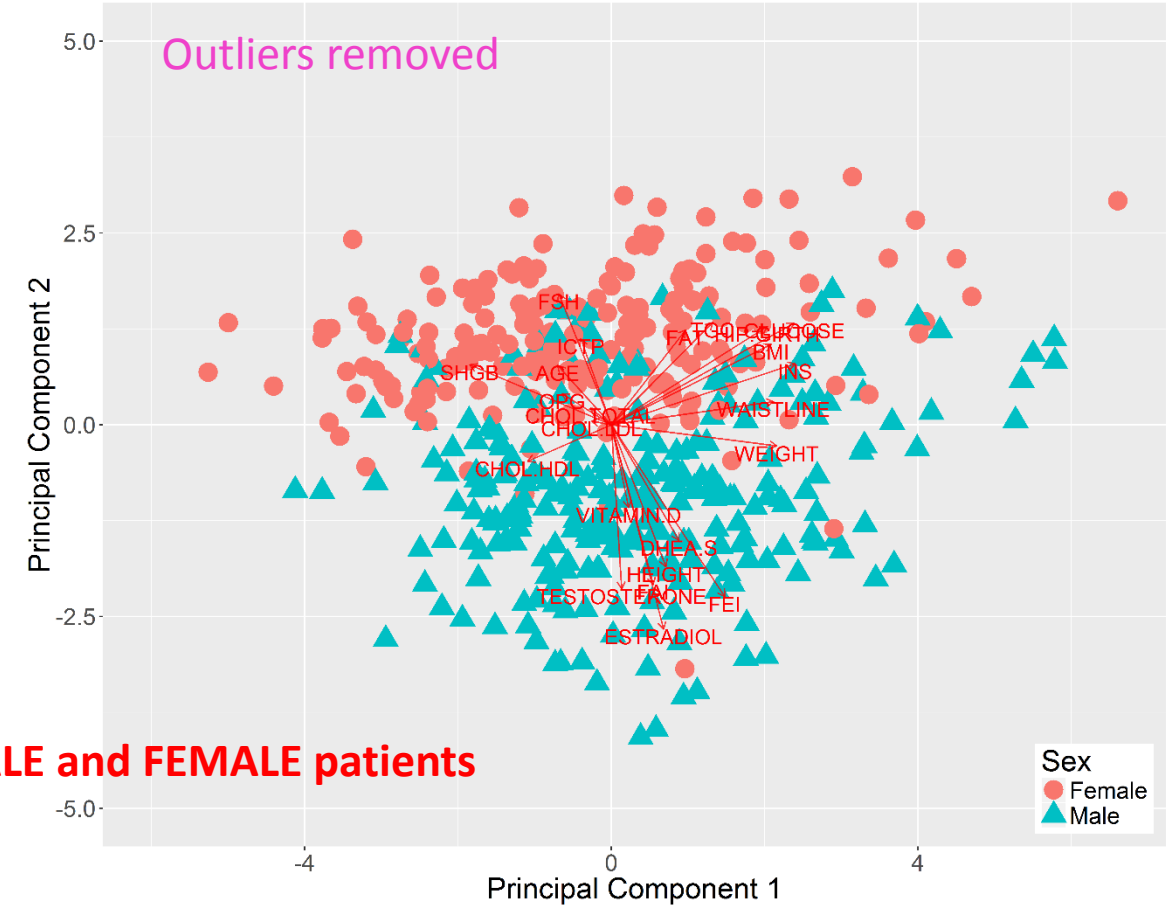


All based on data variance.

(Murtagh, Legendre, 2014)

Introductory analysis





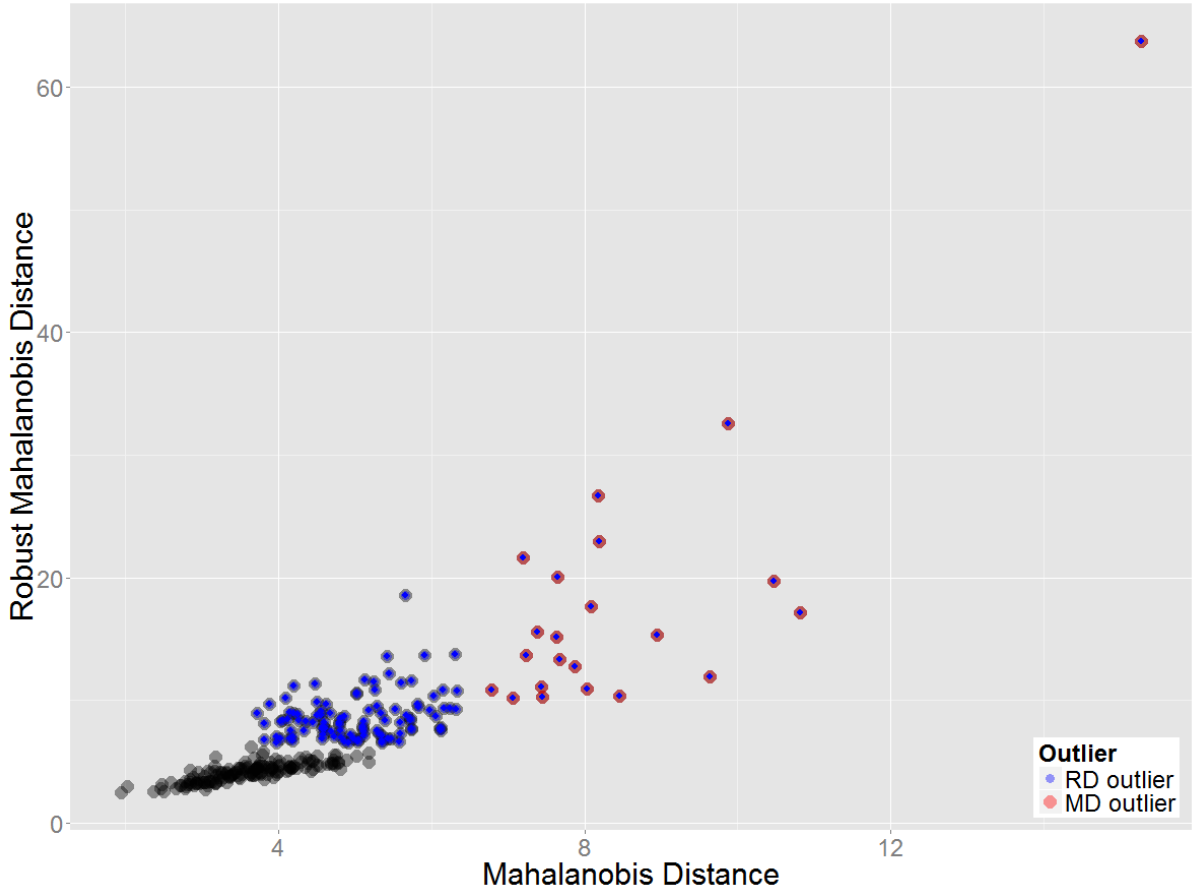
Need to separate MALE and FEMALE patients

Male set analysis - outlier removal

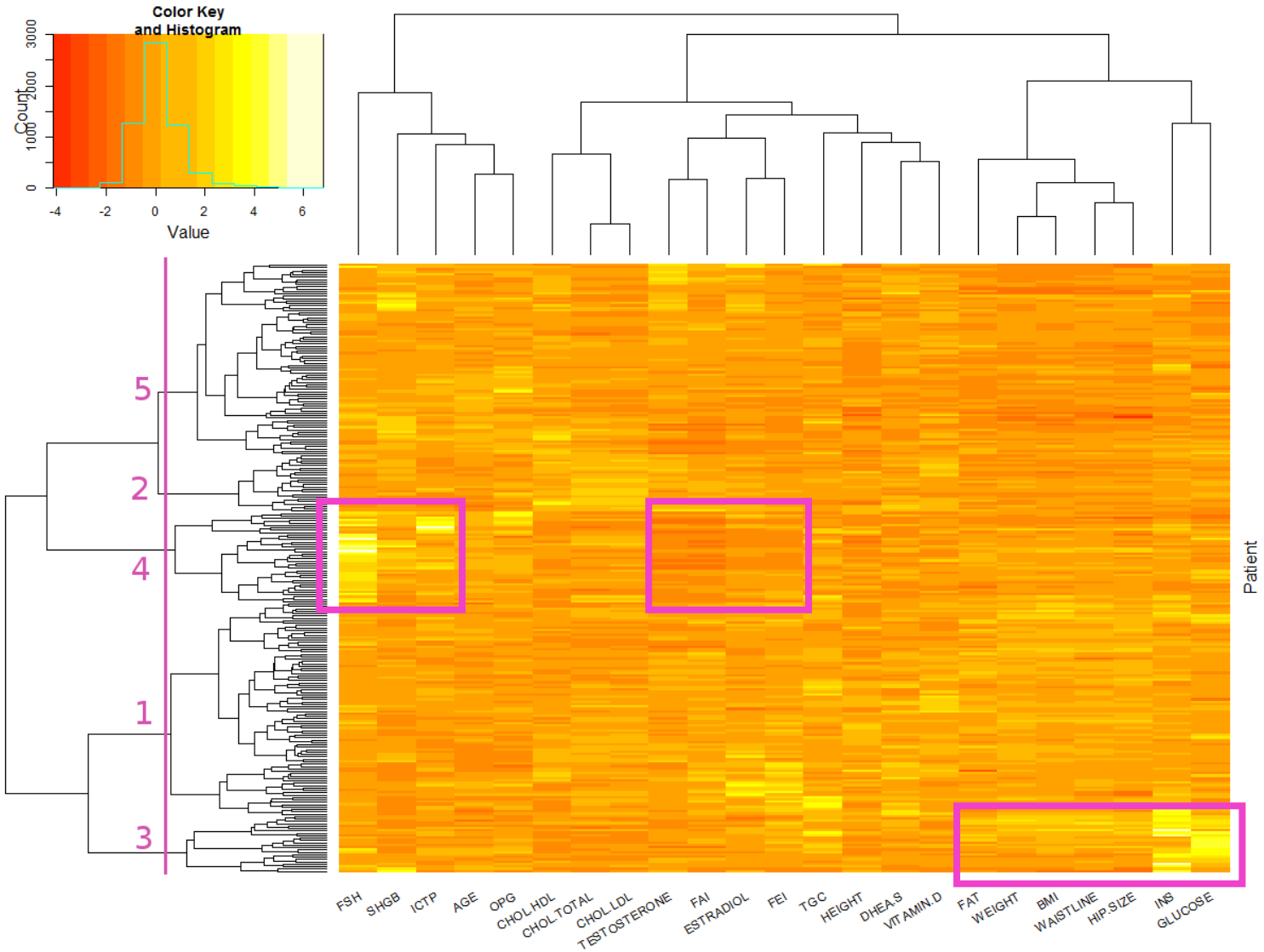
- Outliers based on **Mahalanobis Distance (MD)**:
  - **22 patients**

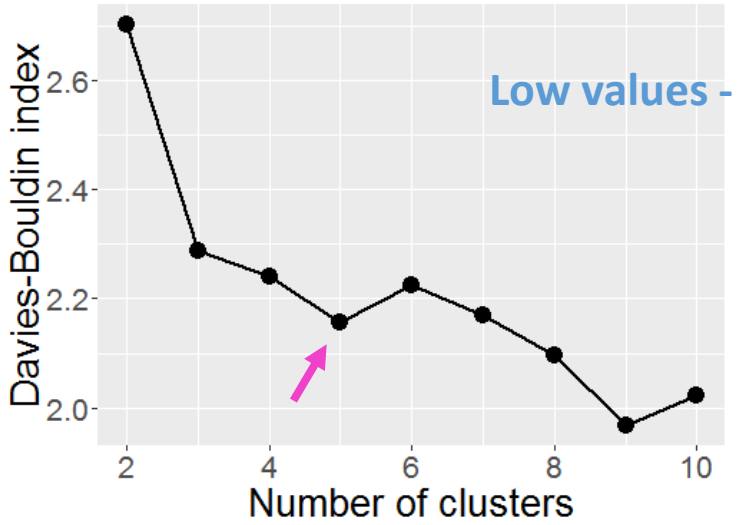
- Outlier detection with **robust Mahalanobis Distance (rMD)**:
  - **124 patients**

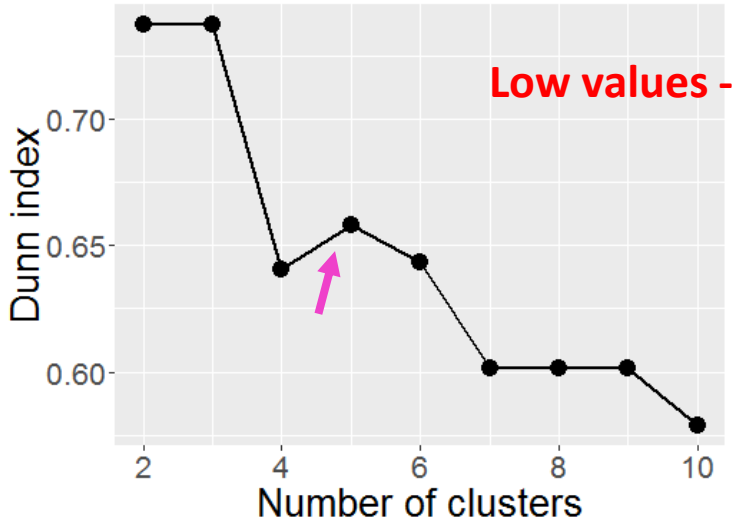
The data set is heterogeneous and may contain subgroups



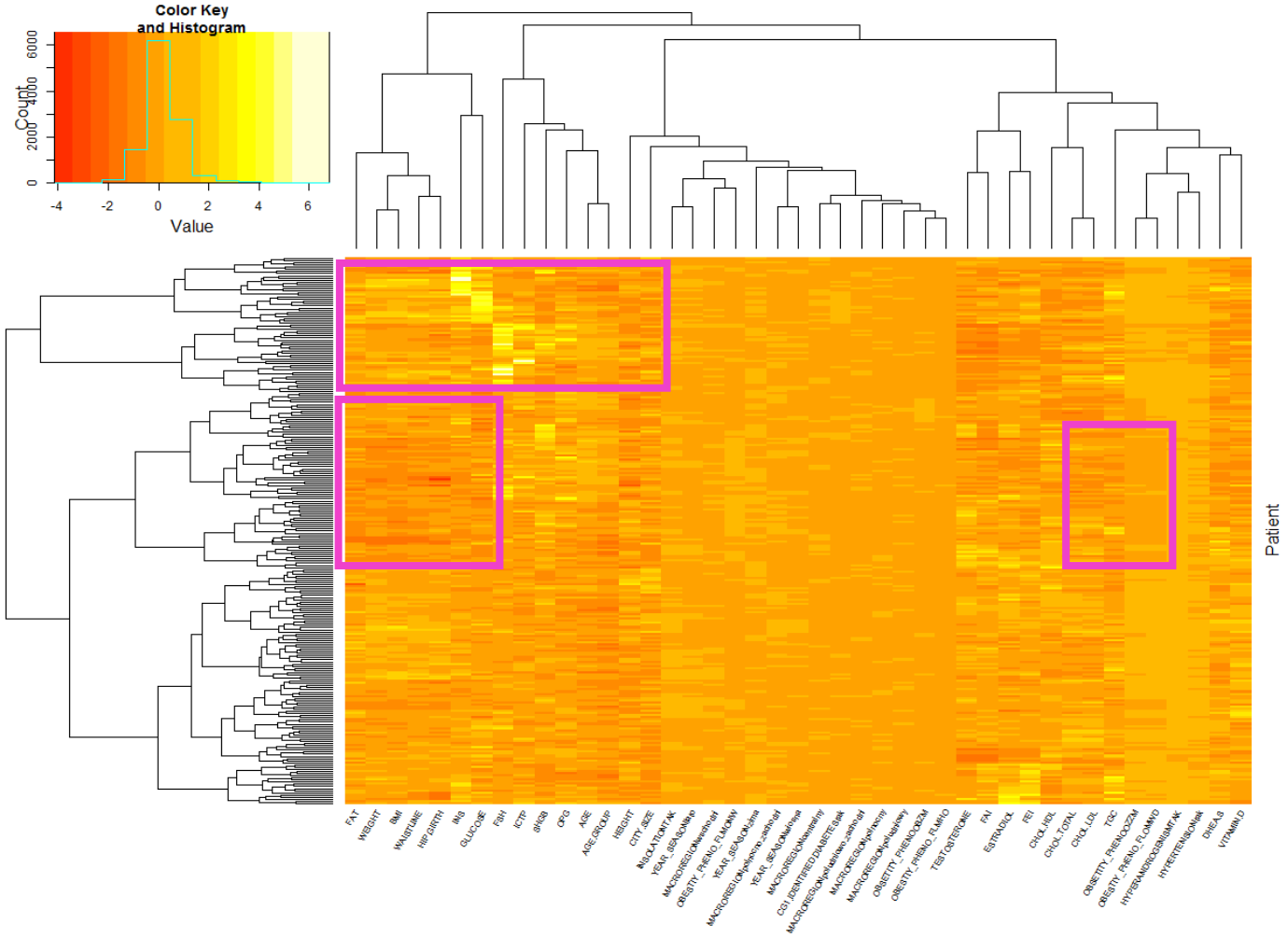
Male set analysis – hierarchical clustering



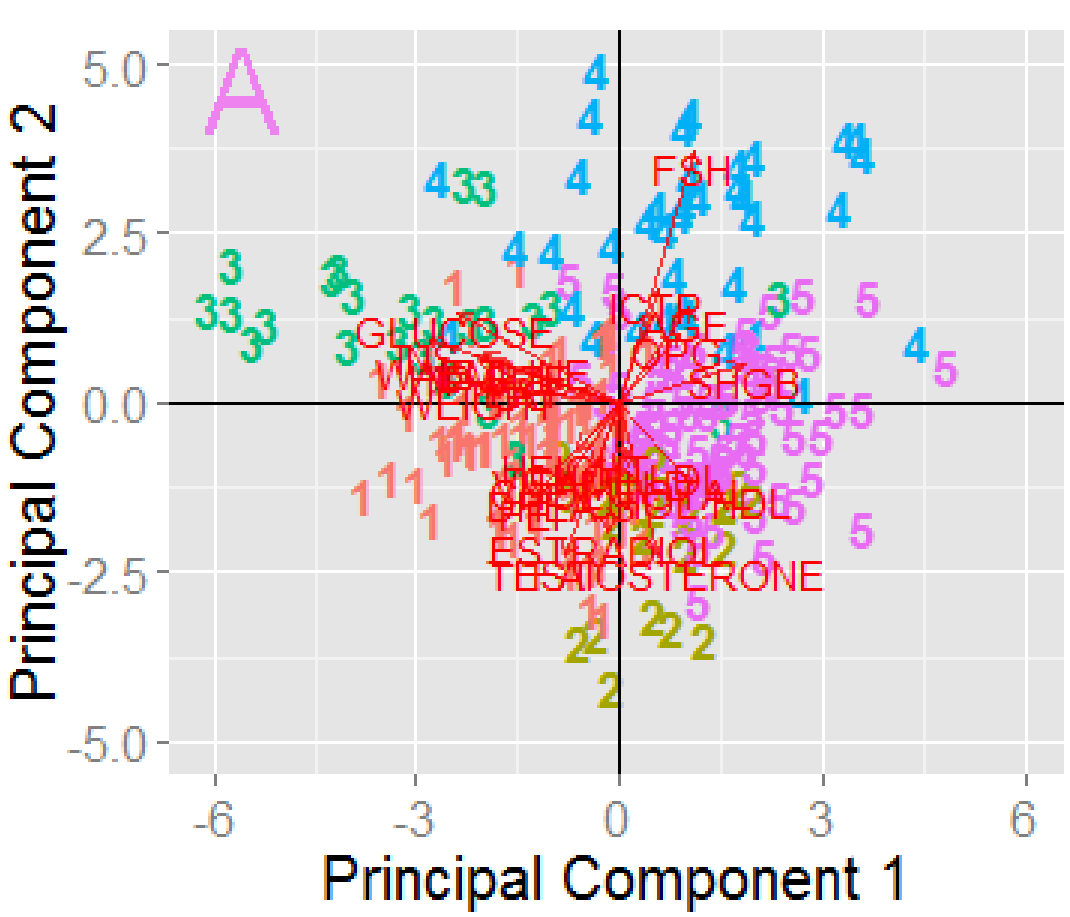


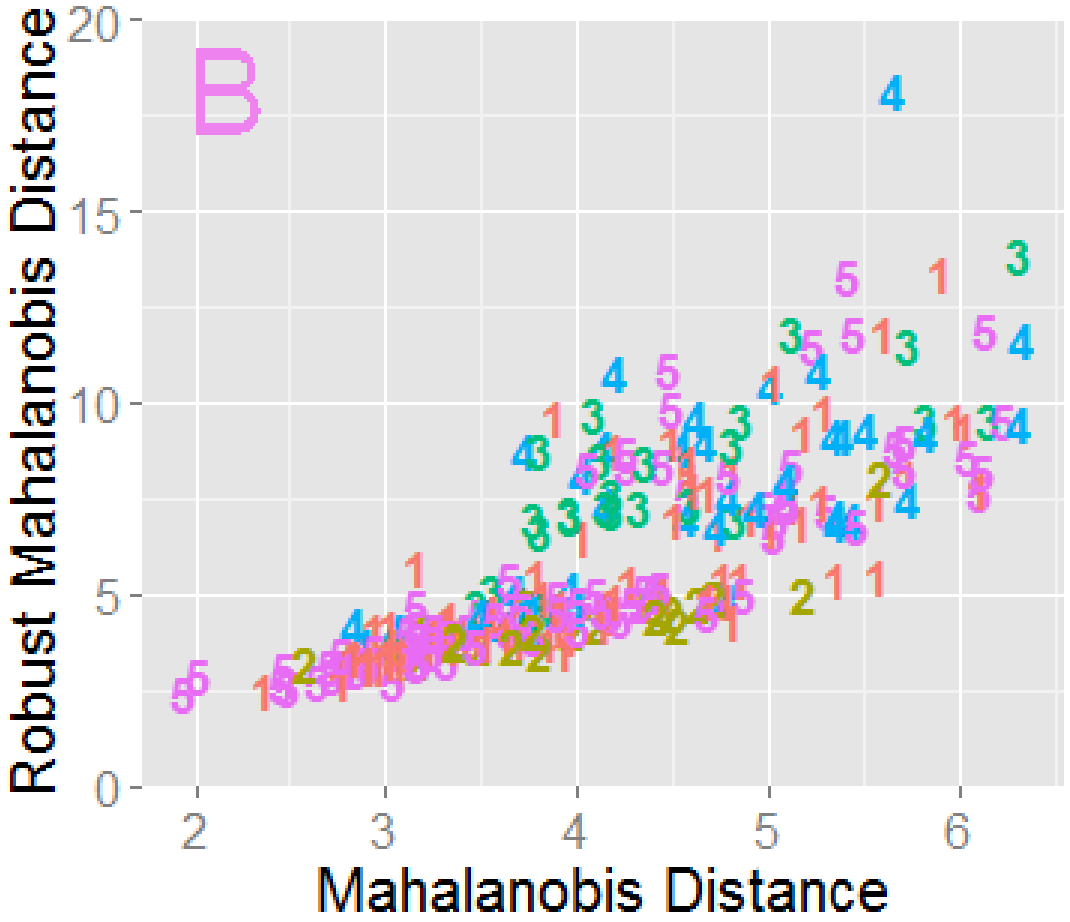


Addition of categorical data



Male set analysis – summarizing data





Methodological conclusions

- MD vs rMD plot
  - Emphasis of the data set heterogeneity
- Ward's hierarchical clustering + PCA
  - Consistent grouping & visualization of patient groups
- PCA with biplot vectors
  - Facilitated biological interpretation of structure of the data

Tutorial paper to be submitted to ***Statistics in Medicine (Wiley)***

Biological conclusions

- There are 5 distinct patient subgroups
  - **Among them patients with elevated FSH and low TESTOSTERONE (hypogonadism)**
- There are three groups of attributes:
  - Age-related attributes
  - Obesity-related attributes
  - Sex-hormones related and other

Department and Clinic of Endocrinology, Diabetology and Isotope Therapy, Wroclaw Medical University

Łukasz Łaczmański

Department of Health Promotion, University School of Physical Education, Wroclaw

Felicja Lwow

bogumil.konopka@pwr.edu.pl

Additional slides

Why use the Ward's algorithm

Outlier detection (Mahalanobis Distance MD)

Hierarchical clustering (Ward's algorithm)

Visualization (PCA)

$$MD(x_i) = \sqrt{(x_i - \bar{X})S^{-1}(x_i - \bar{X})},$$

where :

S $-$ *covariance matrix*

$$Var(I) = V(Q) + \sum_{q \in Q} \frac{m_q}{m_I} V(q)$$

Maximize inter-cluster variance

Minimize within cluster variance

$$V(Q) = \sum_{\bar{x}_q \in Q} \frac{m_q}{m_I} (\bar{x}_q - \bar{X})^2 \qquad V(q) = \frac{1}{N_q} \sum_{x_i \in q} (x_i - \bar{x}_q)^2$$

Q – partitioning

q - cluster

$$T = XW$$
$$X^T X = W \Lambda W^T$$

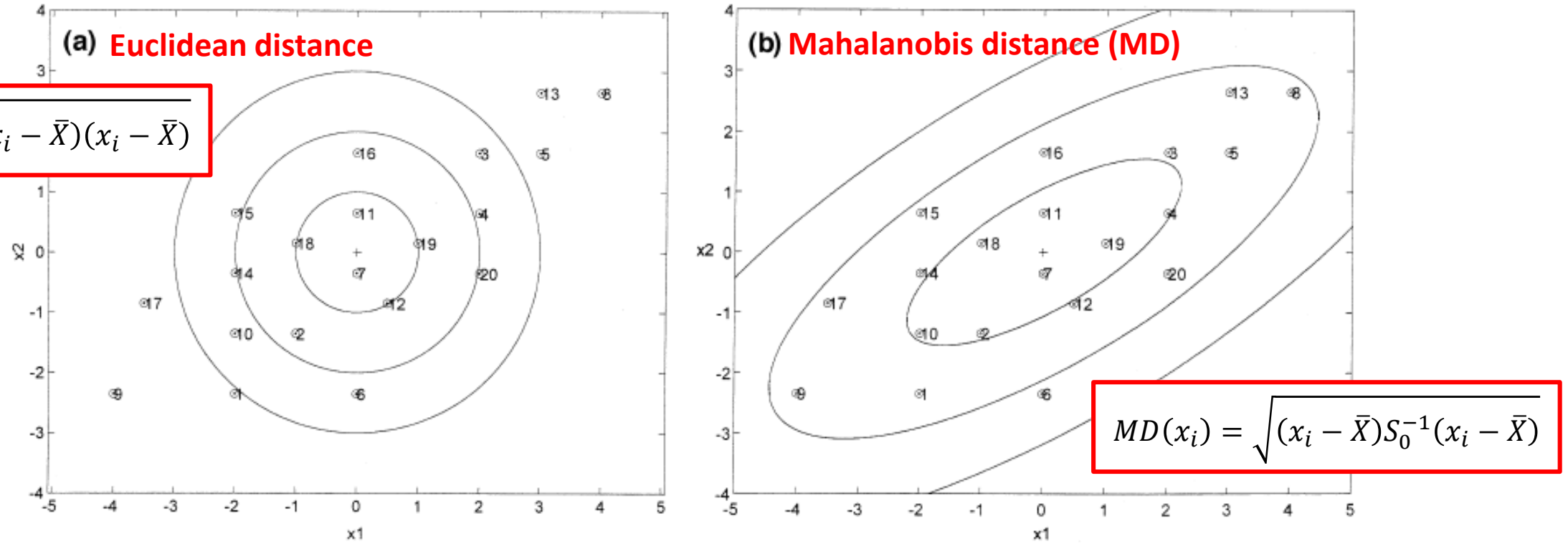
Λ – diagonal matrix of eigenvalues of $X^T X$ (S)

W – p-by-p matrix whose columns are eigenvectors of $X^T X$

Outlier detection with Mahalanobis Distance - intuition

De Maesschalck R, et al., Chemometrics and Intelligent Laboratory Systems 50 2000. 1–18

$$ED(x_i) = \sqrt{(x_i - \bar{X})(x_i - \bar{X})}$$



$$MD(x_i) = \sqrt{(x_i - \bar{X})S_0^{-1}(x_i - \bar{X})}$$

Fig. 1. (a) Plot of the simulated data for two variables x_1 and x_2 together with the circles representing equal EDs towards the center point. (b) Plot of the simulated data for two variables x_1 and x_2 together with the ellipses representing equal MDs towards the center point.

chemometrics: Multivariate Statistical Analysis in Chemometrics
http://cran.r-project.org/web/packages/chemometrics/index.html

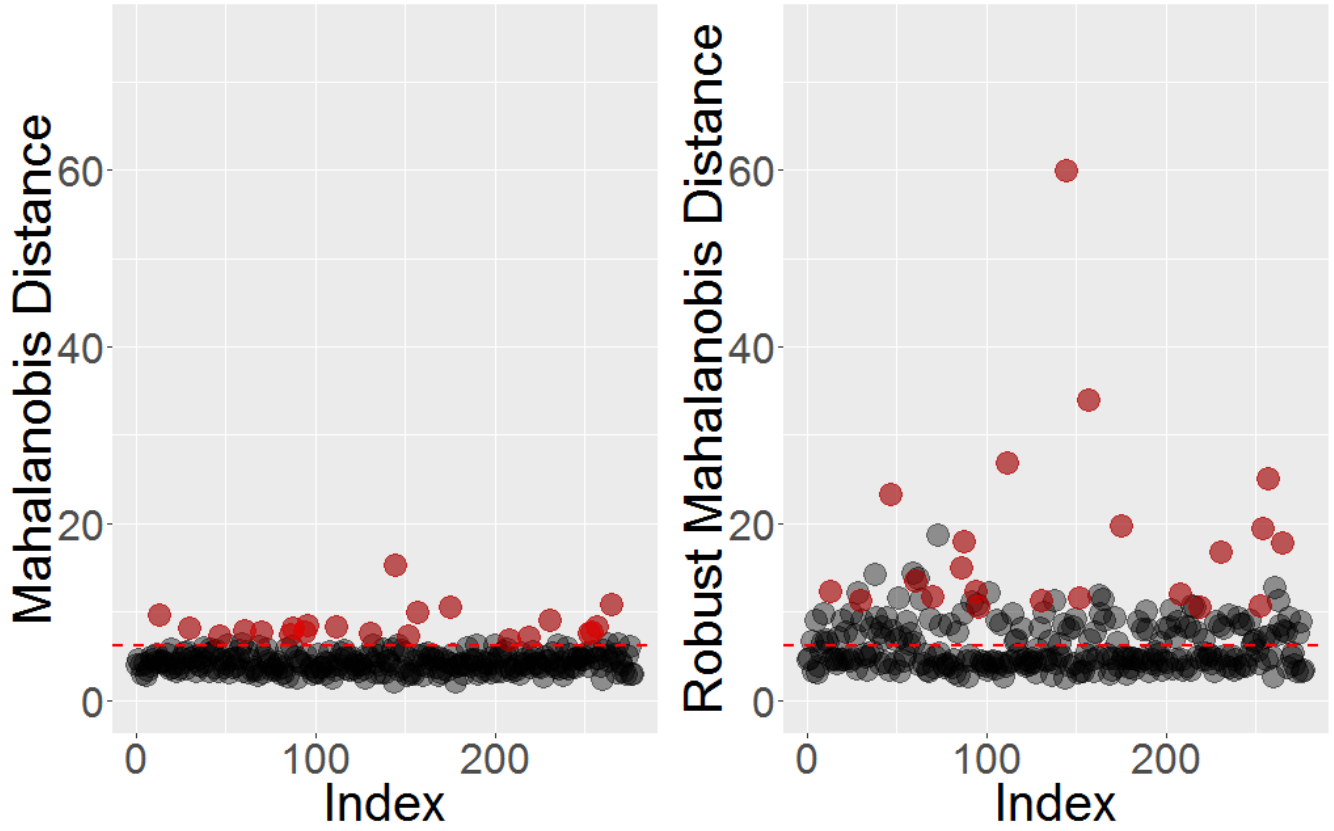
Male set analysis - outlier removal

- Outliers based on Mahalanobis Distance (MD):
  - **22 patients**

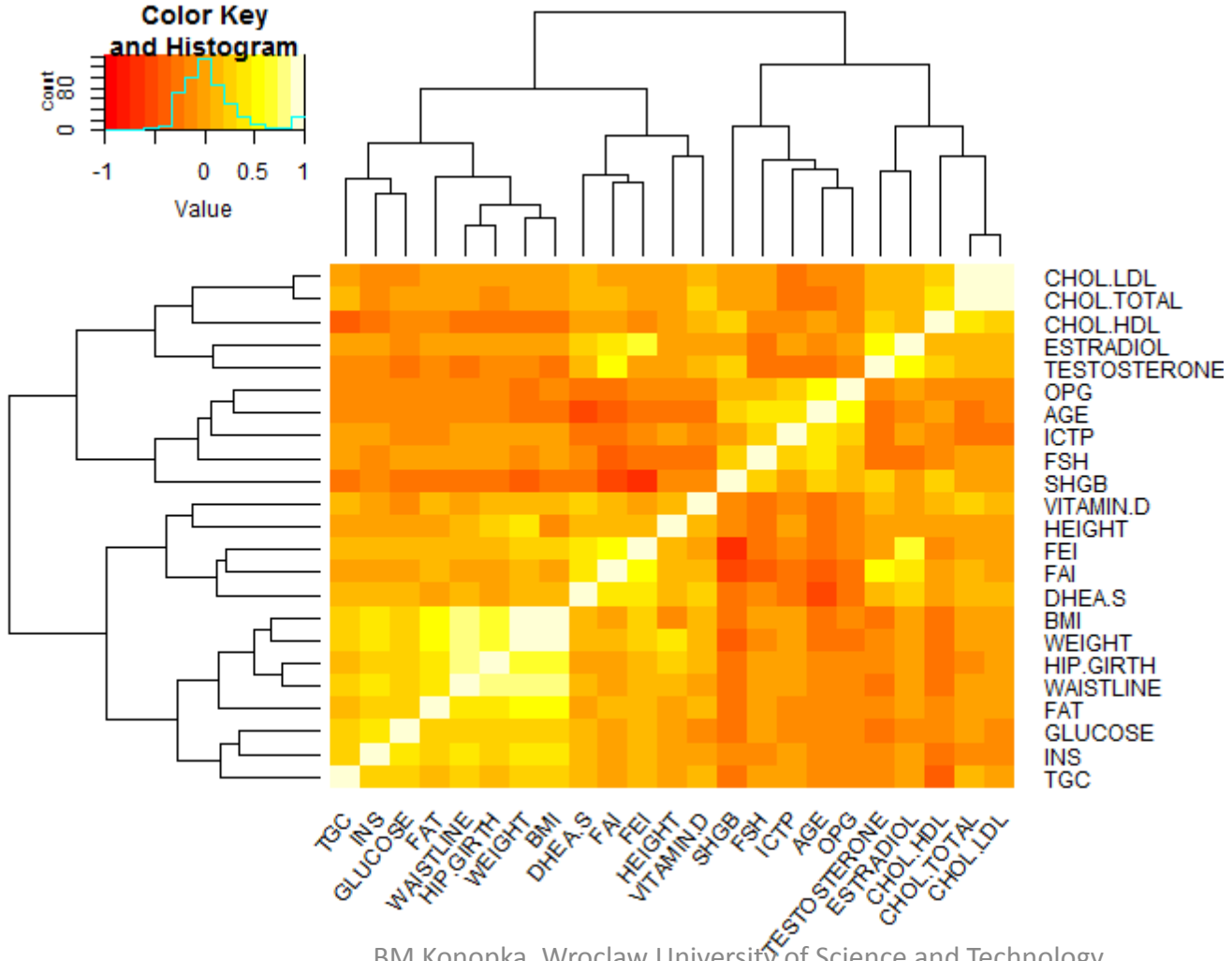
- Outlier detection with robust Mahalanobis Distance (rMD):

$$rMD(x_i) = \sqrt{(x_i - \overline{X_{\mathrm{k}}})S_{\mathrm{k}}^{-1}(x_i - \overline{X_{\mathrm{k}}})}$$

- Outliers based on robust MD:
  - **124 patients**

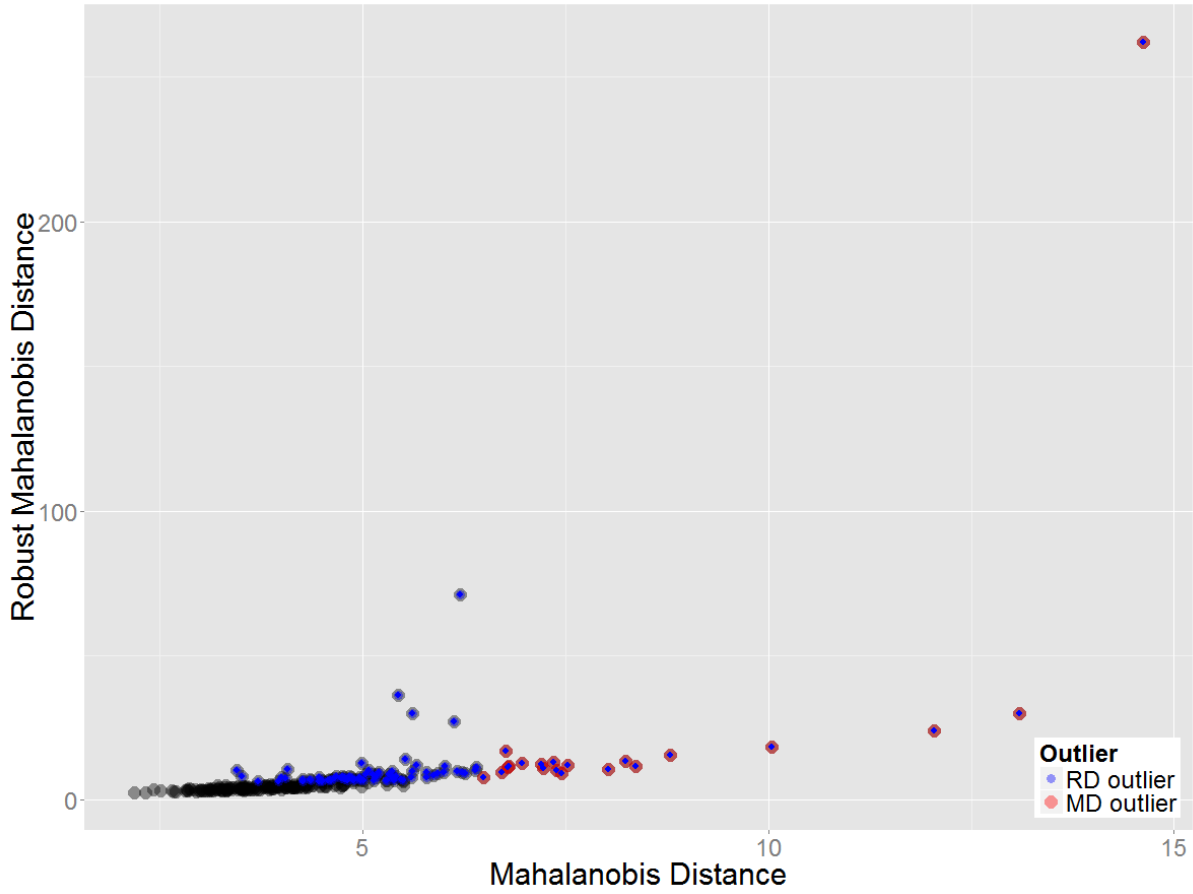


Clustering of attributes based on correlation

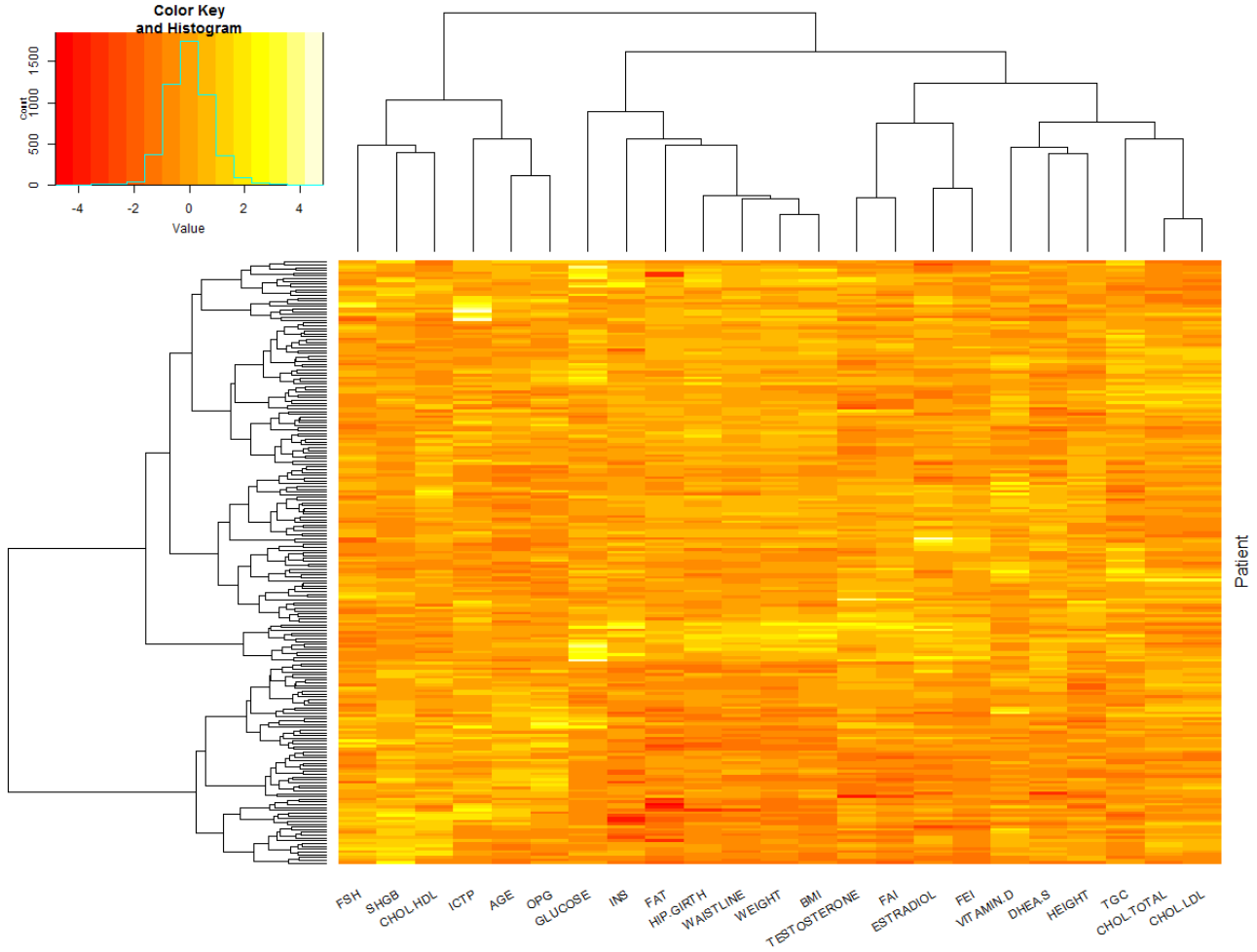


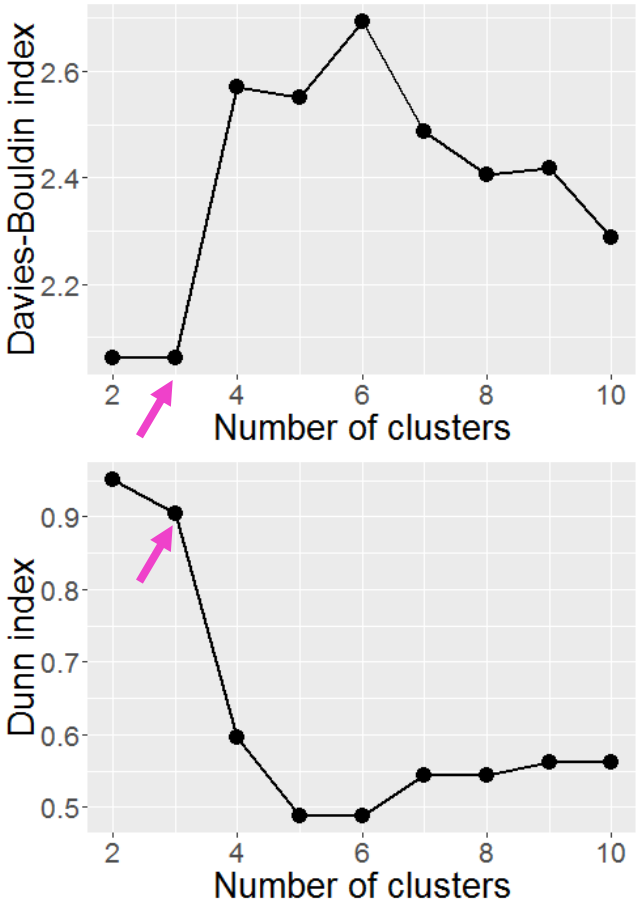
Female set analysis – outlier removal

- MD and rMD suggest removal of 20 and 70 data points
- Data points are more condensed



Female set analysis – hierarchical clustering





Female set analysis – hierarchical clustering

- Three subgroups have been identified
  - Cluster 3 – diabetes
- The set is more homogeneous
- Three groups of attributes have been identified:
  - Age-related attributes
  - Obesity-related attributes
  - Sex-hormones related and other