Statistical Data Analysis – a course map

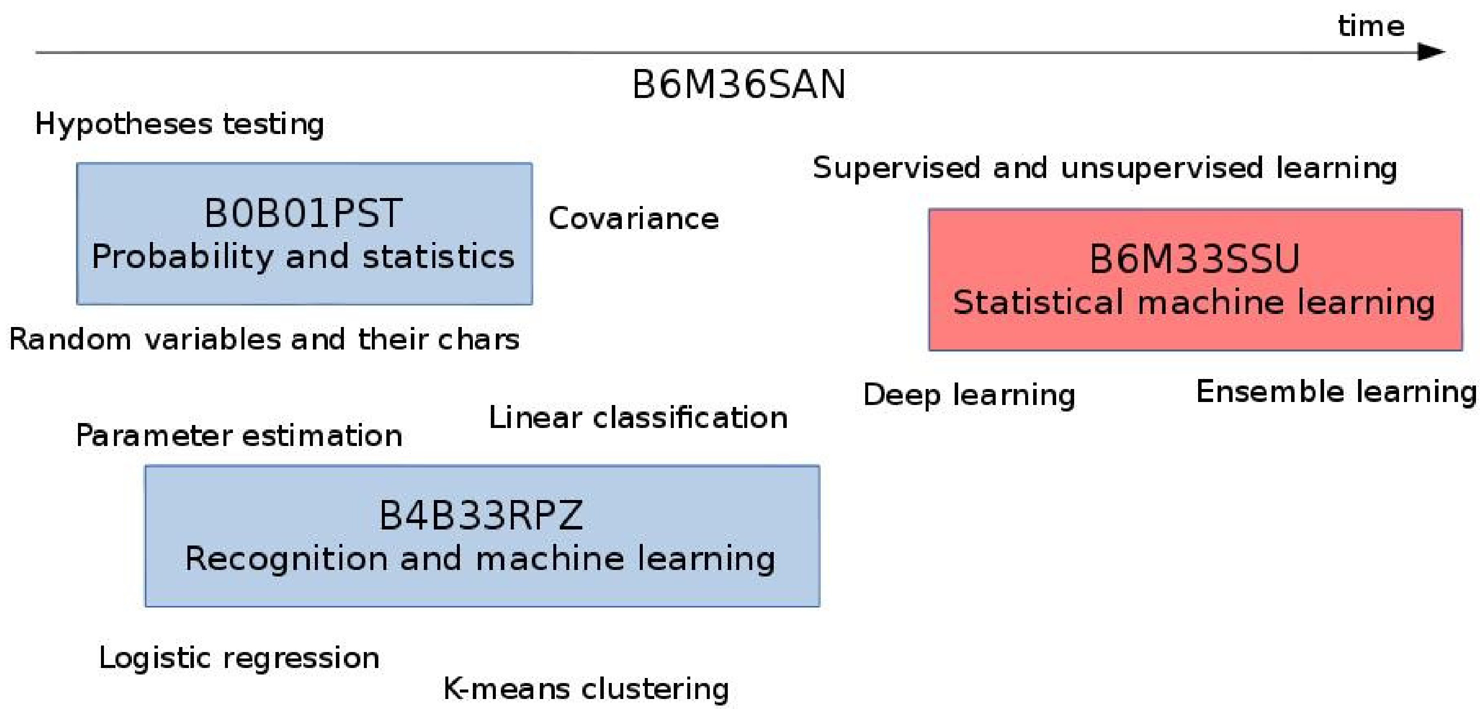
Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

http://cw.felk.cvut.cz/wiki/courses/b4m36san/start

B4M36SAN

- Purpose
  - This course mainly aims at the statistical methods that help to understand, interpret, visualize and model potentially high-dimensional data. It works with R environment.
- Interactions with other courses



Teachers

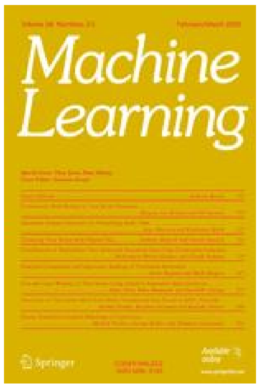
Doc. Jiří Kléma
CTU, Dept. of Computer Science
klema@fel.cvut.cz

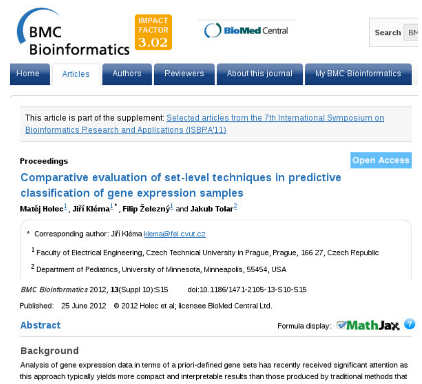
Dr. Tomáš Pevný
CTU, Dept. of Computer Science, CISCO Technical Leader
pevnytom@fel.cvut.cz

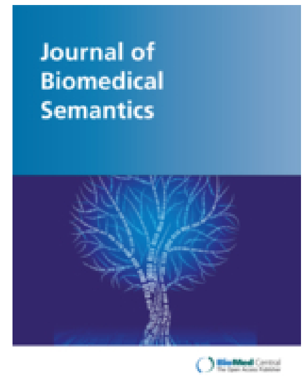
Doc. Zdeněk Míkovec
CTU, Dept. of Computer Graphics and Interactions
xmikovec@fel.cvut.cz

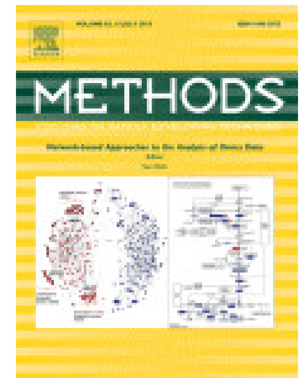
Ing. František Malinka
CTU, Dept. of Computer Science
malinfr1@fel.cvut.cz

IDA Highlights









publications

SUPREME+

SEVENPRO

XGENE.ORG
Cross-Genome Cross-Organism

organizing conferences

software projects

The key terms

- Multivariate statistical analysis
  - concerned with data that consists of **sets** of measurements on a number of individuals,
  - statistical approach based on **stochastic data models**
    * a certain model is assumed (a class of models),
    * its parameters are learned based on data,
  - more than independent testing of the individual variables (i.e., univariate tests known from introductory statistical courses),
  - intertwined variables, **examined simultaneously**,
  - not only the extensions of univariate and bivariate procedures,
  - examples: multivariate analysis of variance,
    multivariate discriminant analysis.

The key terms

- Applied statistics
  - in general, rather a branch of study than a course,
  - in here, the course could be understood as an opportunity to bring the (previously learned) methods to practice,
  - in labs, stress on applications and their implementation in R.
- Statistical inference/learning
  - close interaction with (statistical) machine learning,
  - sometimes it is difficult to distinguished these two fields
    - as their goals are interchangeable,
  - the most striking distinctions
    - different schools – statistics is a subfield of mathematics, machine learning is a subfiled of computer science,
    - different eras – for centuries versus modern,
    - different degree of assumptions – larger versus smaller.

B4M36SAN and quotes

:: Data do not give up their secrets easily. They must be tortured to confess.

Jeff Hopper, Bell Labs

:: All models are wrong, but some models are useful.

George Box, Princeton University

:: There are two kinds of statistics ...

... the kind you look up and the kind you make up.

Rex Stout, writer

Syllabus

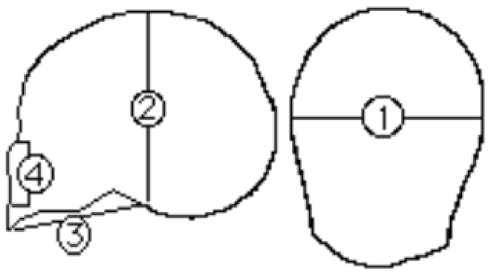
#	Lect	Content
1.	JK	Introduction, course map, review of the basic stat terms/methods.
2.	JK	Dimension reduction (PCA and kernel PCA).
3.	JK	Dimension reduction (other non-linear methods).
4.	JK	Clustering (basic methods, spectral clustering).
5.	JK	Clustering (biclustering, semi-supervised clustering).
6.	JK	Multivariate confirmation analysis (ANOVA and MANOVA).
7.	JK	Discriminant analysis (categorical, LDA, logistic regression).
8.	JK	Multivariate regression (continuous, linear regression, p-vals, overfitting).
9.	JK	Multivariate regression (non-linear, polynomial and local regression).
10.	TP	Anomaly detection.
11.	TP	Robust statistics.
12.	ZM	Empirical studies, their design and evaluation.
13.	ZM	Power analysis.

R package

- **R – the platform selected for labs**
  - the leading tool for statistics,
  - one of the main tools in data analysis and machine learning,
  - it is free, open-source and platform independent,
  - a large community of developers and users
    $\rightarrow$ a great variety of libraries, tutorials, mailing lists,
  - easy to integrate with other languages (C, Java, Python),
  - we actually use it,
  - bottlenecks in memory management, speed, and efficiency,
- alternatives
  - **Python** with its data analysis libraries (more general use),
  - **Matlab** (popular at FEL for its forte in control, Simulink etc.).

An example: male Egyptian skulls [Manly, 1991]

- 150 male Egyptian skulls from 5 different periods (30 skulls per group)
  - the early predynastic period (circa 4000 BC),
  - the late predynastic period (circa 3300 BC),
  - the 12th and 13th dynasties (circa 1850 BC),
  - the Ptolemiac period (circa 200 BC),
  - the Roman period (circa AD 150),
- 4 anthropometric measures collected for each skull
  - the maximum breadth (V1, mb),
  - the basibregmatic height (V2, bh),
  - the basialveolar length (V3, bl),
  - the nasal height (V4, nh).



An example: male Egyptian skulls [Manly, 1991]

- research questions
  - are there any differences in the skull sizes between the time periods?
  - do they show any (gradual) changes with time?
  - how are the four measurements related?
  - note: a change in skull size over time could be an evidence of the interbreeding of the Egyptians with immigrant populations over the years.
- statistical tasks (performed in R)
  - data visualization and understanding (exploratory analysis),
  - are there any differences in the skull sizes between the time periods?
  - do they show any changes with time?
  - discriminant analysis, can we tell the period from the skull measurements?

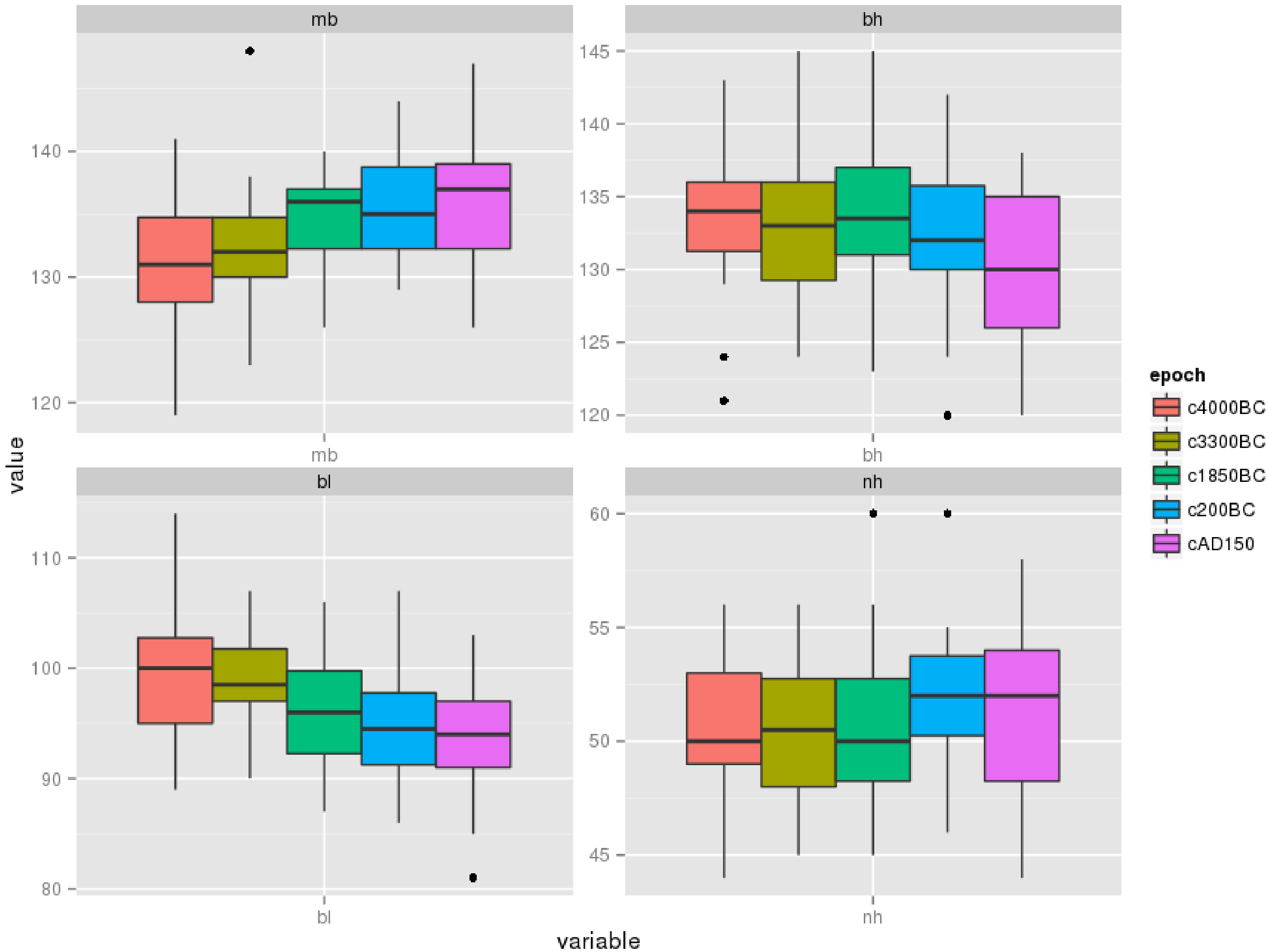
An example: male Egyptian skulls [Manly, 1991]

- upload into R, analysis there too,
- see the skript *skulls.R* at the course page,

```
> library(HSAUR) # load the library, must be installed before
> data("skulls", package = "HSAUR") # load into the workspace
> skulls # see the content of the data.frame

    epoch  mb  bh  bl nh
1 c4000BC 131 138  89 49
2 c4000BC 125 131  92 48
3 c4000BC 131 132  99 50
4 c4000BC 119 132  96 44

> help(skulls) # see the info about the dataset
> summary(skulls) # see the basic univariate stats
```

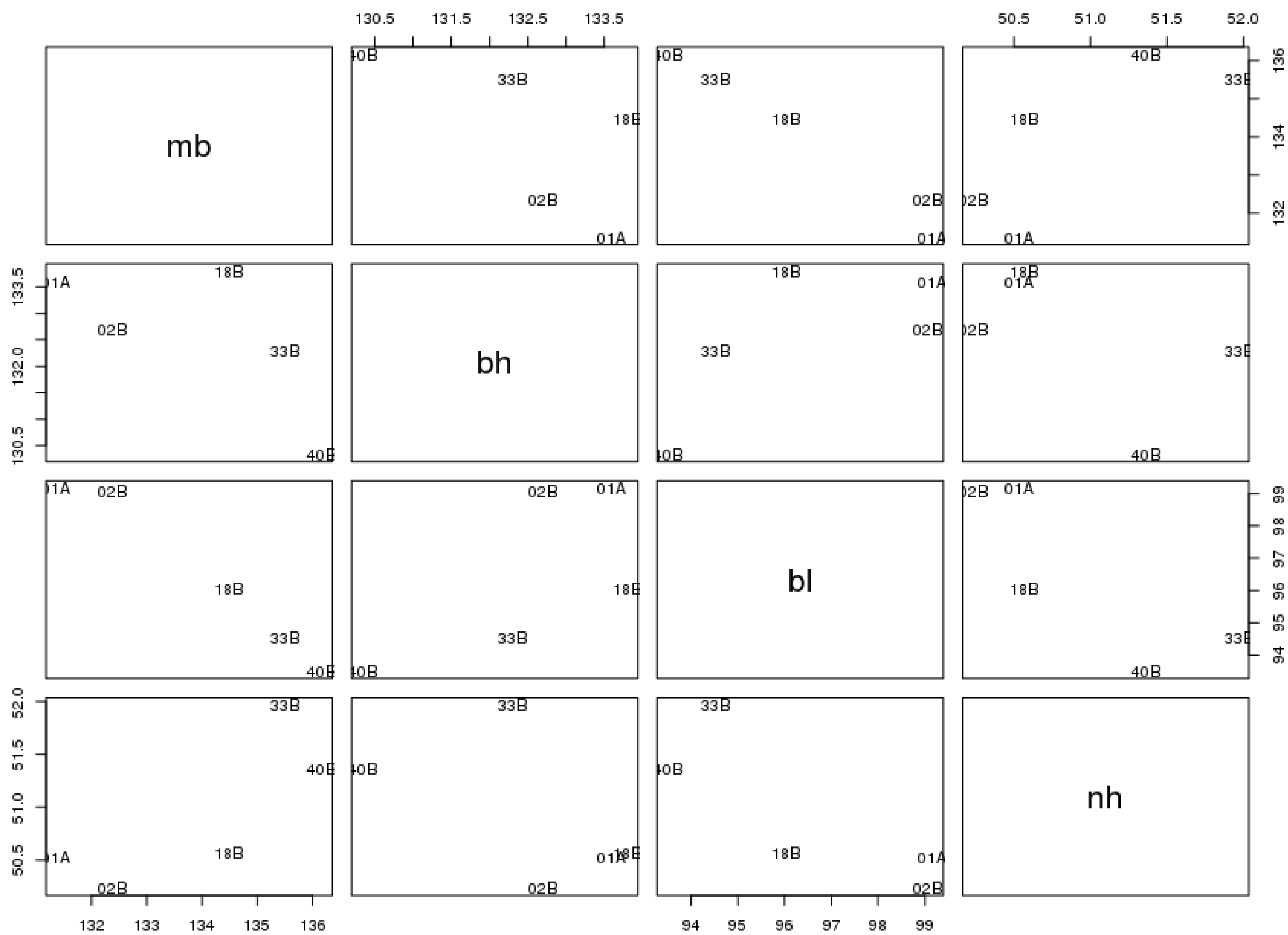
An example: male Egyptian skulls [Manly, 1991]

- use boxplots to visualize the data (univariate analysis)



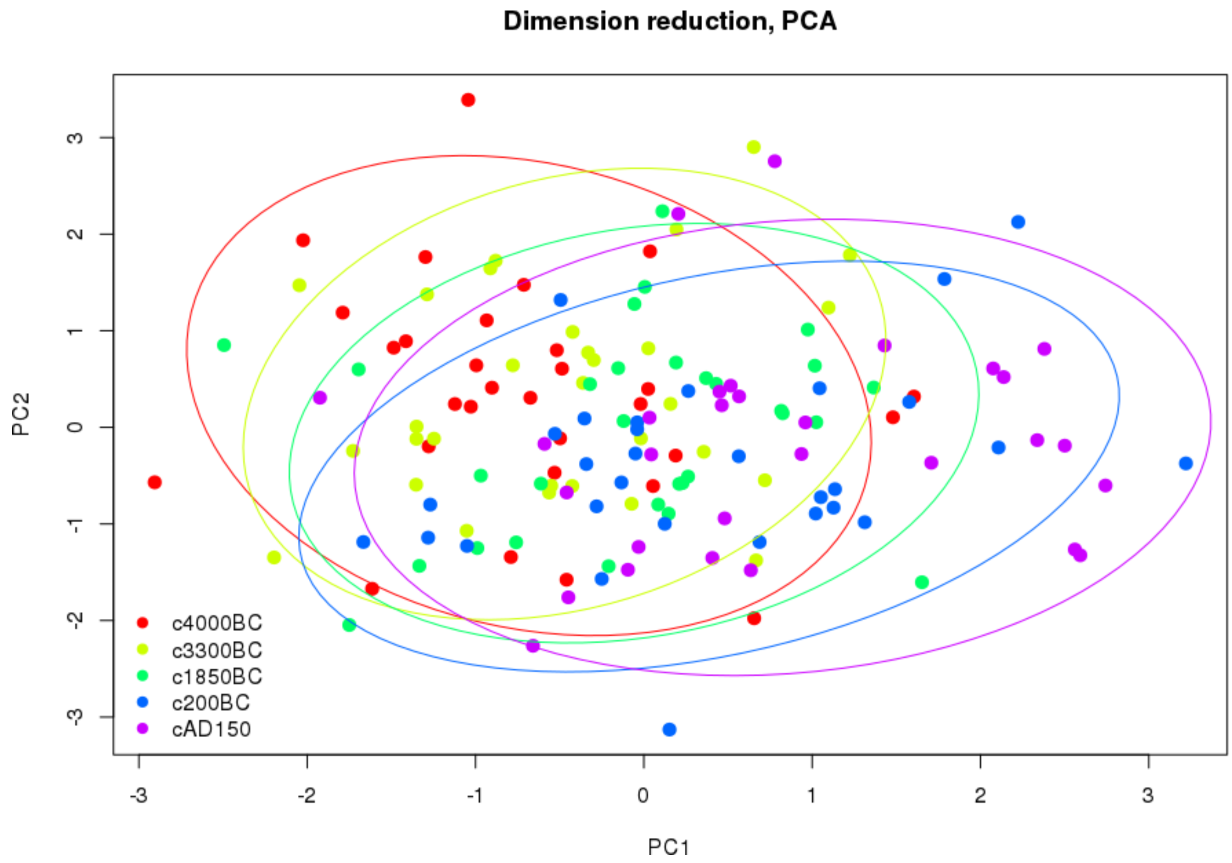
An example: male Egyptian skulls [Manly, 1991]

- use scatterplots to see interactions between variable group means



An example: male Egyptian skulls [Manly, 1991]

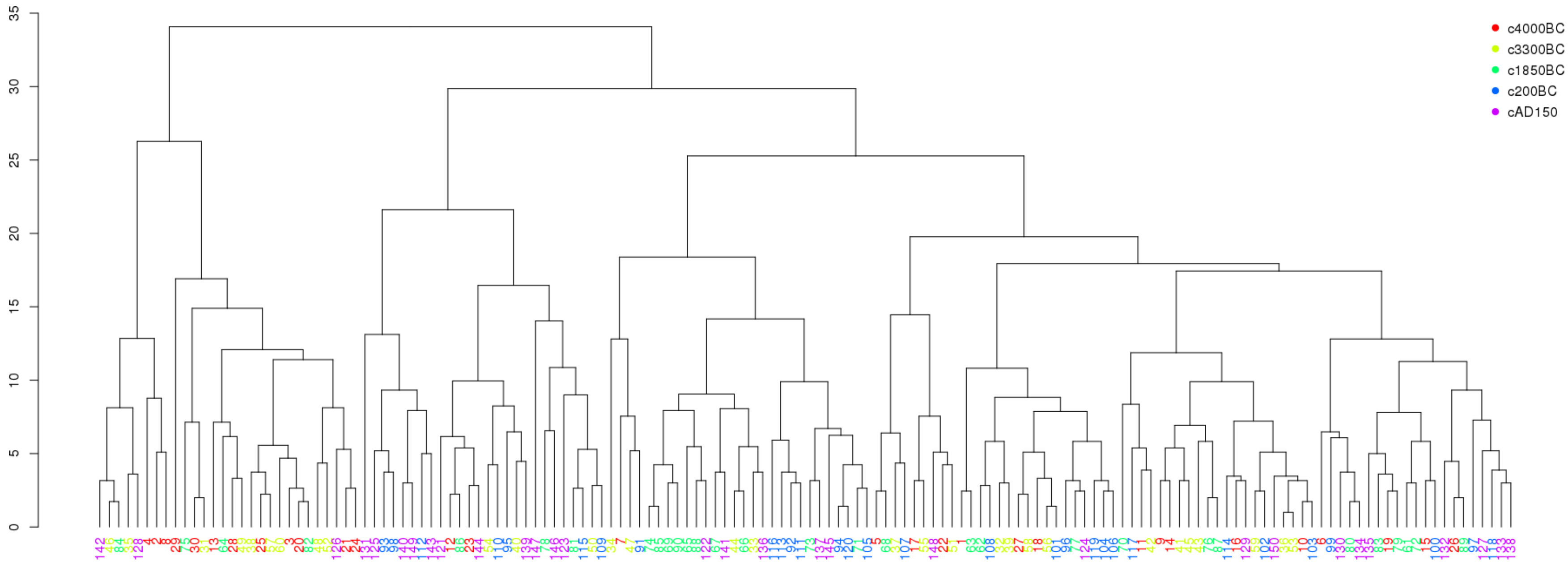
- employ **dimension reduction** to understand the epoch distribution



- only 64% of variance captured in this plot.

An example: male Egyptian skulls [Manly, 1991]

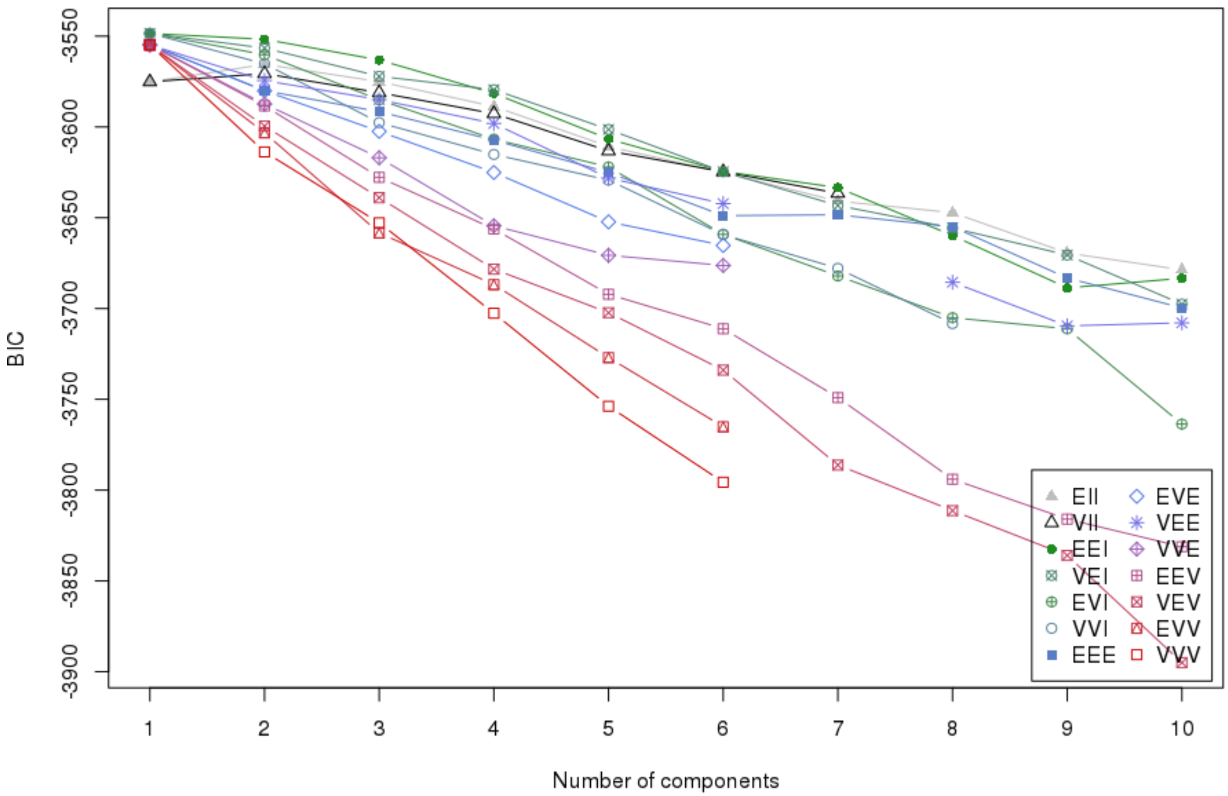
- **cluster analysis** may help to understand object distribution in skulls data,
- run hierarchical clustering and inspect the resulting **dendrogram**,



- conclusion: no evidence for the existence of 5 clusters, however epochs not randomly distributed along x axis.

An example: male Egyptian skulls [Manly, 1991]

- **cluster analysis** may propose the optimal number of clusters in skulls data,
- the expected/hoped-for value is 5, i.e., the number of epochs,



- run gaussian mixture model for different k values, compare in terms of **BIC**.

An example: male Egyptian skulls [Manly, 1991]

- multiple dependent variables $\rightarrow$ multivariate analysis,
- assume that time is a discrete variable, employ **MANOVA** (multivariate analysis of variance)
  - $H_0 : \boldsymbol{\mu}_{4000BC} = \boldsymbol{\mu}_{3300BC} = \cdots = \boldsymbol{\mu}_{AD150}$
  - $H_a : \mu_{ik} \neq \mu_{jk}$ for at least one $i \neq j$ and at least one variable k
  - assumptions: normal within-group distributions, and equal covariance matrices across groups,
- MANOVA outcome

```
           Df   Wilks approx F num Df den Df    Pr(>F)
epoch       4 0.66359   3.9009     16 434.45 7.01e-07 ***
Residuals 145
```

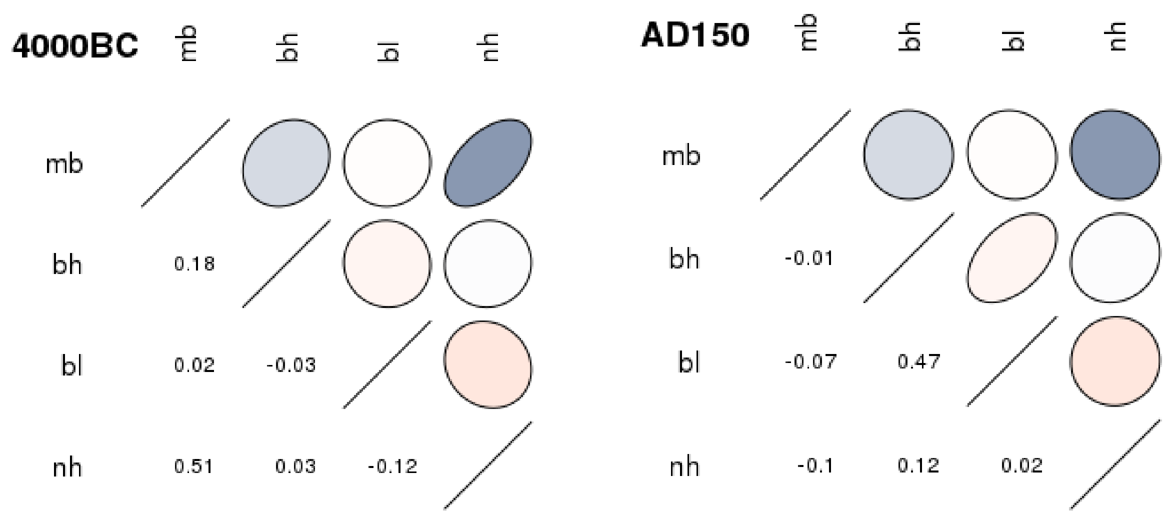
- conclusion: it is very likely that there are differences in the skull sizes between the time periods.

An example: male Egyptian skulls [Manly, 1991]

- test MANOVA assumptions
  - start with the **homogeneity of covariance matrices**, Box's M-test,
  - the clusters for the individual epochs must have a similar shape,

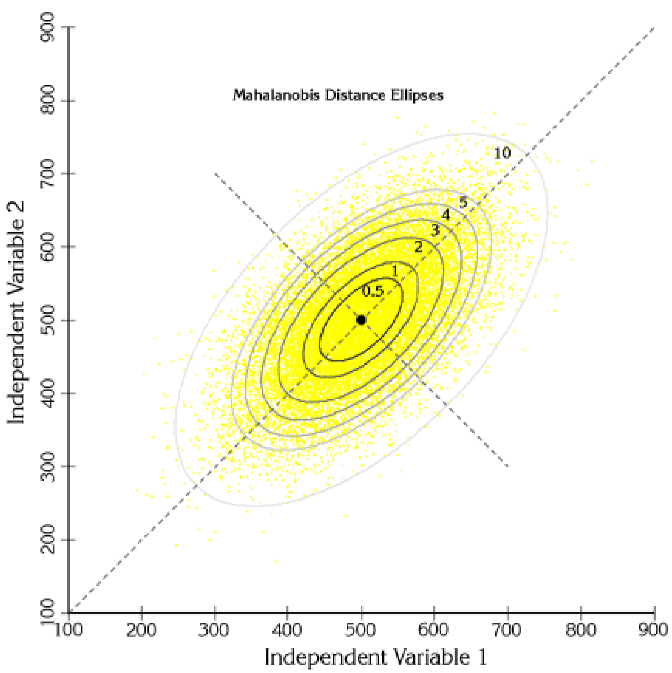
```
> boxM(skulls[c(2:5)], skulls$epoch)
Chi-Sq (approx.) = 45.667, df = 40, p-value = 0.2483
```

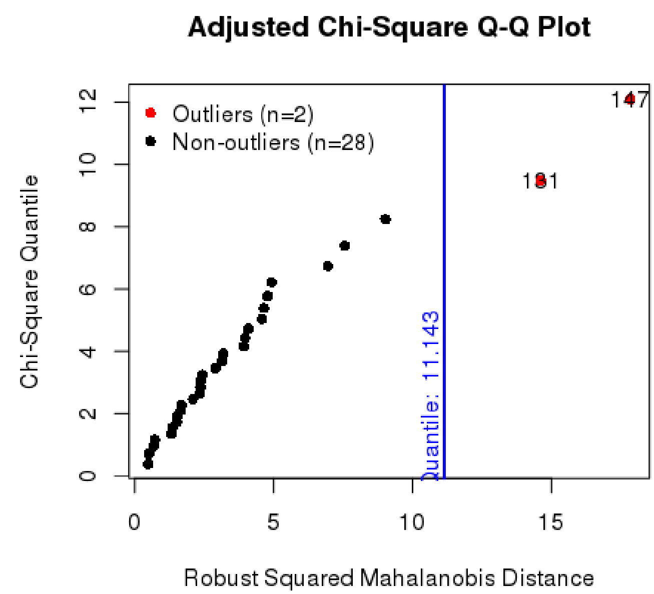
- conclusion: H_0 about homogeneous covariance matrices cannot be rejected.



An example: male Egyptian skulls [Manly, 1991]

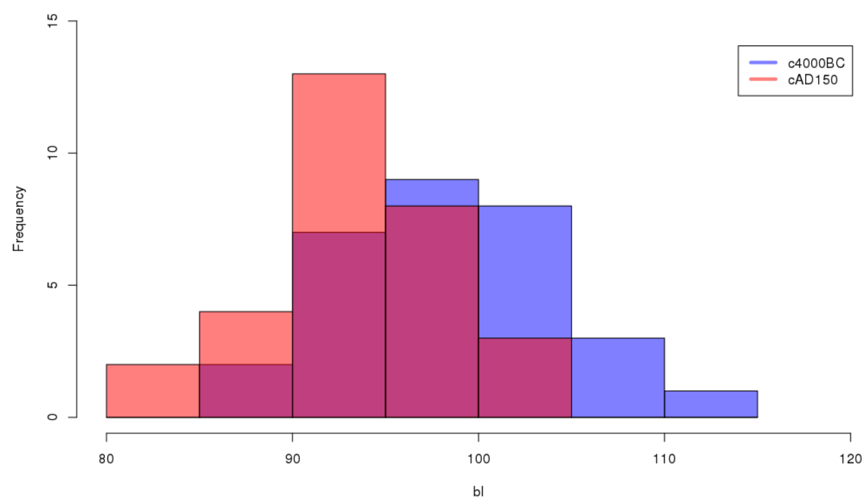
- test MANOVA assumptions
  - continue with the **normal within-group distributions**,
  - MANOVA not sensitive to shape deviations, rather to outliers,
  - see PCA plot to get an overall image,
  - perform an outlier test based on Mahalanobis distance, remove outliers.





An example: male Egyptian skulls [Manly, 1991]

- why MANOVA?
  - it has a higher **statistical power** than (a series of) simple tests,
  - avoids **multiple comparisons** and problems with their corrections,
- in our case, e.g., **t-test** provides a simple alternative
  - the null hypothesis is that the means of two populations are equal,
  - assumes normal distribution, in here Welch's test for unequal variances,
  - experiment: test two most distant epochs and the most promising variable,



An example: male Egyptian skulls [Manly, 1991]

- t-test call and outcome

```
> t.test(skulls$bl[skulls$epoch=="c4000BC"],
         skulls$bl[skulls$epoch=="cAD150"])

t = 4.0004, df = 56.716, p-value = 0.0001852
alt hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 2.83 8.50
sample estimates: mean of x 99.2 mean of y 93.5
```

- conclusion is same as in MANOVA
  - there are differences in mean skull sizes between epochs, namely in bl,
- however, p-value is higher and should further be corrected.

An example: male Egyptian skulls [Manly, 1991]

- assume that time is an **ordinal variable**, perform a trend test,
- in particular, **Jonckheere-Terpstra** test for ordered differences among classes
  - $H_0 : \theta_{4000BC,k} = \theta_{3300BC,k} = \cdots = \theta_{AD150,k}$
  - $H_a : \theta_{4000BC,k} \geq \theta_{3300BC,k} \geq \cdots \geq \theta_{AD150,k}$ (at least 1 strict inequality),
  - non-parametric rank-based test (θ stands for group medians),
  - not multivariate, perform independently for the individual variables,

```
> jonckheere.test(skulls$bl,skulls$epoch)

Jonckheere-Terpstra test

JT = 2989, p-value = 5.281e-07
alternative hypothesis: two.sided
```

- conclusion: bl manifests a monotonic trend.

An example: male Egyptian skulls [Manly, 1991]

- assume that time is a real variable, perform **multiple linear regression**,
  - $year = \beta_0 + \beta_1 mb + \beta_2 bh + \beta_3 bl + \beta_4 nh$
  - in this task, a non-realistic linearity and additivity assumption,
  - multivariate, partial effects of the individual variables on the model,

```
> skulls_lm <- lm(year ~ mb+bh+bl+nh,data=skulls)

Coeffs        Estimate Std. Error t value Pr(>|t|)
(Intercept) -3687.40    4946.86  -0.745 0.457235
mb              96.40      24.19   3.986 0.000106 ***
bh             -29.38      24.40  -1.204 0.230384
bl            -109.03      22.35  -4.877  2.8e-06 ***
nh              65.64      36.85   1.782 0.076918 .

Multiple R-squared:  0.2957, Adjusted R-squared:  0.2763
F-statistic: 15.22 on 4 and 145 DF, p-value: 2.06e-10
```

An example: male Egyptian skulls [Manly, 1991]

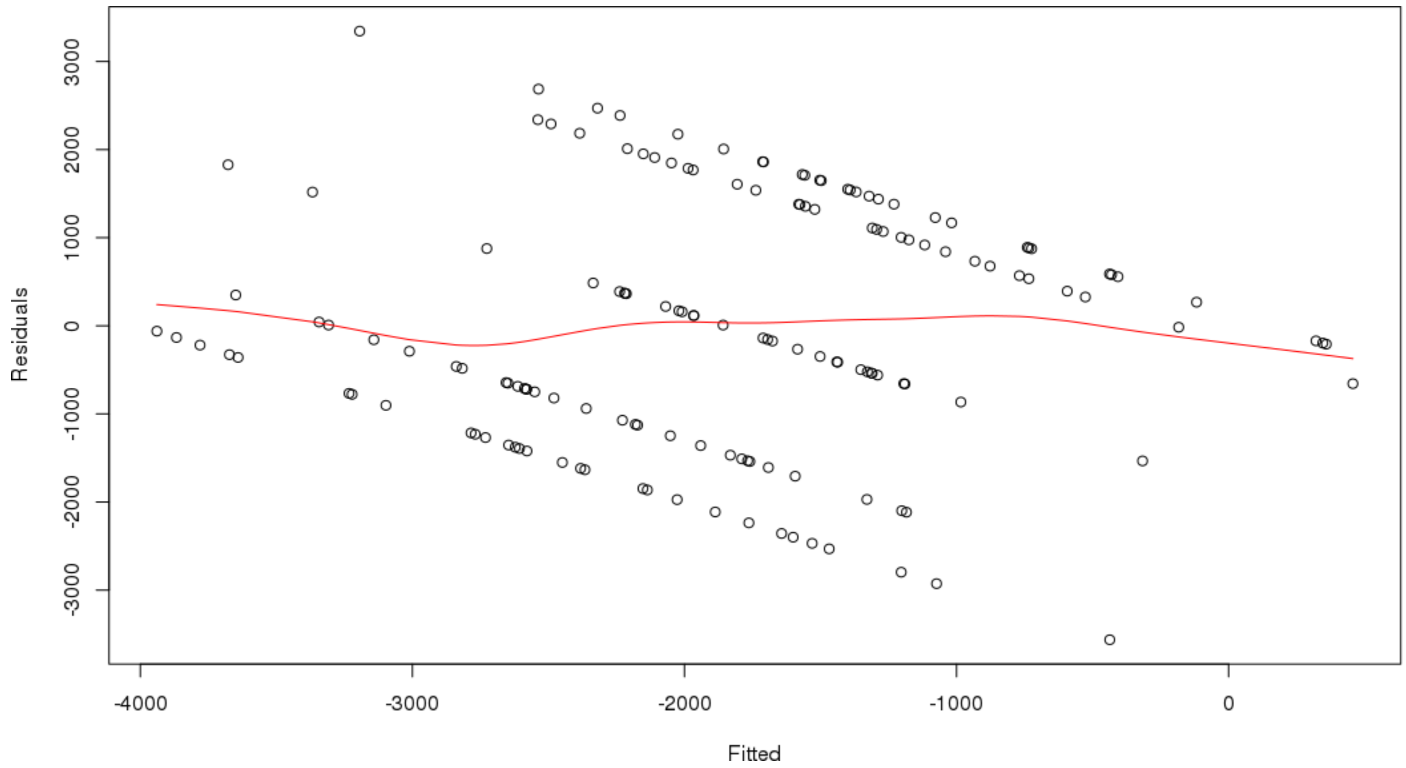
- perform **multiple linear regression** with two independent variables only
  - $year = \beta_0 + \beta_1 mb + \beta_2 bl$
  - take only those with partial effect on the model,

```
> skulls_lm_small <- lm(year ~ mb+bl,data=skulls)

Coeffs       Estimate Std. Error t value Pr(>|t|)
(Intercept) -4802.12    4096.13  -1.172    0.243
bl           -115.18      21.74  -5.298 4.20e-07 ***
mb            105.04      23.91   4.394 2.12e-05 ***

Multiple R-squared: 0.2761, Adjusted R-squared: 0.2662
F-statistic: 28.03 on 2 and 147 DF,  p-value: 4.873e-11
```

- about the same performance in terms of Adjusted R-squared as before.

An example: male Egyptian skulls [Manly, 1991]

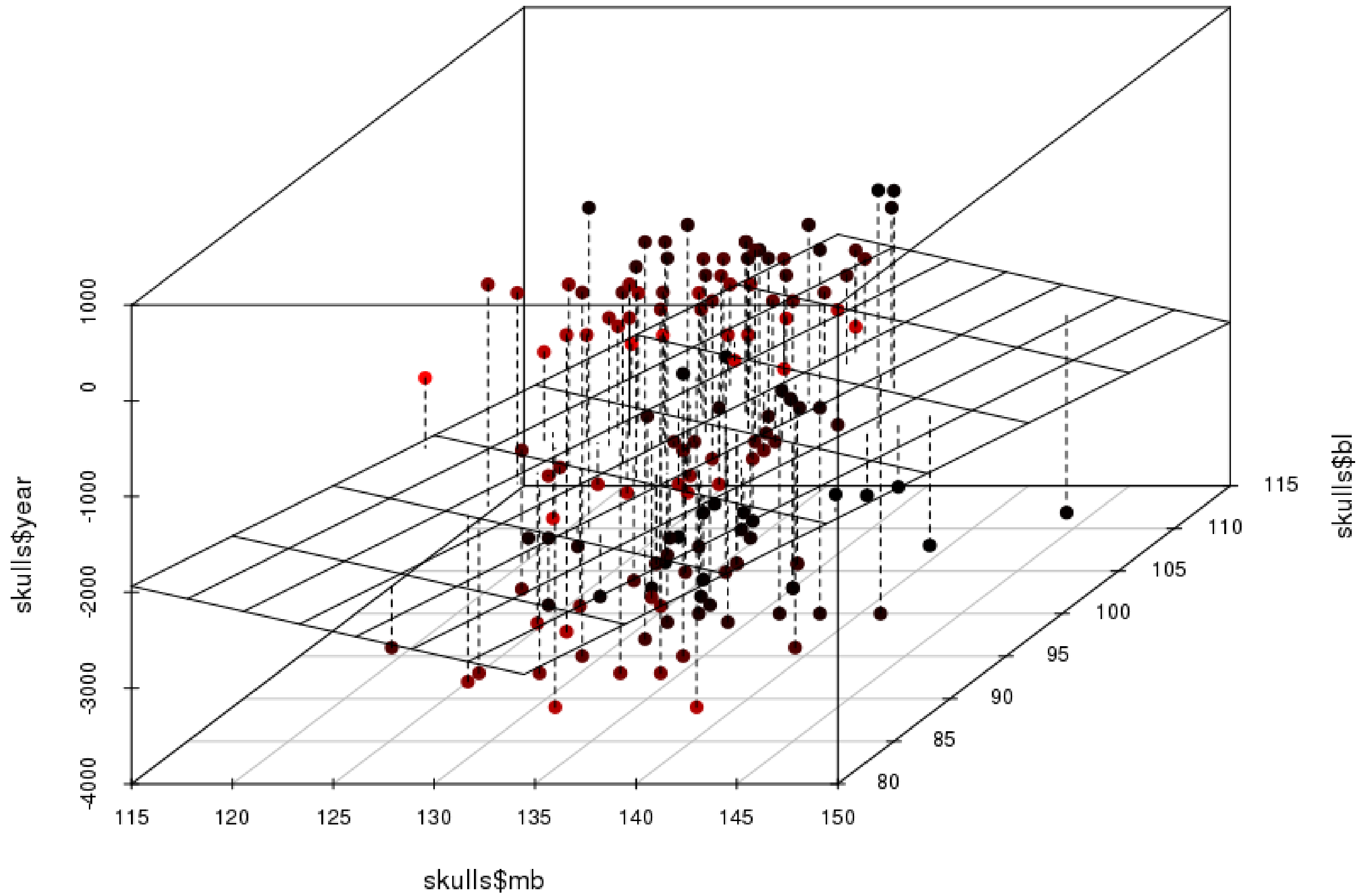
- multiple linear regression, assumption violations through residual analysis



- the model is unbiased, i.e., the error is not dependent on skull measures,
- the relationship is probably not linear as the error is not normally distributed.

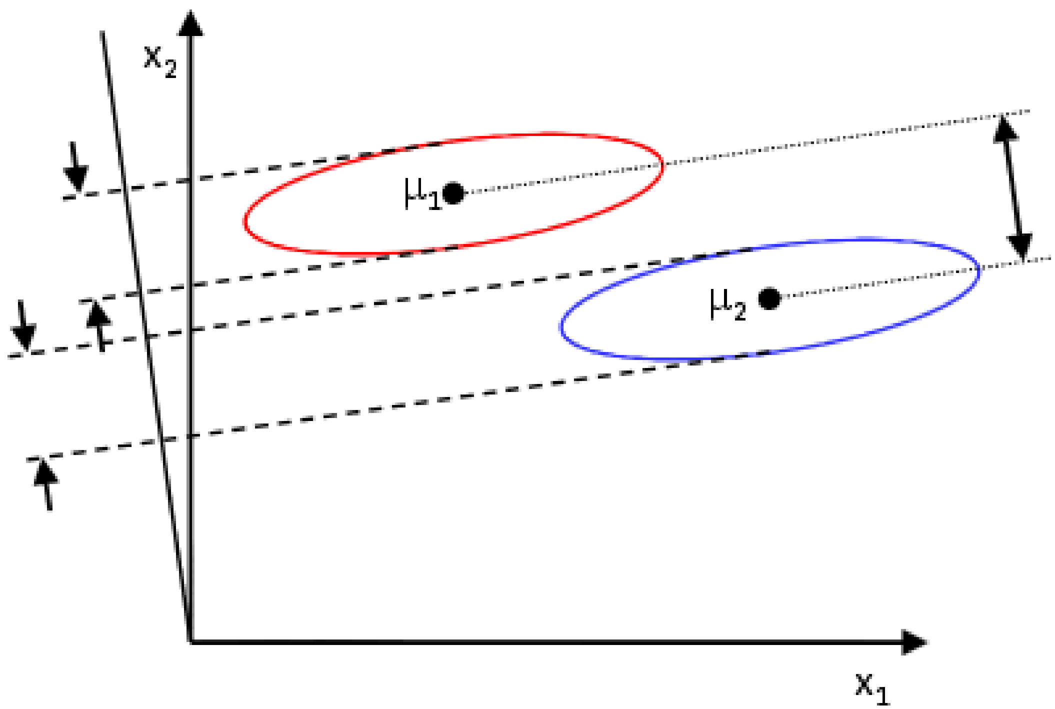
An example: male Egyptian skulls [Manly, 1991]

- multiple linear regression plane,
- for small models brings similar information as the previous residual plot.



An example: male Egyptian skulls [Manly, 1991]

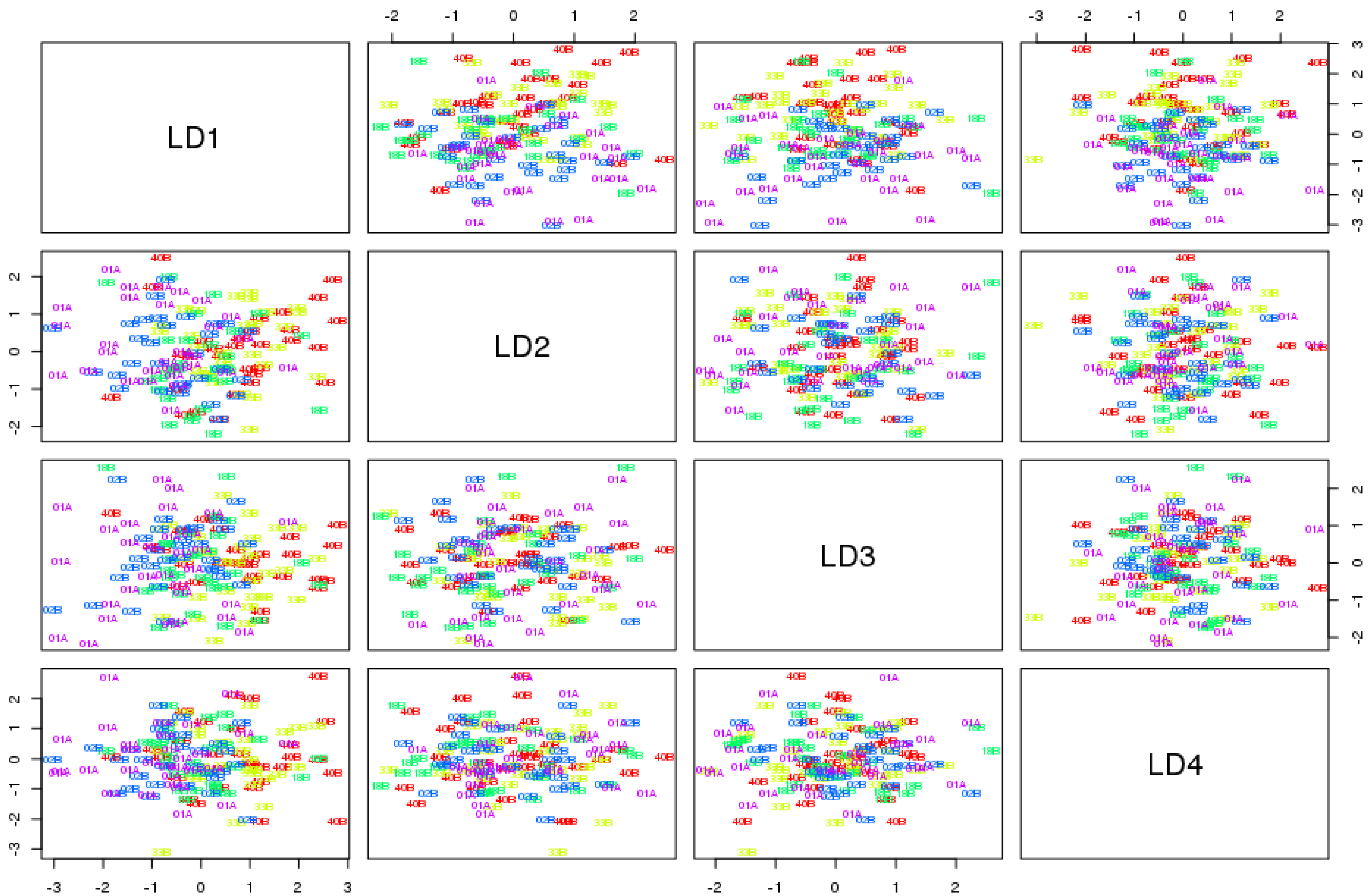
- can we tell the period from the skull measurements?
- perform **linear discriminant analysis**,
- find linear combinations of features that separate two or more classes of objects
  - maximize the difference between class means and minimize class variances,



Gutierrez-Osuna: LDA, Pattern Analysis course.

An example: male Egyptian skulls [Manly, 1991]

- linear discriminant analysis, projections into new bases.



An example: male Egyptian skulls [Manly, 1991]

- linear discriminant analysis, classification accuracy

```
> dis1 <- lda(epochs ~ mb+bh+bl+nh, skulls)
> plot(dis1, col = rainbow(5)[skulls$epoch])
> table(skulls$epochs,predict(dis1, skulls)$epochs)

     01A 02B 18B 33B 40B
01A   11   9   4   4   2
02B   12   5   7   3   3
18B    5   2  15   4   4
33B    3   4   5   8  10
40B    2   4   4   8  12
```

- around 34% accuracy on train data,
- namely intermediate epochs difficult to discriminate.

An example: male Egyptian skulls [Manly, 1991]

- Conclusions, answers to previous questions
  - the skull measures developed during time,
  - maximum breadth and basialveolar length manifest monotonic change with epochs,
  - the epoch cannot be reliably reconstructed solely from the skull measures,
  - there are no obvious additional dependencies between skull measures.

The key prerequisities – a brief review

- probability, independence, conditional probability, Bayes theorem,
- random variables, random vector,
- their description, distribution function, quantile function,
- categorical and continuous random variables,
- characteristics of random variables,
- the most common probability distributions,
- random vector characteristics, covariance, correlation, central limit theorem,
- measures of central tendency and dispersion, sample mean and variance,
- point and interval estimates of population mean and variance,
- maximum likelihood estimation, EM algorithm,
- statistical hypotheses testing,
- parametric and non-parametric tests,
- multiple comparisons problem, family wise error rate and false discovery rate.

The main references

:: Resources (slides, scripts, tasks) and reading

- G. James, D. Witten, T. Hastie and R. Tibshirani: **An Introduction to Statistical Learning with Applications in R.** Springer, 2014.
- A. C. Rencher, W. F. Christensen: **Methods of Multivariate Analysis.** 3rd Edition, Wiley, 2012.
- T. Hastie, R. Tibshirani and J. Friedman: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** Springer, 2009.