Zurich University
of Applied Sciences

School of
Engineering

IDP Institute of
Data Analysis and
Process Design

WBL Statistik 2024 — Robust Fitting

Half-Day 3: Multivariate Analysis Based on Robust Fitting

Andreas Ruckstuhl
Institut für Datenanalyse und Prozessdesign
Zürcher Hochschule für Angewandte Wissenschaften

Outline:

Half-Day 1
- Regression Model and the Outlier Problem
- Measuring Robustness
- Location M-Estimation
- Inference
- Regression M-Estimation
- Example from Molecular Spectroscopy

Half-Day 2
- General Regression M-Estimation
- Regression MM-Estimation
- Example from Finance
- Robust Inference
- Robust Estimation with GLM

Half-Day 3
- Robust Estimation of the Covariance Matrix
- Principal Component Analysis
- Linear Discriminant Analysis
- Baseline Removal: An application of robust fitting beyond theory

4.1 Robust Estimation of the Covariance Matrix

The **multivariate Gaussian distribution**

- plays a key role in **multivariate statistical analysis**
- is given by the mean (expectation) $\underline{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.

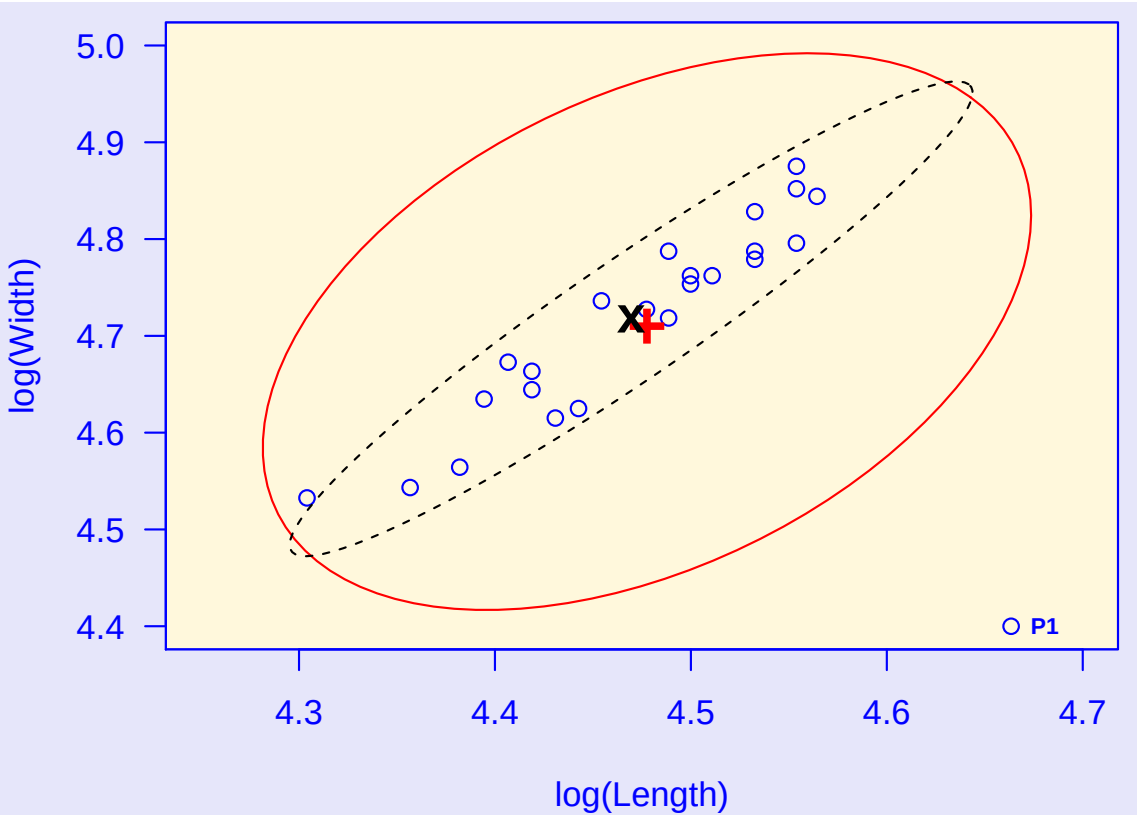
The optimal estimates for the parameters are

- the **(arithmetic) mean** $\underline{\bar{X}}$ and
- the **sample covariance matrix** $\widehat{\boldsymbol{\Sigma}}$

Example Painted Turtles:
Jolicoeur and Mosimann studies the relationship of size and shape for painted turtles. They measured the carapaces of 24 female and 24 male turtles.

The figure shows the estimated covariance matrix for the slightly modified data set: The covariance matrix is represented by the ellipsoid which contains 95% of the mass.

The **standard estimations** are based on the data including (solid, +) or excluding (dotted, x) observation P1.



Mahalanobis Distances to Detect Outliers

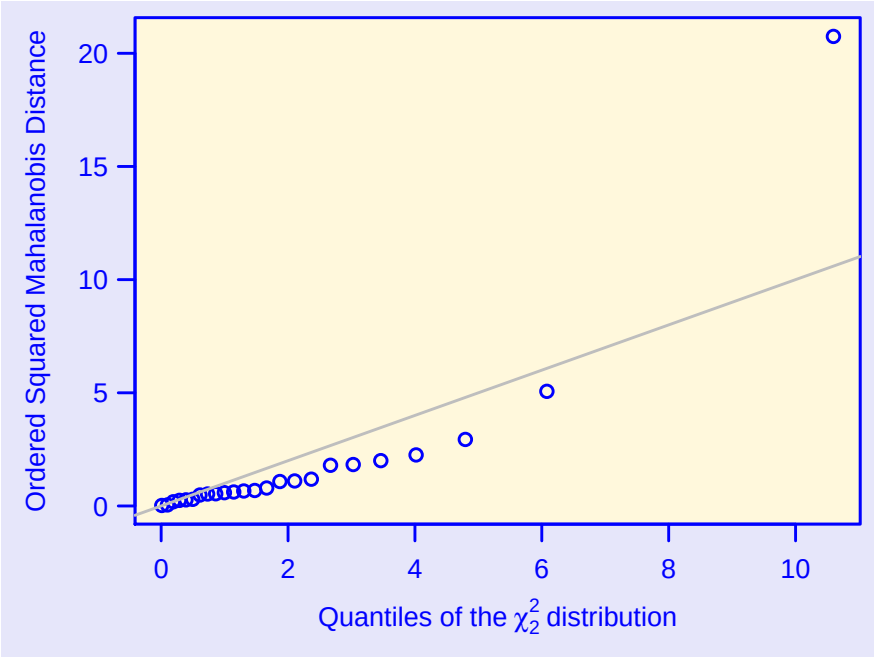
In a classical setting, squared Mahalanobis distances

$$u_i = (\underline{x}_i - \underline{\mu})^T \boldsymbol{\Sigma}^{-1} (\underline{x}_i - \underline{\mu})$$

are used to detect outliers:

Since U_i is χ^2_m distributed (m: number of variables)
use a QQ plot of the squared Mahalanobis distances versus χ^2_m distribution.

For modified Painted Turtles data: $m = 2$



Robust Estimation of the Covariance Matrix $\boldsymbol{\Sigma}$

Estimators of $\underline{\mu}$ and $\boldsymbol{\Sigma}$ based on a robust scale estimator:

Split $\boldsymbol{\Sigma}$ into a scale parameter σ and a shape matrix $\boldsymbol{\Sigma}^*$ with $|\boldsymbol{\Sigma}^*| = 1$:

$$\boldsymbol{\Sigma} = \sigma^2 \cdot \boldsymbol{\Sigma}^*$$

Calculate a scaled version of the squared Mahalanobis distance,

$$d\left\langle \underline{x}_i, \underline{\mu}, \boldsymbol{\Sigma}^* \right\rangle := (\underline{x}_i - \underline{\mu})^T \left(\boldsymbol{\Sigma}^*\right)^{-1} (\underline{x}_i - \underline{\mu}), \quad i = 1, \ldots, n,$$

and collect these elements in a vector $\underline{d}\left\langle \boldsymbol{X}, \underline{\mu}, \boldsymbol{\Sigma}^* \right\rangle$. Then $\mathrm{Var}\left\langle d\left\langle \underline{x}_i, \underline{\mu}, \boldsymbol{\Sigma}^* \right\rangle \right\rangle = \sigma^2 \cdot 2 \cdot m$

The estimates $\widehat{\underline{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}^*$ are defined by minimizing a scale estimator $S\langle\rangle$, i.e.,

$$S\left\langle \underline{d}\left\langle \boldsymbol{X}, \widehat{\underline{\mu}}, \widehat{\boldsymbol{\Sigma}}^* \right\rangle \right\rangle = \min .$$

To obtain **robust** estimation of $\underline{\mu}$ and $\boldsymbol{\Sigma}^*$, use a *robust* scale estimator $S\langle\rangle$

- The simplest approach is to take the median of d_i ($d_i > 0$)
  comparable to the MAV in regression

  This results in the **M**inimum-**V**olume-**E**llipsoid (**MVE**) **estimator**.
  It is the covariance matrix defined by the ellipsoid with minimum volume containing 50% of the data

  It has high breakdown point of 0.5 but is very inefficient.

- Use a trimmed scale estimator of the squared distances:
  $S\langle d_i \rangle = \sum_{i=1}^{h} d_{(i)}$ with $h = \frac{n+m}{2}$ ($m = \#$ variables, $n = \#$ observations).

  ☞ **M**inimum-**C**ovariance-**D**eterminant estimator (**MCD estimator**):
  Minimizes the determinant of the ellipsoid containing at least h data points.

  MCD estimator also has breakdown point of 0.5 but is more efficient than the MVE estimator.

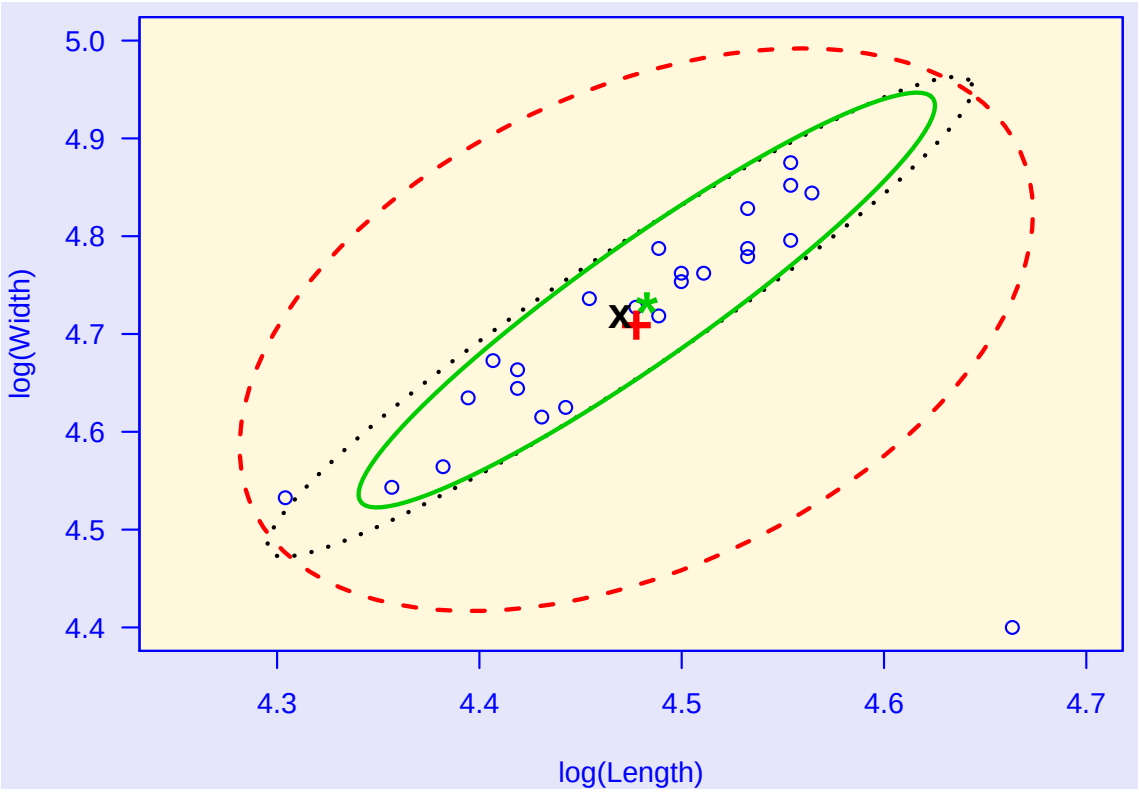
The computation of both estimators is, however, quite intensive as they are based on stochastic resampling algorithms.

Estimated covariance matrices for modified Painted Turtles data:

The estimated covariance matrices are represented by the ellipse containing 95% of the mass:

Classical estimates including (dashed, +) or excluding (dotted, x) observation P1.

The solid line (*) represents the **robust MCD estimation**.



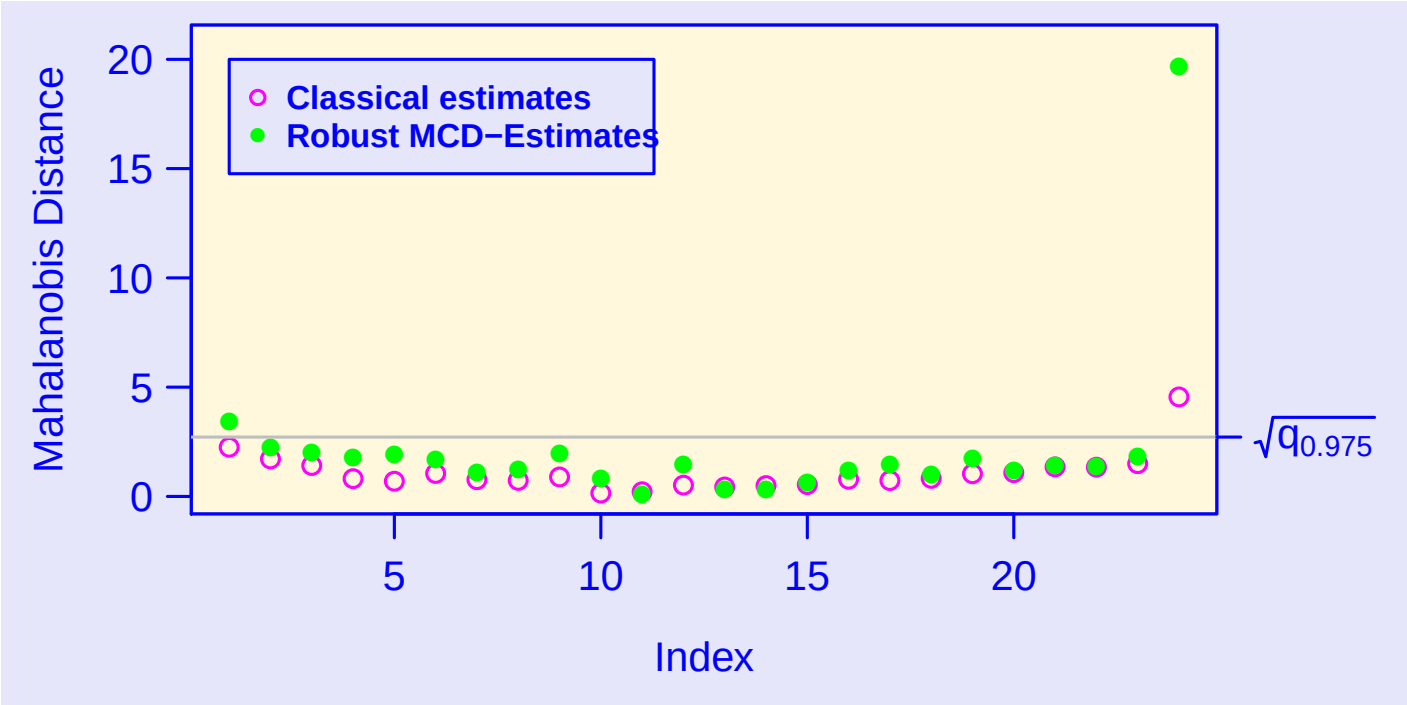
There seems to be a second outlier (see l.h.s.)

Mahalanobis Distances

To visualize the Mahalanobis distances (not the squared Mahalanobis distances), one can plot them versus the observation number.

Observations which are above the square-root of the 97.5%-χ^2_2 quantile line can be regarded as outliers.

Plot for the modified painted turtles data: There is a second outlier!



Other Approaches

There are other approaches like, e.g.,

- The **S-estimator** is also based on a robust scale estimator.
  The scale estimator $S\langle d_i \rangle$ satisfies

$$\frac{1}{n-m}\sum_{i=1}^{n}\rho\left\langle \frac{d_i}{S\langle d_i\rangle}\right\rangle = \frac{1}{2}$$

  where $\rho\langle u\rangle$ is the adequately adjusted bisquare function.

- the **Stahel-Donoho estimator**.
  Idea: A multivariate outlier should also be an outliers in *some* univariate projections
  - ☞ scan all univariate projections for outliers and weight them down.
  - ☞ apply a classical estimator using these weights
  - ☞ No exact algorithm is known; only for approximate solutions

- **Orthogonalized Gnanadesikan-Kettenring (OGK) Estimation**
  For really high dimensional data, the above approaches are far too slow.
  In such chase, an approach based on pairwise covariances may still help.
  Robust Estimates of pairwise covariances: $c^{(x,y)} = \frac{1}{4}\left(\left(S\langle x+y\rangle\right)^2 - \left(S\langle x-y\rangle\right)^2\right)$,
  where $S\langle . \rangle$ is a robust estimation of σ.
  A correction is needed to obtain a semi-definite matrix.

R functions

In practise, use

- `CovRobust(..., control="auto")` from R package `rrcov`

  Using `"auto"` selects an appropriate method according to the size of the dataset:

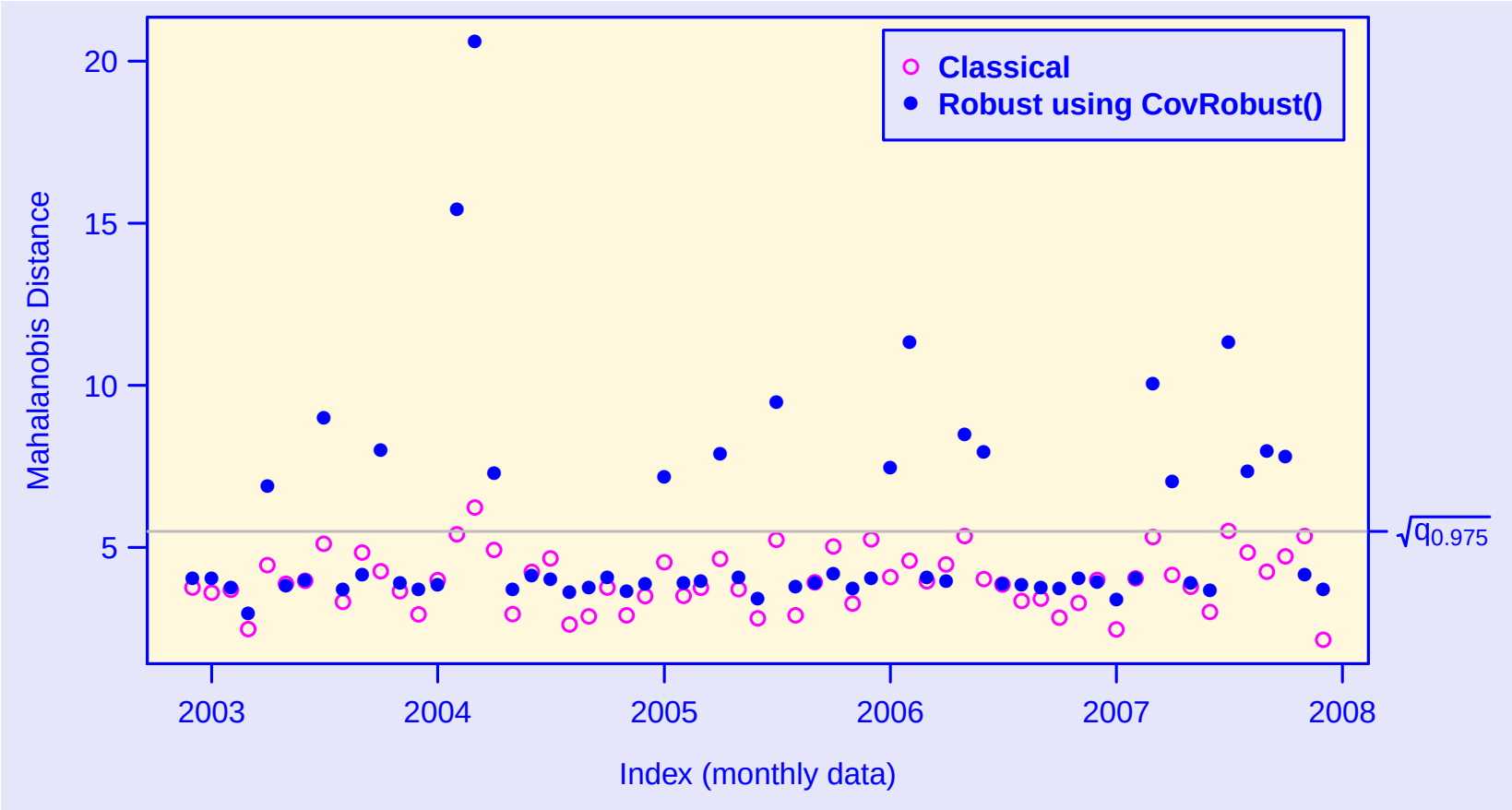
  - Stahel-Donoho estimator if dataset $< n = 1'000 \times p = 10$ or $< 5'000 \times 5$
  - S-estimator if dataset $< 50'000 \times 20$
  - Orthogonalized Quadrant Correlation (=OGK) if $n > 50'000$ and/or $p > 20$

Alternatives:

- `covMcd(...)` and `covOGK(...)` from R package `robustbase`
- `CovRobust(..., control="xxx")` with `xxx="mcd", "ogk", ...` (cf help page) from R package `rrcov`
- `cov.rob(..., method="mcd")` from R package `MASS`

Example Focused Directional FoHF

Monthly returns of 17 funds of hedge funds (FoHF), which according to a self-declaration run a "focused directional" strategy. The Mahalanobis distances of data covering 61 month are analysed.



4.2 Principal Component Analysis (PCA)

The goals of a principal component analysis (PCA) may be manifold;

for example

- reduction of dimensionality by elimination of directions (= linear combination of original variables) of low variability (= low information content).
- Finding structures like subgroups or outliers
- transformation of exploratory variables to avoid collinearity
  ☞ principal regression analysis.

Note

- The main principal components specify uncorrelated directions that account for most of the variability in the sample
- As a descriptive tool, there is no need for an underlying statistical model.

  However, since the analysis is based just on the first two moments, the **multivariate Gaussian model** is somehow nearby.

- To robustify a procedure we rely on a underlying statistical model.
- As there is no underlying model in PCA, it is unclear what **PCA should be robust against**.

But we can construct yet another explorative tool by computing the principal components from a **robustly estimated covariance** matrix.

When using robust methods, we explore a multivariate data set by investigating both

- the scatterplot of the main principal components
  (for finding interesting structures)
- and the QQ-plot of the squared Mahalanobis distances versus χ^2_m distribution for finding outliers.

To perform a PCA based on robust estimated covariance matrix we can use, e.g., the R function `PcaCov` from package `rrcov`

```
## Default S3 method:
> PcaCov(x, k=ncol(x), kmax=ncol(x), cov.control=CovControlMcd(),
        scale=FALSE, signflip=TRUE, crit.pca.distances=0.975,
        trace=FALSE, ...)
```

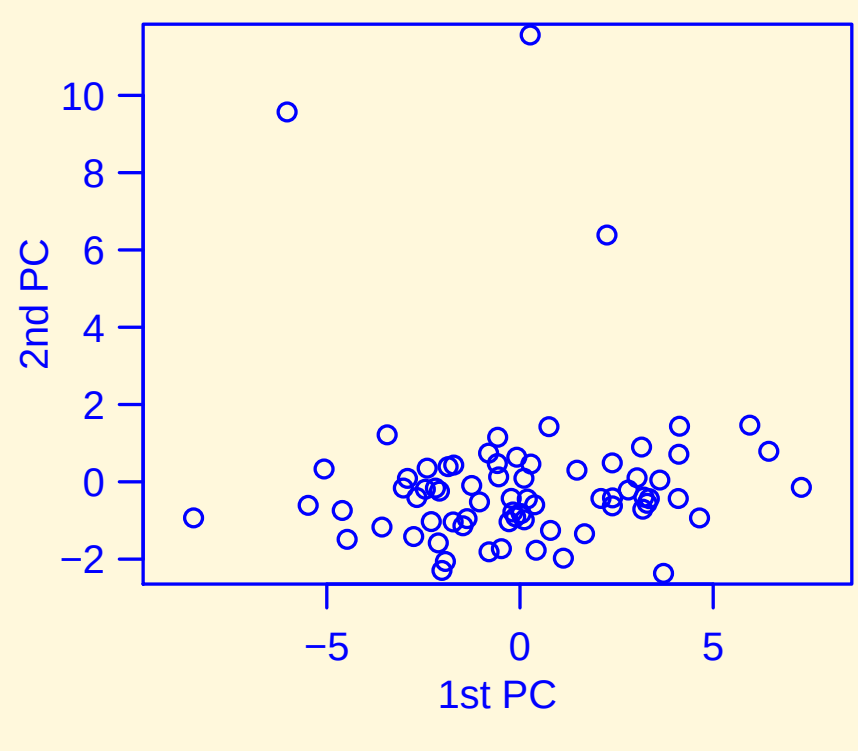
`formula x`	a numeric matrix (or data frame) which provides the data for the principal components analysis.
`k`	number of principal components to compute. If k is missing, or $k = 0$, the algorithm itself will determine the number of components (. . . see help)
`kmax`	maximal number of principal components to compute. Default is kmax=10. If k is provided, kmax does not need to be specified, unless k is larger than 10.
`cov.control`	specifies which covariance estimator to use by providing a CovControl-class object. . . . (see help)
`scale`	a value indicating whether and how the variables should be scaled to have unit variance. If scale=FALSE no scaling is performed . . . (see help)
. . .	

Simulated Example:

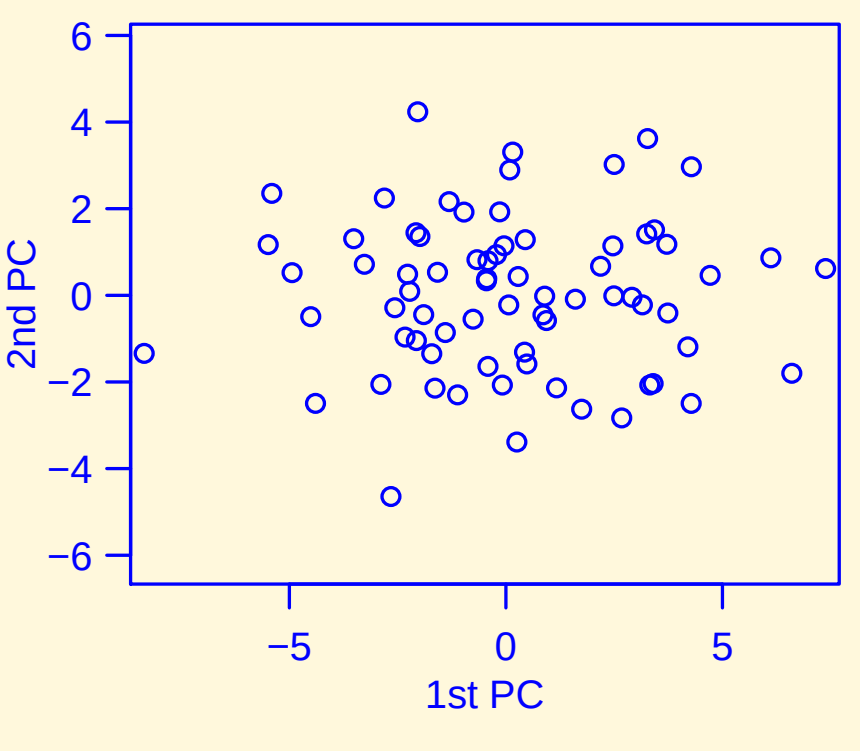
```
> library(rrcov); library(mvtnorm); library(MASS)

## Data Simulation
> set.seed(4711)
> mN <- rmvnorm(n=72, mean = c(-2,0, 1), sigma = diag(c(1,12,3)))
> mN[c(29,30,31),1] <- c(8, 5, 10)

## PCA
> mN.pc <- princomp(mN, cor=FALSE)
> mN.Rpc <- PcaCov(mN, scale=FALSE)
> structure(cbind(loadings(mN.pc)[,1:3], mN.Rpc@loadings[,1:3]), class="loadings")
Loadings:
                 classical                          robust
        Comp.1   Comp.2   Comp.3          PC1      PC2     PC3
 [1,]            0.998                                   0.996
 [2,]    0.977           -0.212        0.978   -0.203
 [3,]    0.212            0.975        0.206    0.976

## Plot results
> mN.pc.p <- predict(mN.pc)
> mN.Rpc.p <- predict(mN.Rpc)
> par(mfrow=c(1,2)), las=1
> eqscplot(mN.pc.p[,1:2])
> eqscplot(mN.Rpc.p[,1:2])
```

Simulatet Example: Data in the first two PCs

classical



robust



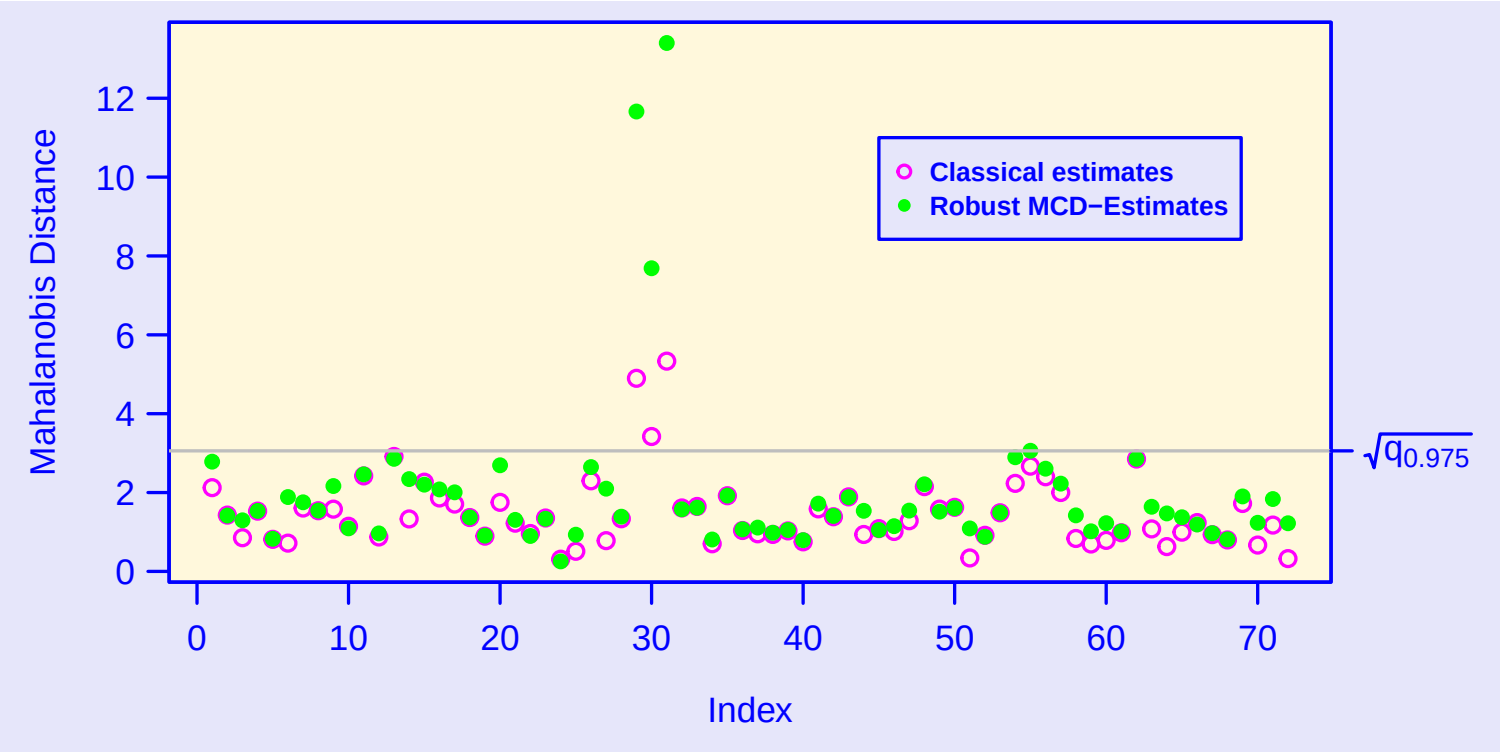
Outliers clearly visible

Outliers not visible

in the first two PCs

Simulatet Example: Mahalanobis Distances

Mahalanobis distances versus the observation number.
It is based either on the classical estimation of the covariance matrix (magenta) or on a robust MCD estimation (green) as in `PcaCov(...)`



Outliers are clearly visible in both cases.

4.3 Linear Discriminant Analysis

Linear Discriminant Analysis is an **explorative** multivariate data analysis technique describing the difference between several groups. These differences can be **visualized** by a scatterplot on the canonical variates.

Based on the result from a linear discriminant analysis, we can subdivide the space spanned by the observations into as many subspaces as there are groups. The partition can then be used to assign new observations to one of the groups ☞ **classification**.

Fisher's Linear Discriminant Analysis

Find the linear combinations of the variables which lead to a maximum separation between the centres of the groups measured with respect to the variability within the groups.

Let $\boldsymbol{W}$ be the covariance matrix within a group and $\boldsymbol{B}$ the covariance matrix of the group centres. The optimal linear combination $\underline{a}_1$ is given by

$$\underline{a}_1 = \arg\max_{\underline{a}} \frac{\underline{a}^T \boldsymbol{B}\, \underline{a}}{\underline{a}^T \boldsymbol{W}\, \underline{a}}\,; \qquad (*)$$

i.e., the solution is $\underline{a}_1 = \boldsymbol{W}^{-1/2}\, \underline{e}_1$, where $\underline{e}_1$ is the eigenvector to the largest eigenvalue of the matrix

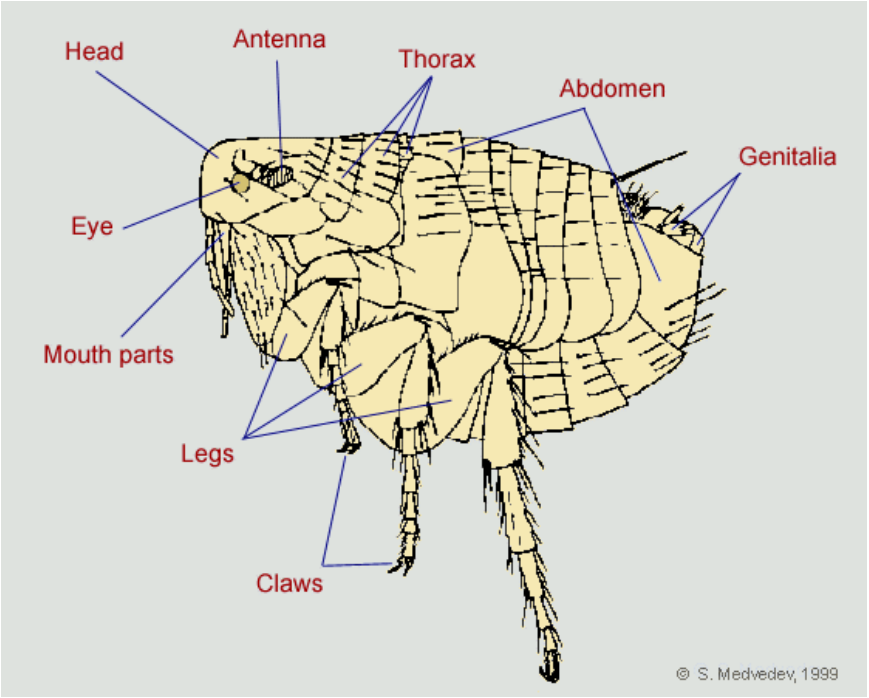
$$\boldsymbol{W}^{-1/2}\, \boldsymbol{B}\, \boldsymbol{W}^{-1/2}\,.$$

Additional vectors $\underline{a}_k$, $k > 1$, are built by optimizing (*) with the constraint that $\underline{a}_k$ is orthogonal to $\underline{a}_1, \ldots, \underline{a}_{k-1}$.

The values $z_i^{(k)} = \underline{a}_k^T\, \underline{x}_i$, $i = 1, 2, \ldots, n$, form the k-th discriminant variable.

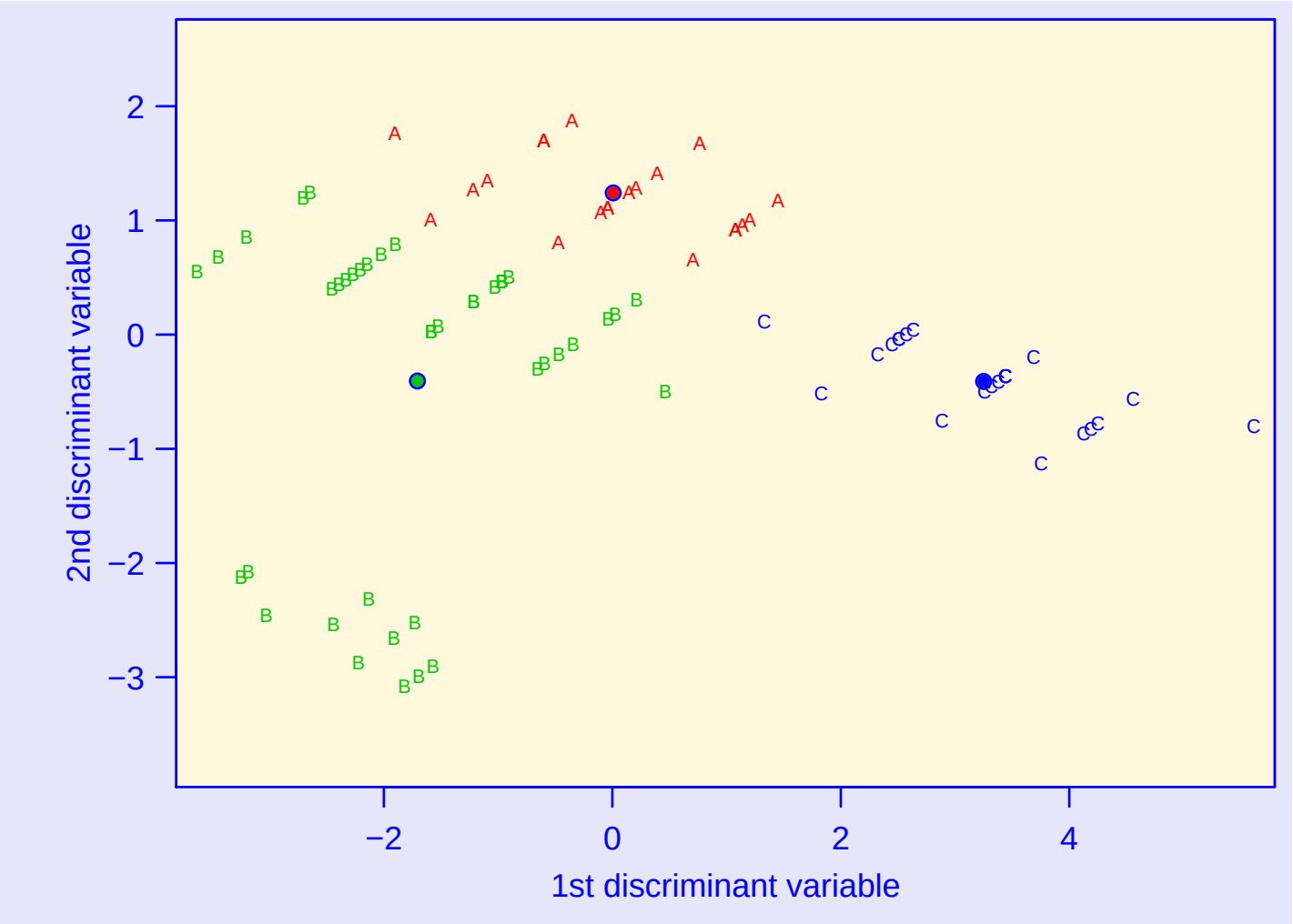
Example Flea

Lubischew (1962) collected data on the genus of flea beetle Chaetocnema, which contains three species: concinna, heikertingeri, and heptapotamica. Measurements were made on the `width` (in microns) and `angle` (in units of 7.5°) of the aedeagus of each beetle. The goal of the original study was to form a classification rule to distinguish the three species.



Example Flea

Plot of the "slightly" modified data in the first two discriminant variates:



- The **covariance matrix** $\boldsymbol{W}$ obviously represents the **Gaussian distribution** of the data within each class
- There is just a **faint idea of a model** how the (usually few) groups centres should scatter ☞ **exploration of their geometric constellation**

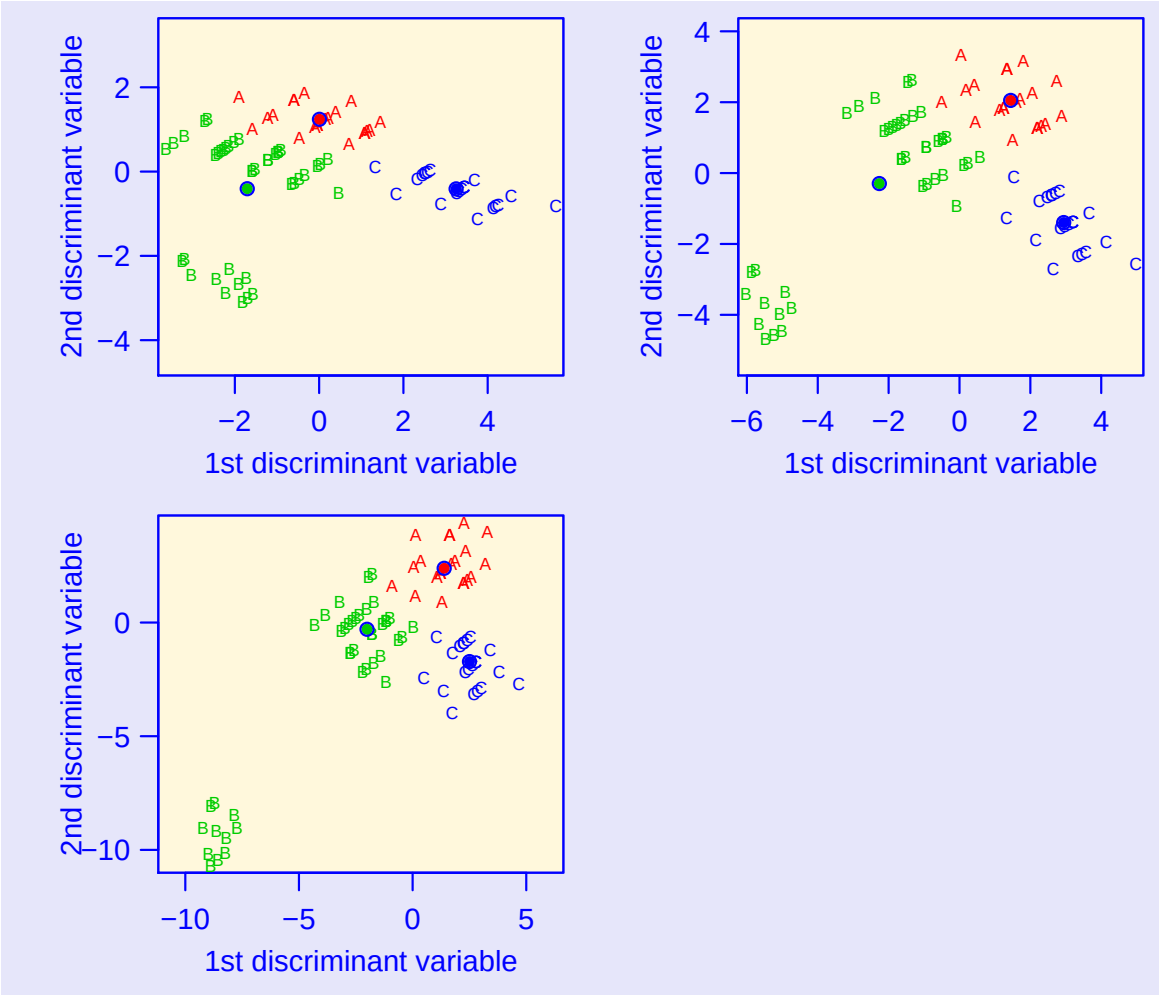
Thus,

Approach A: Estimate the **covariance matrix** $\boldsymbol{W}$ **robustly** and treat the matrix $\boldsymbol{B}$ as in the standard procedure
☞ `lda(..., method="mve")` of R package MASS

Approach B: Estimate both the **covarianz matrix** $\boldsymbol{W}$ **and the locations of the groups robustly**. The matrix $\boldsymbol{B}$ is treated as in Approach A:
☞ `rlda(...)` (own contribution).

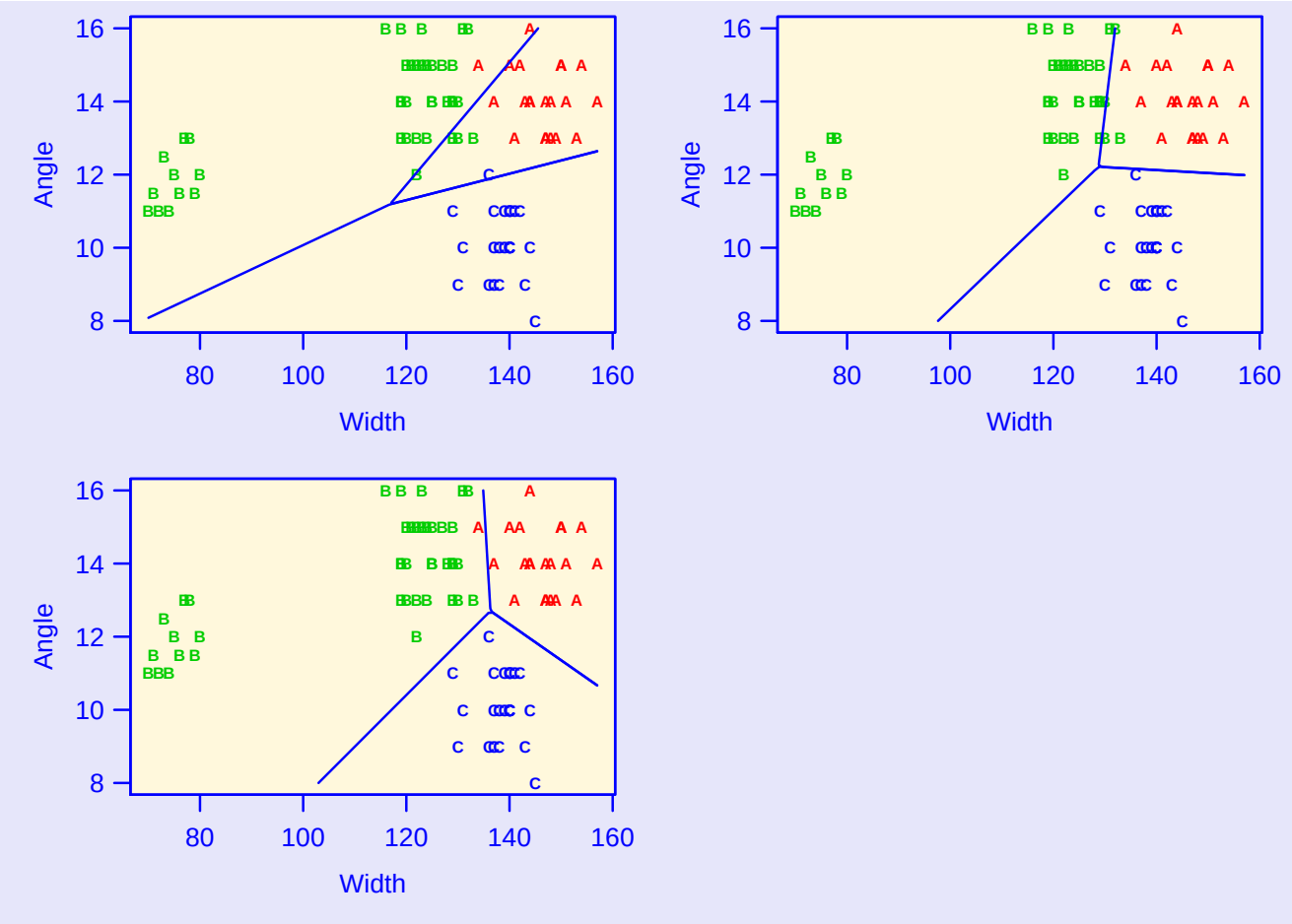
Example Flea

Scatterplot of the data in the canonical variates using the classical method (upper left), Approach A (upper right), and Approach B (lower left).



Example Flea

Plot of the original variables overlaid by the group borders which are based on the classical method (upper left), Approach A (upper right), and Approach B (lower left).

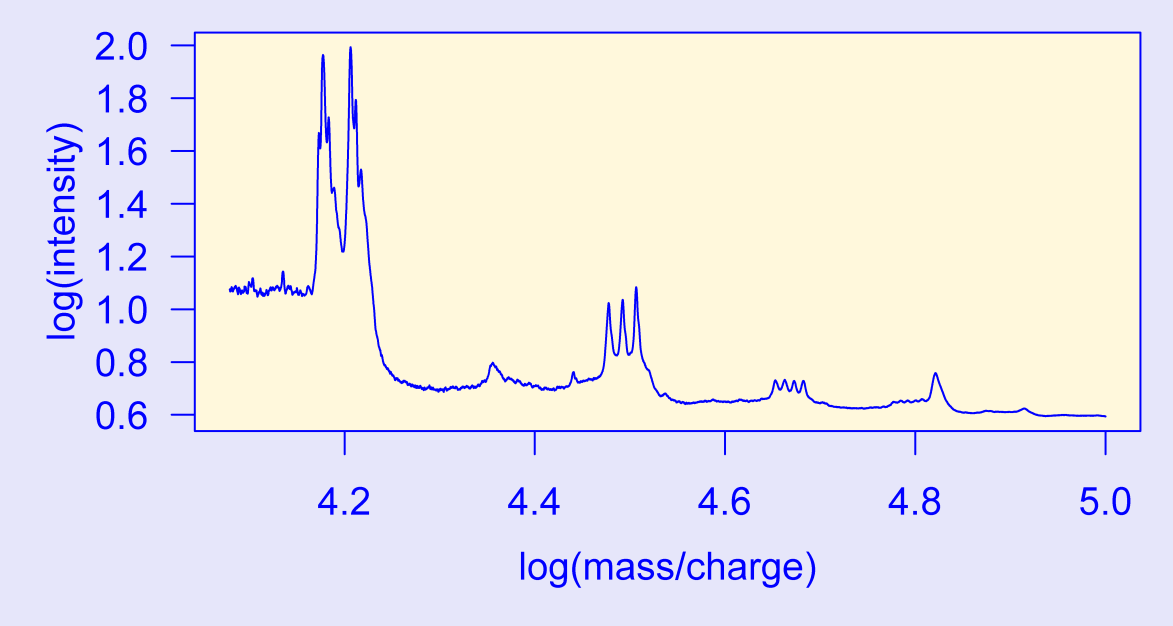


5 Baseline Removal Using Robust Local Regression

5.1 A Motivating Example From Mass Spectroscopy

The spectrum was taken from a sample of sheep blood. The instrument used was a so called SELDI TOF (Surface Enhanced Laser Desorption Ionisation, Time Of Flight) Mass Spectrometer.

The spectrum on the left consists of sharp features superimposed upon a continuous, slowly varying baseline.

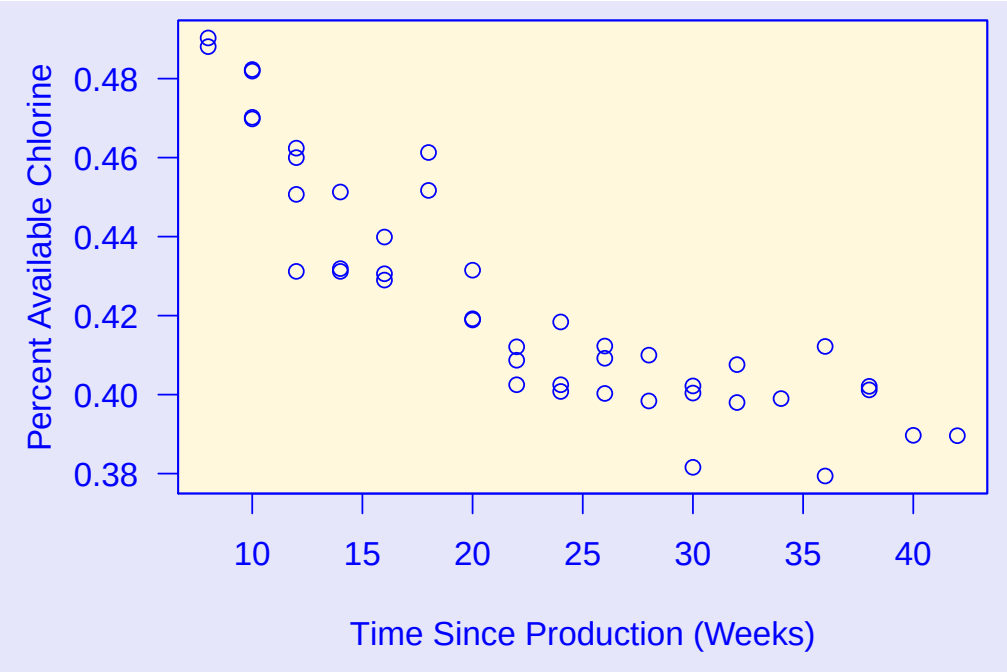


Goal: Remove baseline by robust local regression.

5.2 A Simpler Problem to Start With: Local Regression

Example Chlorine:

The investigation involved a product A, which must have a fraction of 0.50 of available chlorine at the time of manufacture. The fraction of available chlorine in the product decreases with time. Since theoretical calculations are not feasible, a study was run to get some insight into the decrease.



In regression analysis we study

$$Y_i = h\langle x_i;\, \underline{\beta}\rangle + E_i\,.$$

The unstructured deviations from the function h are modelled by random errors E_i which are normally distributed with mean 0 and constant variance.

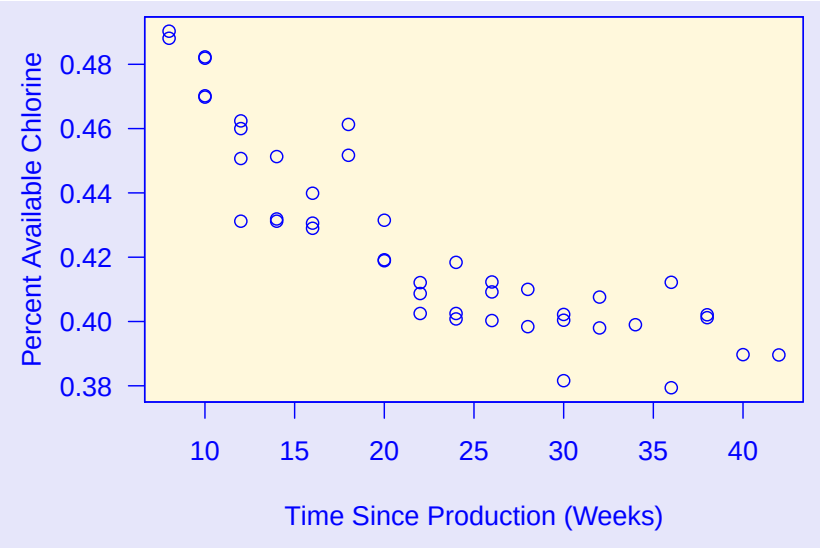
In **linear** regression:

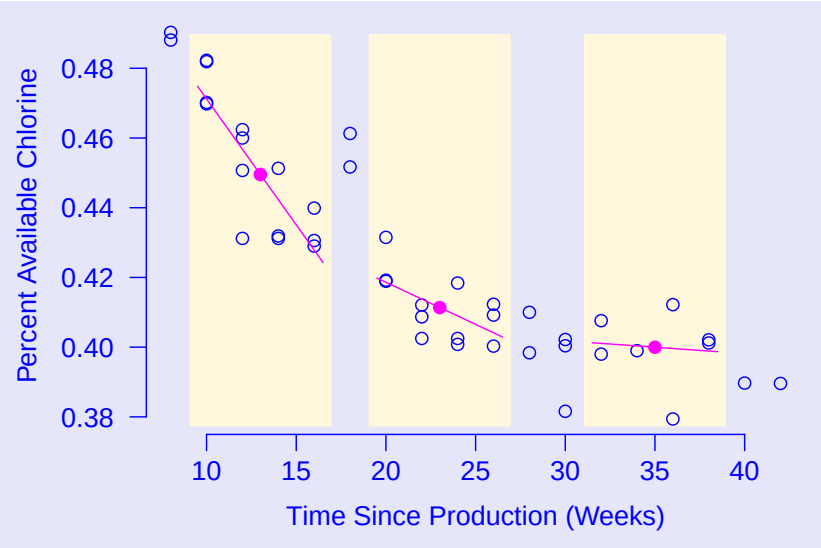
$$h\langle x_i;\, \underline{\beta}\rangle = \beta_0 + \beta_1 \widetilde{x_i}\,.$$

What can be done, if the function h is **non-linear**, also w.r.t. the parameter $\underline{\beta}$?

- ☞ Nonlinear regression (cf. next block course)
- ☞ relationship h is determined from the data by a *smoother*

Local Regression – Basic Idea





- Select a **window** around a point z_1 at which $h(z_1)$ is to be estimated
- Select window **width** so that h is approximated well by a straight line
- **Fit** the straight line to the data within the window and and predict at z_1: ☞ $\widehat{h}(z_1)$.
- These steps are applied to a **grid of points** $z_1, \ldots, z_N$ which covers the range of the exploratory variable: ☞ $\widehat{h}(z_1), \ldots, \widehat{h}(z_N)$.
- To **visualize** the estimated function $\widehat{h}$, the points $(z_k, \widehat{h}_k)$ are connected by line segments to each other.

Local regression – a weighted least-square problem

The estimated function value at z_1 is $\widehat{h}(z_1) = \widehat{\beta}_0$,

where $\widehat{\beta}_0$ is the first component of

$$\underline{\widehat{\beta}}(z_1) = \arg\min_{\underline{\beta}} \sum_{i=1}^{n} w_r \langle x_i \rangle \; K \left\langle \frac{x_i - z_1}{b_w} \right\rangle \; (y_i - (\beta_0 + \beta_1 \, (x_i - z_1)))^2$$

b_w is called the **bandwidth** and $K \langle ((x_i - z_1)/b_w \rangle$ **kernel weights**.

To be specified:

- Choice of bandwidth b_w: adaptive such that, e.g. 2/3 of the points are within the window
- Choice of kernel weight $K \langle (x_i - z_1)/b_w \rangle$

  e.g., Tukey's tricube kernel $\quad K \left\langle \frac{x_i - z_1}{b_w} \right\rangle = \left[\max \left\{ 1 - \left| \frac{x_i - z_1}{b_w} \right|^3, 0 \right\} \right]^3$

  $K \langle \cdot \rangle$ is zero outside $z_1 \pm b_w$.
- $w_r \langle x_i \rangle$ are implicit weights with which robustness can be achieved.
  e.g., Tukey's biweight robustness weights

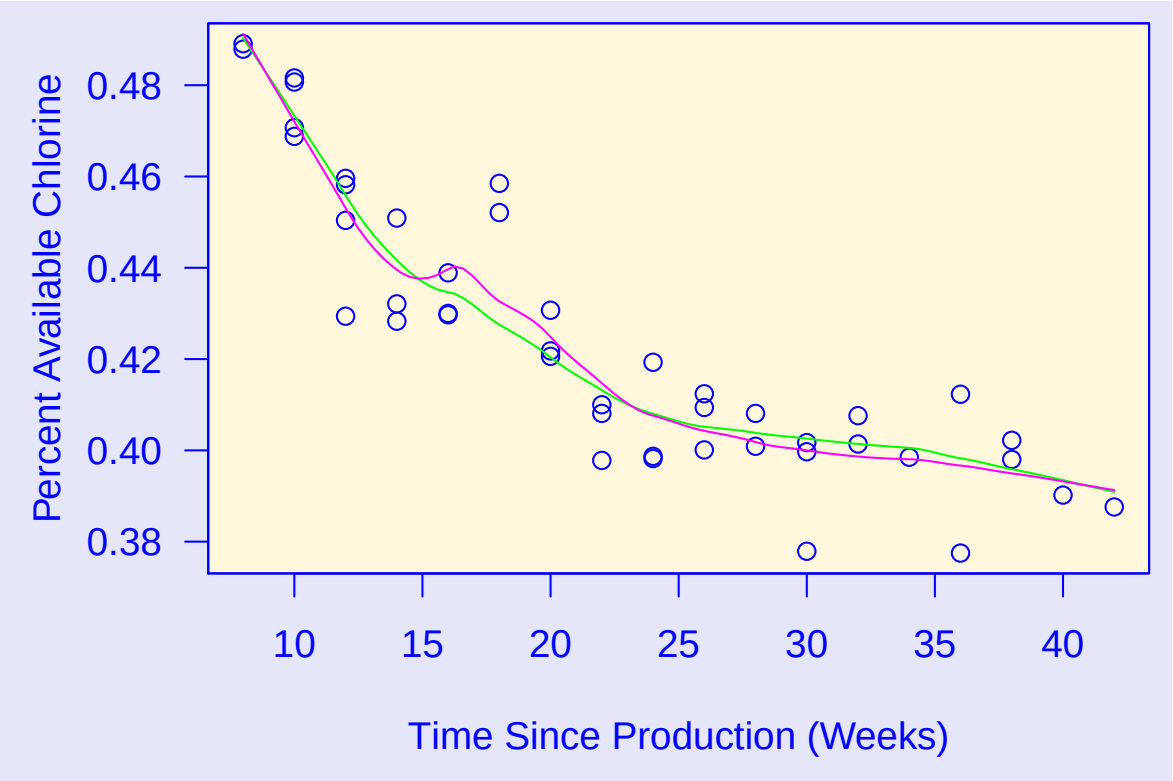
  $$w_r \langle x_i \rangle = \left(\max \left\langle 1 - (\widetilde{r}_i / b)^2, 0 \right\rangle \right)^2 \quad \text{with } \widetilde{r}_i = (y_i - \widehat{h} \langle x_i \rangle) / \widehat{\sigma}_{\text{MAV}} \text{ and } b = 4.05$$

(For more details on the LOWESS procedure see my lecture notes)

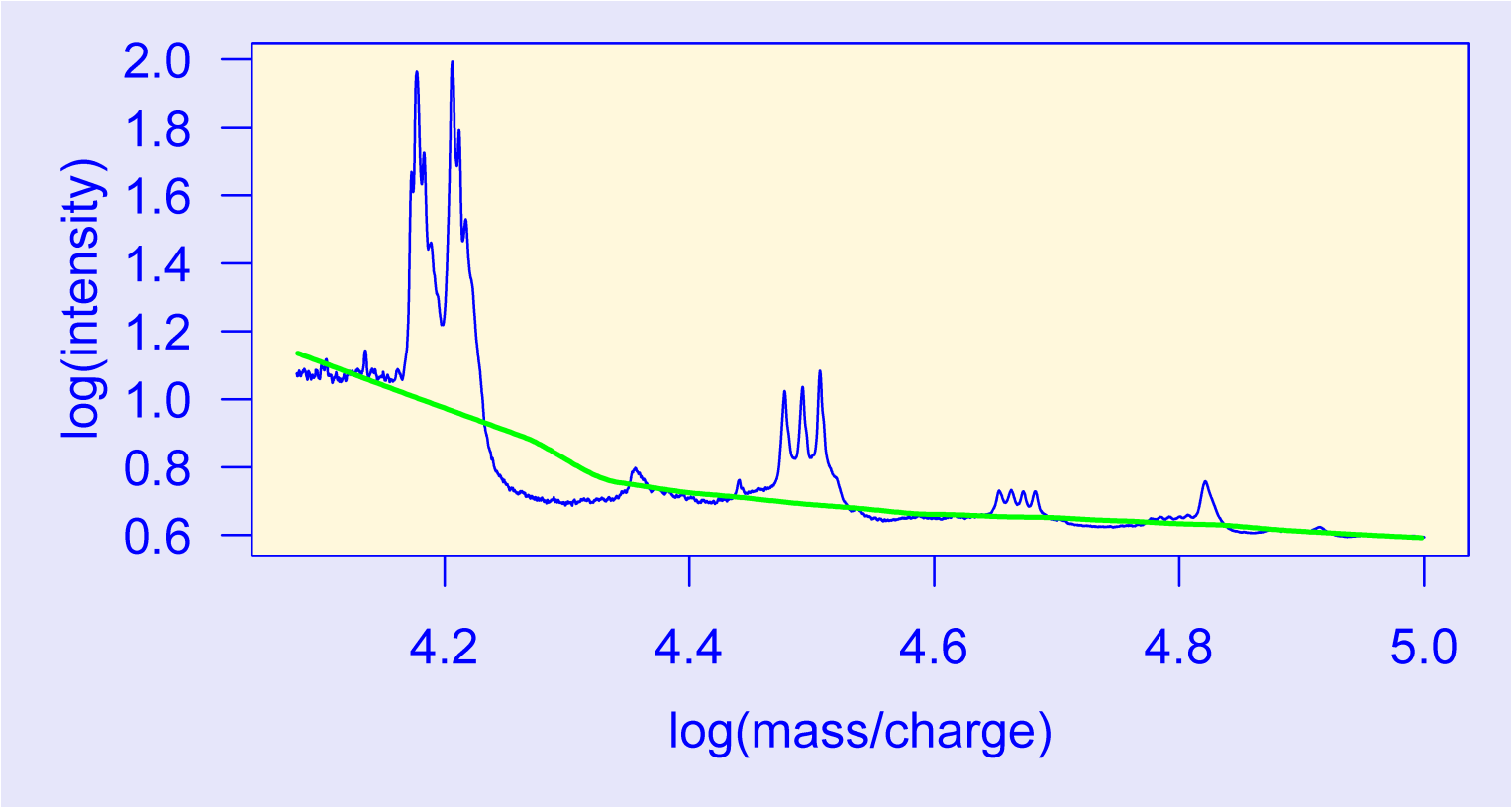
Example Chlorine: LOWESS Fit

Non-robust (magenta) and robust (green)

```
> clr <- loess(YY ~ x, data=Chlor, span=0.35, degree=1, family="gaussian")
> lines(xnew, predict(clr, xnew), col="magenta")
> rlr <- loess(YY ~ x, data=Chlor, span=0.35, degree=1, family="symmetric")
> lines(xnew, predict(rlr, xnew), col="green")
```



Apply LOWESS/LOESS to the Mass Spectroscopy Data



- `loess(I ~ mz, data=MS1, span=0.35, degree=1, family="symmetric")`
- **This is of no use** - The approach does not work in this case

Modify LOWESS/LOESS

New View:
- The baseline is contaminated by the target signal.
- The contamination is one-sided.

☞ Use an asymmetric robustness weight function $w_r(t_i)$ in

$$\widehat{\underline{\beta}}(z_1) = \arg\min_{\underline{\beta}} \sum_{i=1}^{n} w_r(t_i)\, K\left(\frac{t_i - z_1}{b_w}\right) \cdot \left[y_i - \{\beta_0 + \beta_1\,(t_i - z_1)\}\right]^2$$

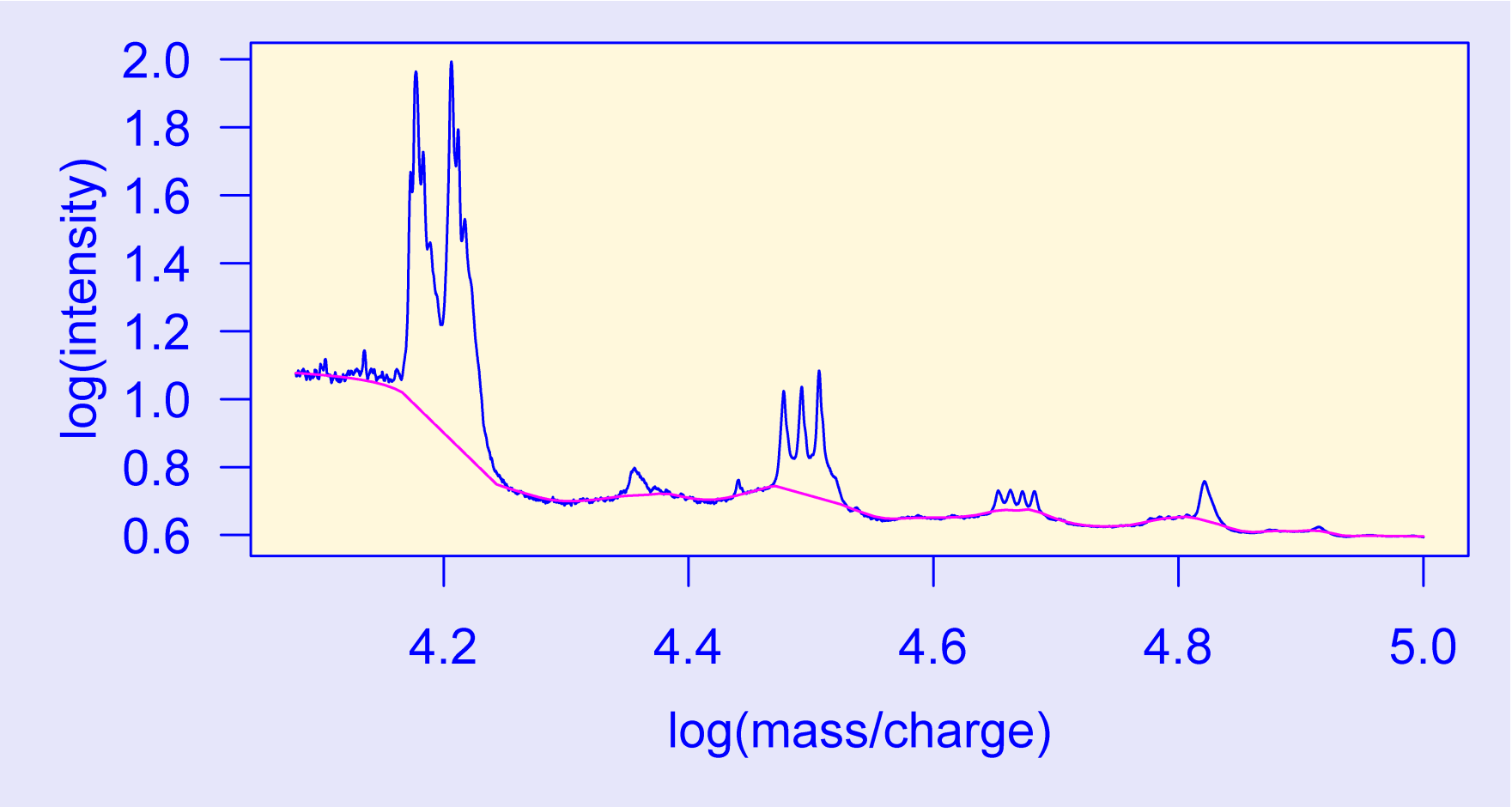
as, e.g.,

$$w_r(x_i) = \begin{cases} 1 & \text{if } r_i < 0 \\ \left[\max\left\{1 - (\widetilde{r}_i/b)^2, 0\right\}\right]^2 & \text{otherwise,} \end{cases}$$

- good choice for b is 3.5 (or any value between 3 and 4).
- Bandwidth b_w: at least 2 $\times$ the longest period in which the baseline is contaminated by the target signal.
- σ is estimated from the negative residuals.

☞ **Robust fitting of baseline with `rfbaseline()` in the R package `IDPmisc`**

Example from Mass Spectroscopy: `rfbaseline()`




```
> library(IDPmisc)
> MS1.rfb4 <- rfbaseline(x=MS1$mz, y=MS1$I, NoXP=1400, maxit=c(5,0),
                         DOT=TRUE, Scale=rfbaselineScale)
```

Take Home Message Half-Day 3

- **Multivariate statistical analysis are often based on the covariance matrix**,

  because the multivariate Gaussian distribution is such a convenient model.

- Robust Estimators of the covariance matrix with breakdown point of 1/2 are able to detect multidimensional outlieres fast and reliably.

- The clearer a procedure is based on a model

  the better the procedure can be robustified

- Principal component analysis (PCA), which is based on a robustly estimated covariance matrix, may yield additional insight.

- If there are outliers, the robustified linear discriminant analysis (LDA) shows the difference between the groups clearer and estimates the class borders more reliable.

- There are useful "misuses" of robust methods . . . ☞ Baseline Removal

Take Home Message from "Robust Fitting"

Suitable robust methods are implemented in R for

linear regression models	`lmrob(...)` in the package `robustbase` `plot(`*`lmrob object`*`)` ☞ residual analysis
GLM	`glmrob(...)` in the package `robustbase`
Model Comparision	`anova(`*`lmrob - or glmrob object`*`)` in the package `robustbase`
covariance matrices	`CovRobust(...)` in the package `rrcov`
PCA	`PcaCov(...)` in the package `rrcov`
linear discriminant analysis	`rlda(...)` (own contribution)
Baseline removal	`rfbaseline(...)` in the package `IDPmisc`
...	

Robust methods are essential

in the daily business of statistical data analysis