Zurich University
of Applied Sciences

zh
aw

School of
Engineering

IDP Institute of
Data Analysis and
Process Design

Half-Day 3: Multivariate Analysis Based on Robust Fitting

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

WBL Statistik 2016 — Robust Fitting

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

3.1 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 $\boldsymbol{\Sigma}$.

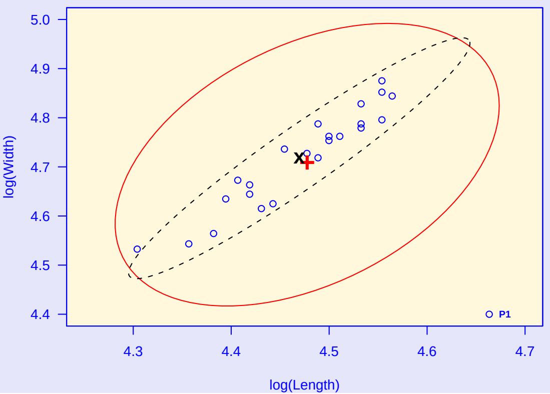
The optimal estimates for the parameters are

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

Example Carapaces:
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.

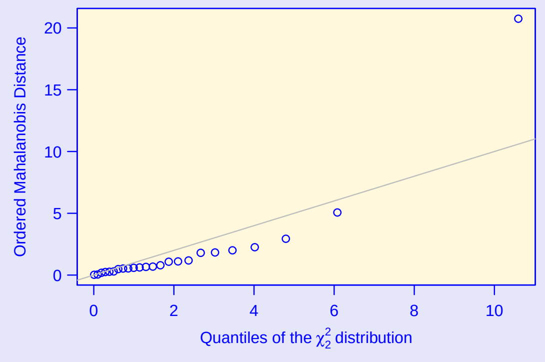
The **standard estimations** are based on the data including (solid, +) or excluding (dotted, x) observation P1.



Mahalanobis Distances to Detect Outliers

In a classical setting, Mahalanobis distances $u_i = (\underline{x}_i - \underline{\mu})^T \boldsymbol{\Sigma}^{-1} (\underline{x}_i - \underline{\mu})$ are used to detect outliers: U_i is χ^2_m distributed; m: number of variables

Q-Q plot of the Mahalanobis distances versus χ^2_2 distribution for modified Carapaces data.



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

Estimators based on a robust scale:

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 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 $\mathrm{Var}\left\langle d\left\langle \underline{x}_i, \underline{\mu}, \boldsymbol{\Sigma}^* \right\rangle \right\rangle = \sigma^2 \cdot m$

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

- 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**.

  Its 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 as $S\langle d_i\rangle = \sum_{i=1}^{h} d_{(i)}$ with $h = \frac{n+m}{2}$ $(m = \#$ variables).

  ☞ **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 and is more efficient than the MVE estimator.

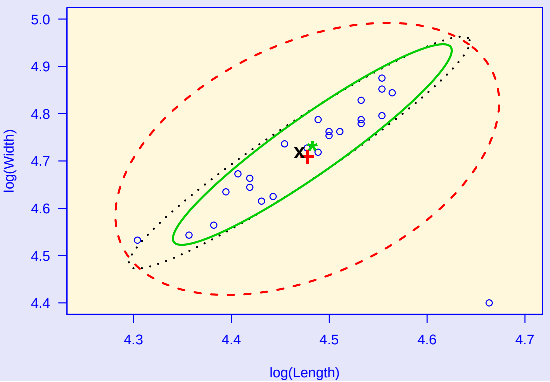
The computation of both estimators is, however, quite intensive as they are based on stochastic resampling algorithms.

Estimated covariance matrices for modified Carapaces 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**.



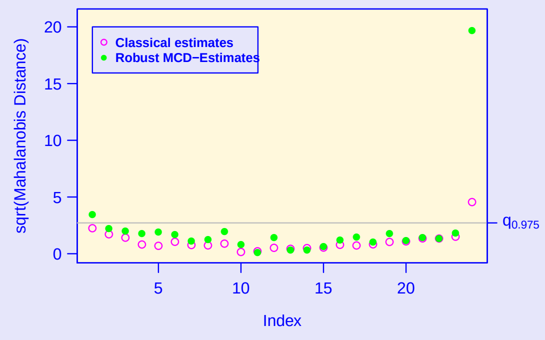
There seems to be a second outlier (see l.h.s.)

Mahalanobis Distances

To visualize the Mahalanobis distance, its square-root transformed value is plotted versus the observation number.

Observations which are above the 97.5%-χ^2_2 quantile ($=q_{0.975}$) line can be identified as outliers.

Plot for the modified Carapaces 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 $S\left\langle d_i \right\rangle$ satisfies

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

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

- the **Stahel-Donoho estimator**.
  Idea: A multivariate outlier should also be an outliers in *some* univariate projection
  ☞ 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: $c^{(x,y)} = \frac{1}{4}\left(\left(S\left\langle x+y \right\rangle\right)^2 - \left(S\left\langle x-y \right\rangle\right)^2\right)$,
  where $S\left\langle . \right\rangle$ is a robust estimation of σ.

  A correction is needed to obtain a semi-definite matrix.

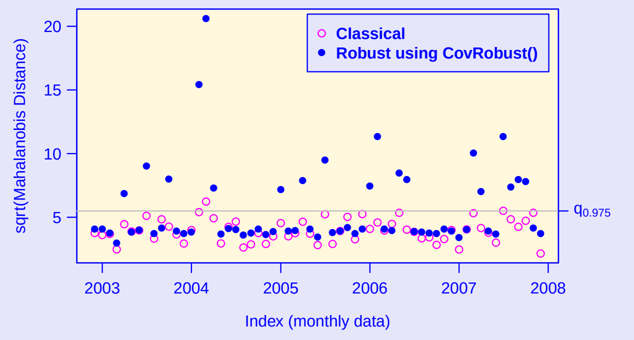
R functions

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 if $n > 50'000$ and/or $p > 20$
- `covMcd(...)` and `covOGK(...)` from R package `robustbase`
- `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.



3.2 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 (= information).
- Finding structures like subgroups or outliers
- transformation of exploratory variables to avoid collinearity
  ☞ principal regression analysis.

- The principal components specify uncorrelated directions (linear combinations of the measured characteristics) 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 for PCA, we **cannot robustify PCA**.

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 Mahalanobis distances for finding outliers.

Example Carapaces:

Classical PCA

```
Importance of components:
                          Comp.1     Comp.2
Standard deviation     0.1219237 0.06720862
Proportion of Variance 0.7669535 0.23304647
Cumulative Proportion  0.7669535 1.00000000
```

PCA based on a robustly estimated covariance matrix

```
Importance of components:
                          Comp.1     Comp.2
Standard deviation     0.1029720 0.01708216
Proportion of Variance 0.9732171 0.02678286
Cumulative Proportion  0.9732171 1.00000000
```

3.3 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

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

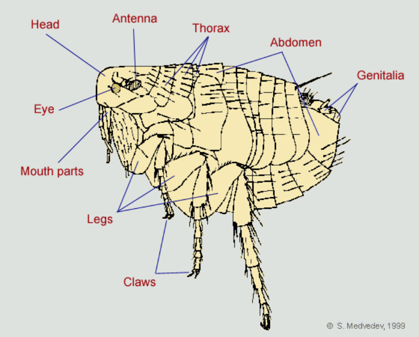
i.e., the solution is $\underline{a}_1 = \boldsymbol{W}^{-1/2} \, \underline{e}_1$, where $\underline{e}_1$ is the largest eigenvalue of the matrix

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

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

Example Flea

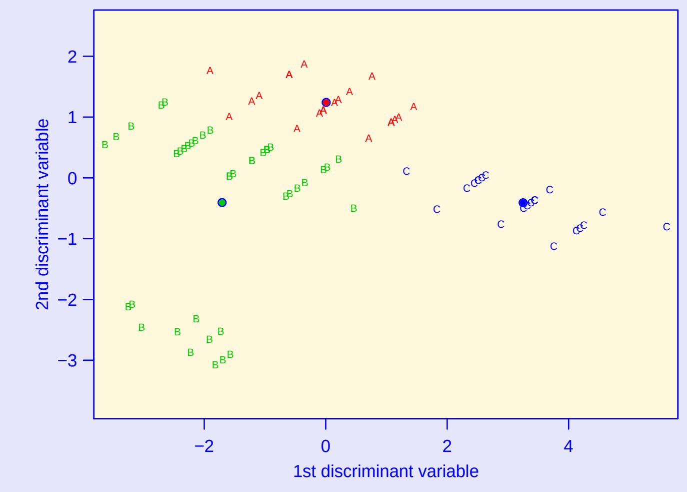
Lubischew (1962) collected data on the genus of flea beetle Chaetocnema, which contains three species: concinna, heikertingeri, and heptapotamica. Measurements were made on the `width` (in microns) and `angle` (in units of 7.5°) of the aedeagus of each beetle. The goal of the original study was to form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



© S. Medvedev, 1999

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**

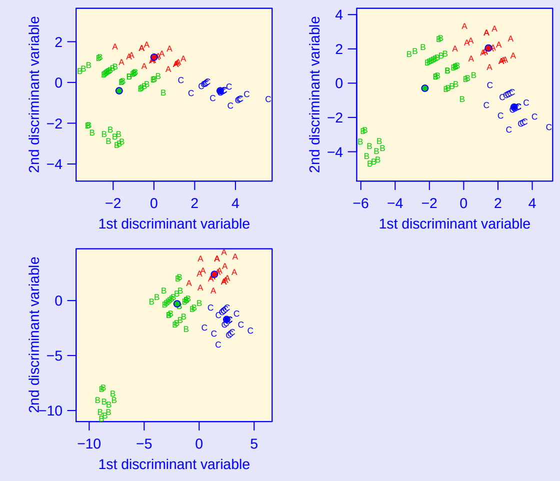
Thus,

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).

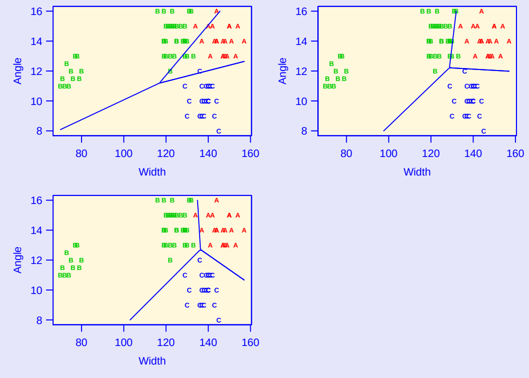
Example Flea

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).

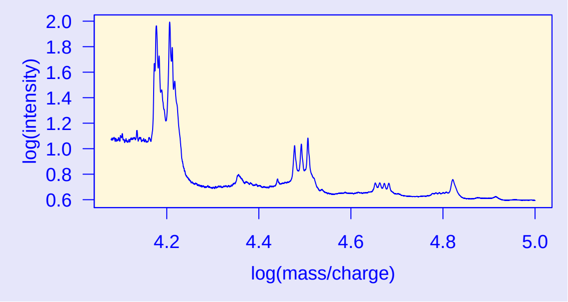


2.4 Baseline Removal Using Robust Local Regression

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.



A Simpler Problem to Start With

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.

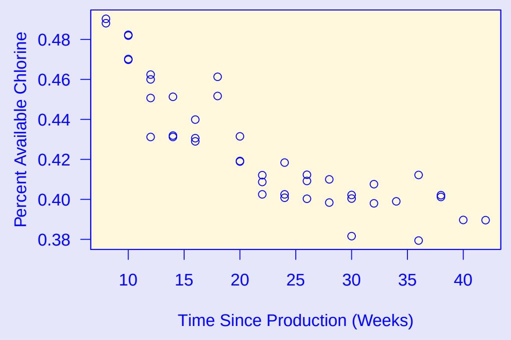
In **linear** regression:

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

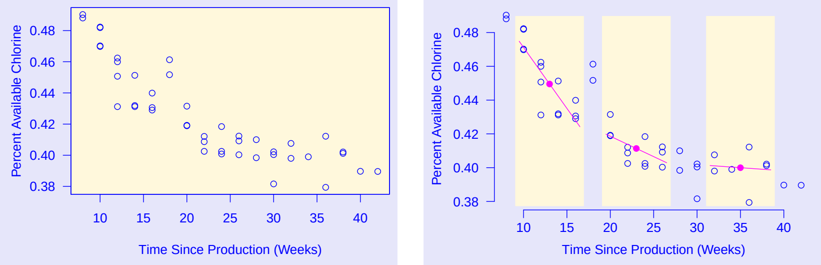
What can be done, if the function h is **nonlinear** w.r.t. the parameter $\underline{\beta}$?

☞ 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

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

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

To be specified:

- Choice of bandwidth b_w
- Choice of kernel weight $K\left\langle(x_i - z_1)/b_w\right\rangle$

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

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

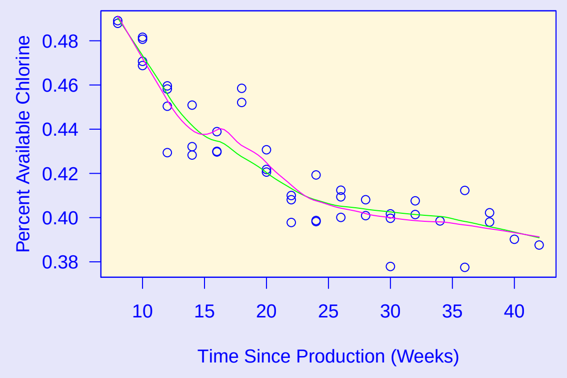
  $w_r\left\langle x_i\right\rangle = \left(\max\left\langle 1 - (\widetilde{r}_i/b)^2, 0\right\rangle\right)^2$ with $\widetilde{r}_i = (y_i - \widehat{h}\left\langle x_i\right\rangle)/\widehat{\sigma}_{\text{MAV}}$ and $b = 4.05$

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

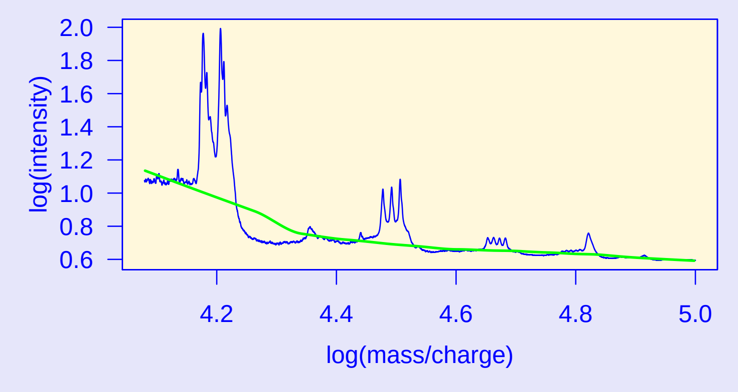
Example Chlorine:

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



This is of no use - My approach is too naive.

Modify LOWESS/LOESS

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

☞ Use an asymmetric robustness weight function in

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

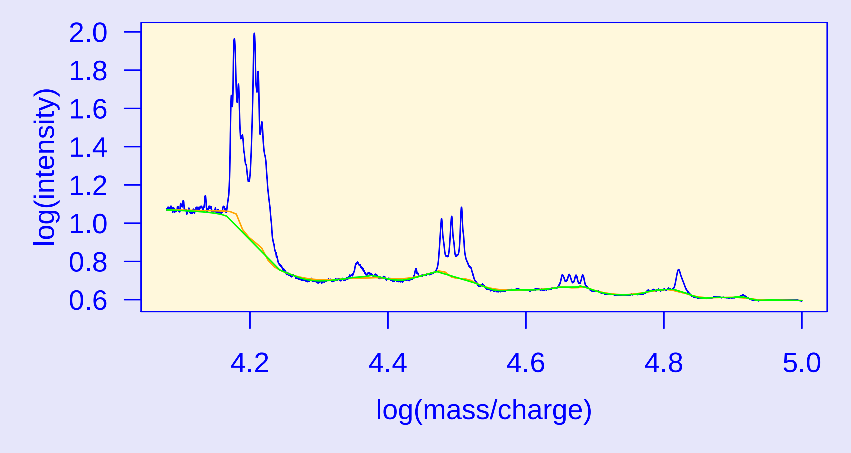
as, e.g.,

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

- good choice for b is 3.5 (or any value between 3 and 4).
- Bandwidth b_w: at least 2 × 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()`



A. Ruckstuhl et Al. (2012), *Robust extraction of baseline signal of atmospheric trace species using local regression*, J. Atmospheric Measurement Techniques.

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 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`
GLM	`glmrob(...)` in the package `robustbase`
Model Comparision	`anova(`*lmrob - or glmrob object*`)` in the package `robustbase`
covariance matrices	`CovRobust(...)` 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