Outlier Detection with Application to Geochemistry

Peter Filzmoser

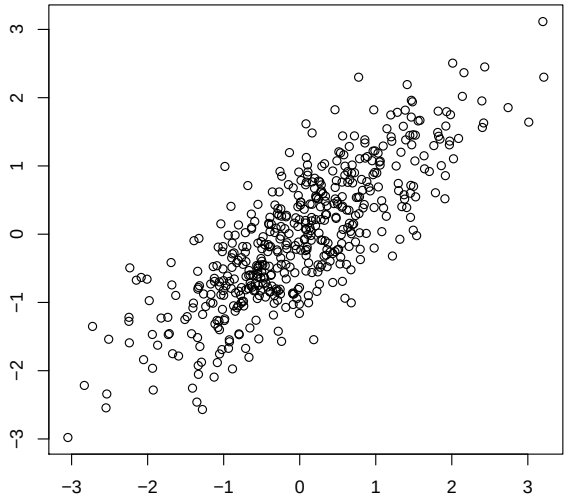
Department of Statistics and Probability Theory
Vienna University of Technology, Austria

Vienna, Austria

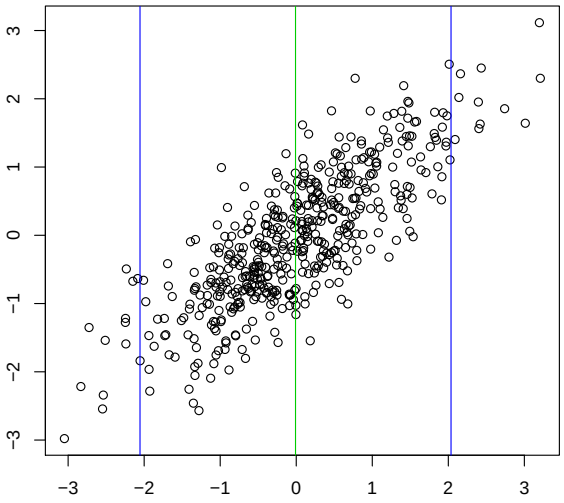
June 16, 2006

Vienna University of Technology

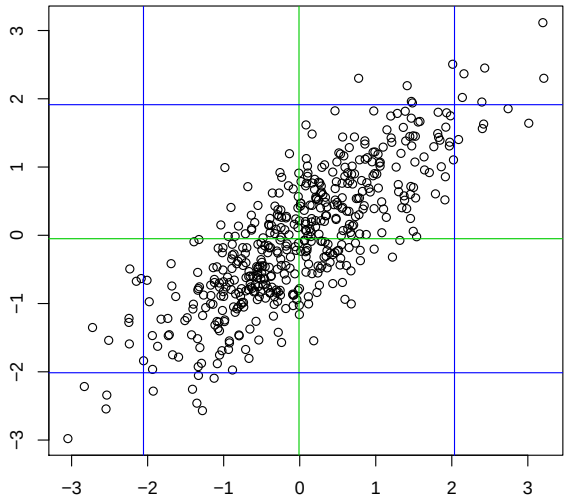
Univariate versus Multivariate Outliers



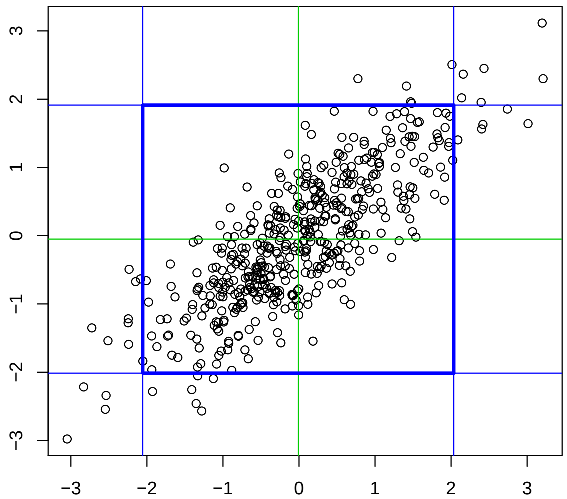
Univariate versus Multivariate Outliers



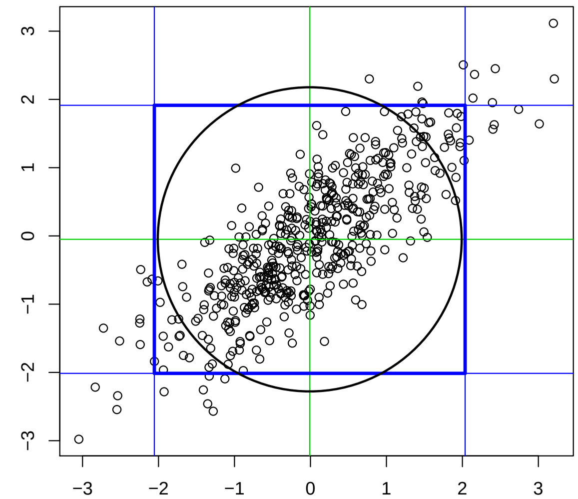
Univariate versus Multivariate Outliers



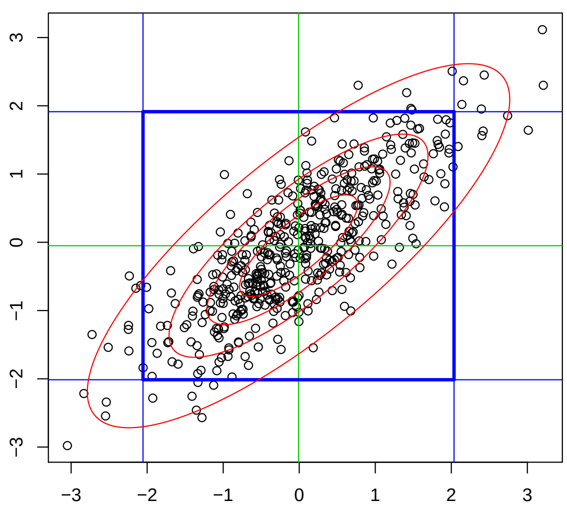
Univariate versus Multivariate Outliers



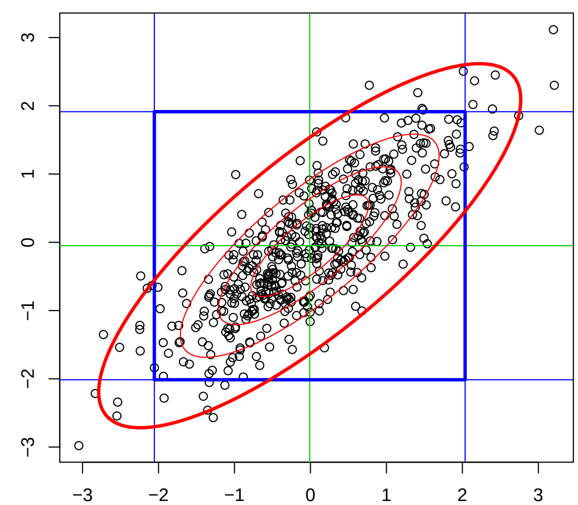
Univariate versus Multivariate Outliers



Univariate versus Multivariate Outliers



Univariate versus Multivariate Outliers



Multivariate Outlier Detection Methods

Standard methods are based on the **Mahalanobis distances** (MD):

$$\text{MD}_i := d(\boldsymbol{x}_i, \boldsymbol{t}, \boldsymbol{C}) = \{(\boldsymbol{x}_i - \boldsymbol{t})^\top \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{t})\}^{1/2}$$

for a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in I\!R^p$ and estimators of location $\boldsymbol{t}$ and covariance $\boldsymbol{C}$.

$\Longrightarrow$ **Robust estimates of location and covariance are needed!**

Outlier detection:
Outliers will typically have large distance. If multivariate normal distribution is assumed, MD_i^2 is approx. χ^2_p distributed.

$\Longrightarrow$ suspect observations: $\text{MD}_i^2 > \chi^2_{p,0.975}$

- does not account for different sample size
- χ^2_p-approximation is poor

Garrett (1989)

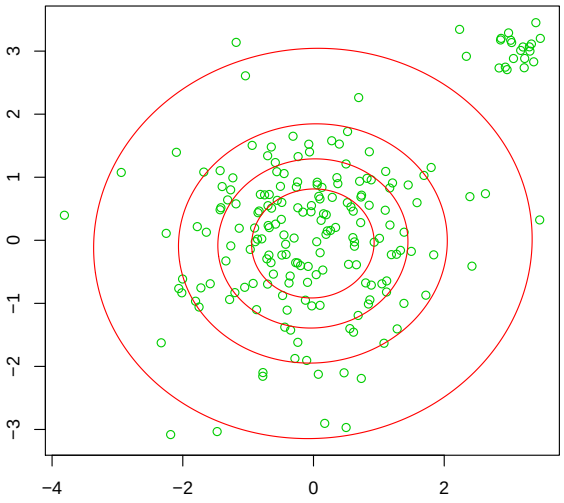
Chi-square plot:

Plot robust MD_i^2 against quantiles of χ^2_p.

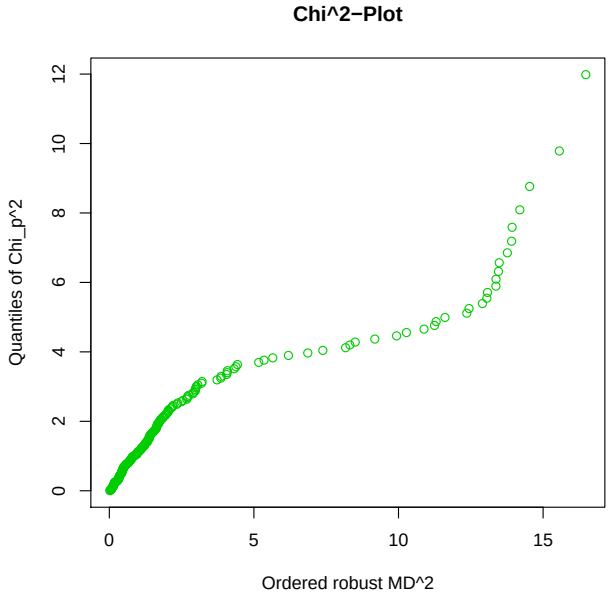
$\Longrightarrow$ iterative deletion of points with large distance until a straight line appears.

Drawback: no automatic procedure, needs user interaction.

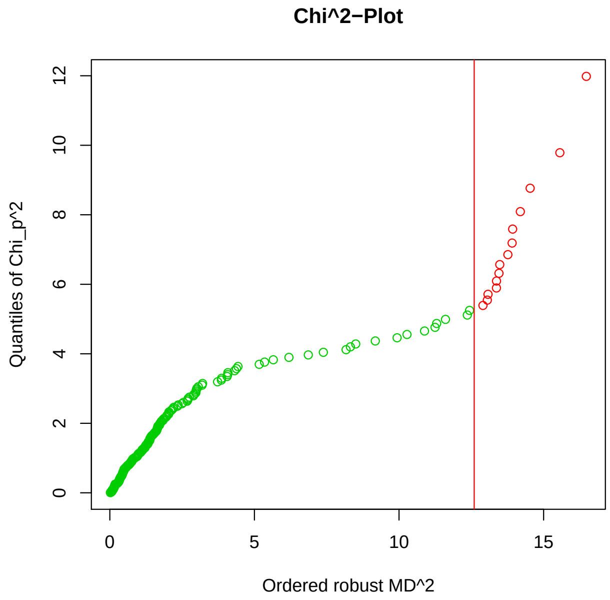
Example: Simulated data with outliers



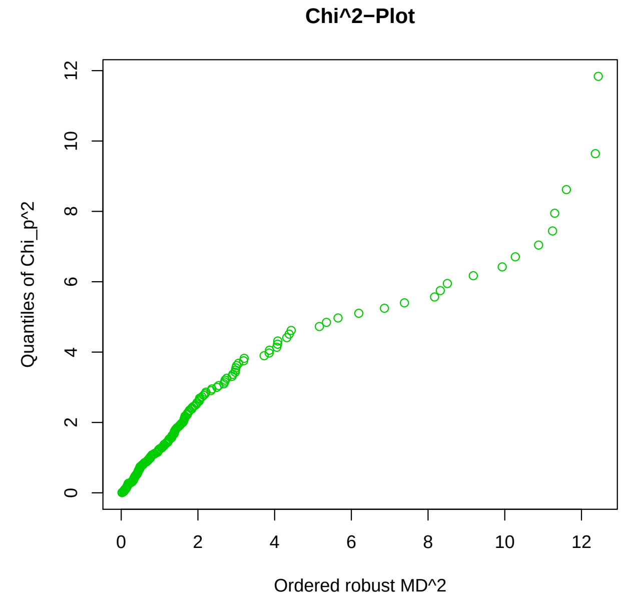
Iterative deletion of outliers:



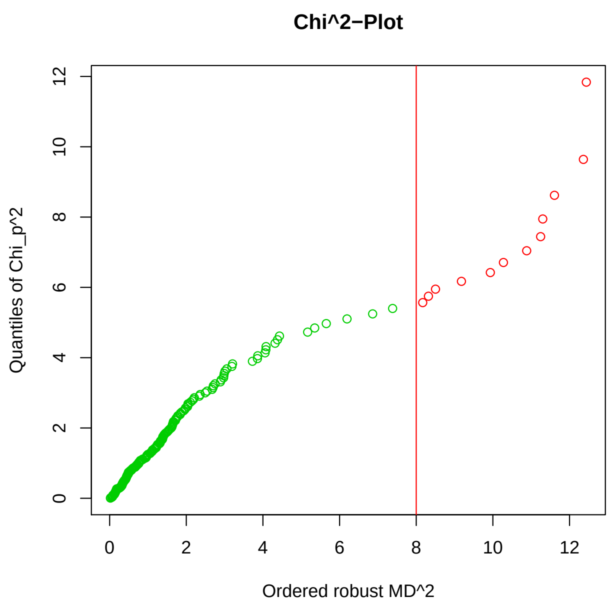
Iterative deletion of outliers:



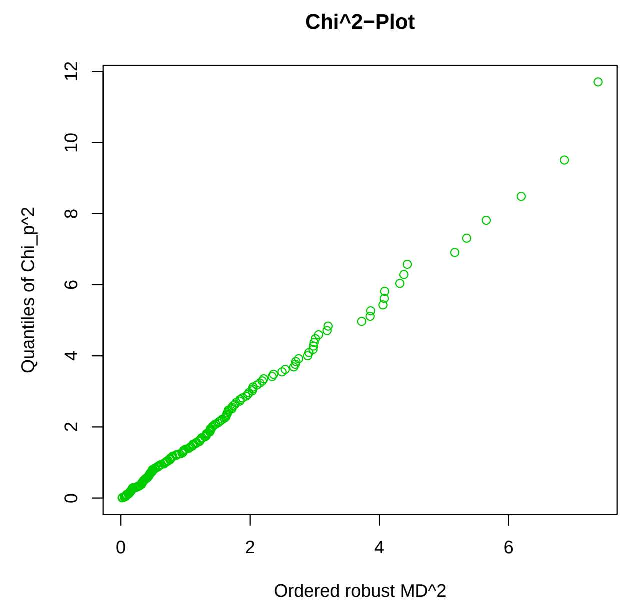
Iterative deletion of outliers:



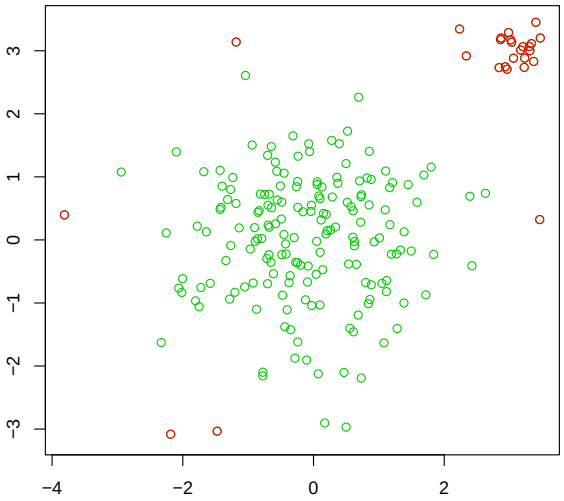
Iterative deletion of outliers:



Iterative deletion of outliers:



Example: Simulated data with outliers



New Proposal

$G(u)$ ... theoretical distribution function of χ^2_p,
$G_n(u)$ ... empirical distribution function of MD_i^2.

For $\eta = \chi^2_{p,1-\alpha}$ define

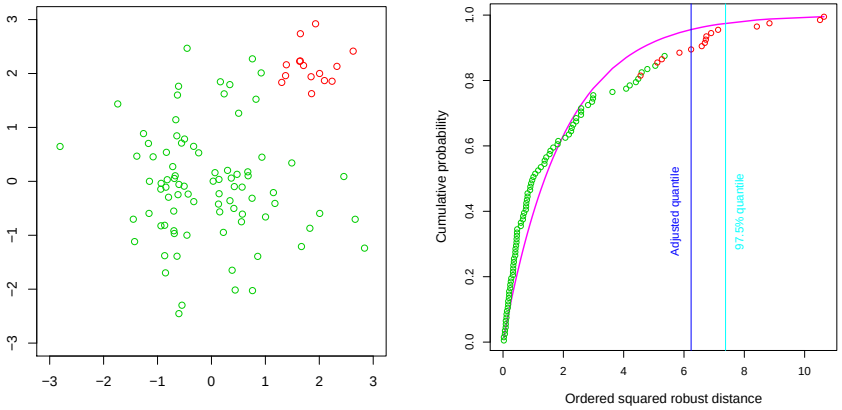
$$p_n(\eta) = \sup_{u \geq \eta} \{G(u) - G_n(u)\}^+.$$

Then a measure of *outliers* in the sample is

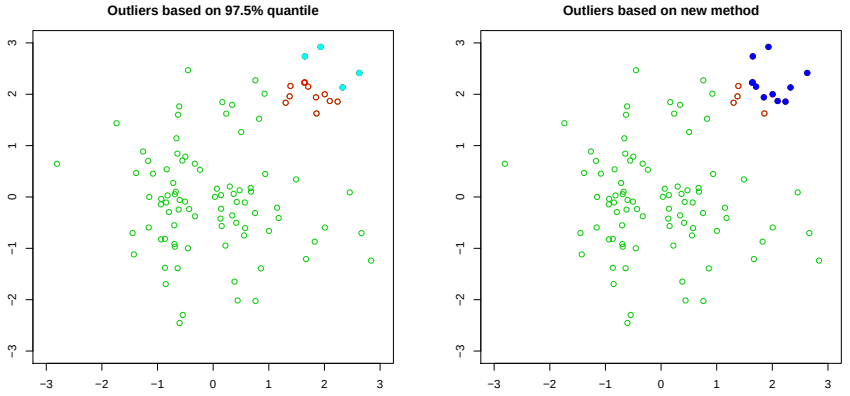
$$\alpha_n(\eta) = \begin{cases} 0 & \text{if } \; p_n(\eta) \leq p_{crit}(\eta, n, p) \\ p_n(\eta) & \text{if } \; p_n(\eta) > p_{crit}(\eta, n, p). \end{cases}$$

$p_{crit}(\eta, n, p)$ can be obtained by simulations.

Simulated Data Example



Example: Simulated Data



Example: Kola Data

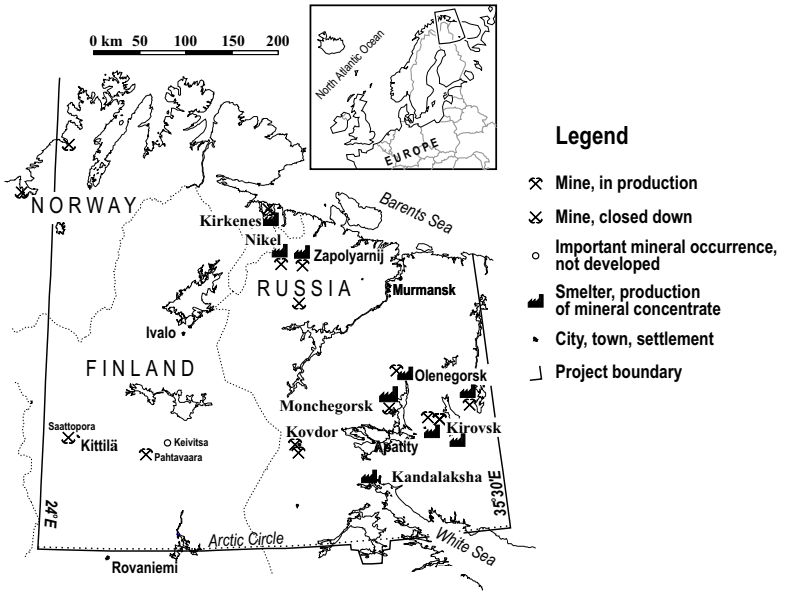
Consider the O-horizon (organic surface soil) of the Kola data set.

Take (more or less) typical elements for "pollution":

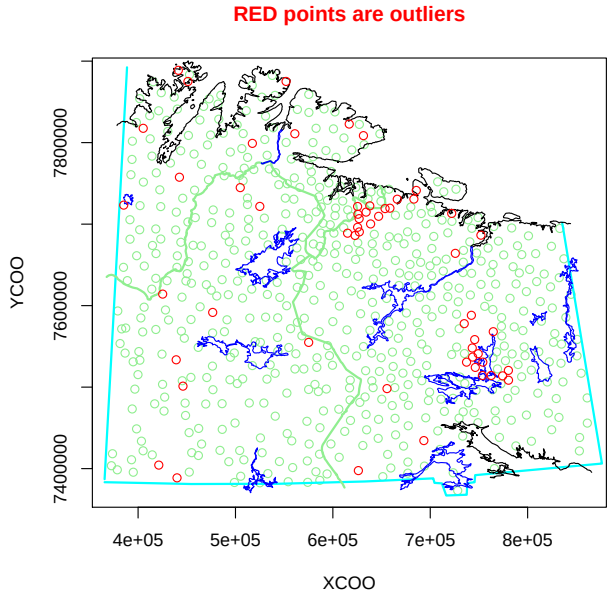
As, Cd, Co, Cu, Mg, Pb, Zn

Question: *Where are the multivariate outliers?*

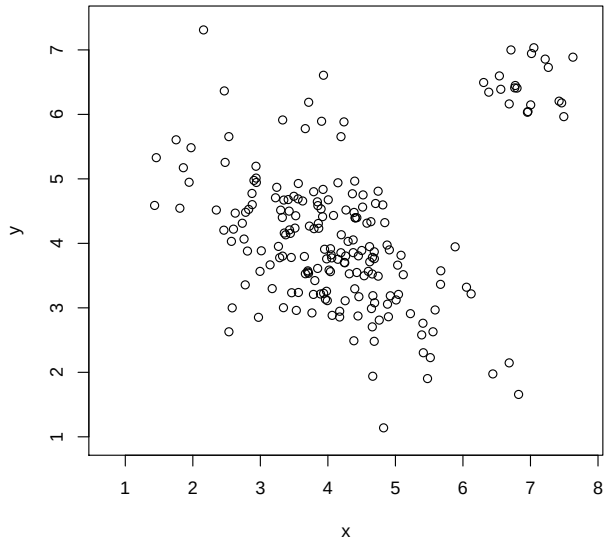
KOLA-Project



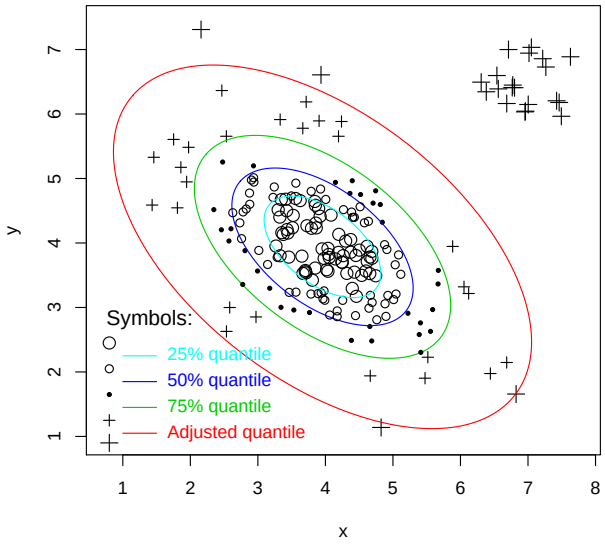
Example: Map showing outliers



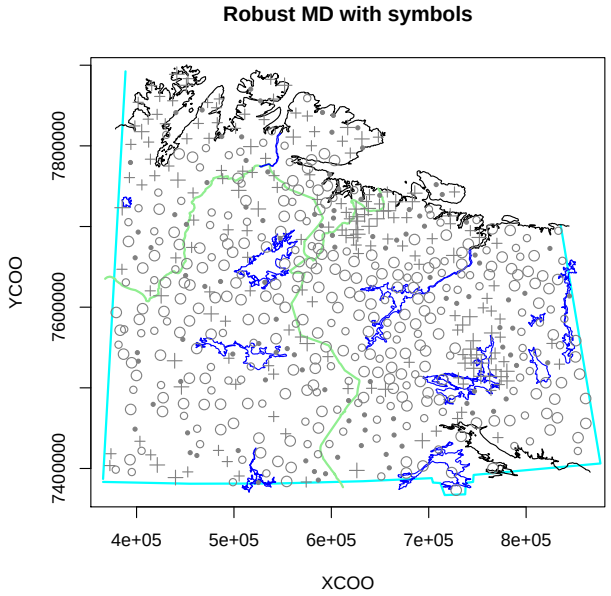
Choice of Symbols



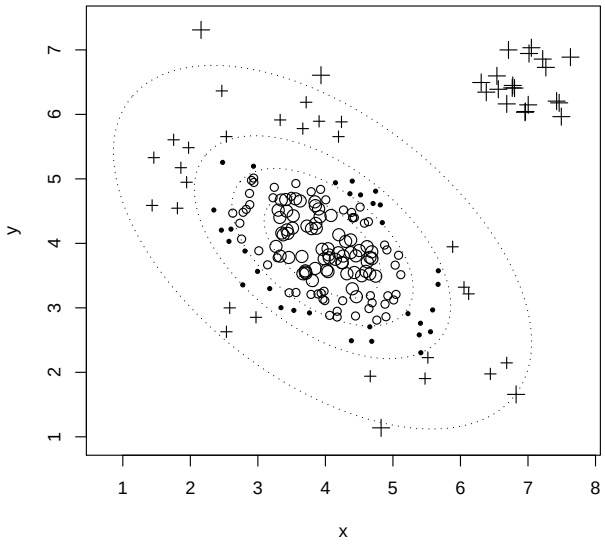
Choice of Symbols



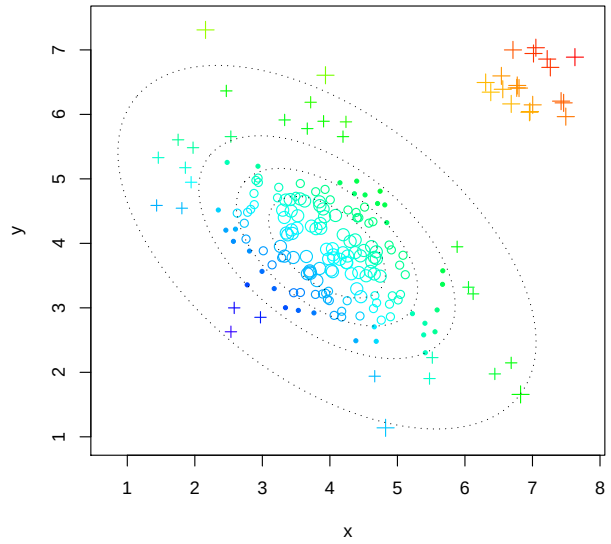
Example: Map showing outliers



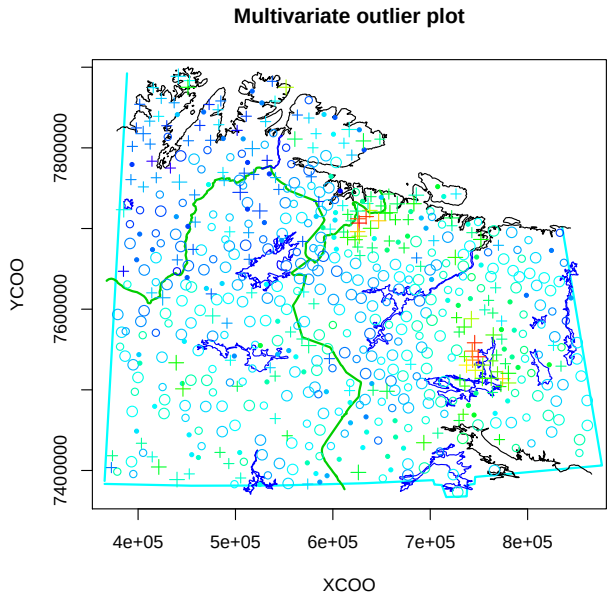
Which Outliers?



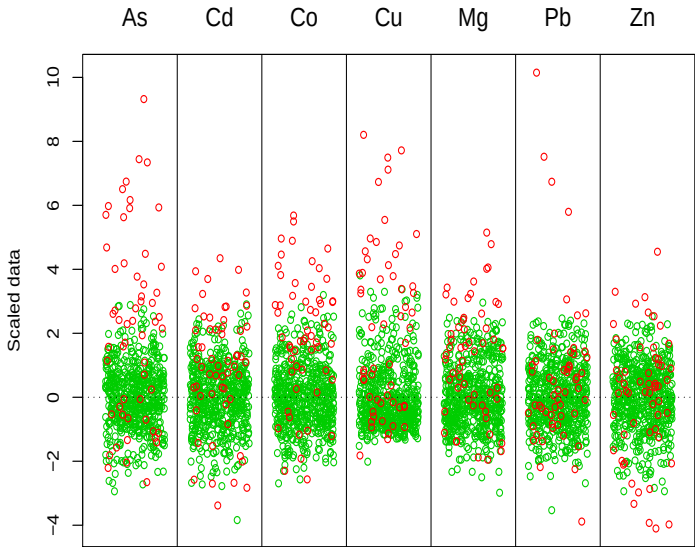
Which Outliers?



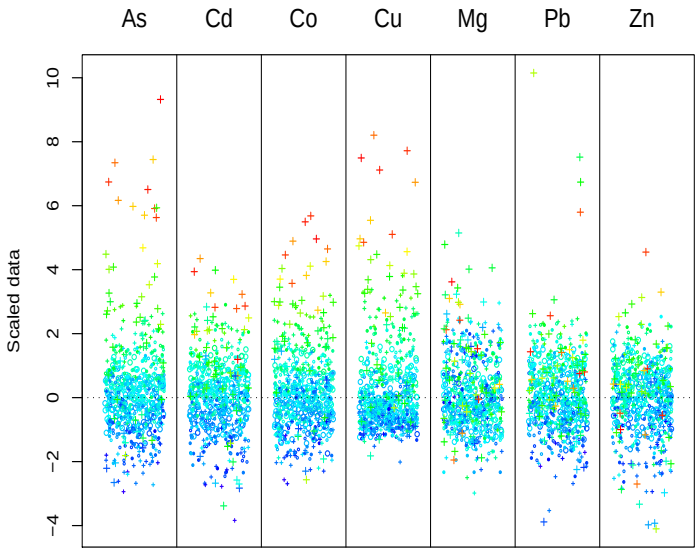
Example: Map showing outliers



Example: From Multivariate to Univariate



Example: Symbols from multivariate plot



Summary

`library(mvoutlier)`

includes

- all routines to generate the presented plots
- Kola data and other interesting geochemical data sets