Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Finding Multivariate Outlier

Applied Multivariate Statistics – Spring 2012

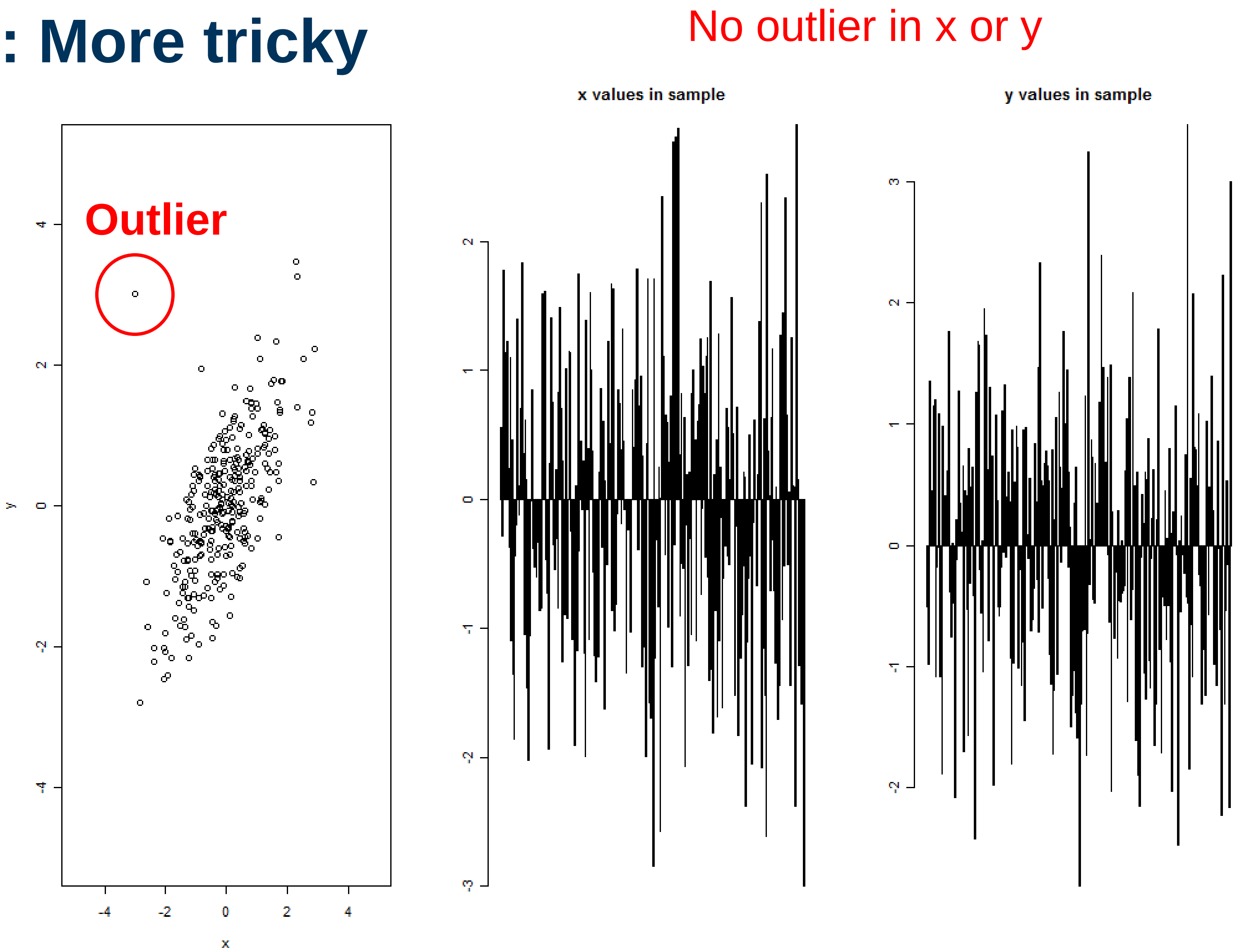
Goals

- Concept: Detecting outliers with (robustly) estimated Mahalanobis distance and QQ-plot
- R: chisq.plot, pcout from package "mvoutlier"

Outlier in one dimension - easy

- Look at scatterplots
- Find dimensions of outliers
- Find extreme samples just in these dimensions
- Remove outlier

2d: More tricky



Recap: Mahalanobis distance

- True Mahalanobis distance:

$$MD(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

- Estimated Mahalanobis distance:

$$\hat{M}D(x) = \sqrt{(x-\hat{\mu})^T \hat{\Sigma}^{-1} (x-\hat{\mu})}$$

Sq. Mahalanobis Distance $MD^2(x)$

=

Sq. distance from mean in

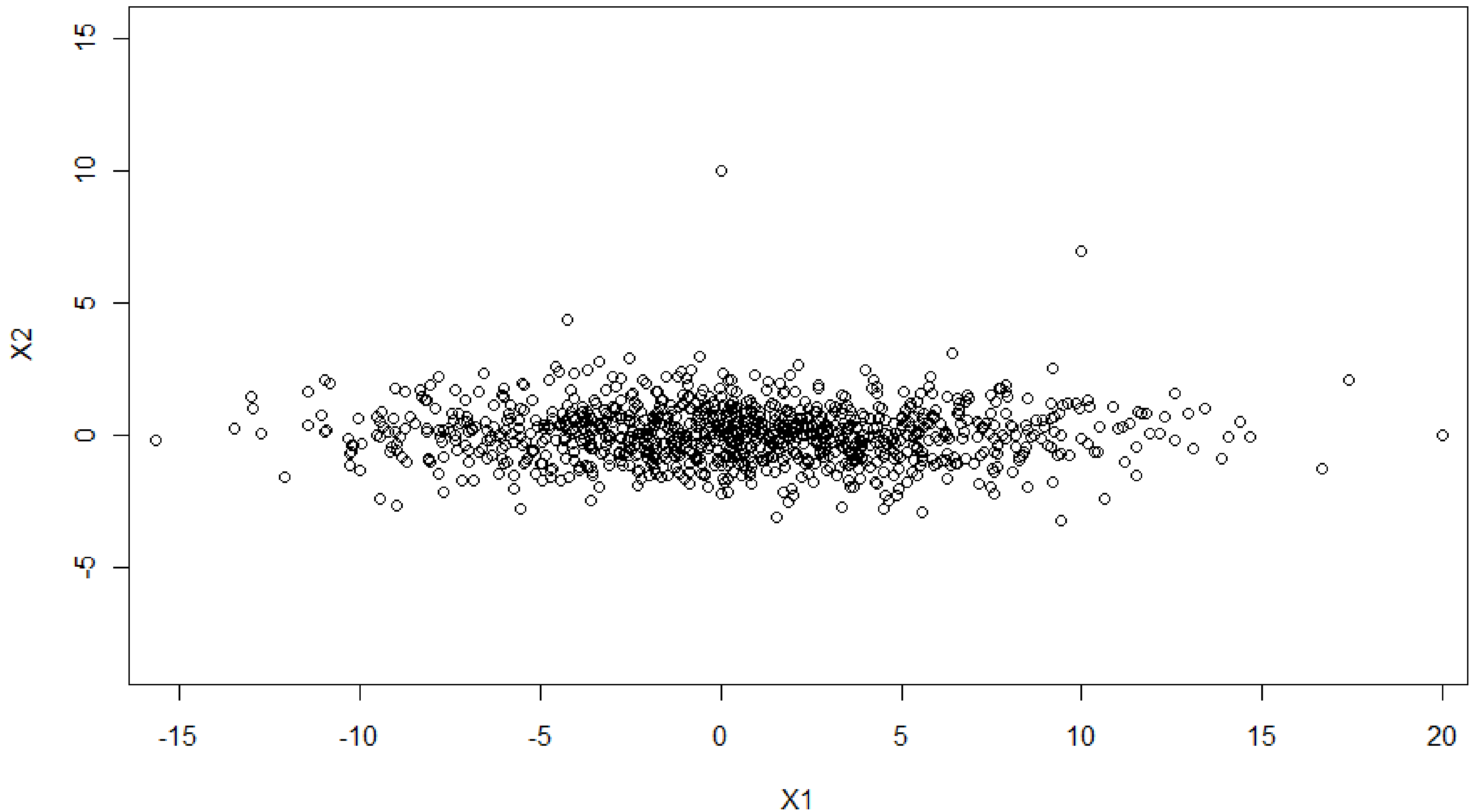
standard deviations

IN DIRECTION OF X

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

Mahalanobis distance: Example

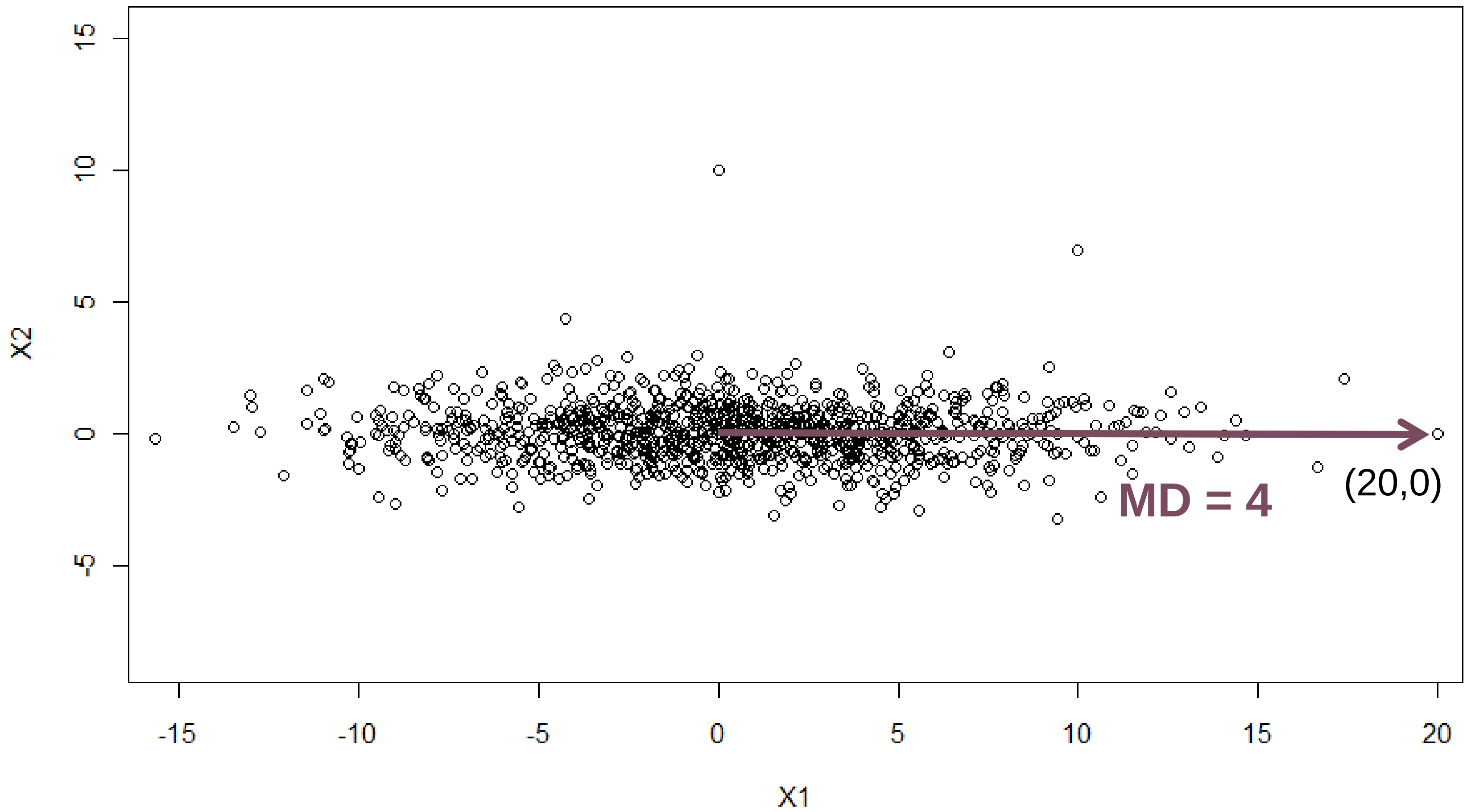
$$\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$



Mahalanobis distance: Example

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

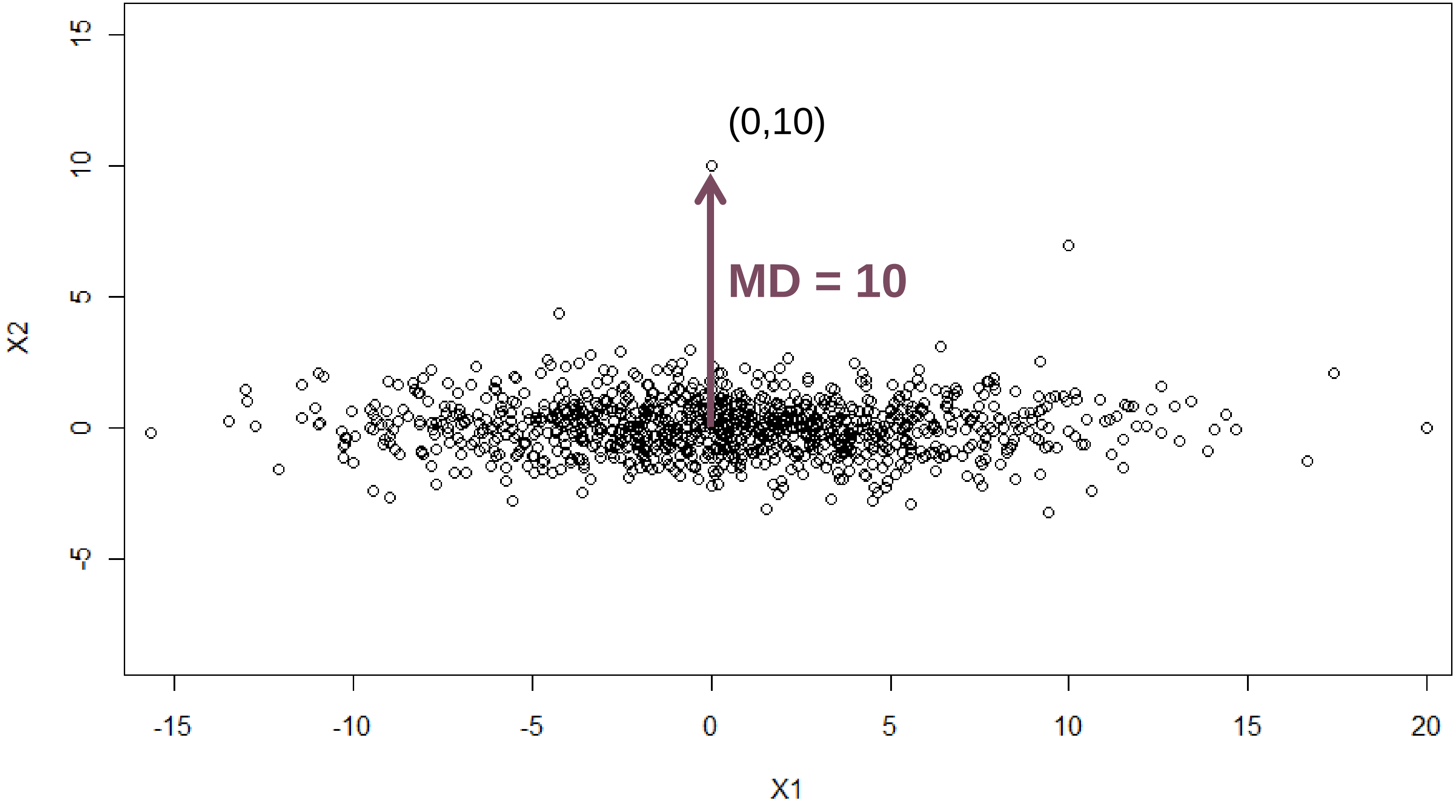
$$\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

Mahalanobis distance: Example

$$\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$

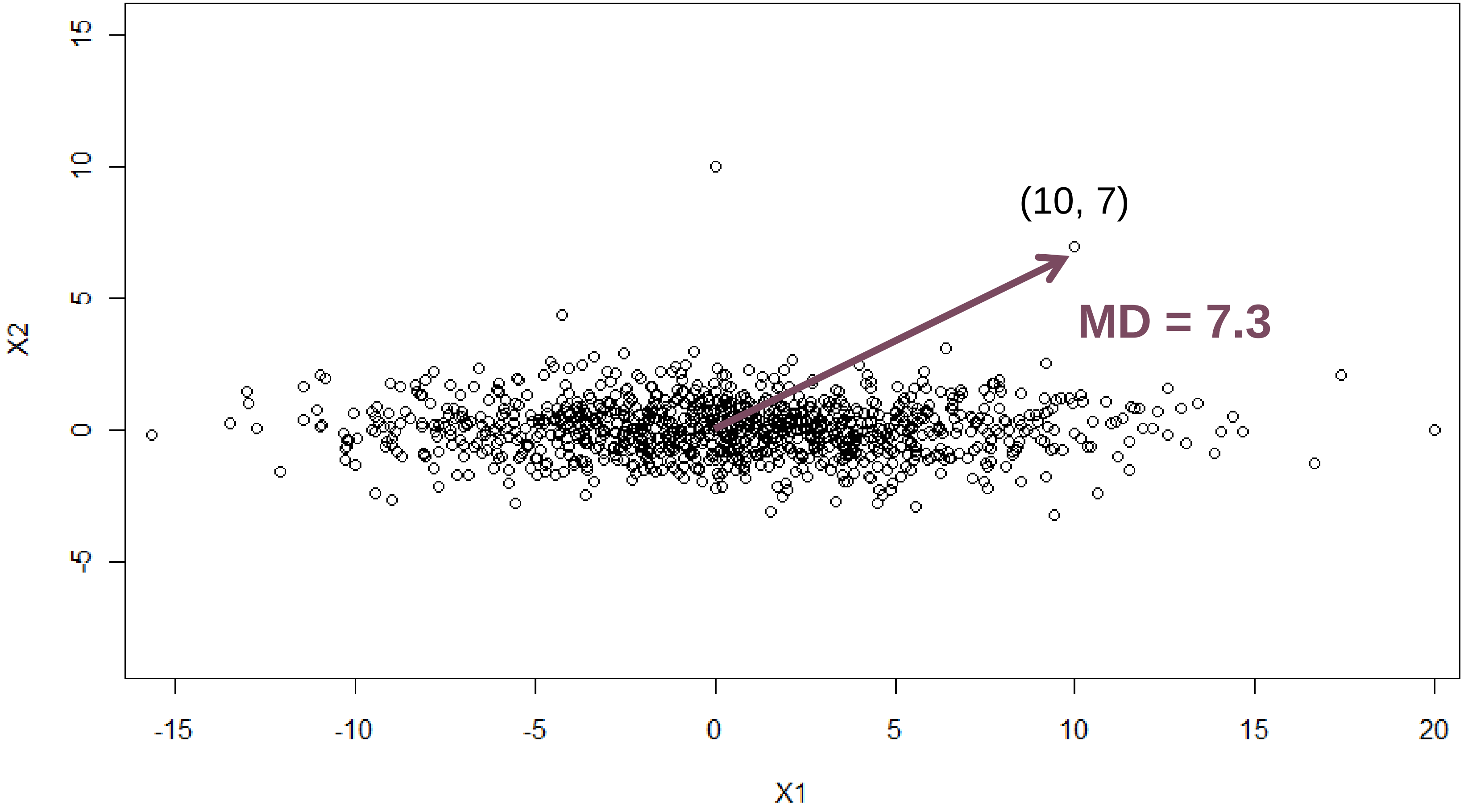


Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Mahalanobis distance: Example

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$



Theory of Mahalanobis Distance

Assume data is multivariate normally distributed
(d dimensions)

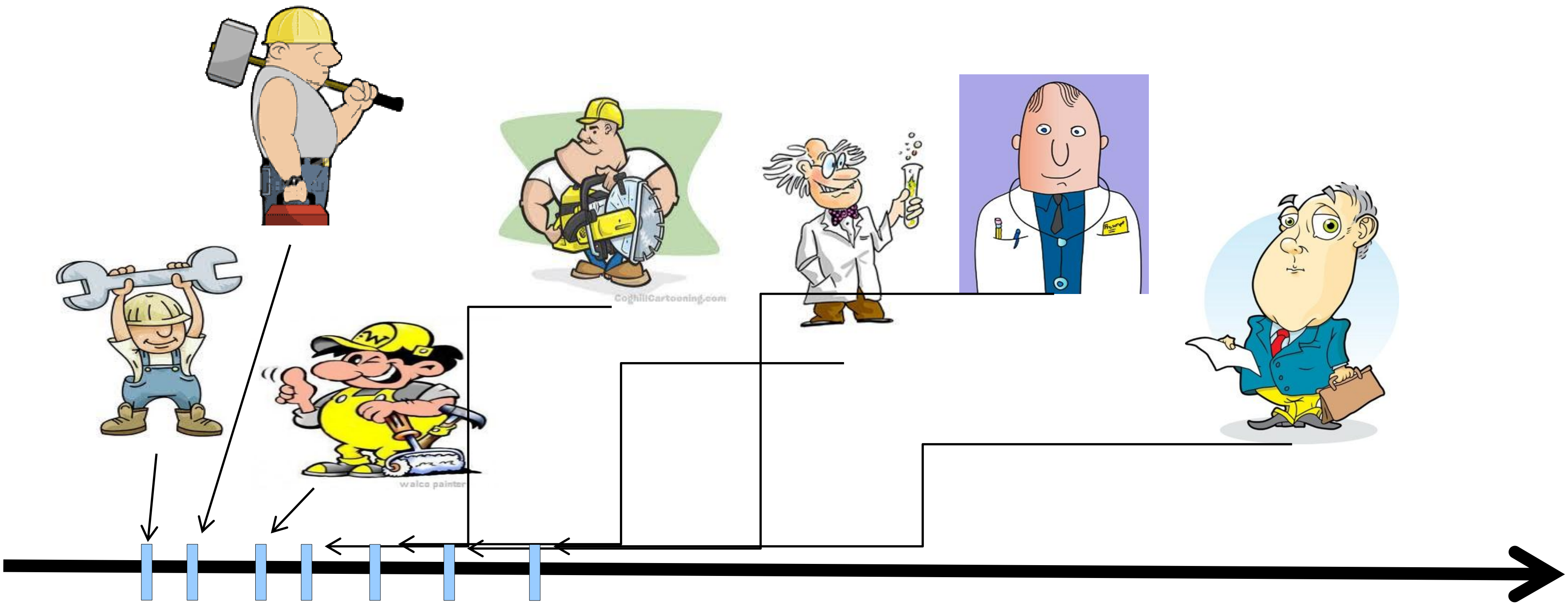
Mahalanobis distance of samples follows a Chi-Square distribution
with d degrees of freedom

("By definition": Sum of d standard normal random variables has
Chi-Square distribution with d degrees of freedom.)

Check for multivariate outlier

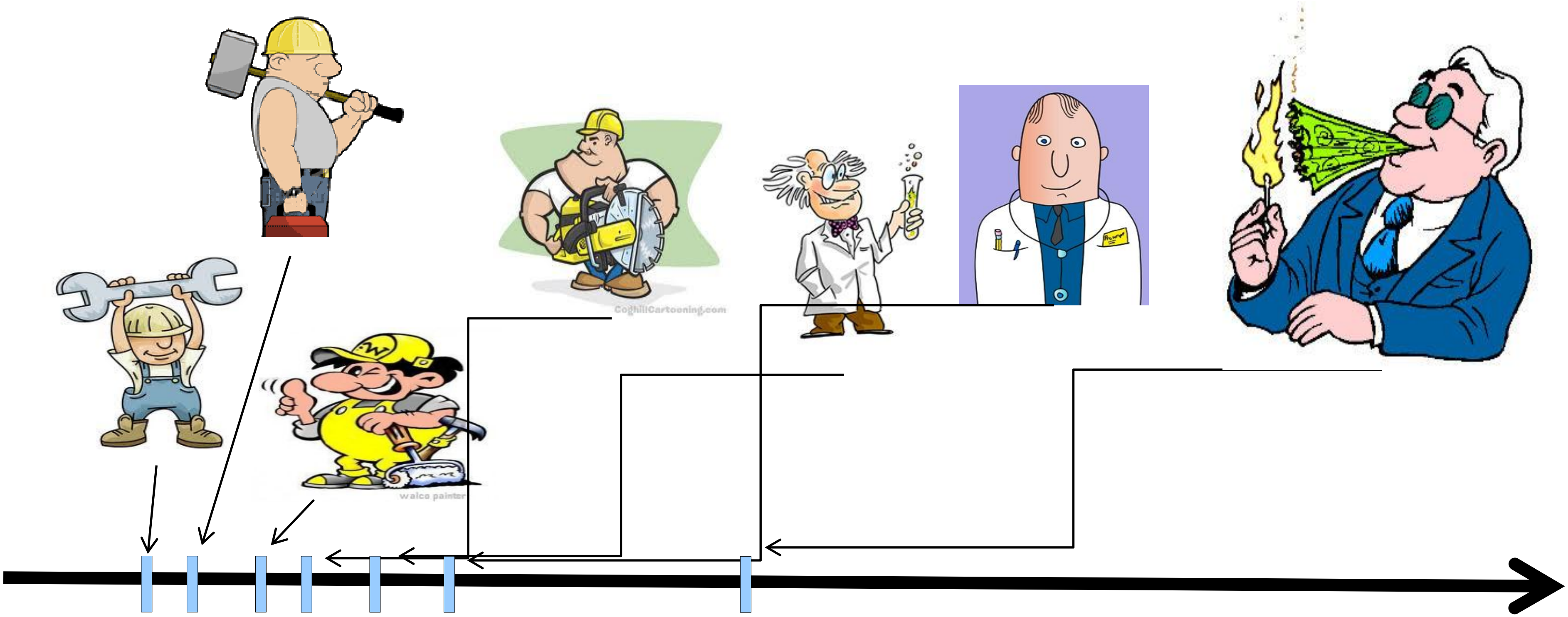
- Are there samples with estimated Mahalanobis distance that don't fit at all to a Chi-Square distribution?
- Check with a QQ-Plot
- Technical details:
  - Chi-Square distribution is still reasonably good for estimated Mahalanobis distance
  - use robust estimates for μ, Σ

Robust Estimates: Income of 7 people



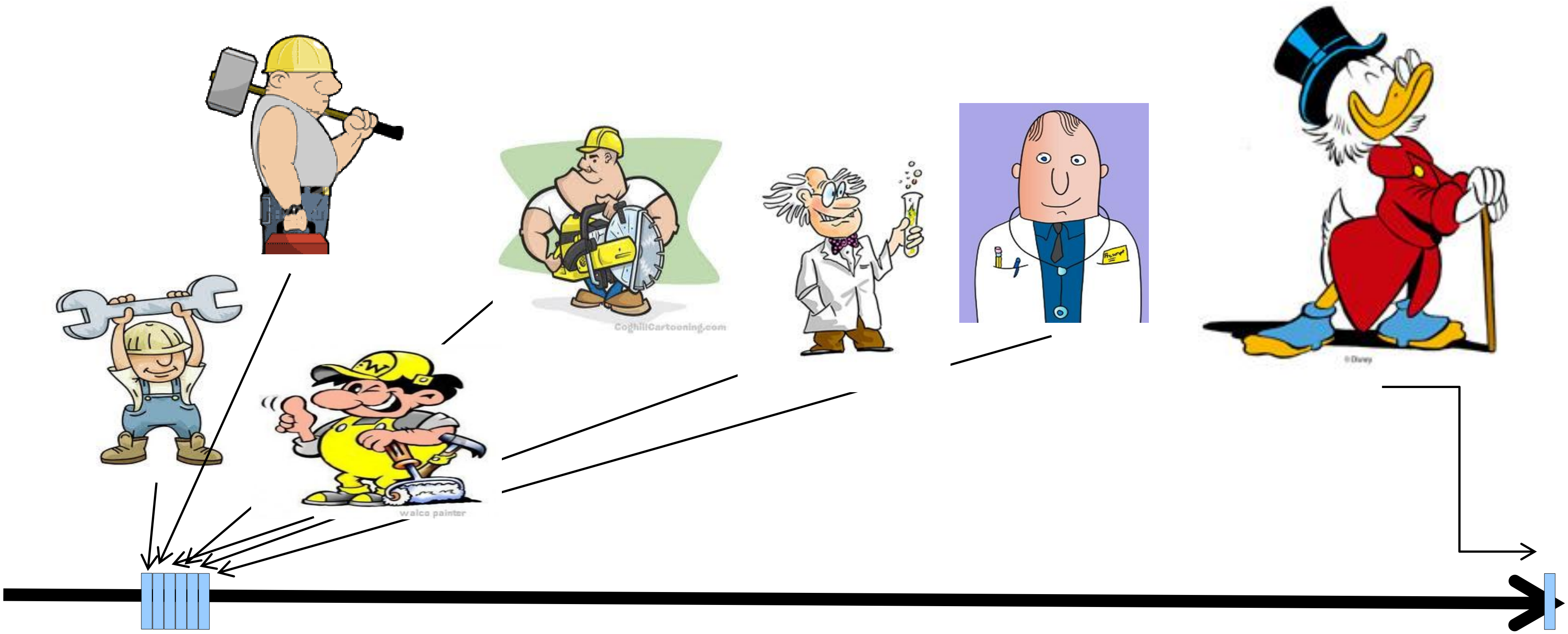
Robust Scatter

Std. Dev.



Robust

Std. Dev.



- Robust

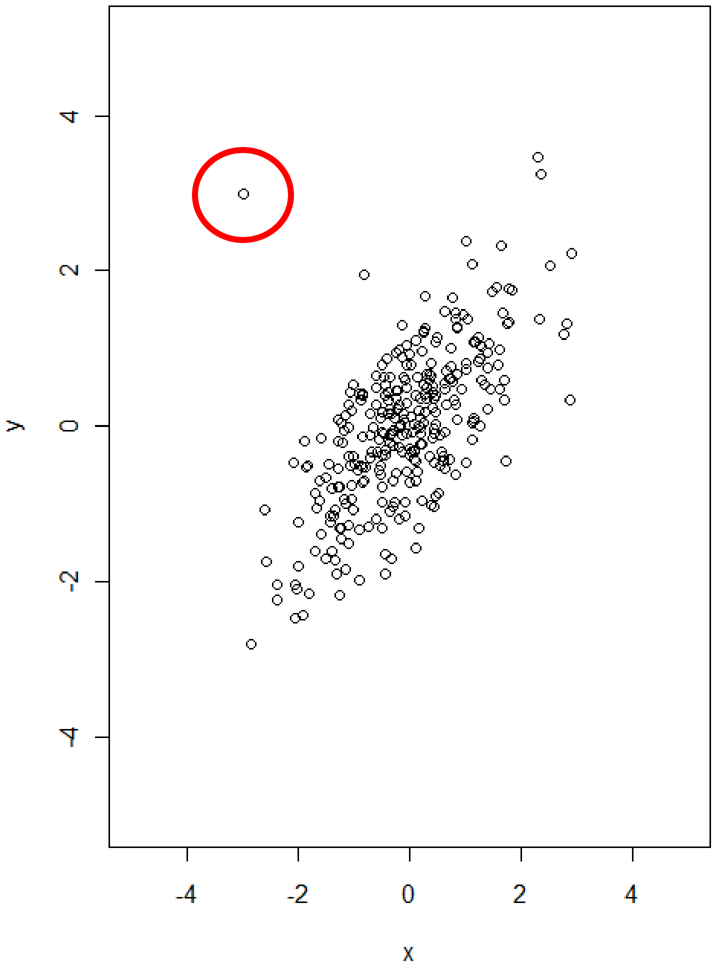
Std. Dev.

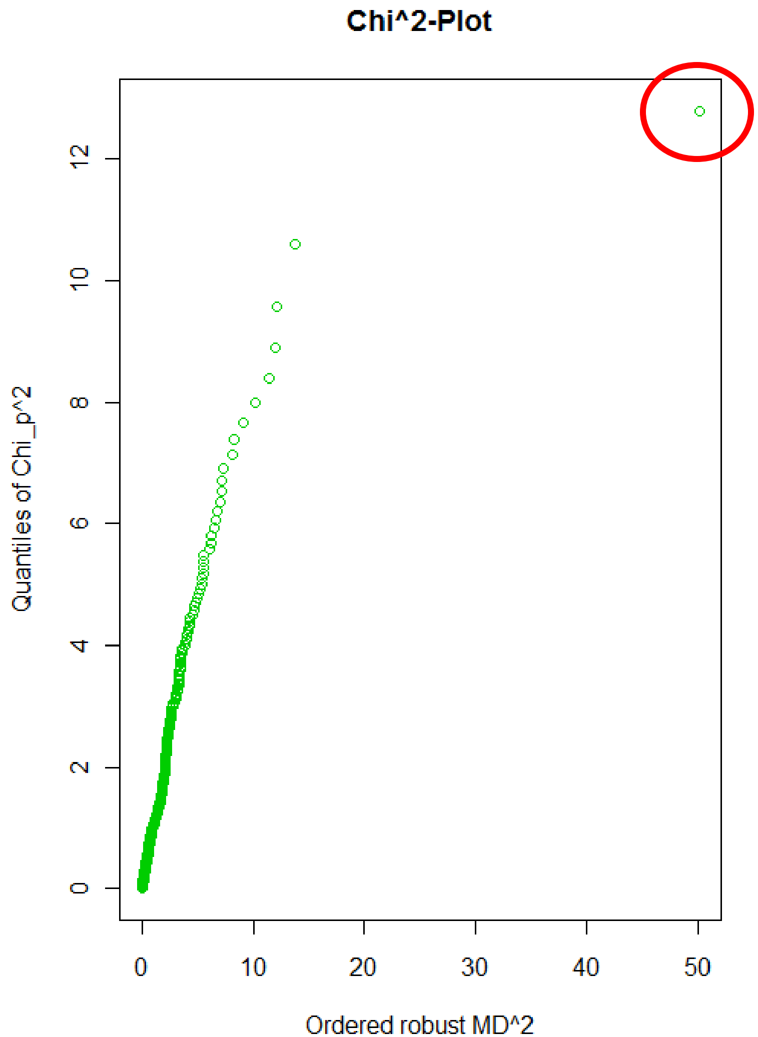
Robust Estimates for outlier detection

- If scatter is estimated robustly, outlier “stick out” much more
- Robust Mahalanobis distance:
  Mean and Covariance matrix estiamted robustly

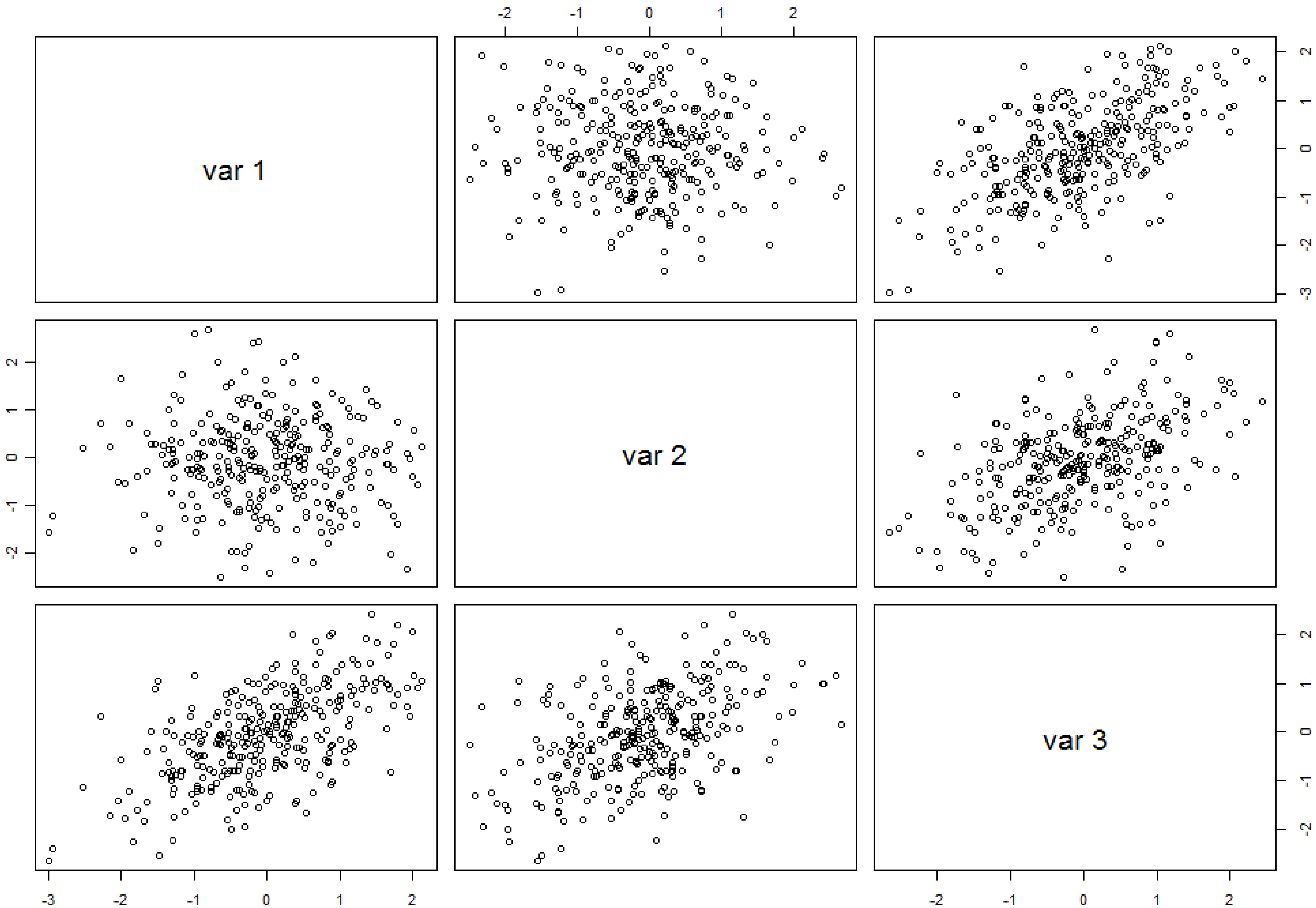
Outlier easily detected !

Example - continued



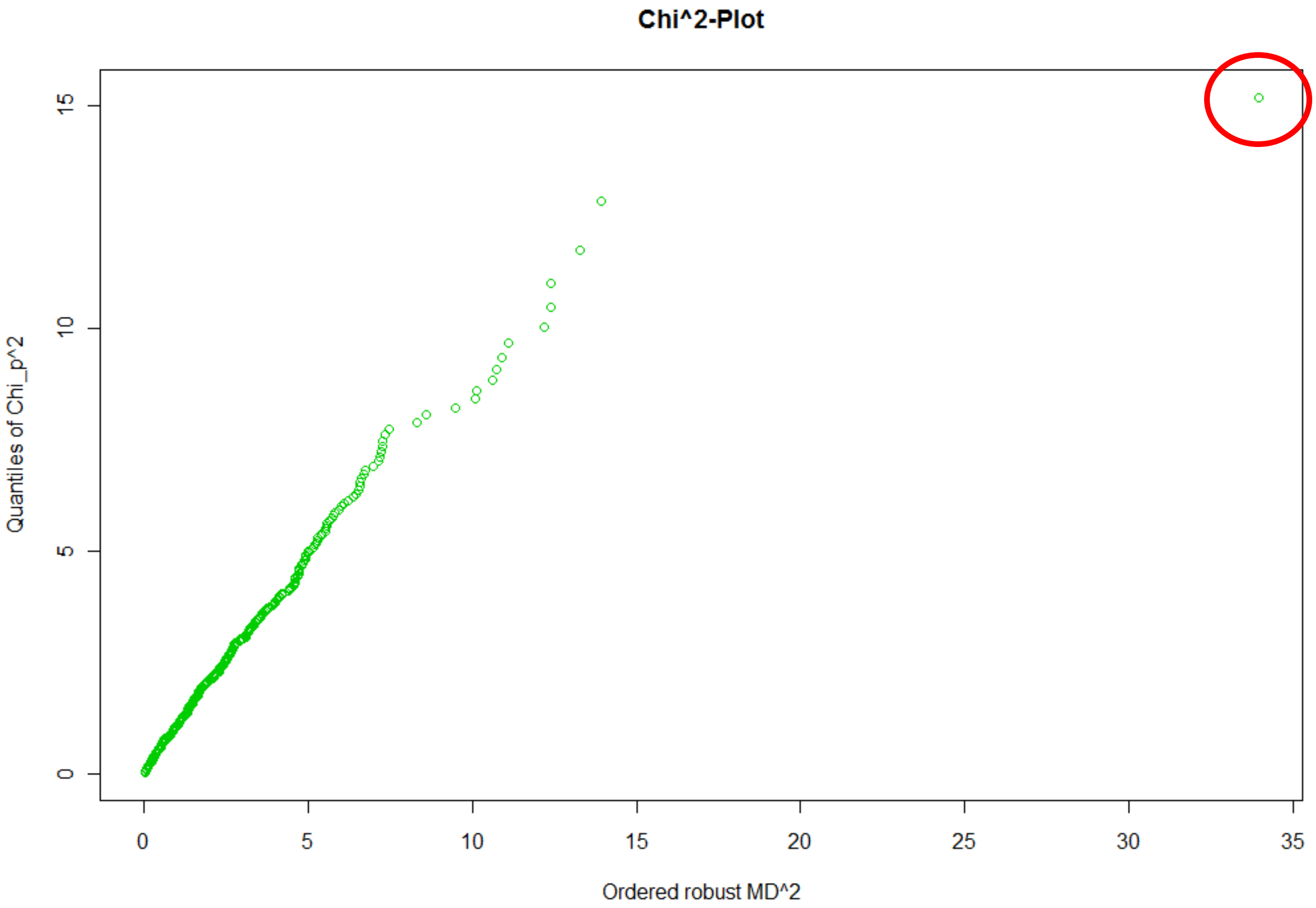


Outliers in >2d can be well hidden !



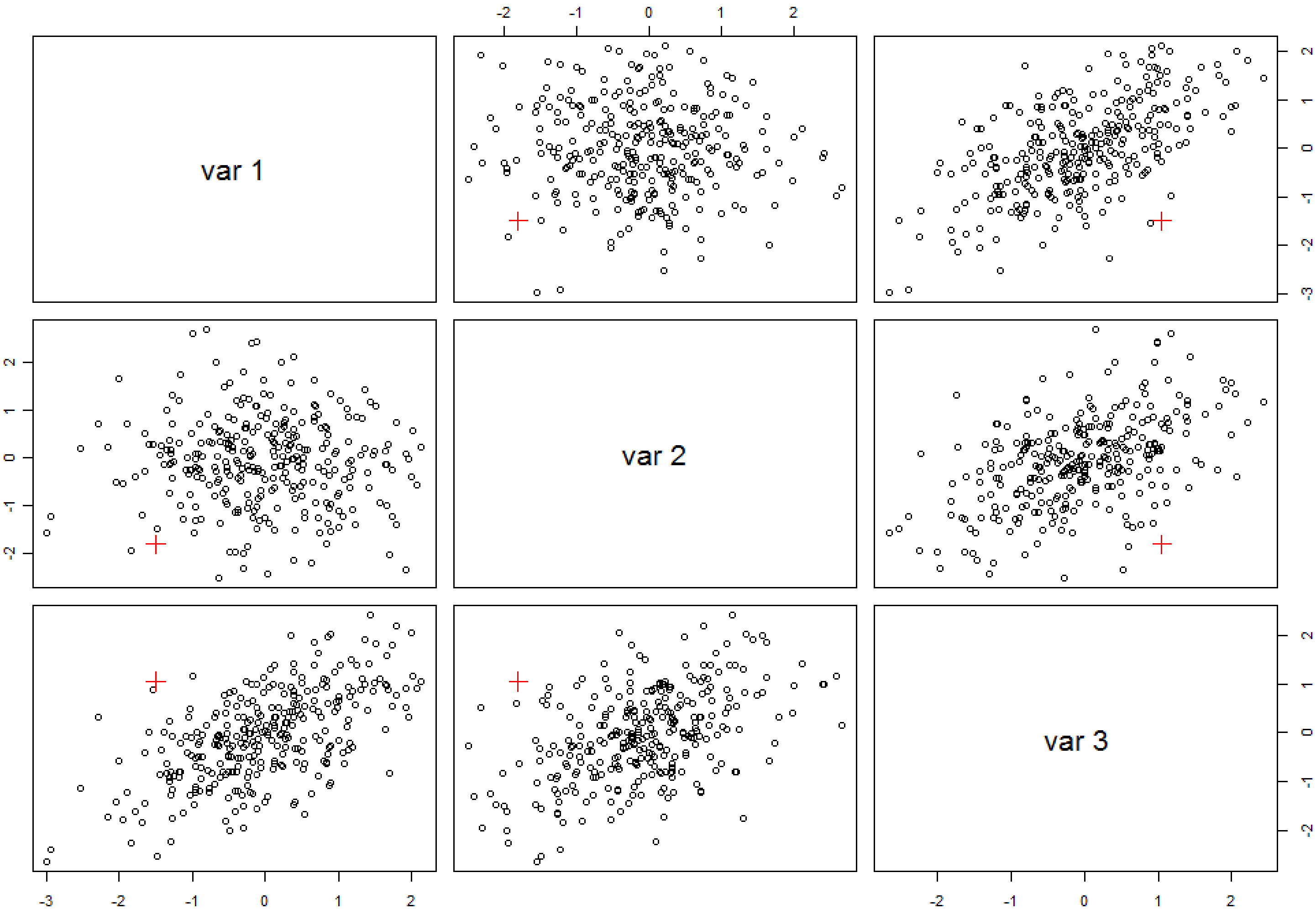
No outlier, right?

Outliers in >2d can be well hidden !



Wrong!

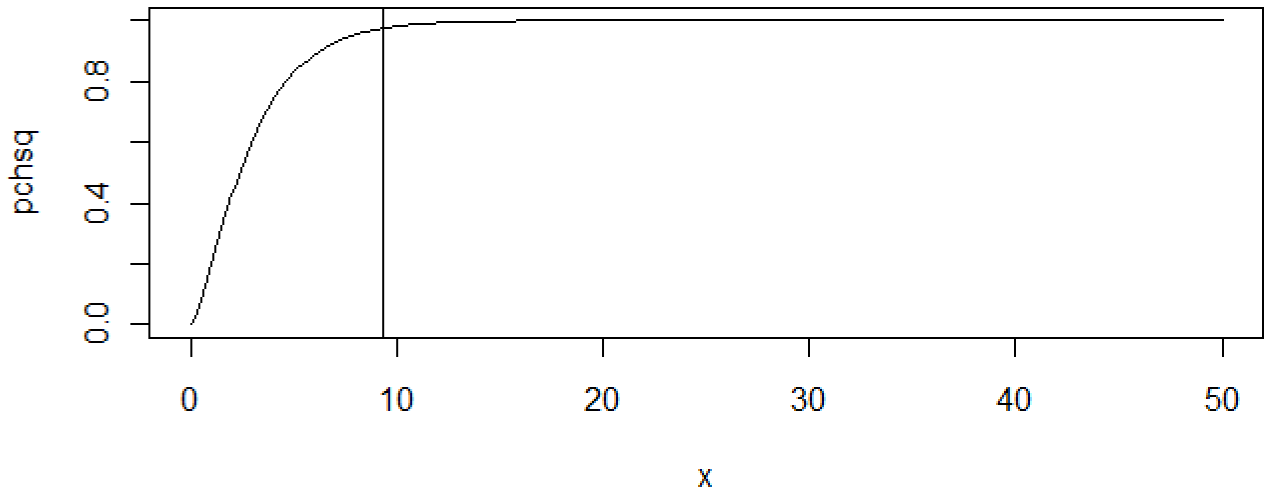
Outliers in >2d can be well hidden !



This outlier can't be seen in the scatterplot-matrix (but in a 3d plot)

Method 1: Quantile of Chi-Sqaure distribution

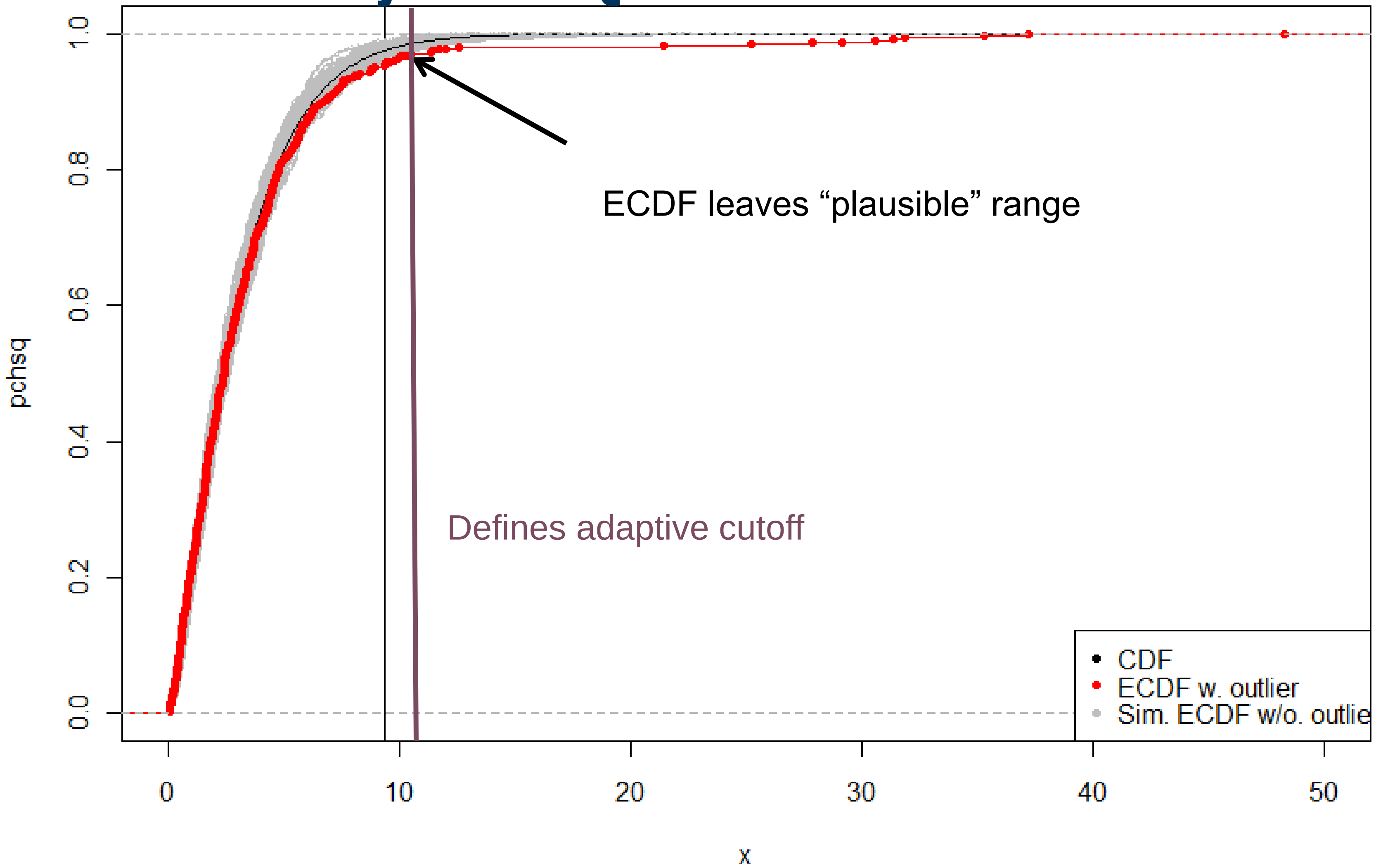
- Compute for each sample (in d dimensions) the robustly estimated Mahalanobis distance $MD(x_i)$
- Compute the 97.5%-Quantile Q of the Chi-Square distribution with d degrees of freedom
- All samples with $MD(x_i) > Q$ are declared outlier



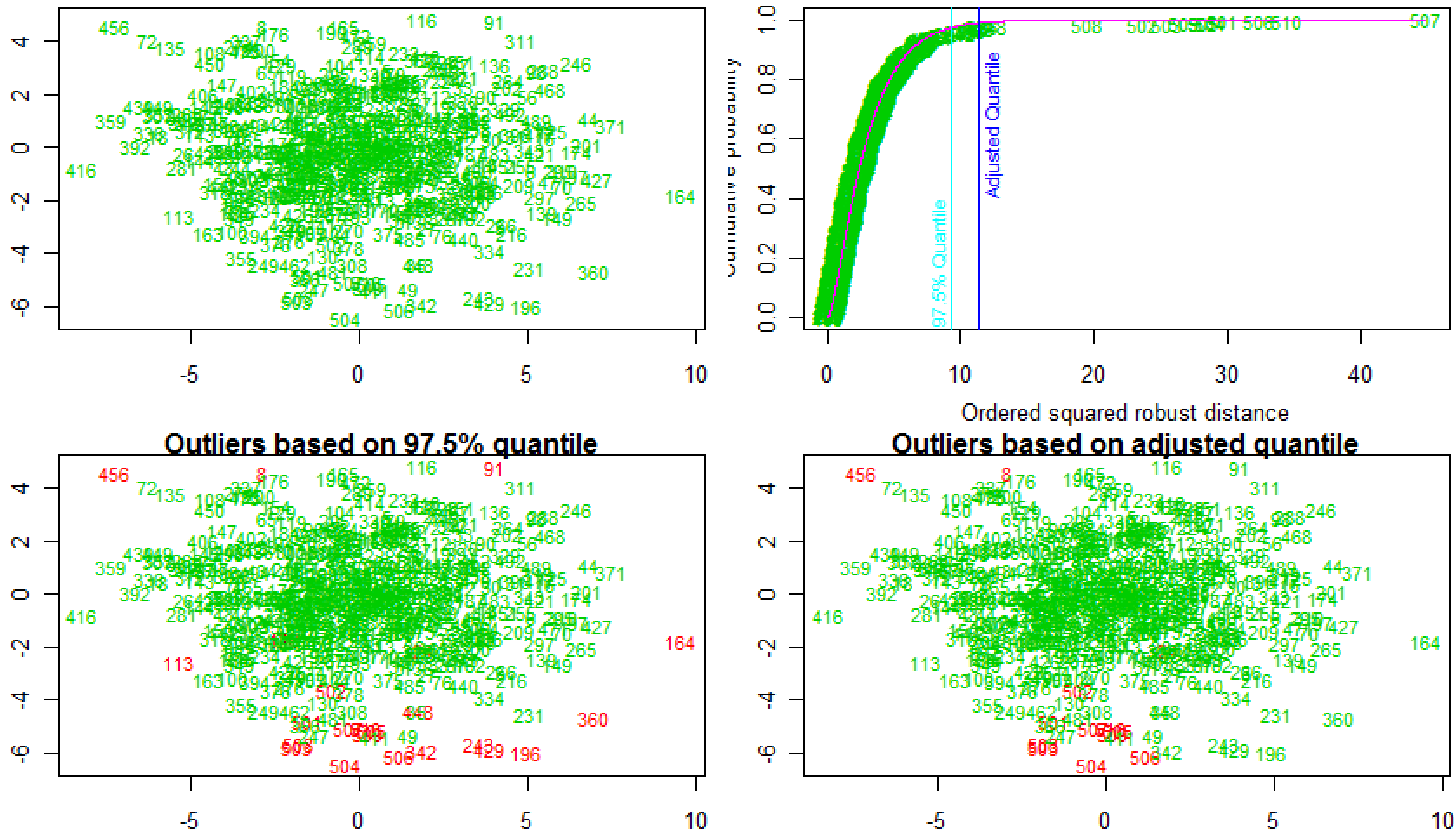
Method 2: Adjusted Quantile

- Adjusted Quantile for outlier: Depends on distance between cdf of Chi-Square and ecdf of samples in tails
- Simulate “normal” deviations in the tails
- Outlier have “abnormally large” deviations in the tails (e.g. more than seen in 100 simulations without outliers)

Method 2: Adjusted Quantile



Method 2: Adjusted Quantile Function "aq.plot"



Method 3: State of the art - pcout

- Complex method based on robust principal components
- Pretty involved methodology
- Very fast – good for high dimensions

- R: Function "pcout" in package "mvoutlier"
- $wfinal01: 0 is outlier
- $wfinal: Small values are more severe outlier

- P. Filzmoser, R. Maronna, M. Werner. Outlier identification in high dimensions, *Computational Statistics and Data Analysis*, 52, 1694-1711, 2008

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Automatic outlier detection

- It is ***always better*** to look at a QQ-plot to find outlier !
  Just find points “sticking out”; no distributional assumption
- If you can’t: Automatic outlier detection
  - finds usually too many or too few outlier depending on parameter settings
  - depends on distribution assumptions
  (e.g. multivariate normality)
  + good for screening of large amounts of data

Concepts to know

- Find multivariate outlier with robustly estimated Mahalanobis distance
- Cutoff
  - by eye (best method)
  - quantile of Chi-Square distribution

R commands to know

- chisq.plot, pcout in package “mvoutlier”

Next week

- Missing values