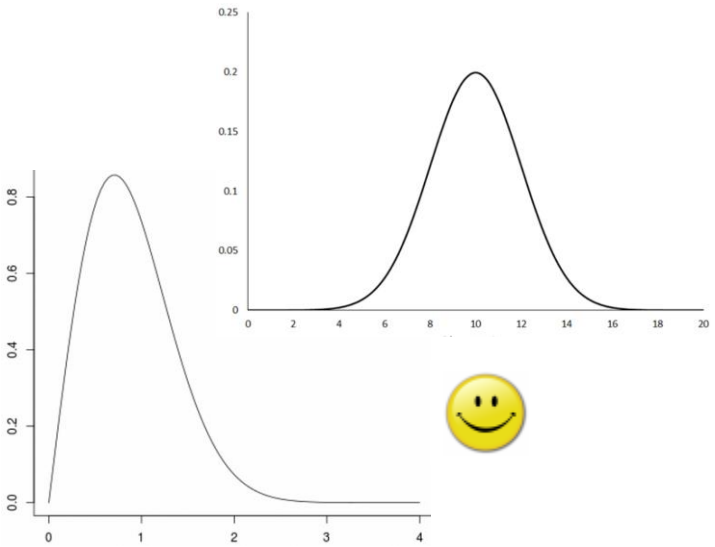
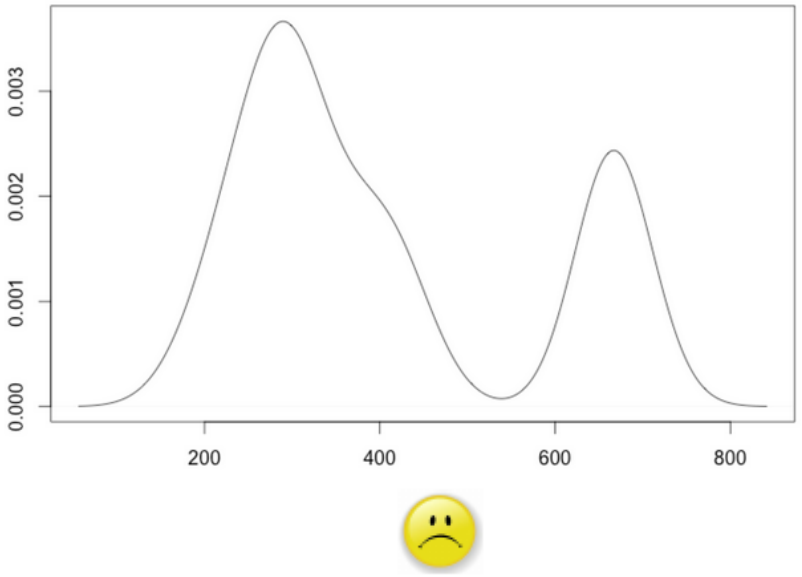
Anomaly Detection

- Motivation and Introduction
- Supervised Methods
- Semisupervised Methods
- Unsupervised Methods:
  - Graphical and Statistical approaches
  - Nearest neighbor based approaches
  - Clustering based approaches
- Evaluation

Unsupervised Methods

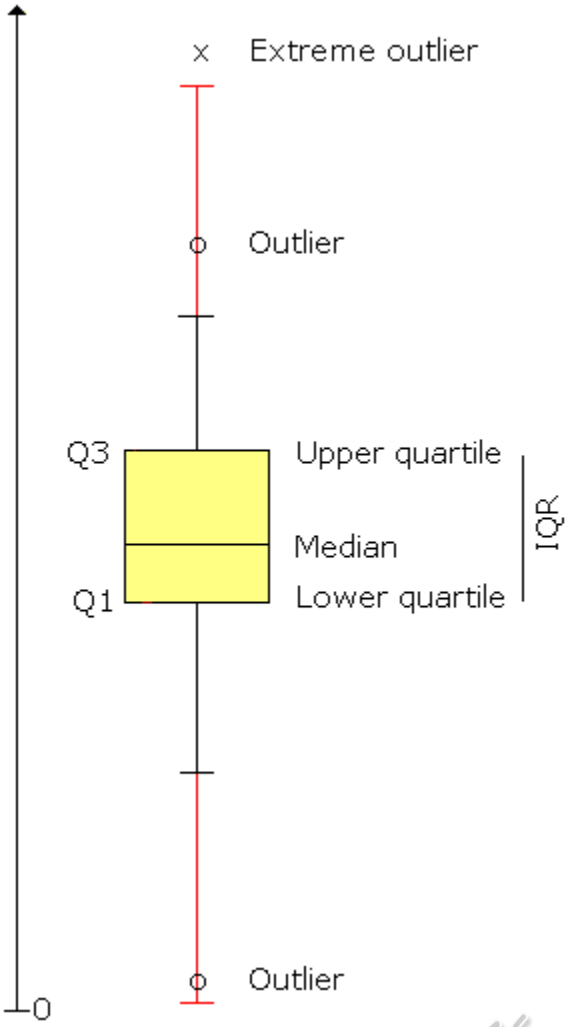
This approach is valid when working with the Normal distribution. It can also be applied to other distributions, whenever they are not very “rare” (for instance, they should have only one mode -not “camel” distributions-)





IQR = Q3 - Q1

P is an Outlier if P > Q3 + 1.5 IQR
P is an Outlier if P < Q1 - 1.5 IQR
P is an Extreme Outlier if P > Q3 + 3 IQR
P is an Extreme Outlier if P < Q1 - 3 IQR



- Assume a parametric model describing the distribution of the data (e.g., normal distribution)

- Given a database D, and a data point $x \in D$, a statistical test determines whether x is an outlier or not, at a significance level p. The test depends on:
  - Data distribution
  - Number of outliers to identify

- 1-variate Normal distribution:

  One outlier: Grubbs' test

  k-Outliers: Tietjen and Moore's test

  Less than k-Outliers: Rosner's test

- Multi-variate Normal distribution:

  Non Robust Methods: Mahalanobis distances

  Robust Methods: Mahalanobis distances with MCD and median instead of Covariance and mean.

Univariate data:

- Grubbs' test for **normal distribution** (R package: "outliers")
  - H_0: There are no outliers in data
  - H_A: There is **exactly one** outlier

Grubbs' test statistic:

$$G = \frac{\max_{i=1..N} \left| X_i - \overline{X} \right|}{S}$$

The outlier itself participates in computing the mean and the standard deviation *S*

The Grubbs' test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation. This is the two sided version. There are also one sided versions for the maximum (minimum).

Reject H_0 if:

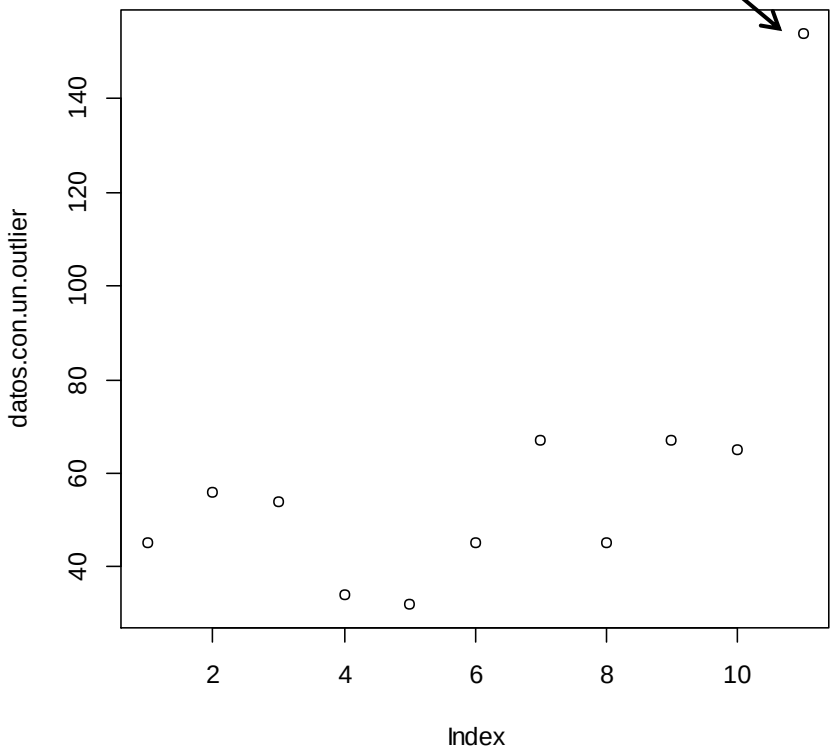
$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/(2N),N-2)}}{N-2+t^2_{(\alpha/(2N),N-2)}}}$$

http://www.graphpad.com/quickcalcs/Grubbs1.cfm

Grubbs' test for

- H_0: There are no outliers in data
- H_A: There is exactly one outlier

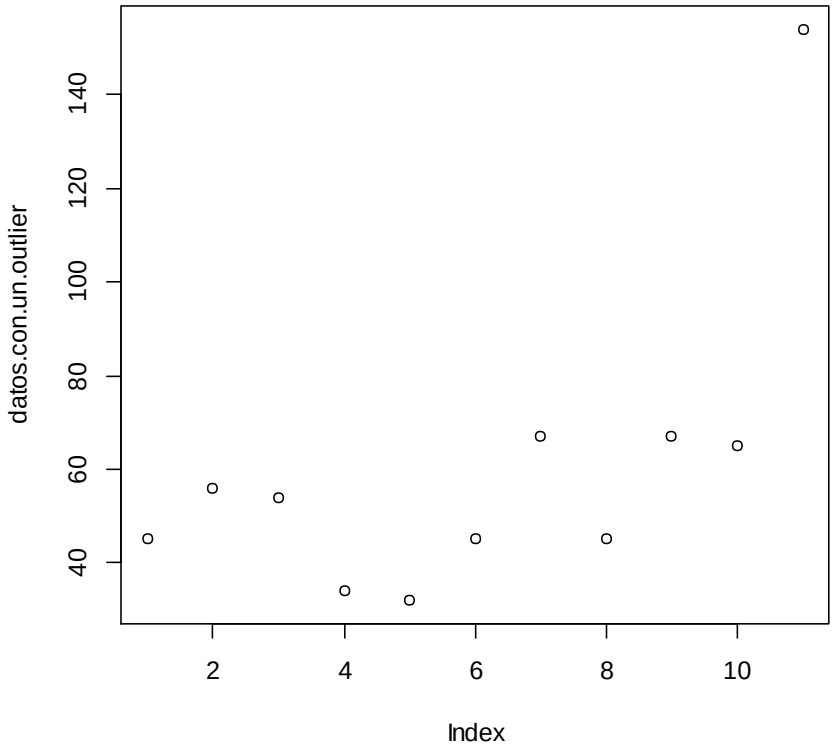
Grubbs' test: 0.00036
→ There's one outlier
☺



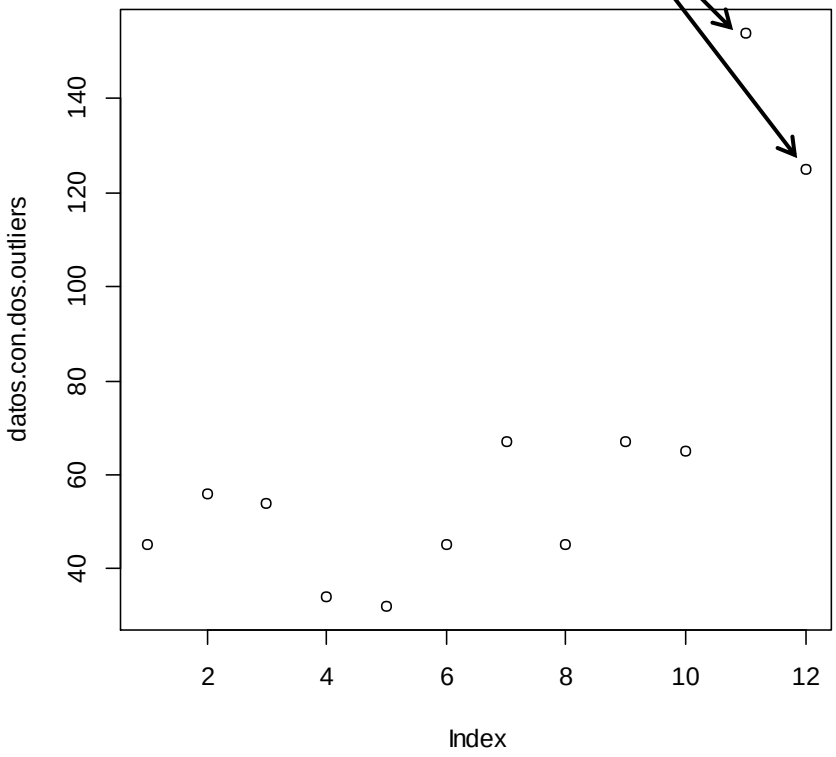
- Grubbs' test is for a ***single*** outlier in 1-variate Normal distribution (there are other proposals as Dixon's test)
- What to do if there are ***multiple outliers***, i.e, when there's more than a single outlier?
- Detect one outlier at a time by applying a statistical test, remove the outlier, and repeat the test. Problem: ***Masking***
- Masking appears when a test for a single outlier fails to detect it because there is another outlier which *hides* the first one.

Masking

These points skews the mean and S toward them, i.e, they become too large: So, these points are not too far from the mean.



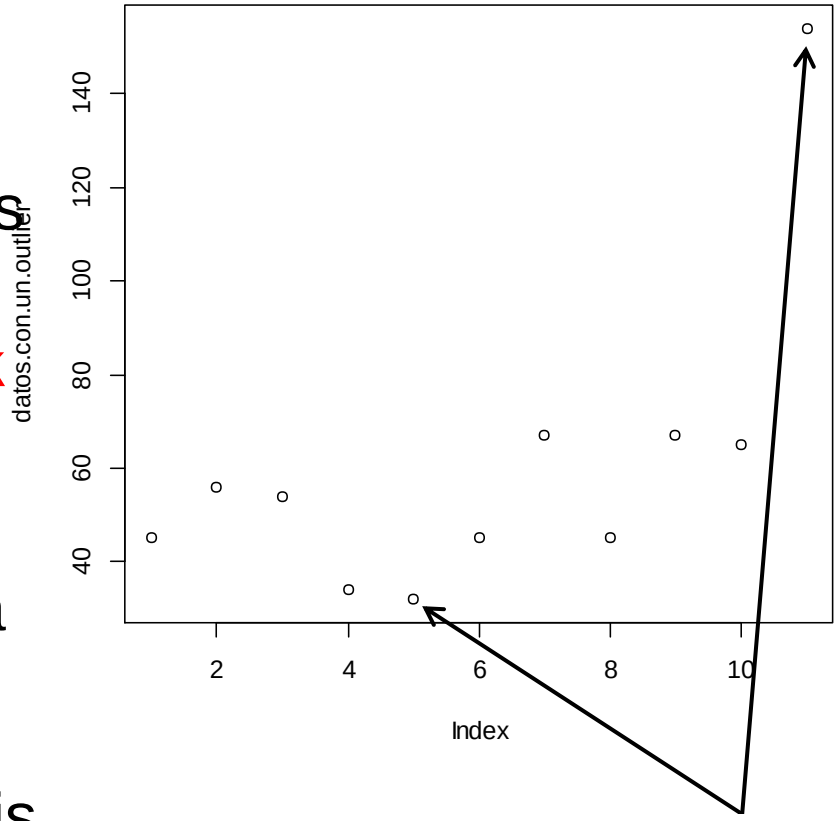
Grubbs' test: 0.00036
→ There's one outlier ☺



Grubb's test: 0.056 → Grubbs' test fail to detect any outlier, when in fact there are two ☹

- Another possibility: Apply a test to check if there are **k** outliers.

- Tietjen and Moore's test:
  - H_0: There are no outliers in the data
  - H_A: There are **exactly k** outliers
- Problem: ***Swamping***.
- Swamping appears when a test for k outliers declares there are k outliers when in fact the number of outliers is lower.

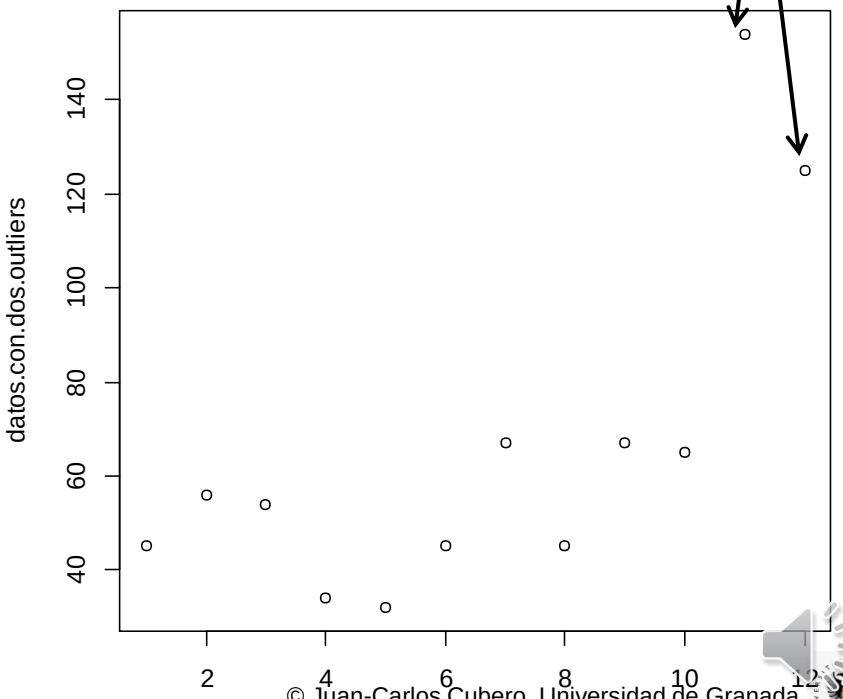


If we apply the test for k=2, it is significative because the most extreme value has an extreme statistic value 😟

- Another possibility: Apply a test to check if there are **less than k** outliers.

- Rosner's test (R package: "EnvStats"):
  - H_0: There are no outliers in the data
  - H_A: There are **less than k** outliers

Rosner's test with k=3, for instance, declares two outliers ☺

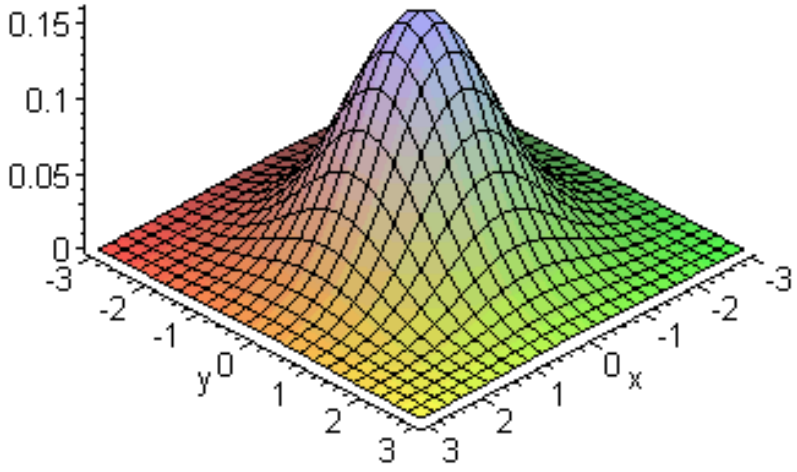


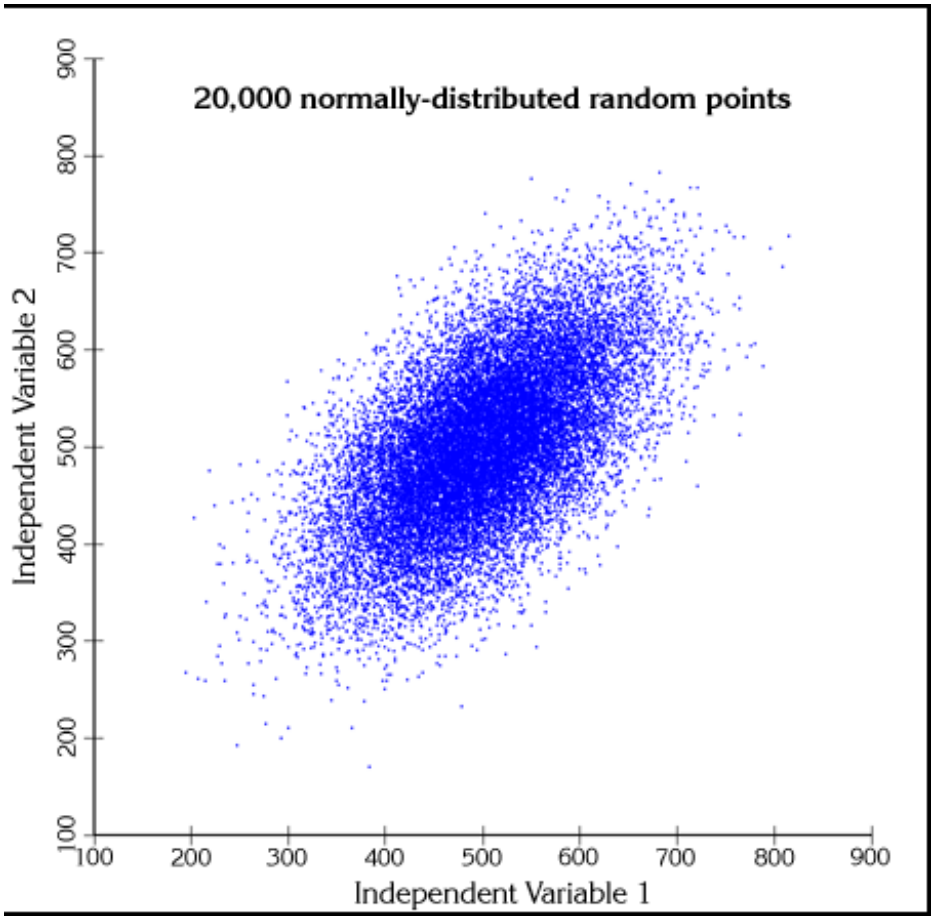
This test applies Grubbs' test sequentially, removing one outlier in each step. So, it performs k tests. In order to control FWER, a step by step correction to the significance level is applied (this is the same idea than Hochberg's correction in multiple comparisons tests)

- Working with several *(p)* dimensions

$$N(x) = \frac{1}{\sqrt{(2\pi)^p \, |\Sigma|}} e^{-\frac{(x-\mu)^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (x-\mu)}{2}}$$

Bivariate Normal



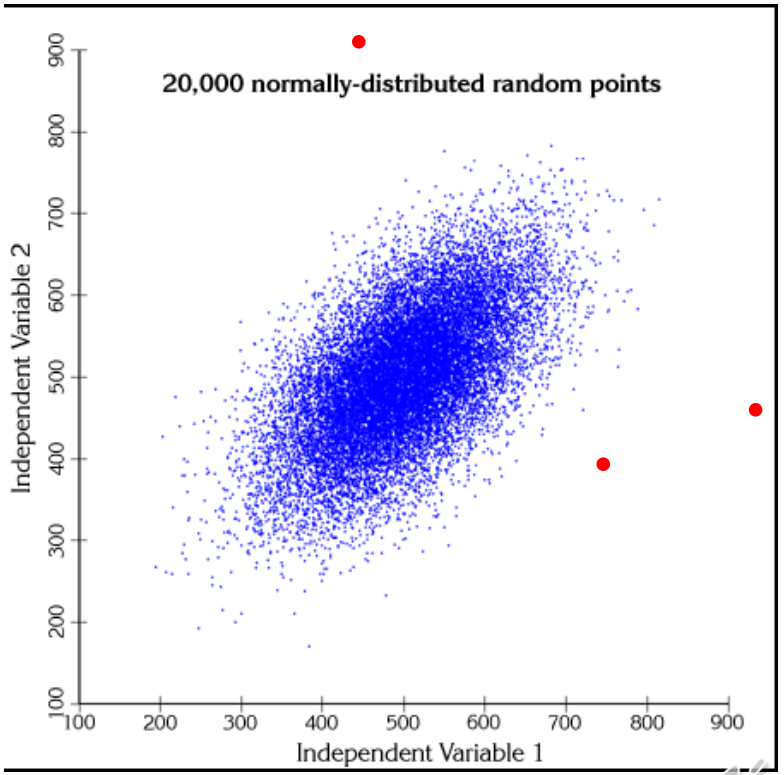


In the Normal distribution we can detect abnormal values in one attribute and abnormal combinations of attribute values by taking into account distance to the mean, variance and covariance

$$s_{jk} = \frac{\Sigma_{i=1}^{n}(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{n-1}$$

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

p attributes



Mahalanobis distance of point *x* **to a Normal sample distribution**

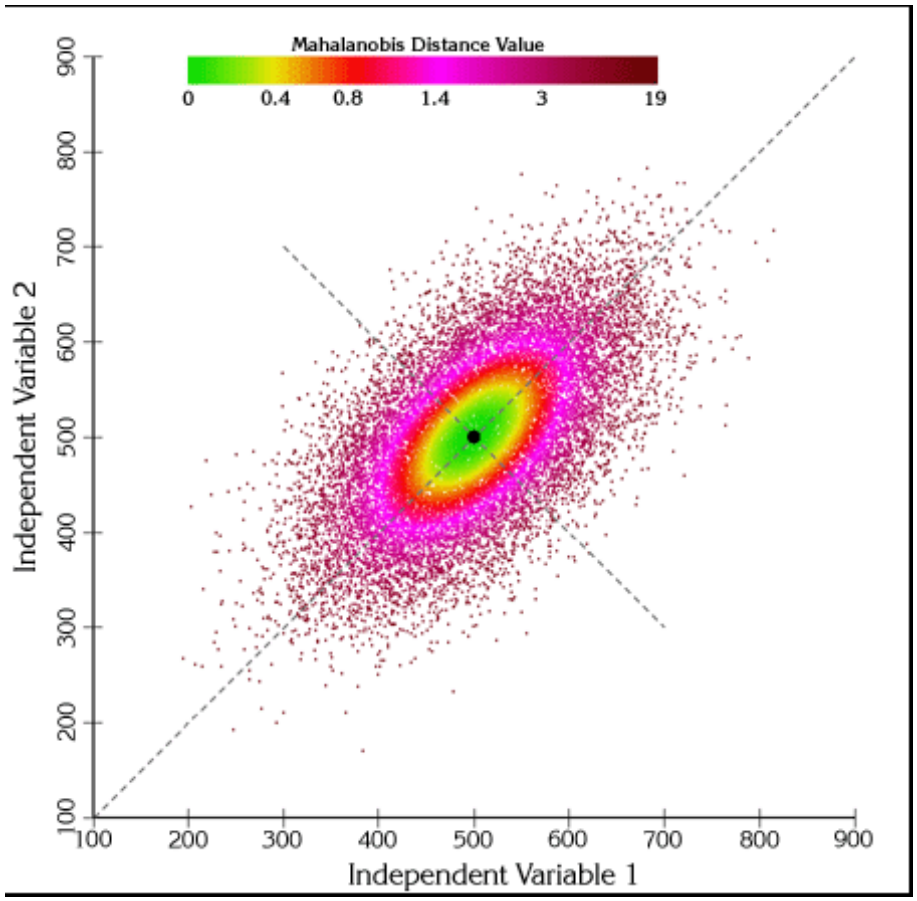
$$d(\mathrm{x}) \equiv d_{S,\bar{\mathrm{x}}}(\mathrm{x}) = \sqrt{(\mathrm{x}-\bar{\mathrm{x}})^{\mathrm{T}} S^{-1} (\mathrm{x}-\bar{\mathrm{x}})}$$

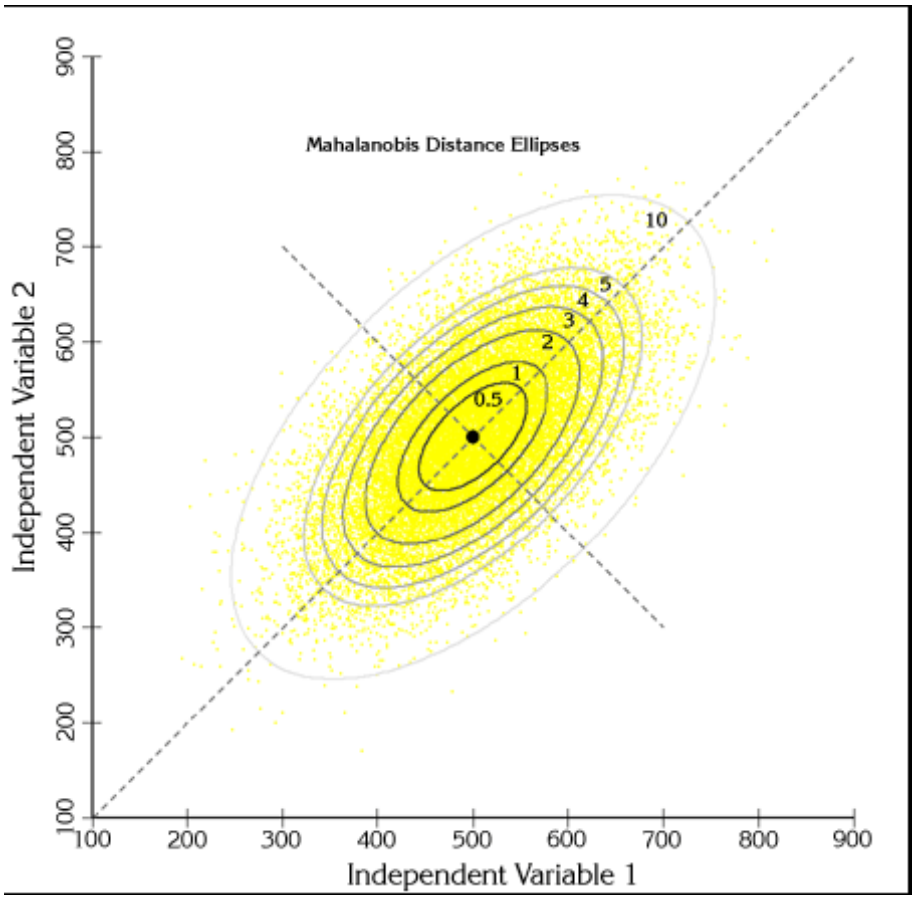
$$d^2(\mathrm{x}) = (\mathrm{x}-\bar{\mathrm{x}})^{\mathrm{T}} S^{-1} (\mathrm{x}-\bar{\mathrm{x}})$$

Mahalanobis distance is a multidimensional version of a z-score (Mah.dist. is applied to the data without previously been normalized). It measures the distance of a case to the centroid of a normal distribution.

Unsupervised Methods

Stat. Approaches: Multivariate



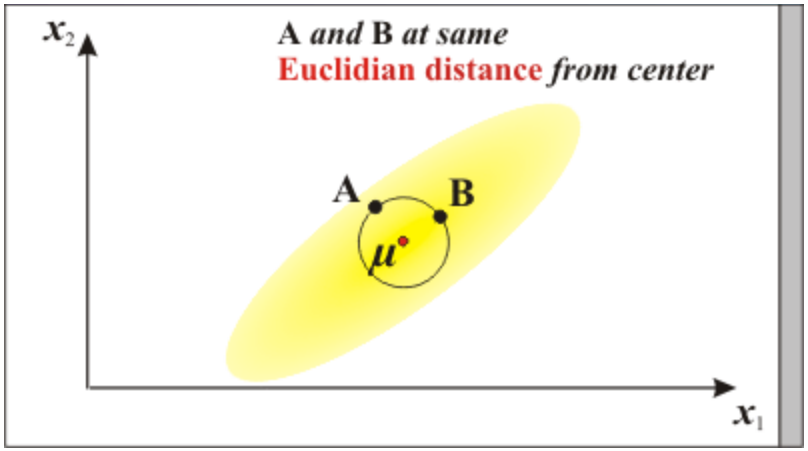


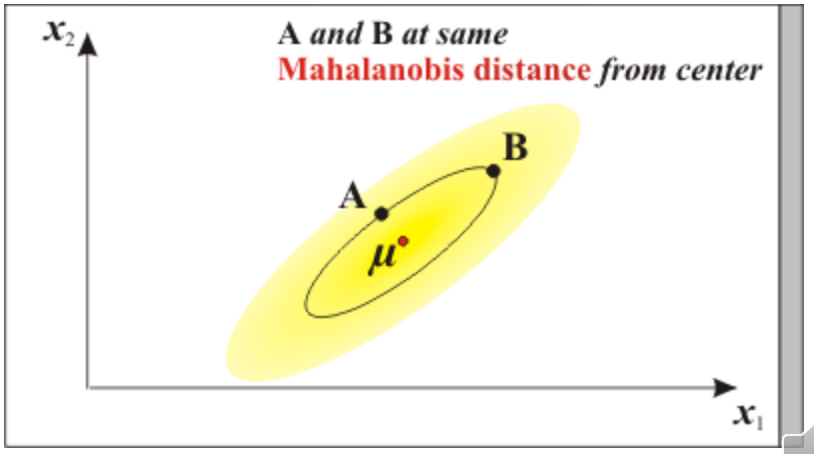
Formally, the Mahalanobis distance measures the distance of a case to the multidimensional mean μ (centroid) of a distribution, given the covariance Σ (multidimensional variance) of the distribution

$$N(x) = \frac{1}{\sqrt{(2\pi)^p \, |\Sigma|}} e^{-\frac{(x-\mu)^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (x-\mu)}{2}}$$

$$d_{\boldsymbol{\Sigma},\mu}(x) = \sqrt{(x-\mu)^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (x-\mu)}$$

$$d^2{}_{\boldsymbol{\Sigma},\mu}(x) = (x-\mu)^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (x-\mu)$$





Theoretical Mahalanobis distances distribution $d^2_{\boldsymbol{\Sigma},\mu}(x_i) \sim \chi^2_p$

The parameters are unknown, so we work with their estimators.

$d^2_{S,\bar{x}}(x_i) = (x_i - \bar{x})^{\mathbf{T}} S^{-1} (x_i - \bar{x})$ is an estimator of $d^2_{\boldsymbol{\Sigma},\mu}(x_i)$

Exact Sample Mahalanobis distances distribution

$$\left(\frac{(N-1)^2}{N}\right) d^2_{S,\bar{x}}(x_i) \sim Beta\left(\frac{p}{2}, \frac{(N-p-1)}{2}\right)$$

Approximate Sample Mahalanobis distances distribution

$$d^2_{S,\bar{x}}(x_i) \approx \chi^2_p \qquad \left(\frac{Np}{N-p}\right) d^2_{S,\bar{x}}(x_i) \approx F_{p,N-p}$$

Which test should be considered?

If we are interested on testing if there is exactly one outlier, a single test should be considered:

$$H_0 = x_{highest} \sim N(\mu, \Sigma)$$

Where $x_{highest}$ is the data value with highest Mah. distance

If we are interested on testing whether there are several outliers, a multiple comparison should be performed on the dataset with size *N*

$$H_{0i} = x_i \sim N(\mu, \Sigma) \quad i = 1...N$$

A correction is used to control the FWER error

So, any method to control FWER error rate like Bonferroni correction or Holm's/Hochberg's procedures should be applied. Traditionally, Sidak correction is used:

$$\alpha_N = 1-(1-\alpha)^{1/N}\,(Sidak) \cong (\geq)\alpha / N\,(Bonferroni)$$

$$\alpha_N = 1-(1-\alpha)^{1/N} \qquad \alpha = 0.05 \rightarrow \alpha_{1000} \approx 0.00005!$$

→ Therefore, it will be very difficult to reject ☹

If we want to test for k or less outliers, we use 0.05/k as significance level

All the previous tests (1 or multivariate) require Normality
They are based on the sample mean and variance/covariances , which are not ***robust*** estimators:
Mean(1,2,3,4,5) = 3 Mean(1,2,3,4,200) = 42

The ***breakdown point*** of an estimator is the proportion of incorrect observations (e.g. arbitrarily large observations) an estimator can handle before giving an incorrect (e.g., arbitrarily large) result. It is a value between 0 and 0.5

Breakdown point of the mean: 0 (just one point may alter the result) → It's not a robust estimator.

Breakdown point of the median: 0.5 → It's a robust estimator.

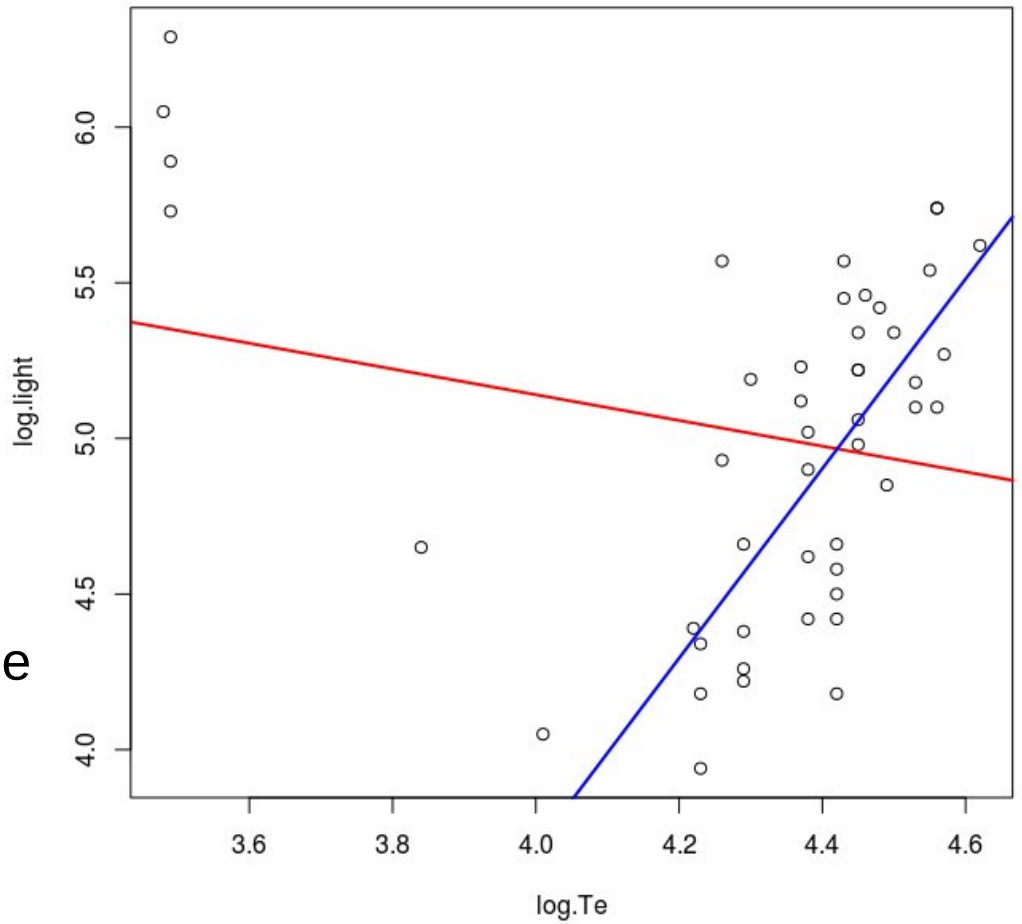
Breakdown point of the X% trimmed mean (X% of the greatest and lowest values are discarded to compute the mean): 0.X

Example of how outliers could affect a Linear Regression (using non robust estimators -red- and robust ones -blue-)

DataSet: starsCYG

log.Te: Logarithm of the effective temperature at the surface of the star (Te).

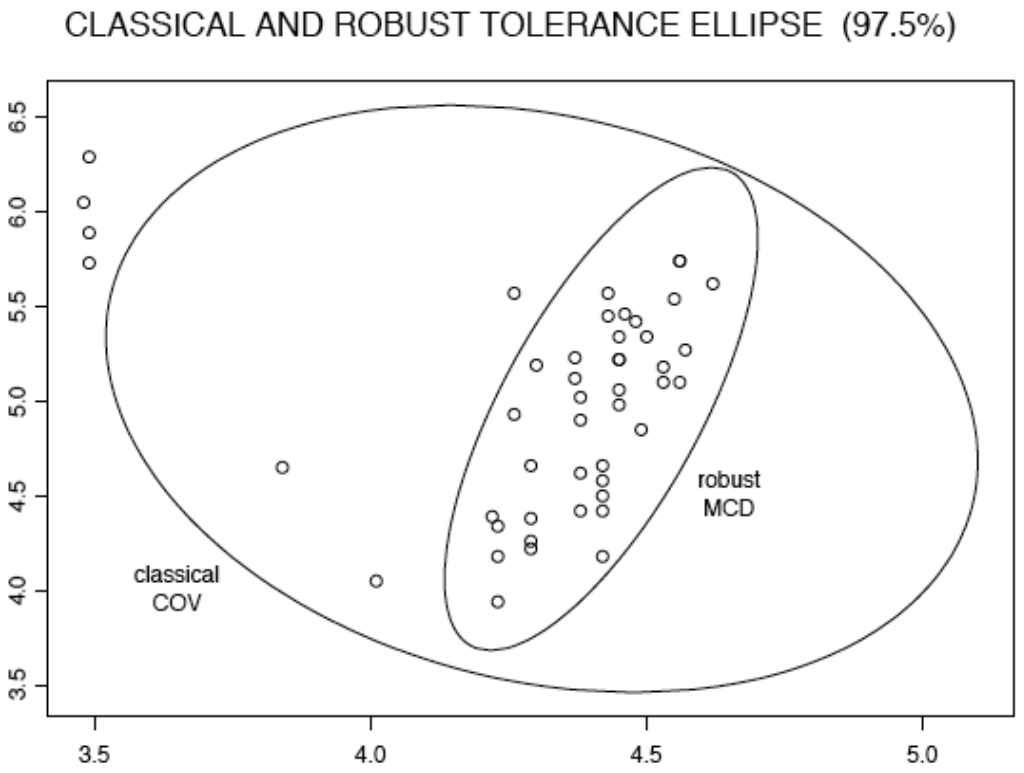
log.light: Logarithm of its light intensity (*L/L_0*)



In the multivariate case robust estimators *S** of the covariance matrix are used → MCD (minimum covariance determinant) is an iterative method introduced by Rousseeuw, (84,99)

The robust estimator MCD of the Cov.Matrix is the sample covariance of the *good* points

The robust estimator of the mean is the sample mean of those points included in the computation of MCD



Squared Mahalanobis distance estimator:

$$d^2_{S,\bar{x}}(x_i) = (x_i - \bar{x})^{\mathbf{T}} S^{-1} (x_i - \bar{x})$$

Robust (MCD) Squared Mahalanobis distance estimator:

$$d^2_{S^*,\bar{x}^*}(x_i) = (x_i - \bar{x}^*)^{\mathbf{T}} S^{*-1} (x_i - \bar{x}^*)$$

The exact distribution of the robust distance estimator is unknown. Hardin and Rocked (2005) showed that:

- The ChiSq approximation is well suited for the points used in the computation of *S**
- The *F* approximation (properly adjusted to take into account the size of the sample used to estimate Σ) is better suited for points which are independent of those used to compute *S** (this is the case of outliers)

Peter J. Rousseeuw and Mia Hubert. Robust statistics for outlier detection. 2011. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 1, pages 73–79, January/February 2011

Hubert, M. and Debruyne, M. (2010), Minimum covariance determinant. WIREs Comp Stat, 2: 36–43

Hardin, J., Rocke, D.M., 2005. The distribution of robust distances. Journal of Computational and Graphical Statistics 14

Cerioli, Andrea, (2010), Multivariate Outlier Detection With High-Breakdown Estimators, Journal of the American Statistical Association, 105, issue 489, p. 147-156.

Andrea Cerioli, Alessio Farcomeni, Error rates for multivariate outlier detection, Computational Statistics & Data Analysis, Volume 55, Issue 1, 1 January 2011, Pages 544-553

Limitations when using Mahalanobis distance estimators.

- The curse of dimensionality
- Non linear relations

Mixture of Normal Distributions