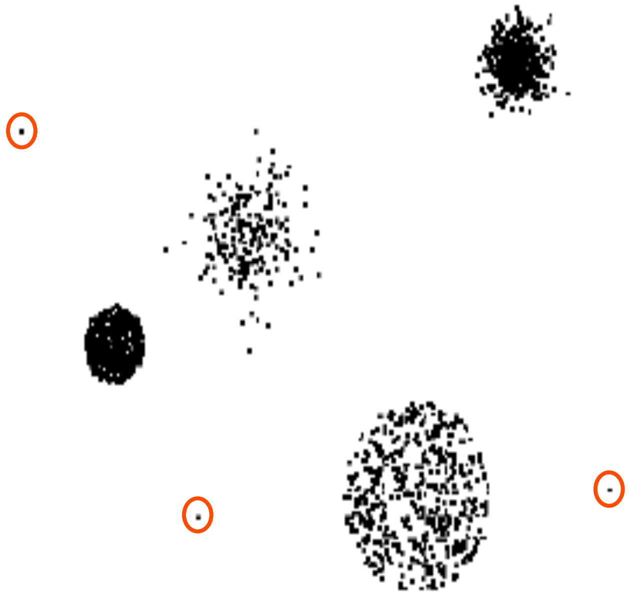
# Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



# What is an outlier

- Definition by Hawkins [Hawkins 1980]

“ An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”

- Definition by Barnet and Lewis [Barnet and Lewis, 1994]

“An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.

- Definition by Johnson [Johnson, 1992]

“An outlier is an observation in a data set which appears to be inconsistent with the remainder of that set of data”

# Outlier Detection Methods

- Taxonomy of outlier detection methods:
  - Univariate and multivariate
  - Parametric (statistical) and non-parametric (model-free)
    - Parametric: assume a known underlying distribution of the observations, or based on statistical estimates of unknown distribution parameters
    - Non-parametric:
      - data-mining methods (distance-based methods): based on local distance measures and are capable of handling large databases
      - Clustering techniques: a cluster of small sizes can be considered as clustered outliers

- **Univariate Outlier detection**
- Multivariate Outlier detection

# statistical method

- For any confidence coefficient alpha, 0<alpha<1, the alpha-outlier region of N(mu, sigma^2) distribution is defined by

$$out(\alpha,\mu,\sigma^2)=\left\{x:|\,\left|x-\mu\right|>z_{1-\alpha/2}\,\sigma\right\}$$

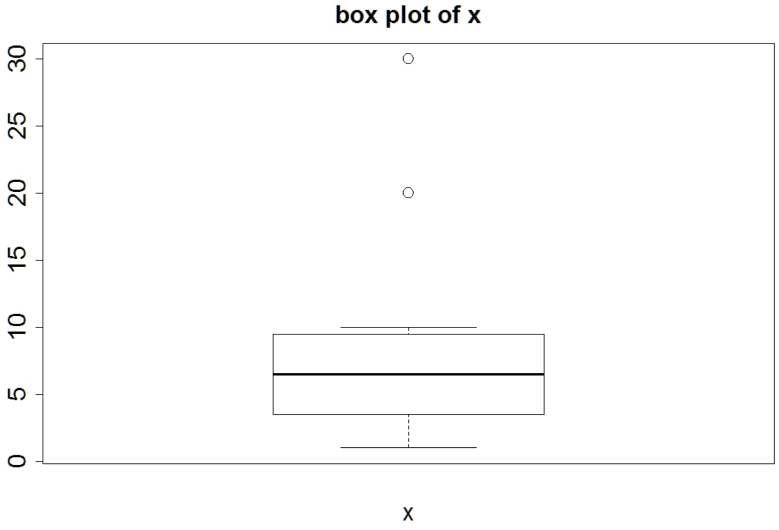
# box-and-whisker method

- Tukey's box-and-whisker method:
  - An observation is considered to be an outlier when it is larger than the "whiskers" of the set of observations
  - Can be detected using box plot
  - Limitation: fails when data are skewed, for examples in cases of exponential, log-normal distribution

Example:

x=[1 2 3 4 5 6 7 8 9 10 20 30]

Outliers: 20 and 30



# Hiridoglou and Berthelot’s method

- Hiridoglou and Berthelot’s method:
  - Suitable for skewed data
  - Find outliers for both side of the distribution

$$h(x) = \max(\frac{x}{x^{*}}, \frac{x^{*}}{x}) \geq r, and\ x > 0$$

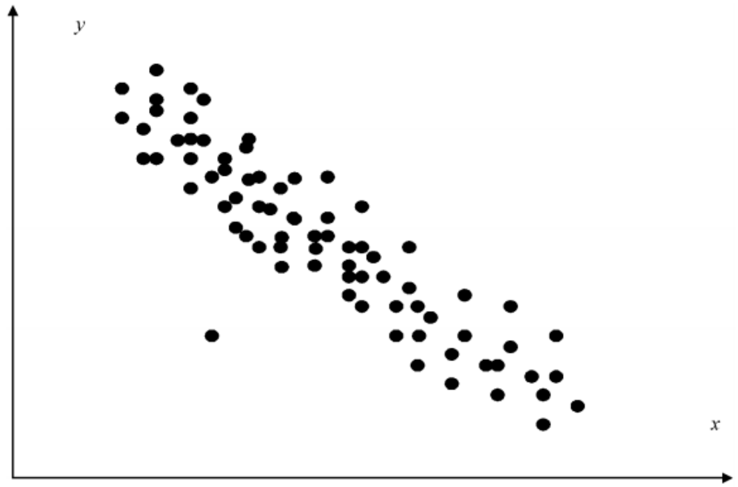
h(x): score function (logical indicator)

x*: median value

r: user defined reference value

# Multivariate outlier

- Relationship among the variables should be considered

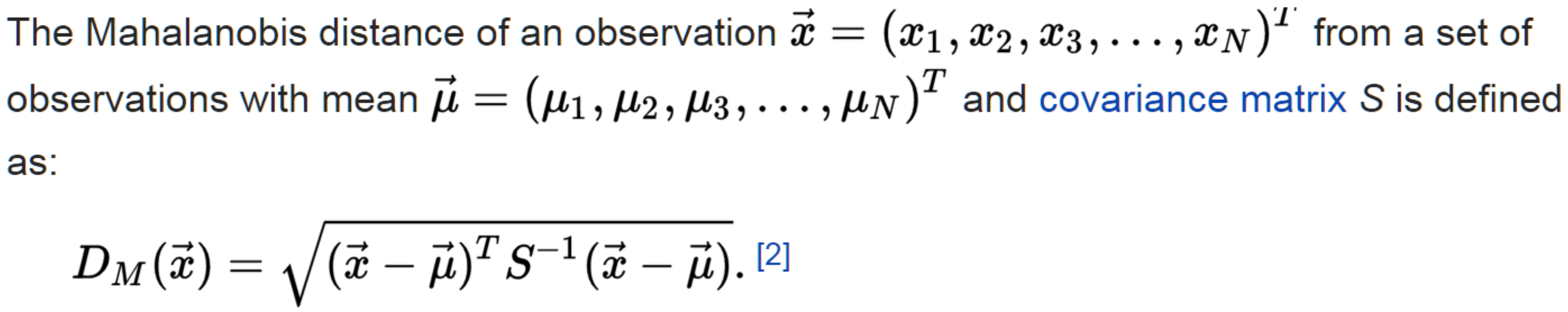


A two-dimensional space with one outlier (Ben-Gal I, 2005)

- When considering  each measure separately, the lower left point falls close to the center of the univariate distributions
- In the two-dimensional case, the lower left point is an outlier

# Multivariate Outlier detection

- Mahalanobis distance



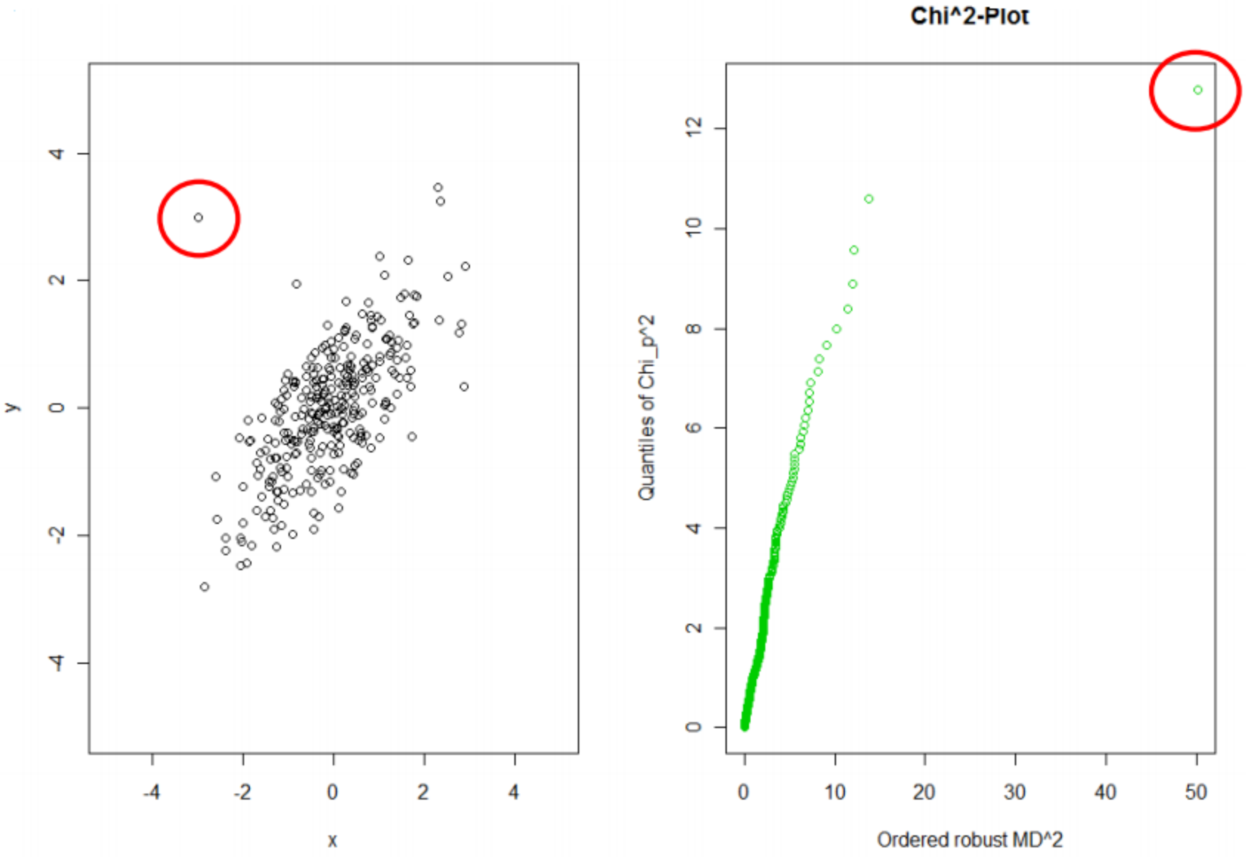
The Mahalanobis distance of an observation $\vec{x} = (x_1, x_2, x_3, \dots, x_N)^T$ from a set of observations with mean $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \dots, \mu_N)^T$ and covariance matrix $S$ is defined as:

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}.$$ [2]

- Assume data is multivariate normally distributed
- Mahalanobis distance of samples follow a Chi-Square distribution with d degree of fredom
- Samples with Mahalanobis distance that don’t fit at all to a Chi-Square distribution are outliers (check with Q-Q plot)

# Mahalanobis distance continued

Samples with Mahalanobis distance that don't fit at all to a Chi-Square distribution are outliers



Multivariate outlier detection (*from applied multivariate statistics-Spring 2012, ETH*)