La méthode ICS pour détecter des atypiques en multivarié

Anne Ruiz-Gazen[1]

Co-authors: Aurore Archimbaud[1]
Klaus Nordhausen[2]

(1) TSE-R, University of Toulouse 1 Capitole, France.
(2) CSTAT, Vienna University of Technology, Austria.

JMS 2018
Paris, 12-14 Juin 2018

Table of Contents

Outlier definition and detection

- Let us consider $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ a p-variate dataset.
- "An outlier is an observation which deviates so much from the other observations as to arouse suspicious that is was generated by a different mechanism." (Hawkins, 1980).
- For one variable x and n observations $x_1, \ldots, x_n$, a simple rule is to look at the large values of:

$$\frac{|x_i - \bar{x}_n|}{\hat{\sigma}_n}, \text{ for } i = 1, \ldots, n$$

- The generalization of this rule to the multivariate case is the Mahalanobis distance of each observation to the mean:

$$\mathrm{MD}^2(\mathbf{x}_i) = ||\mathrm{COV}(\mathbf{X}_n)^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}_n)||^2$$

  where $\bar{\mathbf{x}}_n$ denotes the empirical mean and $\mathrm{COV}(\mathbf{X}_n)$ the empirical covariance matrix.

Table of Contents

The Mahalanobis distance (sample version)

- Classical measure for multivariate outlier detection:

$$\mathrm{MD}^2(\mathbf{x}_i) = ||\mathrm{COV}(\mathbf{X}_n)^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}_n)||^2$$

  where $\bar{\mathbf{x}}_n$ denotes the empirical mean and $\mathrm{COV}(\mathbf{X}_n)$ the empirical covariance matrix.

- An observation $\mathbf{x}_i$ is identified as an outlier if:

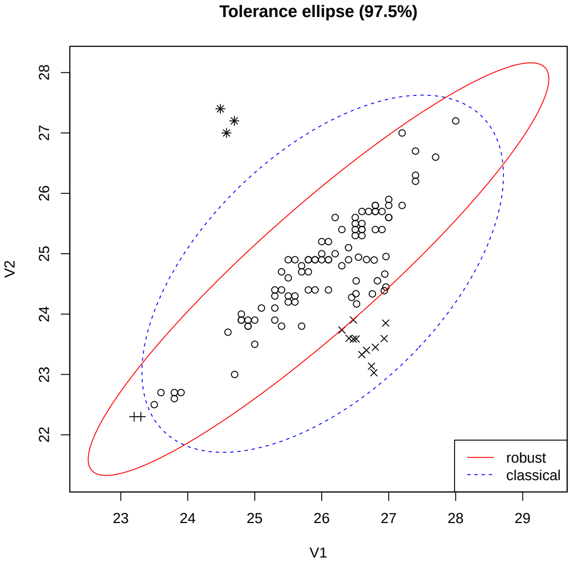
$$\mathrm{MD}^2(\mathbf{x}_i) \geq c_{p,1-\alpha}$$

  with $c_{p,1-\alpha}$ the $(1-\alpha)$-th quantile of a χ^2_p distribution.

- Alternative: use a robust version based on the MCD[1] estimators.

[1] Minimum Covariance Determinant: reweighted empirical mean and covariance estimators of the MCD subset based on the $h \approx \alpha * n$ observations whose covariance matrix has the smallest determinant.

The Mahalanobis distance (sample version)



The Mahalanobis distance (functional version)

- Functional version: $\mathbf{X}$ is a p-variate random vector, $F_{\mathbf{X}}$ its cumulative distribution function and $\mathbf{m}(F_{\mathbf{X}})$ an affine equivariant location estimator.
- Let $\mathcal{P}_p$ be the set of all symmetric positive definite matrices of order p.
- A scatter functional is defined as a matrix $\mathbf{V}(F_{\mathbf{X}}) \in \mathcal{P}_p$, uniquely defined at $F_{\mathbf{X}}$, which is affine equivariant in the sense that:

$$\mathbf{V}(F_{\mathbf{AX}+\gamma}) = \mathbf{AV}(F_{\mathbf{X}})\mathbf{A}',$$

  for all $p \times p$ non-singular matrices $\mathbf{A}$ and all $\gamma \in \Re^p$.
- The Mahalanobis distance:

$$d^2(\mathbf{X}) = (\mathbf{X} - \mathbf{m}(F_{\mathbf{X}}))'\mathbf{V}(F_{\mathbf{X}})^{-1}(\mathbf{X} - \mathbf{m}(F_{\mathbf{X}}))$$

- $d^2(\mathbf{X})$ is affine invariant: $d^2(\mathbf{AX} + \gamma) = d^2(\mathbf{X})$

Limitation of the Mahalanobis distance

- Classical measure for multivariate outlier detection:

$$d^2(\mathbf{X}) = (\mathbf{X} - \mathbb{E}(\mathbf{X}))'\mathrm{COV}(\mathbf{X})^{-1}(\mathbf{X} - \mathbb{E}(\mathbf{X})).$$

- Let us consider the following model (M) which is a mixture of $(q+1)$ Gaussian distributions:

$$\mathbf{X} \sim \underbrace{(1-\epsilon)\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_W)}_{\text{majority of the data}} + \underbrace{\sum_{h=1}^{q} \epsilon_h \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_W)}_{\text{clustered outliers}}, \text{ where } \epsilon = \sum_{h=1}^{q} \epsilon_h < 0.5$$

We have:
$\mathbb{E}(\mathbf{X}) = (1-\epsilon)\,\boldsymbol{\mu}_0 + \sum_{h=1}^{q} \epsilon_h \boldsymbol{\mu}_h$ and $\mathrm{COV}(\mathbf{X}) = \boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_B$, with
$\boldsymbol{\Sigma}_B = (1-\epsilon)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_{\mathbf{X}})' + \sum_{h=1}^{q} \epsilon_h(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\mathbf{X}})'$.

Limitation of the Mahalanobis distance

"Non-outlier" observations $\mathbf{X}_{no} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_W)$,

"Outlier" observations $\mathbf{X}_{o,h} \sim \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_W)$, $\mathbf{X}_{no} \perp \mathbf{X}_{o,h}$, for $h = 1, \ldots, q$

Proposition

Assuming that q is fixed and p becomes large, the distribution of the difference:

$$\frac{1}{2\sqrt{p}} \left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no}) - \mathbb{E}\left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no})\right)\right) \xrightarrow[p\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

where the expectation $\mathbb{E}\left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no})\right)$ does not depend on the dimension p.

Thus, the probability of finding outliers decreases when the dimension p increases.

A similar result is also derived for the robust case (when considering $\boldsymbol{\Sigma}_W$ instead of $\mathrm{COV}(\mathbf{X})$).

20 obs $\sim \mathcal{N}_p(\mu_1, \mathbf{W})$ & 980 obs $\sim \mathcal{N}_p(0, \mathbf{W})$

with $\mu_1 = (6, 0, \ldots, 0)'$, $\mathbf{W} = \text{diag}(1, 4, \ldots, 4)$, $n = 1000$,

2% of outliers and $p = 6, 25, 50$.

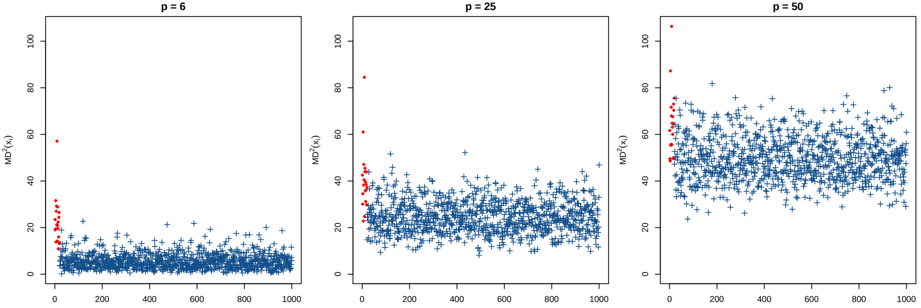


Table of Contents

ICS Tyler et al., 2009

- Simultaneous diagonalization of two scatter matrices $\mathbf{V}_1$ and $\mathbf{V}_2$:

$$\mathbf{V}_1^{-1}\mathbf{V}_2\mathbf{B}' = \mathbf{B}'\mathbf{D}$$

  where the diagonal matrix $\mathbf{D}$ contains the eigenvalues $\rho_1, \ldots, \rho_p$ of $\mathbf{V}_1^{-1}\mathbf{V}_2$ in decreasing order and $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ contains the corresponding eigenvectors as its rows such that: $\mathbf{B}\mathbf{V}_1\mathbf{B}' = \mathbf{I}_p$.
- New components:

$$\mathbf{Z} = \mathbf{B}(\mathbf{X} - \mathbf{m}_1)$$

  with $\mathbf{m}_1$ being a location estimator associated with $\mathbf{V}_1$.

ICS

- Many possibilities for $\mathbf{V}_1$ and $\mathbf{V}_2$.
- For example, $\mathbf{V}_1 = \mathrm{COV}(\mathbf{X})$ and $\mathbf{V}_2 = \mathrm{COV}_4(\mathbf{X})$ with

$$\mathrm{COV}_4(\mathbf{X}) = \frac{1}{p+2}\,\mathbb{E}\left[d^2(\mathbf{X})(\mathbf{X}-\mathbb{E}(\mathbf{X}))(\mathbf{X}-\mathbb{E}(\mathbf{X}))'\right].$$

- M-estimators, MCD estimators,...

ICS

- Focusing only on the first eigenvalue and the first eigenvector, it is equivalent to maximizing the ratio:

$$\mathcal{K}(\mathbf{b}) = \frac{\mathbf{b}'\mathbf{V}_2\mathbf{b}}{\mathbf{b}'\mathbf{V}_1\mathbf{b}}$$

  where ρ_1 is the maximal possible value of $\mathcal{K}(\mathbf{b})$ over $\mathbf{b} \in \Re^p$ which is achieved in the direction of the eigenvector $\mathbf{b}_1$. This ratio can be seen as a generalized measure of kurtosis.
- ICS follows the same "philosophy" as PCA. However, it differs from PCA which maximizes a variance criterion and which is only orthogonally invariant.
- In a different (supervised) context where the groups are known, one can use the between and within covariance matrices as $\mathbf{V}_1$ and $\mathbf{V}_2$ which leads to Discriminant Analysis.

Property of the components

Under the model (M):

$$\mathbf{X} \sim (1-\epsilon)\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_W) + \sum_{h=1}^{q} \epsilon_h \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_W),$$

where $\epsilon = \sum_{h=1}^{q} \epsilon_h < 0.5$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_q - \boldsymbol{\mu}_0$ span some q-dimensional hyperplane.

Theorem (Tyler et al., 2009)

Suppose that the roots $\rho_1, \ldots, \rho_p$ consist of m distinct values, say $\rho_{(1)}, \ldots, \rho_{(m)}$, with $\rho_{(k)}$ having multiplicity p_k for $k = 1, \ldots, m$ and hence $p_1 + \cdots + p_m = p$.
There is at least one root $\rho_{(k)}$ with multiplicity greater than or equal to $p - q$.
If no root has multiplicity greater than $p - q$, then there is a root with multiplicity $p - q$, say $\rho_{(j)}$, such that

$$\mathrm{span}\{\boldsymbol{\Sigma}_W^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0) | k = 1, \ldots, q\} = \mathrm{span}\{\mathbf{B}_q\}$$

where $\mathbf{B}_q = (\mathbf{b}_1, \ldots, \mathbf{b}_{p_1+\cdots+p_{j-1}}, \mathbf{b}_{p_1+\cdots+p_{j+1}}, \ldots, \mathbf{b}_p)$.

$\Rightarrow$ Fisher's Linear Discriminant subspace even though the groups are unknown

If $\mathbf{V}_1 = \mathrm{COV}(\mathbf{X})$ and $\mathbf{V}_2 = \mathrm{COV}_4(\mathbf{X})$

Mean-shift outlier model ($q = 1$)

$$\mathbf{X} \sim (1-\epsilon)\,\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_1) + \epsilon\,\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1), \quad \text{with } \epsilon < 0.5 \text{ and } \boldsymbol{\mu} \neq \mathbf{0}_p \text{ a } p\text{-vector.}$$

The eigenvalues of $\mathrm{COV}^{-1}(\mathbf{X})\mathrm{COV}_4(\mathbf{X})$ are such that either:

(a) $\rho_1 > \rho_2 = \ldots = \rho_p$ if $\epsilon < (3-\sqrt{3})/6$ ($\approx 21\%$),

(b) $\rho_1 = \ldots = \rho_{p-1} > \rho_p$ if $\epsilon > (3-\sqrt{3})/6$,

(c) $\rho_1 = \rho_2 = \ldots = \rho_p$ if $\epsilon = (3-\sqrt{3})/6$.

Moreover, if (a) (resp. (b)) holds then the eigenvector associated with ρ_1 (resp. ρ_p) is proportional to $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}$.

Symmetric contamination of a Gaussian distribution

Proposition

$$\mathbf{X} \sim (1-\epsilon)\,\mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_{21}) + \frac{\epsilon}{2}\,\mathcal{N}(\delta \mathbf{e}_1, \mathbf{\Sigma}_{22}) + \frac{\epsilon}{2}\,\mathcal{N}(-\delta \mathbf{e}_1, \mathbf{\Sigma}_{22})$$

with $\mathbf{\Sigma}_{21} = \mathrm{diag}(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{12}^2)$, $\mathbf{\Sigma}_{22} = \mathrm{diag}(\sigma_{21}^2, \sigma_{22}^2, \ldots, \sigma_{22}^2)$ and $\delta \neq 0$.

The eigenvalues of $\mathrm{COV}^{-1}(\mathbf{X})\mathrm{COV}_4(\mathbf{X})$ are such that either:

$$\begin{array}{lccccccc}
\text{(a)} & \rho_1 & > & \rho_2 & = & \ldots & = & \rho_p\,, \\
\text{(b)} & \rho_1 & = & \ldots & = & \rho_{p-1} & > & \rho_p, \\
\text{(c)} & \rho_1 & = & \rho_2 & = & \ldots & = & \rho_p.
\end{array}$$

with $\rho_1 = \frac{1}{p+2}\left(\frac{3(1-\epsilon)\sigma_{11}^4 + \epsilon(3\sigma_{21}^4 + 6\sigma_{21}^2\delta^2 + \delta^4)}{((1-\epsilon)\sigma_{11}^2 + \epsilon(\sigma_{21}^2 + \delta^2))^2} + p - 1\right)$

and $\rho_2 = \frac{1}{p+2}\left(\frac{3((1-\epsilon)\sigma_{12}^4 + \epsilon\sigma_{22}^4)}{((1-\epsilon)\sigma_{12}^2 + \epsilon\sigma_{22}^2)^2} + p - 1\right)$.

Moreover, if (a) (resp. (b)) holds then the eigenvector associated with ρ_1 (resp. ρ_p) is proportional to $\mathbf{e}_1$.

Corollary: with $\mathbf{\Sigma}_{21} = \mathbf{\Sigma}_{22} = \mathbf{I}_p$, (a) holds if $\epsilon < 1/3$.

Scale-shift outlier model ($q \leq p$)

Proposition

$$\mathbf{X} \sim (1-\epsilon)\mathcal{N}(\mathbf{0}_p, \mathbf{I}_p) + \epsilon\mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_5)$$

with $\epsilon < 0.5$, $\mathbf{\Sigma}_5 = \mathrm{diag}(\alpha\mathbf{I}_q, \mathbf{I}_{p-q})$, $q < p$ and $\alpha > 1$.

The eigenvalues of $\mathrm{COV}^{-1}(\mathbf{X})\mathrm{COV}_4(\mathbf{X})$ are such that:

$$\rho_1(\mathbf{F_X}) = \cdots = \rho_q(\mathbf{F_X}) > \rho_{q+1}(\mathbf{F_X}) = \cdots = \rho_p(\mathbf{F_X})$$

Moreover, the eigenvectors associated with the q largest eigenvalues span the subspace spanned by $\{\mathbf{e}_1, \ldots, \mathbf{e}_q\}$.

Remark: if $q = p$ then all the eigenvalues are equal.

ICS (sample version)

- Simultaneous diagonalization of two scatter matrices $\mathbf{V}_{1,n}$ and $\mathbf{V}_{2,n}$:

$$\mathbf{V}_{1,n}^{-1}\mathbf{V}_{2,n}\mathbf{B}_n' = \mathbf{B}_n'\mathbf{D}_n$$

where the diagonal matrix $\mathbf{D}_n$ contains the eigenvalues $\rho_1, \ldots, \rho_p$ of $\mathbf{V}_{1,n}^{-1}\mathbf{V}_{2,n}$ in decreasing order and $\mathbf{B}_n = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ contains the corresponding eigenvectors as its rows such that: $\mathbf{B}_n\mathbf{V}_{1,n}\mathbf{B}_n' = \mathbf{I}_p$.

- The (affine) invariant coordinates are:

$$\mathbf{z}_i = \mathbf{B}_n(\mathbf{x}_i - \mathbf{m}_{1,n})$$

with $\mathbf{m}_{1,n}$ being the location estimator associated with $\mathbf{V}_{1,n}$.

- For k selected ICS components, we define an ICS distance as:

$$\mathrm{ICSD}^2_{\mathbf{V}_{1,n}^{-1}\mathbf{V}_{2,n}}(\mathbf{x}_i, k) = \mathbf{z}_{i,k}'\mathbf{z}_{i,k}$$

with $\mathbf{z}_{i,k} = \mathbf{B}_{n,k}(\mathbf{x}_i - \mathbf{m}_{1,n})$ and $\mathbf{B}_{n,k}$ contains the first k rows of $\mathbf{B}_n$.

Relationship with the MD^2

For k selected ICS components, we define an ICS distance as:

$$\mathrm{ICSD}^2_{\mathbf{V}_{1,n}^{-1}\mathbf{V}_{2,n}}(\mathbf{x}_i, k) = \mathbf{z}'_{i,k}\mathbf{z}_{i,k}$$

with $\mathbf{z}_{i,k} = \mathbf{B}_{n,k}(\mathbf{x}_i - \mathbf{m}_{1,n})$ and $\mathbf{B}_{n,k}$ contains the first k rows of $\mathbf{B}_n$.

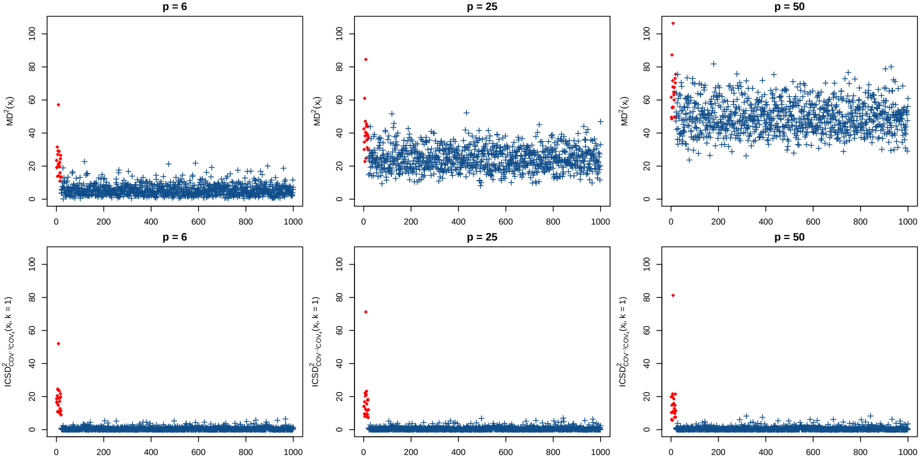
Property

$$\mathrm{ICSD}^2_{\mathrm{COV}(\mathbf{X}_n)^{-1}\mathrm{COV}_4(\mathbf{X}_n)}(\mathbf{x}_i, p) = \mathrm{MD}^2(\mathbf{x}_i)$$

If the structure of outlyingness is contained on a subspace of dimension q less than p, then ICS has an advantage over MD if we select $k = q$ components.

20 obs $\sim \mathcal{N}_p(\mu_1, \mathbf{W})$ & 980 obs $\sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{W})$

with $\mu_1 = (6, 0, \ldots, 0)'$, $\mathbf{W} = \mathrm{diag}(1, 4, \ldots, 4)$, $n = 1000$, 2% of outliers and $p = 6, 25, 50$.



Selection of the Invariant Coordinates

As we look only for a small proportion of outliers, the outliers should be found in the first components.

Approaches:

- Based on the analysis of the eigenvalues:
  - Visually, using a scree plot.
  - Using asymptotic distribution of the eigenvalues.
  - Using quasi inferential procedures (parallel analysis).
- Based on the analysis of the Invariant Components:
  - Using marginal normality tests.

In this context of particular sequential multiple testing, we apply the Bonferroni correction on the significance level: $\alpha_i = \alpha/i$ for $i = 1, \ldots, p$ with $\alpha = 5\%$.

The ICSOutlier and ICSShiny R packages

Detection of a small proportion of outliers via ICS can be easily done using our package ICSOutlier which is available on CRAN or with the ICSShiny application.

There the user can:

- Choose the scatter matrices $\mathbf{V}_{1,n}$ and $\mathbf{V}_{2,n}$.
- Choose the ICS components visually, using parallel analysis or marginal normality testing.
- Explore the invariant components.
- Identify outliers based on a cut-off obtained from simulations.

Simulations & Real Examples

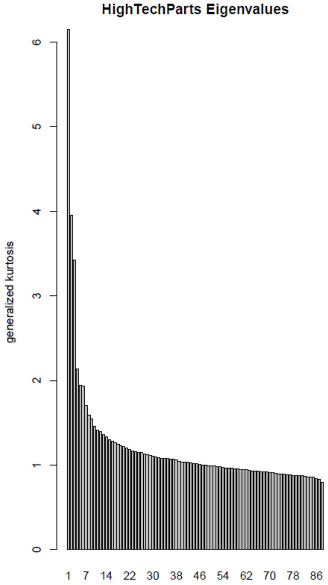
We conducted an extensive simulation study comparing

- MD and robust MD, and PCA based outlier detection methods with ICS
- Concerning ICS, we evaluated:
  - different scatter combinations.
  - different ways to select the ICS components.

Conclusion: ICS and especially the scatter combination $\mathrm{COV}\text{-}\mathrm{COV}_4$ works well and has interesting theoretical properties.

High Tech Parts Example I

- 902 high-tech parts characterized by 88 numerical tests (available in ICSOutlier).
- All parts have been sold (considered flawless) but afterwards two parts have been returned due to malfunctions ⇒ two quality defects.



High Tech Parts Example II

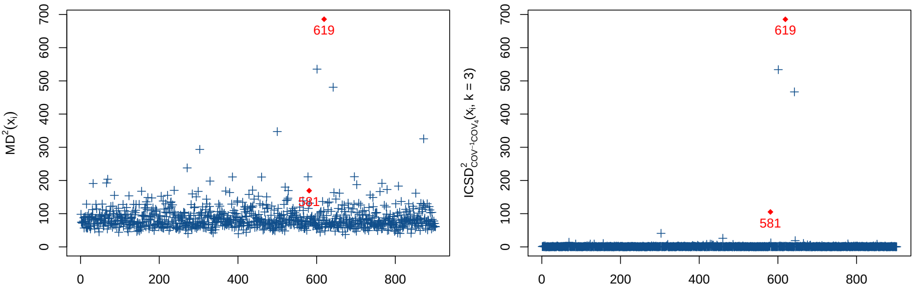


Table of Contents

Conclusion

- Exhibit a limitation of the Mahalanobis distance when p is large and when outliers lie on a subspace.
- Propose a methodology for outlier detection with ICS.
- Generalize ICS to semi-definite positive scatter matrices for data not in general position.
- Perpectives: extend the theory and package to be able to handle also large fractions of outliers (e.g. deriving results for mixtures with three or more components), propose a sparse ICS.