ICSOutlier: An R package for unsupervised multivariate outlier detection with ICS

Aurore Archimbaud[(1)], Klaus Nordhausen[(2)]
and Anne Ruiz-Gazen[(1)]

(1) TSE-R, University of Toulouse 1 Capitole, France
aurore.archimbaud@ut-capitole.fr
(2) University of Turku, Finland.

RJS, April 2017

Table of Contents

Table of Contents

Table of Contents

Table of Contents

Table of Contents

Table of Contents

Context & Objectives

Technical context

Aerospace or Automotive Integrated Circuits (IC) reliability.

Context & Objectives

Technical context

Aerospace or Automotive Integrated Circuits (IC) reliability.
Upstream identification of the few potential Customer Quality Incident (CQI) with a False Alarm Rate (FAR) $< 2\%$.

Context & Objectives

Technical context

Aerospace or Automotive Integrated Circuits (IC) reliability.
Upstream identification of the few potential Customer Quality Incident (CQI) with a False Alarm Rate (FAR) $< 2\%$.

Statistical context

Unsupervised and immediate outlier detection.

Context & Objectives

Technical context

Aerospace or Automotive Integrated Circuits (IC) reliability.
Upstream identification of the few potential Customer Quality Incident (CQI) with a False Alarm Rate (FAR) $< 2\%$.

Statistical context

Unsupervised and immediate outlier detection.
Multivariate numerical data with a small proportion of outliers ($< 2\%$) and many variables.

Context & Objectives

Technical context

Aerospace or Automotive Integrated Circuits (IC) reliability.
Upstream identification of the few potential Customer Quality Incident (CQI) with a False Alarm Rate (FAR) $< 2\%$.

Statistical context

Unsupervised and immediate outlier detection.
Multivariate numerical data with a small proportion of outliers ($< 2\%$) and many variables.

Objectives

- Present the Invariant Coordinate Selection (ICS) method for outlier detection.

Context & Objectives

Technical context

Aerospace or Automotive Integrated Circuits (IC) reliability.
Upstream identification of the few potential Customer Quality Incident (CQI) with a False Alarm Rate (FAR) $< 2\%$.

Statistical context

Unsupervised and immediate outlier detection.
Multivariate numerical data with a small proportion of outliers ($< 2\%$) and many variables.

Objectives

- Present the Invariant Coordinate Selection (ICS) method for outlier detection.
- Present the new ICSOutlier R package (version 0.2-0 on CRAN).

Notations

- $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ a p-variate dataset, with $n > p$.
- The location estimator: $\bar{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$.
- Two symmetric and positive definite scatter estimators:
  - $\mathrm{COV}(\mathbf{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}}_n)(\mathbf{x}_i - \bar{\mathbf{x}}_n)'$
  - $\mathrm{COV}_4(\mathbf{X}_n) = \frac{1}{(p+2)n}\sum_{i=1}^{n} r_i^2(\mathbf{x}_i - \bar{\mathbf{x}}_n)(\mathbf{x}_i - \bar{\mathbf{x}}_n)'$

where $r_i^2 = ||\mathrm{COV}(\mathbf{X}_n)^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}_n)||^2$ is the classical squared Mahalanobis distance.

Table of Contents

The Mahalanobis distance

Classical measure for multivariate outlier detection

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

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.

The Mahalanobis distance

Classical measure for multivariate outlier detection

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

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.

Rousseeuw and Van Zomeren (1990) proposed to use a robust version based on the MCD[1] (Minimum Covariance Determinant) estimators.

[1] Reweighted empirical mean and covariance estimates of the MCD subset based on the $h \approx \alpha * n$ observations whose covariance matrix has the smallest determinant.

The Mahalanobis distance

Classical measure for multivariate outlier detection

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

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.

Rousseeuw and Van Zomeren (1990) proposed to use a robust version based on the MCD[1] (Minimum Covariance Determinant) estimators.

Theroretical Asymptotic Property (Archimbaud et al., 2016)

For mixture of Gaussian distributions, if the dimension of the subspace containing the structure is $q < p$ and p $\nearrow$: the probability of finding outliers $\searrow$ because the variance of the distances increases with p.

[1] Reweighted empirical mean and covariance estimates of the MCD subset based on the $h \approx \alpha * n$ observations whose covariance matrix has the smallest determinant.

High Tech Parts example

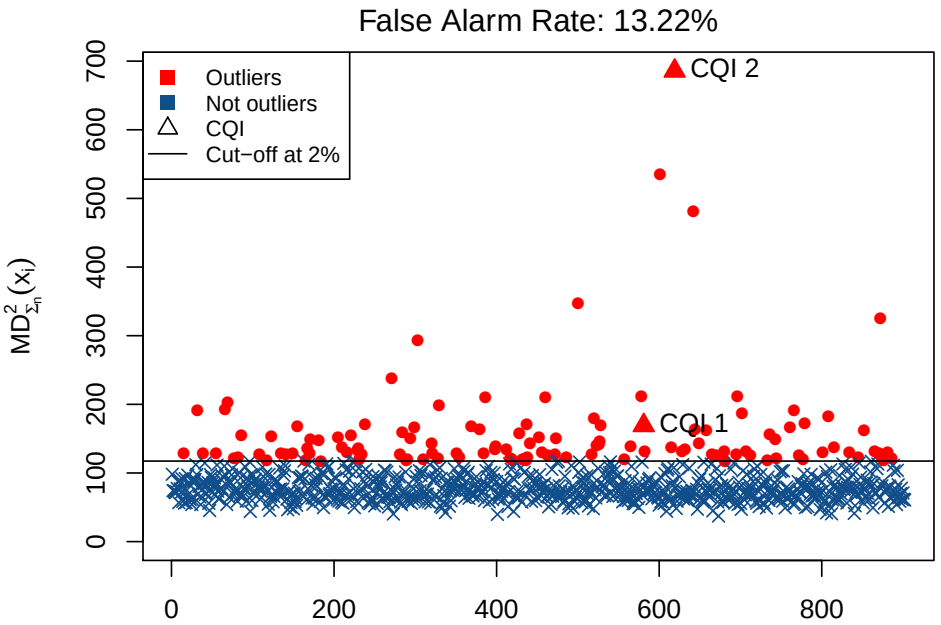
Description

902 high-tech parts characterized by 88 numerical tests. All parts have been considered flawless but later two parts have been classified as CQIs.

High Tech Parts example

Description

902 high-tech parts characterized by 88 numerical tests. All parts have been considered flawless but later two parts have been classified as CQIs.



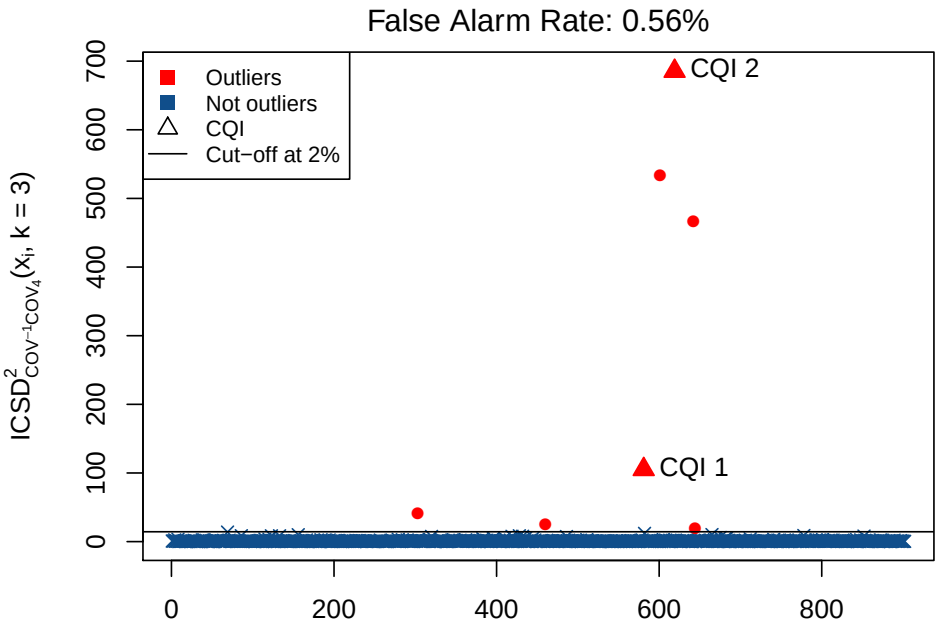


Table of Contents

Principle, (Tyler et al., 2009)

Simultaneous diagonalization of two symmetric and positive definite scatter matrices, $\mathbf{V}_{1,n}$ and $\mathbf{V}_{2,n}$ s.t:

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

where the diagonal matrix $\mathbf{D}_n$ contains the eigenvalues $d_1, \ldots, d_p$ of $\mathbf{V}_{1,n}^{-1}\mathbf{V}_{2,n}$ in decreasing order and $\mathbf{B}_n = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ the corresponding eigenvectors as its rows s.t: $\mathbf{B}_n\mathbf{V}_{1,n}\mathbf{B}_n' = \mathbf{I}_p$ and $\mathbf{B}_n\mathbf{V}_{2,n}\mathbf{B}_n' = \mathbf{D}_n$.

Principle, (Tyler et al., 2009)

Simultaneous diagonalization of two symmetric and positive definite scatter matrices, $\mathbf{V}_{1,n}$ and $\mathbf{V}_{2,n}$ s.t:

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

where the diagonal matrix $\mathbf{D}_n$ contains the eigenvalues $d_1, \ldots, d_p$ of $\mathbf{V}_{1,n}^{-1}\mathbf{V}_{2,n}$ in decreasing order and $\mathbf{B}_n = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ the corresponding eigenvectors as its rows s.t: $\mathbf{B}_n\mathbf{V}_{1,n}\mathbf{B}_n' = \mathbf{I}_p$ and $\mathbf{B}_n\mathbf{V}_{2,n}\mathbf{B}_n' = \mathbf{D}_n$.

The invariant coordinates are $\mathbf{Z}_n = (\mathbf{X}_n - \mathbf{1}_n\mathbf{m}_{1,n}'(\mathbf{X}_n))\mathbf{B}_n'$,
with $\mathbf{1}_n$ a vector of n-ones, $\mathbf{m}_{1,n}(\mathbf{X}_n)$ the location estimator associated with $\mathbf{V}_{1,n}(\mathbf{X}_n)$.
Let $Z_{n,k}$ be the first k components of Z_n.

Principle, (Tyler et al., 2009)

Simultaneous diagonalization of two symmetric and positive definite scatter matrices, $\mathbf{V}_{1,n}$ and $\mathbf{V}_{2,n}$ s.t:

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

where the diagonal matrix $\mathbf{D}_n$ contains the eigenvalues $d_1, \ldots, d_p$ of $\mathbf{V}_{1,n}^{-1}\mathbf{V}_{2,n}$ in decreasing order and $\mathbf{B}_n = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ the corresponding eigenvectors as its rows s.t: $\mathbf{B}_n\mathbf{V}_{1,n}\mathbf{B}_n' = \mathbf{I}_p$ and $\mathbf{B}_n\mathbf{V}_{2,n}\mathbf{B}_n' = \mathbf{D}_n$.

The invariant coordinates are $\mathbf{Z}_n = (\mathbf{X}_n - \mathbf{1}_n\mathbf{m}_{1,n}'(\mathbf{X}_n))\mathbf{B}_n'$,
with $\mathbf{1}_n$ a vector of n-ones, $\mathbf{m}_{1,n}(\mathbf{X}_n)$ the location estimator associated with $\mathbf{V}_{1,n}(\mathbf{X}_n)$.
Let $Z_{n,k}$ be the first k components of Z_n.

Property: Equivalence with the MD^2 if $k = p$: $||Z_{n,k}||^2 = \mathrm{MD}^2_{V_{1,n}}(\mathbf{x}_i)$.

Principle, (Tyler et al., 2009)

Simultaneous diagonalization of two symmetric and positive definite scatter matrices, $\mathbf{V}_{1,n}$ and $\mathbf{V}_{2,n}$ s.t:

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

where the diagonal matrix $\mathbf{D}_n$ contains the eigenvalues $d_1, \ldots, d_p$ of $\mathbf{V}_{1,n}^{-1}\mathbf{V}_{2,n}$ in decreasing order and $\mathbf{B}_n = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ the corresponding eigenvectors as its rows s.t: $\mathbf{B}_n\mathbf{V}_{1,n}\mathbf{B}_n' = \mathbf{I}_p$ and $\mathbf{B}_n\mathbf{V}_{2,n}\mathbf{B}_n' = \mathbf{D}_n$.

The invariant coordinates are $\mathbf{Z}_n = (\mathbf{X}_n - \mathbf{1}_n\mathbf{m}_{1,n}'(\mathbf{X}_n))\mathbf{B}_n'$,
with $\mathbf{1}_n$ a vector of n-ones, $\mathbf{m}_{1,n}(\mathbf{X}_n)$ the location estimator associated with $\mathbf{V}_{1,n}(\mathbf{X}_n)$.
Let $Z_{n,k}$ be the first k components of Z_n.

Property: Equivalence with the MD^2 if $k = p$: $||Z_{n,k}||^2 = \mathrm{MD}^2_{V_{1,n}}(\mathbf{x}_i)$.
The added value of ICS is to select the subspace of $k = q < p$ components in which the structure of the outliers can be detected.

4 steps, (Archimbaud et al., 2016)

Choice of the scatter matrices, $\mathbf{V}_{1,n}$ and $\mathbf{V}_{2,n}$

Maximization of the kurtosis measure (Tyler et al., 2009), Caussinus and Ruiz-Gazen (1993), Caussinus et al. (2003) and Peña and Prieto (2001): $\mathbf{V}_{1,n} = \mathrm{COV}(\mathbf{X}_n)$ and $\mathbf{V}_{2,n} = \mathrm{COV}_4(\mathbf{X}_n)$.

4 steps, (Archimbaud et al., 2016)

Choice of the scatter matrices, $\mathbf{V}_{1,n}$ and $\mathbf{V}_{2,n}$

Maximization of the kurtosis measure (Tyler et al., 2009), Caussinus and Ruiz-Gazen (1993), Caussinus et al. (2003) and Peña and Prieto (2001): $\mathbf{V}_{1,n} = \mathrm{COV}(\mathbf{X}_n)$ and $\mathbf{V}_{2,n} = \mathrm{COV}_4(\mathbf{X}_n)$.

Selection of the Invariant Coordinates

Selecting only the k components of interest for outlier detection.

4 steps, (Archimbaud et al., 2016)

Choice of the scatter matrices, $\mathbf{V}_{1,n}$ and $\mathbf{V}_{2,n}$

Maximization of the kurtosis measure (Tyler et al., 2009), Caussinus and Ruiz-Gazen (1993), Caussinus et al. (2003) and Peña and Prieto (2001): $\mathbf{V}_{1,n} = \mathrm{COV}(\mathbf{X}_n)$ and $\mathbf{V}_{2,n} = \mathrm{COV}_4(\mathbf{X}_n)$.

Selection of the Invariant Coordinates

Selecting only the k components of interest for outlier detection.

Measure of outlierness

For an observation $\mathbf{x}_i$: $ICSD^2_{\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)}(\mathbf{x}_i, k) = ||\mathbf{Z}_{n,k}||^2$

4 steps, (Archimbaud et al., 2016)

Choice of the scatter matrices, $\mathbf{V}_{1,n}$ and $\mathbf{V}_{2,n}$

Maximization of the kurtosis measure (Tyler et al., 2009), Caussinus and Ruiz-Gazen (1993), Caussinus et al. (2003) and Peña and Prieto (2001): $\mathbf{V}_{1,n} = \mathrm{COV}(\mathbf{X}_n)$ and $\mathbf{V}_{2,n} = \mathrm{COV}_4(\mathbf{X}_n)$.

Selection of the Invariant Coordinates

Selecting only the k components of interest for outlier detection.

Measure of outlierness

For an observation $\mathbf{x}_i$: $ICSD^2_{\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)}(\mathbf{x}_i, k) = ||\mathbf{Z}_{n,k}||^2$

Identification of the outliers

If $ICSD^2_{\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)}(\mathbf{x}_i, k) > c_{1-\beta}$, with $c_{1-\beta}$ the cut-off derived from *mDist* Monte Carlo simulations of a standard normal population.

Selection of the Invariant Coordinates

Objectives

Developing an automated test as in Caussinus et al. (2003) which selects only the components of interest for outlier detection for any pair of scatters.

Selection of the Invariant Coordinates

Objectives

Developing an automated test as in Caussinus et al. (2003) which selects only the components of interest for outlier detection for any pair of scatters.

Quite challenging because the distribution of the eigenvalues is unknown for general pair of scatters. (see Nordhausen et al. (2017) for $\mathrm{COV} - \mathrm{COV}_4$.)

Selection of the Invariant Coordinates

Objectives

Developing an automated test as in Caussinus et al. (2003) which selects only the components of interest for outlier detection for any pair of scatters.
Quite challenging because the distribution of the eigenvalues is unknown for general pair of scatters. (see Nordhausen et al. (2017) for $\mathrm{COV} - \mathrm{COV}_4$.)

Approaches

- A test based on a quasi inferential procedure (simulations).
- Some normality tests.

Selection of the Invariant Coordinates

Objectives

Developing an automated test as in Caussinus et al. (2003) which selects only the components of interest for outlier detection for any pair of scatters.
Quite challenging because the distribution of the eigenvalues is unknown for general pair of scatters. (see Nordhausen et al. (2017) for $\mathrm{COV} - \mathrm{COV}_4$.)

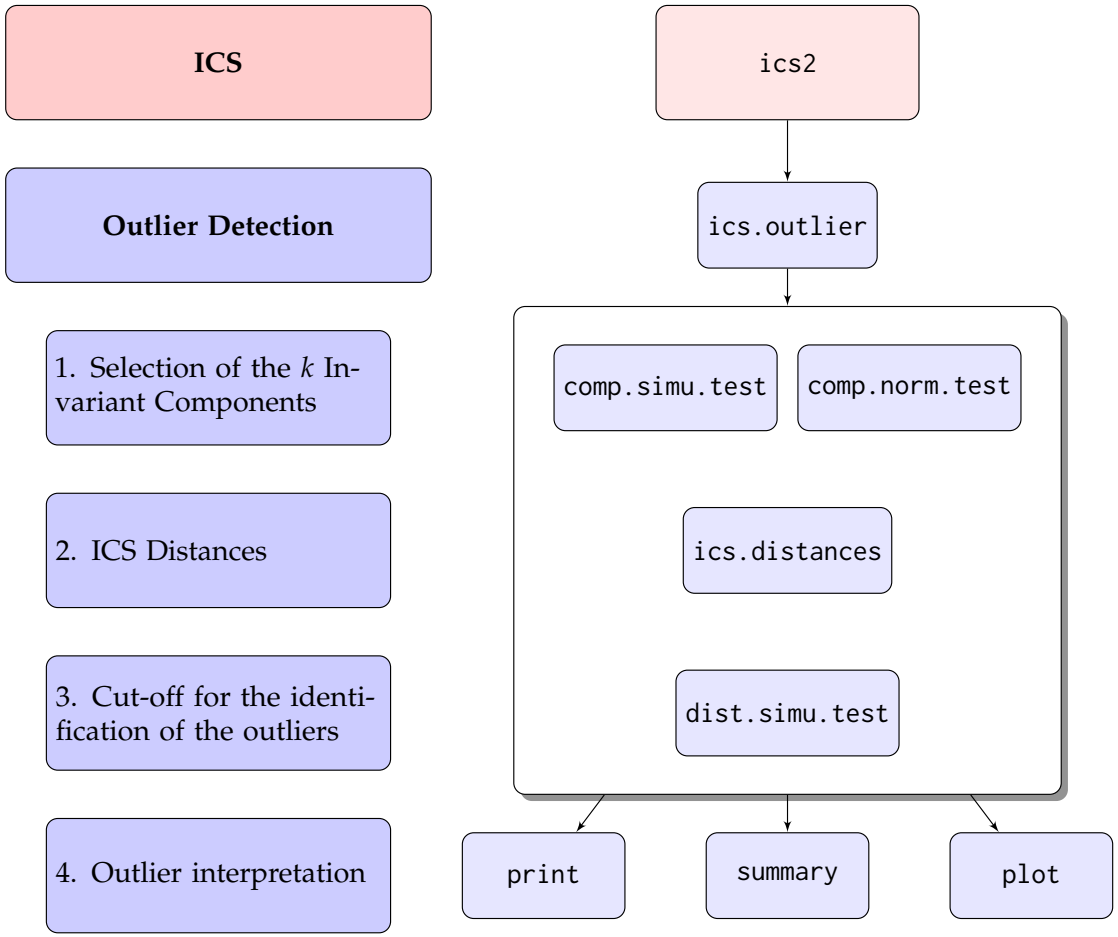
Approaches

- A test based on a quasi inferential procedure (simulations).
- Some 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 $\alpha = 5\%$ (see Dray (2008)).

Table of Contents

OverView of the ICSOutlier package



In pink: function from the ICS package.
In blue: functions from the ICSOutlier package.

Arguments:

- method
- test
- mEig
- level.test
- adjust
- level.dist
- mDist

Example of no outlier

Model:

$$\mathbf{X} \sim \mathcal{N}_2(\mathbf{0}, \mathrm{diag}(0.1)).$$

Example of no outlier

Model:

$$\mathbf{X} \sim \mathcal{N}_2(\mathbf{0}, \mathrm{diag}(0.1)).$$

R:

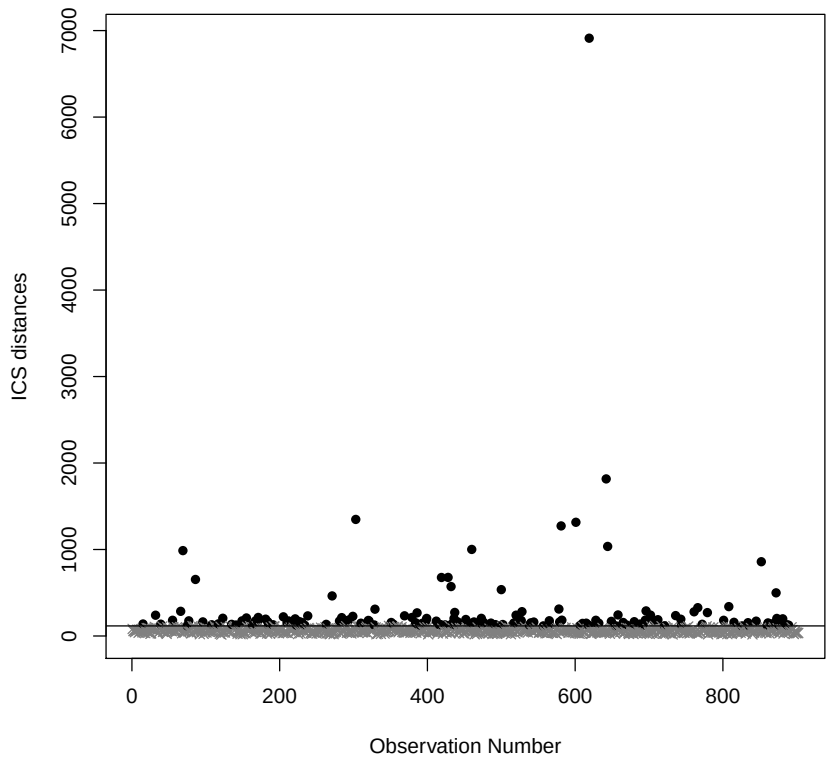
```
library(ICSOutlier)


## Loading required package:  ICS
## Loading required package:  mvtnorm
## Loading required package:  moments

# Data simulation
set.seed(123)
X <- matrix(rnorm(1000, 0, 0.1), 500, 2)

# default ICS
icsX <- ics2(X)

# Outlier Detection
icsOutlierDefault <- ics.outlier(icsX)
print(icsOutlierDefault)


## [1] "0 components were selected and no outliers were detected."
```

Example of the HTP data set: default parameters

```
# HTP dataset
library("ICSOutlier")
set.seed(123)
data(HTP)
outliers <- c(581, 619)

# default ICS
icsHTP <- ics2(HTP)

# Outlier detection with selection of components based on D'Agostino test
# by default it can take quite long as mDist = 10000
icsOutlierDA <- ics.outlier(icsHTP, method = "norm.test", test = "agostino.test",
                            level.test = 0.05, adjust = TRUE,
                            level.dist = 0.025, mDist = 10000)
summary(icsOutlierDA)


##
## ICS based on two scatter matrices and two location estimates
## S1:  MeanCov
## S2:  Mean3Cov4
##
## Searching for a small proportion of outliers
##
## Components selected at nominal level 0.05: 14
## Selection method: norm.test (agostino.test)
## Number of outliers at nominal level 0.025: 43
```

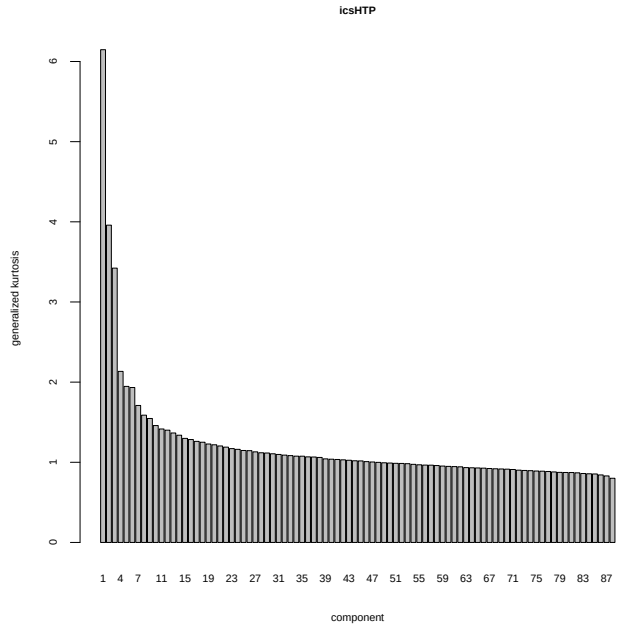
Example of the HTP data set: different parameters

```
# ICS with MLE Cauchy and the Mean-Cov
icsHTP2 <- ics2(HTP, S1 = tM, S2 = MeanCov, S1args = list(df = 1))

# Outlier detection with selection of components based on simulations
icsOutlierPA <- ics.outlier(icsHTP2, method = "simulation", level.test = 0.05, mEig = 5000,
                            level.dist = 0.01, mDist = 5000)
plot(icsOutlierPA)
```



Example of the HTP data set: screeplot I

```
# Choice of scatter based on screeplot
screeplot(icsHTP, cex.lab = 0.7, cex.axis = 0.7, cex.names = 0.7, cex.main = 0.7)
```




```
# ICS Distances
ics.dist.scree <- ics.distances(icsHTP, index = 1:3)
# Cut-off ICS Distances
ics.cutOff <- dist.simu.test(icsHTP, 1:3, m = 10000, level = 0.025)
```

Example of the HTP data set: screeplot II

```
plot(ics.dist.scree, col = colPoints, pch = pchPoints,
     cex.lab = 0.7, cex.axis = 0.7, cex.main = 0.7, ylab = ylabICS)
points(outliers, ics.dist.scree[outliers], pch = 24, bg = 1)
text(outliers, ics.dist.scree[outliers], labels = outliers, pos = 2, cex = 0.7)
abline(h = ics.cutOff)
```

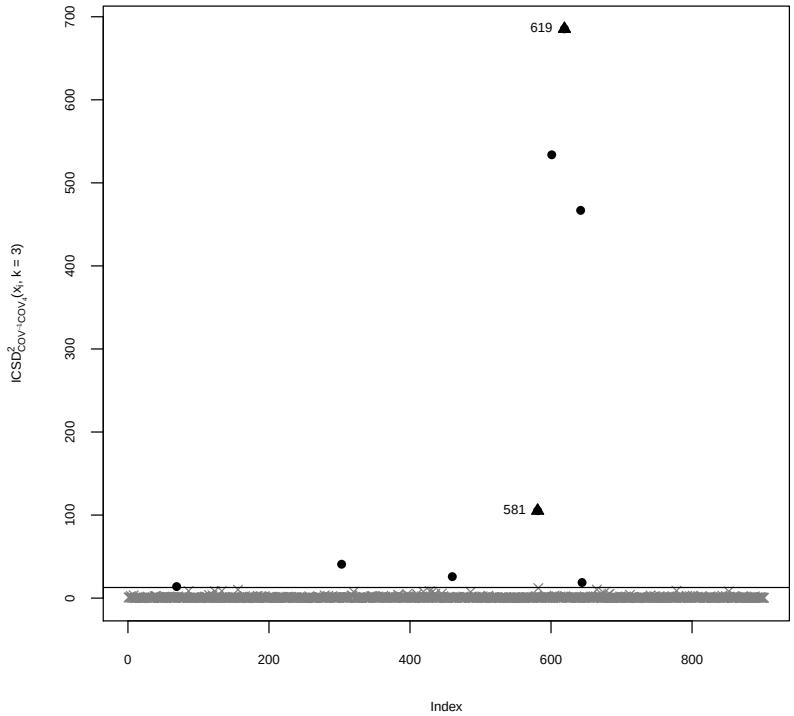


Table of Contents

Conclusion and Perspectives

Conclusion

- New method for identifying outliers lying on a subspace: ICS.
- Good results of the ICS method for outlier detection (Archimbaud et al., 2016) on simulated and real data sets.
- ICSOutlier package provides a full implementation which can be used out of the box.

Conclusion and Perspectives

Conclusion

- New method for identifying outliers lying on a subspace: ICS.
- Good results of the ICS method for outlier detection (Archimbaud et al., 2016) on simulated and real data sets.
- ICSOutlier package provides a full implementation which can be used out of the box.

Perspectives

- Developping a Shiny Application dedicated to exploratory purpose using the ICS and ICSOutlier packages.
- Extending the theory and package to be able to handle also large fractions of outliers.
- Extending the theory and package to be able to handle also high-dimensional/low sample size data.

References

Archimbaud, A., Nordhausen, K., and Ruiz-Gazen, A. (2016). Multivariate outlier detection with ICS. *arXiv preprint arXiv:1612.06118v2*.

Bonett, D. G. and Seier, E. (2002). A test of normality with high uniform power. *Computational Statistics and Data Analysis*, 40(3):435–445.

Caussinus, H., Fekri, M., Hakam, S., and Ruiz-Gazen, A. (2003). A monitoring display of multivariate outliers. *Computational Statistics & Data Analysis*, 44(1):237–252.

Caussinus, H. and Ruiz-Gazen, A. (1993). Projection pursuit and generalized principal component analysis. *New Directions in Statistical Data Analysis and Robustness*, pages 35–46.

Dray, S. (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational Statistics and Data Analysis*, 52(4):2228 – 2237.

Nordhausen, K., Oja, H., Tyler, D. E., and Virta, J. (2017). Asymptotic and bootstrap tests for the dimension of the non-gaussian subspace. *arXiv preprint arXiv:1701.06836*.

Peña, D. and Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43(3).

Peres-Neto, P. R., Jackson, D. A., and Somers, K. M. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, 49(4):974 – 997.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.

Tyler, D. E., Critchley, F., Dümbgen, L., and Oja, H. (2009). Invariant coordinate selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):549–592.

Yazici, B. and Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, 77(2):175–183.

Behavior of the Mahalanobis distance in large dimension

Model: $\mathbf{X} \sim (1-\epsilon)\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_W) + \sum_{h=1}^{q} \epsilon_h \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_W)$, where $\epsilon = \sum_{h=1}^{q} \epsilon_h < \frac{1}{2}$.

Proposition (Archimbaud et al., 2016)

(Using the Lindeberg-Feller central limit.)
Assuming that q is fixed and p becomes large, the distribution of the differences, where E denotes the expectation:

$$\frac{1}{2\sqrt{p}} \left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no}) - E\left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no}) \right) \right),$$
$$\frac{1}{2\sqrt{p}} \left(d_R^2(\mathbf{X}_{o,h}) - d_R^2(\mathbf{X}_{no}) - \boldsymbol{\mu}_h' \boldsymbol{\mu}_h \right)$$

converge to a standard Gaussian distribution.
The expectations $E\left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no}) \right)$ and $\boldsymbol{\mu}_h' \boldsymbol{\mu}_h$ do not depend on p.

Notations: with $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\Sigma}_q, \mathbf{I}_{p-q})$,
$d^2(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) = (\mathbf{X}_q - \boldsymbol{\mu}_{\mathbf{X}_q})' \boldsymbol{\Sigma}_q^{-1} (\mathbf{X}_q - \boldsymbol{\mu}_{\mathbf{X}_q}) + \sum_{i=q+1}^{p} X_i^2$
$d_R^2(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_0)' {\boldsymbol{\Sigma}_W}^{-1} (\mathbf{X} - \boldsymbol{\mu}_0) = \sum_{i=1}^{p} X_i^2$.
"Non-outlier" observations $\mathbf{X}_{no} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$,
"outlier" observations $\mathbf{X}_{o,h} \sim \mathcal{N}(\boldsymbol{\mu}_h, \mathbf{I}_p)$, with the last $p - q$ coordinates of $\boldsymbol{\mu}_h$ equal to 0 and $h = 1, \ldots, q$.
$\mathbf{X}_{no} \perp \mathbf{X}_{o,h}$.

Parallel Analysis (PA)

As in Peres-Neto et al. (2005) for PCA or in Caussinus et al. (2003).

Computation of cut-offs:

- 10 000 simulations of $\mathcal{N}_n(0, I_p)$
- ICS with $\mathrm{COV}(\mathbf{X}_n)^{-1}\mathrm{COV}_4(\mathbf{X}_n)$
- Quantiles of the ICS eigenvalues at level $1 - \frac{\alpha}{i}$ for each component i, with $\alpha = 5\%$ and $i = 1, \ldots, p$.

Parallel Analysis (PA)

As in Peres-Neto et al. (2005) for PCA or in Caussinus et al. (2003).

Computation of cut-offs:

- 10 000 simulations of $\mathcal{N}_n(0, I_p)$
- ICS with $\mathrm{COV}(\mathbf{X}_n)^{-1}\mathrm{COV}_4(\mathbf{X}_n)$
- Quantiles of the ICS eigenvalues at level $1 - \frac{\alpha}{i}$ for each component i, with $\alpha = 5\%$ and $i = 1, \ldots, p$.

Test:

- Sequentially testing if the ICS eigenvalues are higher than corresponding quantiles.
- Stop as soon as one is lower than the cut-off.

Normality Tests

Identifying the last i^{th} invariant coordinates that are non normal using univariate normality tests (Yazici and Yolacan (2007) and Bonett and Seier (2002)).

Normality Tests

Identifying the last i^{th} invariant coordinates that are non normal using univariate normality tests (Yazici and Yolacan (2007) and Bonett and Seier (2002)).

Normality tests:

- The D'Agostino test of skewness (DA),
- The Anscombe-Glynn (AG) test of kurtosis,
- The Bonett-Seier (BS) test of Geary's kurtosis,
- The Jarque-Bera (JB) test for normality which is based on both skewness and kurtosis measures,
- The Shapiro-Wilk (SW) normality test.

Normality Tests

Identifying the last i^{th} invariant coordinates that are non normal using univariate normality tests (Yazici and Yolacan (2007) and Bonett and Seier (2002)).

Normality tests:

- The D'Agostino test of skewness (DA),
- The Anscombe-Glynn (AG) test of kurtosis,
- The Bonett-Seier (BS) test of Geary's kurtosis,
- The Jarque-Bera (JB) test for normality which is based on both skewness and kurtosis measures,
- The Shapiro-Wilk (SW) normality test.

Stopping rule

- Sequentially testing if the Invariant Coordinates are gaussian or not.
- Stop as soon as one is gaussian based on the corrected level of 5%.

Real Applications

	Glass			Reliability			HighTech		
	TP (/3)	FP (/109)	k (/11)	TP (/2)	FP (/518)	k (/55)	TP (/2)	FP (/900)	k (/88)
MD	**3**	**4**		**2**	**52**		**2**	**119**	
RD	3	15					2	243	
Best selection									
ICS $\mathrm{COV} - \mathrm{COV}_4$	3	3	2	2	1	1	2	0	1
PCA	3	9	5	2	41	52	2	21	1
PCA std	3	4	2	2	22	40	2	25	6
ROBPCA	3	13	5				2	50	1
Automated selection									
ICS $\mathrm{COV} - \mathrm{COV}_4$ DA	3	3	2	2	23	12	2	39	14
ICS $\mathrm{COV} - \mathrm{COV}_4$ PA	3	3	2	2	42	28	2	87	50
PCA	1	5	1	0	6	12	2	24	3
PCA std	1	4	1	2	31	20	2	28	4
ROBPCA	3	17	1				2	80	2
Scree plot selection									
ICS $\mathrm{COV} - \mathrm{COV}_4$	**3**	**3**	2	**2**	**1**	2	**2**	**5**	3

Table: TP, FP and number k of selected components for the three real data examples.