

Fortifying Statistical Analyses: Software Tools for Robust Methods

V. Todorov¹

¹United Nations Industrial Development Organization (UNIDO)

ICORS 2025

Stresa, 22 May, 2025

Outline

- 1 Introduction
- 2 R packages on CRAN
- 3 Multivariate location and scatter
- 4 Principal Component Analysis (PCA)
- 5 Other languages
- 6 AI assisted data analysis
- 7 Summary and conclusions

Outline

- 1 Introduction
- 2 R packages on CRAN
- 3 Multivariate location and scatter
- 4 Principal Component Analysis (PCA)
- 5 Other languages
- 6 AI assisted data analysis
- 7 Summary and conclusions

Software for robust statistics

Software for robust statistics

- The practical deployment and success of robust methods is unthinkable without the availability of reliable and easy-to-use software
- The computations involved in applying most robust techniques are considerable, and their routine use requires a suitable software implementation
- This issue was recognized already in the past (Stromberg 2004; Morgenthaler 2007)
- Many of the initially developed software for robust statistics appeared in SAS, S-Plus, and MATLAB

Software for robust statistics

Software for robust statistics–MATLAB

- Currently two MATLAB toolboxes for robust statistics are popular:
 - ▶ **LIBRA**, developed by the research groups in robust statistics of the Katholieke Universiteit Leuven (Department of Mathematics) and the University of Antwerp (Department of Mathematics and Computer Science) (Verboven and Hubert, 2005, 2010) and
 - ▶ **FSDA** developed jointly by the University of Parma and the Joint Research Center (JRC) of the European Commission (Riani, Perrotta, and Torti 2012)

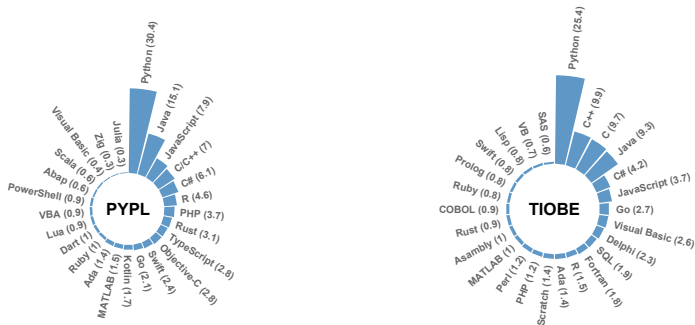
Software for robust statistics–R

Software for robust statistics–R

- It seems that the free software environment for statistical computing and graphics R(R Core Team, 2024) is a viable option giving the developers and users many possibilities to develop and apply robust methods
- The usage of R is extensive and it is becoming one of the most popular software in scientific research.
- It provides interactive data analysis tools as well as high-quality graphics for different fields in scientific research with tools for exploratory data analysis as well as tools for robust diagnostics

Software for robust statistics–R

- According to the **PYPL (Popularity of Programming Language)** index, <https://pypi.github.io/PYPL.html>, R ranked sixth in May 2025
- According to the **TIOBE Popularity Index**, <https://www.tiobe.com>, also for May 2025, R ranked 12th.



Outline

- 1 Introduction
- 2 R packages on CRAN
- 3 Multivariate location and scatter
- 4 Principal Component Analysis (PCA)
- 5 Other languages
- 6 AI assisted data analysis
- 7 Summary and conclusions

About CRAN

CRAN

- The **Comprehensive R Archive Network (CRAN)**: R's central software repository
- Supported by the R Foundation
- Contains an archive of the latest and previous versions of the R distribution, documentation, and contributed R packages.
- As of today, more than **22,400** packages are available.
- CRAN was created by **Kurt Hornik** and **Friedrich Leisch** in 1997 (the name paralleling other early packing systems such as TeX's CTAN and Perl's CPAN).
- Maintained by Hornik and a team of volunteers
- Located at the Vienna University of Economics and Business, mirrored on servers around the world

About CRAN

Importance of CRAN

- The large number of packages available for R, and the ease of installing and using them: a major factor driving the widespread adoption of R
- **Quality assurance:**
 - ▶ CRAN offers guidelines as to package content, structure, documentation, and compatibility: R packages must conform to a relatively strict specification and a standard directory structure
 - ▶ CRAN provides a screening and monitoring process (for initial submission as well as during the package life cycle)
 - ▶ This procedure will remove packages that are no longer up-to-date or not maintained properly.
- **Version management:** A comprehensive history of package versions is maintained , allowing users to access and install specific versions of packages (reproducibility)
- **Community Collaboration:** CRAN fosters a vibrant community of R developers, maintainers, and users

About CRAN

Task Views

- Provide guidance which packages on CRAN are relevant for tasks related to a certain topic
- List a wide range of tasks in fields such as finance, genetics, high performance computing, machine learning, medical imaging, meta-analysis, social sciences, spatial statistics, robust statistics, official statistics and survey methodology
- CRAN Task View: Robust Statistical Methods
 - ▶ Maintained by Martin Maechler
 - ▶ Regression, multivariate analysis, clustering, time series, descriptive statistics and others

Other repositories

Other repositories

- **Bioconductor** project (Gentleman et al. 2004): specialized on collaborative development of software for computational biology and bioinformatics.
- **GitHub**: Developers can also share their R packages through large code-sharing sites like GitHub and GitLab: fewer restrictions on the format or content of shared code compared to CRAN and Bioconductor.

About CRAN

Finding R packages

- Look regularly, say, once a month, at [CRANberries - Dirk Eddelbuettel](#): aggregates information about new, updated and removed packages from the CRAN
- Function `findPackage()` from package [packagefinder](#): allows a keyword search through the metadata of all CRAN packages
- Package [pkgsearch](#): uses an R-hub web service and a careful ranking that puts popular packages before less frequently used one

About CRAN

```
> library(pkgsearch)
> ps("robust", size=25)
- "robust" ----- 722 packages in 0.008 seconds -
```

#	package	version	by	@ title
1	100 jsonlite	2.0.0	Jeroen Ooms	1M A Simple and Robust JSON...
2	46 sandwich	3.1.1	Achim Zeileis	8M Robust Covariance Matrix...
3	44 robustbase	0.99.4.1	Martin Maechler	8M Basic Robust Statistics
4	22 robust	0.7.5	Valentin Todorov	9M Port of the S+ "Robust L...
5	21 golem	0.5.1	Colin Fay	9M A Framework for Robust S...
6	19 pcaPP	2.0.5	Valentin Todorov	9M Robust PCA by Projection...
7	11 rrcov	1.7.7	Valentin Todorov	19d Scalable Robust Estimato...
8	10 RobStatTM	1.0.11	Matias Salibian-Barrera	7M Robust Statistics: Theor...
9	9 clubSandwich	0.6.0	James E. Pustejovsky	1M Cluster-Robust (Sandwich...
10	9 WRS2	1.1.6	Patrick Mair	1y A Collection of Robust S...
11	9 tclust	2.1.0	Valentin Todorov	3M Robust Trimmed Clustering
12	8 varhandle	2.0.6	Mehrad Mahmoudian	2y Functions for Robust Var...
13	8 robustHD	0.8.1	Andreas Alfons	10M Robust Methods for High-...
14	7 robumeta	2.1	Zachary Fisher	2y Robust Variance Meta-Reg...
15	7 RobAStBase	1.2.7	Matthias Kohl	4M Robust Asymptotic Statis...
16	7 mvoutlier	2.1.1	P. Filzmoser	4y Multivariate Outlier Det...
17	7 rdrobust	2.2	Sebastian Calonico	2y Robust Data-Driven Stati...
18	7 ROptEst	1.3.5	Matthias Kohl	4M Optimally Robust Estimation
19	6 Gmedian	1.2.7	Herve Cardot	3y Geometric Median, k-Medi...
20	6 robustX	1.2.7	Martin Maechler	2y 'eXtra' / 'eXperimental'...
21	6 GSE	4.2.1	Claudio Agostinelli	2y Robust Estimation in the...
22	5 genieclust	1.1.6	Marek Gagolewski	9M Fast and Robust Hierarch...
23	5 ramcmc	0.1.2	Jouni Helske	4y Robust Adaptive Metropol...
24	5 robslopes	1.1.3	Jakob Raymaekers	2y Fast Algorithms for Robu...
25	5 robcor	0.1.6.1	Paul Smirnov	3y Robust Correlations

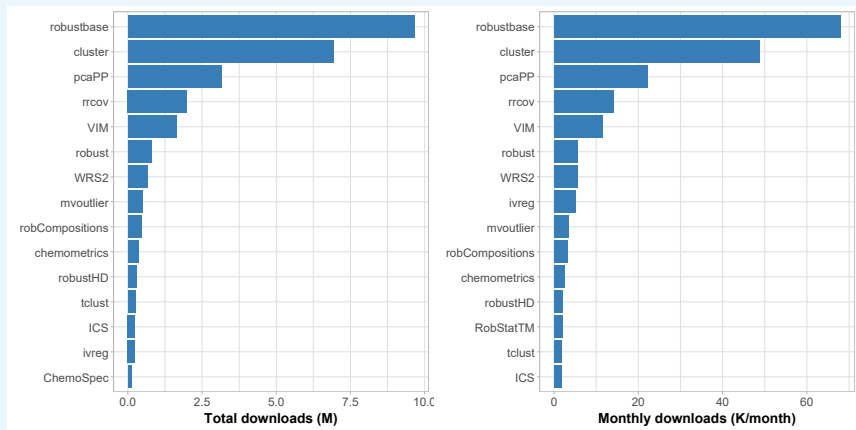
```
> |
```

About CRAN

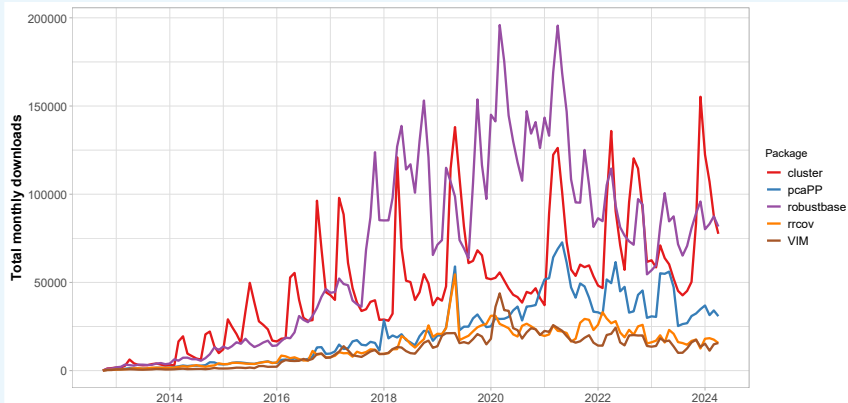
Evaluating R packages

- The package is on CRAN or Bioconductor: It has already passed more or less quality checks
- Availability and quality of help: Beyond the minimal required documentation a package could have interactive tutorials, vignettes, pkgdown website
- Metrics to quantify how established the package is:
 - ▶ Maturity (when was the package published for the first time)
 - ▶ Number of downloads—total downloads and average downloads per month
 - ▶ Number of reverse dependencies
- The evidence of peer acceptance and review
- Who developed it
- How was it developed

Robust statistical methods on CRAN



Robust statistical methods on CRAN



Robust statistical methods on CRAN

Robust statistical methods on CRAN

- Some robust methods were available in the R environment already from the very beginning:
 - ▶ The base package **stats** has functions for
 - ▶ computing robust location and scale (`median()`, `mean(x, trim = alpha)`, `mad()`, `IQR()`)
 - ▶ `fivenum()` computing the statistics behind the box plot
 - ▶ `runmed()` for computing the running median used for robust scatter plot smoothing
 - ▶ The base package **graphics**: boxplots for outlier detection and `lowess()` for robust nonparametric regression
 - ▶ More advanced robust methods are implemented in the recommended package **MASS** (Venables and Ripley 2003):
 - ▶ `rlm()` for robust regression and
 - ▶ `cov.rob()` for robust estimation of multivariate location and covariance

CRAN packages for robust statistical analysis

#	Package	Category	Description	Author(s)	Year	R.Deps
1	cluster	Cluster analysis	"Finding Groups in Data" extended	M. Maechler, P.Rousseeuw, and others	1999	247
2	rrcov	Multivariate	Robust estimators with high breakdown point	V. Todorov	2004	31
3	mvoutlier	Multivariate	Multivariate outlier detection based on robust methods	P. Filzmoser and M. Gschwandtner	2004	1
4	pcaPP	Multivariate	Robust PCA by projection pursuit	P. Filzmoser, H. Fritz, and K. Kalcher	2006	29
5	robustbase	Basics, regression	Basic robust statistics	M. Maechler, P. Rousseeuw, C. Croux, V. Todorov, and others	2006	138
6	robust	Regression, multivariate	Port of the S+ "Robust Library"	J. Wang, R. Zamar, A. Marazzi, and others	2006	6
7	ICS	Multivariate	Tools for exploring multivariate data via ICS/ICA	K. Nordhausen, A. Alfons, and others	2007	6

CRAN packages for robust statistical analysis (II)

8	robCompositions	Compositional	Methods for analysis of compositional data including robust methods	M. Templ and others	2009	2
9	tclust	Cluster analysis	Provides functions for robust trimmed clustering	V. Todorov, A. M. Iscar, L. G. Escudero, and others	2010	3
10	rrcovNA	Missing data	Robust location and scatter estimation and robust multivariate analysis with high breakdown point for incomplete data	V. Todorov	2010	0
11	robustHD	High dimensions	Robust regression with high-dimensional data	A. Alfons	2012	6
12	rrcovHD	High dimensions	Robust multivariate methods for high dimensional data	V. Todorov	2012	0
13	VIM	Missing data	Visualization and imputation of missing values	M. Templ, A. Kowarik, A. Alfons, and others	2012	11

CRAN packages for robust statistical analysis (III)

14	GSE	Missing data	Robust estimation in the presence of cellwise and casewise contamination and missing data	A. Leung, M. Danilov, V. Yohai, and R. Zamar	2013	4
15	WRS2	Collections of several methodologies	A collection of robust statistical methods	P. Mair and R. Wilcox	2014	5
16	ICSOutlier	Multivariate	Outlier detection using invariant coordinate selection	K. Nordhausen, A. Archimbaud, and A. Ruiz-Gazen	2016	1
17	cellWise	Multivariate	Analyzing data with cellwise outliers	J. Raymaekers, P. Rousseeuw, and others	2016	10
18	rrcov3way	Multiway	Provides methods for multiway data analysis by means of (robust) Parafac and Tucker3 models	V. Todorov, V. Simonacci, M. A. Di Palma, and M. Gallo	2017	0
19	pense	High dimensions	Robust penalized (adaptive) elastic net S- and M-estimators for linear regression	D. Kepplinger, M. Salibian-Barrera, and G. C. Freue	2017	0

CRAN packages for robust statistical analysis (IV)

20	rospca	Sparse PCA	Robust sparse PCA using the ROSPCA algorithm	T. Reynkens, V. Todorov, M. Hubert, and others	2017	0
21	fsdaR	Regression and multivariate	Robust data analysis through monitoring and dynamic visualization	V. Todorov, E. Sordini, and others	2017	0
22	mrfDepth	Multivariate	Depth measures in multivariate, regression, and functional settings	P. Segaert, M. Hubert, P. Rousseeuw and others	2017	1
23	ltsspca	Sparse PCA	Sparse principal component based on least trimmed squares	Y. Wang, S. Van Aelst, and others	2019	0
24	RobStatTM	Regression, multivariate	Robust statistics: theory and methods	M. Salibián-Barrera, V. Yohai, R. Maronna, and others	2019	6
25	classmap	Cluster analysis	Visualizing classification results	J. Raymaekers, P. Rousseeuw	2021	0

Robust tools for specific application fields

- Chemometrics:
 - ▶ **chemometrics**, a companion package to the book of Varmuza and Filzmoser (2009), multivariate analysis of chemical data, mostly for regression, classification, and model evaluation.
 - ▶ **ChemoSpec**: implements a function `r_pcaSpectra()` for robust PCA for spectra objects (a wrapper around `PCAGrid()` from the package **pcaPP**)
- Econometrics
 - ▶ **ivreg** implements instrumental variable estimation for linear models by two-stage least-squares (2SLS). Uses MM-regression instead of OLS in both stages of the procedure (based on `rlm()` of **MASS**)
 - ▶ **riv**: robust instrumental variables estimator using a high breakdown point S-estimator of multivariate location and scatter matrix

Robust tools for specific application fields

- Social, behavioral, and medical sciences
 - ▶ **robmed** (Alfons, Ateş, and Groenen 2022) implements a robust procedure for mediation analysis based on the fast-and-robust bootstrap methodology for robust regression estimators
 - ▶ **WRS2** mediation analysis via the bootstrap test after data cleaning
 - ▶ **RobLoxBioC** (Kohl and Deigner 2010) provides functions for the determination of optimally robust influence curves and estimators for preprocessing omics data
- Official statistics
 - ▶ **robsurvey** implements robust estimators of finite population characteristics like means, totals, ratios, regression
 - ▶ **modi** provides algorithms for multivariate outlier detection when missing values occur
 - ▶ **rsae** contains empirical best linear unbiased prediction (EBLUP) and robust prediction of the area-level means under the basic unit-level model
 - ▶ **VIM** (Kowarik and Templ 2016)

Robust tools for specific application fields

- Time series analysis
 - ▶ **robfilter** implements several robust procedures that allow for (online) extraction of the signal of univariate or multivariate time series by applying robust regression techniques to a moving time window
 - ▶ **RobPer**: calculates periodograms based on robust fitting periodic functions to light curves (irregularly observed time series, possibly with measurement accuracies, occurring in astroparticle physics)
 - ▶ **robustarima**: Functions for fitting a linear regression model with ARIMA errors using a filtered tau-estimate. The methodology is described in Maronna et al. (2019)

New packages on CRAN

New packages on CRAN

- Looking for new packages on CRAN: [CRANberries - Dirk Eddelbuettel](#): aggregates information about new, updated and removed packages from the CRAN
- New packages (on CRAN after 1. April 2025)
 - ▶ [RANSAC](#): Robust Model Fitting Using the RANSAC (Random Sample Consensus) Algorithm (Abreu, 2025)
 - ▶ [spareg](#): A combination of variable screening and random projection techniques for fitting ensembles of predictive generalized linear models to high-dimensional data (Vana-Gür L, Parzer R and Filzmoser P, 2025)
 - ▶ [robcat](#): Robust categorical data analysis based on the theory of C-estimation (Welz M, Alfons A and Mair P, 2025)
 - ▶ [robustGarch](#): Modeling robust generalized autoregressive conditional heteroskedasticity (Garch) (1,1) processes (Liu E, Xia D, Martin RD, 2025)

The R Package Ecosystem for Robust Statistics

Wiley Interdisciplinary Reviews: Computational Statistics

WILEY



FOCUS ARTICLE

The R Package Ecosystem for Robust Statistics

Valentin Todorov 

United Nations Industrial Development Organization (UNIDO), Vienna, Austria

Correspondence: Valentin Todorov (valentin@todorov.at)

Received: 31 May 2024 | **Revised:** 14 October 2024 | **Accepted:** 15 October 2024

Edited by: Mia Hubert, Commissioning Editor and David Scott, Review Editor and Co-Editor-in-Chief

Keywords: high dimensions | multivariate | outlier | R | robust

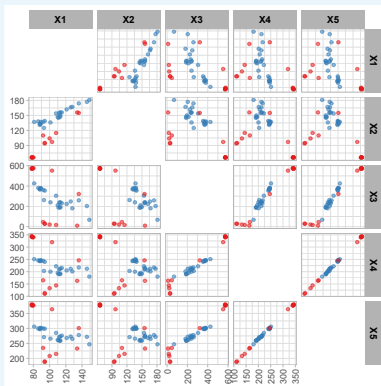
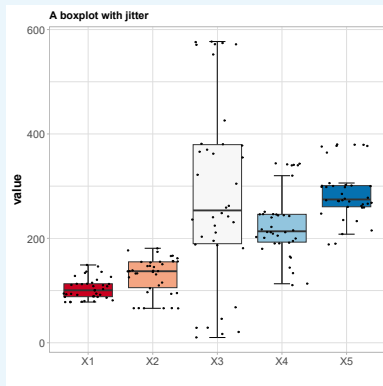
ABSTRACT

In the last few years, the number of R packages implementing different robust statistical methods have increased substantially. There are now numerous packages for computing robust multivariate location and scatter, robust multivariate analysis like principal components and discriminant analysis, robust linear models, and other algorithms dedicated to cope with outliers and other irregularities in the data. This abundance of package options may be overwhelming for both beginners and more experienced R users. Here we provide an overview of the most important 25 R packages for different tasks. As metrics for the importance of each package, we consider its maturity and history, the number of total and average monthly downloads from CRAN (The

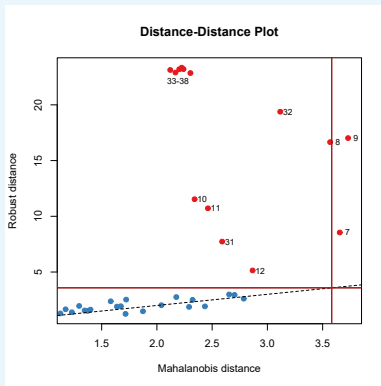
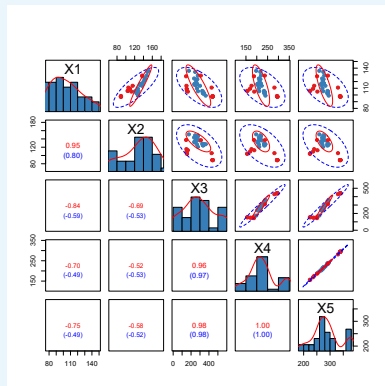
Outline

- 1 Introduction
- 2 R packages on CRAN
- 3 Multivariate location and scatter**
- 4 Principal Component Analysis (PCA)
- 5 Other languages
- 6 AI assisted data analysis
- 7 Summary and conclusions

Multivariate Location and Scatter: Bushfire data



Multivariate Location and Scatter: Bushfire data



Multivariate Location and Scatter in package **rrcov**

- **CovMcd** - Minimum Covariance Determinant Estimator (Rousseeuw, 1985; Rousseeuw and Leroy, 1987)
- **CovOgk** - Pairwise cov estimator (Maronna and Zamar, 2002)
- **CovMve** - Minimum Volume Ellipsoid Estimator (Rousseeuw, 1985; Maronna et al., 2019)
- **CovSde** - Stahel-Donoho estimates (Stahel, 1981; Donoho, 1982; Maronna and Yohai, 1995)
- **CovSest** - S estimates (Davies, 1987; Lopuhaä, 1989; Rousseeuw and Leroy, 1987)
 - ▶ FAST S - (Salibian-Barrera et al., 2006)
 - ▶ SURREAL - (Ruppert, 1992)
 - ▶ Bisquare iteration with HBDP (MVE) start (Maronna et al., 2019)
 - ▶ Rocke type (Maronna et al., 2019)
 - ▶ Deterministic (Hubert et al., 2015)
- **CovMMest** - MM estimates (Tatsuoka and Tyler, 2000)
- **CovMrcd** - Minimum Regularized Covariance Determinant Estimator (Boudt et al., 2020)

CovRobust: a generalized function for robust location and covariance estimation - package `rrcov`

- `CovRobust(x, control, na.action = na.fail)`
- Computes a robust multivariate location and scatter estimate with a high breakdown point, using one of the available estimators.
- Select the estimation method through the argument `control`. It can be:
 - ▶ A control object with estimation options, e.g. an object of class `CovControlMcd` signals MCD estimation
 - ▶ A character string naming the desired method, like "mcd", "ogk", etc.
 - ▶ Empty - then the function will select a method based on the size of the problem
- Demonstrates the power of the OO paradigm - the function is shorter than half screen and has no switch on the method

CovRobust(): Example

```
> getMeth(CovRobust(matrix(rnorm(40),ncol=2)))  
[1] "Stahel-Donoho estimator"  
  
> getMeth(CovRobust(matrix(rnorm(16000),ncol=8)))  
[1] "S-estimates: bisquare"  
  
> getMeth(CovRobust(matrix(rnorm(20000),ncol=10)))  
[1] "S-estimates: Rocke type"  
  
> getMeth(CovRobust(matrix(rnorm(2E5),ncol=2)))  
[1] "Orthogonalized Gnanadesikan-Kettenring Estimator"
```

CovRobust(): Simulation example

```

> ##
> ## Controlling the estimation options: example
> ## of looping through several estimators
> ##
> cc <- list(CovControlMcd(), CovControlMest(), CovControlOgk(),
+           CovControlSest(), CovControlSest(method="rocke"))
> clist <- sapply(cc, reestimate, x=delivery[,1:2])
> sapply(clist, data.class)

[1] "CovMcd" "CovMest" "CovOgk" "CovSest" "CovSest"

> sapply(clist, getMeth)

[1] "Fast MCD(alpha=0.5 ==> h=14); nsamp = 500; (n,k)mini = (300,5)"
[2] "M-Estimates"
[3] "Orthogonalized Gnanadesikan-Kettenring Estimator"
[4] "S-estimates: S-FAST"
[5] "S-estimates: Rocke type"

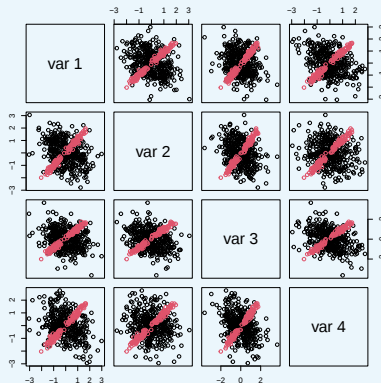
```

Example 1: The Barrow wheel benchmark

- **Barrow wheel**: A benchmark distribution for multivariate tools that aim at good robustness properties. It was introduced in section 5.5a of Hampel et al. (1986) and recalled by Stahel and Maehler in a comment to Tyler et al. (2009).
- This distribution is a mixture of a flat normal distribution contaminated with a portion ε of gross errors concentrated near a one-dimensional subspace.
- A function to generate such distribution is available in the package **robustX**.

Example 1: The Barrow wheel benchmark

```
> library(robustX)
> n <- 500; p <- 4
> r <- rbwheel(n, p)
> n1 <- attr(r, "n1")
> pairs(r, col=1+((1:n) > n1))
```



Example 1: The Barrow wheel benchmark

- We will measure the performance of the different estimators of location and scatter by the condition number of the estimated covariance matrix, i.e.

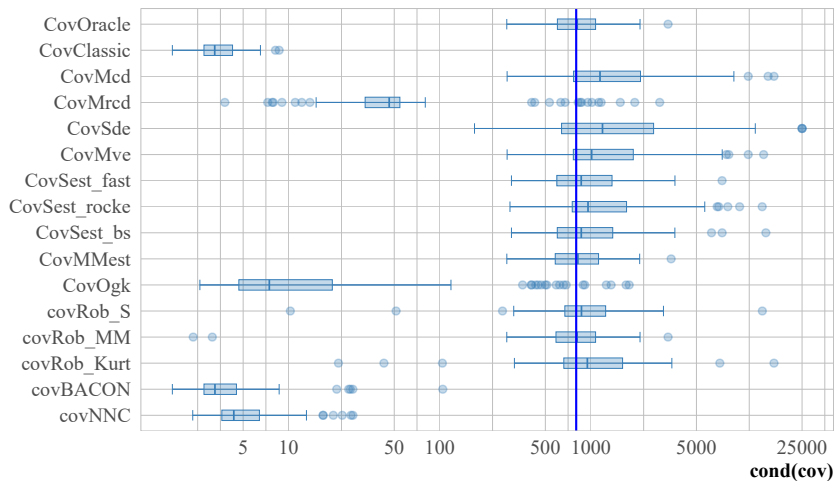
$$\kappa(\hat{\mathbf{S}}) = \text{cond}(\hat{\mathbf{S}}) = \frac{\lambda_{\max}(\hat{\mathbf{S}})}{\lambda_{\min}(\hat{\mathbf{S}})} = \frac{\lambda_1(\hat{\mathbf{S}})}{\lambda_p(\hat{\mathbf{S}})}$$

- The estimators, available in several R packages are listed in the following Table $\Rightarrow$

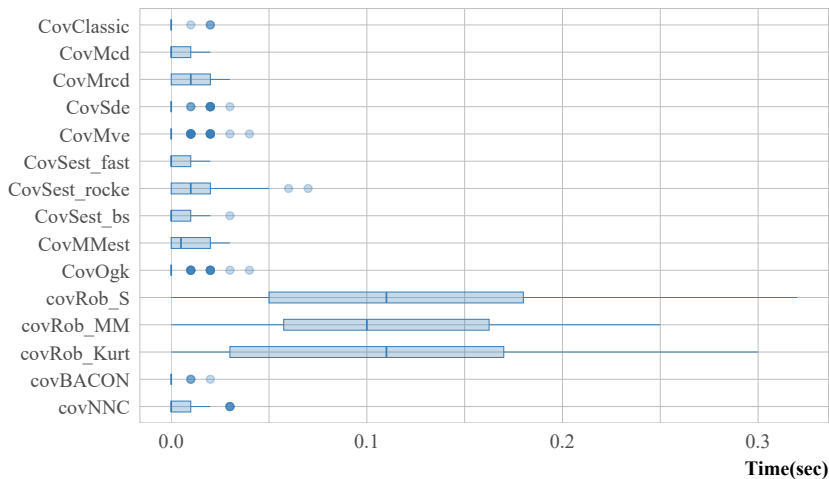
Example 1: The Barrow wheel benchmark

Package	Function	Description
	CovOracle	cov() on the data with the outliers removed
rrcov	CovClassic	Classic (CovClassic in rrcov)
rrcov	CovMcd	FastMCD (Minimum Covariance Determinant)
rrcov	CovMrcd	Minimum Regularized Covariance Determinant
rrcov	CovSde	Stahel-Donoho
rrcov	CovMve	Fast MVE (Minimum Volume Estimator)
rrcov	CovSest_fast	Fast S-estimate (Salibián-Barrera et al., 2006)
rrcov	CovSest_rocke	S-estimate with Rocke ρ function
rrcov	CovSest_bs	S-estimate with bisquare ρ function and HBDP start (MVE)
rrcov	CovMMest	MM-estimator
rrcov	CovOgk	Orthogonalized Gnanadesikan-Kettenring (OGK)
RobStatTM	covRob_S	S-estimate with Rocke function (Maronna et al., 2019)
RobStatTM	covRob_MM	MM-estimate with SHR ρ function (Maronna et al., 2019)
RobStatTM	covRob_Kurt	Kurtosis plus specific directions (Pena and Prieto 2007; Maro
robustX	covBACON	Blocked Adaptive Computationally-Efficient Outlier Nom.
robustX	covNNC	Nearest-Neighbor Cleaning Wang and Raftery (2002)

Example 1: The Barrow wheel benchmark



Example 1: The Barrow wheel benchmark



Example 2: Glass spectra data

Glass spectra data

- A data set originating from 180 16th-17th century archeological glass vessels (Janssens et al., 1998).
- There are 750 characteristics for each vessel, coming from an analysis by an electron-probe X-ray micro-analysis. The data set includes four different materials comprising the vessels, with the larger group being of 145 observations.



Example 2: Glass spectra data

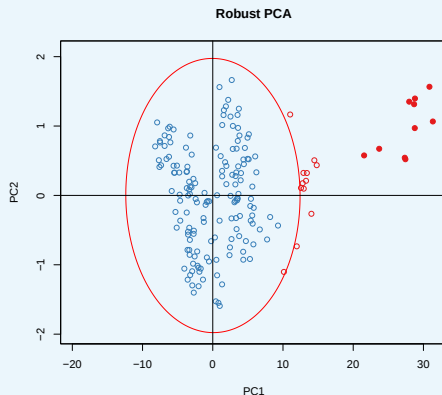
- It is known from other studies on this data set (Serneels et al., 2005; Filzmoser et al., 2020; Filzmoser and Todorov, 2011; Todorov and Filzmoser, 2014) that these 145 observations should form two groups, because during the measurement process the detector efficiency has been changed.
- This example is also considered in detail in Hubert et al. (2005) illustrating the newly proposed ROBPCA algorithm.
- See also Maronna and Yohai (2017)

Example 2: Glass spectra data

- The data are available in the package `rospca` as the data set `Glass`. It is a data frame with 180 observations and 750 variables.
- For this example we will consider only the frequencies 350–385 of this data set, which yield $p = 36$.
- The reason for choosing this interval is that this is a region of the spectra where the variability is highest (Maronna and Yohai, 2017).

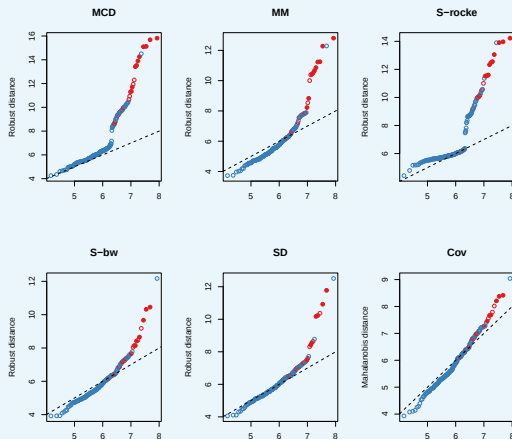
Example 2: Glass spectra example

- Robust principal component analysis of the subset of the data.
- Two groups of outliers are identified: $PC1 \geq 20$ and $10 \geq PC1 > 10$
- The regular observations lie in the zone $PC1 \leq 10$ (inside the 0.975 tolerance ellipse)



Example 2: Glass spectra example

- For several estimators, we show the QQ-plots that compare the ordered squared Mahalanobis distances of the estimators to the quantiles of the χ^2 distribution with p degrees of freedom



Example 2: Glass spectra example

- For each estimator we compute the number of observations corresponding to the largest 22 distances such that $PC1 \geq 20$ and $PC1 \geq 10$



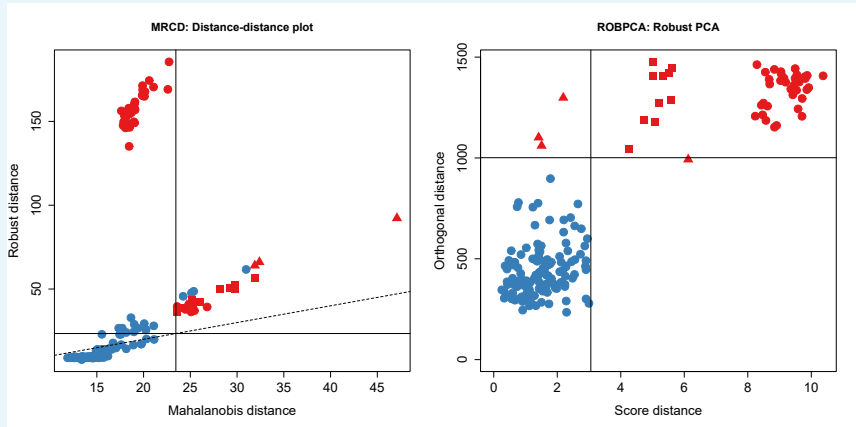
	Number	MCD	MM	S-rocke	S-bw	SDE	COV
$PC1 \geq 20$	10	10	10	10	8	8	6
$PC1 \geq 10$	22	22	16	18	11	11	9

Example 3: Glass spectra II

- Now we turn again to the complete data set.
- Let's remove the variables with 0 MAD (there 13 such variables).
We remain with a 180×487 data set
- And then compute the regularized MCD (MRCD) estimator

```
> X <- Glass[, 1:500]
> X <- X[,-which(apply(X, 2, mad) <= 0)]
> mrcd <- CovMrcd(X)
> plot(mrcd, id.n=0)
```

Example 3: Glass spectra II



Outline

- 1 Introduction
- 2 R packages on CRAN
- 3 Multivariate location and scatter
- 4 Principal Component Analysis (PCA)**
- 5 Other languages
- 6 AI assisted data analysis
- 7 Summary and conclusions

Principal Component Analysis in

Package	Function	Description
rrcov	PcaClassic	Classical PCA (equivalent to <code>prcomp()</code>)
rrcov	PcaCov	PCA based on robust covariance matrix
rrcov	PcaHubert	ROBPCA (Hubert et al., 2005)
rrcov	PcaGrid	PCA based on projection pursuit (Croux et al., 2007)
rrcov	PcaProj	PCA based on PP (Croux and Ruiz-Gazen, 2005)
rrcov	PcaLocantore	Spherical PCA (Locantore et al., 1999)
rrcovHD	SPcaGrid	Sparse PCA based on PP (Croux et al., 2013)
rospca	rospca	ROSPCA (Hubert et al., 2016)
ltsspca	ltsspca	Sparse PCA based on LTS (Wang and Van Aelst, 2020)
pcaPP	PCAggrid	PCA based on projection pursuit (Croux et al., 2007)
pcaPP	PCApj	PCA based on PP (Croux and Ruiz-Gazen, 2005)
pcaPP	sPCAggrid	Sparse PCA based on PP (Croux et al., 2013)
RobStatTM	pcaRobS	Robust PCA based on a robust scale (Maronna, 2005)
RobStatTM	prcompRob	Same as <code>pcaRobS()</code> but returns an object similar to <code>prcomp</code>
cellWise	MacroPCA	ROBPCA for missing values and cellwise outliers (Hubert et al.)

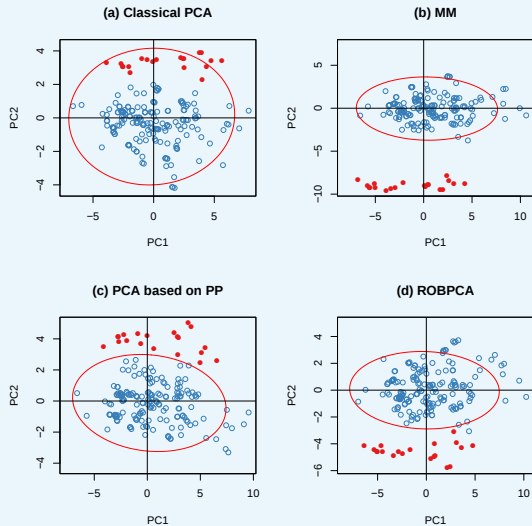
Example: 1985 Auto Imports Database

1985 Auto Imports Database

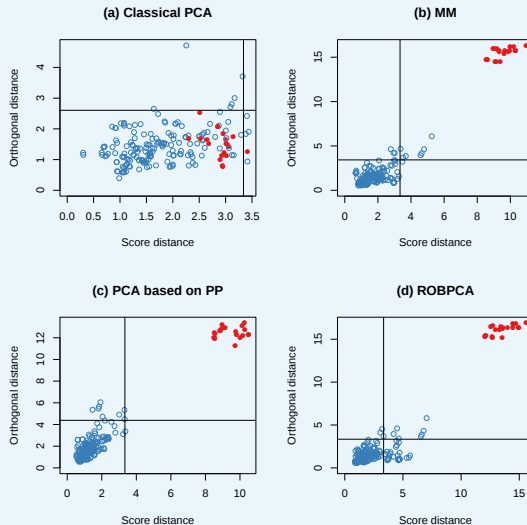
- The original data set (Kibler et al., 1989) contains 205 observations (cars) and 26 variables. Taking only the numerical variables and removing the observations with missing values: 195 observations and 14 variables.
- Twenty of the 195 cars are equipped with diesel engines and we could consider these as outliers compared to the rest 175, running on gasoline
- For other use of this data set see Croux et al. (2013) and Greco and Farcomeni (2016)



Example: 1985 Auto Imports Database



Example: 1985 Auto Imports Database



Sparse Principle Component Analysis

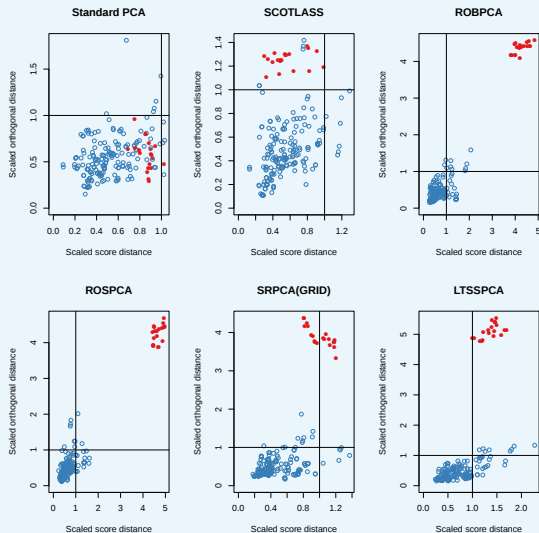
- Obtaining sparse loading matrices yields a sparse PCA, which is especially helpful for analyzing high dimensional data sets
 - SRPCA** Croux et al. (2013): Sparseness is introduced by adding an L_1 constraint and optimizing

$$\mathbf{p}_j = \arg \max_{\mathbf{p}} \text{VAR}(\mathbf{X}\mathbf{p}) - \lambda_j \|\mathbf{p}\|_1$$

where λ_j is a tuning parameter

- Choosing VAR as the sample variance results in the **SCOTLASS** method, which is sparse but not robust
- ROSPCA** Hubert et al. (2016): Based on the ROBPCA algorithm, can be viewed as a two-step procedure. It starts by computing the Stahel-Donoho outlyingness of each observation to find a subset of regular observations and then uses SCOTLASS to generate sparse loadings
- LTSSPCA** Wang and Van Aelst (2020): Based on LTS, adds sparseness to the method of Maronna (2005) by modifying the objective function to include a penalty term

Example: 1985 Auto Imports Database



Outline

- 1 Introduction
- 2 R packages on CRAN
- 3 Multivariate location and scatter
- 4 Principal Component Analysis (PCA)
- 5 Other languages**
- 6 AI assisted data analysis
- 7 Summary and conclusions

Robustness in Matlab

Matlab

- **LIBRA** (Verboven and Hubert, 2005, 2010). Last updated in 2016
- **FSDA**: Flexible Robust Statistics Data Analysis. Last updated April 2025
 - ▶ Robust Regression: Forward search, LMS, LTS, S- and MM estimates; Robust Bayesian and heteroskedastic regression
 - ▶ Robust transformations and model selection
 - ▶ Robust Multivariate analysis: Forward search, MCD, MVE, S- and MM-estimates
 - ▶ Robust Cluster analysis
 - ▶ Dynamic visualization

Example: Fish morphology data

Fish morphology data

- FishMorphology dataset in the R package easyCODA (Greenacre, 2019)
- 26 measurements on a sample of 75 fish of the species Arctic charr (*Salvelinus alpinus*)
- Additionally are recorded: sex (male or female), habitat (littoral, close to the shore) and pelagic (in deeper water far from the shore) and body mass



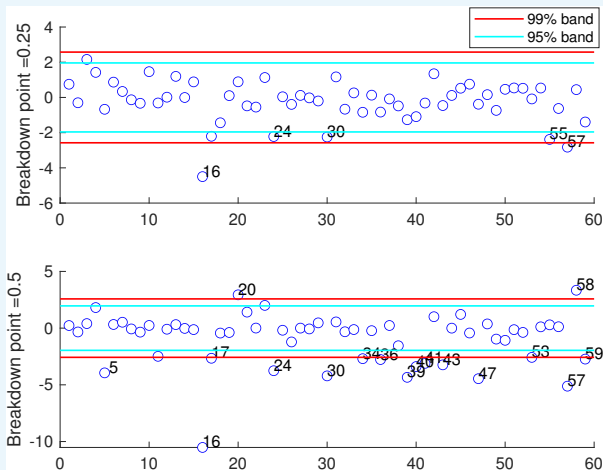
Example: Fish morphology data

Fish morphology data

- Select only the former habitat (59 observations)
- Take the first 10 (out of the 26) measures as explanatory variables (to provide a single population of moderate size):
 - ▶ {Bg, Bd, Bcw, Jw, Jl, Bp, Bac, Bch, Fc, Fdw}
 - ▶ The dependent variable y will be the body mass
- There is one observation which is an “obvious” outlier (obs. 16). This “obvious” outlier can be identified by any outlier detection method and it will appear in all graphs shown
- The data are not subject to a constant sum constraint but, nevertheless, it is compositional in nature

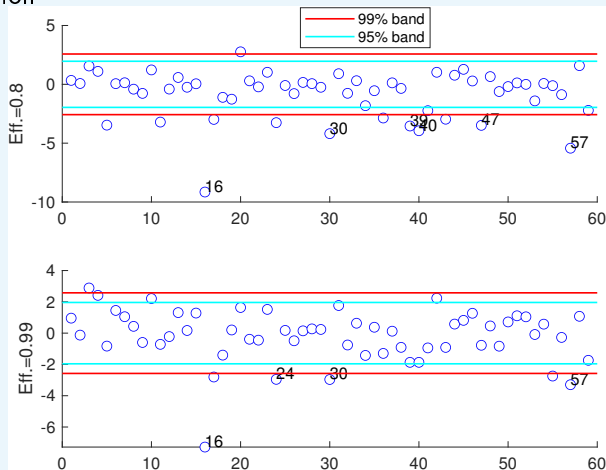
Example: Fish morphology data - S estimates

- S-estimation: Index plots of residuals from Tukey's biweight ρ function



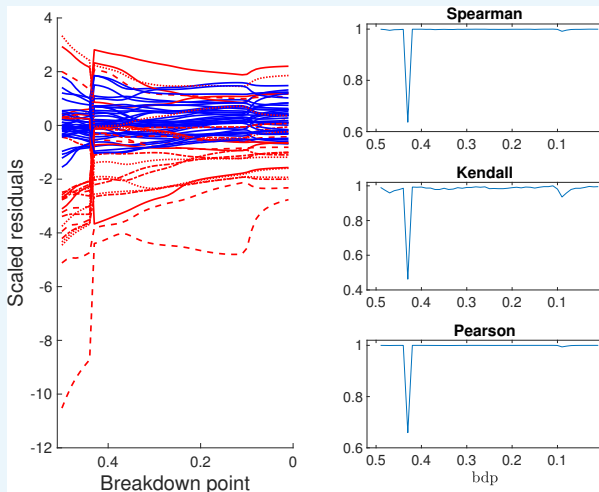
Example: Fish morphology data - MM estimates

- MM-estimation: Index plots of residuals from Tukey's biweight ρ function



Example: Fish morphology data - monitoring S estimates

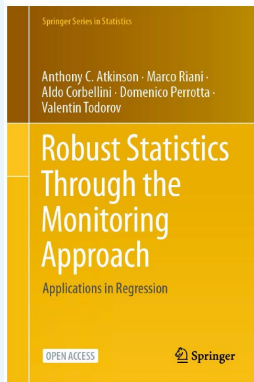
- S-estimation: Plot of scaled residuals (left) and three measures of the correlations of adjacent residuals (right). The abrupt switch at 0.43 is evident



Robustness in MATLAB



springer.com



2025. XX, 404 p. 357 illus., 354 illus. in color.



Special offer / Get 20% off the printed book or eBook!

Enter the following coupon code at checkout on link.springer.com to apply discount.

3JWrMnzewuJnfY / Valid May 19, 2025 – Jun 16, 2025

A. C. Atkinson, M. Riani, A. Corbellini, D. Perrotta, V. Todorov

Robust Statistics Through the Monitoring Approach

Applications in Regression

- This book is open access, which means that you have free and unlimited access
- Presents robust statistical methods and procedures and their applications, with a focus on regression
- Provides numerous data analyses using MATLAB and R

Robustness in MATLAB

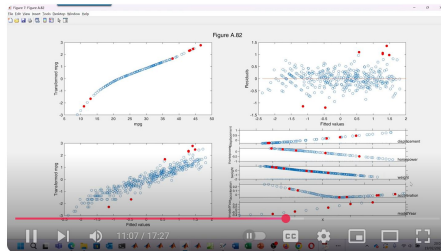
GitHub and YouTube

- GitHub repository with the code for all examples and figures:

<https://github.com/UniprJRC/MonitoringBook>

- More than 20 videos on YouTube:

<https://www.youtube.com/watch?v=4wP0rbELIjE&t=833s>



RSMA: analysis of the auto mpg data (Exercise 10.4)

Robustness in Python

Python

- Function `MinCovDet()` in `scikit-learn`
- Function `RLM()` in `statsmodels`
- Package `Hypotetize` based on Rand R. Wilcox's collection of R functions
- Package `RobPy` (Leyder et al., 2024)

Robustness in Python

RobPy (Leyder et al., 2024)

- Inherits from the popular Python libraries **NumPy**, **SciPy** and **scikit-learn**, ensuring compatibility with these foundational tools
- Many of the algorithms are inspired by implementations found in the R-packages **robustbase**, **rrcov** and **cellWise**
- Includes tools for
 - ▶ Robust preprocessing and univariate estimation (**Qn**, **Tau**, **uniMCD**), adjusted boxplot
 - ▶ Covariance matrices: MCD, OGK, DetMCD and cellMCD
 - ▶ Regression: LTS, S- and MM-estimators
 - ▶ Principal Component Analysis: ROBPCA and spherical PCA
 - ▶ Detecting of cellwise outliers
- Data sets (from R packages)
- Provides specialized diagnostic plots for visualizing outliers in both casewise and cellwise contexts.

Example: MM Regression of the Stackloss data

```
>>> import matplotlib.pyplot as plt
>>> import numpy as np
>>> import pandas as pd
>>> from robpy.regression import MMRegression

>>> stackloss = pandas.read_csv('stacklos.csv')
>>> X = stackloss.drop(columns=['stackloss'])
>>> y = stackloss['stackloss']

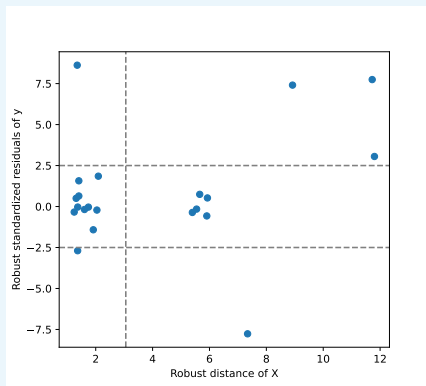
>>> estimator = MMRegression().fit(X, y)
>>> estimator.model.coef_

[ 0.07827451  0.03288496 -0.06428978]
```

Example: MM Regression of the Stackloss data

- Create a diagnostic plot where robust residuals are plotted against the robust mahalanobis distances of the training data

```
>>> resid, std_resid, distances, vt, ht = estimator.outlier_map(X, y.to_numpy(),
```



Julia

Julia

- High-performance, open-source programming language designed for technical computing, data science, and machine learning
- Combines the ease of use of languages like R and Python with the speed of C and Fortran
- Just-In-Time (JIT) compiler that translates code to machine instructions on the fly
- Strong interactive capabilities, allowing users to execute code and explore data in real-time
- Designed for parallel and distributed computing
- Supports multiple dispatch (functions can have different behavior based on the types of their arguments)
- Built-in package manager

The two language problem

The two language problem

- Prototyping with one slow dynamic language (Python) and rewrite it with a fast static language (C++) for the final product
- Programming in a high-level language such as R and Python, the performance-critical parts have to be rewritten in C/C++ for performance
- This is hugely inefficient, because
 - ▶ Introduces human error
 - ▶ Wasted effort
 - ▶ Hinders parallelism
 - ▶ Difficult to maintain
 - ▶ Difficult reproducibility and contribution
- Julia promises to solve the two language problem

The two language problem

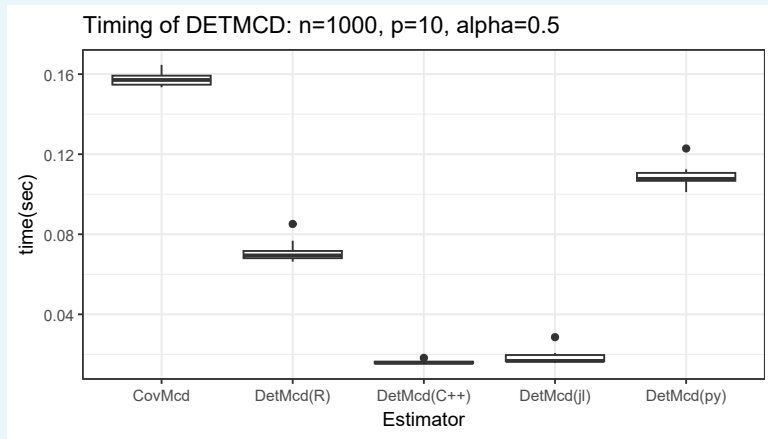
Only a few (7) packages of the 25 listed earlier are programmed in pure R

#	Package	Compiled language
1	cluster	C
2	rrcov	FORTRAN, C, C++
3	mvoutlier	R
4	pcaPP	C++
5	robustbase	FORTRAN, C
6	robust	FORTRAN, C
7	ICS	R
8	robCompositions	C++
9	tclust	C++, Rcpp, RcppArmadillo
10	rrcovNA	FORTRAN, C
11	robustHD	C++
12	rrcovHD	R
13	VIM	C++, Rcpp

#	Package	Compiled language
14	GSE	C++
15	WRS2	R
16	ICSOutlier	R
17	cellWise	C++, Rcpp, RcppArmadillo
18	rrcov3way	R
19	pense	C++, Rcpp, RcppArmadillo
20	rospca	R
21	fsdaR	R, Matlab
22	mrfDepth	FORTRAN, C++
23	Itsspca	C++, Rcpp, RcppArmadillo
24	RobStatTM	R, C
25	classmap	R
26		

The two language problem

Lets compare different implementation of the Deterministic MCD algorithm



Robustness in Julia

Julia packages for robust statistics

- **RobustStats.jl**: A collection of robust statistical tests in Julia
 - ▶ Estimators of location and dispersion, standard error and the confidence intervals for the location or dispersion estimators
 - ▶ Based on the R package WRS by Rand Wilcox
- **LinRegOutliers.jl**: A Julia package for detecting outliers in linear regression
 - ▶ Robust regression: Least Absolute Deviations, LMS, LTS and others
 - ▶ Multivariate methods: Hadi-1992, Hadi-1994, MVE, MCD
 - ▶ Visual methods: BCH plot, MVE-LTS Plot, Stalactite Plot
- **OutlierDetection.jl**: Julia toolkit for detecting outlying objects
 - ▶ Provides a unified API for outlier detection
 - ▶ Provides access to state-of-the-art outlier detection algorithms
 - ▶ Seamlessly integrates with Julia's existing machine learning ecosystem

Robustness in Julia

`rrcov.jl`

- Many of the algorithms are inspired by implementations found in the R-packages `robustbase` and `rrcov`
- Includes tools for
 - ▶ Robust preprocessing and univariate estimation (`Qn`, `Tau`, `uniMCD`)
 - ▶ Covariance matrices: MCD, OGK, DetMCD
 - ▶ Regression: LTS, S- and MM-estimators
 - ▶ Principal Component Analysis: ROBPCA and spherical PCA
- Data sets (from R packages)
- Provides specialized diagnostic plots for visualizing outliers

Example: Outlier detection with `CovMed()` in the `wood` data set

```
julia> using rrcov                # load the rrcov package
julia> using RData                # used to read R data sets

julia> objs = load("data/wood.rda"); # load the wood data set
julia> X = objs["wood"][:,1:5]
```

20x5 DataFrame

Row	x1	x2	x3	x4	x5
	Float64	Float64	Float64	Float64	Float64
1	0.573	0.1059	0.465	0.538	0.841
2	0.651	0.1356	0.527	0.545	0.887
3	0.606	0.1273	0.494	0.521	0.92
4	0.437	0.1591	0.446	0.423	0.992
5	0.547	0.1135	0.531	0.519	0.915
...	...				

15 rows omitted

Example: Outlier detection with `CovMcd()` in the `wood` data set

```
julia> mcd = CovMcd(correct_covariance=true, reweighting=true);
julia> fit!(mcd, X);
```

```
julia> display(mcd)
```

```
-> Method: Fast MCD Estimator: (alpha=0.51 ==> h=77)
```

```
Robust estimate of location:
```

```
[6.17368, 2.91711, 4.79868, 1.66184]
```

```
Robust estimate of covariance:
```

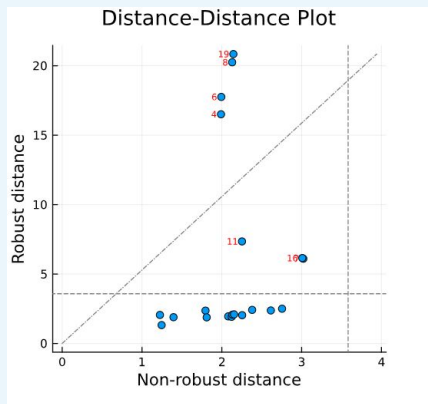
```
4×4 Matrix{Float64}:
```

```
0.221432  0.0721895  0.208232  0.106582
0.0721895  0.0705035  0.0921561  0.0615947
0.208232  0.0921561  0.408932  0.237549
0.106582  0.0615947  0.237549  0.167458
```

Example: Outlier detection with `CovMcd()` in the `wood` data set

- Create a diagnostic plot where robust mahalanobis distances are plotted against the classical mahalanobis distances

```
julia> dd_plot(mcd)
```



Outline

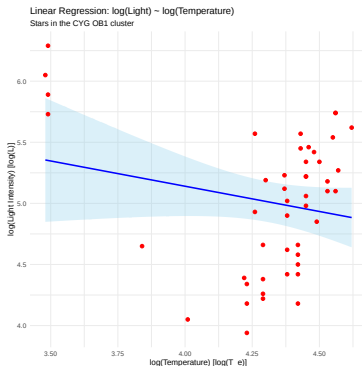
- 1 Introduction
- 2 R packages on CRAN
- 3 Multivariate location and scatter
- 4 Principal Component Analysis (PCA)
- 5 Other languages
- 6 AI assisted data analysis**
- 7 Summary and conclusions

Example: AI assisted data analysis

Prompt(s)

- Please show me a linear regression model of $\log(\text{light})$ vs. $\log(\text{Te})$ in starsCYG, complete with 95% confidence bands
- Please note that $\log(\text{light})$ and $\log(\text{Te})$ are already log-transformed

Linear Regression Model with 95% Confidence Bands:



Example: AI assisted data analysis

Prompt(s)

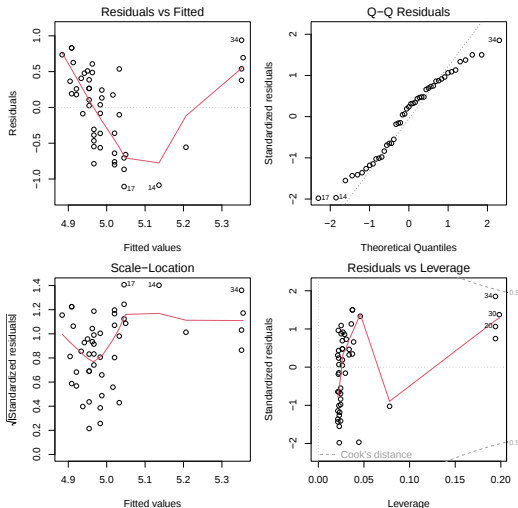
- Please check whether the assumptions of the linear regression model hold and report the corresponding regression diagnostics.

We need to check:

- Linearity (relationship between predictors and response is linear)
- Independence (residuals are not autocorrelated)
- Homoscedasticity (constant variance of residuals)
- Normality (residuals are normally distributed)
- No influential outliers (no points unduly affecting the model)

Example: AI assisted data analysis

Diagnostic plots (base R):



Example: AI assisted data analysis

Additional checks:

```
library(car)  # For advanced diagnostics

# Normality test (Shapiro-Wilk)
shapiro.test(residuals(model))

# Breusch-Pagan test for heteroscedasticity
ncvTest(model)

# Durbin-Watson test for autocorrelation
durbinWatsonTest(model)

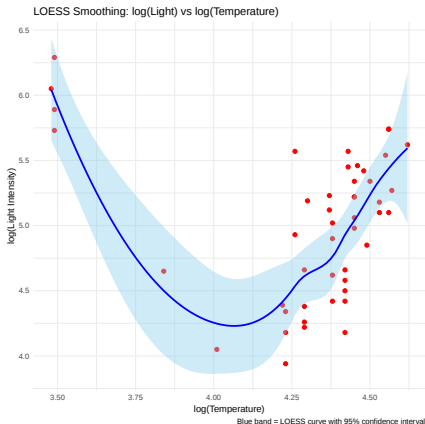
# Influence metrics (Cook's distance, leverage)
influence.measures(model)
```

Example: AI assisted data analysis

Prompt(s)

- Please show me a LOESS curve instead.

Fit and Plot the LOESS Curve:

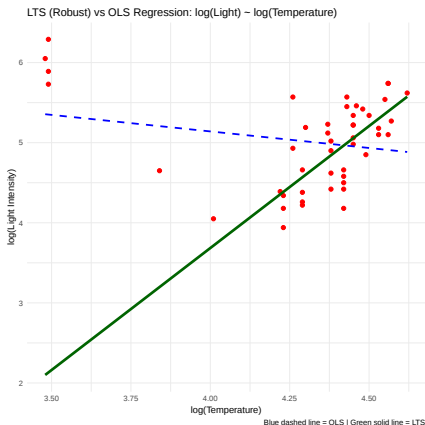


Example: AI assisted data analysis

Prompt(s)

- Please show me an LTS line now

Plot LTS Line with OLS and Data Points:



Example: AI assisted data analysis

Prompt(s)

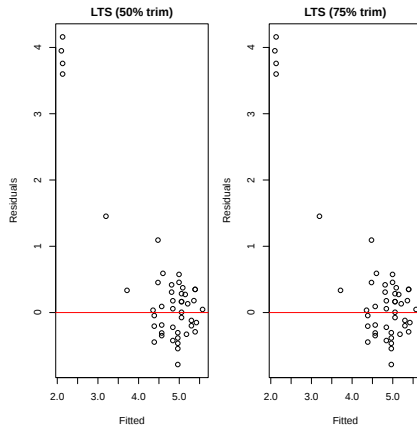
- Please adjust the LTS trimming and analyze the residuals

Adjusting LTS Trimming and Refitting the Model:

- The `ltsReg()` function in `robustbase` allows you to control the trimming proportion via the `alpha` argument. By default, $\alpha = 0.5$ (50% trimming), but we can adjust it to include more or fewer observations in the fitting process.
- Interpretation:
 - ▶ Higher alpha (e.g., 0.75): More data included $\rightarrow$ closer to OLS but still robust.
 - ▶ Lower alpha (e.g., 0.5): More trimming $\rightarrow$ highly robust but potentially less efficient.

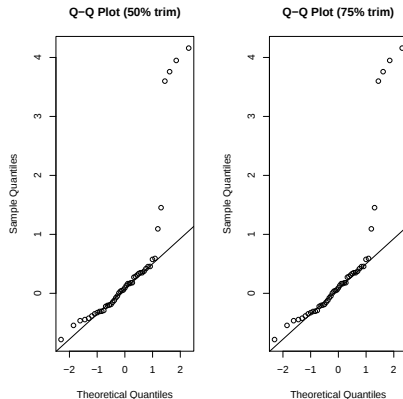
Example: AI assisted data analysis

Plot residuals vs fitted values:



Example: AI assisted data analysis

Q-Q plots for normality:



ChatGPT ecosystem in

Incorporating AI technology into the R language workflow

- Important things to keep in mind:
 - ▶ Everything you ask gets sent to OpenAI's servers. Don't use ChatGPT tools to process sensitive information.
 - ▶ ChatGPT may confidently return incorrect answers.
 - ▶ ChatGPT can generate different responses to the same query - reproducibility?
 - ▶ LLMs have training data cutoff dates
 - ▶ Need own OpenAI API key, and the API isn't free to use.

ChatGPT ecosystem in

The most notable R packages for using OpenAI's GPT-3.5 and GPT-4 LLMs

- **ellmer**: Supports chat with LLMs from a range of providers. Probably the most important package - developed by Hadley Wickham, supported by Posit (formerly RStudio)
- **tidyllm**: A tidy interface for integrating large language model (LLM) APIs
- **ragnar** (RAG for R): Helps implement Retrieval-Augmented Generation (RAG) workflows. Only on GitHub
- **tidyprompt**: Easily construct prompts and associated logic for interacting with LLMs
- **air**: An R console utility that lets you ask R related questions to the 'OpenAI' LLM
- **TheOpenAIR**: Integrates 'OpenAI' LLM into R Workflows
- **RTutor**: Available on the web. It is currently the only app or package listed that does not require a ChatGPT API key to use

Outline

- 1 Introduction
- 2 R packages on CRAN
- 3 Multivariate location and scatter
- 4 Principal Component Analysis (PCA)
- 5 Other languages
- 6 AI assisted data analysis
- 7 Summary and conclusions

Summary and conclusions

- We explored robust statistical methods and the software tools available to support them — with a strong focus on R packages, comparative implementations in MATLAB, Python and Julia and emerging AI-assisted tools.
- Most of the information in this presentation is available at the GitHub repository:
<https://github.com/valentint/robust-R-ecosystem-WIREs>
- Regularly updated
- Comments and suggestions are very welcome:
valentin@todorov.at

References I

- R. A. Maronna, D. Martin, V. Yohai, M. Salibian-Barrera, Robust Statistics: Theory and Methods (with R): Second edition, John Wiley & Sons, New York, 2019.
- R. A. Maronna, R. H. Zamar, Robust estimation of location and dispersion for high-dimensional datasets, *Technometrics* 44 (2002) 307–317.
- K. Boudt, P. J. Rousseeuw, S. Vanduffel, T. Verdonck, The minimum regularized covariance determinant estimator, *Statistics and Computing* 30 (2020) 113–128.
- M. Salibian-Barrera, S. Van Aelst, G. Willems, Principal components analysis based on multivariate MM estimators with fast and robust bootstrap, *Journal of the American Statistical Association* 101 (2006) 1198–1211.

References II

- M. Hubert, P. Rousseeuw, K. Vanden Branden, ROBPCA: A new approach to robust principal component analysis, *Technometrics* 47 (2005) 64–79.
- R. A. Maronna, V. J. Yohai, Robust and efficient estimation of multivariate scatter and location, *Computational Statistics & Data Analysis* 109 (2017) 64–75.
- C. Croux, P. Filzmoser, M. Oliveira, Algorithms for projection-pursuit robust principal component analysis, *Chemometrics and Intelligent Laboratory Systems* 87 (2007) 218–225.
- C. Croux, A. Ruiz-Gazen, High breakdown estimators for principal components: The projection-pursuit approach revisited, *Journal of Multivariate Analysis* 95 (2005) 206–226.

References III

- N. Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang, K. Cohen, Robust principal components for functional data, *Test* 8 (1999) 1–28.
- C. Croux, P. Filzmoser, H. Fritz, Robust sparse principal component analysis, *Technometrics* 55 (2013) 202–214.
- M. Hubert, T. Reynkens, E. Schmitt, T. Verdonck, Sparse PCA for high-dimensional data with outliers, *Technometrics* 58 (2016) 424–434.
- Y. Wang, S. Van Aelst, Sparse principal component analysis based on least trimmed squares, *Technometrics* 62 (2020) 473–485.
doi:10.1080/00401706.2019.1671234.
- R. A. Maronna, Principal components and orthogonal regression based on robust scales, *Technometrics* 47 (2005) 264–273.

References IV

- M. Hubert, P. J. Rousseeuw, W. Van den Bossche, MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers, *Technometrics* 61 (2019) 459–473.
- L. Greco, A. Farcomeni, A plug-in approach to sparse and robust principal component analysis, *TEST* 25 (2016) 449–481.
- M. Greenacre, *Compositional Data Analysis in Practice*, Chapman & Hall / CRC Press, 2019.
- S. Leyder, J. Raymaekers, P. J. Rousseeuw, T. Servotte, T. Verdonck, Robpy: a python package for robust statistical methods, 2024. [arXiv:2411.01954](https://arxiv.org/abs/2411.01954).