OUTLIER DETECTION IN SURVEY DATA

Valentin Todorov

United Nations Industrial Development Organization (UNIDO)

Conference on Indicators and Survey Methodology 2010
Vienna, Austria, 24-26 February 2010

Outline

What is an Outlier

" ... whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways."

Bacon, F. (1620) *Novum Organum*

Hadi, Imon and Werner (2009) *Detection of Outliers*

What is an Outlier

- **Bacon, F. (1620)** *Novum Organum*
- **Legendre, A.M. (1848)** On the method of least squares
- **Edgeworth, F.Y. (1887)** The choice of means. *Philosophical Magazine*
- **Hawkins, D. (1980)** An outlier is an observation that deviates so much from other observations as to arose suspicion that it was generated by different mechanism
- **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. They provide more than 100 outlier detection tests $\Rightarrow$ most are univariate and distribution-based

Outliers in Sample Surveys

- "Rule based" approach - identification by data specific edit rules developed by subject matter experts followed by deletion and imputation $\longleftarrow$ strictly deterministic, ignore the probabilistic component, extremely labor intensive
- Univariate methods - favored for their simplicity. These are informal graphical methods like histograms, box plots, dot plots; quartile methods to create allowable range for the data; robust methods like medians, Winsorized means, etc.
- Multivariate methods - rarely used although most of the surveys collect multivariate data

Outliers in Sample Surveys: Multivariate methods

- Statistics Canada (Franklin *et al.*, 2000) - Annual Wholesale and Retail Trade Survey (AWRTS)
  - Based on PCA and Stahel-Donoho estimator of multivariate location and scatter
  - Easily run and interpreted by the subject matter experts
  - Limited data set size
  - Only complete data
  - No sampling weights
- The EUREDIT project of the EU (Charlton 2004)
  - Handling of missing values
  - Sampling weights

Outlier detection and Robust estimation are closely related

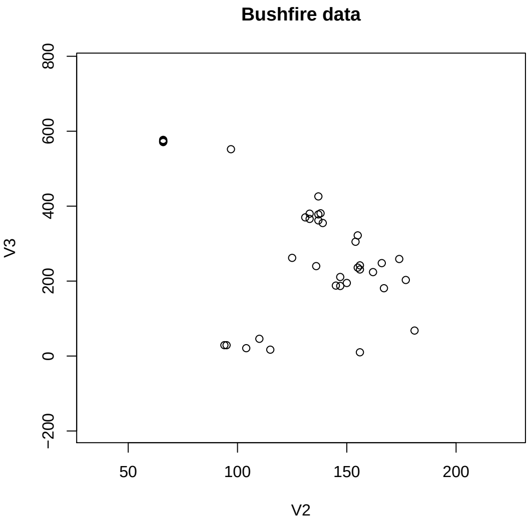
1. **Robust estimation:** find an estimate which is not influenced by the presence of outliers in the sample
2. **Outlier detection:** find all outliers, which could distort the estimate

- If we have a solution to the first problem we can identify the outliers using robust residuals or distances
- If we know the outliers we can remove or downweight them and use classical estimation methods
- For the purposes of official statistics the second approach is more appropriate

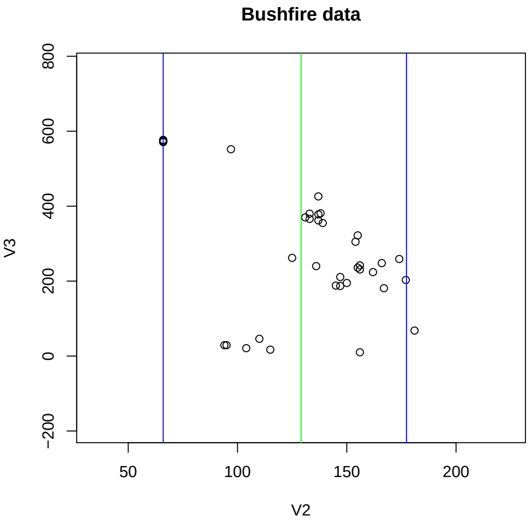
Example: Bushfire data

- A data set with 38 observations in 5 variables - Campbell (1989)
- Contains satellite measurements on five frequency bands, corresponding to each of 38 pixels
- Used to locate bushfire scars
- Very well studied (Maronna and Yohai, 1995; Maronna and Zamar, 2002)
- 12 clear outliers: **33-38**, **32**, **7-11**; 12 and 13 are suspect
- Available in the R package `robustbase`

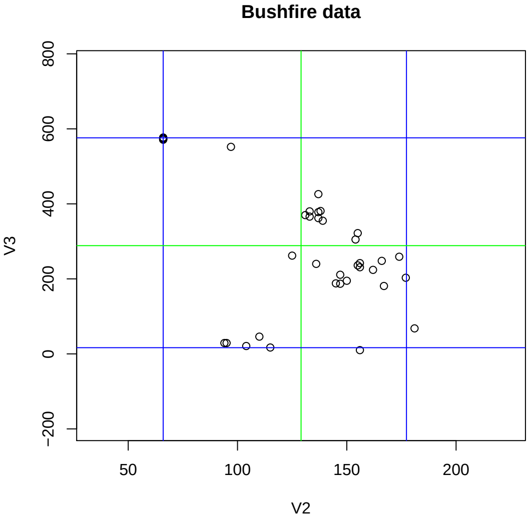
Example: Bushfire data



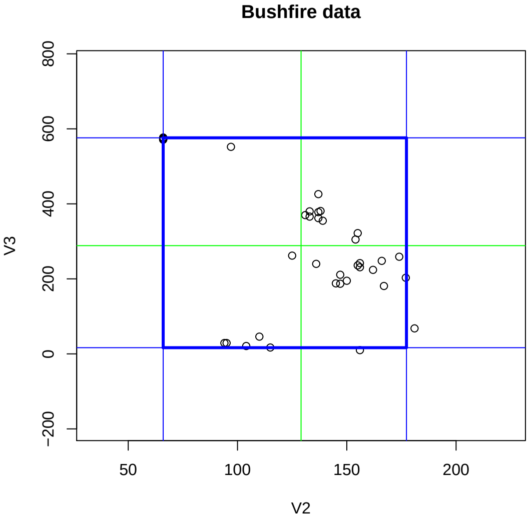
Example: Bushfire data



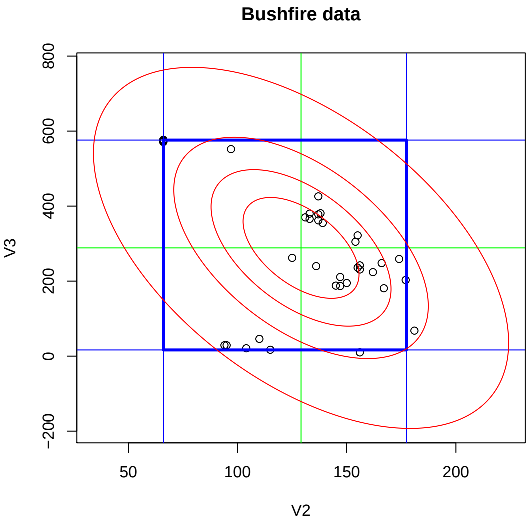
Example: Bushfire data



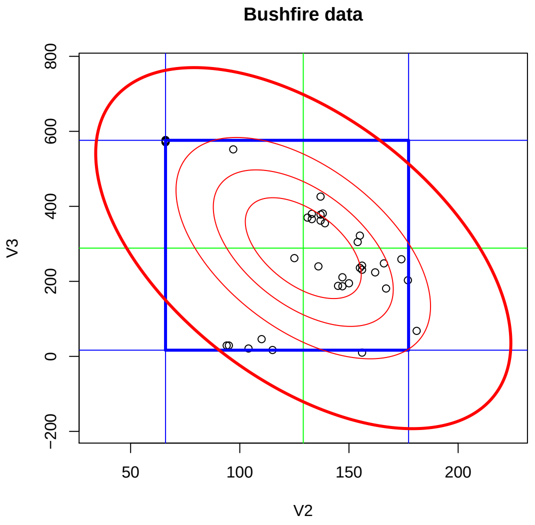
Example: Bushfire data



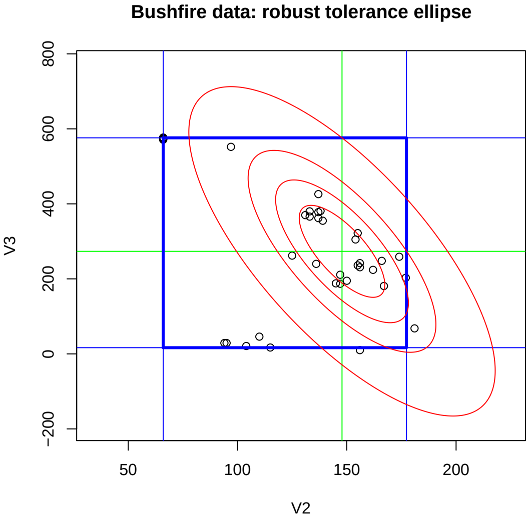
Example: Bushfire data



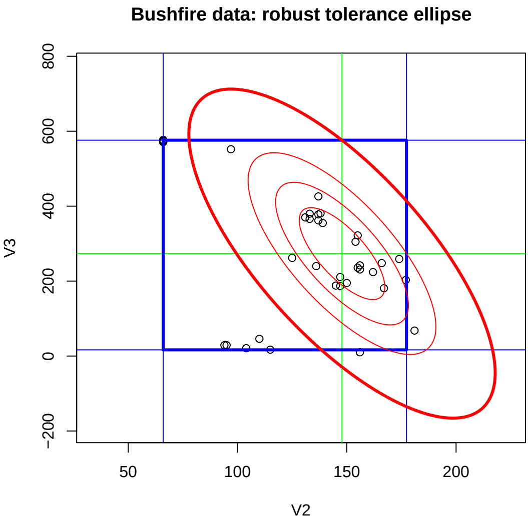
Example: Bushfire data



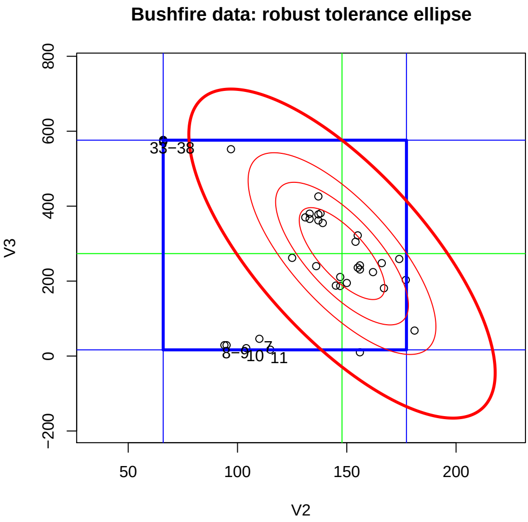
Example: Bushfire data



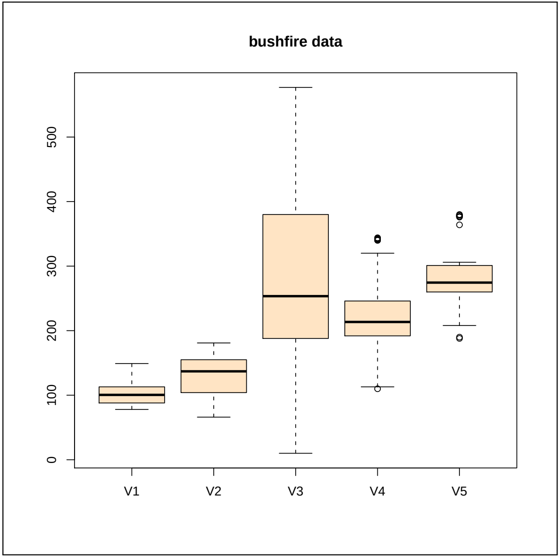
Example: Bushfire data



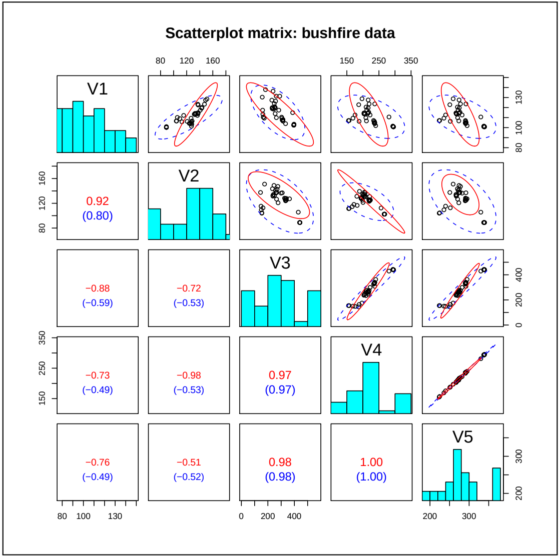
Example: Bushfire data



Example: Bushfire data - Boxplots



Example: Bushfire data - Scatterplot matrix



Detection of Multivariate Outliers in Sample Surveys: The Challenges

- The methods must be able to work with moderate to **large data sets** (hundreds of variables and tens of thousands of observations) - therefore we consider computational speed a very important criterion
- Survey data often contain **missing values**, therefore the methods must be able to work with incomplete data
- The survey data are often **skewed** - use appropriate transformations or special robust methods for skewed data (Hubert *et al.*, 2008)
- The methods must be able to cope with the complex sample design of a survey using **sampling weights**

DIFFICULT SET-UP:

LARGE MULTIVARIATE INCOMPLETE SAMPLE SURVEY DATA

General Framework for Multivariate Outliers

Two phases (Rocke and Woodruff, 1996)

1. Calculate **Robust Distances**
   - Obtain robust estimates of location **T** and scatter **C**
   - Calculate robust Mahalanobis-type distance

$$RD_i = \sqrt{((\mathbf{x}_i - \mathbf{T})^t \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{T}))}$$

2. **Cutoff point**: Determine separation boundary Q.

Declare points with $RD_i > Q$, i.e. points which are sufficiently far from the robust center as outliers.

Usually $Q = \chi^2_p(0.975)$ but see also Hardin and Rocke (2005), Filzmoser, Garrett, and Reimann (2005), Cerioli, Riani, and Atkinson (2008).

Robust multivariate location and scatter

- M-ESTIMATES - Maronna (1976) $\leftarrow$ zero breakdown point
- STAHEL-DONOHO - Stahel (1981), Donoho (1982) $\leftarrow$ computationally feasible only for small data sets
- MCD - Minimum Covariance Determinant - (Rousseeuw, 1985; Rousseeuw and Van Driessen, 1999) $\leftarrow$ efficient computational algorithm exists
- OGK - Orthogonalized Gnanadesikan-Kettenring - (Maronna and Zamar, 2002) $\leftarrow$ even faster than MCD
- S-ESTIMATORS - (Rousseeuw and Leroy 1987; Davies 1987) $\leftarrow$ fast algorithms available

None of them can handle missing values

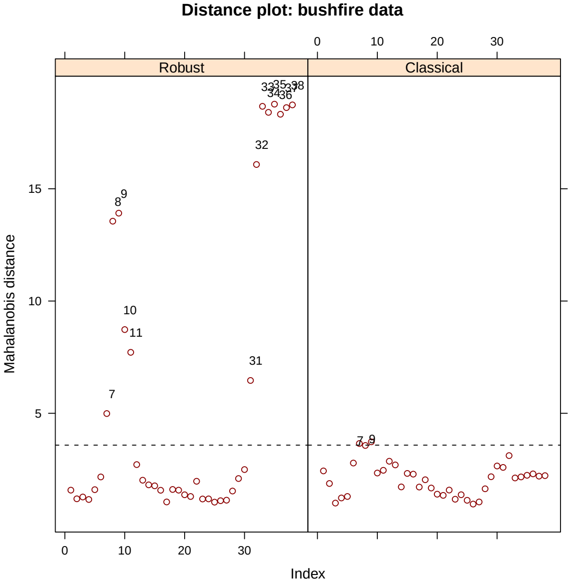
Robust multivariate location and scatter in R - rrcov

- **CovMcd** - Minimum Covariance Determinant Estimator (Rousseeuw, 1985; Rousseeuw and van Driessen, 1999)
- **CovOgk** - Pairwise cov estimator (Maronna and Zamar, 2002)
- **CovMve** - Minimum Volume Ellipsoid Estimator (Rousseeuw, 1985; Maronna et al., 2006)
- **CovMest** - M estimates (Rocke, 1996)
- **CovSest** - S estimates (Rousseeuw and Leroy, 1987; Davies, 1987; Lopuhaä, 1989)
  - FAST S - (similar to the regression estimator of Salibián-Barrera and Yohai, 2006)
  - SURREAL - (Ruppert, 1992)
  - Bisquare iteration with HBDP (MVE) start
  - Rocke type (Maronna et al., 2006)
- **CovSde** - Stahel-Donoho estimator (Stahel,1981;Donoho,1982;Maronna and Yohai, 1995.)

CovRobust: a generalized function for robust location and covariance estimation

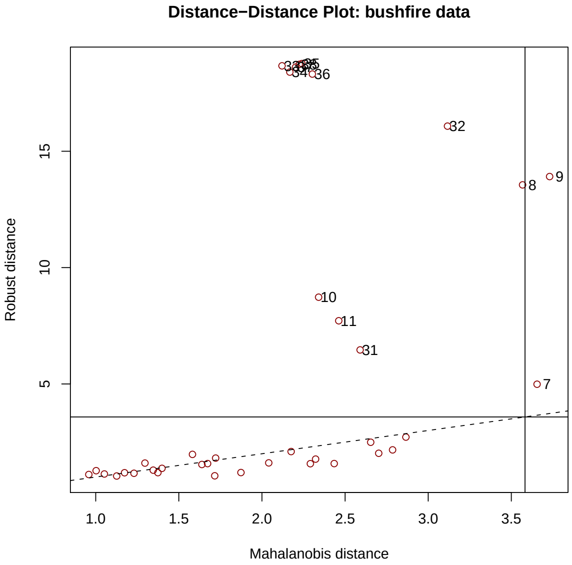
- `CovRobust(x, control, na.action = na.fail)`
- Computes a robust multivariate location and scatter estimate with a high breakdown point, using one of the available estimators.
- Select the estimation method through the argument **`control`**. It can be:
  - A control object with estimation options, e.g. an object of class **`CovControlMcd`** signals MCD estimation
  - A character string naming the desired method, like "mcd","ogk", etc.
  - Empty - than 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

Multivariate Location and Scatter: Plots I



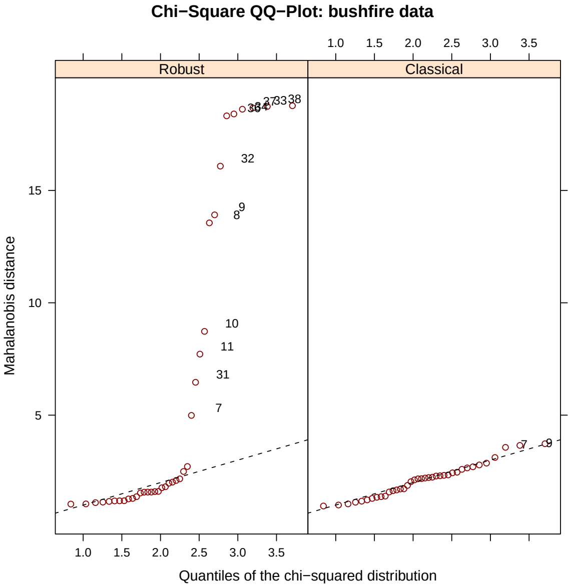
- Both robust and classical Mahalanobis distances are shown in parallel panels - the outliers have large MD_i
- A line is drawn at $y = cutoff = \sqrt{\chi^2_{p,0.975}}$
- The observations with $MD_i \geq cutoff = \sqrt{\chi^2_{p,0.975}}$ are identified by their index

Multivariate Location and Scatter: Plots II



- Robust distances versus Mahalanobis distances
- The dashed line is $RD_i = MD_i$
- The horizontal and vertical lines are $y = \sqrt{\chi^2_{p,0.975}}$ and $x = \sqrt{\chi^2_{p,0.975}}$

Multivariate Location and Scatter: Plots III



- A Quantile-Quantile comparison plot of the Robust distances and the Mahalanobis distances versus the square root of the quantiles of the chi-squared distribution

Algorithms: Handling Missing Values

- ER-ALGORITHM - Little (1988) ← zero breakdown point (based on M-estimates)
- PM-MCD - imputation under MVN model followed by MCD (R package `norm` and fast MCD implementation in package `rrcov`)
- PM-OGK - imputation under MVN model followed by OGK (fast OGK implementation in package `rrcov`)
- PM-S - same as above
- EM-MCD - Victoria-Feser and Copt (2004) ← cannot attain high breakdown point
- ERTBS - Victoria-Feser and Copt (2004) ← same as above

Algorithms: Handling Missing Values

- TRC - Transformed Rank Correlations - Béguin and Hulliger (2004)
- EA - Epidemic Algorithm - Béguin and Hulliger (2004)
- BACON-EEM - Béguin and Hulliger (2008) - a combination of BACON algorithm (Billor, Hadi and Vellemann 2000) and EM

All three algorithms can handle sampling weights

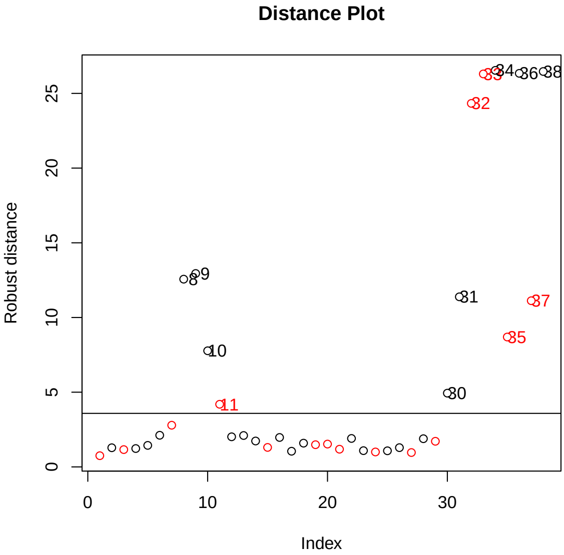
Robust Sequential Imputation followed by HBDP estimation

- SEQIMPUTE - Sequential Imputation - Verboven *et al* (2007): start from a complete subset $\boldsymbol{X}_c$ and impute the missing values in one observation at a time by minimizing the determinant of the augmented data set $\boldsymbol{X}^* = [\boldsymbol{X}_c; (\boldsymbol{x}^*)^t]$
- RSEQ - Robust Sequential Imputation - Vanden Branden and Verboven (2009): replace the sample mean and covariance by robust estimators; use the outlyingness measure proposed by Stahel (1981) and Donoho(1982)

Robust location and scatter for incomplete data in R - rrcovNA

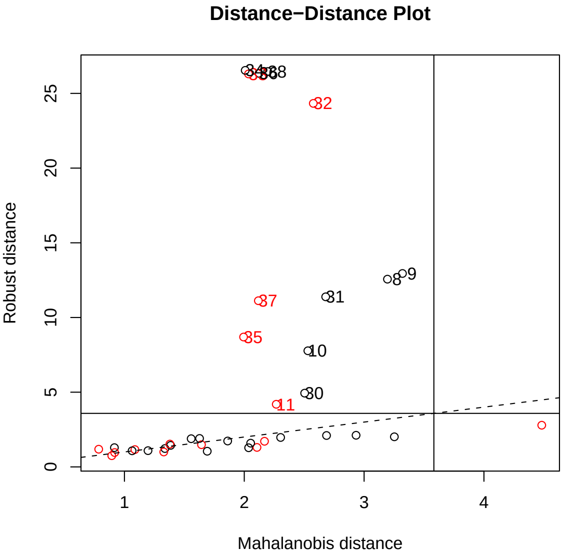
- **CovNAMcd** - Minimum Covariance Determinant
  - no imputation: Victoria-Feser and Copt (2004) or
  - normal imputation or
  - robust sequential imputation or
  - "other" robust imputation
- **CovNAOgk** - Pairwise cov estimator
  - same imputation methods as in CovNAMcd
- **CovNASest** - S estimates
  - same imputation methods as in CovNAMcd
  - several estimation methods FAST S, SURREAL, Bisquare, Rocke type
- **CovNASde** - Stahel-Donoho estimator
  - same imputation methods as in CovNAMcd
- **CovNABacon** - BACON-EEM algorithm as described in Béguin and Hulliger (2008)

Incomplete Data Plots in `rrcovNA` I



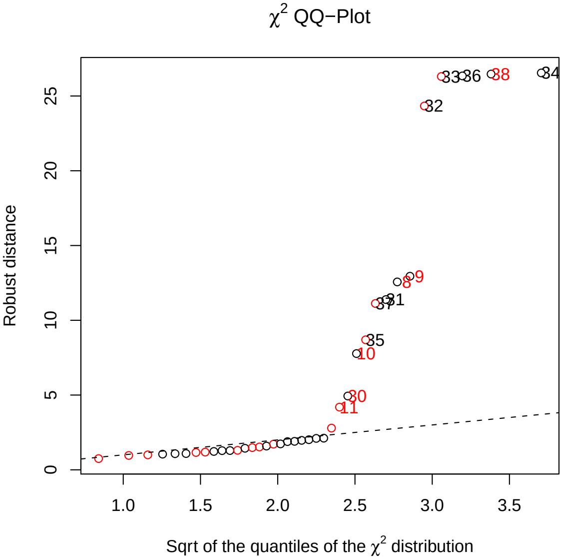
- Robust distances vs observation index - the outliers have large RD_i
- Observations with missing values are in red color
- A line is drawn at $y = cutoff = \sqrt{\chi^2_{p,0.975}}$
- The observations with $RD_i \geq cutoff = \sqrt{\chi^2_{p,0.975}}$ are identified by their index

Incomplete Data Plots in `rrcovNA` II



- Robust distances versus Mahalanobis distances
- Observations with missing values are in red color
- The dashed line is $RD_i = MD_i$
- The horizontal and vertical lines are $y = \sqrt{\chi^2_{p,0.975}}$ and $x = \sqrt{\chi^2_{p,0.975}}$

Incomplete Data Plots in `rrcovNA` III



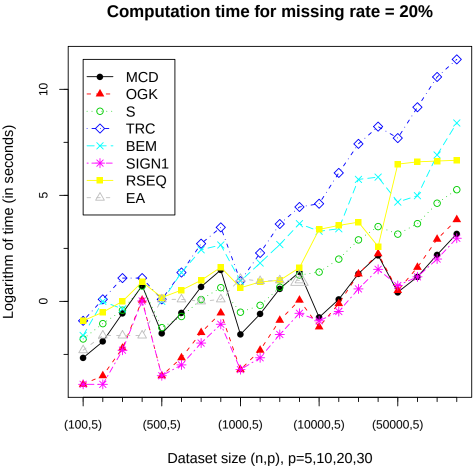
- A Quantile-Quantile comparison plot of the Robust distances versus the square root of the quantiles of the chi-squared distribution
- Observations with missing values are in red color

Implementation summary

All computations in this work were done in the `R` programming language using the following packages:

- MCD, OGK, S, SDE, RSEQ - `rrcovNA`, covariance estimation with `rrcov` - `CovMcd, CovOgk and CovSest`, normal imputation with `norm`
- EA, TRC and BACON-EEM - `R` code provided by the authors (Béguin and Hulliger, 2004)
- SIGN1 - package `mvoutlier`. In this package are available also the methods SIGN2 and PCOUT. The normal imputation was done with `norm`.
- EM-MCD and ERTBS - `R` code with compiled DLL library from the authors (Victoria-Feser and Copt, 2004)

Computation time of the algorithms.



- Large data sets - $n = 100 - 50000$, $p = 5 - 30$.
- 40% shift outliers and 20% MCAR; Average over 100 runs
- SIGN 1 is fastest, followed closely by OGK and MCD; MCD is faster than OGK for large n and p
- EA is fast but does not work when $n > 5000$
- TRC is very slow especially for n > 1000
- BEM and RSEQ - slightly better then TRC

Example: Bushfire data

SIMPLE EXPERIMENT WITH THE BUSHFIRE DATA

Béguin and Hulliger (2004)

- 12 outliers: **33-38**, **32**, **7-11**; 12 and 13 are suspect
- Missing values added with an MCAR mechanism
- Created 4 data sets: with 10%, 20%, 30% and 40% missing data
- For each method and data set the known outliers are indicated as detected or not
- Non-outliers that were classified as outliers, or swamped non-outliers are given too (**FP**=false positives)

Example: Bushfire data

OUTLIERS DETECTED BY SOME OF THE METHODS

MCD	7	8	9	10	11	12	13	32	33	34	35	36	37	38	FP
0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	3
0.1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
0.2	1	1	1	1	0	0	0	1	1	1	1	1	1	1	0
0.3	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0
0.4	1	0	1	1	1	0	1	0	0	0	0	0	0	0	2

TRC	7	8	9	10	11	12	13	32	33	34	35	36	37	38	FP
0	0	1	1	0	0	0	0	1	1	1	1	1	1	1	1
0.1	1	0	0	0	1	0	0	1	1	1	1	1	1	1	2
0.2	1	1	1	0	1	1	0	1	1	1	1	1	1	1	3
0.3	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1

S	7	8	9	10	11	12	13	32	33	34	35	36	37	38	FP
0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	3
0.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
0.2	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0
0.3	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0
0.4	1	1	1	1	1	0	0	1	0	1	1	1	1	0	0

Example: Bushfire data

- Repeat $m = 100$ times for each method and missingness rate
- Average the number of non-identified outliers and the number of regular observations declared outliers
- Methods: MCD, OGK, S, EA, TRC, BEM, SIGN1, SDE and RSEQ

Example: Bushfire data

AVERAGE PERCENTAGE OF OUTLIERS THAT WERE NOT IDENTIFIED

	0	10	20	30	40
MCD	0.00	2.00	5.42	19.58	24.58
OGK	0.00	4.17	14.83	27.75	39.42
S	0.00	6.75	13.50	29.25	38.67
EA	67.83	74.25	76.50	86.92	87.83
TRC	25.00	16.92	22.58	21.75	21.08
BEM	0.00	1.25	2.00	3.00	4.17
SIGN1	0.00	6.00	15.83	34.25	37.50
SDE	0.00	6.23	16.15	25.23	37.80
RSEQ	0.00	2.10	5.12	18.12	23.28

AVERAGE PERCENTAGE OF NON-OUTLIERS THAT WERE CLASSIFIED AS OUTLIERS

	0%	10%	20%	30%	40%
MCD	15.38	6.73	4.42	3.81	2.46
OGK	19.23	8.62	5.50	4.31	3.69
S	11.54	6.38	4.77	5.73	3.23
EA	1.88	1.27	1.58	1.08	1.73
TRC	3.85	10.81	10.96	8.42	7.54
BEM	7.69	11.39	10.42	12.92	17.27
SIGN1	23.08	15.58	10.85	7.58	6.46
SDE	1.72	3.28	1.68	1.72	1.60
RSEQ	15.22	6.03	3.82	3.80	2.40

Austrian Structural Business Statistics Data 2006

- More than 320.000 enterprises. Available raw data set: 21669 observations in 90 variables, structured according NACE revision 1.1 with 3891 missing values
- We investigate the following 10 variables of NACE 52.42 - "Retail sale of clothing"

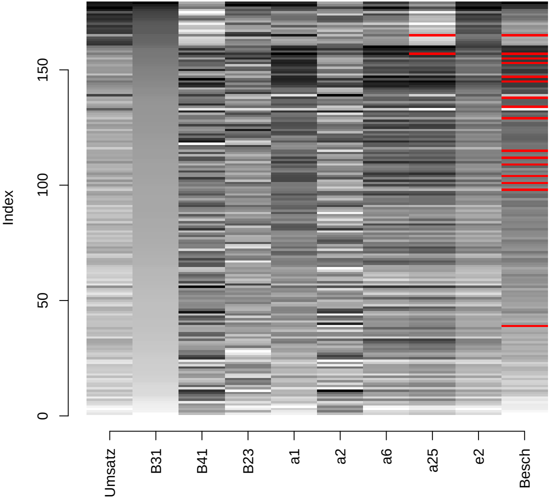
TURNOVER	Total turnover
B31	Number of white-collar employees
B41	Number of blue-collar workers
B23	Part-time employees
EMP	Number of employees
A1	Wages
A2	Salaries
A6	Supply of trade goods for resale
A25	Intermediate inputs
E2	Revenues from retail sales

Synthetic SBS data, NACE 5244

Missing value patterns analyzed with the R package **VIM**.

DATA MATRIX PLOT:

- **Missing values are red colored**

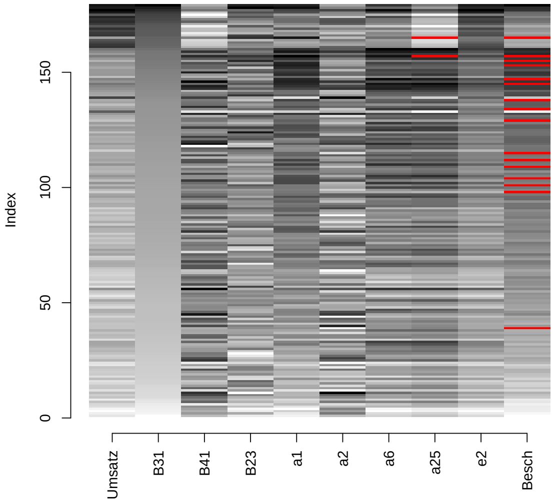


Synthetic SBS data, NACE 5244

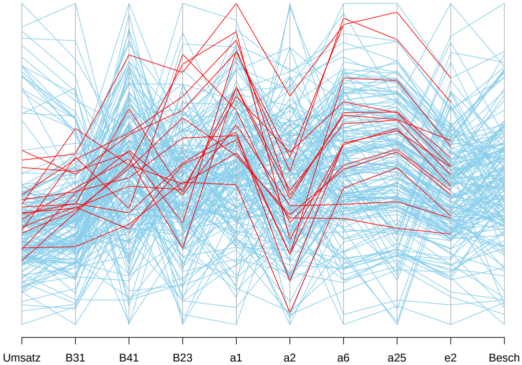
Missing value patterns analyzed with the R package **VIM**.

DATA MATRIX PLOT:

- Missing values are red colored
- **The darker a line the higher the value of an observation**



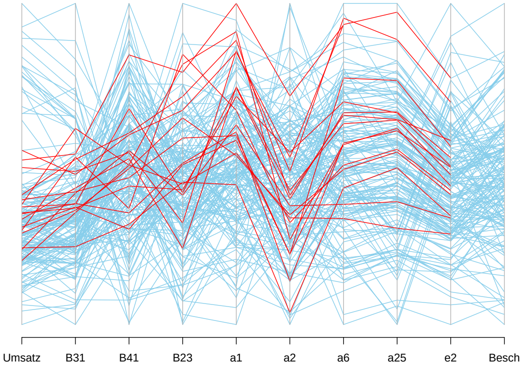
Synthetic SBS data, NACE 5244



PARALLEL COORDINATE PLOT:

- **Observations with Missing values in *EMP* are red colored**

Synthetic SBS data, NACE 5244



Parallel Coordinate Plot:

- Observations with Missing values in *EMP* are red colored
- → **MAR situation.**

Simulation Setup I

SIMULATION SETTINGS

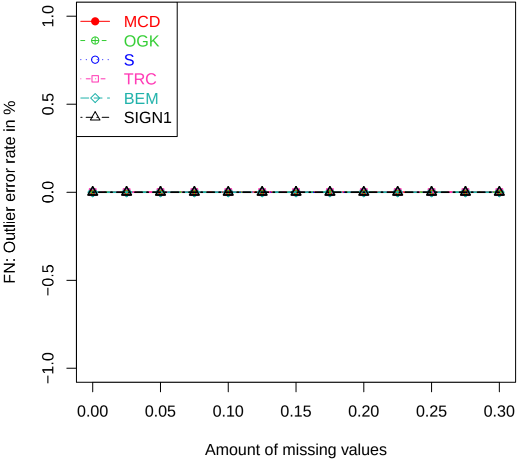
- Log-normal data generated according to the structure $(\mathbf{T}, \mathbf{C})$ and size of the original data.
- Two experiments:
  1. Fixed fraction of outliers = 0.1 and missing rates = $0.0, \ldots, 0.3$ with step 0.025
  2. Fixed missing rate = 0.1 and fractions of outliers = $0.0, \ldots$..., 0.25 with step 0.025
- Methods: **MCD, OGK, S, SIGN1, BEM, TRC**
- m=400 repeated for all data sets and methods

Simulation Setup II

WE COMPARE

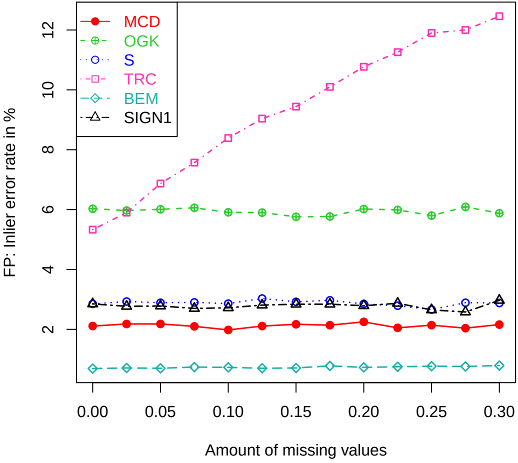
- The average percentage of **false negatives (FN)** - the outliers that were not identified, or masked outliers (outlier error rate)
- The average percentage of **false positives (FP)** - non-outliers that were classified as outliers, or swamped non-outliers (non-outlier error rate)
- Average computation time

Simulation results I



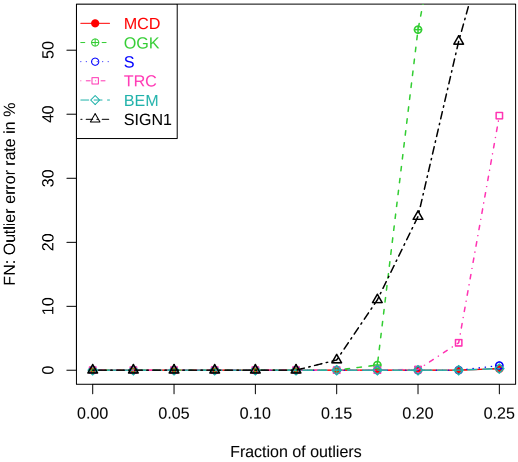
- **False Negatives (FN)** or outlier error rate
- Fixed fraction of outliers: 10%
- Varying percent of missingness
- Average over 400 runs
- All methods perform excellent and identify all outliers

Simulation results II



- **False Positives (FP)** or non-outlier error rate
- Fixed fraction of outliers: 10%
- Varying percent of missingness
- BEM, MCD, S and SIGN1 perform best (uniformly less than 3%) followed by OGK (6%). TRC diverges.

Simulation results III



- **False Negatives (FN)** or outlier error rate
- Fixed missingness rate: 10%
- Varying fraction of outliers
- BEM, MCD and S are best
- OGK and SIGN1 break down by less than 20% of outliers, TRC breaks down by 25%.

Simulation results IV

OUTLIER DETECTION IN SURVEY DATA

Todorov

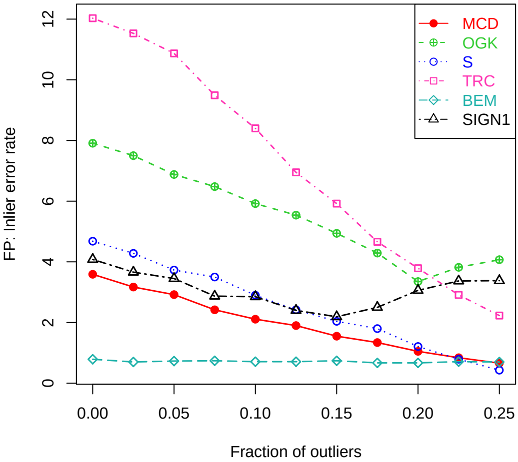
Multivariate Outliers

Algorithms for Outlier Detection

Examples

Simulation Study - SBS DATA

Summary and Conclusions



- **False Positives (FP)** or non-outlier error rate
- Fixed missingness rate: 10%
- Varying fraction of outliers
- In terms of non-outlier error rate BEM, MCD, S and SIGN1 perform best (uniformly less than 5%) followed by OGK (less than 8%).

Conclusions and Outlook

- We considered methods for identification of outliers in **large multivariate incomplete sample survey data**
- In this context the following new methods were proposed: **PM-OGK, PM-S, RSEQ and SIGN1**
- The methods were compared in terms of computation time and identification performance on examples and simulation study based on real data
- **ADVERTISING:** The considered methods are implemented in an `R` package `rrcovNA` based on `rrcov` and `robustbase` $\Rightarrow$ soon available on CRAN.
- **Outlook**
  - Sampling weights for MCD and S estimators.
  - What to do after the outliers are found? $\Rightarrow$ Development of a practical procedure for handling of multivariate outliers.