

SOFTWARE FOR MULTIVARIATE OUTLIER DETECTION IN SURVEY DATA

V. Todorov¹ M. Templ² P. Filzmoser³

¹United Nations Industrial Development Organization (UNIDO)

²Statistics Austria

³Vienna University of Technology

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Outline

- 1 Multivariate Outliers
- 2 Multivariate Location and Scatter
- 3 Handling of incomplete data
- 4 Principal Component Analysis for incomplete data
- 5 Summary

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 arouse suspicion that it was generated by different mechanism
- **Barnett 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 ← 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

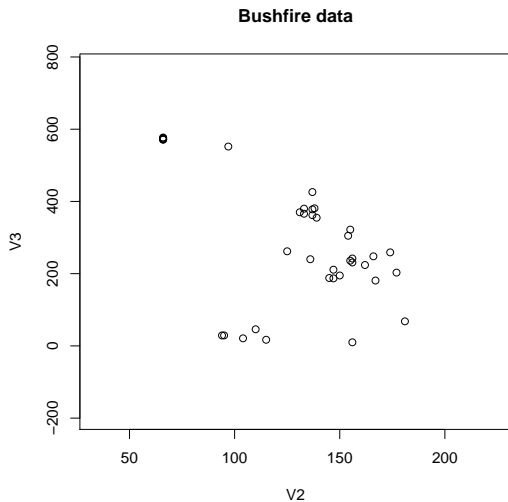
Outliers in Sample Surveys: Multivariate methods

- 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
- Todorov, Templ and Filzmoser (2011)
 - Investigated and compared many different methods for detection of multivariate outliers based on robust estimators
 - Example application to Structural Business Statistics data

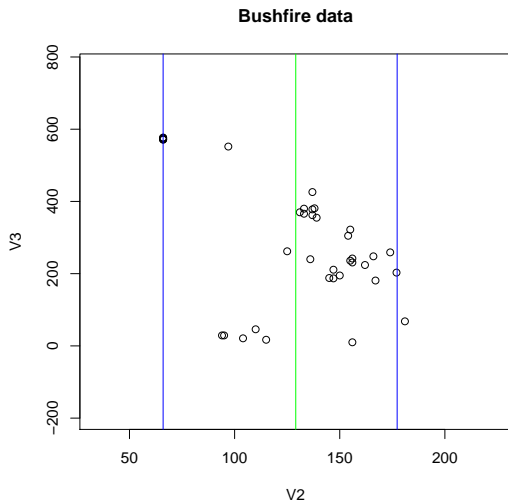
Example: Bushfire data

- 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`

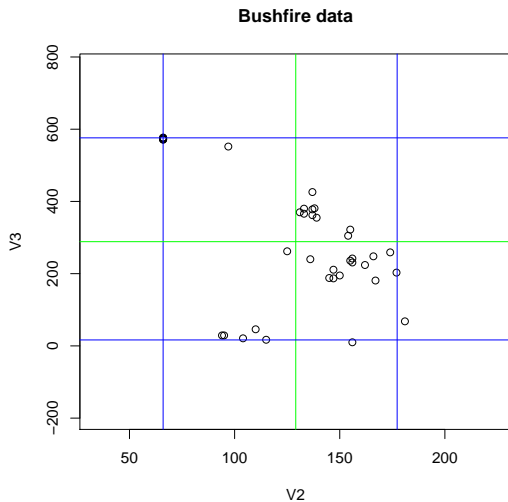
Example: Bushfire data



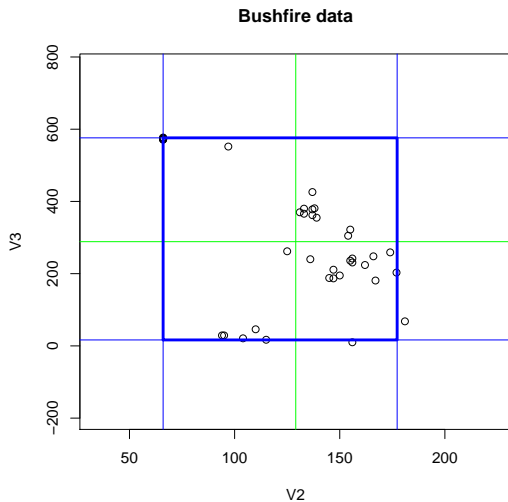
Example: Bushfire data



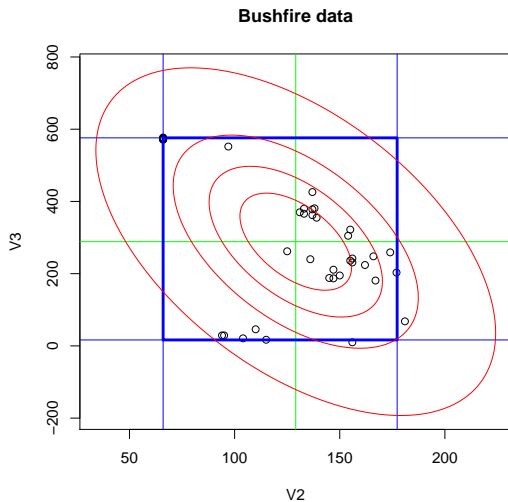
Example: Bushfire data



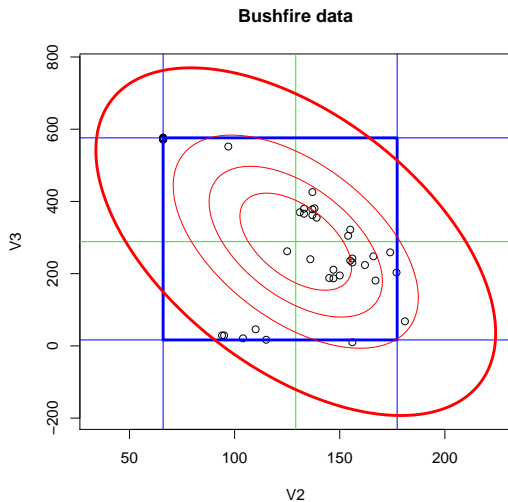
Example: Bushfire data



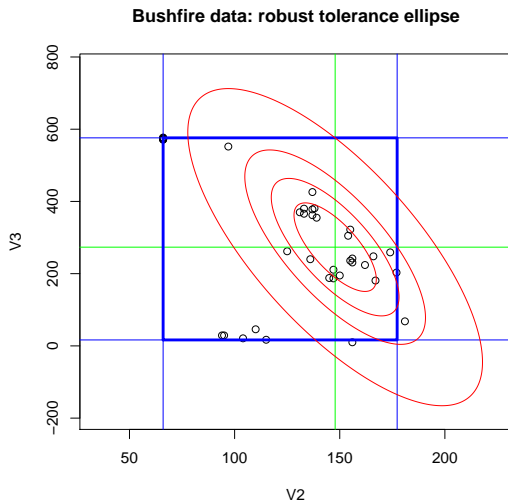
Example: Bushfire data



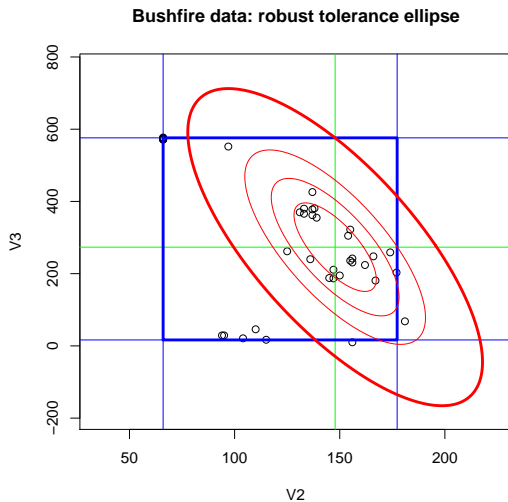
Example: Bushfire data



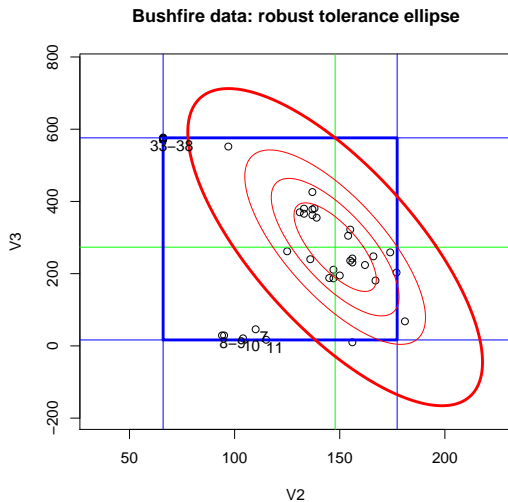
Example: Bushfire data



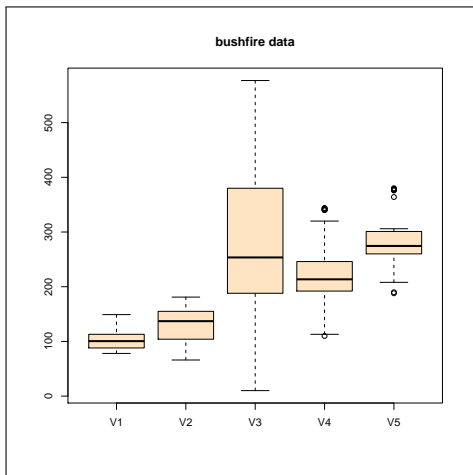
Example: Bushfire data



Example: Bushfire data



Example: Bushfire data - Boxplots



Outliers and Robustness

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 Robustness is "... insensitivity against small deviations from the assumptions" (Huber, 1987)
- ③ **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
 - It depends on the particular research, on which problem to set the focus

Multivariate Location and Scatter

- **Location:** coordinate-wise mean
- **Scatter:** covariance matrix
 - Variances of the variables on the diagonal
 - Covariance of two variables as off-diagonal elements
- Optimally estimated by the sample mean and sample covariance matrix at any multivariate normal model
- Essential to a number of multivariate data analysis methods
- But extremely sensitive to outlying observations

MCD Estimator

MCD-Estimator - *Minimum Covariance Determinant* (Rousseeuw, 1985)

- Find the subset of h observations out of n whose classical covariance matrix has a smallest determinant
- The MCD location estimator \mathbf{T} is defined by the mean of that subset and the MCD scatter estimator \mathbf{C} is a multiple of its covariance matrix.
- $h = \frac{(n+p+1)}{2}$ yields maximal breakdown point
- Fast algorithm to compute the MCD - (Rousseeuw and Van Driessen, 1999)

MVE-Estimator - *Minimum Volume Ellipsoid* (Rousseeuw, 1985)

- looks for the minimal volume ellipsoid covering at least half of the points

OGK Estimator

OGK-Estimator - *Orthogonalized Gnanadesikan-Kettenring* estimator (Maronna and Zamar, 2002; Gnanadesikan and Kettenring, 1972)

- Compute robust covariance matrix pairwise by

$$U_{jk} = \text{cov}(X_j, X_k) = \frac{1}{4}(\sigma(X_j + X_k)^2 - \sigma(X_j - X_k)^2)$$
 using a robust scale as σ .
- Use eigenvector/-value decomposition of $U = [U_{jk}]$ to construct robust principal components
- Obtain robust estimates of location and scatter
- Not affine equivariant, but extremely fast

Other estimators - M estimator, Stahel-Donoho, S estimator, MM-estimator

Multivariate Location and Scatter: Detection of Multivariate Outliers

Two phases (Rocke and Woodruff, 1996)

① Calculate **Robust Distances**

- Obtain robust estimates of location \mathbf{T} and scatter \mathbf{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 $Q = \chi_p^2(0.975)$ but see also Hardin and Rocke (2005), Filzmoser, Garrett, and Reimann (2005), Cerioli, Riani, and Atkinson (2008).

Algorithms: Handling Missing Values

Normal Imputation followed by HBDP estimation

- ER-algorithm - Little (1988) ← zero breakdown point (based on M-estimates)
- MCD - imputation under MVN model followed by MCD (R package `norm` and fast MCD implementation in package `rrcov`)
- OGK - imputation under MVN model followed by OGK (fast OGK implementation in package `rrcov`)
- 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 Velleman 2000) and EM

All three algorithms can handle sampling weights

Algorithms: Handling Missing Values

Robust Sequential Imputation followed by HBDP estimation

- SEQImpute - Sequential Imputation - Verboven *et al* (2007): start from a complete subset \mathbf{X}_c and impute the missing values in one observation at a time by minimizing the determinant of the augmented data set $\mathbf{X}^* = [\mathbf{X}_c; (\mathbf{x}^*)^t]$
- 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)

To start with: CovNAMcd example session

```
R> ##  
R> ## Load the 'rrcovNA' package and the data sets to be  
R> ## used throughout the examples  
R> ##  
R> library("rrcovNA")  
  
Scalable Robust Estimators with High Breakdown Point (version 1.2-02)  
Scalable Robust Estimators with High Breakdown Point for  
Incomplete Data (version 0.4-00)  
  
R> data("bush10")
```

To start with: CovNAMcd example session

```
R> ## Compute MCD estimates for the modified bushfire data set
R> ## - show() and summary() examples
R> mcd <- CovNAMcd(bush10)
R> mcd
```

```
Call:
CovNAMcd(x = bush10)
-> Method: Minimum Covariance Determinant Estimator for incomplete data.
```

```
Robust Estimate of Location:
      V1      V2      V3      V4      V5
109.5 149.5 257.9 215.0 276.9
```

```
Robust Estimate of Covariance:
      V1      V2      V3      V4      V5
V1    697.6   489.3 -3305.1 -671.4 -550.5
V2    489.3   424.5 -1889.0 -333.5 -289.5
V3   -3305.1 -1889.0 18930.9 4354.2 3456.4
V4   -671.4  -333.5  4354.2 1100.1  856.0
V5   -550.5  -289.5  3456.4  856.0  671.7
```

Example session: summary method of CovNAMcd

```
R> summary(mcd)
```

```
Call:
```

```
CovNAMcd(x = bush10)
```

```
Robust Estimate of Location:
```

V1	V2	V3	V4	V5
109.5	149.5	257.9	215.0	276.9

```
Robust Estimate of Covariance:
```

	V1	V2	V3	V4	V5
V1	697.6	489.3	-3305.1	-671.4	-550.5
V2	489.3	424.5	-1889.0	-333.5	-289.5
V3	-3305.1	-1889.0	18930.9	4354.2	3456.4
V4	-671.4	-333.5	4354.2	1100.1	856.0
V5	-550.5	-289.5	3456.4	856.0	671.7

```
Eigenvalues of covariance matrix:
```

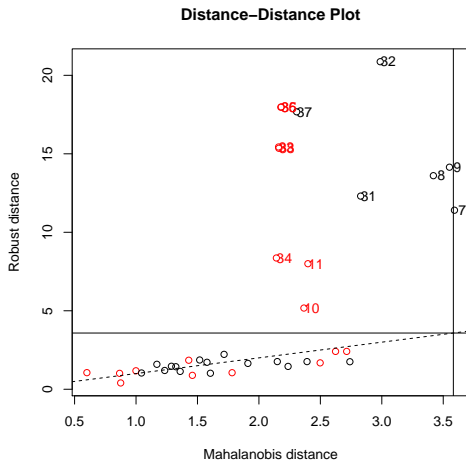
[1]	21334.429	428.703	56.662	3.701	1.263
-----	-----------	---------	--------	-------	-------

```
Robust Distances:
```

[1]	3.1071	1.1127	1.3864	1.1215	2.1500	3.0780	130.1256
[8]	185.1492	200.1491	26.7795	63.9884	5.8178	2.8298	4.9464
[15]	2.1220	3.1128	1.0421	2.7172	2.9548	2.0638	1.4335

Example session: plot method of CovNAMcd

```
R> plot(mcd)
```

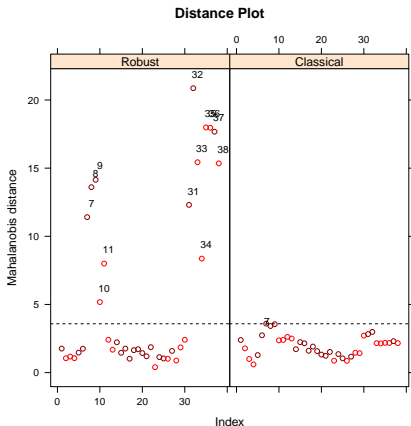


A propos - R Graphics

- R offers three main types of graphics: traditional (or base), **lattice**, and **ggplot2**. The latter two have as underlying graphics system the **grid** package.
 - The base graphics functions come with the main R installation and include high-level functions such as bar plots, histograms, and scatter plots - brief, with good default settings; Also include low-level functions for drawing points, lines, axes, etc which add flexibility and control for creating new types of graphs. **Paul Murrell (2005)**
 - The lattice graphics package, written by Deepayan Sarkar implements W. S. Cleveland's Trellis graphics system and now is part of the base R distribution. **Deepayan Sarkar (2007)**
 - The third graphics package, ggplot2, written by Hadley Wickham, is based on the grammar of graphics and offers an excellent balance between power and ease of use. **Hadley Wickham (2009)**
- All plots in **rrcov** are implemented in base graphics but many plots have already **lattice** alternatives

Example session: more plots for CovNAMcd: distance plot

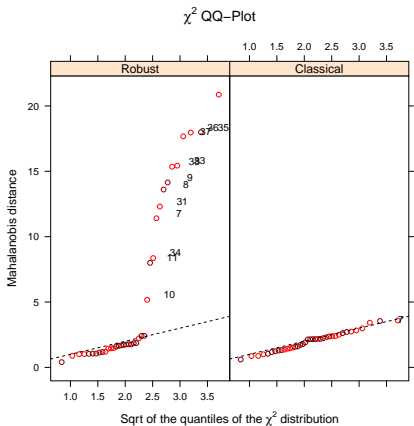
```
R> plot(mcd, which="xydistance")
```



- The robust distances are plotted versus their index - the outliers have large RD_i
- A line is drawn at $y = cutoff = \sqrt{\chi_{p,0.975}^2}$
- The observations with $MD_i \geq cutoff = \sqrt{\chi_{p,0.975}^2}$ are identified by their index
- The observations which have a **missing value** in any of the coordinates are shown in **red**.

Example session: more plots ... χ^2 QQ-plot

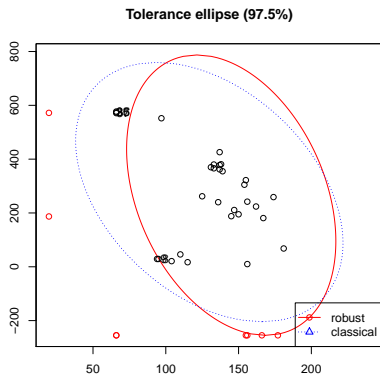
```
R> plot(mcd, which="xyqqchi2")
```



- 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
- The observations which have a **missing value** in any of the coordinates are shown in **red**.

Example session: more plots ... tolerance ellipses

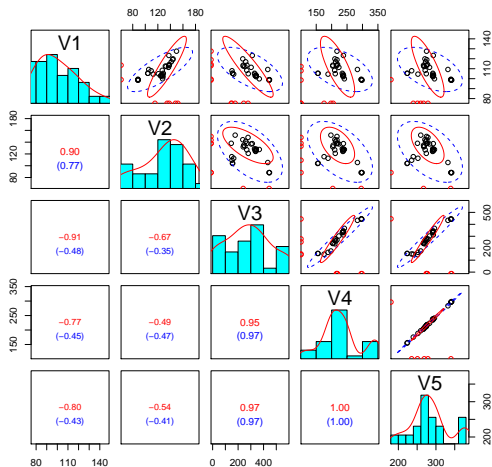
```
R> mcd2 <- CovNAMcd(bush10[,2:3])
R> plot(mcd2, which="tolEllipse", class=TRUE)
```



- Consider the case of bivariate data (variables V2 and V3 of bush10 data set) - see next slide for more than two variables
- Scatter plot of the data with superimposed 97.5% robust and classical tolerance ellipses
- The observations with $MD_i \geq \text{cutoff} = \sqrt{\chi_{p,0.975}^2}$ are identified by their subscript
- The observations which have a **missing value** in any of the coordinates are projected on the axis and are shown in **red**.

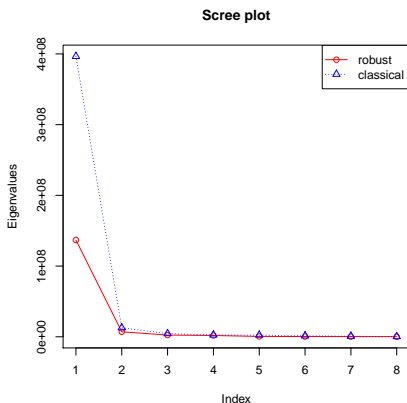
Example session: more plots ... pairs

```
R> plot(mcd, which="pairs")
```



Example session: more plots ... screeplot

```
R> data(ces)
R> plot(CovNAMcd(ces), which="screeplot")
```



- Robust and classical Scree plot of the data
- Eigenvalues comparison plot for the **Consumer Expenditure Survey Data** - Hubert et al. (2009).
- Find out if there is much difference between the classical and robust covariance (or correlation) estimates.

Handling of incomplete data in - package `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)
- **PcaNA** - Different robust PCA methods for incomplete data - see Serneels (2008) - to be discussed later in this presentation.

CovNARobust: a generalized function for robust location and covariance estimation for incomplete data - package `rrcovNA`

- `CovNARobust(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

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

Principal Component Analysis

- Data matrix \mathbf{X} with n rows (observations) and p columns (variables)
- Goal: Dimension reduction - the information contained in the multivariate data should be expressed by few components with minimal loss of information.
- Solution: Spectral decomposition of the covariance matrix of \mathbf{X} :

$$\hat{\Sigma} = \Gamma \Lambda \Gamma^t$$

- The (estimated) principal components:

$$\mathbf{Z} = (\mathbf{X} - \mathbf{1}\mu) \Gamma$$

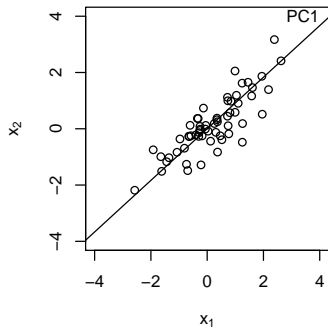
- Estimates (arithmetic mean and sample covariance matrix)

$$\hat{\mu} = \bar{\mathbf{x}} \text{ and } \hat{\Sigma} = \mathbf{S}$$

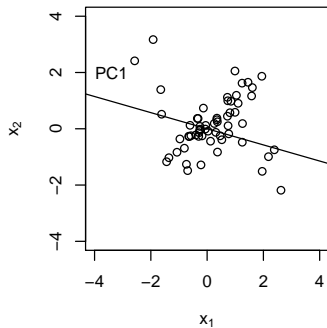
Example: Principal Component Analysis

- A bivariate data set with: $n = 60$, $\mu = (0, 0)$ and $\rho = 0.8$
- sample correlation: 0.84
- sort the data by the first coordinate x_1 and modify the first four observations with smallest x_1 and the last four with largest x_1 by interchanging their first coordinates
- thus (less than) 15% of outliers are introduced which are undistinguishable on the univariate plots of the data
- the sample correlation changes even its sign and becomes -0.05

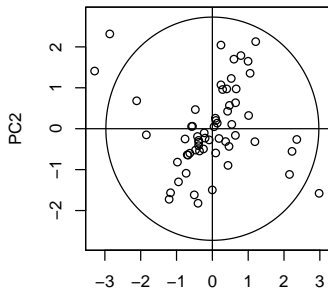
(a) Clean data



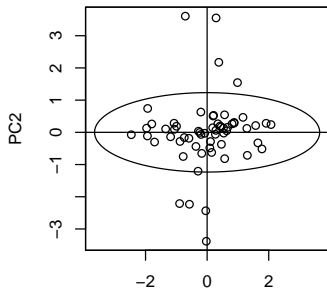
(b) Data with 15% outliers



(c) Classical



(d) Robust (MCD)



Principal Component Analysis for Incomplete data

- **Walczak and Massart (2001)** - use the EM approach applied to classical PCA.
- **Serneels and Verdonck (2008)** - use the EM approach applied to any robust PCA.
- Most of the available robust PCA methods are implemented in the R package **rrcov**
- The corresponding methods for dealing with incomplete data are available in the package **rrcovNA**

Principal Component Analysis for Incomplete data

- `PcaNA(x, k, method, cov.control, ...)`
- `PcaNA(formula, data = NULL, ...)`
- **method="cov"** - PCA based on a robust covariance matrix. Different estimators of the covariance matrix can be used.
- **method="proj"** - Projection pursuit approach (Croux and Ruiz-Gazen, 1996, 2005)
- **method="grid"** - Projection pursuit approach, enhanced search algorithm (Croux, Filzmoser, Oliveira, 2007)
- **method="hubert"** - ROBPCA, (Hubert, Rousseeuw and Vanden Branden, 2005)
- **method="locantore"** - Spherical PCA (Locantore et al., 1999)
- **method="class"** - Classical PCA.

Example session: PcaNA

```
R> data(bush10)
R> pca <- PcaNA(bush10)    # by default the MCD cov matrix will be used
R> pca
```

```
Call:
PcaNA.default(x = bush10)
```

Standard deviations:

```
[1] 145.627942  21.836183  10.869330   3.366705   1.574500
```

Loadings:

	PC1	PC2	PC3	PC4	PC5
V1	-0.1472945	0.2971574	0.4885216	0.80001612	-0.10640793
V2	-0.1015296	0.4235344	0.6815787	-0.57745246	0.11094823
V3	0.9573892	-0.1098694	0.2607165	0.05576422	-0.01587048
V4	0.1762042	0.6946248	-0.3738399	-0.07496570	-0.58401396
V5	0.1426675	0.4875872	-0.2984421	0.13339174	0.79689627

Example session: PcaNA

```
R> pca <- PcaNA(bush10, cov.control=CovControlSde()) # use SDE
R> pca
```

Call:

```
PcaNA.default(x = bush10, cov.control = CovControlSde())
```

Standard deviations:

```
[1] 167.012473  18.768804  11.356572   3.869779   1.389190
```

Loadings:

	PC1	PC2	PC3	PC4	PC5
V1	-0.13638664	0.2524196	0.4289942	0.85133076	-0.09424951
V2	-0.08679023	0.6449877	0.5753147	-0.48267359	0.11179214
V3	0.95054687	-0.1030613	0.2906412	0.03509577	-0.01161905
V4	0.20871804	0.6031312	-0.4970402	0.04013959	-0.58652306
V5	0.16359576	0.3819506	-0.3917342	0.19854281	0.79653954

Example session: summary method of PcaNA

```
R> summary(pca)
```

```
Call:
```

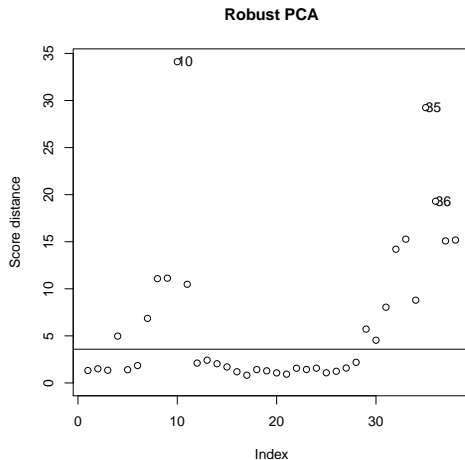
```
PcaNA.default(x = bush10, cov.control = CovControlSde())
```

```
Importance of components:
```

	PC1	PC2	PC3	PC4	PC5
Standard deviation	167.0125	18.76880	11.35657	3.86978	1.38919
Proportion of Variance	0.9825	0.01241	0.00454	0.00053	0.00007
Cumulative Proportion	0.9825	0.99486	0.99940	0.99993	1.00000

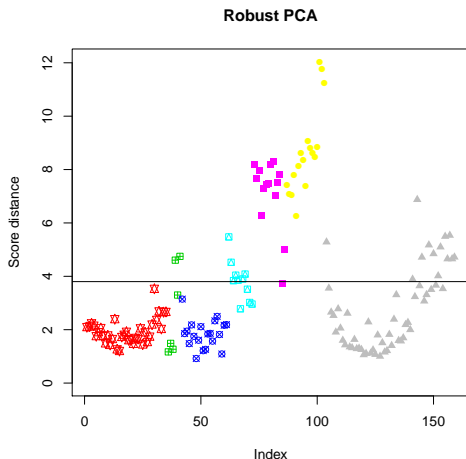
Example session: plot method of PcaNA

```
R> plot(pca, id.n.sd=3) # score distance plot
```



Example session: plot method of PcaNA

```
R> ## Score distance plot for the fish data set
R> data(fish);          ## obs. 14 has a missing value
R> plot(PcaNA(fish[,1:6]), col=fish[,7]+1, pch=fish[,7]+10, id.n.sd=0)
```



Score distances and orthogonal distances

- An outlier in the context of PCA is characterized by the following:
 - Lies far from the subspace spanned by the first k eigenvectors (large orthogonal distance) and/or
 - The projected observation lies far from the bulk of the data within this space (large score distance)
- The **score distances** are given by:

$$SD_i = \sqrt{\sum_{j=1}^k \frac{t_{ij}^2}{l_j}}, i = 1, \dots, n$$

where l_j are the eigenvalues

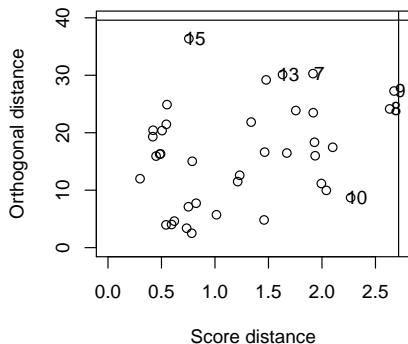
- The **orthogonal distances** of each observation to the subspace spanned by the first k principal components are defined by:

$$OD_i = \|\mathbf{x}_i - \boldsymbol{\mu}_i - \mathbf{P}\mathbf{t}_i\|, i = 1, \dots, n$$

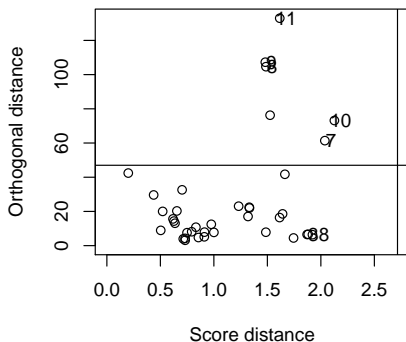
Outlier map (Hubert et al, 2005)

```
R> opar <- par(mfrow=c(1,2))
R> plot(PcaNA(bush10,k=2, method="class"), main="Classical PCA")
R> plot(PcaNA(bush10,k=2, method="hubert"), main="Robust PCA")
R> par(opar)
```

Classical PCA






Robust PCA



Summary and Conclusions

- We considered methods for identification of outliers in **large multivariate incomplete sample survey data** and their implementation in the statistical software environment R.
- Practical application of the methods and examples
- Available visualization and diagnostic tools
- The considered methods are implemented in the R package **rrcovNA** based on **rrcov** and **robustbase**
- These packages are available on CRAN:
<http://cran.r-project.org/>
- **Outlook**
 - Implementation of sampling weights for MCD and S estimators.
 - Handling of semi-continuous variables
 - What to do after the outliers are found? ⇒ Development of a practical procedure for handling of multivariate outliers.

References I

-  V. Todorov, M. Templ and P. Filzmoser
Detection of multivariate outliers in business survey data with incomplete information,
Advances in Data Analysis and Classification, **5**, 37–56, 2011
-  V. Todorov and P. Filzmoser
An object oriented framework for robust multivariate analysis,
Journal of Statistical Software, **32**(3), 1–47, 2009
<http://www.jstatsoft.org/v32/i03/>
-  M. Hubert, P.J. Rousseeuw and S. van Aelst
High-Breakdown Robust Multivariate Methods,
Statistical Science, **23**, 92–119, 2008.