Recent trends in robust statistics

Jānis Valeinis[1]

[1]University of Latvia, Riga

28th of May, 2011

Research at the University of Latvia in statistics

1. Nonparametric statistical procedures (Students - S.Vucāne, M.Vēliņa, L.Pahirko, E.Cers)
   - **Empirical likelihood method**;
   - Smooth nonparametric regression estimation;
   - Smooth nonparmetric density estimation;
   - Bootstrap methods.

2. Other research in mathematical statistics
   - Empirical process theory (J. Cielēns);
   - Goodness-of-fit tests for dependent data (A. Ločmelis);
   - Long memory processes (I. Dasmane);
   - Change-point analysis (A. Vaselāns).

Why robust statistics?

1. Collaboration with Prof.dr. George Lutta from the University of Georgetown.
2. Recent research activities in nonparametric robust statistical methods.
3. Valeinis, Velina and Lutta (2011). *Empirical likelihood-based inference for the difference of smoothed Huber estimators*. An abstract in International Conference on Robust Statistics (ICORS) in Valladolid, Spain.

Thanks to MMA conference in 2009 and publication:

- J.Valeinis, E.Cers, J.Cielēns (2010). Two-sample problems in statistical data modelling. *Mathematical modelling and analysis*, **15**(1), 137-151.

What is robust statistics?

Definition

Robust statistics provides an alternative approach to classical statistical methods. The motivation is to produce estimators that are not unduly affected by small departures from model assumptions (e.g. of normality)

- Classical statistical procedures are typically sensitive to "longtailedness" and "outliers"
- Goal: to obtain *distributionally robust* (or outlier-resistant) procedures.
- Foundators: John Tuckey (1960, 1962), Peter Huber (1964, 1967) and Frank Hampel (1971, 1974)

Data example: copper content in wholemeal flour

Table: Data example: copper content in wholemeal flour (Analytical Methods Committee, 1989)

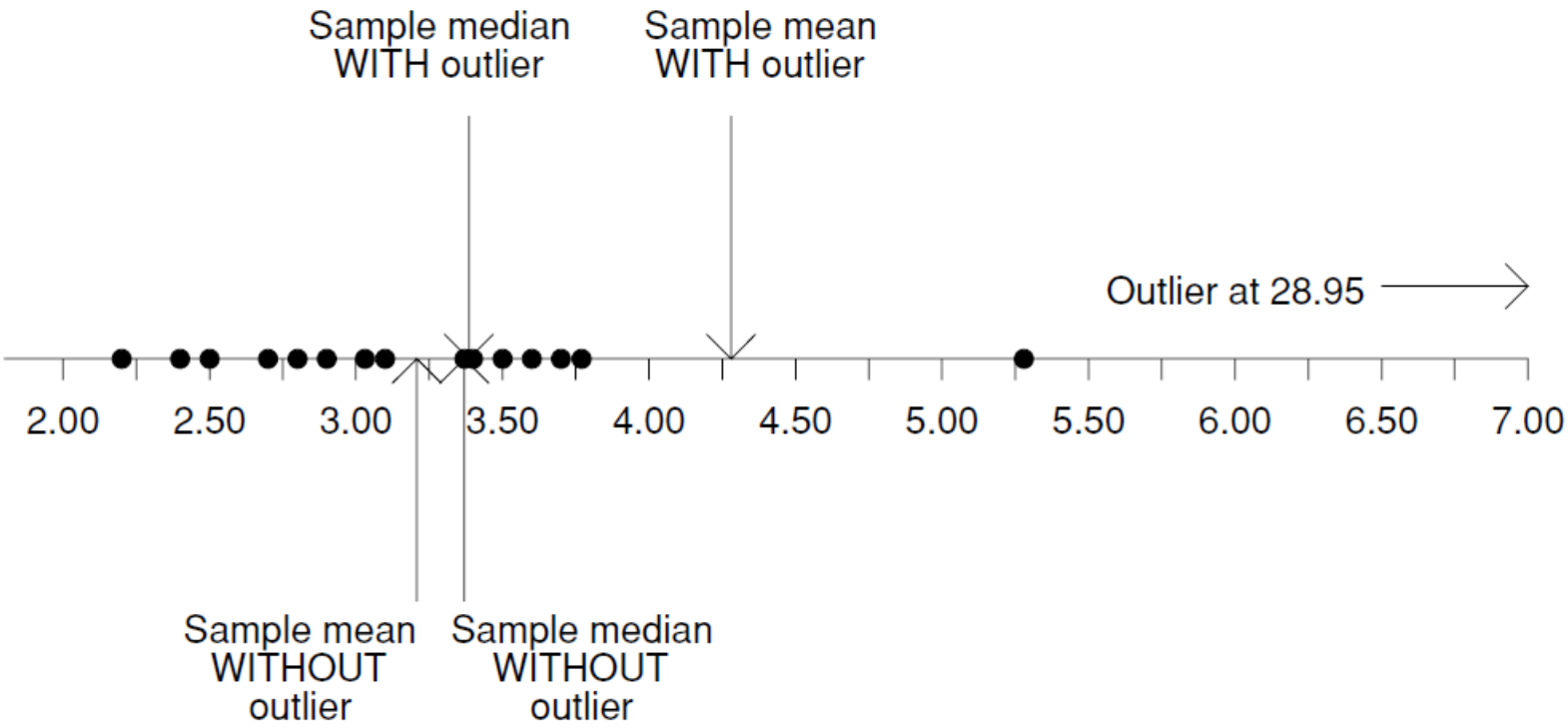
2.20	2.20	2.40	2.40	2.50	2.70	2.80	2.90
3.03	3.03	3.10	3.37	3.40	3.40	3.40	3.50
3.60	3.70	3.70	3.70	3.70	3.77	5.28	28.95

- The value 28.95 - an outlier!?

	with outlier	without outlier
$Med(X)$	3.38	3.37
$\bar{X}$	4.28	3.21
S	5.30	0.69
t-test CI	(2.05, 6.51)	(2.91, 3.51)

- Problem: A single outlier has an unbounded influence on statistics $\bar{X}$ and S (not on the sample median).

Data example: copper content in wholemeal flour



- Sample mean $\bar{X}$ is not a robust location estimate.
- Sample median $Med(X)$ is a robust location estimate!

Data example: copper content in wholemeal flour

Why not just delete the outlier?

Kandel (1991): "The discovery of the ozone hole was announced in 1985 by a British team working on the ground with "conventional" instruments and examining its observations in detail. Only later, after reexamining the data transmitted by the TOMS instrument on NASA's Nimbus 7 satellite, was it found that the hole had been forming for several years. Why had nobody noticed it?

The reason was simple: the systems processing the TOMS data, designed in accordance with predictions derived from models, which in turn were established on the basis of what was thought to be "reasonable", had rejected the very ("excessively") low values observed above the Antarctic during the Southern spring. As far as the program was concerned, there must have been an operating defect in the instrumemt"!

M-estimates of location: motivation

Definition

The *location model*

$$X_i = \mu + U_i,$$

where $U_1, \ldots, U_n$ are iid with $U_1 \sim F_0$.

- $X_1, \ldots, X_n$ are iid with $F(x) = F_0(x - \mu)$.
- The likelihood function

$$L(X_1, \ldots, X_n; \mu) = \prod_{i=1}^{n} f_0(X_i - \mu).$$

- Estimate of μ:

$$\hat{\mu} = \arg\max_{\mu} L(X_1, \ldots, X_n; \mu),$$

M-estimates of location: motivation

Definition

Equivalently an estimate of μ with $\rho = -\log f_0$.

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(X_i - \mu),$$

- If $F_0 = N(0,1)$, then $\rho(x) = x^2/2$ and

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2.$$

- If F_0 is the double exponential distribution, then $\rho(x) = |x|$ and

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} |X_i - \mu|.$$

M-estimates of location

Definition

Given a function ρ, an M-estimate of location is a solution of

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(X_i - \mu).$$

or equivalently if ρ is differentiable, then

$$\sum_{i=1}^{n} \psi(X_i - \hat{\mu}) = 0.$$

M-estimates of location

Definition

Huber's (1964) ρ and ψ functions

$$\rho_k(x) = \begin{cases} x^2 & \text{if } |x| \leq k \\ 2k|x| - k^2 & \text{if } |x| > k, \end{cases}$$

$$\psi_k(x) = \begin{cases} x & \text{if } |x| \leq k \\ sgn(x)k & \text{if } |x| > k, \end{cases}$$

- If $k \to \infty$ we obtain the sample mean;
- If $k \to 0$ we obtain the sample median;
- For Normal distribution standard choice is $k = 1.28$ (0.9th quantile of $N(0, 1)$).

Other robust estimates of location and scale

Definition

Median absolute deviation (MAD) about the median estimator of the scale

$$MAD(\mathbf{X}) = MAD(X_1, X_2, \ldots, X_n) = Med\{|\mathbf{X} - Med(\mathbf{X})|\}.$$

Normalized MAD

$$MADN(\mathbf{X}) = \frac{MAD(\mathbf{X})}{0.6745}.$$

Definition

Simultaneous M-estimate of location and dispersion

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - \hat{\mu}}{\hat{\sigma}}\right) = 0.$$

Data example: copper content in wholemeal flour

Table: Data example: copper content in wholemeal flour

	with outlier	without outlier
$\bar{X}$	4.28	3.21
$Med(X)$	3.38	3.37
$HuberM$	3.22	3.19
S	5.30	0.69
$MADN$	0.53	0.50
t-test CI	(2.05, 6.51)	(2.91, 3.51)

- Why not always use $Med(X)$ and $MADN(X)$?
- Answer - if there are no outliers these statistics have poorer behavior than usual ones.

Measuring robustness

Definition

The sensitivity curve of the estimate $\hat{\mu}$ for the sample $X_1, X_2, \ldots, X_n$ is the difference

$$\hat{\mu}(X_1, X_2, \ldots, X_n, x_0) - \hat{\mu}(X_1, X_2, \ldots, X_n)$$

as a function of the location x_0 of the outlier.

- Sensitivity curves show how any statistic is affected by an additional observation having value x_0.
- For robust statistics it should be bounded!

Sensitivity curves of location estimates

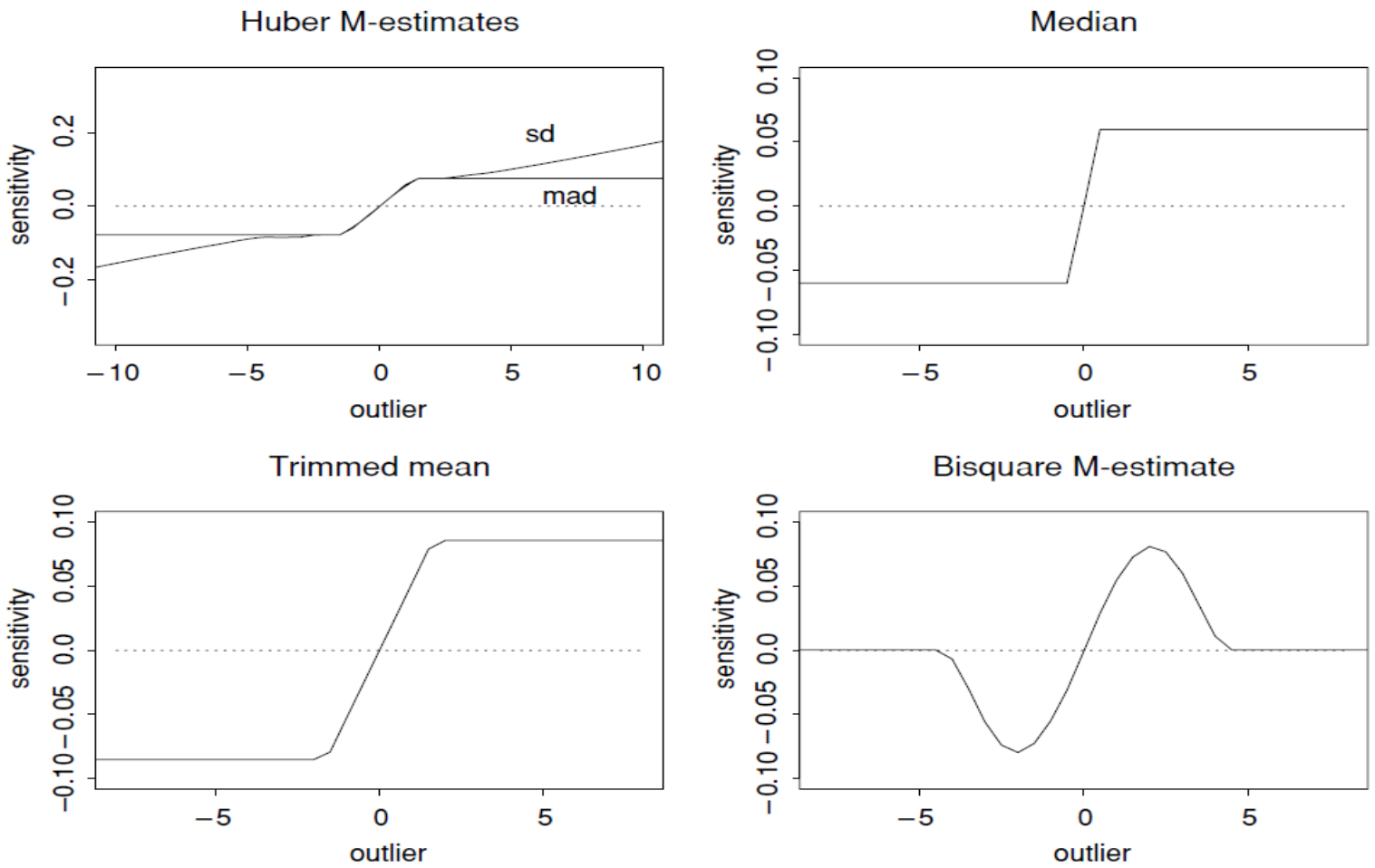


Figure: Artificial data set: simulated data with $n = 20$ from $N(0, 1)$.

Sensitivity curves of dispersion estimates

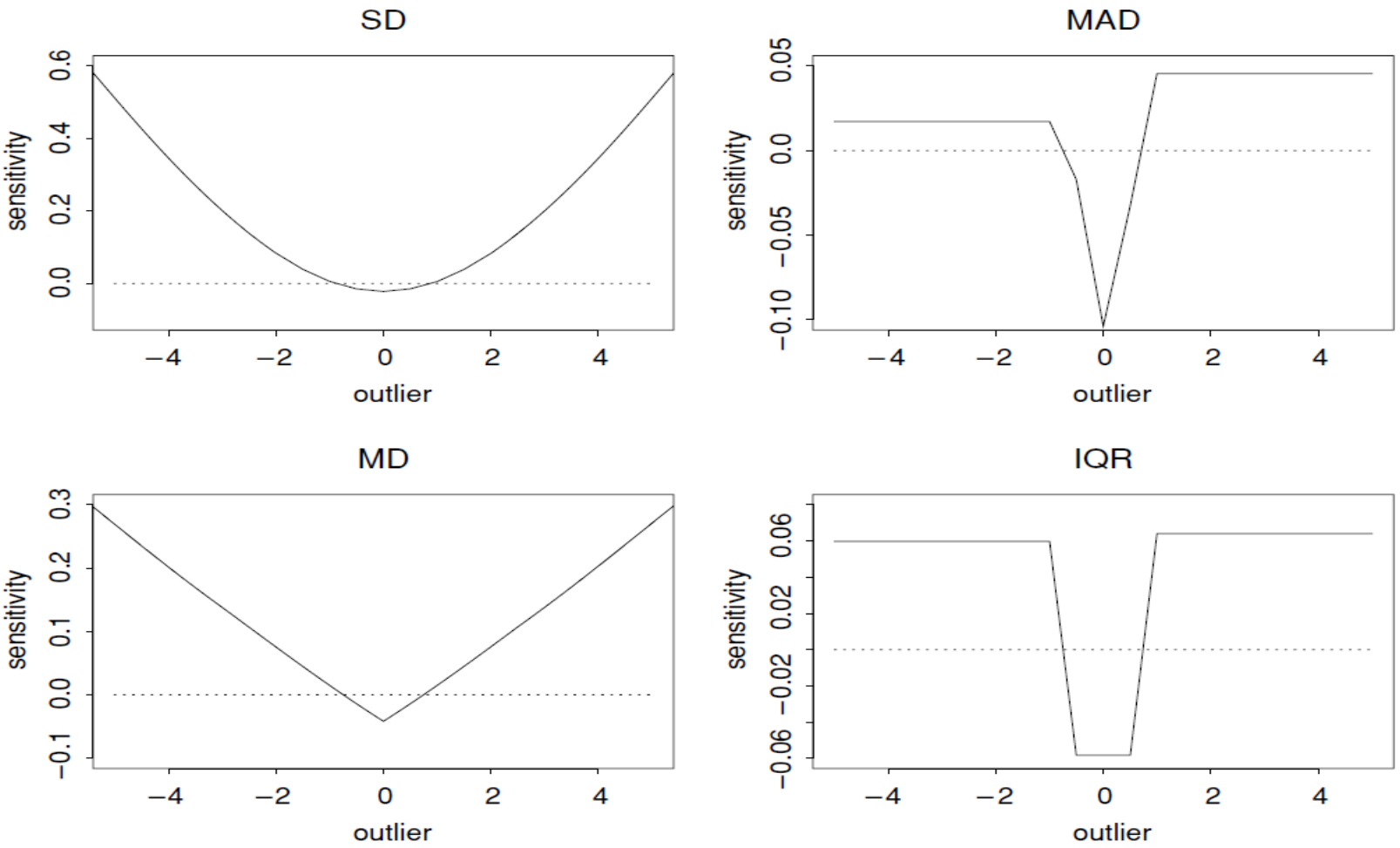


Figure: Artificial data set: simulated data with $n = 20$ from $N(0, 1)$ ($MD = E(|X - EX|)$, IQR - interquartile range).

Influence functions

Definition

For a statistical functional $T(F)$ when a sample contains a small fraction ϵ of identical outliers Influence function is defined as

$$IF(x_0) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon\delta_{x_0}) - T(F)}{\epsilon}$$

- $T(F)$ is said to have infinitesimal robustness if IF is bounded.
- The influence function (IF) is an asymptotic version of its sensitivity curve.
- For location M-estimate

$$IF(x_0) = \frac{\psi(x_0 - \hat{\mu})}{E\psi'(x_0 - \hat{\mu})}.$$

Contaminated models $F = (1-\epsilon)G + \epsilon H$

- Location model $X_i = \mu + U_i$, where $U_i \sim N(0, \sigma^2)$
  1. $\bar{X} \sim N(\mu, \sigma^2/n)$.
  2. $Med(X) \sim N(\mu, 1.57\sigma^2/n)$.

- The median has a 57% increase in variance relative to sample mean. It has a *low efficiency* at the normal distribution.

- For heavy tailed distributions characterized by contamination, the median will have high efficiency at the normal distribution.

$$F = (1-\epsilon)N(\mu, 1) + \epsilon N(\mu, \tau^2)$$

$$Var(\bar{X}) = \frac{1-\epsilon+\epsilon\tau^2}{n}, \quad Var(Med(X)) = \frac{\pi}{2n(1-\epsilon+\epsilon/\tau)^2}.$$

Contaminated models $F = (1-\epsilon)N(\mu, 1) + \epsilon N(\mu, \tau^2)$

Table: variances ($\times n$) of mean and median for large n.

ε	0.05		0.10	
τ	$n\,\mathrm{Var}(\overline{x})$	$n\,\mathrm{Var}(\mathrm{Med})$	$n\,\mathrm{Var}(\overline{x})$	$n\,\mathrm{Var}(\mathrm{Med})$
3	1.40	1.68	1.80	1.80
4	1.75	1.70	2.50	1.84
5	2.20	1.70	3.40	1.86
6	2.75	1.71	4.50	1.87
10	5.95	1.72	10.90	1.90
20	20.9	1.73	40.90	1.92

- Tradeoff between robustness and efficiency!

Robust regression

Definition

Parametric linear regression M-estimate $\hat{\beta}$ is the solution to

$$\sum_{i=1}^{n} \psi\left(\frac{Y_i - \hat{Y}_i(\hat{\beta})}{\hat{\sigma}}\right) X_i = 0.$$

Definition

Nonparametric regression M-estimate $\hat{m}$ is the solution to

$$\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \psi(Y_i - \hat{m}(x)) = 0.$$

Recent results

1. Empirical likelihood and robust statistics
   - F. Hampel *et al.* (2011). A smoothing principle for the Huber and other location *M*-estimators. *Computational Statistics and Data Analysis*, **55**(1), pages 324-337.
   - J. Valeinis, M. Velina and G. Lutta (2011). Empirical likelihood-based inference for the difference of smoothed Huber estimators. *ICORS conference abstract*.
2. Nonparametric regression and robust statistics
   - G. Boente *et al.* (2010). On a robust local estimator for the scale function in heteroscedastic nonparametric regression. *Statistics & Probability Letters*, **80**(15-16), pages 1185–1195.
   - H. Dette and M. Marchlewski (2010). A robust test for homoscedasticity in nonparametric regression. *Journal of Nonparametric Statistics*, **22**(6), pages 723–736.