Robust Tools for Statistical Data Editing and Imputation

WADA Kazumi (Tsuda University) and **TSUBAKI Hiroe** (The Institute of Statistical Mathematics (ISM)

I. M-estimators for regression (with two weight functions and scale parameters)
II. M-estimators for generalised ratio model (with Tukey's biweight function and average absolute deviation (AAD) for scale parameter)
III. Modified Stahel-Donoho estimators

I. M-estimators for regression:

$$y = \alpha + \boldsymbol{\beta} x + \varepsilon, \quad \varepsilon \sim N(0, \sigma), \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$$

Features

- IRLS (Iteratively Reweighted Least Squares) based on Bienias et al. (1997) in *Statistical Data Editing 2 –Methods and Techniques* by Tukey's biweight function with AAD scale.
- R functions by Huber's weight function, and also with MAD scale are implemented for comparison
- Tukey's biweight function is more suitable with AAD scale
- AAD scale converge faster than MAD scale

R functions available at https://github.com/kazwd2008/IRLS

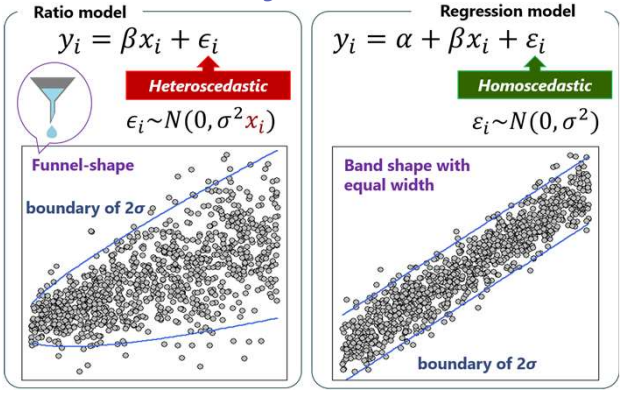
File name	R function	Weight function	Scale parameter
Tirls.r	Tirls.aad	Tukey	AAD
	Tirls.mad		MAD
Hirls.r	Hirls.aad	Huber	AAD
	Hirls.mad		MAD

* AAD: average absolute deviation $\mathrm{mean}(|r_i - \mathrm{mean}(r_i)|)$
* MAD: median absolute deviation $\mathrm{median}(|r_i - \mathrm{median}(r_i)|)$
* r_i: residual

II. M-estimators for generalised ratio model:

$$y = \beta x + \varepsilon x^{\gamma}, \varepsilon \sim N(0, \sigma)$$

Deference between the regression model and the ratio model



Robustification and generalisation

1. **Making the error term homoscedastic**

$\epsilon \sim N(0, x\sigma^2)$: $y_i = \beta x_i + \epsilon_i$ $\quad \because \varepsilon_i = \frac{\epsilon_i}{\sqrt{x_i}}$

$\varepsilon \sim N(0, \sigma^2)$: $y_i = \beta x_i + \varepsilon_i \sqrt{x_i}$

2. **Robustification**

$$\hat{\beta}_{rob} = \frac{\sum w_i y_i}{\sum w_i x_i}$$

Quasi-residual: $\tilde{\varepsilon}_i = \frac{y_i \sqrt{w_i}}{\sqrt{x_i}} - \hat{\beta}_{rob} \sqrt{w_i x_i}$

Weight function: Tukey's biweight (c=8)

Scale parameter: $\hat{\sigma}_{\mathrm{AAD}} = \frac{1}{n} \sum_{i=1}^{n} |\tilde{\varepsilon}_i|$

* AAD: average absolute deviation

3. **Generalisation**

Error term proportional to $x_i^{1/2} \Rightarrow x_i^{\gamma}$

model $\frac{y_i}{x_i^{\gamma}} = \beta x_i^{(1-\gamma)} + \varepsilon_i$

estimator $\hat{\beta} = \frac{\sum y_i x_i^{1-2\gamma}}{\sum x_i^{2(1-\gamma)}}$

robustification $\hat{\beta}_{rob} = \frac{\sum w_i y_i x_i^{1-2\gamma}}{\sum w_i x_i^{2(1-\gamma)}}$

* γ : an arbitral constant

R functions available at https://github.com/kazwd2008/IRLS

File name	R function		Feature	Weight function	Scale parameter
RrT.r*	RrTa.aad	γ=1	Robust estimation for generalized ratio model with a fixed γ value (All functions in RrT.r and RrH.r are included in the REGRM package at https://github.com/kazwd2008/REGRM.)	Tukey	AAD
	RrTb.aad	γ=1/2			
	RrTc.aad	γ=0			
	RrTa.mad	γ=1			MAD
	RrTb.mad	γ=1/2			
	RrTc.mad	γ=0			
RrH.r*	RrHa.aad	γ=1		Huber	AAD
	RrHb.aad	γ=1/2			
	RrHc.aad	γ=0			
	RrHa.mad	γ=1			MAD
	RrHb.mad	γ=1/2			
	RrHc.mad	γ=0			
RBreds.r	Rbred	Robust estimation for generalized ratio model (γ and β are simultaneously estimated)		Tukey	AAD
	Bred	Non robust estimation for generalized ratio model (γ and β are simultaneously estimated)			

III. Modified Stahel-Donoho (MSD) estimator

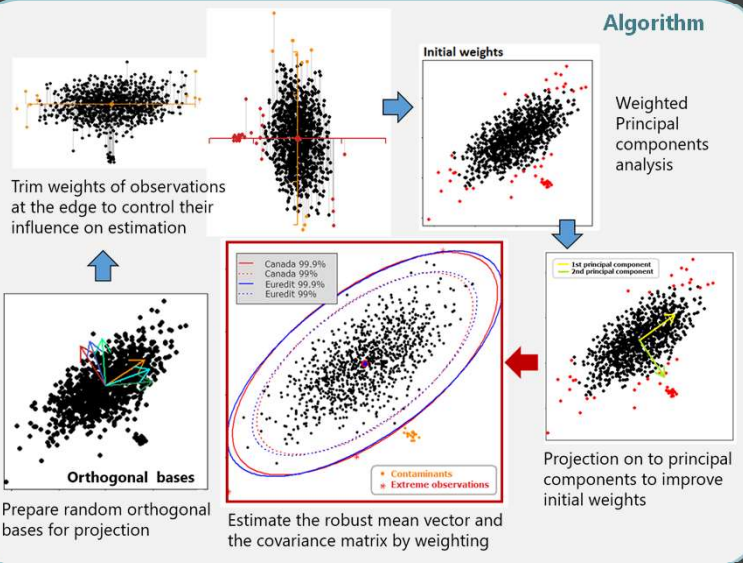
Features

- Multivariabe outlier detection method estimating mean vector and covariance matrix
- Adapted for Annual Wholesale and Retail Trade Survey (AWRTS) in Statistics Canada (Franklin & Brodeur, 1997)
- Suggestions by Béguin & Hulliger (2003) improve the performance while having a problem of the curse of dimensionality due to the increase of orthogonal bases (Wada, 2010)
- Higher performance compared with BACON, NNVE and MCD (Wada, Kawano & Tsubaki, 2020), while computationally expensive with high dimensional data

Parallelisation

- Paralleled version to cope with higher dimensional data is implemented (Wada & Tsubaki, 2013)

Algorithm



Manufacturing industry (Unincorporated Enterprise Survey)

Upper triangular matrix: square root transformation
Lower triangular matrix: biquadratic root transformation

Practical application

2016 Economic Census for Business Activities
- Robust estimator for generalised ratio model (γ=1/2)

2019 Unincorporated Enterprise Survey
- Robust estimator for generalised ratio model (γ=1/2)
- MSD estimators (single core version)

Statistics Japan