Robust linear and logistic regression in high dimension

Fatma Sevinc Kurnaz[1], Irene Hoffmann[2], Peter Filzmoser[2]

[1] Department of Statistics
Yildiz Technical University
Istanbul, Turkey

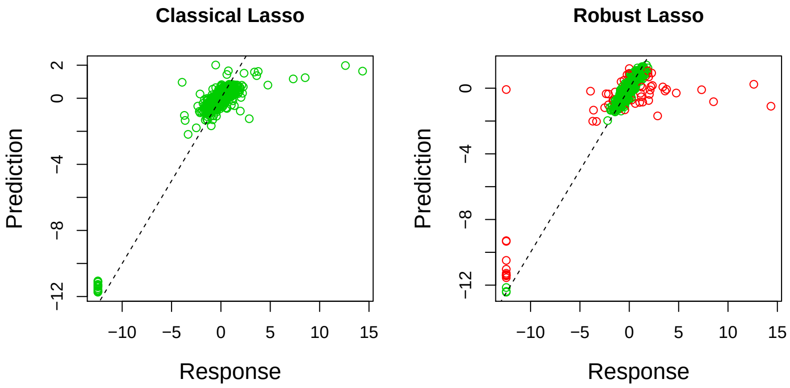
[2] Institute of Statistics and Mathematical Methods in Economics
Vienna University of Technology
Vienna, Austria

June 18, 2018, MedUni Wien

Outline

Chemical production process

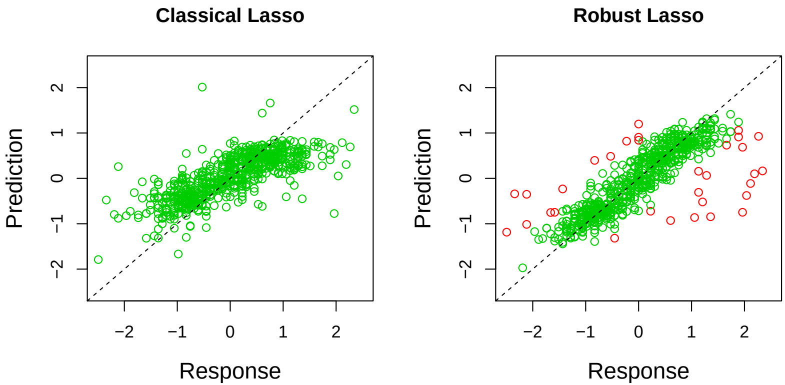
Response "Quality" (continuous variable – regression setting) is modeled with 468 features, 684 observations.



Left: 16 active variables, right: 75 active variables

Chemical production process

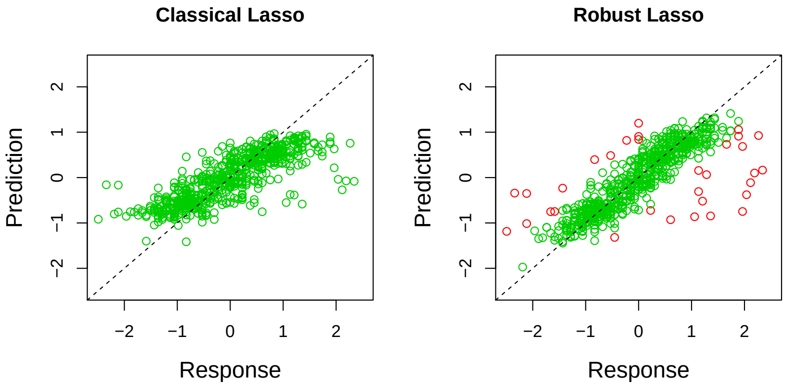
Response "Quality" (continuous variable – regression setting) is modeled with 468 features, 684 observations.



Zoom into the main data part.

Chemical production process

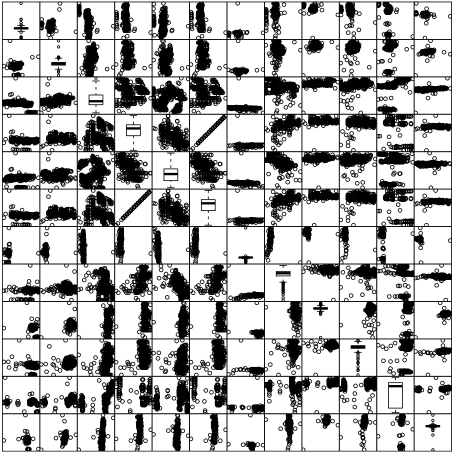
Omit obvious outliers in the response for Classical Lasso – but impossible to remove outliers in explanatory variables.



Left: 12 active variables, right: 75 active variables

Chemical production process

12 active variables from classical Lasso regression:



The setting

Linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{X} \in \mathbb{R}^{n \times p}$ predictor data matrix (centered, scaled) with observations $\mathbf{x}_i$, $i = 1, \ldots, n$, and p variables

$\mathbf{y} = (y_1, \ldots, y_n)^T$ response (centered)

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ coefficient vector

$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ error term

The setting

Esimating β in the linear regression model:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

by ordinary least squares (OLS) has problems with:

- many predictors: $n < p$
- multicollinearity
- uninformative (noise) variables
- outliers

Linear regression

Ordinary Least-Squares (OLS) regression: minimize sum of squared residuals

$$\sum_{i=1}^{n} r_i^2(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2$$

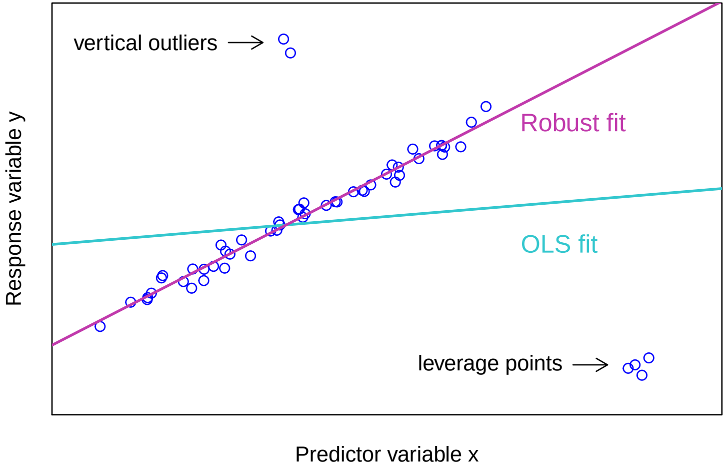
Robust alternative: LTS (Least Trimmed Squares) regression:

Sort squared residuals: $r_{(1)}^2(\beta) \leq \ldots \leq r_{(h)}^2(\beta) \leq \ldots \leq r_{(n)}^2(\beta)$
Minimize trimmed sum:

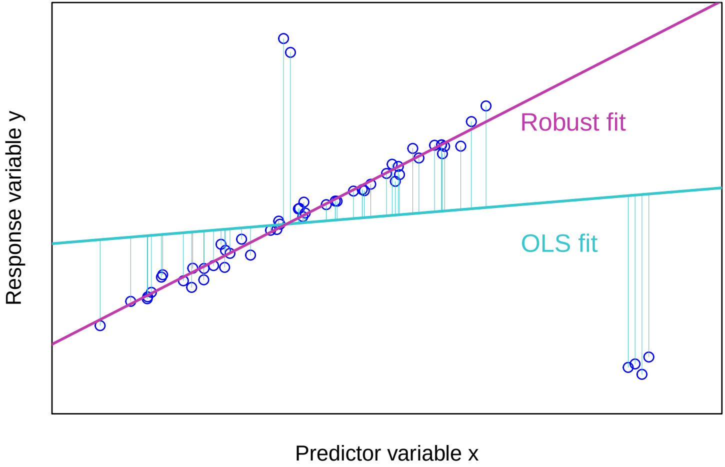
$$\sum_{i=1}^{h} r_{(i)}^2(\beta)$$

for some h in $[n/2, n]$.

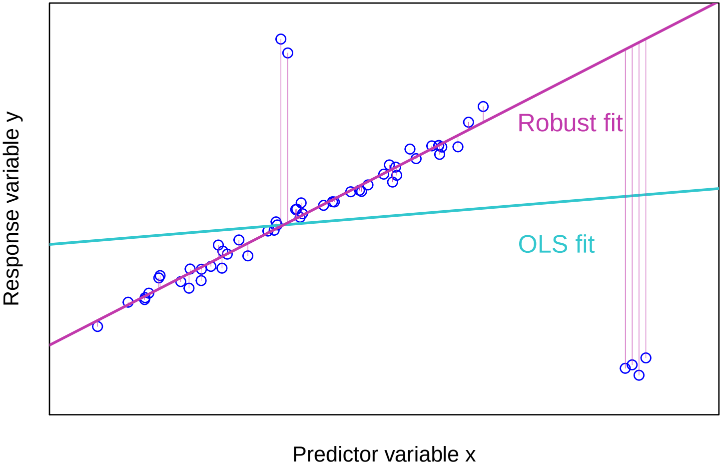
Outliers in linear regression



Outliers in linear regression



Outliers in linear regression



Fast-LTS algorithm

Key feature: *concentration steps* (C-steps)

1. Select a subset of $h \leq n$ observations.
2. Compute the OLS solution with the subset.
3. Construct next subset of size h from the observations corresponding to the h smallest squared residuals.

Value of LTS objective function gets successively smaller (until convergence).

Start the algorithm with random subsets of size p (here $p < n$).

P.J. Rousseeuw, K. Van Driessen, **Computing LTS regression for large data sets**, *Data Mining and Knowledge Discovery*, 12(1) (2006) 29–45.

Logistic regression

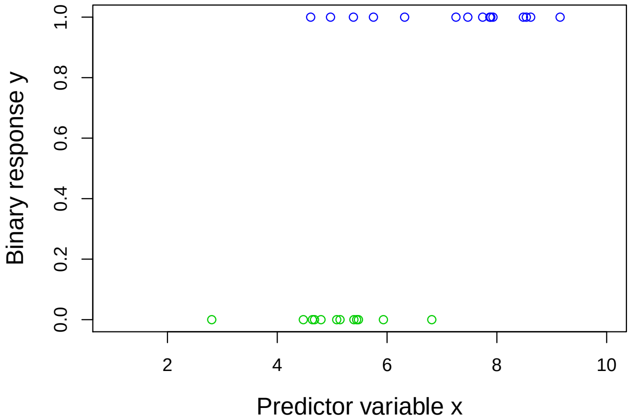
Logistic regression model (response y is binary 0/1): minimize the sum of deviances

$$\sum_{i=1}^{n} d_i(\beta) = \sum_{i=1}^{n} (-y_i \log \pi_i - (1 - y_i) \log(1 - \pi_i))$$

with conditional probability for i-th observation

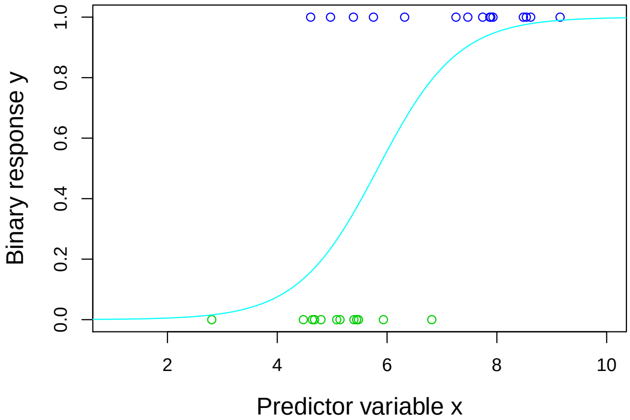
$$\pi_i = \Pr(y_i = 1|\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$$

Outliers in logistic regression



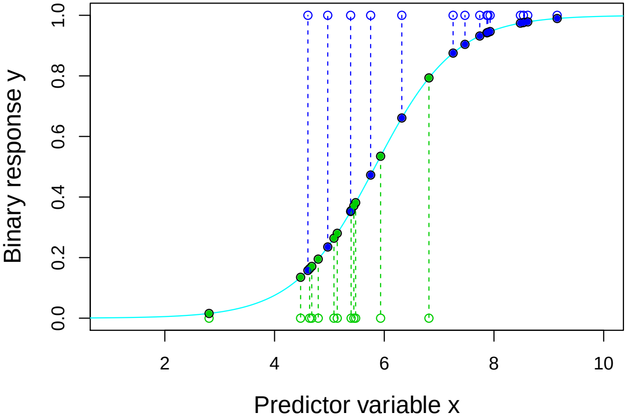
Response is coded with 0 (green group) and 1 (blue group).

Outliers in logistic regression



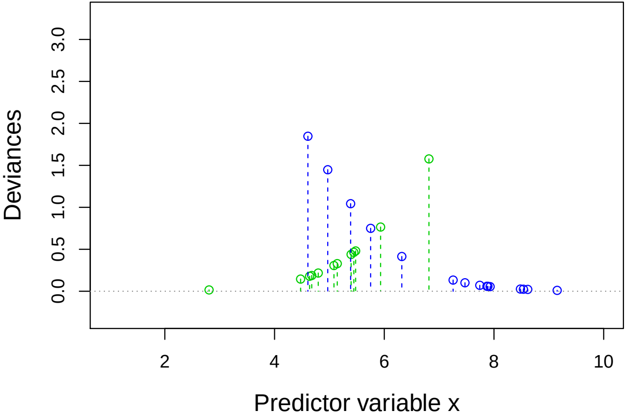
Logistic function (estimated cond. prob. $\hat{\pi}_i$) for the blue group.

Outliers in logistic regression



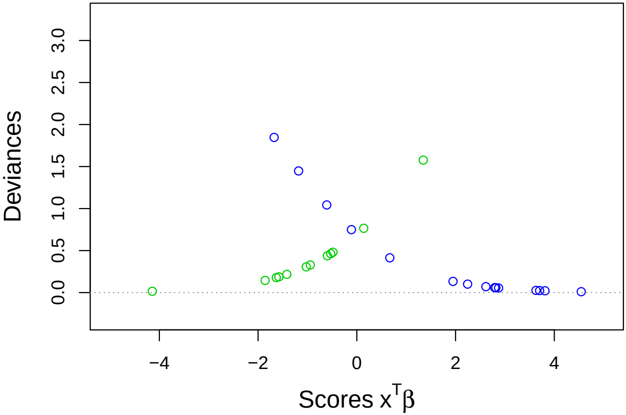
The logarithms of these “residuals” are the deviances.

Outliers in logistic regression



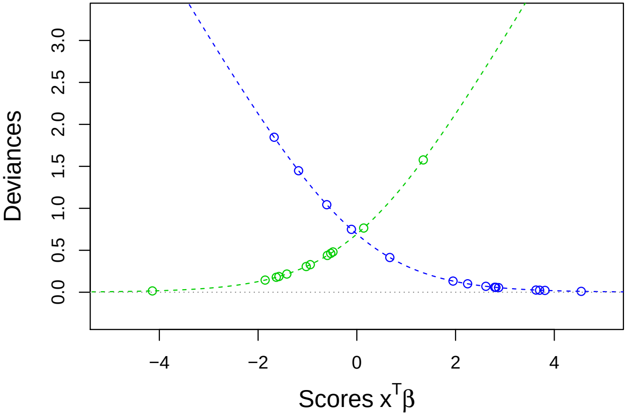
Deviances get larger for points on the wrong side.

Outliers in logistic regression



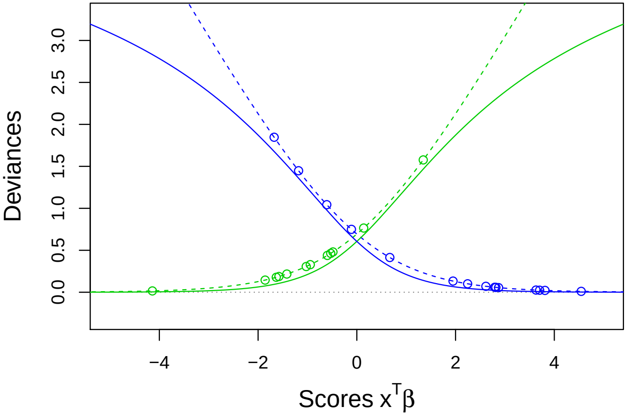
Note that the scores are always univariate!

Outliers in logistic regression



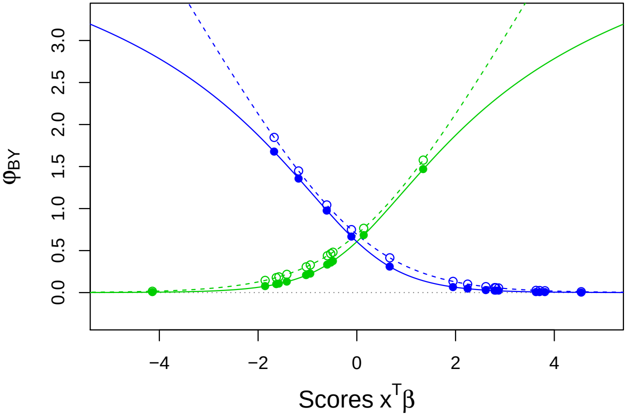
Deviances get larger for points on the wrong side.

Outliers in logistic regression



More robust by reducing the effect of large deviances.

Outliers in logistic regression



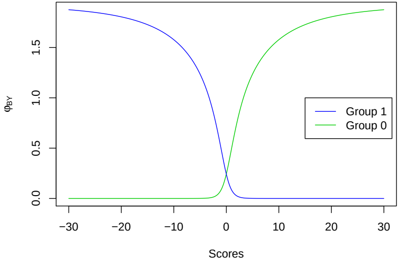
φ_{BY} is the Bianco-Yohai function for robust logistic regression.

Robust logistic regression

Croux and Haesbroeck (CSDA, 2003) replaced the deviance function by the function φ_{BY} to define the Bianco-Yohai (BY) estimator, a highly robust logistic regression estimator.

Function φ_{BY} for $y_i = 1$ (blue):

- Positive scores $\boldsymbol{x}_i^T \boldsymbol{\beta}$
  $\Rightarrow$ correct classification
  $\Rightarrow$ low values of φ_{BY}
- Negative scores $\boldsymbol{x}_i^T \boldsymbol{\beta}$
  $\Rightarrow$ wrong classification
  $\Rightarrow$ high values of φ_{BY}



- For incorrectly classified outliers: bounded influence.

Instead of minimizing $\sum_{i=1}^{n} d_i(\boldsymbol{\beta})$ they minimize $\sum_{i=1}^{n} \varphi_{BY}(\boldsymbol{\beta})$.

Elastic net regression

$$\hat{\beta}_{enet} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} r_i^2(\beta) + \lambda P_\alpha(\beta) \right\}$$

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1$$

- Elastic net penalty: combines L_1 and L_2 norm of β.
- Penalty with two tuning parameters: $\lambda \in [0, \lambda_{max}]$, $\alpha \in [0, 1]$.
- L_1 norm induces sparsity: excludes uninformative variables.
- L_2 norm favours similar coefficients for correlated variables.
- Penalized regression: feasible when $n \ll p$.

Sparse LTS regression

- Start with many (e.g. 500) random subsets of size 3 (*elemental subsets*).
- Compute the lasso fit (L_1 penalty) for each subset.
- Perform two C-steps for all subsets.
- Retain the best (e.g. 10) subsamples with lowest value of the objective function.
- For those, perform C-steps until convergence.
- Reweighting to increase statistical efficiency.

A. Alfons, C. Croux, S. Gelper, **Sparse least trimmed squares regression for analyzing high-dimensional large data sets**, *The Annals of Applied Statistics*, 7(1) (2013) 226–248.

Robust elastic net regression

Find an index set H with $|H| = h$ that minimizes

- for linear regression,

$$Q(H, \beta) = \sum_{i \in H} r_i^2(\beta) + h\lambda P_\alpha(\beta)$$

- for logistic regression,

$$Q(H, \beta) = \sum_{i \in H} d_i(\beta) + h\lambda P_\alpha(\beta)$$

Restrictions for logistic regression:

- Elemental subsets have size 4 (2 from each group)
- H includes the same proportion of observations from both groups as the full data set.

Algorithm for logistic regression

1. Start with 500 elemental subsets of size 4: $\tilde{H}_1, \ldots, \tilde{H}_{500}$.
2. Estimate classical logistic models with elastic net penalty for each subset.
3. Take from each model the h observations with smallest deviances (proportional from both groups) to form the updated subsets $H_1, \ldots, H_{500}$.
4. Estimate a new model for each subset $H_1, \ldots, H_{500}$.
5. Repeat 3-4.
6. Take the 10 subsets with maximum value of

$$Q^*(H, \boldsymbol{\beta}) = \sum_{i \in H} \varphi_{BY}(\boldsymbol{x}_i^T \boldsymbol{\beta}; y_i)$$

7. For those subsets repeat 3-4 till convergence.
8. Choose the final subset H with maximum value of $Q^*(H, \boldsymbol{\beta})$.

Tuning parameter selection

2 tuning parameters → We need a good strategy or a lot of time!

- For each combination of α and λ calculate the best subset H.
- Perform 5-fold cross-validation on each best subset.
- Repeat the 5-fold cross-validation for more stable results.
- Select that couple of tuning parameters which minimizes

$$\mathrm{tSUMd}(\alpha, \lambda) = \frac{1}{h} \sum_{i=1}^{h} d_i(\hat{\beta}_{\alpha,\lambda}),$$

where d_i are test set deviances from the models estimated on the training data for a specific α and λ.

Reweighting step

Goal: improve the efficiency of the estimator

- Linear regression: same procedure as in Alfons et al. (2013)
- Logistic regression: compute Pearson residuals

$$r_i^s = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi} i (1 - \hat{\pi}_i)}},$$

which are approximately distributed as $N(0,1)$.

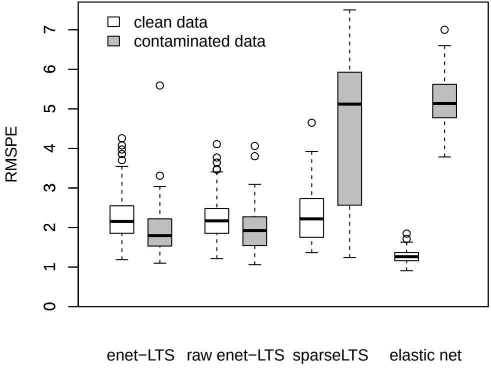
Define weights as

$$w_i = \begin{cases} 1, & \text{if } |r_i^s| \leq \Phi^{-1}(1-\delta) \\ 0, & \text{if } |r_i^s| > \Phi^{-1}(1-\delta) \end{cases} \quad i = 1, 2, \ldots, n,$$

where $\delta = 0.0125$ (gives 2.5% outliers in the normal model). Then:

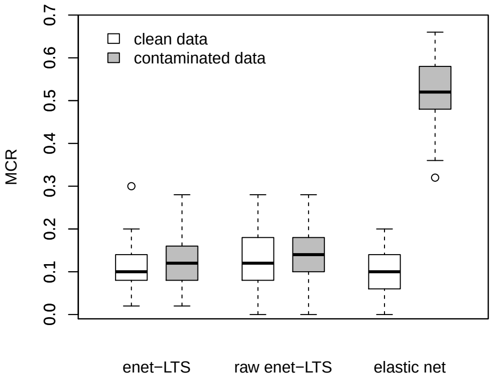
$$\hat{\beta}_{reweighted} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} w_i d_i(\beta) + \lambda_{upd} \left(\sum_{i=1}^{n} w_i \right) P_{\alpha_{opt}}(\beta) \right\},$$

Simulation results: linear regression



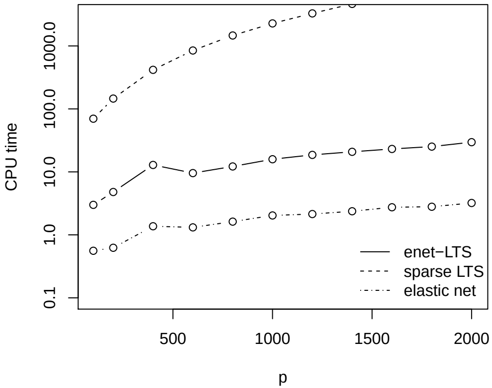
Root mean squared prediction error: $n = 50$ and $p = 100$; 100 rep.

Simulation results: logistic regression



Misclassification rate: $n = 50$ and $p = 100$; 100 repetitions

Time comparison



averaged over 5 replications, for fixed $n = 150$

Mass spectra from meteorites: Renazzo and Ochansk

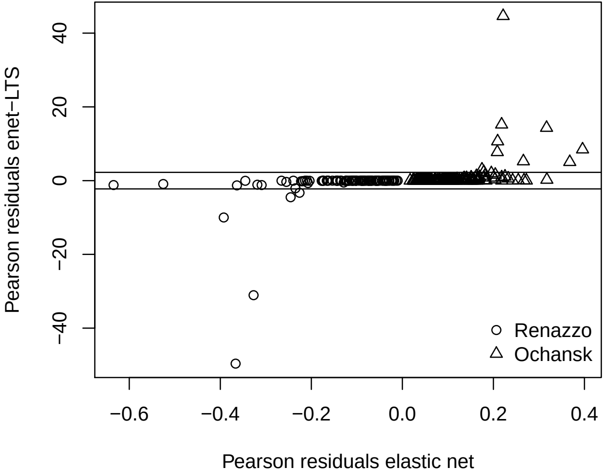
- Renazzo ($n_1 = 110$)
- Ochansk ($n_0 = 160$)

	# variables	tSUMd
elastic net	136	0.00866
enet-LTS raw	294	0.00030
enet-LTS	397	0.00014

Table: Number of variables (out of 1540) in the optimal models, and trimmed mean deviance from leave-one-out cross-validation of the optimal models.

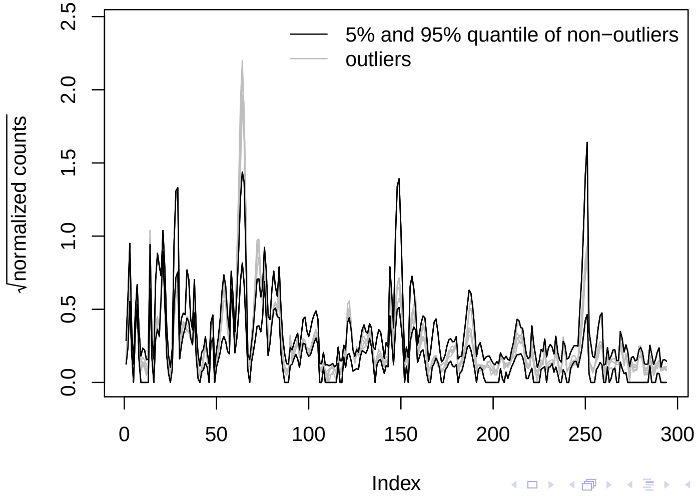
Mass spectra from meteorites: Renazzo and Ochansk

Pearson residuals with standard normal quantiles ± 2.5.



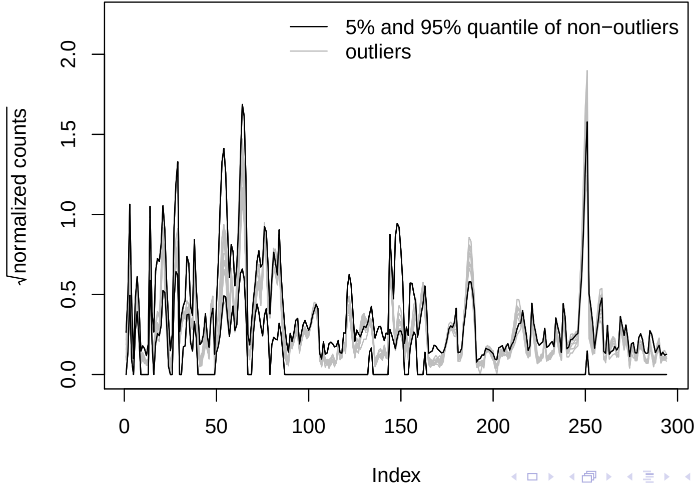
Mass spectra from meteorites: Renazzo and Ochansk

Outliers in the spectra of Renazzo samples:



Mass spectra from meteorites: Renazzo and Ochansk

Outliers in the spectra of Ochansk samples:



Glass vessels

Two groups are selected for demonstration:

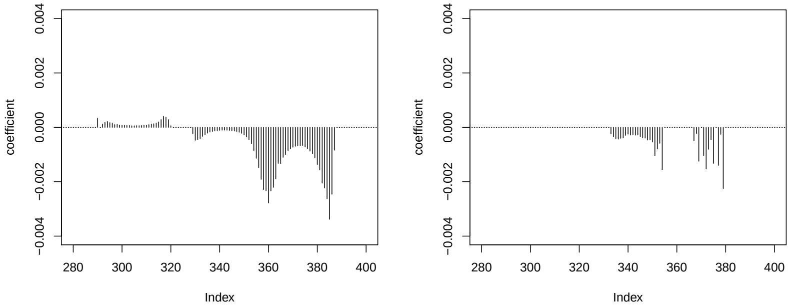
- potassic group ($n_1 = 15$)
- potasso-calcic group ($n_0 = 10$)

	# variables	tSUMd
elastic net	50	0.004290
enet-LTS raw	375	0.000345
enet-LTS	448	0.000338

Table: Number of variables (out of 1905) in the optimal models, and trimmed deviance from leave-one-out cross validation of the optimal models.

Glass vessels

Enet penalty favors similar coefficients for correlated variables.
Left: enet-LTS; Right: elastic net



Left: pos.: assoc. with potassium, neg.: assoc. with calcium

Summary

- robust procedure for linear and logistic regression using the elastic net penalty
- suitable for regression with many covariates
- robustness through trimming; reweighting step to increase efficiency
- cross-validation for selecting the tuning parameters
- R package enetLTS, freely available on https://cran.r-project.org/

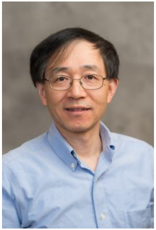
DSSV 2018

Data Science, Statistics and Visualisation

July 9-11, 2018, TU Wien, Austria

http://iasc-isi.org/dssv2018/

Keynote speakers:



Xuming He
Univ. of Michigan

Helwig Hauser
Univ. of Bergen