Texte

# Robust and Sparse Estimators for Linear Regression Models

Ezequiel Smucler and Víctor J. Yohai

Instituto de Cálculo - Universidad de Buenos Aires, CONICET

November 11, 2015

1. Linear Regression
   - Least Squares
   - Robust estimators
   - Sparsity and high-dimension

2. Existent methods for sparse models
   - What about robustness?

3. Our proposal
   - Computation
   - A real data set
   - Simulations

## Section 1

# Linear Regression

# Linear Regression

- A problem common to all branches of science and technology is to explain a variable $y$ as a function of other variables $\mathbf{x} = (x_1, \ldots, x_p)$.
- The simplest way to model the relation between $y$, the response, and $\mathbf{x}$, the covariates, is via a linear model.

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + u_i \text{ for } i = 1, ..., n.$$

- $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is to be estimated and $u_i$ is a random error term.

Subsection 1

# Least Squares

# Least Squares

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2.$$

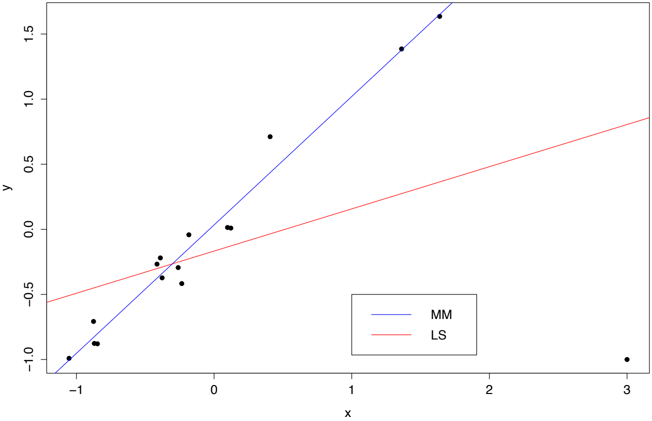
# Least Squares

- It is well known that $\hat{\beta}_{LS}$ is optimal when the errors are normal.
- However, it is also well known that it is not robust: it is very sensitive to the presence of outlying (atypical) observations.
- Informally, we say that an estimator is robust if it is not much affected by a small fraction of outlying observations in the data.

# Robust estimators

- A measure of an estimators robustness is the finite sample breakdown point, introduced by Donoho and Huber (1982).
- Loosely speaking, the finite sample breakdown point of an estimator is the minimum fraction of outliers that may take the estimator beyond any limit.
- $\hat{\beta}_{LS}$ has a breakdown point equal to $1/n$.

# Least Squares



Subsection 2

## Robust estimators

# Robust estimators of regression

$$\hat{\beta}_{LS} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2.$$

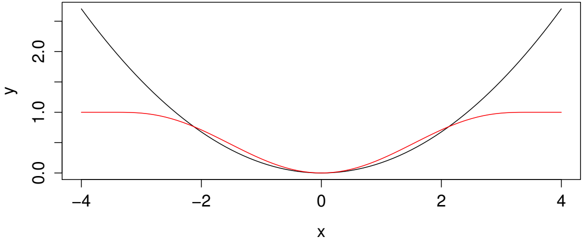
$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \beta).$$

# Robust estimators of regression

- Many robust regression estimators are based on bounded loss functions.
- We will say that $\rho$ is a $\rho$-function if it is a bounded, continuous, symmetric and non-decreasing loss function.

# Robust estimators of regression

Figure : In red: Tukey's $\rho$ function, in black: $x^2$ .



# Robust estimators of regression

- We need a preliminary estimation of the scale of the errors.
- Observations with residuals that are large when divided by robust estimate of scale of the errors will be considered outliers.

# Robust estimators of regression

- Given a $\rho$-function $\rho_1$, Yohai (1987) defines the MM-estimator of regression as

$$\hat{\beta}_{MM} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_1\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{s_n}\right),$$

where $s_n$ is a robust estimate of scale of the errors.

# MM-estimators of regression

- He shows that the loss function $\rho_1$ can be chosen so that the resulting MM-estimator has simultaneously the two following properties:
- High breakdown point.
- Arbitrarily high efficiency at the normal distribution.

Subsection 3

## Sparsity and high-dimension

# Sparsity and high-dimension

- Suppose we have a large number (compared to the sample size) of candidate covariates, but we believe that a small number of them are actually useful to predict the response $y$: most of the coefficients of $\beta_0$ are either zero or very small.
- We do not know in advance the set of indices corresponding to predictor variables that are relevant, and it may be of interest to estimate it.
- These type of scenarios have become increasingly common in areas such as bioinformatics and chemometrics, among others.

## Goals

- We aim at constructing estimating procedures that simultaneously

i) Have a high prediction accuracy.

ii) Do variable selection. We would like our estimator to set most of the coefficients corresponding to the non-relevant covariates variables to zero.

Section 2

# Existent methods for sparse models

# Existent methods for sparse models

- Standard regression estimators such as LS can have very poor prediction properties when $p/n$ is close to 1 and they are not defined for $p > n$. Moreover, they do not produce sparse models. The same is true of robust regression estimators, such as MM-estimators.

# Existent methods for sparse models

- A popular approach to sparse estimation in linear models is to use penalized estimators.
- $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_{\lambda_n}(\beta_j)$.
- $p_{\lambda_n}()$ is a penalization function that depends on some penalization parameter $\lambda_n$.
- We penalize the model's complexity as measured by $\sum_{j=1}^{p} p_{\lambda_n}(\beta_j)$.
- Penalized estimators can be calculated for $p > n$.

# Existent methods for sparse models

- $p_{\lambda_n}(\beta) = \lambda_n|\beta|$, gives the LS-Lasso of Tibshirani (1996). Note that in this case,

  $$\sum_{j=1}^{p} p_{\lambda_n}(\beta_j) = \lambda_n \sum_{j=1}^{p} |\beta_j| = \lambda_n \|\boldsymbol{\beta}\|_1,$$

  so that the penalization term is proportional to the $\ell_1$ norm of the coefficients.

# Existent methods for sparse models

- The optimization program that defines the LS-Lasso estimator,

$$\hat{\boldsymbol{\beta}}_{LASSO} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|,$$

  is convex.
- There exist very efficient algorithms to solve it, for example the LARS algorithm by Efron et al. (2004), or Coordinate Descent Optimization.
- For other penalization or loss functions, the corresponding optimization program may no longer be convex and optimization becomes harder.

# Asymptotic properties of penalized estimators

- We will say that an estimator has the oracle property if, simultaneously
  - The estimated coefficients corresponding to zero coefficients of the true regression parameter are set to zero with probability tending to one.
  - The coefficients corresponding to non-zero coefficients of the true regression parameter are estimated with the same accuracy we would have if an Oracle told us which were the relevant covariates and we had applied the non-penalized procedure to the relevant covariates only.

# Asymptotic properties of penalized estimators

- The LS-Lasso estimator does not have the oracle property.
- Moreover, the LS-Lasso estimator has a bias problem: it can excessively shrink large coefficients.
- To remedy these issues, Zou (2006) introduced the adaptive LS-Lasso.
- Let $\hat{\boldsymbol{\beta}}_{ini}$ is an initial estimate (such as the Lasso).
- $$\hat{\boldsymbol{\beta}}_{ADA} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{ini,j}|}.$$
- The adaptive LS-Lasso has oracle property and it can be calculated using any algorithm that solves the Lasso problem.

## Subsection 1

## What about robustness?

# What about robustness?

- All of the aforementioned penalized estimators are based on the quadratic loss function and hence are not expected to be robust.
- Alfons et al. (2013) showed that the breakdown point of the LS-Lasso estimators is $1/n$.
- In this talk, we propose robust versions of LS-Lasso and adaptive LS-Lasso estimators that are based on MM-estimators.

Section 3

# Our proposal

# Lasso type robust estimators of regression

- Given a $\rho$-function $\rho_1$, $\lambda_n > 0$, and a robust scale estimate $s_n$ we define the MM-Lasso estimator of regression as

$$\hat{\boldsymbol{\beta}}_B = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_1 \left( \frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{s_n} \right) + \lambda_n \sum_{j=1}^{p} |\beta_j|.$$

# Lasso type robust estimators of regression

- Let $\hat{\boldsymbol{\beta}}_{ini}$ be a robust initial estimate of $\boldsymbol{\beta}_0$, such as the MM-Lasso.
- We define the adaptive MM-Lasso estimator of regression as

$$\hat{\boldsymbol{\beta}}_A = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_1 \left( \frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{s_n} \right) + \lambda_n \sum_{i=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{ini,j}|},$$

# Lasso type robust estimators of regression

- We proved that for any positive $\lambda_n$ both the MM-Lasso and the adaptive MM-Lasso have a high breakdown point.

# The oracle property for the adaptive MM-Lasso

**Theorem**

*Under regularity assumptions, if $p < n$, the adaptive MM-Lasso has the oracle property.*

Subsection 1

## Computation

# Computation of MM-Lasso estimators

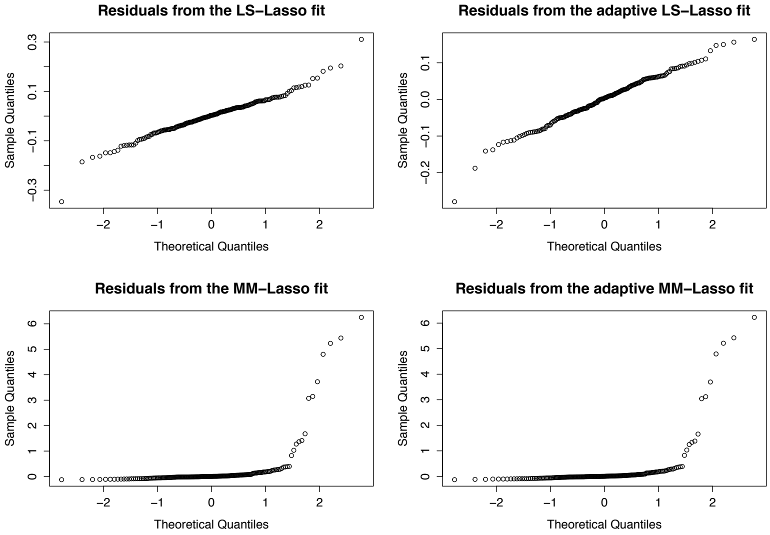
- Since the optimization program used to define the MM-Lasso is not convex, calculating the estimator is much more time consuming than calculating the ordinary LS-Lasso.
- We calculate the MM-Lasso by iteratively solving a standard weighted LS-Lasso problem. The same procedure allows us to calculate adaptive MM-Lasso estimators.
- We use Tukey's loss function and the penalization parameters are chosen over a set of candidates via 5-fold cross validation.
- An R package that includes the functions to calculate the MM-Lasso and adaptive MM-Lasso estimators is available at `http://esmucler.github.io/mmlasso/`.

Subsection 2

## A real data set

# Electron-probe X-ray microanalysis of glass vessels

- We consider a data set corresponding to electron-probe X-ray microanalysis of glass vessels, where each of the $n = 180$ glass vessels is represented by a spectrum on $p = 486$ frequencies. For each vessel the contents of the chemical compound PbO are registered.
- We fit a linear model where the response variable is the content of PbO and the covariates are the frequencies measured on each glass vessel.
- It is generally believed that only a small number of frequencies are needed to accurately predict the contents of the chemical compound.



Residuals from the LS–Lasso fit

Residuals from the adaptive LS–Lasso fit

Residuals from the MM–Lasso fit

Residuals from the adaptive MM–Lasso fit

# Electron-probe X-ray microanalysis of glass vessels

- The MM-Lasso selects seven variables.
- The adaptive MM-Lasso selects four variables.
- The LS-Lasso selects 70 variables, the adaptive LS-Lasso selects 49.
- To asses the prediction accuracy of the estimators, we used 5-fold cross-validation. The criterion used was a robust and efficient scale of the residuals, called a $\tau$-scale, introduced in Yohai and Zamar (1988).

| | $\tau$-scale |
|---|---|
| MM-Lasso | 0.086 |
| adaptive MM-Lasso | 0.083 |
| LS-Lasso | 0.131 |
| adaptive LS-Lasso | 0.138 |

Table : Cross-validated $\tau$-scale of the residuals of each of the estimators for the electron-probe X-ray microanalysis data.

## Subsection 3

## Simulations

# Competing methods

- We compare the performance with regards to prediction accuracy and variable selection properties of
  1. The MM-Lasso estimator described in the previous subsection.
  2. The adaptive MM-Lasso estimator described in the previous subsection.
  3. The LS-Lasso estimator.
  4. The adaptive LS-Lasso estimator.
  5. The Least Squares Oracle estimator, that is, the LS estimator applied only to the relevant covariates.
  6. The Oracle MM estimator: an MM-estimator applied to the relevant covariates only.

# Simulation scenarios

- To evaluate the estimators we generate two independent samples of size $n$ of the model $y = \mathbf{x}^T \boldsymbol{\beta}_0 + u$, with $u \sim N(0, 1.5^2)$.
- We take $p = 30$, $n = 100$ and $\boldsymbol{\beta}_0$ given by: components 1-5 are 2.5, components 6-10 are 1.5, components 11-15 are 0.5 and the rest are zero.
- We take $\mathbf{x} \backsim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.95$.
- The first sample is used to fit the estimates and the second sample is used to evaluate the prediction accuracy of the estimates using the root mean squared prediction error (RMSE).
- We also evaluate the variable selection performance of the estimators by calculating the false negative ratio (FNR) and the false positive ratio (FPR).

# Simulation scenarios

1. To evaluate the robustness of the estimators we contaminate the samples used to fit the estimators as follows. We take $m = [0.1n]$ and for $i = 1, .., m$ we set $y_i = 5y_0$ and $\mathbf{x}_i = (5, ..., 0)$. We moved $y_0$ in a grid between 0 and 10.
2. To summarize the results for the contaminated scenarios we report for each estimator the maximum RMSE, FNR and FPR over all outlier sizes $y_0$.
3. The number of Montecarlo replications for the uncontaminated scenario was $M = 500$. The number of Montecarlo replications for the contaminated scenario was $M = 100$.

| | RMSE | FNR | FPR |
|---|---|---|---|
| MM-Lasso | 1.69 | 0.13 | 0.21 |
| adaptive MM-Lasso | 1.77 | 0.26 | 0.09 |
| LS-Lasso | 1.74 | 0.11 | 0.22 |
| adaptive LS-Lasso | 1.74 | 0.21 | 0.13 |
| Oracle | 1.63 | 0 | 0 |

Table : Results for the simulated scenario, with normal distributed errors. RMSE, FNR and FPR, averaged over 500 replications are reported for each estimator.

| | Max. RMSE | Max. FNR | Max. FPR |
|---|---|---|---|
| MM-Lasso | 2.02 | 0.20 | 0.35 |
| adaptive MM-Lasso | 2.11 | 0.36 | 0.21 |
| LS-Lasso | 3.05 | 0.25 | 0.26 |
| adaptive LS-Lasso | 3.24 | 0.41 | 0.15 |
| Oracle MM | 2.09 | 0 | 0 |

Table : Results with normal errors and 10% contaminated observations. Maximum RMSEs, FNRs and FPRs over all outlier sizes are averaged over 100 replications.

# Conclusions

- We proposed robust versions of the LS-Lasso and the adaptive LS-Lasso that are based on MM-estimators.
- We applied the estimators to a real data set and ran a small simulation study that we hope showed that the proposed estimators are a valuable addition to the statisticians toolbox.
- A lot more details and full bibliographical references can be found in: `http://arxiv.org/abs/1508.01967`.

# Acknowledgements

Thank you