# Robust and Sparse Estimators for Linear Regression Models

Ezequiel Smucler and Víctor J. Yohai

Instituto de Cálculo - Universidad de Buenos Aires, CONICET

November 19, 2015

Section 1

# Linear Regression

# Linear Regression

- Consider $n$ independent observations from a linear regression model:

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

- $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is to be estimated and $u_i$ is a random error term. We considering the possibility of $p$ growing with the sample size, i.e. $p = p_n$.

# Robust estimators of regression

- A popular family of robust estimators for the linear regression model is that of MM-estimators.
- We will say that $\rho$ is a $\rho$-function if it is a bounded, continuous, symmetric and non-decreasing loss function.
- Given a $\rho$-function $\rho_1$, Yohai (1987) defines the MM-estimator of regression as

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

where $s_n$ is an M-estimate scale of the residuals of a robust initial estimator of regression, for example, the scale provided by an S-estimator of regression (Rousseeuw and Yohai (1984)).

# MM-estimators of regression

- It can be shown (Yohai (1987), Salibian-Barrera (2006)) that under regularity assumptions

$$\sqrt{n}(\hat{\beta}_{MM} - \beta_0) \rightarrow^d N_p(\mathbf{0}, s_0^2 c(\psi_1) \mathbf{V}_{\mathbf{x}}^{-1}),$$

where $\mathbf{V}_{\mathbf{x}} = \lim 1/n \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$, $s_0$ is the limit in probability of $s_n$ and $c(\psi)$ depends on the loss function and the distribution of the errors.

# MM-estimators of regression

- Yohai (1987) shows that the loss function $\rho_1$ and the scale estimate $s_n$ 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 1

# Sparsity and high-dimension

# Sparsity and high-dimension

- Suppose we have a large number of covariates, but we believe that a small number of them are actually useful to predict the response $y$: most of the coordinates of $\beta_0$ are either zero or very small.
- We do not know in advance the set of indices corresponding to covariates that are relevant, i.e. have non-zeros coefficients in $\beta_0$, 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.

# Sparsity and high-dimension

- For simplicity, we will write $\beta_0 = (\beta_{0,I}, \beta_{0,II})$, where $\beta_{0,I} \in \mathbb{R}^s$, $\beta_{0,II} \in \mathbb{R}^{p-s}$, all the coordinates of $\beta_{0,I} \in \mathbb{R}^s$ are non-zero and all the coordinates of $\beta_{0,II} \in \mathbb{R}^{p-s}$ are zero. Note that $s$ may also depend on $n$.
- Let $\mathbf{x}_I$ be the first $s$ coordinates of $\mathbf{x}$ , i.e. the relevant covariates.

# 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 $p - s$ 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.
- $\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \sum_{j=1}^{p} p_{\lambda_n}(\beta_j)$.
- $p_{\lambda_n}()$ is a penalization function that depends on some penalization parameter $\lambda_n$, that may be chosen via a data-drive procedure such as cross validation.
- 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

- $p_{\lambda_n}(\beta) = \lambda_n |\beta|^q$, with $q > 0$, gives the LS-Bridge estimators of Frank and Friedman (1993). Note that in this case,

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

  so that the penalization term is proportional to the $q$-th power of the $\ell_q$ "norm" of the coefficients.
- These include as very important special cases the LS-Ridge estimator, $q = 2$, of Hoerl and Kennard (1970) and the LS-Lasso estimator, $q = 1$, of Tibshirani (1996).

# Existent methods for sparse models

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

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

  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

- An interesting theoretical property of penalized estimators is the so called oracle property, introduced for the fixed dimensional case by Fan and Li (2001) and extended to the case of a diverging number of parameters by Fan and Peng (2004).

# Asymptotic properties of penalized estimators

- Informally, 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 penalization methods

- 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 $\hat{\beta}_{ini}$ be an initial estimate (such as the LS-Lasso).
- $$\hat{\beta}_{ADA} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{ini,j}|}.$$
- The adaptive LS-Lasso has oracle property (Zou (2006) and Huang et al. (2008)) 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 robust.
- There have been several proposals of penalized estimators which are robust to outliers in the response variable, for example, LAD-Lasso of Wang (2007) or the general penalized M-estimators of Li et. al. (2011). However, since these estimators are defined using convex loss functions, they are not robust to outliers in the covariates.
- In Alfons et al. (2013), the authors introduced the Sparse-LTS, a least trimmed squares estimator (Rousseeuw (1984)) with an $\ell_1$ penalty term.
- In this talk, we introduce Bridge and adaptive Bridge versions of MM-estimators.

Section 3

# Our proposal

# Bridge type robust estimators of regression

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

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

- When $q = 1$ we will call the resulting estimator MM-Lasso.
- When $q = 2$ the resulting estimator is the MM-Ridge introduced in Maronna (2012).
- In practice, we take the scale estimate $s_n$ to be the scale of the residuals provided by an S-Ridge estimator, i.e. an S-estimator with an $\ell_2$ penalty, introduced in Maronna (2012).

# Bridge type robust estimators of regression

- Let $\hat{\beta}_{ini}$ be a robust initial estimate of $\beta_0$, such as the MM-Lasso.
- Given a penalization parameter $\iota_n > 0$ and $t > 0$ we define the adaptive MM-Bridge estimator of regression as

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

When $t = 1$ we will call the resulting estimator adaptive MM-Lasso. In practice we will take $t = 1$ and $\hat{\beta}_{ini}$ to be an MM-Lasso estimator.

# Bridge type robust estimators of regression

- One can easily show that for any fixed penalization parameters $\lambda_n$ and $\iota_n$, both the MM-Bridge and the adaptive MM-Bridge have a breakdown point equal to $1 - 1/n$.
- In practice, the penalization parameters $\lambda_n$ and $\iota_n$ may be chosen via some data-driven procedure such as cross-validation. The robustness of the resulting estimators will depend sorely on the robustness of the cross-validation scheme, and hence the use of robust residual scales as objective functions, instead of the classical root mean squared error, is crucial.

Subsection 1

# Some asymptotic results

# A consistency result

## Theorem

*Under regularity assumptions, if* $(p_n \log p_n)/n \to 0$, *there exists sequences* $(\lambda_n)_n$ *and* $(\iota_n)_n$ *such that the resulting MM-Bridge* $(\hat{\boldsymbol{\beta}}_B)$ *and adaptive MM-Bridge* $(\hat{\boldsymbol{\beta}}_A)$ *estimators satisfy*

$$\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$$

$$\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$$

# The oracle property for adaptive MM-Bridge

- Let $\hat{\boldsymbol{\beta}}_{A,I}$ and $\hat{\boldsymbol{\beta}}_{A,II}$ respectively stand for the first $s$ and the last $p - s$ coordinates of an adaptive MM-Bridge estimator $\hat{\boldsymbol{\beta}}_A$.

# The oracle property for adaptive MM-Bridge

## Theorem

*Under regularity assumptions, if* $t \leq 1$ *and* $p_n^2/n \to 0$, *there exists a sequence* $(\iota_n)_n$ *such that the resulting adaptive MM-Bridge has the oracle property:*

- $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{A,II} = \mathbf{0}_{p-s}\right) \to 1.$
- *For any sequence* $(\mathbf{a}_n)_n$ *such that* $\|\mathbf{a}_n\| \leq 1$ *for all* $n$,

$$\sqrt{n} r_n^{-1} \mathbf{a}_n^T \left(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I}\right) \xrightarrow{d} N\left(0, s_0^2 c(\psi_1)\right),$$

*where* $\boldsymbol{\Sigma}_{1,n} = 1/n \sum_{i=1}^{n} \mathbf{x}_{i,I} \mathbf{x}_{i,I}^T$ $r_n^2 = \mathbf{a}_n^t \boldsymbol{\Sigma}_{1,n}^{-1} \mathbf{a}_n.$

In particular, the adaptive MM-Lasso has the oracle property.
An entirely analogous result can be proved for MM-Bridge estimators with $q < 1$.

Subsection 2

# Computation

# Computation of MM-Lasso estimators

- We calculate the MM-Lasso by iteratively solving a standard weighted Lasso problem, using Maronna's S-Ridge as a starting point. The same procedure allows us to calculate adaptive MM-Lasso estimators.
- The penalization parameters are chosen over a set of candidates via 5-fold cross validation, using a $\tau$-scale of the residuals (Yohai and Zamar (1988)) as the objective function. We used Tukey's bisquare loss function as $\rho_1$.
- The estimators can be calculated for $p > n$.

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 Sparse-LTS of Alfons et. al (2013).
  4. The LS-Lasso estimator.
  5. The adaptive LS-Lasso estimator.
  6. The Least Squares Oracle estimator, that is, the LS estimator applied only to the relevant covariates.
  7. The Oracle MM estimator, that is, 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 $u \sim N(0, 3^2)$.
- 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).
- We take two possible combinations of $p$ and $n$: $(p, n) = (8, 40)$ and $(p, n) = (250, 50)$.
- We take $\boldsymbol{\beta}_0$ given by: component 1 is 3, component 2 is 1.5, component 6 is 2 and the rest are zero.
- We take $\mathbf{x} \backsim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.5$.

# 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 $y_0$.

$(p, n) = (8, 40)$

| | RMSE | FNR | FPR |
|---|---|---|---|
| MM-Lasso | 3.42 | 0.04 | 0.52 |
| adaptive MM-Lasso | 3.43 | 0.09 | 0.27 |
| Sparse-LTS | 3.92 | 0.03 | 0.82 |
| LS-Lasso | 3.33 | 0.02 | 0.43 |
| adaptive LS-Lasso | 3.28 | 0.06 | 0.26 |
| Oracle | 3.15 | 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.

$(p, n) = (8, 40)$

| | Max. RMSE | Max. FNR | Max. FPR |
|---|---|---|---|
| MM-Lasso | 4.38 | 0.11 | 0.57 |
| adaptive MM-Lasso | 4.43 | 0.25 | 0.32 |
| Sparse-LTS | 4.92 | 0.07 | 0.95 |
| LS-Lasso | 5.78 | 0.27 | 0.49 |
| adaptive LS-Lasso | 6.14 | 0.36 | 0.33 |
| Oracle MM | 3.71 | 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.

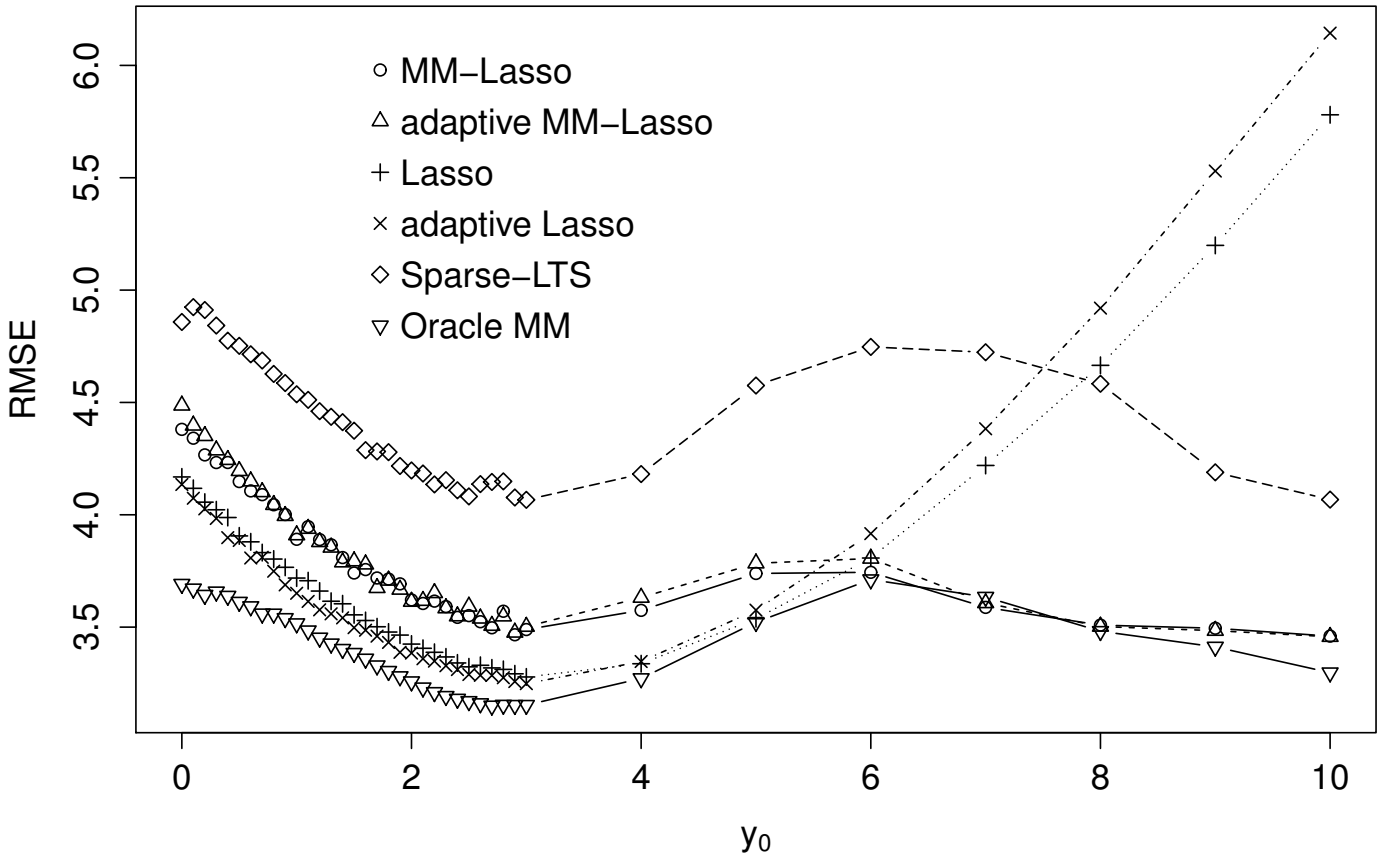



Figure : RMSEs as a function of outlier sizes for each of the estimators, with $p = 8$, $n = 40$, normal errors and 10% contamination.

$(p, n) = (250, 50)$

| | RMSE | FNR | FPR |
|---|---|---|---|
| MM-Lasso | 4.05 | 0.12 | 0.07 |
| adaptive MM-Lasso | 3.99 | 0.18 | 0.03 |
| Sparse-LTS | 4.72 | 0.26 | 0.12 |
| LS-Lasso | 3.67 | 0.05 | 0.07 |
| adaptive LS-Lasso | 3.97 | 0.06 | 0.06 |
| Oracle | 3.13 | 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.

$(p, n) = (250, 50)$

| | Max. RMSE | Max. FNR | Max. FPR |
|---|---|---|---|
| MM-Lasso | 4.97 | 0.36 | 0.08 |
| adaptive MM-Lasso | 5.08 | 0.45 | 0.04 |
| Sparse-LTS | 5.40 | 0.47 | 0.11 |
| LS-Lasso | 6.04 | 0.42 | 0.07 |
| adaptive LS-Lasso | 7.89 | 0.45 | 0.06 |
| Oracle MM | 3.68 | 0 | 0 |

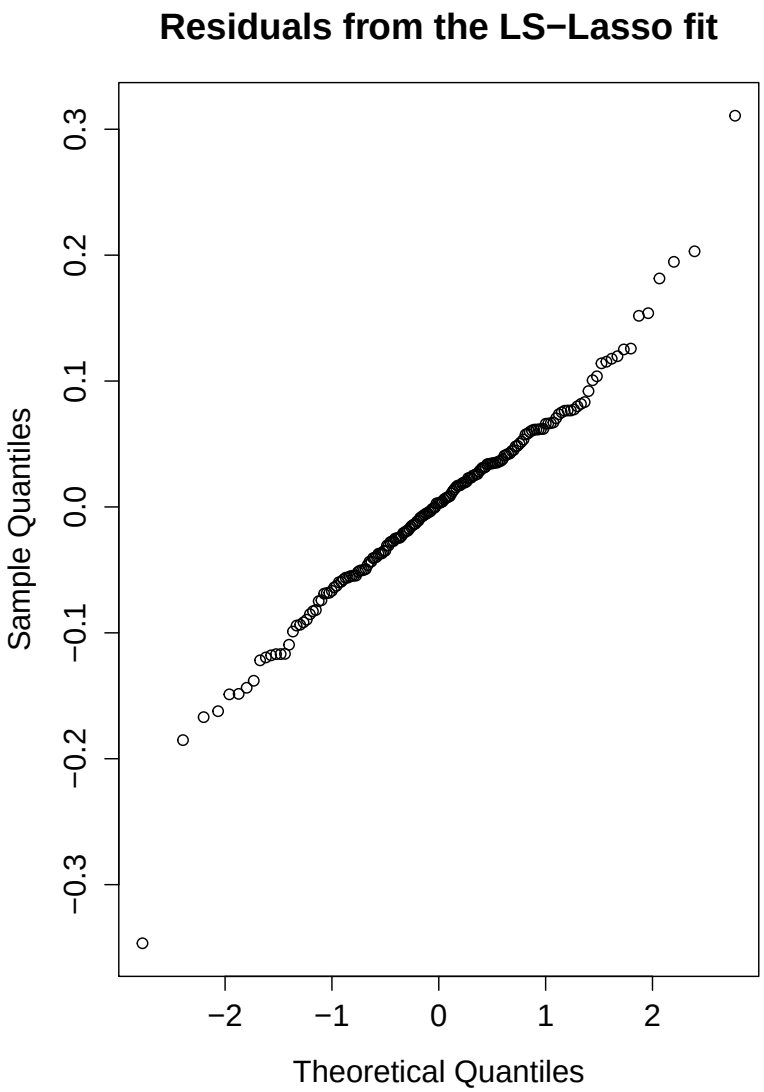
Table : Results with normal errors and 10% contaminated observations. Maximum RMSEs, FNRs and FPRs over all outlier sizes are averaged over 100 replications.

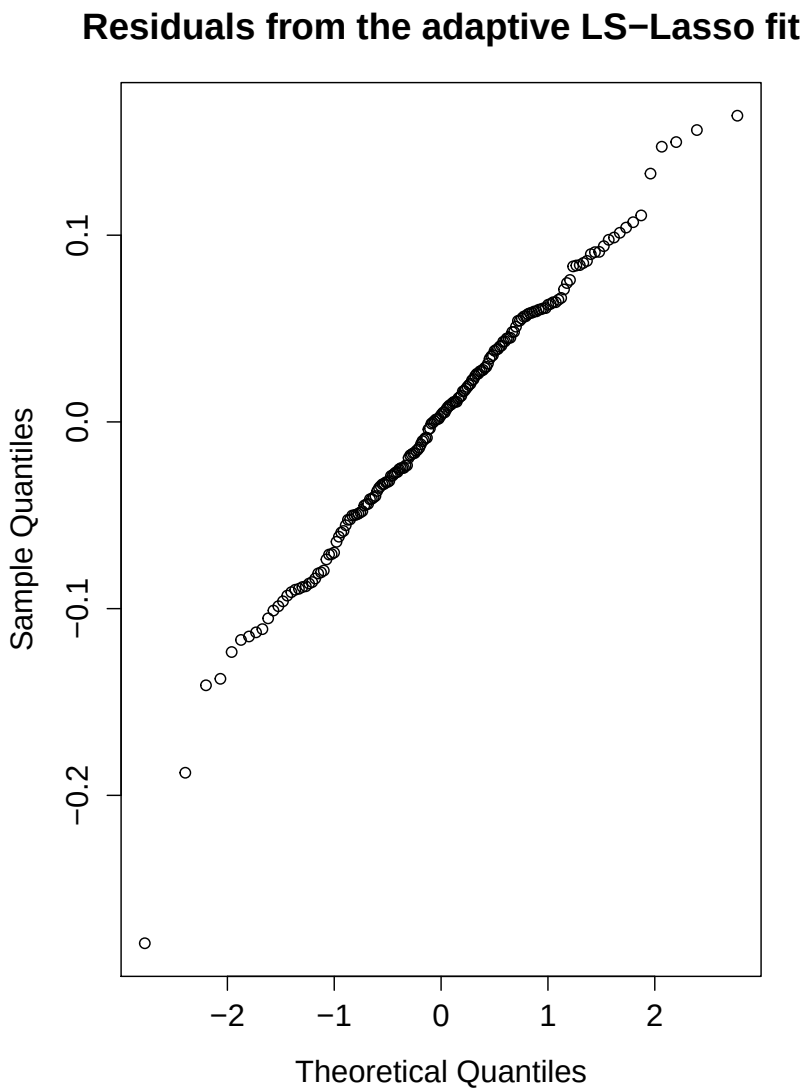
Subsection 4

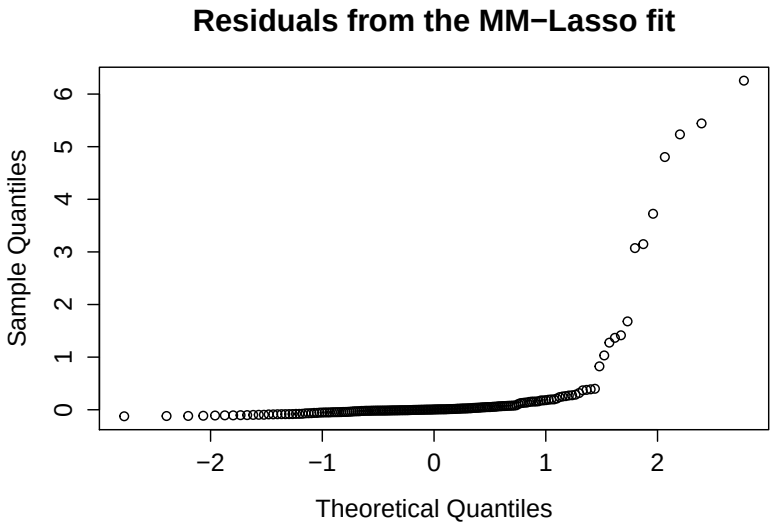
# A real data set

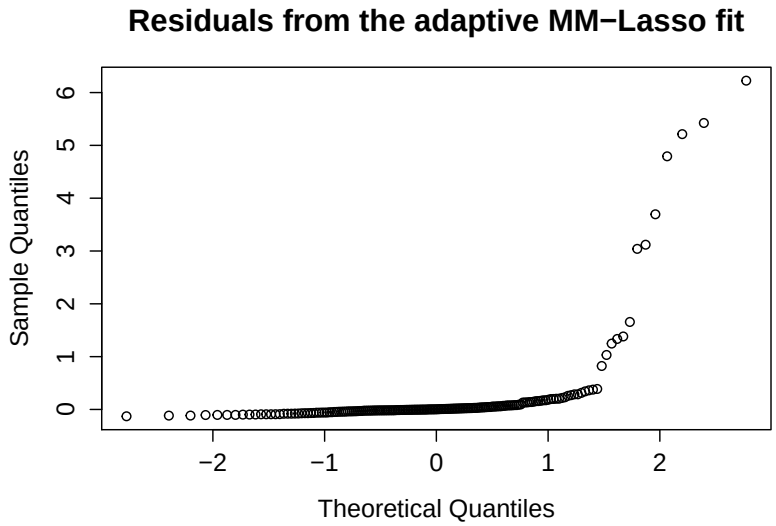
# Electron-probe X-ray microanalysis of glass vessels

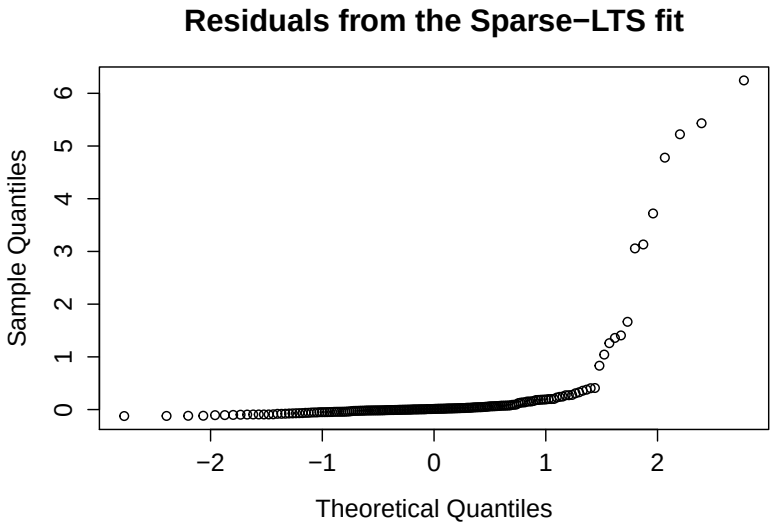
- We consider a data set corresponding to electron-probe X-ray microanalysis of glass vessels, where each of the $n = 180$ glass vessels is represented by a spectrum on $p = 486$ frequencies. For each vessel the contents of the chemical compound PbO are registered.
- We fit a linear model where the response 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 predict the contents of the chemical compound.


Residuals from the LS−Lasso fit
Sample Quantiles
Theoretical Quantiles
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
−2
−1
0
1
2


Residuals from the adaptive LS−Lasso fit
Sample Quantiles
Theoretical Quantiles
−0.2
−0.1
0.0
0.1
−2
−1
0
1
2


Residuals from the MM−Lasso fit
Sample Quantiles
Theoretical Quantiles
0
1
2
3
4
5
6
−2
−1
0
1
2


Residuals from the adaptive MM−Lasso fit
Sample Quantiles
Theoretical Quantiles
0
1
2
3
4
5
6
−2
−1
0
1
2


Residuals from the Sparse−LTS fit
Sample Quantiles
Theoretical Quantiles
0
1
2
3
4
5
6
−2
−1
0
1
2

# Electron-probe X-ray microanalysis of glass vessels

- The MM-Lasso selects seven variables.
- The adaptive MM-Lasso selects four variables.
- The Sparse-LTS selects three variables.
- The Lasso selects 70 variables, the adaptive Lasso selects 49.
- To asses the prediction accuracy of the estimators, we used 5-fold cross-validation. The criterion used was a $\tau$-scale of the residuals.

| | $\tau$-scale |
|---|---|
| MM-Lasso | 0.086 |
| adaptive MM-Lasso | 0.083 |
| Sparse-LTS | 0.329 |
| 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.

Section 4

# Conclusions

# Conclusions

- We proposed robust estimators for sparse linear models that are based on MM-estimators of regression.
- We analysed the asymptotic properties of the proposed estimators for the case of $p << n$.
- The merits of our proposed estimators were illustrated with a simulation study and the analysis of a real high-dimensional data set.

# Acknowledgements

Thank you