Challenging the Least Squares: A New Perspective on Robustness

Wenxin Zhou
UC San Diego

November, 2018
School of Mathematics and Statistics
University of Melbourne
Robustness and accuracy of methods for high dimensional data analysis based on Student's $t$-statistic

Aurore Delaigle, Peter Hall, Jiashun Jin

First published: 18 January 2011 | https://doi.org/10.1111/j.1467-9868.2010.00761.x

Challenging the empirical mean and empirical variance:
A deviation study

Olivier Catoni

CNRS – UMR 8553, Département de Mathématiques et Applications, Ecole Normale Supérieure, 45 rue d’Ulm, F75230 Paris cedex 05, and INRIA Paris-Rocquencourt – CLASSIC team. E-mail: olivier.catoni@ens.fr
Why Challenges are Needed?

**Motivating example:** Generate data \((Y_i, X_i)\) from a multi-response linear model

\[
Y_{ij} = \mu_j + \sum_{k=1}^{3} X_{ik} \beta_{jk} + \epsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, p,
\]

where

- \(n = 100, p = 5000\);
- \(\mu_j = \pm 1.5, j = 1, \ldots, 500 \) & \(\mu_j = 0, j = 501, \ldots, 5000\);
- \(X_{ik} \overset{iid}{\sim} \text{Unif}(-1,1), \beta_j = (2, -1,2)\top\);
- \(\epsilon_{ij} \overset{iid}{\sim} 0.25F_1 + 0.75F_2, F_1\text{—centered Lognormal}(0,1)\) and \(F_2\sim t_{1.5}\).
Histograms of Estimated Signals

OLS

Data–Adaptive Huber
Histograms of Estimated $\mu_j$'s

<table>
<thead>
<tr>
<th>Type</th>
<th>Alternative</th>
<th>Null</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data–Adaptive Huber</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Background and Motivation
Mean Estimation

Given $X_1, \ldots, X_n$, an iid sample with mean $\mu = \mathbb{E}X_1$.

Natural estimator: empirical/sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$  

If $\sigma^2 = \text{var}(X_1) < \infty$, by CLT,

$$\lim_{n \to \infty} \mathbb{P}\left(\sqrt{n} | \bar{X}_n - \mu | \geq \sigma \sqrt{2 \log \frac{2}{\delta}} \right) \leq \delta.$$  

Do non-asymptotic inequalities of a similar form hold?
Finite-Sample Performance

• If $X_1$ is sub-Gaussian $- \mathbb{E} e^{\lambda X_1} \leq e^{\lambda^2 \sigma^2 / 2}$, $\lambda \in \mathbb{R}$, then
  $$\mathbb{P} \left( \sqrt{n} | \bar{X}_n - \mu | \geq \sigma \sqrt{2 \log \frac{2}{\delta}} \right) \leq \delta, \ n \geq 1.$$  

• If $X_1$ only has finite variance, then
  $$\mathbb{P} \left( \sqrt{n} | \bar{X}_n - \mu | \geq \sigma \sqrt{\frac{1}{\delta}} \right) \leq \delta, \ n \geq 1.$$  

Exponentially weaker bound: hurts when many means are estimated simultaneously.
Lower Bound

Catoni (2012) proved that for each $\delta$, there exists a dist. with finite variance $\sigma^2$ such that

$$\mathbb{P}\left(\sqrt{n} | \bar{X}_n - \mu | \geq \sigma \sqrt{\frac{c}{\delta}} \right) \geq \delta,$$

where $c > 0$ is a constant.

There is no room of improvement for the sample mean if the sampling distribution is heavy-tailed.
Sub-Gaussian Estimator

Question:

Under what conditions do there exist sub-Gaussian mean estimators?

Let $\mathcal{P}$ be a class of prob. dist. on $\mathbb{R}$ with finite variance.

Does there exist a mean estimator $\hat{\mu}_n$ s.t. for all dist. in $\mathcal{P}$, it holds that, with probability $\geq 1 - \delta$,

$$|\hat{\mu}_n - \mu| \leq L\sigma \sqrt{\frac{\log 2/\delta}{n}}$$

for some constant $L = L(\mathcal{P})$?
Literature

• **Median-of-means**: Nemirovski & Yudin (83); Jerrum, Valiant & Vazirani (86); Alon, Matias & Szegedy (02); …; Lerasle & Oliveira (12); Minsker (15); Hsu & Sabato (16); Devroye et al. (16); Lugosi & Mendelson (17); etc.

• **M-estimation**: Huber (64, 73); …; Catoni (12, 16); Brownlees, July & Lugosi (15); Minsker (16); Fan, Li & Wang (17); Zhou et al. (18); etc.
Tuning Parameters

Median-of-means: number of subsamples/blocks

Median-of-means based methods can be numerically unstable and empirically unsatisfactory.

$M$-estimation: scale parameter in the loss/influence func.

Cross-validation can always be applied. However, if there are many parameters to be estimated, CV is computationally intensive.
Huber’s Method Revisited

Huber’s loss: $\ell_{\tau}(x) = \begin{cases} 
  \frac{x^2}{2} & \text{if } |x| \leq \tau \\
  \tau |x| - \frac{\tau^2}{2} & \text{if } |x| > \tau
\end{cases}$.
**Huber’s estimator:** For a given $\tau > 0$,

$$
\hat{\mu}_\tau = \min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \ell_\tau(X_i - \theta).
$$

**Norm:** $\tau = 1.345$ or $\tau = 1.345\sigma$.

**95% efficiency rule** (Huber, 81): The resulting estimator is asymptotically 95% as efficient as the OLS if the data are normal, and generally is more resistant to outliers.

**Bias:** For asymmetric distribution, a fixed $\tau$ brings bias.
Non-asymptotic Viewpoint

**Sub-Gaussian estimator** (Fan, Li & Wang, 17): For any $t > 0$ and $v \geq \sigma$, Huber’s estimator $\hat{\mu}_\tau$ with $\tau = v \sqrt{\frac{n}{t}}$ satisfies

$$|\hat{\mu}_\tau - \mu| \leq 4v \sqrt{\frac{t}{n}}$$

with probability at least $1 - 2e^{-t}$ as long as $n \geq 8t$.

From a non-asymptotic viewpoint, $\tau$ is chosen to balance bias and robustness.
Today’s Themes:

1. Data-driven calibration of $\tau$
2. Linear regression
3. High-dimensional sparse regression
4. Covariance estimation
Mean Estimation
Truncated Mean

Consider iid sample $X_1, \ldots, X_n$ with mean $\mu$ and variance $\sigma^2 > 0$. The truncated mean is

$$m_\tau = \frac{1}{n} \sum_{i=1}^{n} \min(|X_i|, \tau) \text{sign}(X_i) = \frac{1}{n} \sum_{i=1}^{n} \psi_\tau(X_i),$$

where $\psi_\tau = \ell'_\tau$ and $\tau > 0$ is a tuning parameter.

**Proposition:** For $t > 0$ and $\nu \geq (\mathbb{E}X^2)^{1/2}$, the truncated mean $m_\tau$ with $\tau = \nu(n/t)^{1/2}$ satisfies

$$\mathbb{P}\{ |m_\tau - \mu| \geq 2\nu(t/n)^{1/2} \} \leq 2e^{-t}.$$
The theoretically “optimal” $\tau$ should be of the form

$$\tau^* = (\mathbb{E}X^2)^{1/2} \sqrt{\frac{n}{t}}.$$

**Question:** How to choose $\tau^*$ from the data?

**Naive choice:**

$$\tilde{\tau} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2\right)^{1/2} \sqrt{\frac{n}{t}}.$$

**Leptokurtosis:** When $X_i$ are heavy-tailed, $X_i^2$ are extremely right-skewed so that the sample average tends to overestimate $\mathbb{E}X^2$. 
A Data-Driven Method

**Motivation**: A properly chosen $\tau$ produces truncated data

$$\psi_\tau(X_1), \ldots, \psi_\tau(X_n),$$

so that

$$\frac{1}{n} \sum_{i=1}^{n} \psi_\tau(X_i)$$

is a good estimator of $\mu$. Meanwhile, we expect that

$$\frac{1}{n} \sum_{i=1}^{n} \psi_\tau(X_i)^2$$

should also be a reasonable approximation of $\mathbb{E}X_1^2$. 

In practice, we solve

\[ \tau^2 = \frac{1}{n} \sum_{i=1}^{n} \psi_\tau(X_i)^2 \times \frac{n}{t}, \]

or equivalently,

\[ f(\tau) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \frac{\min(X_i^2, \tau^2)}{\tau^2} - \frac{t}{n} = 0. \]

Let \( \hat{\tau} \) be the solution to this equation. The data-driven mean estimator is

\[ \hat{m} = \frac{1}{n} \sum_{i=1}^{n} \min\{ |X_i|, \hat{\tau} \} \text{ sign}(X_i). \]
**Existence**: It can be shown that the equation

\[ f(\tau) = 0, \quad \tau > 0 \]

has a unique solution as long as \( t \leq \sum_{i=1}^{n} I(X_i \neq 0) \).

**Sub-optimality**: The truncated mean is not an ideal estimator. The upper bound depends on the second moment \( \mathbb{E}X_1^2 \) instead of variance \( \text{var}(X_1) \).
Recall Huber’s estimator

\[ \hat{\mu}_\tau = \arg \min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \ell_\tau(X_i - \theta). \]

**Deviation inequality** (Fan, Li & Wang, 17): For any \( t > 0 \), Huber’s estimator \( \hat{\mu}_\tau \) with

\[ \tau = \sigma \sqrt{\frac{n}{t}} \]

satisfies that, with probability at least \( 1 - 2e^{-t} \),

\[ |\hat{\mu}_\tau - \mu| \leq 4\sigma \sqrt{\frac{t}{n}}. \]
**Bahadur representation** (Zhou et al., 18): With above $\tau$, Huber’s estimator $\hat{\mu}_\tau$ satisfies

$$\hat{\mu}_\tau = \mu + \frac{1}{n} \sum_{i=1}^{n} \psi_\tau(X_i - \mu) + \text{Rem},$$

where the remainder term $\text{Rem}$ is of order $\sigma \frac{t}{n}$ with probability at least $1 - Ce^{-t}$.

This suggests that $\hat{\mu}_\tau$ behaves like

$$\mu + \frac{1}{n} \sum_{i=1}^{n} \psi_\tau(X_i - \mu).$$
A properly chosen $\tau$ should make

$$\frac{1}{n} \sum_{i=1}^{n} \psi_{\tau}^2(X_i - \mu)$$

a robust approximation of $\sigma^2$. Ideally, we solve

$$\tau^2 = \frac{1}{n} \sum_{i=1}^{n} \psi_{\tau}^2(X_i - \mu) \times \frac{n}{t}, \quad \tau > 0$$

or equivalently,

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\psi_{\tau}^2(X_i - \mu)}{\tau^2} = \frac{t}{n}, \quad \tau > 0.$$
Data-Adaptive Huber Estimator

We estimate $\mu$ and tune $\tau$ simultaneously by solving

$$
\begin{align*}
  f_1(\theta, z) & \overset{\text{def}}{=} \sum_{i=1}^{n} \min\{ |X_i - \theta|, z \} \operatorname{sign}(X_i - \theta) = 0 \\
  f_2(\theta, z) & \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \min\{ (X_i - \theta)^2, z^2 \} - \frac{t}{n} = 0
\end{align*}
$$

for $\theta \in \mathbb{R}$ and $z > 0$.

Given $z > 0$, $f_1(\cdot, z) = 0$ has a unique solution; given $\theta \in \mathbb{R}$, $f_2(\theta, \cdot)$ has a unique root if $t < \sum_{i} I(X_i \neq \theta)$. 
Remarks on $t$

1. The confidence level is $1 - O(e^{-t})$.

2. The deviation error scales as $\sigma \sqrt{\frac{t}{n}}$.

3. Let $t = t_n$ grow with $n$ to gain robustness, but not too fast to sacrifice bias.

4. We recommend $t = \log n$, a typical slowly growing function of $n$. 
Linear Regression
Let \((Y_1, X_1), \ldots, (Y_n, X_n)\) be iid data vectors that follow the regression model
\[
Y_i = \mu^* + X_i^\top \beta^* + \varepsilon_i, \quad i = 1, \ldots, n,
\]
where \(\varepsilon_i\) are regression errors satisfying \(\mathbb{E}(\varepsilon_i | X_i) = 0\).

Given \(\tau > 0\), Huber’s \(M\)-estimator is defined as
\[
\hat{\theta}_\tau = \arg\min_{\theta \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \ell_{\tau}(Y_i - Z_i^\top \theta),
\]
where \(Z_i = (1, X_i^\top)^\top\).
Asymmetry/Skewness Creates Bias

For any $\tau > 0$, $\hat{\theta}_\tau = (\hat{\mu}_\tau, \hat{\beta}_\tau^\top)^\top$ is a natural $M$-estimator of

$$\theta_\tau = (\mu_\tau, \beta_\tau^\top) \overset{\text{def}}{=} \text{argmin}_{\theta} \sum_{i=1}^{n} \mathbb{E} \mathcal{L}_\tau(Y_i - Z_i^\top \theta).$$

The true parameters are identified as

$$\theta^* = (\mu^*, \beta^*^\top) \overset{\text{def}}{=} \text{argmin}_{\theta} \sum_{i=1}^{n} \mathbb{E}(Y_i - Z_i^\top \theta)^2.$$

**Question:** $\theta_\tau = \theta^*$?
Proposition (fix design): For a fixed $\tau > 0$, assume that

- $\alpha \mapsto \mathbb{E} \ell_\tau(\varepsilon - \alpha)$ has a unique minimizer, denoted $\alpha_\tau$;
- $(Z_1, \ldots, Z_n)^T$ has full column rank.

Then it holds

$$
\mu_\tau = \mu^* + \alpha_\tau, \quad \beta_\tau = \beta^*.
$$

If the distribution of $\varepsilon$ is symmetric, $\alpha_\tau = 0$ for any $\tau > 0$. When $\varepsilon$ is asymmetric, the intercept estimation can be biased with the standard choice of $\tau$. As a result, the prediction would be biased and unreliable.
Adaptive Huber Regression

**Theorem** (Sun, Zhou & Fan, 17): Assume that

\[ \| S^{-1/2} Z \|_{\psi_2} \leq A_0, \quad S = \mathbb{E} Z Z^\top. \]

For \( t > 0 \), \( \hat{\theta}_\tau \) with \( \tau = \sigma \sqrt{n/(d + t)} \) satisfies, with probability \( \geq 1 - 3e^{-t} \),

\[ \| \hat{\theta}_\tau - \theta^* \|_2 \lesssim \sigma \sqrt{\frac{d + t}{n}} \]

as long as \( n \gtrsim d + t \).
A Data-Driven Method

We estimate $\theta^*$ and tune $\tau$ simultaneously by solving

\[
\begin{cases}
    g_1(\theta, z) \overset{\text{def}}{=} \sum_{i=1}^{n} \min \{ |Y_i - Z_i^\top \theta|, z \} \text{sign}(Y_i - Z_i^\top \theta) Z_i = 0_{d+1} \\
    g_2(\theta, z) \overset{\text{def}}{=} \frac{1}{n - d} \sum_{i=1}^{n} \frac{\min \{ (Y_i - Z_i^\top \theta)^2, z^2 \}}{z^2} - \frac{d + \log n}{n} = 0
\end{cases}
\]

for $\theta \in \mathbb{R}^{d+1}$ and $z > 0$.

Starting with an initial value $\theta^{(0)}$ (OLS), we iteratively obtain $\tau^{(k)}$ by solving $g_2(\theta^{(k-1)}, \cdot) = 0$, and obtain $\theta^{(k)}$ by solving $g_1(\cdot, \tau^{(k)}) = 0$ for $k = 1, 2, \ldots$
Numerical Experiment I

- $n = 500$, $d = 5$
- $X_i$ consists of iid $\text{Unif}(-1.5,1.5)$ entries
- $\theta^* = (5,1,-1,1,-1,1)^T$

**Error distribution:**

1. Normal distribution $\mathcal{N}(0,\sigma^2)$
2. Skewed generalized $t$
3. Lognormal distribution
4. Pareto distribution
Intercept Estimation

(a) $N(0, 2.5)$

(b) sgt$(0, 3, 2, 3, 0.75)$

(c) $LN(0, 1)$

(d) $P(1, 2)$
Average $\ell_2$-error vs Tail Parameter

(e) Normal

(f) Skewed generalized $t$

(g) Lognormal

(h) Pareto
95% Quantile of $\ell_2$-error vs Tail Parameter

(a) Normal

(b) Skewed generalized $t$

(c) Lognormal

(d) Pareto
Two-Stage Robust Regression

Model and data: \( Y_i = \mu^* + X_i^T \beta^* + \varepsilon_i, i = 1, \ldots, n. \)

Idea: Estimate coefficients and intercept separately.

Procedure:
I. In the first stage, solve

\[
\tilde{\theta}_\tau = (\tilde{\mu}_\tau, \tilde{\beta}_\tau^T)^T = \arg\min_{\theta \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \ell_\tau(Y_i - Z_i^T \theta)
\]

for a fixed \( \tau \), say \( \tau = 1.345 \sigma \).

II. Note that \( \mu^* = \mathbb{E}(Y_i - X_i^T \beta^*) \). Define fitted residuals

\[
\tilde{Z}_i = Y_i - X_i^T \tilde{\beta}_\tau, \quad i = 1, \ldots, n.
\]
Run the data-adaptive mean estimation algorithm on \( \{\tilde{Z}_i\}_{i=1}^N \), we obtain an estimator of \( \mu^* \), denoted by \( \hat{\mu} \).

III. The final estimator is \( \hat{\theta} = (\hat{\mu}, \tilde{\beta}_\tau^\top)^\top \).

**Calibration of \( \sigma \):** Apply median absolute deviations (MAD) estimator to the residuals at each iteration.

**Advantage:** Improved statistical accuracy on estimating both coefficients and intercept.
Numerical Experiment II

- $n = 500$, $d = 5$
- $X_i$ consists of iid $\text{Unif}(-1.5,1.5)$ entries
- $\theta^* = (5,1,-1,1,-1,1)^T$

**Error distribution:**
1. Normal distribution $\mathcal{N}(0,\sigma^2)$
2. Skewed generalized $t$
3. Lognormal distribution
4. Pareto distribution
Boxplots of Estimation Error (Intercept)

(a) Normal

(b) Skewed generalized $t$

(c) Lognormal

(d) Pareto
Average Error ( Intercept) vs Tail Parameter

(a) $\mathcal{N}(0, 1)$

(b) $\text{sgt}(0, 5, 0.75, 2, 2.5)$

(c) $\text{LN}(0, 1.5)$

(d) $\text{Par}(1, 2)$
Boxplots of Total Error ($\ell_2$-norm)

(a) $\mathcal{N}(0, 1)$

(b) sgt($0, 5, 0.75, 2, 2.5$)

(c) LN($0, 1.5$)

(d) Par($1, 2$)
Extensions
★ High-Dimensional Sparse Regression

$\beta^*$ is sparse: $s$, number of nonzero elements, is small.

- $\ell_1$-regularized least squares regression (Lasso):

$$
\min_{\theta \in \mathbb{R}^{d+1}} \left\{ \frac{\| Y - \mathcal{Z} \theta \|^2_2}{2n} + \lambda \| \beta \|_1 \right\},
$$

where $\mathcal{Z} = (Z_1, \ldots, Z_n)^\top$.

- $\ell_2$ loss: combination of its rapid growth with heavy-tailed sampling inevitably leads to outliers;

- $\ell_1$ penalty: introduce non-negligible estimation bias.
- $\ell_1$-regularized Huber regression (Huber-Lasso):

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell_\tau(Y_i - Z_i^\top \theta) + \lambda \|\beta\|_1.$$

- Non-convex regularized Huber regression:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell_\tau(Y_i - Z_i^\top \theta) + \sum_{j=1}^{d} p_\lambda(\beta_j),$$

where $p_\lambda : \mathbb{R} \mapsto [0, \infty)$ is a concave penalty function with a regularization parameter $\lambda$. 
- **TAC (Tightening After Contraction) algorithm**: Starting with an initial estimate \( \hat{\theta}^{(0)} = (\hat{\mu}^{(0)}, \hat{\beta}_1^{(0)}, \ldots, \hat{\beta}_d^{(0)})^T \), for \( \ell = 1, 2, \ldots \), solve a sequence of convex programs

\[
\min_{\theta} \frac{1}{n} \sum_{i=1}^n \ell_{\tau}(Y_i - Z_i^T \theta) + \sum_{j=1}^d p_{\lambda}^{'}(\hat{\beta}_{(\ell-1)}^j) |\beta_j|
\]

to obtain \( \{\hat{\theta}^{(\ell)}\}_{\ell \geq 1} \). After a few iterations, \( \hat{\theta}^{(\ell)} \) achieves optimal convergence rate \( \sqrt{\frac{s}{n}} \) & satisfies oracle property.

- **R-package under construction...**
This picture is borrowed from

★ Covariance Matrix Estimation

Observe iid $\mathbb{R}^d$-valued vectors $X_1, \ldots, X_n$ with mean $\mu$ and covariance matrix $\Sigma = (\sigma_{k\ell})_{1 \leq k, \ell \leq d}$.

- Sample covariance matrix $\hat{\Sigma}^S = (\hat{\sigma}_{k\ell})_{1 \leq k, \ell \leq d}$:

$$
\hat{\sigma}_{k\ell} = \frac{1}{\binom{n}{2}} \sum_{i \neq j} (X_{ik} - X_{i\ell})(X_{jk} - X_{j\ell}).
$$

- Truncated covariance estimator $\hat{\Sigma}^T = (\hat{\sigma}^T_{k\ell})_{1 \leq k, \ell \leq d}$:

1. Let $N = n(n - 1)/2$, and define paired data
\{Y_1, \ldots, Y_N\} = \{X_1 - X_2, X_1 - X_3, \ldots, X_{n-1} - X_n\}.

Note that \(\hat{\Sigma}^\mathcal{S} = (1/N) \sum_{i=1}^{N} Y_i Y_i^\top / 2\).

II. For \(1 \leq k \leq \ell \leq d\), we robustify \(\hat{\sigma}_{k\ell}\) as follows:

\[
\hat{\sigma}^\mathcal{T}_{k\ell} = \frac{1}{N} \sum_{i=1}^{N} \psi_\tau(Y_{ik} Y_{i\ell} / 2), \quad \tau = \tau_{k\ell} > 0.
\]

- Huber-type covariance estimator \(\hat{\Sigma}^\mathcal{H} = (\hat{\sigma}^\mathcal{H}_{k\ell})_{1\leq k, \ell \leq d}\):

\[
\hat{\sigma}^\mathcal{H}_{k\ell} = \arg\min_\theta \sum_{i=1}^{N} \ell_\tau(Y_{ik} Y_{i\ell} / 2 - \theta), \quad 1 \leq k, \ell \leq d.
\]
To evaluate empirical performance, consider relative error:

$$
\frac{\| \hat{\Sigma}^H - \Sigma \|_2, \text{max,F}}{\| \hat{\Sigma}^{\mathcal{S}} - \Sigma \|_2, \text{max,F}}.
$$

Table 1: Comparison of relative mean errors for \((n, d) = (50, 200)\)

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>t_3</th>
<th>t_3</th>
<th>Normal</th>
<th>t_3</th>
<th>t_3</th>
<th>Normal</th>
<th>t_3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Max</td>
<td>Frob</td>
<td></td>
<td>Max</td>
<td>Frob</td>
<td></td>
<td>Max</td>
</tr>
<tr>
<td>CV-Huber</td>
<td>0.98</td>
<td>0.95</td>
<td>0.98</td>
<td>0.97</td>
<td>0.94</td>
<td>0.98</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>DA-Huber</td>
<td>0.97</td>
<td>0.95</td>
<td>0.96</td>
<td>0.98</td>
<td>0.96</td>
<td>0.97</td>
<td>0.99</td>
<td>0.98</td>
</tr>
</tbody>
</table>

(b) Equal correlation structure

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>t_3</th>
<th>t_3</th>
<th>Normal</th>
<th>t_3</th>
<th>t_3</th>
<th>Normal</th>
<th>t_3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Max</td>
<td>Frob</td>
<td></td>
<td>Max</td>
<td>Frob</td>
<td></td>
<td>Max</td>
</tr>
<tr>
<td>CV-Huber</td>
<td>0.97</td>
<td>0.94</td>
<td>0.98</td>
<td>0.77</td>
<td>0.21</td>
<td>0.76</td>
<td>0.67</td>
<td>0.34</td>
</tr>
<tr>
<td>DA-Huber</td>
<td>0.98</td>
<td>0.96</td>
<td>0.97</td>
<td>0.77</td>
<td>0.22</td>
<td>0.72</td>
<td>0.62</td>
<td>0.32</td>
</tr>
</tbody>
</table>

(c) Power decay structure

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>t_3</th>
<th>t_3</th>
<th>Normal</th>
<th>t_3</th>
<th>t_3</th>
<th>Normal</th>
<th>t_3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Max</td>
<td>Frob</td>
<td></td>
<td>Max</td>
<td>Frob</td>
<td></td>
<td>Max</td>
</tr>
<tr>
<td>CV-Huber</td>
<td>0.98</td>
<td>0.95</td>
<td>0.97</td>
<td>0.58</td>
<td>0.30</td>
<td>0.71</td>
<td>0.48</td>
<td>0.29</td>
</tr>
<tr>
<td>DA-Huber</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.58</td>
<td>0.28</td>
<td>0.72</td>
<td>0.48</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Average Running Time vs Dimension

- CV–Huber
- DA–Huber

R-package under construction…
★ General Robust Loss Functions

Consider a general loss function $\ell_\tau$:

$$\ell_\tau(x) = \tau^2 \ell(x/\tau), \ x \in \mathbb{R},$$

where $\ell : \mathbb{R} \mapsto [0, \infty)$ satisfies

• (Lipschitz Continuity) $\ell'(0) = 0, \ |\ell'(x)| \leq c_1$;

• (Locally Strong Convexity) $\ell''(0) = 1, \ \ell''(x) \geq c_2$ for $|x| \leq c_3$;

• (Smoothness) $|\ell'(x) - x| \leq c_4 x^2$.

Here $c_1 - c_4$ are absolute positive constants.
Examples:

A. (Huber loss)

\[
\ell(x) = \begin{cases} 
  \frac{x^2}{2} & \text{if } |x| \leq 1, \\
  |x| - 1/2 & \text{if } |x| > 1.
\end{cases}
\]

B. (Pseudo-Huber loss I)

\[
\ell(x) = \sqrt{1 + x^2} - 1;
\]

C. (Pseudo-Huber loss II)

\[
\ell(x) = \log\{(e^x + e^{-x})/2\};
\]
D. (Smoothed Huber loss I)

\[ \ell(x) = \begin{cases} 
  x^2/2 - |x|^3/6 & \text{if } |x| \leq 1, \\
  |x|/2 - 1/6 & \text{if } |x| > 1.
\end{cases} \]

E. (Smoothed Huber loss II)

\[ \ell(x) = \begin{cases} 
  x^2/2 - x^4/24 & \text{if } |x| \leq \sqrt{2}, \\
  (2\sqrt{2}/3)|x| - 1/2 & \text{if } |x| > \sqrt{2}.
\end{cases} \]
(a) Loss function $\ell$

(b) First derivative $\ell'$

(c) Second derivative $\ell''$

(d) Third derivative $\ell'''$
References


