Outlier-robust additive matrix decomposition and robust matrix completion

Philip Thompson

FGV EMAp

COLMEA, PUC-Rio
November, 2023

Least-squares trace-regression

TRACE (LINEAR) REGRESSION: Given an iid sample $\{(y_i^\circ, \mathbf{X}_i)\}$ from a random label-feature pair $(y, \mathbf{X}) \in \mathbb{R} \times \mathbb{R}^{d_1 \times d_2}$, estimate the parameter

$$\mathbf{\Theta}^* \in \operatorname*{argmin}_{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}} \mathbb{E}\left(y - \operatorname{tr}(\mathbf{X}^\top \mathbf{\Theta})\right)^2.$$

Equivalently,

$$y = \operatorname{tr}(\mathbf{X}^\top \mathbf{\Theta}^*) + \xi,$$

for some ξ satisfying $\mathbb{E}[\xi \mathbf{X}] = 0$. We may also assume that $(\mathbf{X}, \xi)$ is centered.

Some notation

Definition

Define the **design operator** $\mathfrak{X} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^n$ with coordinates

$$\mathbf{V} \mapsto \mathfrak{X}_i(\mathbf{V}) := \mathrm{tr}(\mathbf{X}_i^\top \mathbf{V}).$$

Some notation

Definition

Define the **design operator** $\mathfrak{X} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^n$ with coordinates

$$\mathbf{V} \mapsto \mathfrak{X}_i(\mathbf{V}) := \mathrm{tr}(\mathbf{X}_i^\top \mathbf{V}).$$

Notation:

- Let $\boldsymbol{y} := (y_i)_{i \in [n]}$ and $\boldsymbol{\xi} := (\xi_i)_{i \in [n]}$ for which $y_i = \mathrm{tr}(\mathbf{X}_i^\top \boldsymbol{\Theta}^*) + \xi_i$.
- $\langle\!\langle \mathbf{A}, \mathbf{B} \rangle\!\rangle := \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$.
- $A^{(n)} := A/\sqrt{n}$.
- $\mathbb{R}^p := \mathbb{R}^{d_1 \times d_2}$.

Least-squares estimator

It is well known that, under "subgaussian distributions", the *Empirical Risk Minimization* (ERM) methodology justifies why the *least-squares estimator*

$$\hat{\mathbf{\Theta}} \in \underset{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \operatorname{tr}(\mathbf{X}_i^\top \mathbf{\Theta}) \right)^2$$

is the "optimal" choice among other possible estimators.

Least-squares estimator

It is well known that, under "subgaussian distributions", the *Empirical Risk Minimization* (ERM) methodology justifies why the *least-squares estimator*

$$\hat{\mathbf{\Theta}} \in \underset{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \operatorname{tr}(\mathbf{X}_i^\top \mathbf{\Theta}) \right)^2$$

is the "optimal" choice among other possible estimators.

In fact, without further assumptions on $\mathbf{\Theta}^*$, nothing is lost by seeing it simply as a vector (linear regression).

High-dimensional statistics

- In case $p \gg n$, further assume $\boldsymbol{\Theta}^*$ belongs to a parsimonious class $\mathcal{F}$.

High-dimensional statistics

- In case $p \gg n$, further assume $\boldsymbol{\Theta}^*$ belongs to a parsimonious class $\mathcal{F}$.
- *Compressive sensing*: $\boldsymbol{\theta}^* \in \mathbb{R}^p$, $\|\boldsymbol{\theta}^*\|_0 \leq s$ and $\boldsymbol{\xi} = \mathbf{0}$.

High-dimensional statistics

- In case $p \gg n$, further assume $\boldsymbol{\Theta}^*$ belongs to a parsimonious class $\mathcal{F}$.
- *Compressive sensing*: $\boldsymbol{\theta}^* \in \mathbb{R}^p$, $\|\boldsymbol{\theta}^*\|_0 \leq s$ and $\boldsymbol{\xi} = \mathbf{0}$.
- *Sparse linear regression or noisy compressive sensing*: $\boldsymbol{\theta}^* \in \mathbb{R}^p$, $\|\boldsymbol{\theta}^*\|_0 \leq s$.

High-dimensional statistics

- In case $p \gg n$, further assume $\boldsymbol{\Theta}^*$ belongs to a parsimonious class $\mathcal{F}$.
- *Compressive sensing*: $\boldsymbol{\theta}^* \in \mathbb{R}^p$, $\|\boldsymbol{\theta}^*\|_0 \leq s$ and $\boldsymbol{\xi} = \mathbf{0}$.
- *Sparse linear regression or noisy compressive sensing*: $\boldsymbol{\theta}^* \in \mathbb{R}^p$, $\|\boldsymbol{\theta}^*\|_0 \leq s$.
- **Trace regression**: $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$, $\text{rank}(\boldsymbol{\Theta}^*) \leq r$.

High-dimensional statistics

- In case $p \gg n$, further assume $\boldsymbol{\Theta}^*$ belongs to a parsimonious class $\mathcal{F}$.
- *Compressive sensing*: $\boldsymbol{\theta}^* \in \mathbb{R}^p$, $\|\boldsymbol{\theta}^*\|_0 \leq s$ and $\boldsymbol{\xi} = \mathbf{0}$.
- *Sparse linear regression or noisy compressive sensing*: $\boldsymbol{\theta}^* \in \mathbb{R}^p$, $\|\boldsymbol{\theta}^*\|_0 \leq s$.
- **Trace regression**: $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$, $\operatorname{rank}(\boldsymbol{\Theta}^*) \leq r$.
- **Additive matrix decomposition**: $\boldsymbol{\Theta}^* = \mathbf{B}^* + \boldsymbol{\Gamma}^*$, $\operatorname{rank}(\mathbf{B}^*) \leq r$, $\|\boldsymbol{\Gamma}^*\|_0 \leq s$.
  - Identity designs: $\mathfrak{X} = \mathbf{I}_n$
    - *Robust PCA.*
  - Multi-task learning: $\boldsymbol{y}_i = \underbrace{\boldsymbol{x}_i^\top \mathbf{V}}_{\mathfrak{X}_i(\mathbf{V})} + \boldsymbol{\xi}_i$, where the *label is a vector of d_2 "tasks" and feature is a d_1-dimensional vector.*
- (...)

Penalized least-squares estimator

Under "subgaussian distributions and proper class $\mathcal{F}$", a *penalized least-squares estimator*

$$\hat{\mathbf{\Theta}} \in \underset{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \operatorname{tr}(\mathbf{X}_i^\top \mathbf{\Theta}) \right)^2 + \lambda \mathcal{R}(\mathbf{\Theta})$$

is "optimal" for properly chosen penalization hyper-parameter $\lambda > 0$ and regularization norm $\mathcal{R}$.

Penalized least-squares estimator

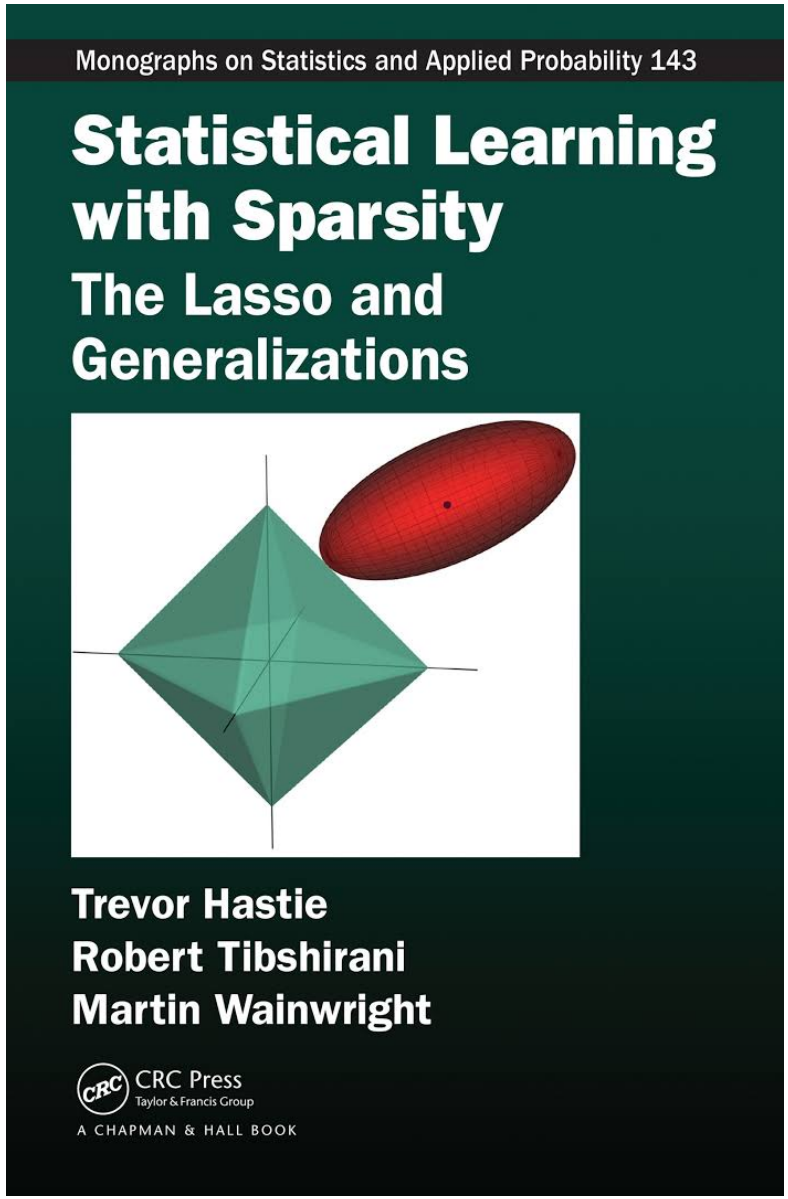


Figure: Chapter 7 on Additive Matrix Decomposition

In this work we are concerned about ...

1. **Additive decomposition in trace regression**: $\mathbf{\Theta}^* = \mathbf{B}^* + \mathbf{\Gamma}^*$ where $\text{rank}(\mathbf{B}^*) \leq r$ and $\|\mathbf{\Gamma}^*\|_0 \leq s$.

In this work we are concerned about ...

1. **Additive decomposition in trace regression**: $\mathbf{\Theta}^* = \mathbf{B}^* + \mathbf{\Gamma}^*$ where $\text{rank}(\mathbf{B}^*) \leq r$ and $\|\mathbf{\Gamma}^*\|_0 \leq s$.
2. **Label contamination** (Robust Statistics/Learning): the observed label sample $\{y_i\}_{i=1}^n$ differs from the iid sample $\{y_i^\circ\}_{i=1}^n$ by o **arbitrary** outliers. The "sample contamination fraction" is

$$\epsilon := \frac{o}{n}.$$

How much ϵ affects the estimation?

In this work we are concerned about ...

1. **Additive decomposition in trace regression**: $\mathbf{\Theta}^* = \mathbf{B}^* + \mathbf{\Gamma}^*$ where $\text{rank}(\mathbf{B}^*) \leq r$ and $\|\mathbf{\Gamma}^*\|_0 \leq s$.
2. **Label contamination** (Robust Statistics/Learning): the observed label sample $\{y_i\}_{i=1}^n$ differs from the iid sample $\{y_i^\circ\}_{i=1}^n$ by o **arbitrary** outliers. The "sample contamination fraction" is

$$\epsilon := \frac{o}{n}.$$

   How much ϵ affects the estimation?
3. **Feature-dependent noise**: "Nothing is assumed" beyond marginal sub-gaussianity of $(\mathbf{X}, \xi)$ and $\mathbb{E}[\xi\mathbf{X}] = \mathbf{0}$.
   - For instance, ξ can be centered non-symmetric and have zero mass around the origin.

In this work we are concerned about ...

1. **Additive decomposition in trace regression**: $\mathbf{\Theta}^* = \mathbf{B}^* + \mathbf{\Gamma}^*$ where $\text{rank}(\mathbf{B}^*) \leq r$ and $\|\mathbf{\Gamma}^*\|_0 \leq s$.
2. **Label contamination** (Robust Statistics/Learning): the observed label sample $\{y_i\}_{i=1}^n$ differs from the iid sample $\{y_i^\circ\}_{i=1}^n$ by o **arbitrary** outliers. The "sample contamination fraction" is

$$\epsilon := \frac{o}{n}.$$

   How much ϵ affects the estimation?
3. **Feature-dependent noise**: "Nothing is assumed" beyond marginal sub-gaussianity of $(\mathbf{X}, \xi)$ and $\mathbb{E}[\xi\mathbf{X}] = \mathbf{0}$.
   - For instance, ξ can be centered non-symmetric and have zero mass around the origin.
4. Dependence on **failure probability** δ: optimal rate and adaptivity.

The classical "LASSO proof"

M-estimation with decomposable regularizers

Let us consider trace-regression assuming

- no matrix decomposition: $\boldsymbol{\Theta}^* = \mathbf{B}^*$ with $\operatorname{rank}(\mathbf{B}^*) \leq r$.
- no label contamination.
- $\boldsymbol{\xi}$ is independent of $\{\mathbf{X}_i\}_{i=1}^n$. E.g. the Gaussian model.

The classical "LASSO proof"

M-estimation with decomposable regularizers

Let us consider trace-regression assuming

- no matrix decomposition: $\mathbf{\Theta}^* = \mathbf{B}^*$ with $\text{rank}(\mathbf{B}^*) \leq r$.
- no label contamination.
- $\boldsymbol{\xi}$ is independent of $\{\mathbf{X}_i\}_{i=1}^n$. E.g. the Gaussian model.

In this case,

$$\hat{\mathbf{B}} \in \min_{\mathbf{B} \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^n (y_i - \langle\langle \mathbf{X}_i, \mathbf{B} \rangle\rangle)^2 + \lambda \mathcal{R}(\mathbf{B}),$$

with $\mathcal{R} = \|\cdot\|_N$ the nuclear norm.

The classical "LASSO proof"

for M-estimation with decomposable regularizers

The previous problem is nonsmooth convex so it is equivalent to its first condition:

$$\exists \mathbf{V} \in \partial \mathcal{R}(\hat{\mathbf{B}}), \forall \mathbf{B} \in \mathbb{R}^p,$$

$$\sum_{i \in [n]} \left[y_i^{(n)} - \mathfrak{x}_i^{(n)}(\widehat{\mathbf{B}}) \right] \langle\!\langle \mathbf{X}_i^{(n)}, \hat{\mathbf{B}} - \mathbf{B} \rangle\!\rangle \geq \lambda \langle\!\langle \mathbf{V}, \hat{\mathbf{B}} - \mathbf{B} \rangle\!\rangle.$$

The classical "LASSO proof"

for M-estimation with decomposable regularizers

Using that

$$\boldsymbol{y}^{(n)} = \mathfrak{X}^{(n)}(\mathbf{B}^*) + \boldsymbol{\xi}^{(n)},$$

and defining $\boldsymbol{\Delta}_{\mathbf{B}^*} := \hat{\mathbf{B}} - \mathbf{B}^*$ one gets

$$\|\mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*})\|_2^2 \leq \langle\!\langle \boldsymbol{\xi}^{(n)}, \mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*}) \rangle\!\rangle - \lambda \langle\!\langle \mathbf{V}, \boldsymbol{\Delta}_{\mathbf{B}^*} \rangle\!\rangle.$$

The classical "LASSO proof"

for M-estimation with decomposable regularizers

Using that

$$\boldsymbol{y}^{(n)} = \mathfrak{X}^{(n)}(\mathbf{B}^*) + \boldsymbol{\xi}^{(n)},$$

and defining $\boldsymbol{\Delta}_{\mathbf{B}^*} := \hat{\mathbf{B}} - \mathbf{B}^*$ one gets

$$\|\mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*})\|_2^2 \leq \langle\!\langle \boldsymbol{\xi}^{(n)}, \mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*}) \rangle\!\rangle - \lambda \langle\!\langle \mathbf{V}, \boldsymbol{\Delta}_{\mathbf{B}^*} \rangle\!\rangle.$$

Lemma

$-\langle\!\langle \mathbf{V}, \boldsymbol{\Delta}_{\mathbf{B}^*} \rangle\!\rangle \leq \mathcal{R}(\mathbf{B}^*) - \mathcal{R}(\hat{\mathbf{B}})$.

The classical "LASSO proof"

for M-estimation with decomposable regularizers

We obtain the "recursion"

$$\|\mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*})\|_2^2 \leq \langle\!\langle \boldsymbol{\xi}^{(n)}, \mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*}) \rangle\!\rangle - \lambda(\mathcal{R}(\mathbf{B}_*) - \mathcal{R}(\hat{\mathbf{B}})).$$

The classical "LASSO proof"

for M-estimation with decomposable regularizers

We obtain the "recursion"

$$\|\mathfrak{X}^{(n)}(\Delta_{\mathbf{B}^*})\|_2^2 \leq \langle\!\langle \boldsymbol{\xi}^{(n)}, \mathfrak{X}^{(n)}(\Delta_{\mathbf{B}^*}) \rangle\!\rangle - \lambda(\mathcal{R}(\mathbf{B}_*) - \mathcal{R}(\hat{\mathbf{B}})).$$

We will need to bound two random processes:

- **upper bound**: the *multiplier process*

$$\mathbf{V} \mapsto \langle\!\langle \boldsymbol{\xi}^{(n)}, \mathfrak{X}^{(n)}(\mathbf{V}) \rangle\!\rangle = \frac{1}{n} \sum_{i=1}^{n} \xi_i \langle\!\langle \mathbf{X}_i, \mathbf{V} \rangle\!\rangle.$$

- **lower bound**: the *quadratic process*

$$\mathbf{V} \mapsto \|\mathfrak{X}^{(n)}(\mathbf{V})\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} \langle\!\langle \mathbf{X}_i, \mathbf{V} \rangle\!\rangle^2.$$

Upper bound: "classical approach"

for M-estimation with decomposable regularizers

The first design property we need is a suitable **upper bound on the multiplier process**. By the **dual-norm inequality**,

$$\langle\!\langle \boldsymbol{\xi}^{(n)}, \mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*}) \rangle\!\rangle \leq \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\xi}_i \mathbf{X}_i \right\|_{\text{op}} \mathcal{R}(\boldsymbol{\Delta}_{\mathbf{B}^*}).$$

Upper bound: "classical approach"

for M-estimation with decomposable regularizers

The first design property we need is a suitable **upper bound on the multiplier process**. By the **dual-norm inequality**,

$$\langle\!\langle \boldsymbol{\xi}^{(n)}, \mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*}) \rangle\!\rangle \leq \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\xi}_i \mathbf{X}_i \right\|_{\mathrm{op}} \mathcal{R}(\boldsymbol{\Delta}_{\mathbf{B}^*}).$$

Lemma (MP)

Assuming "$(\xi, \mathbf{X})$ are independent and marginally 1-subgaussian and $\mathbf{X}$ is isotropic", for any $\delta \in (0,1)$, with probability $\geq 1 - \delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\xi}_i \mathbf{X}_i \right\|_{\mathrm{op}} \leq C \sqrt{\frac{(d_1 + d_2) + \log(1/\delta)}{n}}.$$

Upper bound: "classical approach"

for M-estimation with decomposable regularizers

The first design property we need is a suitable **upper bound on the multiplier process**. By the **dual-norm inequality**,

$$\langle\!\langle \boldsymbol{\xi}^{(n)}, \mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*}) \rangle\!\rangle \leq \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\xi}_i \mathbf{X}_i \right\|_{\mathrm{op}} \mathcal{R}(\boldsymbol{\Delta}_{\mathbf{B}^*}).$$

Lemma (MP)

Assuming "$(\xi, \mathbf{X})$ are independent and marginally 1-subgaussian and $\mathbf{X}$ is isotropic", for any $\delta \in (0,1)$, with probability $\geq 1-\delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\xi}_i \mathbf{X}_i \right\|_{\mathrm{op}} \leq C \sqrt{\frac{(d_1 + d_2) + \log(1/\delta)}{n}}.$$

NOTE: actually, independence is not required.

Upper bound: "classical approach"

for M-estimation with decomposable regularizers

Hence, asking the tuning to be

$$\lambda \geq 2C\sqrt{\frac{(d_1 + d_2) + \log(1/\delta)}{n}},$$

we get

$$\|\mathfrak{X}^{(n)}(\mathbf{\Delta}_{\mathbf{B}^*})\|_2^2 \leq \frac{\lambda}{2}\mathcal{R}(\mathbf{\Delta}_{\mathbf{B}^*}) + \lambda(\mathcal{R}(\mathbf{B}^*) - \mathcal{R}(\hat{\mathbf{B}})).$$

Upper bound: "classical approach"

for M-estimation with decomposable regularizers

Definition (Decomposable norms)

A norm $\mathcal{R}$ over $\mathbb{R}^p$ is said to be decomposable if, for all $\mathbf{B} \in \mathbb{R}^p$, there exists linear map $\mathbf{V} \mapsto \mathcal{P}_{\mathbf{B}}^{\perp}(\mathbf{V})$ such that, for all $\mathbf{V} \in \mathbb{R}^p$, defining $\mathcal{P}_{\mathbf{B}}(\mathbf{V}) := \mathbf{V} - \mathcal{P}_{\mathbf{B}}^{\perp}(\mathbf{V})$,

- $\mathcal{P}_{\mathbf{B}}^{\perp}(\mathbf{B}) = 0$,
- $\langle\!\langle \mathcal{P}_{\mathbf{B}}(\mathbf{V}), \mathcal{P}_{\mathbf{B}}^{\perp}(\mathbf{V}) \rangle\!\rangle = 0$,
- $\mathcal{R}(\mathbf{V}) = \mathcal{R}(\mathcal{P}_{\mathbf{B}}(\mathbf{V})) + \mathcal{R}(\mathcal{P}_{\mathbf{B}}^{\perp}(\mathbf{V}))$.

Upper bound: "classical approach"

for M-estimation with decomposable regularizers

Example (Nuclear norm)

Let $\mathbf{B} \in \mathbb{R}^p$ with rank $r := \text{rank}(\mathbf{B})$, singular values $\{\sigma_j\}_{j\in[r]}$ and singular vector decomposition

$$\mathbf{B} = \sum_{j\in[r]} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^\top .$$

Here $\{\boldsymbol{u}_j\}_{j\in[r]}$ are the left singular vectors spanning the subspace $\mathcal{U}$ and $\{\boldsymbol{v}_j\}_{j\in[r]}$ are the right singular vectors spanning the subspace $\mathcal{V}$. The pair $(\mathcal{U}, \mathcal{V})$ is sometimes referred as the *low-rank support* of $\mathbf{B}$. Given subspace $S \subset \mathbb{R}^\ell$ let $\mathbf{P}_{S^\perp}$ denote the matrix defining the orthogonal projection onto $S^\perp$. Then, the map

$$\mathbf{V} \mapsto \mathcal{P}_{\mathbf{B}}^{\perp}(\mathbf{V}) := \mathbf{P}_{\mathcal{U}^\perp} \mathbf{V} \mathbf{P}_{\mathcal{V}^\perp}^\top$$

satisfy the decomposability condition for the nuclear norm $\|\cdot\|_N$.

Upper bound: "classical approach"

for M-estimation with decomposable regularizers

Lemma

Let $\mathcal{R}$ be a decomposable norm. Let $\mathbf{B}, \hat{\mathbf{B}} \in \mathbb{R}^p$ and $\mathbf{V} := \hat{\mathbf{B}} - \mathbf{B}$. Then

$$\frac{1}{2}\mathcal{R}(\mathbf{V}) + \mathcal{R}(\mathbf{B}) - \mathcal{R}(\hat{\mathbf{B}}) \leq \frac{3}{2}\mathcal{R}(\mathcal{P}_{\mathbf{B}}(\mathbf{V})) - \frac{1}{2}\mathcal{R}(\mathcal{P}_{\mathbf{B}}^{\perp}(\mathbf{V})).$$

Upper bound: "classical approach"

for M-estimation with decomposable regularizers

Lemma

Let $\mathcal{R}$ be a decomposable norm. Let $\mathbf{B}, \hat{\mathbf{B}} \in \mathbb{R}^p$ and $\mathbf{V} := \hat{\mathbf{B}} - \mathbf{B}$. Then

$$\frac{1}{2}\mathcal{R}(\mathbf{V}) + \mathcal{R}(\mathbf{B}) - \mathcal{R}(\hat{\mathbf{B}}) \leq \frac{3}{2}\mathcal{R}(\mathcal{P}_{\mathbf{B}}(\mathbf{V})) - \frac{1}{2}\mathcal{R}(\mathcal{P}_{\mathbf{B}}^{\perp}(\mathbf{V})).$$

In conclusion, for $\mathcal{R} = \|\cdot\|_N$, we get

$$\|\mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*})\|_2^2 \leq \frac{3\lambda}{2}\mathcal{R}(\mathcal{P}_{\mathbf{B}^*}(\boldsymbol{\Delta}_{\mathbf{B}^*})) - \frac{\lambda}{2}\mathcal{R}(\mathcal{P}_{\mathbf{B}^*}^{\perp}(\boldsymbol{\Delta}_{\mathbf{B}^*})).$$

Upper bound: "classical approach"

for M-estimation with decomposable regularizers

In particular, the estimation error $\mathbf{\Delta}_{\mathbf{B}^*} = \hat{\mathbf{B}} - \mathbf{B}^*$ belongs to the **dimension-reduction cone**

$$\mathcal{C}_{\mathbf{B}^*}(3) := \left\{ \mathbf{V} : \mathcal{R}(\mathcal{P}^{\perp}_{\mathbf{B}^*}(\mathbf{V})) \leq 3\mathcal{R}(\mathcal{P}_{\mathbf{B}^*}(\mathbf{V})) \right\}.$$

Upper bound: "classical approach"

for M-estimation with decomposable regularizers

In particular, the estimation error $\mathbf{\Delta}_{\mathbf{B}^*} = \hat{\mathbf{B}} - \mathbf{B}^*$ belongs to the **dimension-reduction cone**

$$\mathcal{C}_{\mathbf{B}^*}(3) := \left\{ \mathbf{V} : \mathcal{R}(\mathcal{P}^{\perp}_{\mathbf{B}^*}(\mathbf{V})) \leq 3\mathcal{R}(\mathcal{P}_{\mathbf{B}^*}(\mathbf{V})) \right\}.$$

This is good because for $\mathbf{V} \in \mathcal{C}_{\mathbf{B}^*}(3)$,

$$\mathcal{R}(\mathbf{V}) \leq 4\mathcal{R}(\mathcal{P}_{\mathbf{B}^*}(\mathbf{V})) \leq 4\sqrt{r}\|\mathbf{V}\|_F.$$

Lower bound

for M-estimation with decomposable regularizers

The second design property we need is to ask for **strong-convexity restricted to the dimension cone**. Turns out that this a consequence of

Definition (RSC)

$\mathfrak{X}$ satisfies $\mathrm{RSC}_{\mathcal{R}}(a_1, a_2)$ if for all $\mathbf{V} \in \mathbb{R}^p$,

$$\|\mathfrak{X}^{(n)}(\mathbf{V})\|_2 \geq a_1 \|\mathbf{V}\|_\Pi - a_2 \mathcal{R}(\mathbf{V}).$$

Here:

$$\|\mathbf{V}\|_\Pi^2 := \mathbb{E}[\langle\langle \mathbf{X}, \mathbf{V} \rangle\rangle^2].$$

In case, $\mathbf{X}$ is isotropic, $\|\mathbf{V}\|_\Pi = \|\mathbf{V}\|_F$.

Lower bound

for M-estimation with decomposable regularizers

Lemma (RSC)

"Assume $\mathbf{X}$ *is* 1*-subgaussian and isotropic". Suppose that* $n \gtrsim 1 + \log(1/\delta)$. *Then, with probability* $\geq 1 - \delta$, $\mathrm{RSC}_{\mathcal{R}}(\mathrm{a}_1, \mathrm{a}_2)$ *holds with constants* $\mathrm{a}_1 \in (0, 1)$ *and*

$$\mathrm{a}_2 \lesssim \sqrt{\frac{r}{n}}.$$

Lower bound

for M-estimation with decomposable regularizers

Lemma (RSC)

"Assume $\mathbf{X}$ *is* 1*-subgaussian and isotropic". Suppose that* $n \gtrsim 1 + \log(1/\delta)$. *Then, with probability* $\geq 1 - \delta$, $\mathrm{RSC}_{\mathcal{R}}(\mathrm{a}_1, \mathrm{a}_2)$ *holds with constants* $\mathrm{a}_1 \in (0, 1)$ *and*

$$\mathrm{a}_2 \lesssim \sqrt{\frac{r}{n}}.$$

In conclusion, if

$$n \gtrsim r \vee (1 + \log(1/\delta)),$$

then restricted strong convexity holds:

$$\inf_{\boldsymbol{v} \in \mathcal{C}_{\mathbf{B}^*}(3)} \frac{\|\mathfrak{X}^{(n)}\mathbf{V}\|_2}{\|\mathbf{V}\|_2} \geq \frac{\mathrm{a}_1}{2}.$$

Conclusion of "LASSO proof"

for M-estimation with decomposable regularizers

Recall:

- $\boldsymbol{\Delta}_{\mathbf{B}^*} \in \mathcal{C}_{\mathbf{B}^*}(3)$.
- $\|\mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*})\|_2^2 \leq \frac{3\lambda}{2}\mathcal{R}(\mathcal{P}_{\mathbf{B}^*}(\boldsymbol{\Delta}_{\mathbf{B}^*})) - \frac{\lambda}{2}\mathcal{R}(\mathcal{P}^{\perp}_{\mathbf{B}^*}(\boldsymbol{\Delta}_{\mathbf{B}^*}))$.

In conclusion,

$$\frac{a_1}{2}\|\boldsymbol{\Delta}_{\mathbf{B}^*}\|_2^2 \leq \|\mathfrak{X}^{(n)}(\boldsymbol{\Delta}_{\mathbf{B}^*})\|_2^2 \leq \frac{3\lambda}{2}\mathcal{R}(\mathcal{P}_{\mathbf{B}^*}(\boldsymbol{\Delta}_{\mathbf{B}^*})) \leq \frac{3\lambda}{2}\sqrt{r}\|\boldsymbol{\Delta}_{\mathbf{B}^*}\|_2,$$

so, "on the event that both design properties hold"

with probability at least $\geq 1 - \delta$,

$$\|\boldsymbol{\Delta}_{\mathbf{B}^*}\|_2 \leq \frac{3\lambda\sqrt{r}}{a_1} \lesssim \sqrt{\frac{r(d_1 + d_2)}{n}} + \sqrt{\frac{r\log(1/\delta)}{n}}.$$

Failure probability δ

Some observations:

- One can show that

$$\sqrt{\frac{r(d_1 + d_2)}{n}}$$

is the **optimal rate in average**.

Failure probability δ

Some observations:

- One can show that

$$\sqrt{\frac{r(d_1 + d_2)}{n}}$$

  is the **optimal rate in average**.

- The penalization is r-adaptive. Still, the approach using the dual norm inequality leads to the δ-**dependent tuning**

$$\lambda \asymp \sqrt{\frac{(d_1 + d_2) + \log(1/\delta)}{n}}.$$

Failure probability δ

Some observations:

- One can show that
$$\sqrt{\frac{r(d_1 + d_2)}{n}}$$
is the **optimal rate in average**.
- The penalization is r-adaptive. Still, the approach using the dual norm inequality leads to the δ-**dependent tuning**
$$\lambda \asymp \sqrt{\frac{(d_1 + d_2) + \log(1/\delta)}{n}}.$$
- What is the **optimal rate in probability**? Under a reasonable regime, is δ-dependent tuning necessary?

Failure probability δ

- Bellec-Guillaume-Tsybakov (2018) was the first to answer these questions in the case of a sparse parameter. The changes for the case of a low-rank parameter are trivial.

Failure probability δ

- Bellec-Guillaume-Tsybakov (2018) was the first to answer these questions in the case of a sparse parameter. The changes for the case of a low-rank parameter are trivial.
- From their bounds, the optimal rate in probability is

$$\sqrt{\frac{r(d_1 + d_2)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}},$$

showing the previous **multiplicative term** $r \log(1/\delta)$ is suboptimal.

Failure probability δ

- Bellec-Guillaume-Tsybakov (2018) was the first to answer these questions in the case of a sparse parameter. The changes for the case of a low-rank parameter are trivial.
- From their bounds, the optimal rate in probability is

$$\sqrt{\frac{r(d_1+d_2)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}},$$

  showing the previous **multiplicative term** $r\log(1/\delta)$ is suboptimal.
- This is achieved with an δ-**adapted** tuning

$$\lambda \asymp \sqrt{\frac{(d_1+d_2)}{n}}.$$

Rephrasing MP in Bellec et al (2018)

Definition (MP)

We will say $(\mathfrak{X}, \boldsymbol{\xi})$ satisfies $\mathrm{MP}_{\mathcal{R}}(f_1, f_2)$ if for all $\mathbf{V} \in \mathbb{R}^p$,

$$|\langle \boldsymbol{\xi}^{(n)}, \mathfrak{X}^{(n)}(\mathbf{V}) \rangle| \leq f_1 \|\mathbf{V}\|_\Pi + f_2 \mathcal{R}(\mathbf{V}).$$

Rephrasing MP in Bellec et al (2018)

Observations:

- Via dual-norm inequality, one proves MP with

$$f_1 = 0, \quad f_2 \asymp \sqrt{\frac{d_1 + d_2 + \log(1/\delta)}{n}}.$$

Rephrasing MP in Bellec et al (2018)

Observations:

- Via dual-norm inequality, one proves MP with

$$\mathsf{f}_1 = 0, \quad \mathsf{f}_2 \asymp \sqrt{\frac{d_1 + d_2 + \log(1/\delta)}{n}}.$$

- Bellec et al (2018) proved MP with **non-null** f_1, namely,

$$\mathsf{f}_1 \asymp \sqrt{\frac{1 + \log(1/\delta)}{n}}, \quad \mathsf{f}_2 \asymp \sqrt{\frac{d_1 + d_2}{n}}.$$

Rephrasing MP in Bellec et al (2018)

Observations:

- Via dual-norm inequality, one proves MP with

$$\mathsf{f}_1 = 0, \quad \mathsf{f}_2 \asymp \sqrt{\frac{d_1 + d_2 + \log(1/\delta)}{n}}.$$

- Bellec et al (2018) proved MP with **non-null** f_1, namely,

$$\mathsf{f}_1 \asymp \sqrt{\frac{1 + \log(1/\delta)}{n}}, \quad \mathsf{f}_2 \asymp \sqrt{\frac{d_1 + d_2}{n}}.$$

- This is the technical reason for the previous improvement in terms of the **failure probability** δ.

A caveat in Bellec et al (2018)

Their approach in proving MP with non-null constant f_1 works in case $(\xi, \mathbf{X})$ is independent:

Define the random norm

$$\hat{T}(\mathbf{V}) := \frac{\|\mathfrak{X}^{(n)}(\mathbf{V})\|_2}{L} \vee \|\mathbf{V}\|_N$$

with $L := \sqrt{n/\log(1/\delta)}/\sigma$. In case $\mathfrak{X}$ is fixed, they bound the multiplier process concentrating the **linear process**

$$\sup_{\mathbf{V}\in\mathbb{B}_{\hat{T}}} \langle \boldsymbol{\xi}^{(n)}, \mathbf{V}\rangle = \frac{1}{\sqrt{n}} \sum_{i\in[n]} \xi_i \mathbf{V}_i.$$

— see Proposition 9.2 in Bellec et al (2018). The proof of this elegant result follows from a simple application of a tail symmetrization-comparison argument and the gaussian concentration inequality. Peeling is not necessary — as homogeneity of norms suffices.

A sharp Multiplier Process Inequality

The *multiplier process* over functions $f \in F$ is defined as

$$M(f) := \frac{1}{n} \sum_{i \in [n]} (\xi_i f(X_i) - \mathbb{E}[\xi f(X)]).$$

"Assume iid sample, ξ is subgaussian and $F \ni f \mapsto f(X)$ is subgaussian".

Theorem (Multiplier process)

There exists universal constant $c > 0$, such that for all $f_0 \in F$, $n \geq 1$, $u \geq 1$ and $v \geq 1$, with probability at least $1 - ce^{-u/4} - ce^{-nv}$,

$$\sup_{f \in F} |M(f) - M(f_0)| \lesssim \left(\sqrt{v} + 1\right) \|\xi\|_{\psi_2} \frac{\gamma_2(F)}{\sqrt{n}} + \left(\sqrt{\frac{2u}{n}} + \frac{u}{n} + \sqrt{\frac{uv}{n}}\right) \|\xi\|_{\psi_2} \bar{\Delta}(F).$$

A sharp Multiplier Process Inequality

- "Nothing is assumed" beyond marginal subgaussianity of (ξ, X).
- The "sharpness" lies in the fact that the **confidence parameter** $u > 0$ does not appear in the **effective dimension** term.
- The proof follows from generic chaining bounds, pioneered by Talagrand (for the empirical process). Precisely, we follow a method from Dirksen (2015) showing sharp bounds for the **quadratic process**.
- NOTE: Mendelson (2014) already obtains impressive MP inequalities **that holds for heavy-tailed processes**. Particularizing for the subgaussian class, the confidence parameter still multiplies the "effective dimension".

Corollary

- We can prove MP with

$$\mathsf{f}_1 \asymp \sqrt{\frac{1 + \log(1/\delta)}{n}}, \quad \mathsf{f}_2 \asymp \sqrt{\frac{d_1 + d_2}{n}},$$

  assuming only marginal subgaussianity of $(\xi, \mathbf{X})$ and $\mathbb{E}[\xi \mathbf{X}] = 0$.
- Without independence assumption, obtain optimal rate in probability with δ-adapted tuning.

Additive decomposition

- It is now well understood that it is hopeless to “separate two low-rank and sparse components” without imposing a identifiability condition.

Additive decomposition

- It is now well understood that it is hopeless to "separate two low-rank and sparse components" without imposing a identifiability condition.
- Since Candès et al (2011), multiple works have studied matrix decomposition or matrix completion under the notion of **incoherence**.
- Since Wainwright and collaborators, e.g. Agarwal et al (2012), other works use the notion of **low-spikeness**

$$\|\mathbf{B}^*\|_\infty \leq \frac{a}{\sqrt{d_1 d_2}}.$$

Additive decomposition (no contamination)

- One of the issues with additive decomposition is to ensure restricted strong convexity. Agarwal et al (2012) proposes a suitable notion of RSC to ensure convergence of a penalized least-squares estimator.

Additive decomposition (no contamination)

- One of the issues with additive decomposition is to ensure restricted strong convexity. Agarwal et al (2012) proposes a suitable notion of RSC to ensure convergence of a penalized least-squares estimator.
- In the concrete applications considered, showing this property holds with high probability is easy.
- For instance, in multi-task learning, the design components are $\mathfrak{X}_i(\mathbf{B}) := \boldsymbol{x}_i^\top \mathbf{B}$. When $n \gtrsim d_1$, standard concentration inequalities imply $\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{b})^2 \geq c\|\boldsymbol{b}\|_\Pi^2$ for all $\boldsymbol{b} \in \mathbb{R}^{d_1}$ with high probability for some absolute constant $c \in (0,1)$. Thus, $\frac{1}{n}\|\mathfrak{X}(\mathbf{B})\|_F^2 \geq c\|\mathbf{B}\|_\Pi^2$ for all $\mathbf{B} \in \mathbb{R}^p$.

Additive decomposition (no contamination)

- One of the issues with additive decomposition is to ensure restricted strong convexity. Agarwal et al (2012) proposes a suitable notion of RSC to ensure convergence of a penalized least-squares estimator.
- In the concrete applications considered, showing this property holds with high probability is easy.
- For instance, in multi-task learning, the design components are $\mathfrak{X}_i(\mathbf{B}) := \boldsymbol{x}_i^\top \mathbf{B}$. When $n \gtrsim d_1$, standard concentration inequalities imply $\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{b})^2 \geq c\|\boldsymbol{b}\|_\Pi^2$ for all $\boldsymbol{b} \in \mathbb{R}^{d_1}$ with high probability for some absolute constant $c \in (0,1)$. Thus, $\frac{1}{n}\|\mathfrak{X}(\mathbf{B})\|_F^2 \geq c\|\mathbf{B}\|_\Pi^2$ for all $\mathbf{B} \in \mathbb{R}^p$.
- IMPORTANT: in **trace-regression**, the design operator is singular when $n \leq d_1 d_2$, so ensuring restricted strong convexity under additive decomposition is harder.

Label contamination in sparse regression

In sparse linear regression with label contamination,

$$\boldsymbol{y} = \mathbb{X}(\boldsymbol{b}^*) + \sqrt{n}\boldsymbol{\theta}^* + \boldsymbol{\xi},$$

with an **arbitrary** o-sparse vector $\boldsymbol{\theta}^*$.

Prior work has considered the estimator

$$[\hat{\boldsymbol{b}}, \hat{\boldsymbol{\theta}}] \in \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p, \boldsymbol{\theta}\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{x}_i, \boldsymbol{b}\rangle - \sqrt{n}\boldsymbol{\theta}\right)^2 + \lambda\|\boldsymbol{b}\|_1 + \tau\|\boldsymbol{\theta}\|_1.$$

This is in fact penalized Huber regression:

$$\hat{\boldsymbol{b}} \quad \in \quad \operatorname{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \tau^2 \textstyle\sum_{i=1}^{n} \Phi\left(\frac{y_i - \langle \boldsymbol{x}_i, \boldsymbol{b}\rangle}{\tau\sqrt{n}}\right) + \lambda\|\boldsymbol{b}\|_1.$$

Label contamination in sparse regression

- Dalalyan-Thompson (2019): identified and proved a design property (IP) enabling to show near-optimality (up to $\log n$ and $\log(1/\epsilon)$ factors) of sparse Huber regression.
  - Chevet's inequality.

Label contamination in sparse regression

- Dalalyan-Thompson (2019): identified and proved a design property (IP) enabling to show near-optimality (up to $\log n$ and $\log(1/\epsilon)$ factors) of sparse Huber regression.
  - Chevet's inequality.
- Chinot (2019): removed both logs and allows heavy-tailed noise.
  - Unlike Dalalyan-Thompson (2019), allows feature-dependent noise and optimal rate in δ.
  - BUT for the **oblivious model** and **assuming knowledge of** (s, o) **and convexity constant**.
  - Mild additional conditions: symmetric noise with positivy mass around the origin.
  - PROOF METHOD: "localization + sparsity equation" (Lecué-Mendelson) instead of "M-estimation with decomposable regularizers".

Label contamination in sparse regression

- Dalalyan-Thompson (2019): identified and proved a design property (IP) enabling to show near-optimality (up to $\log n$ and $\log(1/\epsilon)$ factors) of sparse Huber regression.
  - Chevet's inequality.
- Chinot (2019): removed both logs and allows heavy-tailed noise.
  - Unlike Dalalyan-Thompson (2019), allows feature-dependent noise and optimal rate in δ.
  - BUT for the **oblivious model** and **assuming knowledge of** (s, o) **and convexity constant**.
  - Mild additional conditions: symmetric noise with positivy mass around the origin.
  - PROOF METHOD: "localization + sparsity equation" (Lecué-Mendelson) instead of "M-estimation with decomposable regularizers".
- Both works use δ-non-adapted estimators.

Our estimators

Definition (Slope norm)

Given nonincreasing positive sequence $\boldsymbol{\omega} := \{\omega_i\}_{i\in[n]}$, the Slope norm at a point $\boldsymbol{u} \in \mathbb{R}^n$ is defined by $\|\boldsymbol{u}\|_\sharp := \sum_{i\in[n]} \omega_i \boldsymbol{u}_i^\sharp$, where $\boldsymbol{u}_1^\sharp \geq \ldots \geq \boldsymbol{u}_n^\sharp$ denotes the nonincreasing rearrangement of the absolute coordinates of $\boldsymbol{u}$. Unless otherwise stated, $\boldsymbol{\omega} \in \mathbb{R}^n$ will be the sequence with coordinates $\omega_i = \sqrt{\log(An/i)}$ for some $A \geq 2$.

(2015) BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J.

Our estimators

Definition (Sorted Huber-type losses)

Define the functions $\rho_1(\boldsymbol{u}) := \|\boldsymbol{u}\|_2$ and $\rho_2(\boldsymbol{u}) := \frac{1}{2}\|\boldsymbol{u}\|_2^2$ over $\mathbb{R}^n$. For $q \in \{1, 2\}$ and $\tau > 0$, let $\rho_{\tau\boldsymbol{\omega},q} : \mathbb{R}^n \to \mathbb{R}_+$ be the infimal convolution of ρ_q and $\tau\|\cdot\|_\sharp$, i.e.,

$$\rho_{\tau\boldsymbol{\omega},q}(\boldsymbol{u}) := \min_{\boldsymbol{z}\in\mathbb{R}^n} \rho_q(\boldsymbol{u} - \boldsymbol{z}) + \tau\|\boldsymbol{z}\|_\sharp.$$

Finally, define the loss

$$\mathcal{L}_{\tau\boldsymbol{\omega},q}(\mathbf{B}) := \rho_{\tau\boldsymbol{\omega},q}\left(\boldsymbol{y} - \mathfrak{X}(\mathbf{B})/\sqrt{n}\right).$$

When $\omega_1 = \ldots = \omega_n = 1$,

$$\mathcal{L}_{\tau\boldsymbol{\omega},2}(\mathbf{B}) = \tau^2 \sum_{i=1}^{n} \Phi\left(\frac{y_i - \mathfrak{X}_i(\mathbf{B})}{\tau\sqrt{n}}\right),$$

where $\Phi : \mathbb{R} \to \mathbb{R}$ is the Huber's function.

Our estimators

In case of **matrix decomposition**,

$$\begin{aligned}[\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}] \quad \in \quad & \operatorname{argmin}_{[\mathbf{B},\mathbf{\Gamma}]\in(\mathbb{R}^p)^2} \mathcal{L}_{\tau\omega,2}(\mathbf{B}+\mathbf{\Gamma}) + \lambda\mathcal{R}(\mathbf{B}) + \chi\mathcal{S}(\mathbf{\Gamma}) \\ & \text{s.t.} \quad \|\mathbf{B}\|_\infty \leq a,\end{aligned}$$

or equivalently,

$$\begin{aligned}& \min_{[\mathbf{B},\mathbf{\Gamma},\boldsymbol{\theta}]\in(\mathbb{R}^p)^2\times\mathbb{R}^n} \frac{1}{2n}\sum_{i=1}^n \left(y_i - \langle\!\langle \mathbf{X}_i, \mathbf{B}+\mathbf{\Gamma}\rangle\!\rangle + \sqrt{n}\boldsymbol{\theta}_i\right)^2 + \Psi(\mathbf{B},\mathbf{\Gamma},\boldsymbol{\theta}) \\ & \text{s.t.} \quad \|\mathbf{B}\|_\infty \leq a,\end{aligned}$$

where

$$\Psi(\mathbf{B},\mathbf{\Gamma},\boldsymbol{\theta}) := \lambda\mathcal{R}(\mathbf{B}) + \chi\mathcal{S}(\mathbf{\Gamma}) + \tau\|\boldsymbol{\theta}\|_\sharp.$$

Optimization: alternated proximal gradient between ℓ_∞-constrained/ℓ_1-norm (ℓ_∞-constrained soft-thresholding) and nuclear-norm (soft-thresholded SVD).

Our estimators

Desconsidering matrix decomposition, we consider, for $q \in \{1, 2\}$, estimators of the form

$$\hat{\mathbf{B}} \quad \in \quad \operatorname{argmin}_{\mathbf{B} \in \mathbb{R}^p} \mathcal{L}_{\tau\omega, q}(\mathbf{B}) + \lambda \mathcal{R}(\mathbf{B}).$$

Equivalently, for $q = 2$,

$$\min_{[\mathbf{B}, \theta] \in \mathbb{R}^p \times \mathbb{R}^n} \frac{1}{2n} \sum_{i=1}^n \left(y_i - \langle\!\langle \mathbf{X}_i, \mathbf{B} \rangle\!\rangle + \sqrt{n}\theta_i \right)^2 + \lambda \mathcal{R}(\mathbf{B}) + \tau \|\theta\|_\sharp,$$

and, for $q = 1$,

$$\min_{[\mathbf{B}, \theta] \in \mathbb{R}^p \times \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^n \left(y_i - \langle\!\langle \mathbf{X}_i, \mathbf{B} \rangle\!\rangle + \sqrt{n}\theta_i \right)^2 \right\}^{\frac{1}{2}} + \lambda \mathcal{R}(\mathbf{B}) + \tau \|\theta\|_\sharp.$$

When $\theta \equiv \mathbf{0}$ and $\mathcal{R} = \|\cdot\|_1$, the above estimator corresponds to the *square-root lasso estimator*.

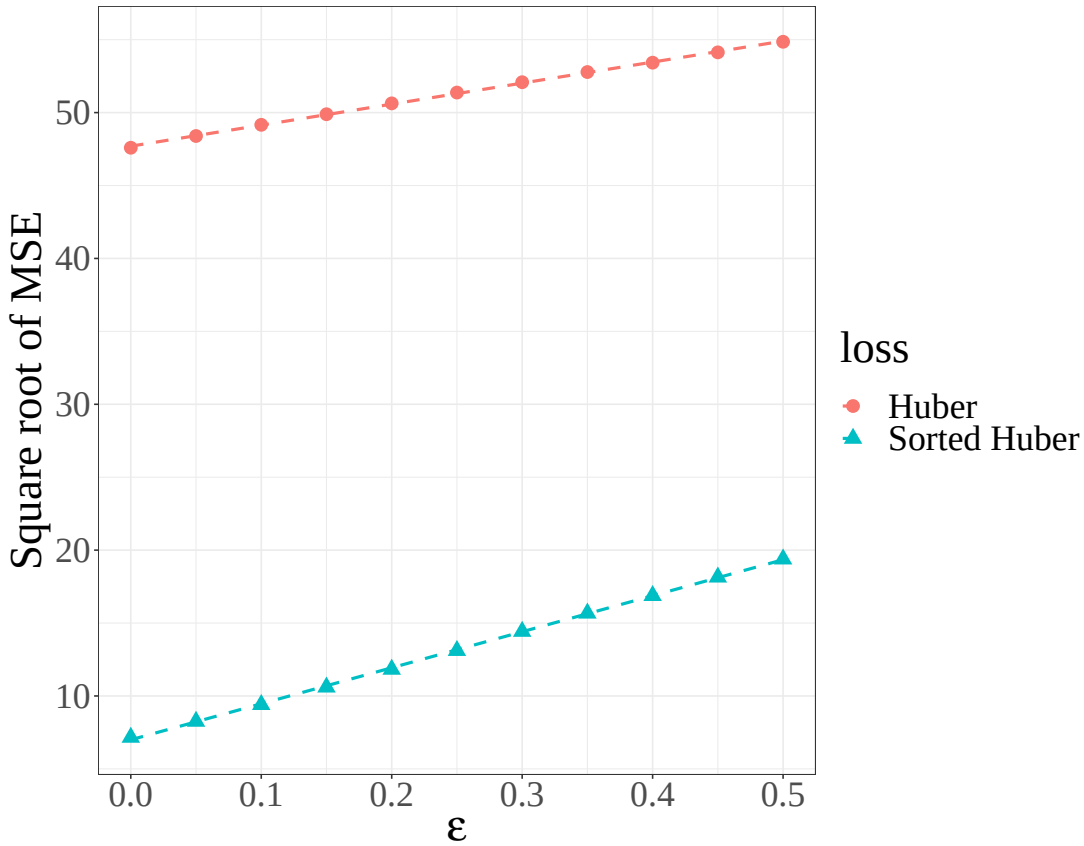


Figure: Huber vs "Sorted" Huber losses in sparse regression: $\sqrt{\mathrm{MSE}}$ versus ϵ.

Upper bound

with additive decomposition, contamination & feature-dependent noise

We will need two design properties to upper bound the "**perturbed multiplier process**"

$$[\mathbf{V}, \mathbf{W}, \boldsymbol{u}] \mapsto \langle \boldsymbol{\xi}^{(n)} - \boldsymbol{u}, \mathfrak{X}^{(n)}(\mathbf{V} + \mathbf{W}) \rangle$$
$$= \frac{1}{n} \sum_{i=1}^{n} \xi_i \langle\!\langle \mathbf{X}_i, \mathbf{V} + \mathbf{W} \rangle\!\rangle - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i \langle\!\langle \mathbf{X}_i, \mathbf{V} + \mathbf{W} \rangle\!\rangle .$$

Moreover, we **wish to avoid mere use of dual-norm inequalities**: optimality in δ.

Upper bound

with additive decomposition, contamination & feature-dependent noise

Definition

- $(\mathfrak{X}, \boldsymbol{\xi})$ satisfies $\mathrm{MP}_{\mathcal{R},\mathcal{S},\mathcal{Q}}(\mathrm{f}_1, \mathrm{f}_2, \mathrm{f}_3, \mathrm{f}_4)$ if for all $[\mathbf{V}, \mathbf{W}, \boldsymbol{u}] \in (\mathbb{R}^p)^2 \times \mathbb{R}^n$,

$$\left|\left\langle \boldsymbol{\xi}^{(n)}, \mathfrak{X}^{(n)}(\mathbf{V} + \mathbf{W}) + \sqrt{n}\boldsymbol{u} \right\rangle\right| \leq \mathrm{f}_1 \|[\mathbf{V}, \mathbf{W}, \boldsymbol{u}]\|_\Pi + \mathrm{f}_2 \mathcal{R}(\mathbf{V}) + \mathrm{f}_3 \mathcal{S}(\mathbf{W}) + \mathrm{f}_4 \mathcal{Q}(\boldsymbol{u}).$$

- $\mathfrak{X}$ satisfies $\mathrm{IP}_{\mathcal{R},\mathcal{S},\mathcal{Q}}(\mathrm{b}_1, \mathrm{b}_2, \mathrm{b}_3, \mathrm{b}_4)$ if for all $[\mathbf{V}, \mathbf{W}, \boldsymbol{u}] \in (\mathbb{R}^p)^2 \times \mathbb{R}^n$,

$$\begin{aligned}\left|\left\langle \boldsymbol{u}, \mathfrak{X}^{(n)}(\mathbf{V} + \mathbf{W}) \right\rangle\right| \leq\ & \mathrm{b}_1 \|[\mathbf{V}, \mathbf{W}]\|_\Pi \|\boldsymbol{u}\|_2 \\ & + \mathrm{b}_2 \mathcal{R}(\mathbf{V}) \|\boldsymbol{u}\|_2 + \mathrm{b}_3 \mathcal{S}(\mathbf{W}) \|\boldsymbol{u}\|_2 \\ & + \mathrm{b}_4 \|[\mathbf{V}, \mathbf{W}]\|_\Pi \mathcal{Q}(\boldsymbol{u}).\end{aligned}$$

Lower bound

with additive decomposition, contamination & feature-dependent noise

With additive decomposition and label contamination, "standard restricted convexity is not enough". We need a lower bound in terms of a "**augmented restricted convexity**" for the augmented design

$$[\mathbf{V}, \mathbf{W}, \boldsymbol{u}] \mapsto \mathfrak{M}^{(n)}(\mathbf{V}, \mathbf{W}, \boldsymbol{u}) := \mathfrak{X}^{(n)}(\mathbf{V} + \mathbf{W}) + \boldsymbol{u}.$$

A third design property we will need is

Definition (ARSC)

$\mathfrak{X}$ satisfies $\mathrm{ARSC}_{\mathcal{R},\mathcal{S},\mathcal{Q}}(\mathrm{d}_1, \mathrm{d}_2, \mathrm{d}_3, \mathrm{d}_4)$ if for all $[\mathbf{V}, \mathbf{W}, \boldsymbol{u}] \in (\mathbb{R}^p)^2 \times \mathbb{R}^n$,

$$\left\{ \|\mathfrak{X}^{(n)}(\mathbf{V} + \mathbf{W}) + \sqrt{n}\boldsymbol{u}\|_2^2 - 2\langle\!\langle \mathbf{V}, \mathbf{W} \rangle\!\rangle_\Pi \right\}_+^{\frac{1}{2}} \geq \mathrm{d}_1 \|[\mathbf{V}, \mathbf{W}, \boldsymbol{u}]\|_\Pi - \mathrm{d}_2 \mathcal{R}(\mathbf{V}) - \mathrm{d}_3 \mathcal{S}(\mathbf{W}) - \mathrm{d}_4 \mathcal{Q}(\boldsymbol{u}).$$

Lower bound

with additive decomposition, contamination & feature-dependent noise

We can show that ARSC holds if both IP and a fourth design property holds:

Definition (PP)

$\mathfrak{X}$ satisfies $\mathrm{PP}_{\mathcal{R},\mathcal{S}}(c_1, c_2, c_3, c_4)$ if for all $[\mathbf{V}, \mathbf{W}] \in (\mathbb{R}^p)^2$,

$$\begin{aligned}|\langle\!\langle \mathbf{V}, \mathbf{W} \rangle\!\rangle_n - \langle\!\langle \mathbf{V}, \mathbf{W} \rangle\!\rangle_\Pi| \leq c_1 \|\mathbf{V}\|_\Pi \|\mathbf{W}\|_\Pi \\ + c_2 \mathcal{R}(\mathbf{V}) \|\mathbf{W}\|_\Pi + c_3 \|\mathbf{V}\|_\Pi \mathcal{S}(\mathbf{W}) \\ + c_4 \mathcal{R}(\mathbf{V}) \mathcal{S}(\mathbf{W}).\end{aligned}$$

A sharp Product Process Inequality

The *product process* is defined as

$$A(f,g) := \frac{1}{n}\sum_{i\in[n]}\left\{f(X_i)g(X_i) - \mathbb{E}f(X_i)g(X_i)\right\},$$

over two distinct classes F and G of measurable functions. When $F = G$, the correspondent process is often termed the *quadratic process*.

"Assume iid sample, and the maps $F \ni f \mapsto f(X)$ and $G \ni g \mapsto G(X)$ are subgaussian".

Theorem (Product process)

Let F, G be subclasses of L_{ψ_2}. There exist universal constants $c, C > 0$, such that for all $n \geq 1$ and $u \geq 1$, with probability at least $1 - e^{-u}$,

$$\sup_{(f,g)\in F\times G}|A(f,g)| \leq C\left[\frac{\gamma_2(F)\gamma_2(G)}{n} + \bar{\Delta}(F)\frac{\gamma_2(G)}{\sqrt{n}} + \bar{\Delta}(G)\frac{\gamma_2(F)}{\sqrt{n}}\right]$$
$$+ c\sup_{(f,g)\in F\times G}\|fg - \mathbf{P}fg\|_{\psi_1}\left(\sqrt{\frac{u}{n}} + \frac{u}{n}\right).$$

Results with additive matrix decomposition and label contamination

Theorem ($q = 2$)

Assume that the covariate is 1*-subgaussian with identity covariance and the noise is* 1*-subgaussian (for simplicity). Assume the low-spikeness condition*

$$\|\mathbf{B}^*\|_\infty \leq \frac{\mathrm{a}^*}{\sqrt{n}}.$$

Assume that $\epsilon \leq c < 0.5$. *Assume* $\mathbf{B}^* \in \mathbb{R}^{d_1 \times d_2}$ *has rank* r *and* $\mathbf{\Gamma}^*$ *is* s*-sparse. Take tuning* $\mathrm{a} := \mathrm{a}^*/\sqrt{n}$ *and*

$$\lambda \asymp \sqrt{(d_1 + d_2)/n}, \quad \chi \asymp \sqrt{\frac{\log(d_1 d_2)}{n}} + \frac{\mathrm{a}^*}{\sqrt{n}}, \quad \tau \asymp \frac{1}{\sqrt{n}}.$$

For any failure probability $\delta \in (0, 1)$ *assume that*

$$n \geq Cr(d_1 + d_2) + Cs\log(d_1 d_2) + C(\mathrm{a}^*)^2 s,$$
$$\delta \geq \exp\left(-c_0 n\right).$$

Results with additive matrix decomposition

Theorem ($q = 2$)

Then with probability at least $1 - \delta$*, the square root of MSE is*

$$\sqrt{r(d_1 + d_2)/n} + \sqrt{s \log p/n} + \mathrm{a}^* \sqrt{s/n} + \sqrt{\log(1/\delta)/n} + \epsilon \log(1/\epsilon).$$

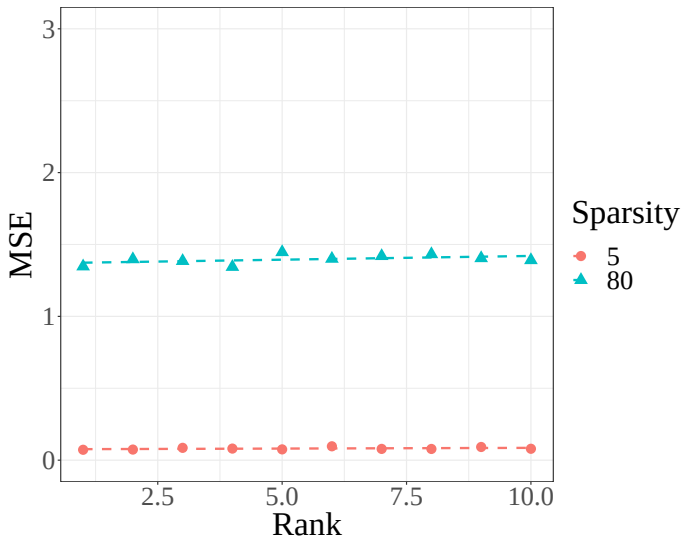
OBSERVATIONS:

- Rate is optimal up to log factors $\log(d_1 d_2)$ and $\log(1/\epsilon)$.
- Optimal rate in δ, δ-uniformity and δ-adaptivity.
- For simplicity, I do not present the formal rate in case of **miss-specification**: it holds as long as there is $[\mathbf{B}, \mathbf{\Gamma}]$ such that

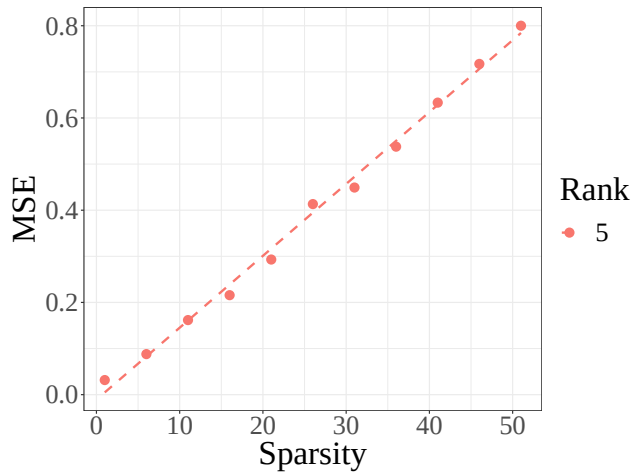
$$\frac{1}{n} \|\mathfrak{X}(\mathbf{B} + \mathbf{\Gamma}) - \boldsymbol{f}\|_2^2 \lesssim \sigma^2.$$

- The "machinery" required for additive matrix decomposition and label contamination imply as "corollary" the optimal rates for the particular cases of sparse linear regression and low-rank

Results with additive matrix decomposition and label contamination



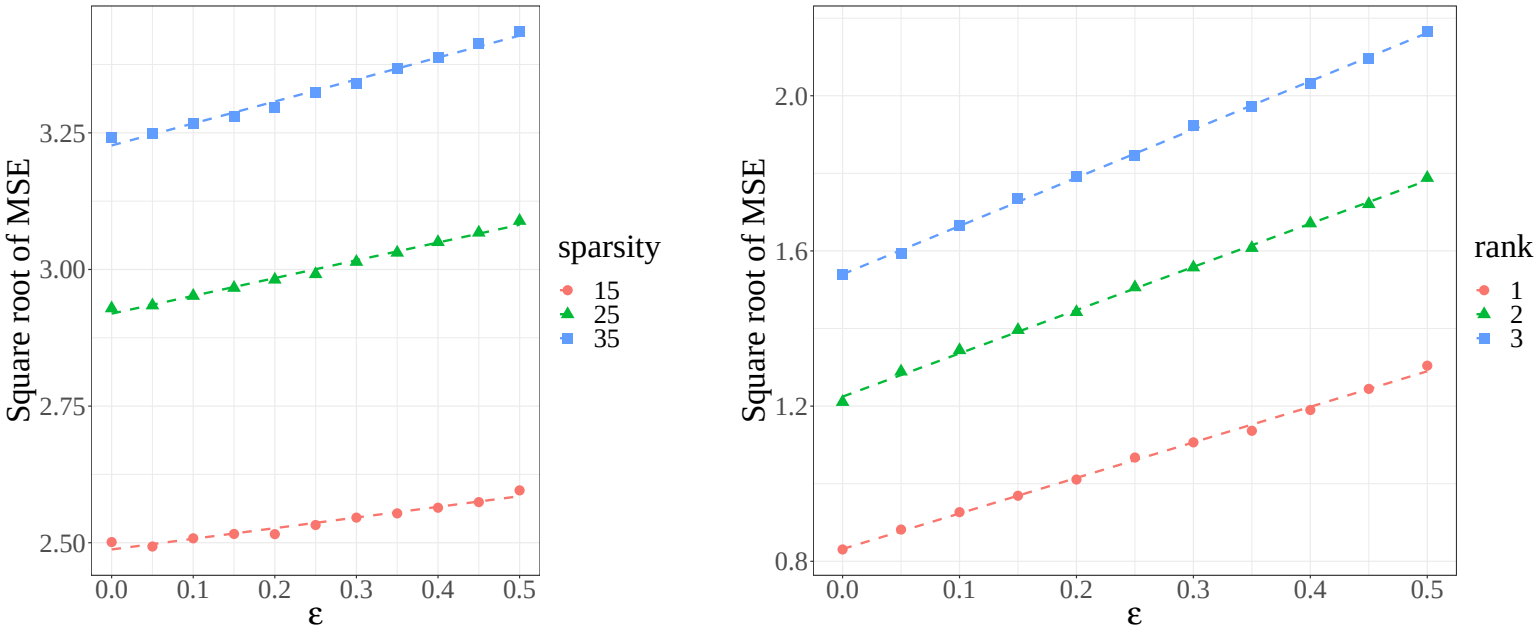
(a) Fixed sparsity.



(b) Fixed rank.

Figure: Trace regression with additive matrix decomposition **with no label contamination**: plot of MSE.

Results for sparse or low-rank and label contamination



(a) Fixed sparsity.

(b) Fixed rank.

Figure: Sparse or low-rank linear regression **with label contamination**: plot of $\sqrt{\text{MSE}}$.

σ-adaptive results (with no matrix decomposition)

Theorem ($q = 1$)

Assume that the covariate is 1*-subgaussian (for simplicity) and the noise is* σ*-subgaussian. Assume that* $\epsilon \leq c < 0.5$*. Assume* $\mathbf{B}^* \in \mathbb{R}^{d_1 \times d_2}$ *is s-sparse. Take tuning*

$$\lambda \asymp \sqrt{\frac{\log p}{n}}, \quad \tau \asymp \frac{1}{\sqrt{n}}.$$

Take $\mathcal{R} := \|\cdot\|_1$*. For any failure probability* $\delta \in (0, 1)$ *assume that*

$$n \geq Cs \log p,$$
$$\delta \geq \exp\left(-c_0 \frac{n}{\sigma^2}\right).$$

σ-adaptive results (with no matrix decomposition)

Theorem ($q = 1$)

Then with probability at least $1 - \delta$*, the square root of MSE is*

$$\sqrt{s \log p/n} + \sqrt{\log(1/\delta)/n} + \epsilon \log(1/\epsilon).$$

OBSERVATIONS:

- Taking $\mathcal{R} := \|\cdot\|_\sharp$, the Slope norm in $\mathbb{R}^p$, we can obtain the optimal rate $\sqrt{s \log(p/s)/n}$.
- Analogous bound when $\text{rank}(\mathbf{B}^*) \leq r$ and $\mathcal{R} := \|\cdot\|_N$.
- Rate is optimal up to log factor $\log(1/\epsilon)$.
- Optimal rate in δ, δ-uniformity and δ-adaptivity.
- **Miss-specification**: it holds as long as there is $[\mathbf{B}, \mathbf{\Gamma}]$ such that

$$\frac{1}{n}\|\mathfrak{X}(\mathbf{B}) - \boldsymbol{f}\|_2^2 \lesssim MSE^2(n, d_{\text{eff}}, \epsilon, \delta).$$

- "Machinery" and proofs: adaptations of additive matrix decomposition and label contamination.

Comments on the proof

- The traditional proof for LASSO does not work.
- New design properties to handle, jointly, additive decomposition, label contamination and feature-dependent noise.
  - PP -> Product Process Inequality
  - IP -> Chevet's Inequality
  - MP -> Multiplier Process Inequality
  - Combination of them to handle sharp dependence on regularization norms for low-rank, sparse and corruption structures.
- NEW: application of Product Process Inequality in low-rank + sparse estimation.

Future work?

"Possible" generalizations of this "theory" with **contaminated scalar labels**:

- The approach is using convex penalization.
- Several recent work has studied the optimization/statistical landscape of gradient descent methods using Burer-Monteiro *matrix factorization*.
- Benefit: under suitable initialization, the computation is faster and the MSE seems to improve in constants when compared to convex relaxations.
- "General idea": *non-convexity* but with "hidden convexity".
- Uses the incoherence condition.
- Is there a corresponding analysis for robust trace regression with additive decomposition?

- Inference?
- Outlier-robust *Generalized Linear Models* (with Robert Basset):
  - Logistic-regression and general exponential families. This may include more complicated models based on this family of distributions ?
  - single-index models ?
- Outlier-robust non-parametric least-squares with *Reproducing kernel Hilbert spaces* with decomposable regularizers.
- Online trace regression ?
- Tensor regression?
- Times Series?

THANK YOU!