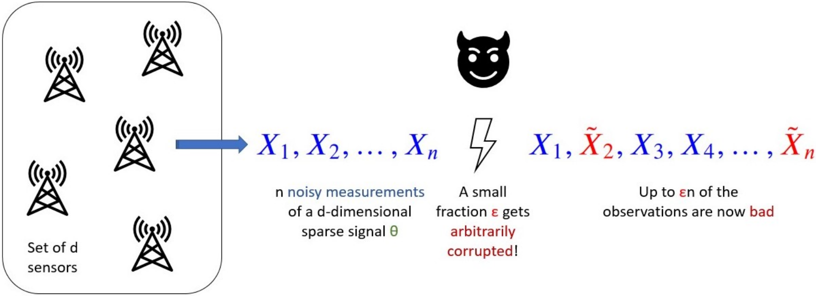
Robust Testing in High-Dimensional Sparse Models

Anand Jerry George
EPFL

Clément Canonne
The University of Sydney

NeurIPS, 2022

Overview of the problem



Is strength (norm) of θ large enough?

Robust Sparse Gaussian Mean Testing

We consider the following model:

$$X_i = \theta + Z_i \quad \text{for } 1 \leq i \leq n,$$

where

Robust Sparse Gaussian Mean Testing

We consider the following model:

$$X_i = \theta + Z_i \quad \text{for } 1 \leq i \leq n,$$

where

- $\theta \in \mathbb{R}^d$ is s-sparse, i.e. $\|\theta\|_0 \leq s$,
- $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$.

Robust Sparse Gaussian Mean Testing

We consider the following model:

$$X_i = \theta + Z_i \quad \text{for } 1 \leq i \leq n,$$

where

- $\theta \in \mathbb{R}^d$ is s-sparse, i.e. $\|\theta\|_0 \leq s$,
- $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$.

We want to find the minimum n required to distinguish between the hypotheses:

$$\mathcal{H}_0 : \|\theta\|_2 = 0$$
$$\mathcal{H}_1 : \|\theta\|_2 \geq \gamma$$

from the observations $(X_1, X_2, \tilde{X}_3, X_4, \cdots, \tilde{X}_{n-2}, X_{n-1}, X_n)$.
Here, $(\tilde{X}_i)$ denote an ε-fraction of X_i's that are arbitrarily corrupted– known as the ε-*corruption* model.

Robust Testing in Sparse Linear Regression Model

Another well studied model:

$$y_i = \langle X_i, \theta \rangle + z_i \quad \text{for } 1 \leq i \leq n,$$

where

Robust Testing in Sparse Linear Regression Model

Another well studied model:

$$y_i = \langle X_i, \theta \rangle + z_i \quad \text{for } 1 \leq i \leq n,$$

where

- $\theta \in \mathbb{R}^d$ is s-sparse,
- $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$,
- $z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, independent of X_i's.

Robust Testing in Sparse Linear Regression Model

Another well studied model:

$$y_i = \langle X_i, \theta \rangle + z_i \quad \text{for } 1 \leq i \leq n,$$

where

- $\theta \in \mathbb{R}^d$ is s-sparse,
- $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$,
- $z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, independent of X_i's.

Again, we want to distinguish between the hypotheses:

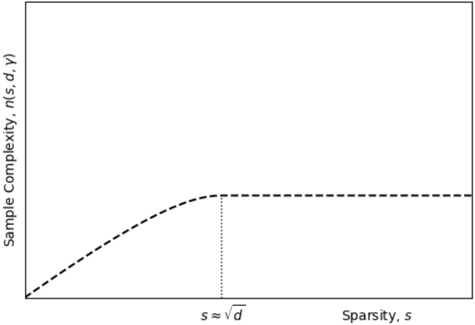
$$\mathcal{H}_0 : \|\theta\|_2 = 0$$
$$\mathcal{H}_1 : \|\theta\|_2 \geq \gamma$$

from the observations
$\Big((X_1, y_1), (X_2, y_2), (\tilde{X}_3, \tilde{y}_3), (X_4, y_4), \cdots, (\tilde{X}_{n-2}, \tilde{y}_{n-2}), (X_{n-1}, y_{n-1}), (X_n, X_n)\Big).$

Sample Complexity in Non-Robust Setting

It is known [Collier-Comminges-Tsybakov 17, Carpentier et. al. 19] that, in the *non-robust* setting, these problems have sample complexity

$$n(s,d) = \begin{cases} \Theta\left(s \log\left(1 + \frac{d}{s^2}\right)\right) & \text{if } s < \sqrt{d} \\ \Theta\left(\sqrt{d}\right) & \text{if } s \geq \sqrt{d}. \end{cases}$$



Both the problems exhibit a phase transition at $s \approx \sqrt{d}$.

Our Results

Theorem 1 (Robust sparse Gaussian mean testing)

Sample complexity of robust sparse Gaussian mean testing under ε-corruption model is

$$\Omega\left(s \log \frac{ed}{s}\right).$$

Our Results

Theorem 1 (Robust sparse Gaussian mean testing)

Sample complexity of robust sparse Gaussian mean testing under ε-corruption model is

$$\Omega\left(s \log \frac{ed}{s}\right).$$

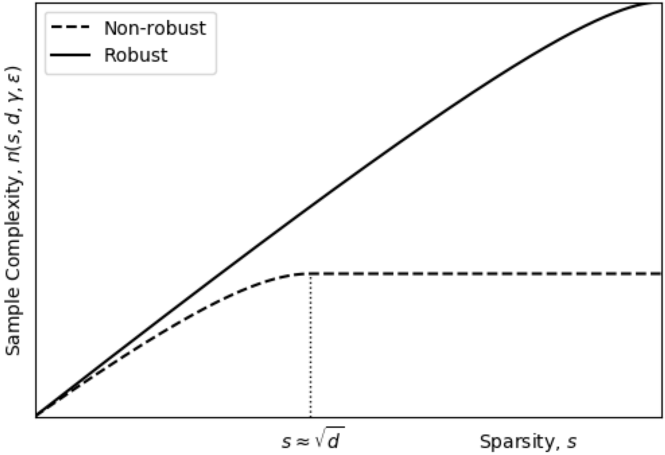
Theorem 2 (Robust testing in sparse linear regression model)

Sample complexity of robust testing in sparse linear regression under ε-corruption model is

$$\Omega\left(\min\left(s \log d, \frac{1}{\gamma^4}\right)\right).$$

These lower bounds are tight and are achieved by already known estimation algorithms.

Our Results



We observe that the phase transition disappear and the testing become much harder in the dense regime.

When θ is s-sparse in l_q norm

We also present the sample complexity of robust sparse Gaussian mean testing when θ is s-sparse in l_q norm instead of l_0 norm, where $q \in (0,2)$.

Theorem 3 (Robust sparse (l_q) Gaussian mean testing)

For $q \in (0,2)$, the sample complexity of robust sparse Gaussian mean testing, where θ is s-sparse in l_q norm, is

$$\Theta\left(m \log \frac{ed}{m}\right),$$

where $m = \max\{u \in [d] : \gamma^2 u^{\frac{2}{q}-1} \leq s^2\}$ is called the effective sparsity.

Conclusions and Future work

Conclusions:

- Ensuring robustness in these testing problems come at a cost, which is in contrast to common estimation problems.
- These problems don't exhibit phase transition anymore!

Conclusions and Future work

Conclusions:

- Ensuring robustness in these testing problems come at a cost, which is in contrast to common estimation problems.
- These problems don't exhibit phase transition anymore!

Future work:

- How to make the sample complexity tight w.r.t. γ and ε?
- What is the sample complexity when the covariance of the noise is not identity?

Conclusions and Future work

Conclusions:

- Ensuring robustness in these testing problems come at a cost, which is in contrast to common estimation problems.
- These problems don't exhibit phase transition anymore!

Future work:

- How to make the sample complexity tight w.r.t. γ and ε?
- What is the sample complexity when the covariance of the noise is not identity?

Thank you!

https://arxiv.org/abs/2205.07488

anand.george@epfl.ch