Robust Statistics: From Information Theory to Algorithms

Ilias Diakonikolas (USC)

ISIT 2019 Tutorial

July 2019

Can we develop learning algorithms that are ***robust*** to a ***constant*** fraction of ***corruptions*** in the data?

Outline Of This Tutorial

Part I: Introduction

Part II: High-Dimensional Robust Mean and Covariance Estimation

Part III: Extensions: Beyond Robust Statistics

Part IV: Computational Limits and Future Directions

PART I: INTRODUCTION

MOTIVATION

- **Model Misspecification/Robust Statistics**: Any model only approximately valid. Need *stable* estimators [Fisher 1920, Huber 1960s, Tukey 1960s]

- **Outlier Removal**: Natural outliers in real datasets (e.g., biology). Hard to detect in several cases [Rosenberg *et al.*, Science'02; Li *et al.*, Science'08; Paschou *et al*., Journal of Medical Genetics'10]

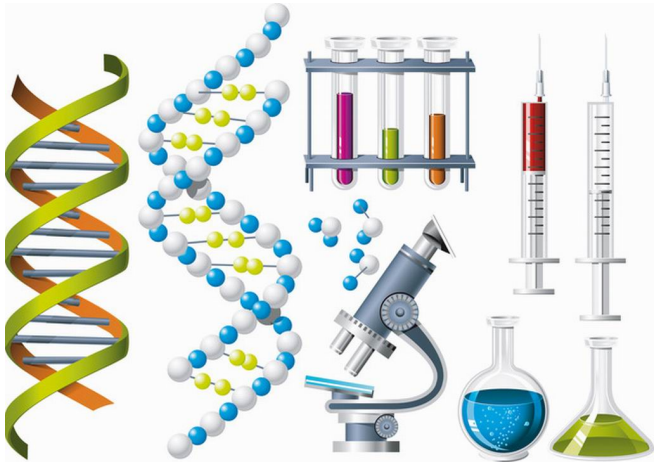
- **Reliable/Adversarial/Secure ML**: Data poisoning attacks (e.g., crowdsourcing) [Biggio *et al.* ICML'12, …]

Detecting Outliers in Real Datasets

- High-dimensional datasets tend to be inherently noisy.

Biological Datasets: POPRES project, HGDP datasets

[November *et al.*, Nature'08];
[Rosenberg *et al.*, Science'02];
[Li *et al.*, Science'08];
[Paschou *et al.*, Medical Genetics'10]

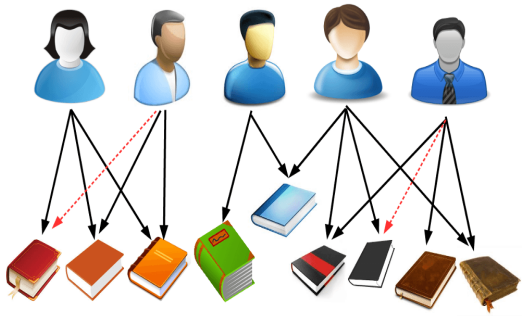


- Outliers: either interesting or can contaminate statistical analysis

DATA POISONING

Fake Reviews [Mayzlin et al. ‘14]

Recommender Systems:



[Li et al. ‘16]

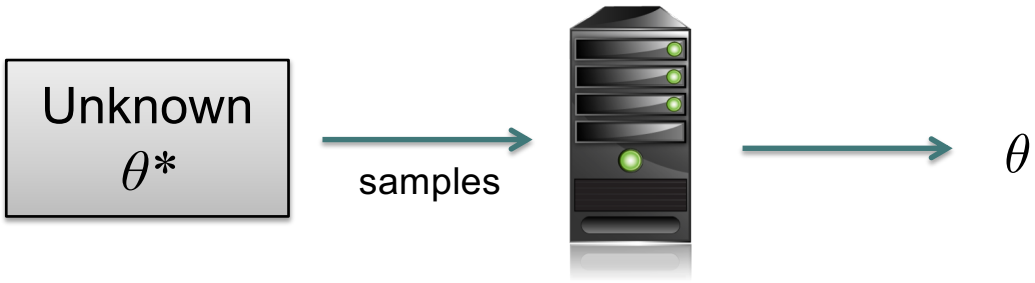
Crowdsourcing:

[Wang et al. ‘14]

Malware/spam:

[Nelson et al. ‘08]

The Statistical Learning Problem



- *Input*: sample generated by a **probabilistic model** with unknown θ^*
- *Goal*: estimate parameters θ so that $\theta \approx \theta^*$

Question 1: Is there an *efficient* learning algorithm?

Main performance criteria:

- Sample size
- Running time
- **Robustness**

Question 2: Are there *tradeoffs* between these criteria?

(Outlier-) Robustness in a Generative Model

Contamination Model:

Let $\mathcal{F}$ be a family of probabilistic models.
We say that a set of N samples is ϵ-corrupted from $\mathcal{F}$ if it is generated as follows:

- N samples are drawn from an unknown $F \in \mathcal{F}$
- An omniscient adversary inspects these samples and changes arbitrarily an ϵ-fraction of them.

cf. Huber's contamination model [1964]

Models of Robustness

- Oblivious/Adaptive Adversary
- Adversary can: add corrupted samples, subtract uncorrupted samples or both.
- Six Distinct Models:

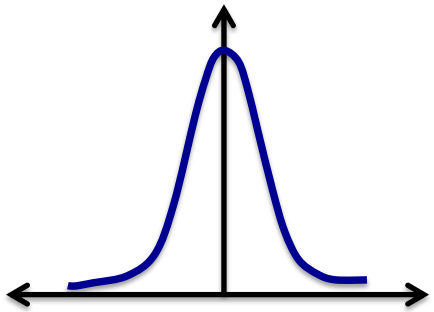
	Oblivious	**Adaptive**
Additive Errors	Huber's Contamination Model $P = (1-\epsilon)G + \epsilon B$	Additive Contamination ("Data Poisoning")
Subtractive Errors	$P = (1-\epsilon)G - \epsilon L$	Subtractive Contamination
Additive and Subtractive Errors	Hampel's Contamination $\mathrm{d_{TV}}(P, G) \leq \epsilon$ $P = G - \epsilon L + \epsilon B$	Strong Contamination ("Nasty Learning Model")

EXAMPLE: PARAMETER ESTIMATION

Given samples from an unknown distribution:

e.g., a 1-D Gaussian

$$\mathcal{N}(\mu, \sigma^2)$$



how do we accurately estimate its parameters?

empirical mean:

$$\frac{1}{N}\sum_{i=1}^{N} X_i \rightarrow \mu$$

empirical variance:

$$\frac{1}{N}\sum_{i=1}^{N} (X_i - \overline{X})^2 \rightarrow \sigma^2$$

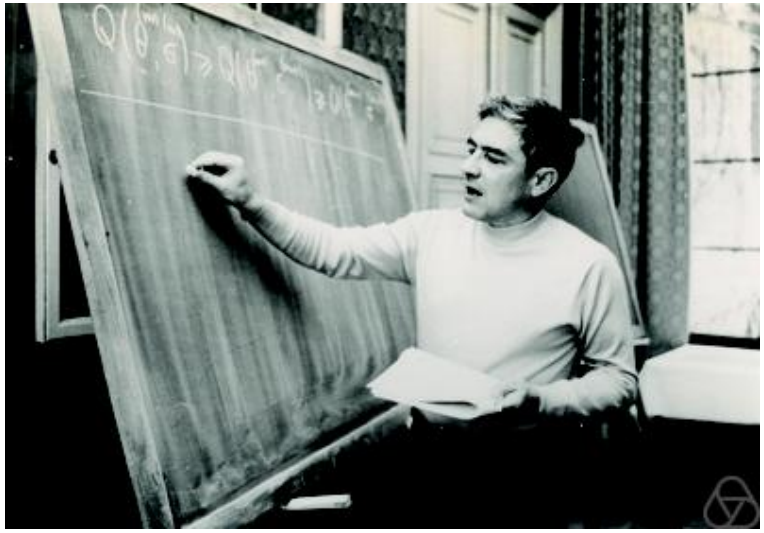
R. A. Fisher

The **maximum likelihood estimator** is asymptotically efficient (1910-1920)

J. W. Tukey

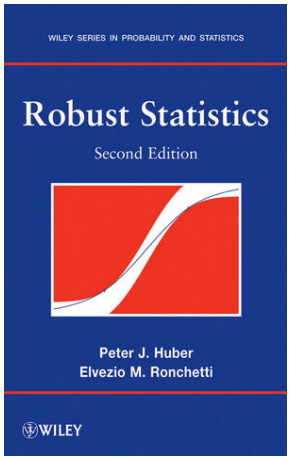
What about **errors** in the model itself? (1960)

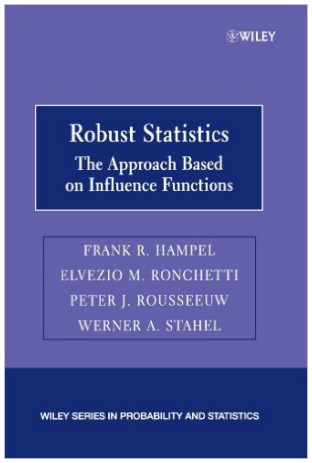
Peter J. Huber

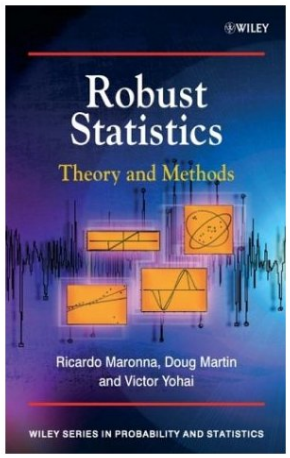


“Robust Estimation of a Location Parameter”
Annals of Mathematical Statistics, 1964.

ROBUST STATISTICS







What estimators behave well in a **neighborhood** around the model?

Robust Estimation: One Dimension

Given **corrupted** samples from a *one-dimensional* Gaussian, can we accurately estimate its parameters?

- A single corrupted sample can arbitrarily corrupt the empirical mean and variance.
- But the **median** and **interquartile range** work.

Fact [Folklore]: Given a set S of N ϵ-corrupted samples from a one-dimensional Gaussian

$$\mathcal{N}(\mu, \sigma^2)$$

with high constant probability we have that:

$$|\widehat{\mu} - \mu| \leq O\left(\epsilon + \sqrt{1/N}\right) \cdot \sigma$$

where $\widehat{\mu} = \mathrm{median}(S)$.

What about robust estimation in high-dimensions?

High-Dimensional Gaussian Robust Mean Estimation

Robust Mean Estimation: Given an ϵ - corrupted set of samples from an **unknown mean**, identity covariance Gaussian $\mathcal{N}(\mu, I)$ in d dimensions, recover $\widehat{\mu}$ with

$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon) \ .$$

Remark: Optimal rate of convergence with N samples is

$$O(\epsilon) + O\left(\sqrt{d/N}\right)$$

[Tukey'75, Donoho'82]

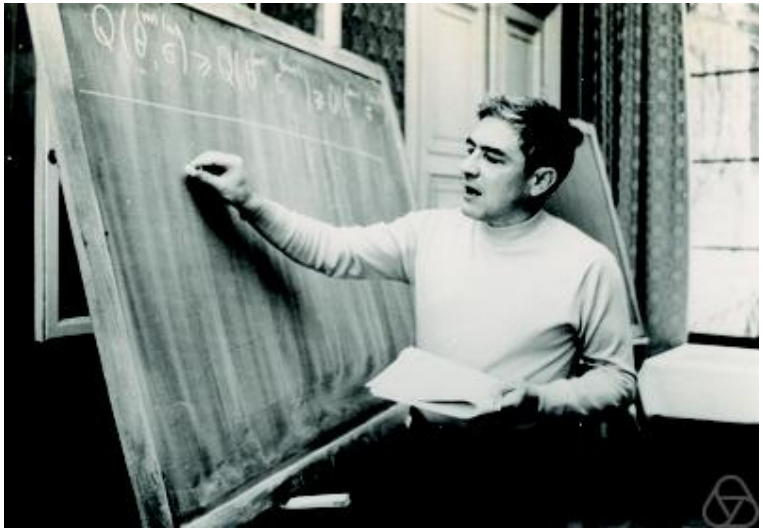
Previous Approaches: Robust Mean Estimation

Unknown Mean	Error Guarantee	Running Time
Pruning	$\Theta(\epsilon\sqrt{d})$ ✗	$O(dN)$ ✓
Coordinate-wise Median	$\Theta(\epsilon\sqrt{d})$ ✗	$O(dN)$ ✓
Geometric Median	$\Theta(\epsilon\sqrt{d})$ ✗	$\mathrm{poly}(d, N)$ ✓
Tukey Median	$\Theta(\epsilon)$ ✓	NP-Hard ✗
Tournament	$\Theta(\epsilon)$ ✓	$N^{O(d)}$ ✗

All known estimators are either **hard to compute** or
can tolerate a **negligible fraction of corruptions**.

Is robust estimation algorithmically possible in high-dimensions?

Peter J. Huber, 1975



“[…] Only simple algorithms (i.e., with **a low degree of computational complexity**) will survive the onslaught of huge data sets. This runs counter to recent developments in computational robust statistics**. It appears to me that none of the above problems will be amenable to a treatment through theorems and proofs.** They will have to be attacked by heuristics and judgment, and by alternative “what if” analyses.[…]”

Robust Statistical Procedures, 1996*, Second Edition.*

Robust estimation in high-dimensions is algorithmically possible!

- First computationally efficient robust estimators that can tolerate a ***constant*** fraction of corruptions.
- General methodology to detect outliers in high dimensions.

Meta-Theorem (Informal): Can obtain *dimension-independent* error guarantees, as long as good data has nice concentration.

First Algorithmic Progress In Unsupervised Setting

[D-Kamath-Kane-Li-Moitra-Stewart, FOCS'16]

Can tolerate a ***constant*** fraction of corruptions:

- Mean and Covariance Estimation
- Mixtures of Spherical Gaussians, Mixtures of Balanced Product Distributions

[Lai-Rao-Vempala, FOCS'16]

Can tolerate a ***mild sub-constant*** (*inverse logarithmic*) fraction of corruptions:

- Mean and Covariance Estimation
- Independent Component Analysis, SVD

Basic Result: Robust Mean Estimation

Theorem: There are polynomial time algorithms with the following behavior:

Given $\epsilon > 0$ and a set of N ϵ-corrupted samples from a d-dimensional Gaussian $\mathcal{N}(\mu, I)$, they output $\widehat{\mu} \in \mathbb{R}^d$ that with high probability satisfies:

- **[LRV'16]:**

$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon\sqrt{\log d}) + \tilde{O}(\sqrt{d/N})$$

  in *additive* contamination model.

- **[DKKLMS'16]:**

$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)}) + O(\sqrt{d/N})$$

  in *strong* contamination model.

Part II: Robust Mean and Covariance Estimation

OUTLINE

Part II: High-Dimensional Robust Mean and Covariance Estimation

- **Statements of Results**
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Experiments
- Extensions

Robust Mean Estimation: Gaussian Case

Problem: Given data $x_1, \ldots, x_N \in \mathbb{R}^d$, of which $(1-\epsilon)N$ come from some distribution D, estimate mean μ of D.

Theorem: Let $\epsilon < 1/2$. If $D = \mathcal{N}(\mu, I)$, there is an efficient algorithm that outputs an estimate $\widehat{\mu}$ that with high probability satisfies

$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon) + O(\sqrt{d/N})$$

in **additive contamination** model.

Error Guarantee Independent of d !

[D-Kamath-Kane-Li-Moitra-Stewart, SODA'18]

ROBUST MEAN ESTIMATION: *SUB*-GAUSSIAN CASE

Problem: Given data $x_1, \ldots, x_N \in \mathbb{R}^d$, of which $(1-\epsilon)N$ come from some distribution D, estimate mean μ of D.

Theorem: Let $\epsilon < 1/2$. If D is isotropic and *sub-Gaussian*, there is an efficient algorithm that outputs an estimate $\widehat{\mu}$ such that with high probability we have:

$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)}) + O(\sqrt{d/N}) \ .$$

Information-theoretically optimal error, even in one-dimension.

[D-Kamath-Kane-Li-Moitra-Stewart, FOCS'16, ICML'17]

Robust Mean Estimation: General Case

Problem: Given data $x_1, \ldots, x_N \in \mathbb{R}^d$, of which $(1-\epsilon)N$ come from some distribution D, estimate mean μ of D.

Theorem: Let $\epsilon < 1/2$. If D has covariance $\Sigma \preceq \sigma^2 \cdot I$, there is an efficient algorithm that outputs an estimate $\widehat{\mu}$ such that with high probability we have

$$\|\widehat{\mu} - \mu\|_2 = O(\sigma\sqrt{\epsilon}) + O(\sqrt{d/N}) .$$

- Sample-optimal, even without corruptions.
- Information-theoretically optimal error, even in one-dimension.

[D-Kamath-Kane-Li-Moitra-Stewart, ICML'17; Steinhardt, Charikar, Valiant, ITCS'18]

Robust Covariance Estimation

Problem: Given data $x_1, \ldots, x_N \in \mathbb{R}^d$, of which $(1-\epsilon)N$ come from some distribution D, estimate covariance Σ of D.

Theorem: Let $\epsilon < 1/2$. If $N = \Omega(d^2/\epsilon^2)$,then can efficiently recover $\widehat{\Sigma}$ such that

$$\|\Sigma^{-1/2}(\widehat{\Sigma} - \Sigma)\Sigma^{-1/2}\|_F = f(\epsilon) \ ,$$

where f depends on the concentration of D.

OUTLINE

Part II: High-Dimensional Robust Mean and Covariance Estimation

- Statements of Results
- **Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal**
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Experiments
- Extensions

High-Dimensional Gaussian Mean Estimation (I)

Fact: Let $X_1, \ldots, X_N$ be IID samples from $\mathcal{N}(\mu, I)$. The empirical estimator $\widehat{\mu}$ satisfies $\|\widehat{\mu} - \mu\|_2 \leq \delta$ with probability at least 9/10 for $N = \Omega(d/\delta^2)$.
Moreover, *any* estimator with this guarantee requires $\Omega(d/\delta^2)$ samples.

Proof:

By definition, $\widehat{\mu} = (1/N) \sum_{i=1}^{N} X_i$, where $X_i \sim \mathcal{N}(\mu, I)$.

Then,

$$\widehat{\mu} \sim \mathcal{N}\left(\mu, (1/N) I\right).$$

We have

$$\mathbf{E}[\|\widehat{\mu} - \mu\|_2^2] = \sum_{j=1}^{d} \mathbf{E}\left[(\widehat{\mu}_j - \mu_j)^2\right] = \sum_{j=1}^{d} \mathbf{Var}\left[\widehat{\mu}_j\right] = d/N$$

Therefore,

$$\mathbf{E}[\|\widehat{\mu} - \mu\|_2] \leq \mathbf{E}[\|\widehat{\mu} - \mu\|_2^2]^{1/2} = \sqrt{\frac{d}{N}}$$

and Markov's inequality gives the upper bound.

High-Dimensional Gaussian Mean Estimation (II)

Fact: Let $X_1, \ldots, X_N$ be IID samples from $\mathcal{N}(\mu, I)$. The empirical estimator $\widehat{\mu}$ satisfies $\|\widehat{\mu} - \mu\|_2 \leq \delta$ with probability at least 9/10 for $N = \Omega(d/\delta^2)$.
Moreover, *any* estimator with this guarantee requires $\Omega(d/\delta^2)$ samples.

Proof:

For the lower bound, consider the following family of distributions:

$$\{\mathcal{N}(\mu, I)\}_{\mu \in \mathcal{M}}$$

where

$$\mathcal{M} = \left\{\mu \in \mathbb{R}^d : \mu_j = -\delta/\sqrt{d} \text{ or } \mu_j = \delta/\sqrt{d}, j \in [d]\right\} .$$

Apply Assouad's lemma to show that learning an unknown distribution in this family within error $\delta/2$ requires $\Omega(d/\delta^2)$ samples.

■

Information-Theoretic Limits On Robust Estimation

Proposition: Any robust mean estimator for $\mathcal{N}(\mu, 1)$ has error $\Omega(\epsilon)$, even in Huber's model.

Claim: Let P_1, P_2 be such that $\mathrm{d_{TV}}(P_1, P_2) = \epsilon/(1-\epsilon)$. There exist noise distributions B_1, B_2 such that $(1-\epsilon)P_1 + \epsilon B_1 = (1-\epsilon)P_2 + \epsilon B_2$.

- Use $\mathrm{d_{TV}}\left(\mathcal{N}(\mu_1, 1), \mathcal{N}(\mu_2, 1)\right) \leq |\mu_1 - \mu_2|/2$
- Under different assumptions on good data, we obtain different functions of ϵ

Sample Efficient Robust Mean Estimation (I)

Proposition: There is an algorithm that uses $N = O(d/\epsilon^2)$ ϵ- corrupted samples from $\mathcal{N}(\mu, I)$ and outputs $\widetilde{\mu} \in \mathbb{R}^d$ that with probability at least $9/10$ satisfies $\|\widetilde{\mu} - \mu\|_2 = O(\epsilon)$.

Main Idea: To robustly learn the mean of $\mathcal{N}(\mu, I)$, it suffices to learn the mean of *all* its 1-dimensional projections (cf. Tukey median).

Basic Fact: $\|\widetilde{\mu} - \mu\|_2 = \max_{v:\|v\|_2=1} |v \cdot \widetilde{\mu} - v \cdot \mu|$

Claim 1: Suppose we can estimate $v \cdot \mu$ for each $v \in \mathbb{R}^d, \|v\|_2 = 1$, i.e., find $\{\widehat{\mu}_v\}_v$ such that for all $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$ we have $|\widehat{\mu}_v - \mu \cdot v| \leq \delta$. Then, we can learn μ within error 2δ .

Proof:

Consider *infinite size* LP: Find $x \in \mathbb{R}^d$ such that *for all* $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$: $|\widehat{\mu}_v - v \cdot x| \leq \delta$.

Let x^* be any feasible solution. Then

$$\|x^* - \mu\|_2 = \max_{v:\|v\|_2=1} |v \cdot x^* - v \cdot \mu| \leq \max_{v:\|v\|_2=1} |v \cdot x^* - \widehat{\mu}_v| + \max_{v:\|v\|_2=1} |v \cdot \mu - \widehat{\mu}_v| \leq 2\delta \ . \qquad \blacksquare$$

Sample Efficient Robust Mean Estimation (II)

Main Idea: To robustly learn the mean of $\mathcal{N}(\mu, I)$, it suffices to learn the mean of "all" its 1-dimensional projections.

Claim 2: Suffices to consider a γ-net C over all directions, where γ is a small positive constant.

Proof:

This gives the following *finite* LP:

Find $x \in \mathbb{R}^d$ such that for all $v \in C$, we have $|\widehat{\mu}_v - v \cdot x| \leq \delta$.

Let x^* be any feasible solution. Let $u \in C$ such that $\|u - \frac{\mu - x^*}{\|\mu - x^*\|_2}\|_2 \leq \gamma$.

Then

$$\|x^* - \mu\|_2 = \left| \left(\left(\frac{\mu - x^*}{\|\mu - x^*\|_2} - u \right) + u \right) \cdot (x^* - \mu) \right| \leq \gamma \|x^* - \mu\|_2 + 2\delta$$

or

$$\|x^* - \mu\|_2 \leq \frac{2\delta}{1 - \gamma} .$$

■

Sample Efficient Robust Mean Estimation (III)

Main Idea: To robustly learn the mean of $\mathcal{N}(\mu, I)$, it suffices to learn the mean of "all" its 1-dimensional projections.

So, for $\gamma = 1/2$, any feasible solution to the LP has $\|x^* - \mu\|_2 \leq 4\delta$.

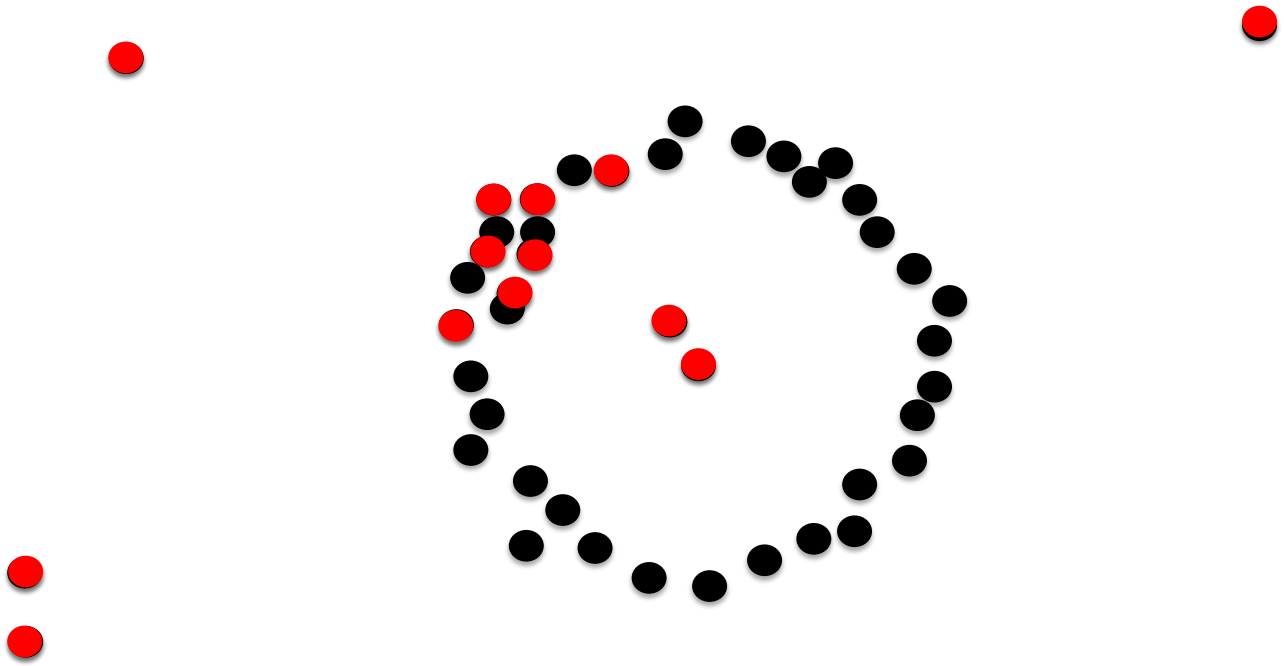
Sample Complexity: Note that the empirical median satisfies $\delta = O(\epsilon)$ with probability at least $1 - \tau$ after $O((1/\epsilon^2)\log(1/\tau))$ samples.

We need union bound over all $v \in C$. Since $|C| = (1/\gamma)^{O(d)} = 2^{O(d)}$, for $\tau = 1/(10|C|)$ our algorithm works with probability at least $9/10$.
Thus, sample complexity will be $N = O(d/\epsilon^2)$.

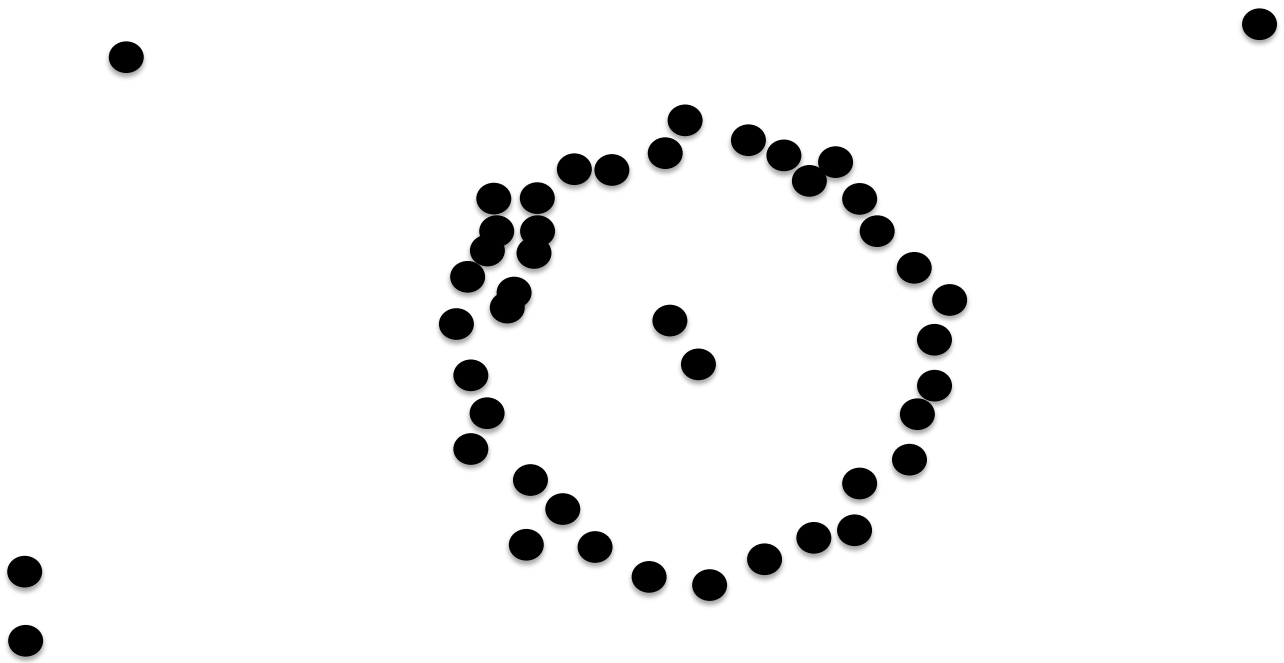
Runtime: $\mathrm{poly}(N, 2^d)$.

■

OUTLIER DETECTION ?



Naïve Outlier Removal (Naïve Pruning)



Gaussian Annulus Theorem: $\mathbf{Pr}_{X\sim\mathcal{N}(\mu,I)}\left[\left|\|X\|_2^2 - d\right| > t\right] \leq 2e^{-\Omega\left(\min\left\{\frac{t^2}{d}, t\right\}\right)}$

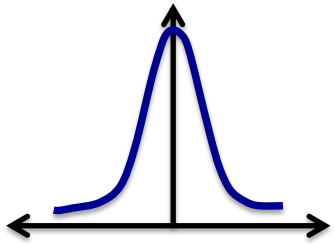
On The Effect of Corruptions

Question: What is the effect of additive and subtractive corruptions?

Let's study the simplest possible example of $\mathcal{N}(\mu, 1)$.

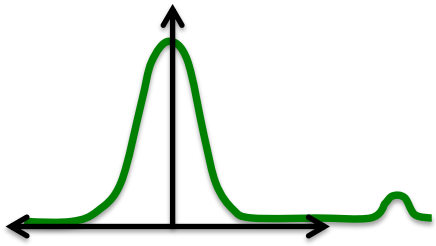
Subtractive errors at rate ϵ can:

- Move the mean by at most $O(\epsilon\sqrt{\log(1/\epsilon)})$
- Increase the variance by $O(\epsilon)$ and decrease it by at most $O(\epsilon \log(1/\epsilon))$



Additive errors at rate ϵ can:

- Move the mean arbitrarily
- Increase the variance arbitrarily and decrease it by at most $O(\epsilon)$



OUTLINE

Part II: High-Dimensional Robust Mean and Covariance Estimation

- Statements of Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- **Overview of Algorithmic Approaches**
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Experiments
- Extensions

High-Level Goal: Reduce “structured” high-dimensional problem to a collection of “low-dimensional” problems.

Three Approaches: Overview and Comparison

Three Algorithmic Approaches:

- Recursive Dimension-Halving [LRV'16]
- Iterative Filtering [DKKLMS'16]
- Soft Outlier Removal [DKKLMS'16]

Commonalities:

- Rely on Spectrum of Empirical Covariance to Robustly Estimate the Mean
- Certificate of Robustness for the Empirical Estimator

Exploiting the Certificate:

- Recursive Dimension-Halving: Find "good" large subspace.
- Iterative Filtering: Check condition on entire space. If violated, filter outliers.
- Soft Outlier Removal: Convex optimization via approximate separation oracle.

OUTLINE

Part II: High-Dimensional Robust Mean and Covariance Estimation

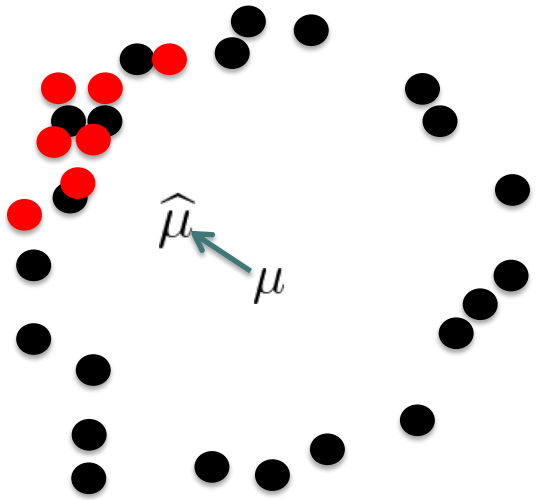
- Statements of Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- **Certificate of Robustness**
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Experiments
- Extensions

Certificate Of Robustness For Empirical Estimator

Idea #1 [DKKLMS'16, LRV'16]: If the empirical covariance is "close to what it should be", then the empirical mean works.

Certificate Of Robustness For Empirical Estimator

Detect when the empirical estimator *may* be compromised



$$\widehat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i$$

= uncorrupted

= corrupted

There is *no* direction of large variance

Key Lemma: Let $X_1, X_2, \ldots, X_N$ be an ϵ-corrupted set of samples from $\mathcal{N}(\mu, I)$ and $N = \Omega(d/\epsilon^2)$, then for

$$\text{(1)}\ \ \widehat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \text{(2)}\ \ \widehat{\Sigma} \triangleq \frac{1}{N}\sum_{i=1}^{N}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^T$$

with high probability we have:

- **[LRV'16]:**

$$\|\widehat{\Sigma}\|_2 \leq 1 + O(\epsilon) \quad \rightarrow \quad \|\widehat{\mu} - \mu\|_2 \leq O(\epsilon)$$

  in ***additive*** contamination model

- **[DKKLMS'16]:**

$$\|\widehat{\Sigma}\|_2 \leq 1 + O(\epsilon \log(1/\epsilon)) \quad \rightarrow \quad \|\widehat{\mu} - \mu\|_2 \leq O(\epsilon\sqrt{\log(1/\epsilon)})$$

  in ***strong*** contamination model

Key Lemma: Let $X_1, X_2, \ldots, X_N$ be an ϵ-corrupted set of samples from $\mathcal{N}(\mu, I)$ and $N = \Omega(d/\epsilon^2)$, then for

(1) $\widehat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i$ (2) $\widehat{\Sigma} \triangleq \frac{1}{N}\sum_{i=1}^{N}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^T$

with high probability we have:

- **[LRV'16]:**

$$\|\widehat{\Sigma}\|_2 \leq 1 + \delta \quad \rightarrow \quad \|\widehat{\mu} - \mu\|_2 \leq O(\sqrt{\delta\epsilon} + \epsilon)$$

in ***additive*** contamination model

- **[DKKLMS'16]:**

$$\|\widehat{\Sigma}\|_2 \leq 1 + \delta \quad \rightarrow \quad \|\widehat{\mu} - \mu\|_2 \leq O(\sqrt{\delta\epsilon} + \epsilon\sqrt{\log(1/\epsilon)})$$

in ***strong*** contamination model

Idea #2 [DKKLMS'16]: Removing *any* small constant fraction of good points does not move the empirical mean and covariance by much.

Remarks on Key Lemma (Strong Corruptions)

- Statement holds for any isotropic distribution with sub-Gaussian tails.
- Essentially same argument goes through if covariance is *approximately* known.
- Argument extends for (approximately known) covariance and weaker concentration.

If $\mathcal{D}$ is isotropic with *sub-exponential* tails:

$$\|\widehat{\Sigma}\|_2 \leq 1+\delta \longrightarrow \|\widehat{\mu}-\mu\|_2 \leq O(\sqrt{\delta\epsilon}+\epsilon\log(1/\epsilon)) \ .$$

- If *only* assumption is that $\mathcal{D}$ has $\Sigma \preceq I$:

$$\|\widehat{\Sigma}\|_2 \leq 1+\delta \longrightarrow \|\widehat{\mu}-\mu\|_2 \leq O(\sqrt{\delta\epsilon}+\sqrt{\epsilon}) \ .$$

OUTLINE

Part II: High-Dimensional Robust Mean and Covariance Estimation

- Statements of Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- **Recursive Dimension Halving**
- Iterative Filtering
- Soft Outlier Removal
- Experiments
- Extensions

Idea #3 [LRV'16]: Additive corruptions can move the covariance in *some* directions, but *not in all* directions simultaneously.

RECURSIVE DIMENSION-HALVING [LRV'16]

Recursive Procedure:

Step #1: Find large subspace where "standard" estimator works.
Step #2: Recurse on complement.
(If dimension is small, use brute-force.)

Combine Results.

Can reduce dimension by factor of 2 in each recursive step.

FINDING A GOOD SUBSPACE (I)

"Good subspace ***G***" = one where the empirical mean works

By **Key Lemma**, sufficient condition is:

> Projection of empirical covariance on ***G*** has no large eigenvalues.

- Also want ***G*** to be "high-dimensional".

Question: How do we find such a subspace?

FINDING A GOOD SUBSPACE (II)

Good Subspace Lemma: Let $X_1, X_2, \ldots, X_N$ be an *additively* ϵ-corrupted set of $N = \Omega(d \log d/\epsilon^2)$ samples from $\mathcal{N}(\mu, I)$. ***After naïve pruning***, we have that

$$\lambda_{d/2}(\widehat{\Sigma}) \leq 1 + O(\epsilon)$$

Corollary: Let W be the span of the bottom $d/2$ eigenvalues of $\widehat{\Sigma}$. Then W is a good subspace.

RECURSIVE DIMENSION-HALVING ALGORITHM [LRV'16]

Algorithm works as follows:

- Remove gross outliers (e.g., naïve pruning).
- Let W, V be the span of bottom $d/2$ and upper $d/2$ eigenvalues of $\widehat{\Sigma}$ respectively .
- Use empirical mean on W.
- Recurse on V (If the dimension is one, use median).

Error Analysis:

$O(\log d)$ levels of the recursion → final error of $O(\epsilon\sqrt{\log d})$

OUTLINE

Part II: High-Dimensional Robust Mean and Covariance Estimation

- Statements of Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- **Iterative Filtering**
- Soft Outlier Removal
- Experiments
- Extensions

Idea #4 [DKKLMS'16]: Iteratively "remove outliers" in order to "fix" the empirical covariance.

ITERATIVE FILTERING [DKKLMS'16]

Iterative Two-Step Procedure:

Step #1: Find certificate of robustness of "standard" estimator

Step #2: If certificate is violated, detect and remove outliers

Iterate on "cleaner" dataset.

General recipe that works for fairly general settings.

Let's see how this works for robust mean estimation.

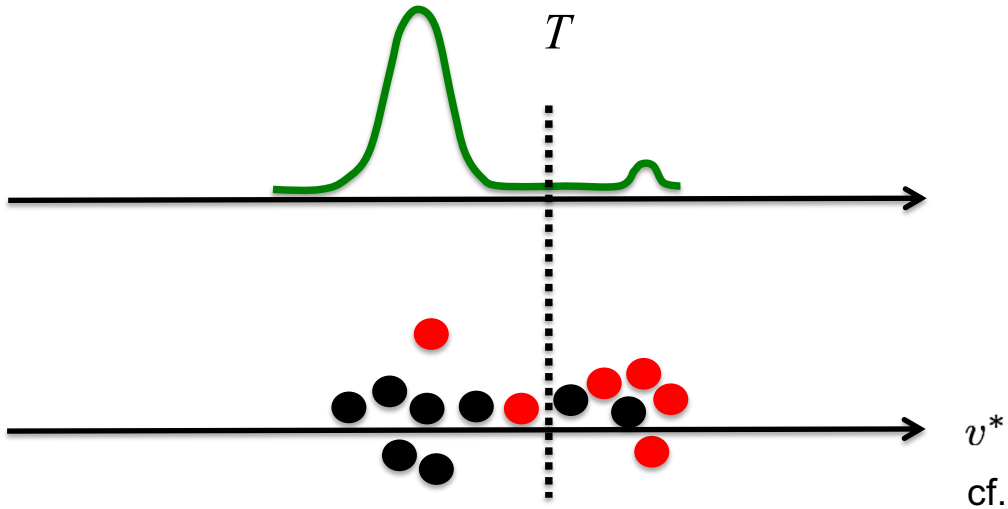
FILTERING SUBROUTINE

Either output empirical mean, or remove many outliers.

Filtering Approach: Suppose that:

$$\|\widehat{\Sigma}\|_2 \geq 1 + \Omega(\epsilon \log(1/\epsilon))$$

Let v^* be the direction of maximum variance.



cf. [Klivans-Long-Servedio'09, Lai-Rao-Vempala'16]

Filtering Subroutine

Either output empirical mean, or remove many outliers.

Filtering Approach: Suppose that:

$$\|\widehat{\Sigma}\|_2 \geq 1 + \Omega(\epsilon \log(1/\epsilon))$$

Let v^* be the direction of maximum variance.

- Project all the points on the direction of v^*.
- Find a threshold T such that

$$\mathbf{Pr}_{X \sim_U S}[|v^* \cdot X - \text{median}(\{v^* \cdot x, x \in S\})| > T + 1] \geq 8 \cdot e^{-T^2/2}.$$

- Throw away all points x such that

$$|v^* \cdot x - \text{median}(\{v^* \cdot x, x \in S\})| > T + 1$$

- Iterate on new dataset.

Filtering Subroutine: Analysis Sketch

Either output empirical mean, or remove many outliers.

Filtering Approach: Suppose that:

$$\|\widehat{\Sigma}\|_2 \geq 1 + \Omega(\epsilon \log(1/\epsilon))$$

Claim 1: In each iteration, we remove more corrupted than uncorrupted points.

After a number of iterations, we stop removing points.

Eventually the empirical mean works

FILTERING SUBROUTINE: PSEUDO-CODE

Input: ϵ-corrupted set S from $\mathcal{N}(\mu, I)$

Output: Set $S' \subseteq S$ that is ϵ'-corrupted, for some $\epsilon' < \epsilon$

OR robust estimate of the unknown mean μ

1. Let $\widehat{\mu}_S, \widehat{\Sigma}_S$ be the empirical mean and covariance of the set S.
2. **If** $\|\widehat{\Sigma}_S\|_2 \leq 1 + C\epsilon \log(1/\epsilon)$, for an appropriate constant $C > 0$:
   Output $\widehat{\mu}_S$
3. **Otherwise**, let (λ^*, v^*) be the top eigenvalue-eigenvector pair of $\widehat{\Sigma}_S$.
4. Find $T > 0$ such that

$$\mathbf{Pr}_{X \sim_U S}[|v^* \cdot X - \text{median}(\{v^* \cdot x, x \in S\})| > T + 1] \geq 8 \cdot e^{-T^2/2}.$$

5. **Return**

$$S' = \{x \in S : |v^* \cdot x - \text{median}(\{v^* \cdot x, x \in S\})| \leq T + 1\}.$$

Sketch of Correctness

Claim 2: Can always find a threshold satisfying the Condition of Step 4.

Proof Sketch:
By contradiction. Suppose that for all $T > 0$ we have

$$\mathbf{Pr}_{X \sim_U S}[|v^* \cdot X - \text{median}(\{v^* \cdot x, x \in S\})| > T + 1] < 8 \cdot e^{-T^2/2}.$$

Can use this to show that $\lambda^* = \|\widehat{\Sigma}_S\|_2$ is smaller than it was assumed to be.

Main Idea: Exploit concentration.

Summary: Robust Mean Estimation via Filtering

Certificate for Robustness:

"Spectral norm of empirical covariance is *close* to what it should be."

Exploiting the Certificate:

- Check if certificate is satisfied.
- If violated, find "subspace" where behavior of outliers different than behavior of inliers.
- Use it to detect and remove outliers.
- Iterate on "cleaner" dataset.

Remarks on Filtering Method(s)

- For known covariance sub-Gaussian case, filter relied on violation of concentration.

- This extends to weaker concentration, as long as covariance is (approximately) known.
- For example, for *sub-exponential* concentration, filter would be:

> Find $T > 0$ such that $\mathbf{Pr}_{X \sim_U S}[|v^* \cdot (X - \widehat{\mu})| > T] \geq 8 \cdot e^{-T}$.

- For *the bounded covariance* setting, need *randomized* filtering.

> Remove point x with probability proportional to $(v^* \cdot (x - \widehat{\mu}))^2$.

- Analogue of Claim 1: Remove more corrupted than good points *in expectation*.

Optimal Gaussian Robust Mean Estimation: *Additive* Errors

Theorem [DKKLMS, SODA'18] There is a polynomial time algorithm with the following behavior: Given $\epsilon > 0$ and $N = \mathrm{poly}(d/\epsilon)$ corrupted samples from an unknown mean, identity covariance Gaussian distribution on $\mathbb{R}^d$, the algorithm finds a hypothesis mean $\widehat{\mu}$ that satisfies

$$\|\mu - \widehat{\mu}\|_2 \leq \sqrt{\pi} \cdot \epsilon + o(\epsilon)$$

in ***additive*** contamination model.

- Robustness guarantee optimal up to $\sqrt{2}$ factor.
- For any univariate projection, mean robustly estimated by median.

Generalized Filtering: Additive Corruptions

- *Univariate* filtering based on tails not sufficient to remove the incurred $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$ error, even for additive errors.

- **Generalized Filtering Idea**: Filter using *top - k eigenvectors* of empirical covariance.

- **Key Observation**: Suppose that $\|\mu - \widehat{\mu}\|_2 \geq \epsilon$. Then either

(1) $\widehat{\Sigma}$ has k eigenvalues at least $1 + \Omega(\epsilon)$, or

(2) The error comes from a k-dimensional subspace.

- Choose $k = \Theta(\log(1/\epsilon))$.

OUTLINE

Part II: High-Dimensional Robust Mean and Covariance Estimation

- Statements of Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- **Soft Outlier Removal**
- Experiments
- Extensions

Soft Outlier Removal

Let

$$S_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : 0 \leq w_i \leq \frac{1}{(1-2\epsilon)N} \right\}$$

Let $\delta = \Theta(\epsilon \log(1/\epsilon))$. Consider the convex set

$$\mathcal{C}_\delta = \left\{ w \in S_{N,\epsilon} : \left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \leq \delta \right\}$$

Algorithm:

- Find $w^* \in \mathcal{C}_\delta$
- Output $\widehat{\mu}_{w^*} = \sum_{i=1}^N w_i^* X_i$.

Main Issue: μ unknown.

Soft Outlier Removal

Let

$$S_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : 0 \leq w_i \leq \frac{1}{(1-2\epsilon)N} \right\}$$

Let $\delta = \Theta(\epsilon \log(1/\epsilon))$. Consider the convex set

$$\mathcal{C}_\delta = \left\{ w \in S_{N,\epsilon} : \left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \leq \delta \right\}$$

Algorithm:

- Find $w^* \in \mathcal{C}_\delta$
- Output $\widehat{\mu}_{w^*} = \sum_{i=1}^N w_i^* X_i$.

- Adaptation of key lemma gives: For all $w \in \mathcal{C}_\delta$, we have:

$$\|\widehat{\Sigma}_w\|_2 \leq 1 + \delta \quad \rightarrow \quad \|\widehat{\mu}_w - \mu\|_2 \leq O(\epsilon\sqrt{\log(1/\epsilon)})$$

Approximate Separation Oracle

Input: ϵ -corrupted set S and weight vector w

Output: Separation oracle for $\mathcal{C}_\delta$

- Let $\delta = \Theta(\epsilon \log(1/\epsilon))$
- Let $\widehat{\mu}_w = \sum_{i=1}^N w_i X_i$ and $\widehat{\Sigma}_w = \sum_{i=1}^N w_i X_i X_i^T - \widehat{\mu}_w \widehat{\mu}_w^T$

- Let (λ^*, v^*) be the top eigenvalue-eigenvector pair of $\widehat{\Sigma}_w$.
- If $\lambda^* \leq 1 + \delta$, return "YES".
- Otherwise, return the hyperplane $L : \mathbb{R}^d \to \mathbb{R}$ with

$$L(u) = \sum_{i=1}^N u_i ((X_i - \widehat{\mu}_w) \cdot v^*)^2 - \lambda^* .$$

OUTLINE

Part II: High-Dimensional Robust Mean and Covariance Estimation

- Statements of Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- **Experiments**
- Extensions

EXPERIMENTS

Being Robust (in High Dimensions) Can Be Practical
D., Kamath, Kane, Li, Moitra, Stewart, ICML'17

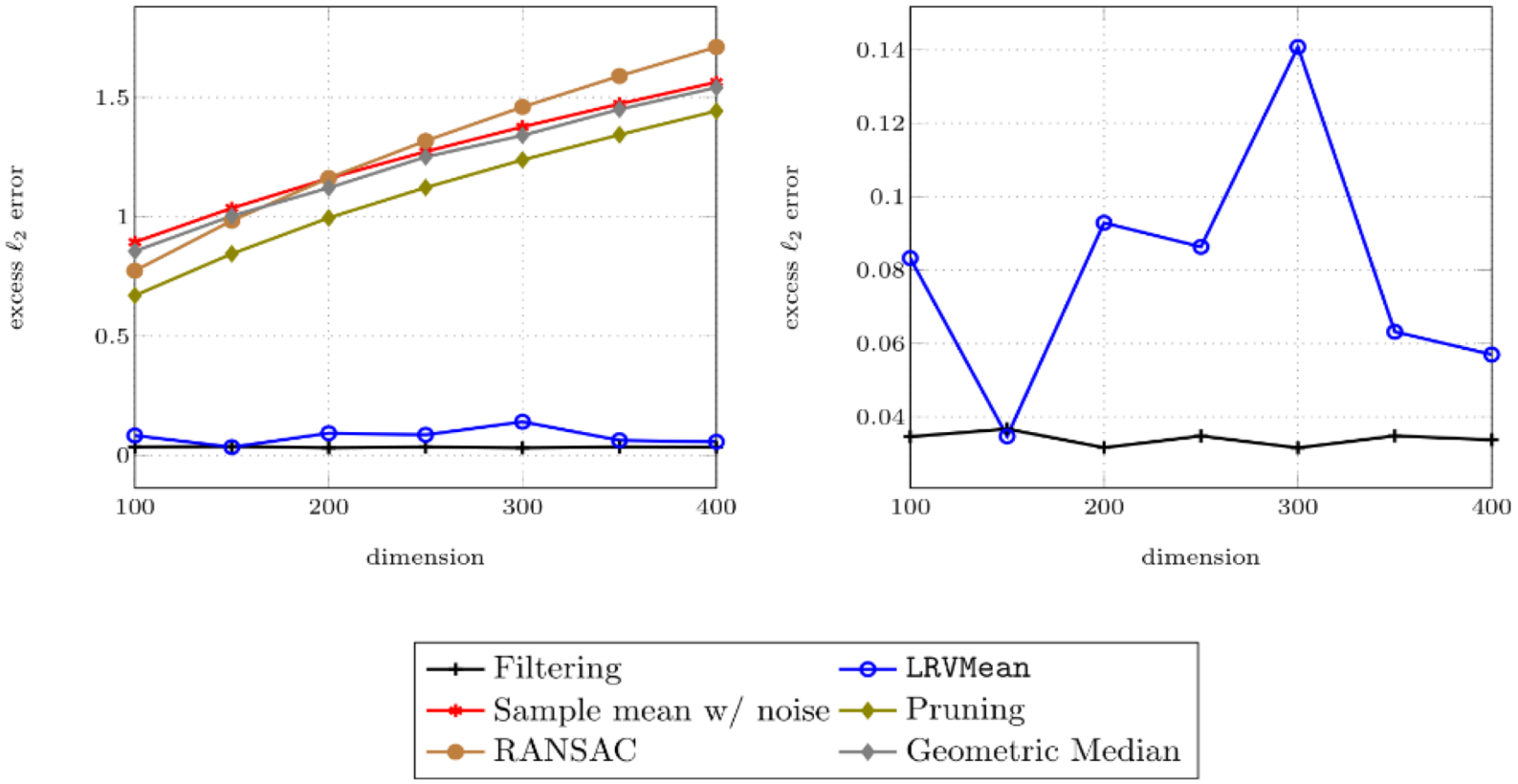
Synthetic Experiments: Unknown Mean

Error rates on synthetic data (**unknown mean**):

$$\mathcal{N}(\mu, I) \quad \textbf{+ 10\% noise}$$

Synthetic Experiments: Unknown Mean

Error rates on synthetic data (**unknown mean**):



Synthetic Experiments: Unknown Covariance (I)

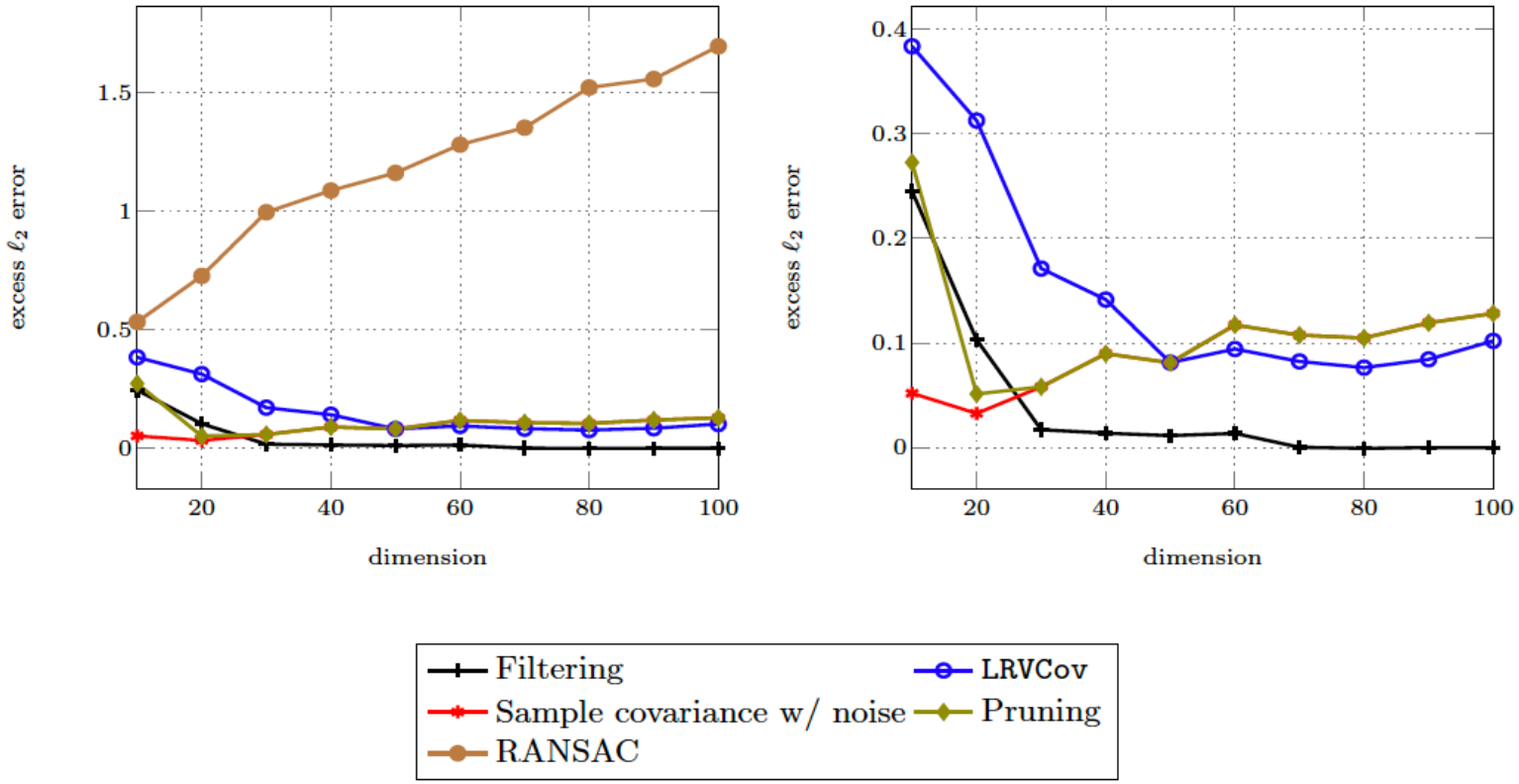
Error rates on synthetic data (**unknown covariance, isotropic**):

$\mathcal{N}(0, \Sigma)$ **+ 10% noise**

close to identity

Synthetic Experiments: Unknown Covariance (I)

Error rates on synthetic data (**unknown covariance, isotropic**):



Synthetic Experiments: Unknown Covariance (II)

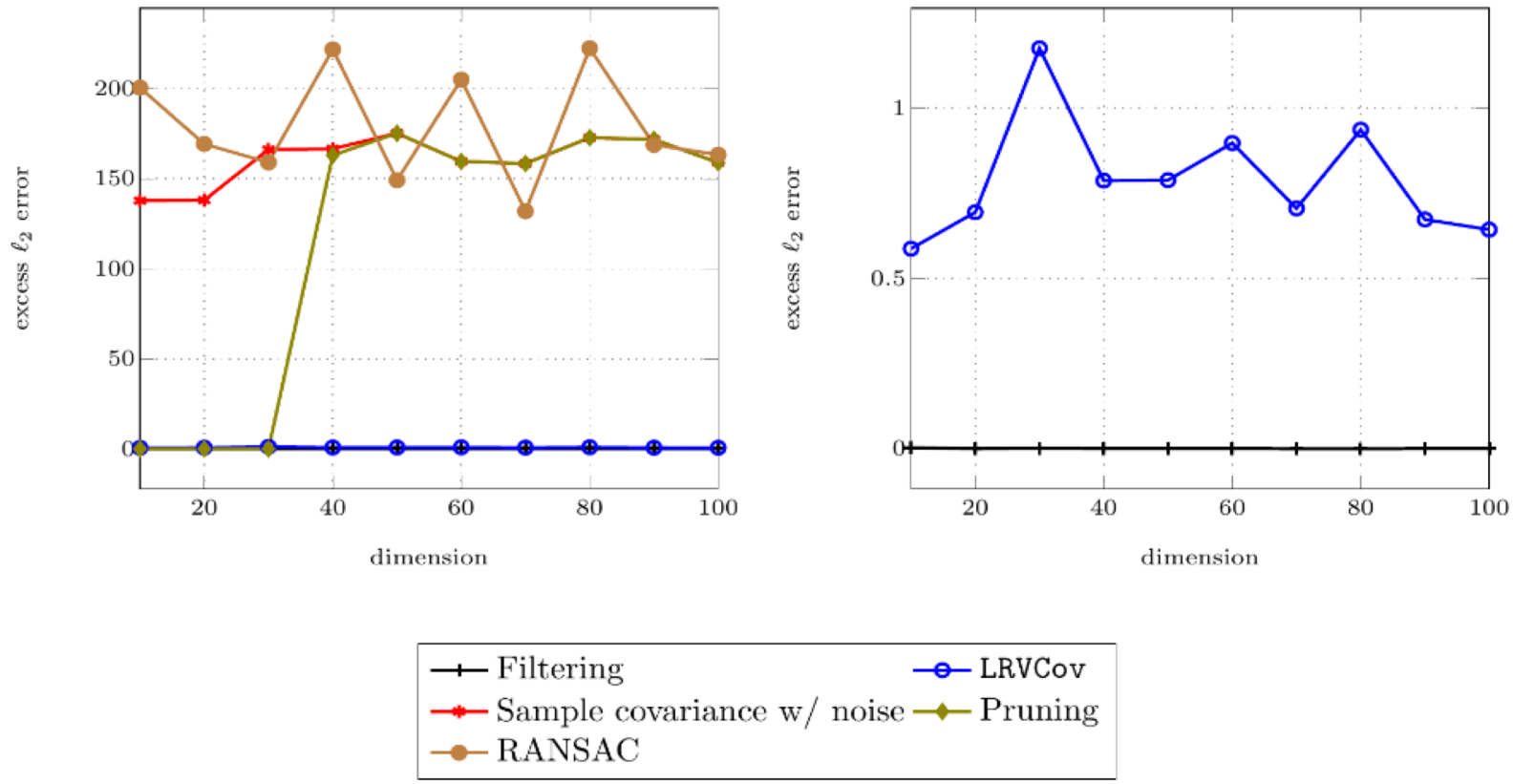
Error rates on synthetic data (**unknown covariance**):

$\mathcal{N}(0, \Sigma)$ **+ 10% noise**

far from identity

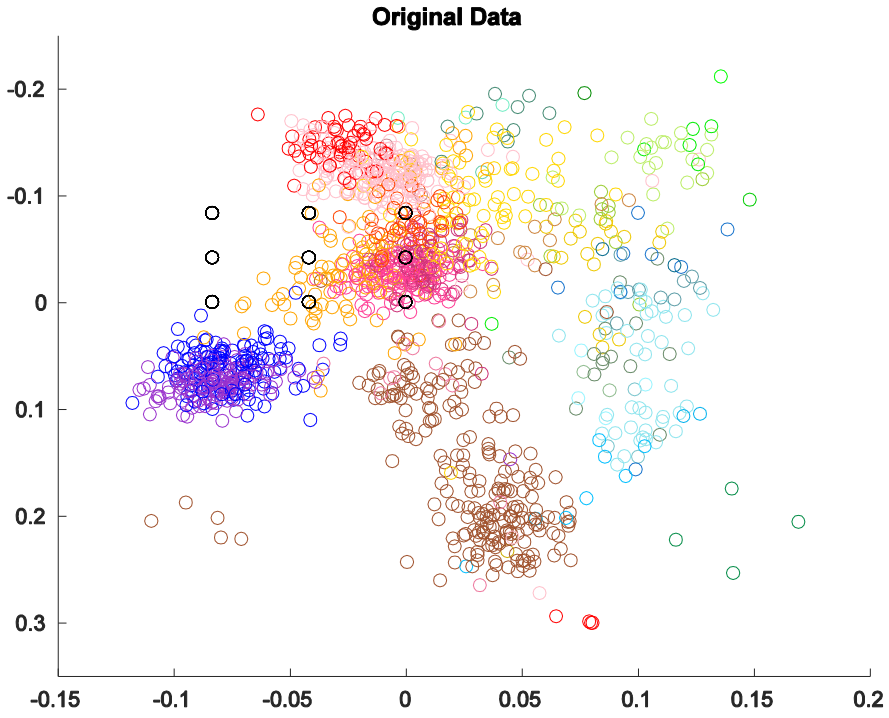
SYNTHETIC EXPERIMENTS: UNKNOWN COVARIANCE (II)

Error rates on synthetic data (**unknown covariance, anisotropic**):



REAL DATA EXPERIMENTS

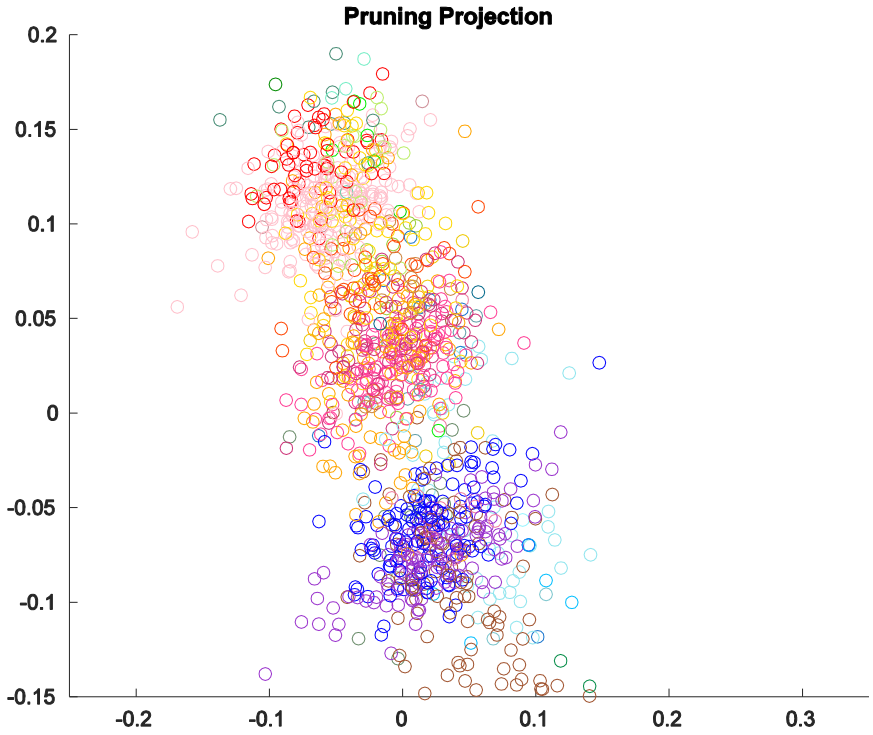
[Novembre et al. ’08]: Take top two singular vectors of people x SNP matrix (POPRES)



“Genes Mirror Geography in Europe”

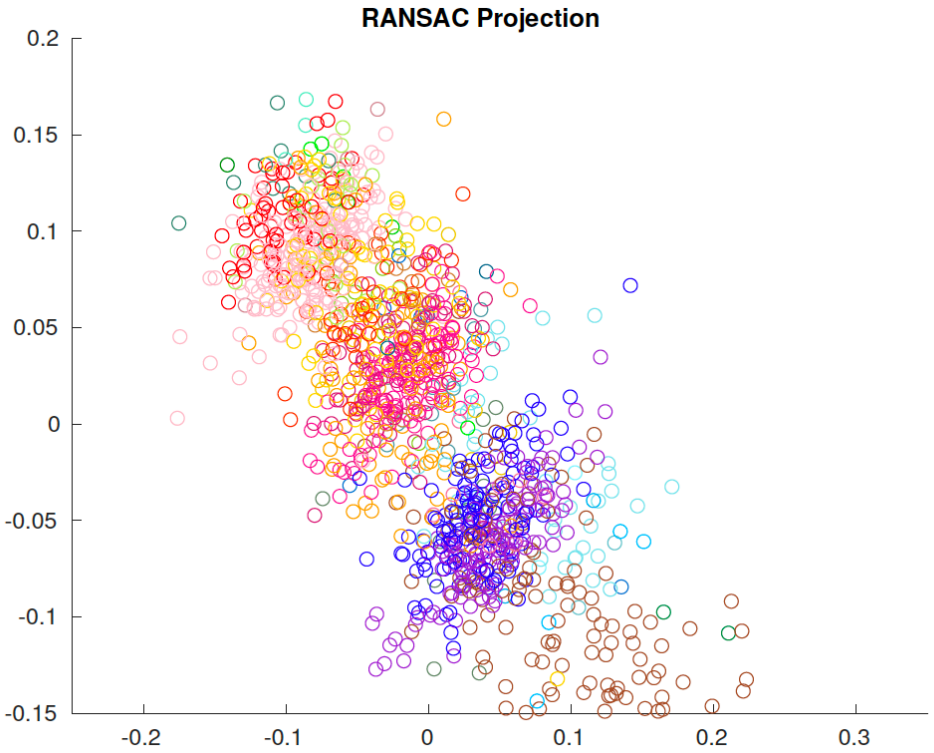
EXPERIMENTS: PRUNING PROJECTION

A comparison of error rate on semi-synthetic data:



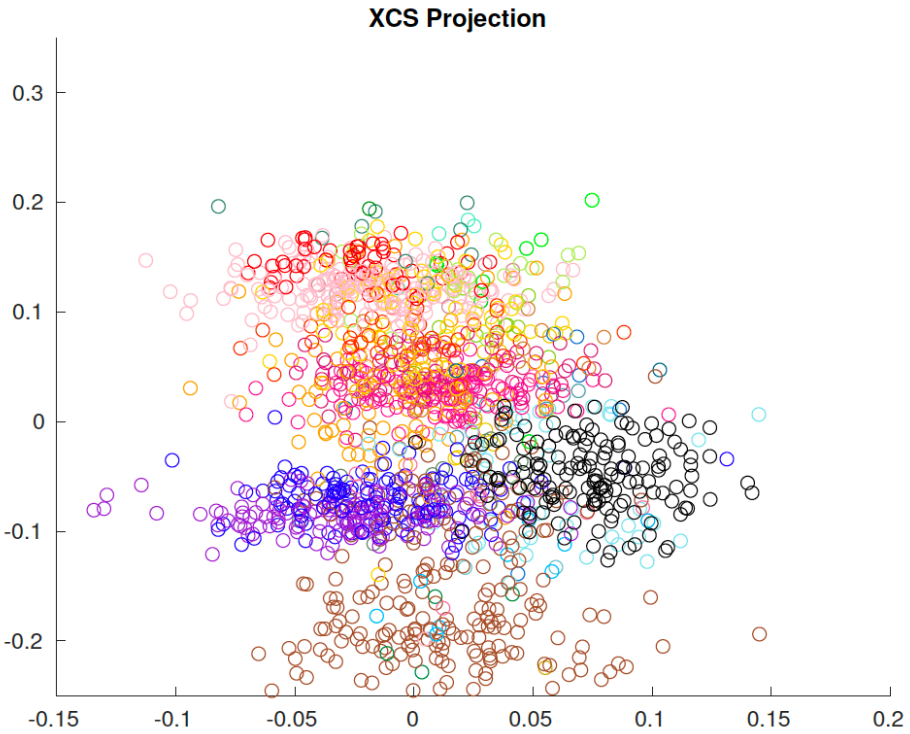
EXPERIMENTS: RANSAC PROJECTION

A comparison of error rate on semi-synthetic data:

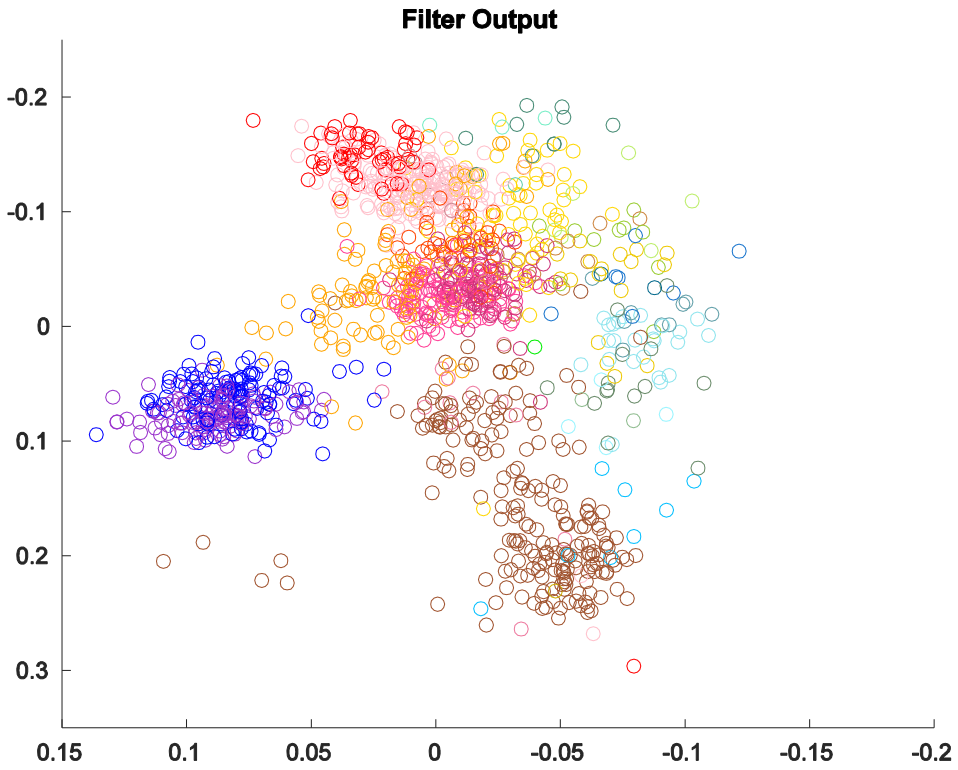


EXPERIMENTS: ROBUST PCA (XCS)

A comparison of error rate on semi-synthetic data:



EXPERIMENTS: FILTER PROJECTION

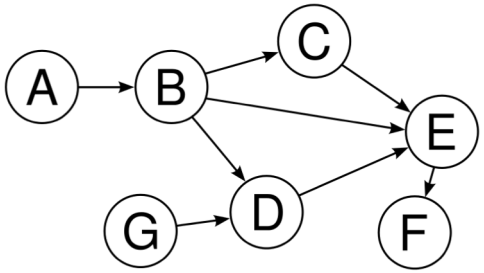


OUTLINE

Part II: High-Dimensional Robust Mean and Covariance Estimation

- Statements of Results
- Basics: Sample Complexity of Robust Estimation, Naïve Outlier Removal
- Overview of Algorithmic Approaches
- Certificate of Robustness
- Recursive Dimension Halving
- Iterative Filtering
- Soft Outlier Removal
- Experiments
- **Extensions**

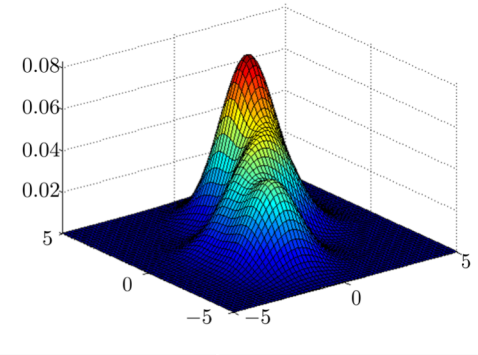
BEYOND ROBUST STATISTICS: ROBUST *UNSUPERVISED* LEARNING



Robustly Learning Graphical Models
[Cheng-**D-**Kane-Stewart'16,
D-Kane-Stewart'18]

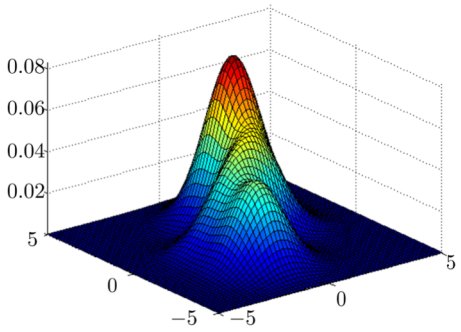
Computational/Statistical-Robustness Tradeoffs
[**D-**Kane-Stewart'17, **D**-Kong-Stewart'18]

Clustering in Mixture Models
[Charikar-Steinhardt-Valiant'17,
D-Kane-Stewart'18,
Hopkins-Li'18,
Kothari-Steinhardt-Steurer'18]

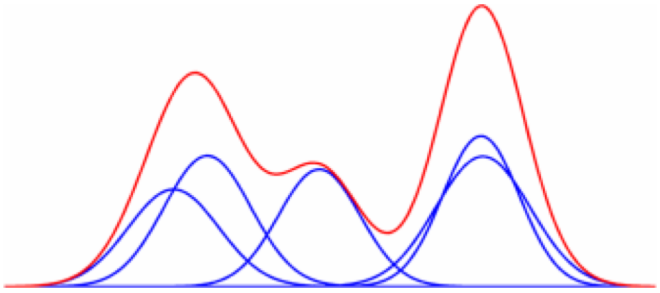


ROBUST *UNSUPERVISED* LEARNING

Clustering in Mixture Models
[**D-**Kane-Stewart'18]



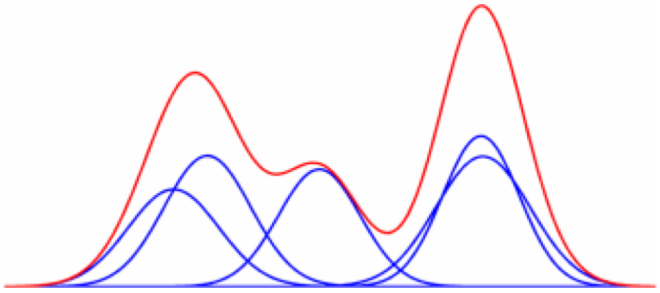
CLUSTERING IN MIXTURE MODELS



Under what assumptions can we disentangle mixture models?

Assumption	**Separation**	**Robustness?**	**Reference**
Spherical Gaussians	$k^{1/4}$	NO	[VW02]
Second Moments	$k^{1/2}$	NO	[AM05]
…			
Second Moments	$k^{1/2}$	YES	[CSV'17, DKS'18]
Spherical Gaussians	$\sqrt{\log k}$	YES	[DKS'18, HL'18, KSS'18]

Robustness As A Lens



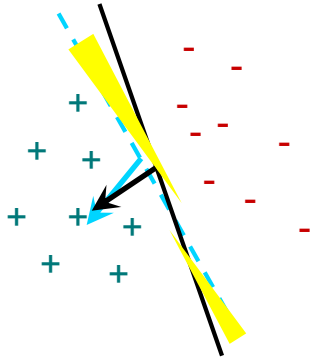
Under what assumptions can we disentangle mixture models?

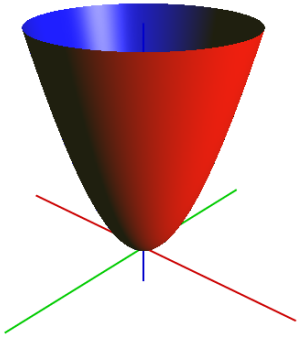
Assumption	Separation	Robustness?	Reference
Second Moments	$k^{1/2}$	YES	[DKS'18]
Spherical Gaussians	$\sqrt{\log k}$	YES	[DKS'18]

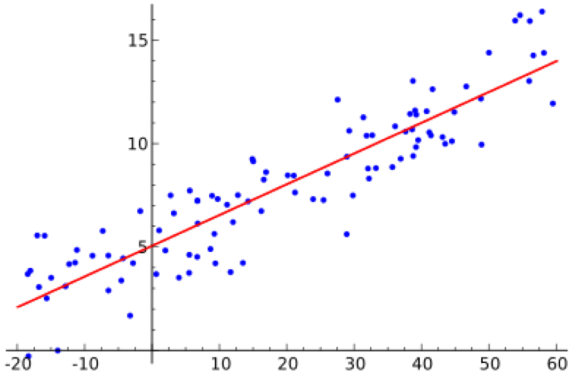
Main Idea: Remove Outliers from a dataset when the *majority* of the points are corrupted (**List-Decodable Learning**).

BEYOND ROBUST STATISTICS: ROBUST *SUPERVISED* LEARNING

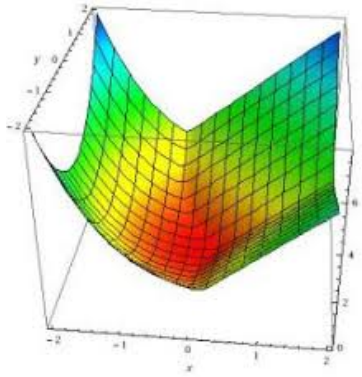
Malicious PAC Learning
[Klivans-Long-Servedio'10,
Awasthi-Balcan-Long'14,
D-Kane-Stewart'18]







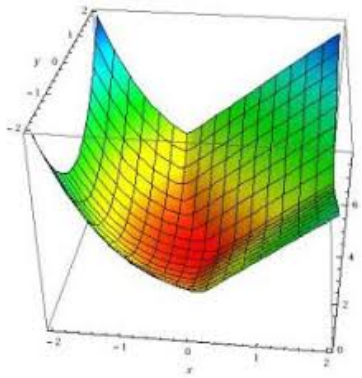
Robust Linear Regression
[**D-**Kong-Stewart'18,
Klivans-Kothari-Meka'18]



Stochastic (Convex) Optimization
[Prasad-Suggala-Balakrishnan-Ravikumar'18,
D-Kamath-Kane-Li-Steinhardt-Stewart'18]

ROBUST *SUPERVISED* LEARNING

Sever: A Robust Meta-Algorithm for Stochastic Optimization.
[D-Kamath-Kane-Li-Steinhardt-Stewart, ICML'19]



Robust Stochastic Convex Optimization

Problem: Given loss function $\ell(X, w)$ and ϵ- corrupted samples from a distribution $\mathcal{D}$ over X, minimize $f(w) = \mathbb{E}_{X \sim \mathcal{D}}[\ell(X, w)]$

Difficulty: Corrupted data can move the gradients.

Theorem: Suppose ℓ is convex and $\mathrm{Cov}_{\mathrm{X} \sim \mathcal{D}}[\nabla \ell(\mathrm{X}, \mathrm{w})] \preceq \sigma^2 \cdot I$.
Under mild assumptions on $\mathcal{D}$, can recover a point such that

$$f(\hat{w}) - \min_{w} f(w) \leq O(\sigma \sqrt{\epsilon}) \ .$$

Main Idea: Filter at minimizer of empirical risk.

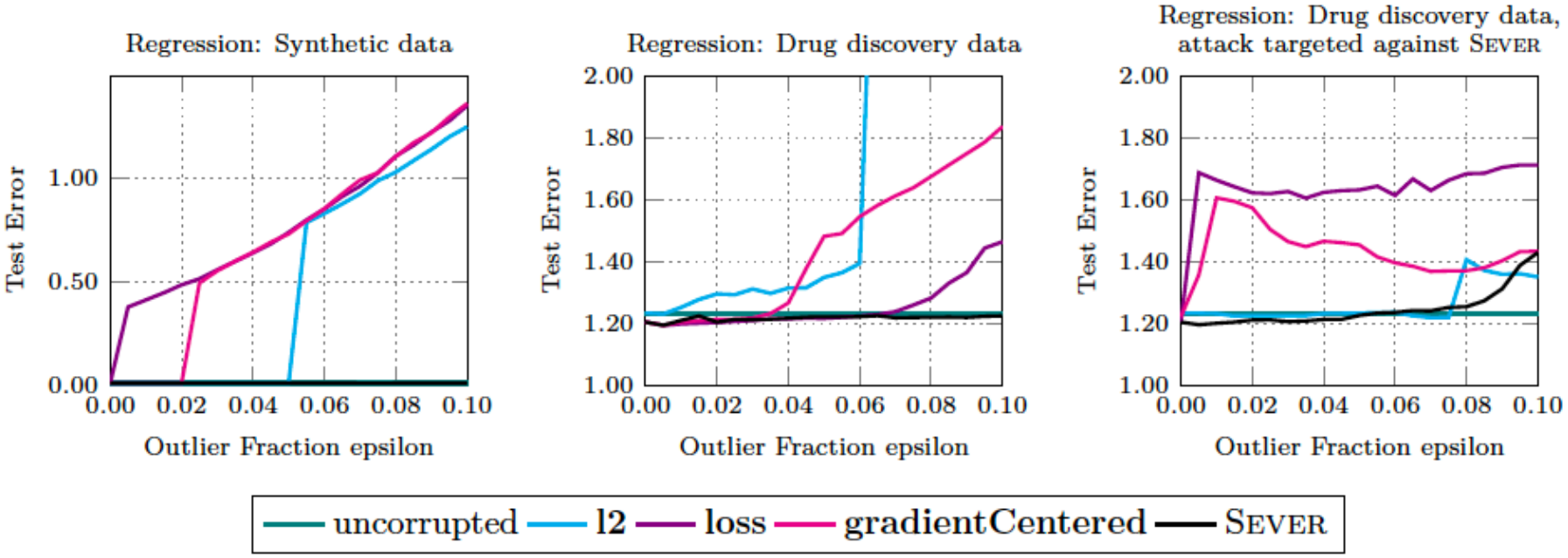
SPECIFIC APPLICATIONS

Corollary: Outlier-robust learning algorithms with dimension-independent error guarantees for:

- SVMs
- Linear Regression
- Logistic Regression
- GLMs
- Experimental Performance Against Data Poisoning Attacks.

Concurrent works obtained tighter guarantees in terms of either sample complexity or error, by focusing on specific tasks and distributional assumptions [Klivans-Kothari-Meka'18, Diakonikolas-Kong-Stewart'18, …].

EXPERIMENTS: RIDGE REGRESSION



Subsequent Related Works

- Graphical Models [Cheng-D-Kane-Stewart'16, D-Kane-Stewart'18]
- Sparse models (e.g., sparse PCA, sparse regression) [Li'17, Du-Balakrishan-Singh'17, Liu-Shen-Li-Caramanis'18, D-Karmalkar-Kane-Price-Stewart'18]
- List-Decodable Learning [Charikar-Steinhardt-Valiant '17, Meister-Valiant'18, D-Kane-Stewart'18, Karmalkar-Klivans-Kothari'19, Raghavendra-Yau'19]
- Robust PAC Learning [Klivans-Long-Servedio'10, Awasthi-Balcan-Long'14, D-Kane-Stewart'18]
- Robust estimation of higher moments via SoS [Hopkins-Li'18, Kothari-Steinhardt-Steurer'18, …]
- "SoS Free" robust estimation of higher moments [D-Kane-Stewart'18]
- Robust Regression [Klivans-Kothari-Meka'18, D-Kong-Stewart'18, …]
- Robust Stochastic Optimization [Prasad-Suggala-Balakrishnan-Ravikumar'18, D-Kamath-Kane-Li-Steinhard-Stewart'18]
- Near-Linear Time Robust Estimators [Chen-D-Ge'18, Cheng-D-Ge-Woodruff'19, Depersin-Lecue'19, Dong-Hopkins-Li'19, …]

Part III: Computational Limits In Robust Estimation

OUTLINE

Part III: Computational Limits to Robust Estimation

- Statistical Query Learning Model
- Our Results
- Generic Lower Bound Technique
- Applications: Robust Mean Estimation & Learning GMMs

OUTLINE

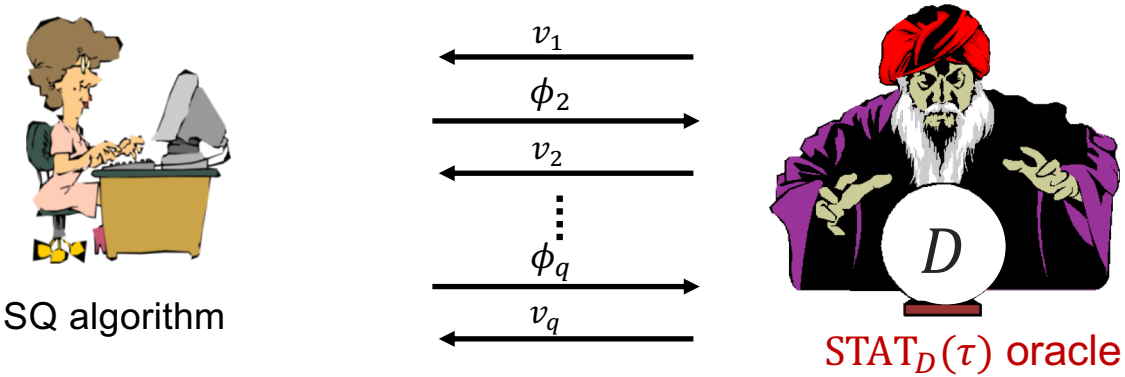
Part III: Computational Limits to Robust Estimation

- **Statistical Query Learning Model**
- Our Results
- Generic Lower Bound Technique
- Applications: Robust Mean Estimation & Learning GMMs

STATISTICAL QUERIES [KEARNS'93]

$x_1, x_2, \ldots, x_m \sim D$ over X

STATISTICAL QUERIES [KEARNS'93]



SQ algorithm

$\text{STAT}_D(\tau)$ oracle

$$\phi_1 : X \to [-1,1] \qquad |v_1 - \mathbf{E}_{x\sim D}[\phi_1(x)]| \le \tau$$

τ is tolerance of the query; $\tau = 1/\sqrt{m}$

Problem $P \in \text{SQCompl}(q, m)$:
If exists a SQ algorithm that solves P using q queries to $\text{STAT}_D(\tau = 1/\sqrt{m})$

Power of SQ Learning Algorithms

- **Restricted Model**: Hope to prove unconditional computational lower bounds.
- **Powerful Model**: Wide range of algorithmic techniques in ML are implementable using SQs*:
- PAC Learning: AC^0, decision trees, linear separators, boosting.
- Unsupervised Learning: stochastic convex optimization, moment-based methods, *k*-means clustering, EM, …
  [Feldman-Grigorescu-Reyzin-Vempala-Xiao/JACM'17]
- **Only known exception**: Gaussian elimination over finite fields (e.g., learning parities).
- For all problems in this talk, strongest known algorithms are SQ.

Methodology for Proving SQ Lower Bounds

Statistical Query Dimension:

- Fixed-distribution PAC Learning
  [Blum-Furst-Jackson-Kearns-Mansour-Rudich'95; …]

- General Statistical Problems
  [Feldman-Grigorescu-Reyzin-Vempala-Xiao'13, …, Feldman'16]

- Pairwise correlation between D_1 and D_2 with respect to D:

$$\chi_D(D_1, D_2) := \int_{\mathbb{R}^d} D_1(x) D_2(x)/D(x) dx - 1$$

- **Fact**: Suffices to construct a large set of distributions that are *nearly* uncorrelated.

OUTLINE

Part III: Computational Limits to Robust Estimation

- Statistical Query Learning Model
- **Our Results**
- Generic Lower Bound Technique
- Applications: Robust Mean Estimation & Learning GMMs

GENERIC SQ LOWER BOUND CONSTRUCTION

General Technique for SQ Lower Bounds:
Leads to Tight Lower Bounds
for a range of High-dimensional Estimation Tasks

Concrete Applications of our Technique:

- Robustly Learning Mean and Covariance
- Learning Gaussian Mixture Models (GMMs)
- Statistical-Computational Tradeoffs (e.g., sparsity)
- Robustly Testing a Gaussian

SQ Lower Bound for Robust Mean Estimation

Theorem: Suppose $d \geq \mathrm{polylog}(1/\epsilon)$. Any SQ algorithm that learns an ϵ - corrupted Gaussian $\mathcal{N}(\mu, I)$ in the strong contamination model within error

$$O(\epsilon\sqrt{\log(1/\epsilon)}/M)$$

requires either:

- SQ queries of accuracy $d^{-M/6}$

or

- At least $d^{\Omega(M^{1/2})}$ many SQ queries.

Take-away: Any asymptotic improvement in error guarantee over prior work requires super-polynomial time.

SQ Lower Bounds for Learning *Separated* GMMs

Theorem: Suppose that $d \geq \mathrm{poly}(k)$. Any SQ algorithm that learns *separated* k-GMMs over $\mathbb{R}^d$ to constant error requires either:

- SQ queries of accuracy $d^{-k/6}$

or

- At least $2^{\Omega(d^{1/8})} \geq d^{2k}$ many SQ queries.

Take-away: Computational complexity of learning GMMs is inherently exponential in **number of components**.

Applications: Concrete SQ Lower Bounds

Learning Problem	Upper Bound	SQ Lower Bound
Robust Gaussian Mean Estimation	Error: $O(\epsilon \log^{1/2}(1/\epsilon))$ [DKKLMS'16]	Runtime Lower Bound: $d^{\mathrm{poly}(M)}$ for factor M improvement in error.
Robust Gaussian Covariance Estimation	Error: $O(\epsilon \log(1/\epsilon))$ [DKKLMS'16]	
Learning k-GMMs (without noise)	Runtime: $d^{g(k)}$ [MV'10, BS'10]	Runtime Lower Bound: $d^{\Omega(k)}$
Robust k-Sparse Mean Estimation	Sample size: $\tilde{O}(k^2 \log d)$ [BDLS'17]	If sample size is $O(k^{1.99})$ runtime lower bound: $d^{k^{\Omega(1)}}$
Robust Covariance Estimation in Spectral Norm	Sample size: $\tilde{O}(d^2)$ [DKKLMS'16]	If sample size is $O(d^{1.99})$ runtime lower bound: $2^{d^{\Omega(1)}}$

OUTLINE

Part III: Computational Limits to Robust Estimation

- Statistical Query Learning Model
- Our Results
- **Generic Lower Bound Technique**
- Applications: Robust Mean Estimation & Learning GMMs

General Recipe For SQ Lower Bounds

- **Step #1:** Construct distribution $\mathbf{P}_v$ that is standard Gaussian in all directions except v.
- **Step #2:** Construct the univariate projection in the v direction so that it matches the first m moments of $\mathcal{N}(0,1)$
- **Step #3:** Consider the family of instances $\mathcal{D} = \{\mathbf{P}_v\}_v$

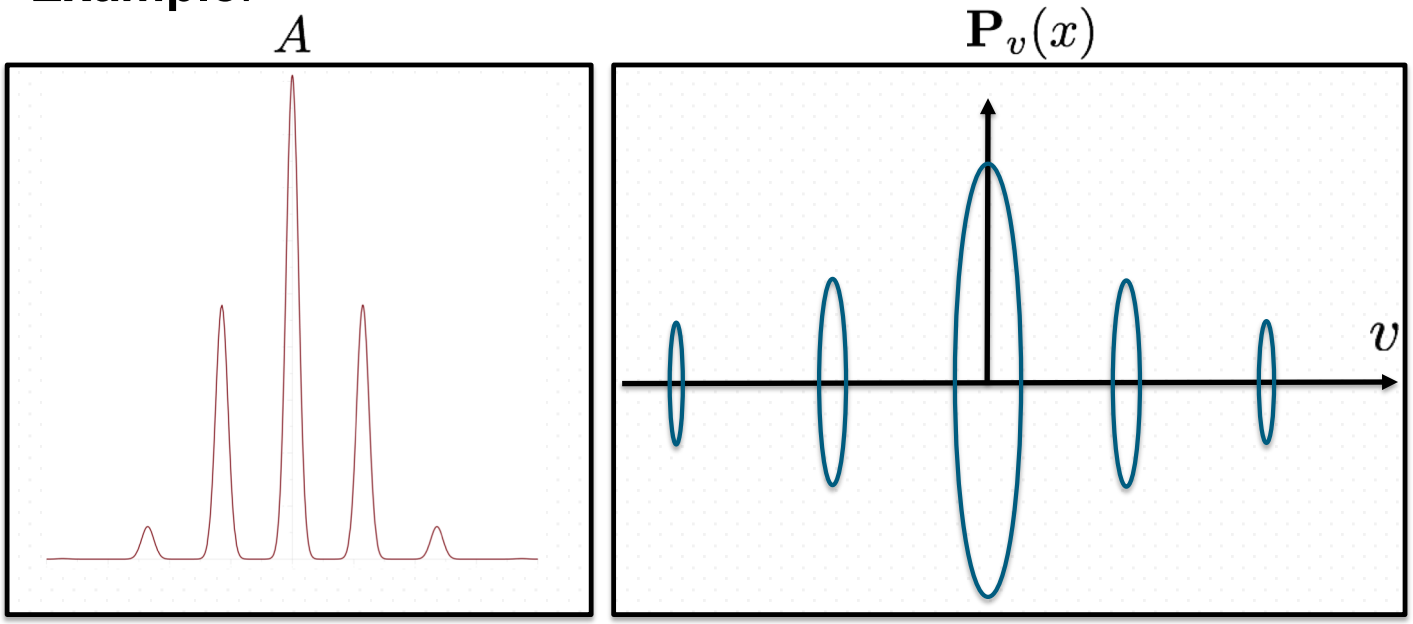
Non-Gaussian Component Analysis [Blanchard *et al.* 2006]

Hidden Direction Distribution

Definition: For a unit vector v and a univariate distribution with density A, consider the high-dimensional distribution

$$\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right) / (2\pi)^{(d-1)/2}.$$

Example:



Generic SQ Lower Bound

Definition: For a unit vector v and a univariate distribution with density A, consider the high-dimensional distribution

$$\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right) / (2\pi)^{(d-1)/2}.$$

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

Why is Finding a Hidden Direction Hard

Observation: Low-Degree Moments do not help.

- A matches the first m moments of $\mathcal{N}(0,1)$
- The first m moments of $\mathbf{P}_v$ are identical to those of $\mathcal{N}(0,I)$
- Degree-$(m+1)$ moment tensor has $\Omega(d^m)$ entries.

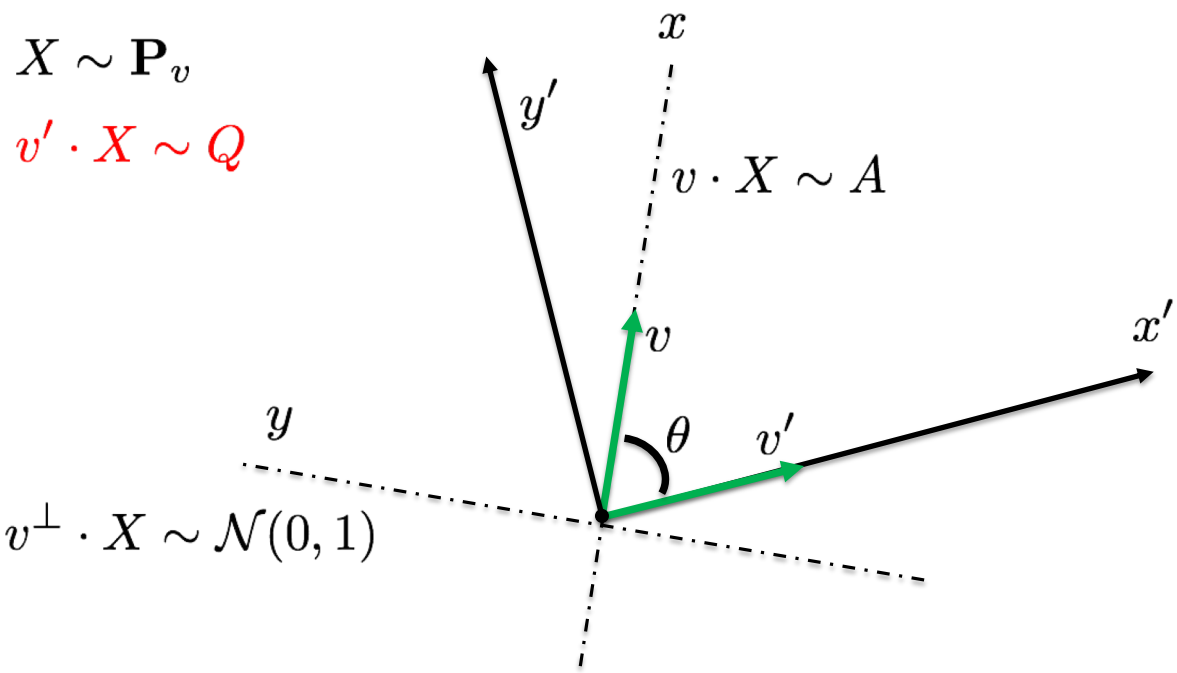
Claim: Random projections do not help.

- To distinguish between $\mathbf{P}_v$ and $\mathcal{N}(0,I)$, would need exponentially many random projections.

1-D Projections Are *Almost* Standard Gaussians

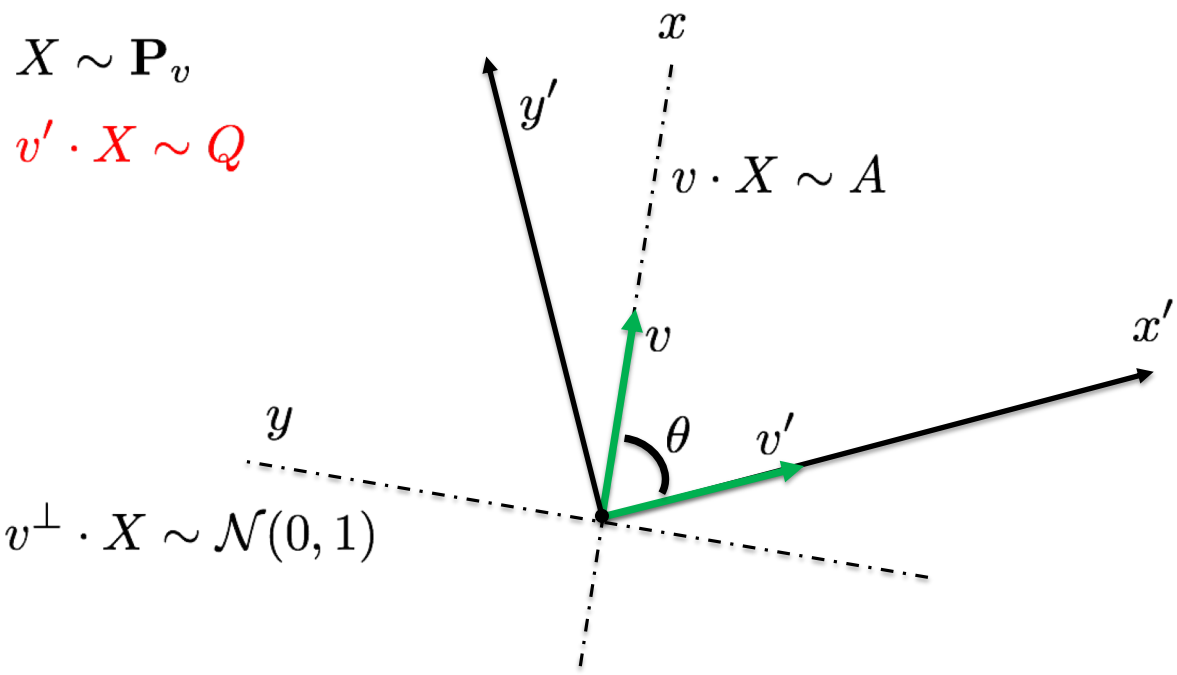
Key Lemma: Let Q be the distribution of $v' \cdot X$, where $X \sim \mathbf{P}_v$. Then, we have that:

$$\chi^2(Q, \mathcal{N}(0,1)) \leq (v \cdot v')^{2(m+1)} \chi^2(A, \mathcal{N}(0,1))$$



Proof of Key Lemma (I)

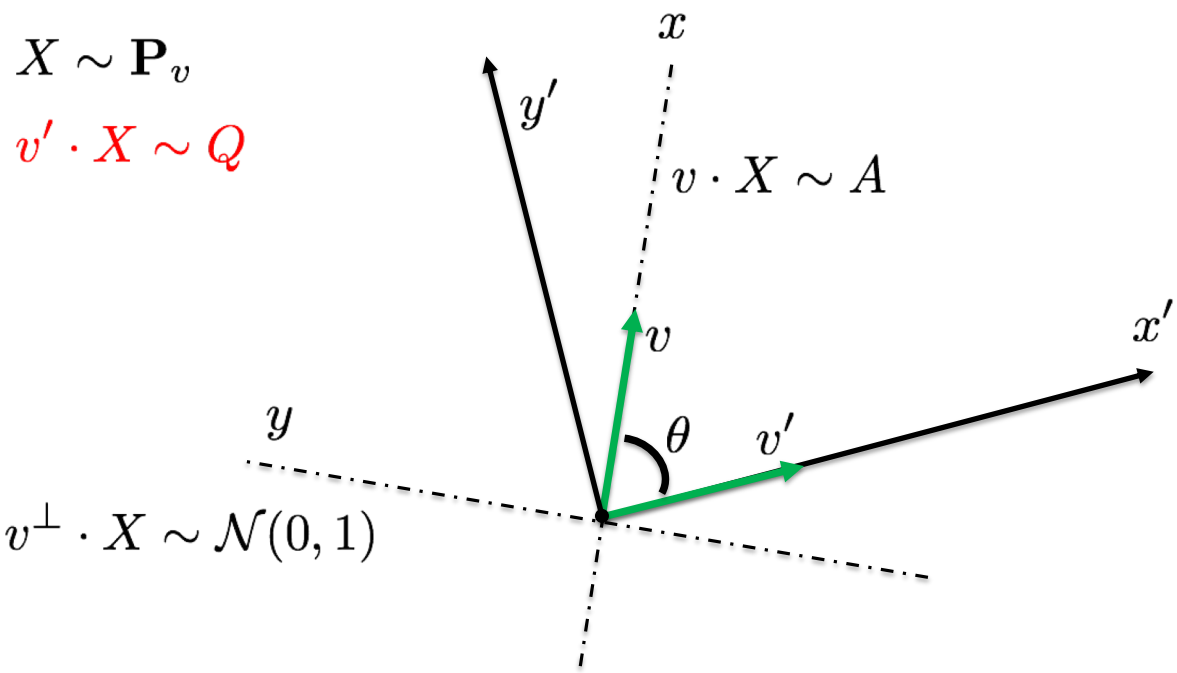
$$Q(x') = \int_{\mathbb{R}} A(x)G(y)dy'$$



Proof of Key Lemma (I)

$$Q(x') = \int_{\mathbb{R}} A(x)G(y)dy'$$

$$= \int_{\mathbb{R}} A(x' \cos\theta + y' \sin\theta)G(x' \sin\theta - y' \cos\theta)dy'$$



Proof of Key Lemma (II)

$$Q(x') = \int_{\mathbb{R}} A(x'\cos\theta + y'\sin\theta)G(x'\sin\theta - y'\cos\theta)dy'$$

$$= (U_\theta A)(x')$$

where U_θ is the operator over $f : \mathbb{R} \to \mathbb{R}$

$$U_\theta f(x) := \int_{y\in\mathbb{R}} f(x\cos\theta + y\sin\theta)G(x\sin\theta - y\cos\theta)dy$$

Gaussian Noise (Ornstein-Uhlenbeck) Operator

Eigen-decomposition of Ornstein-Uhlenbeck Operator

Linear Operator U_θ acting on functions $f : \mathbb{R} \to \mathbb{R}$

$$U_\theta f(x) := \int_{y \in \mathbb{R}} f(x \cos\theta + y \sin\theta) G(x \sin\theta - y \cos\theta) dy$$

Fact (Mehler'66): $U_\theta(He_i G)(x) = \cos^i(\theta) He_i(x) G(x)$

- $He_i(x)$ denotes the degree-i Hermite polynomial.
- Note that $\{He_i(x)G(x)/\sqrt{i!}\}_{i \geq 0}$ are orthonormal with respect to the inner product

$$\langle f, g \rangle = \int_{\mathbb{R}} f(x) g(x) / G(x) dx$$

Generic SQ Lower Bound

Definition: For a unit vector v and a univariate distribution with density A, consider the high-dimensional distribution

$$\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right) / (2\pi)^{(d-1)/2}.$$

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within δ error requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

OUTLINE

Part III: Computational Limits to Robust Estimation

- Statistical Query Learning Model
- Our Results
- Generic Lower Bound Technique
- **Applications: Robust Mean Estimation & Learning GMMs**

SQ Lower Bound for Robust Mean Estimation (I)

Want to show:

Theorem: Any SQ algorithm that learns an ϵ- corrupted Gaussian in the strong contamination model within error $\epsilon\sqrt{\log(1/\epsilon)/M}$ requires either SQ queries of accuracy $d^{-M/6}$ or at least $d^{\Omega(M^{1/2})}$ many SQ queries.

by using our generic proposition:

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

SQ Lower Bound for Robust Mean Estimation (II)

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as *v, v'* are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

Lemma: There exists a univariate distribution A that is ϵ - close to $\mathcal{N}(\mu, 1)$ such that:

- A agrees with $\mathcal{N}(0,1)$ on the first M moments.
- We have that $\mu = \Omega(\epsilon\sqrt{\log(1/\epsilon)/M^2})$
- Whenever v and v' are nearly orthogonal $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) = \Omega(\mu)$.

SQ Lower Bound for Learning GMMs (I)

Want to show:

Theorem: Any SQ algorithm that learns separated k-GMMs over $\mathbb{R}^d$ to constant error requires either SQ queries of accuracy $d^{-k/6}$ or at least $2^{\Omega(d^{1/8})} \geq d^{2k}$ many SQ queries.

by using our generic proposition:

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

SQ Lower Bound for Learning GMMs (II)

Proposition: Suppose that:

- *A* matches the first *m* moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as *v, v'* are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

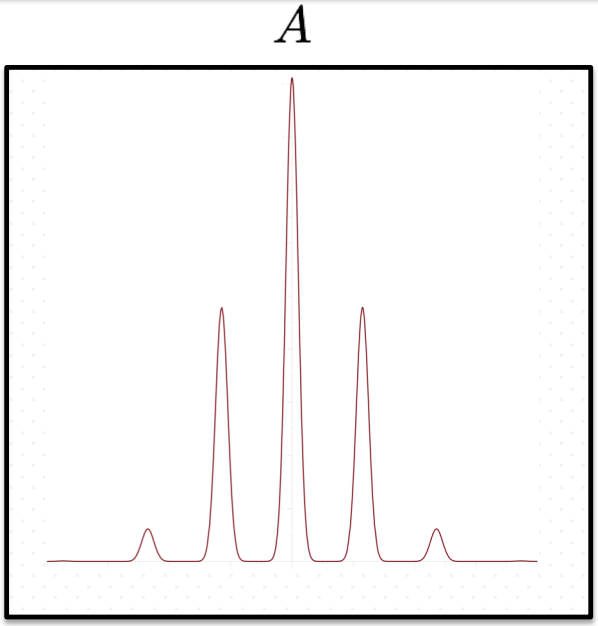
Lemma: There exists a univariate distribution A that is a k-GMM with components A_i such that:

- A agrees with $\mathcal{N}(0,1)$ on the first $2k$-1 moments.
- Each pair of components are separated.
- Whenever v and v' are nearly orthogonal $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) \geq 1/2$.

SQ Lower Bound for Learning GMMs (III)

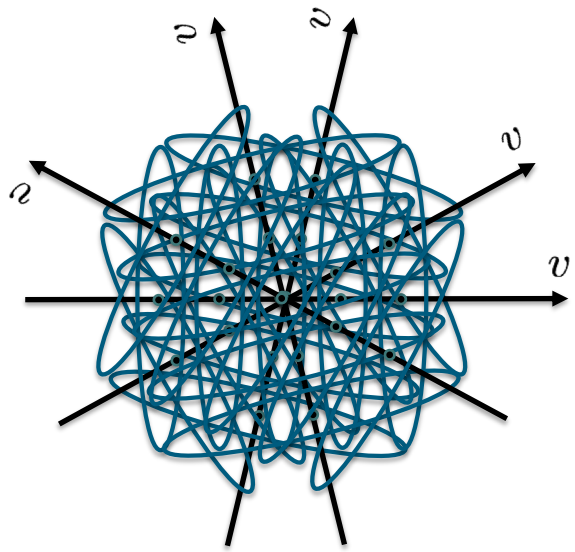
Lemma: There exists a univariate distribution A that is a k-GMM with components A_i such that:

- A agrees with $\mathcal{N}(0,1)$ on the first $2k$-1 moments.
- Each pair of components are separated.
- Whenever v and v' are nearly orthogonal $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) \geq 1/2$.



SQ Lower Bound for Learning GMMs (IV)

High-Dimensional Distributions $\mathbf{P}_v$ look like "parallel pancakes":



Efficiently learnable for k=2. [Brubaker-Vempala'08]

Part IV: Future Directions

Future Directions: Computational Lower Bounds

- General Technique to Prove SQ Lower Bounds
- Robustness can make high-dimensional estimation harder computationally and information-theoretically.

Future Directions:

- Further Applications of this Framework
  - List-Decodable Mean Estimation [D-Kane-Stewart'18]
  - Robust Regression [D-Kong-Stewart'18]
  - Adversarial Examples [Bubeck-Price- Razenshteyn'18]
  - Discrete Distributions [D-Gouleakis-Kane-Stewart'19]

- Alternative Evidence of Computational Hardness?
  - ❖ SoS Lower Bounds
  - ❖ Reductions from Average-Case Problems (e.g., Planted Clique, R-3SAT)
  - ❖ Reductions from Worst-case Problems? First step: [Hopkins-Li, COLT'19]

Future Directions: Algorithms

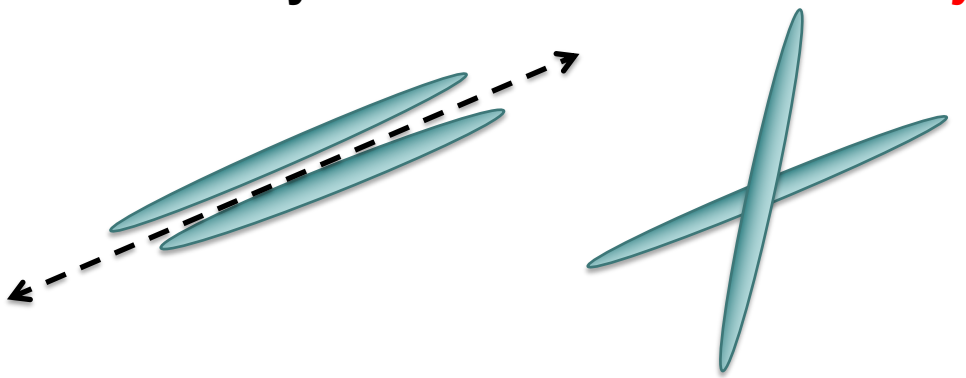
- Pick your favorite high-dimensional probabilistic model for which a (non-robust) efficient learning algorithm is known.
- Make it robust!

Concrete Algorithmic Open Problems

Open Problem 1: Robustly Estimating Gaussian *Covariance* Within Error $O(\epsilon)$ in Additive Contamination Model (Huber's Model)

Currently Best Known Algorithm [DKKLMS'18] runs in time $\mathrm{poly}(d) \cdot (1/\epsilon)^{\mathrm{polylog}(1/\epsilon)}$.

Open Problem 2: Robustly Learn a Mixture of 2 *Arbitrary* Gaussians



Spherical components: [Diakonikolas-Kane-Stewart'18, Hopkins-Li'18, Kothari-Steinhardt'18]

FAST / NEAR-LINEAR TIME ALGORITHMS

Filtering for robust mean estimation is practical, but runtime is *super-linear* $\tilde{\Theta}(Nd^2)$.

Question: Can we design near-linear time algorithms?

- Robust Mean Estimation:
  - ❖ [Cheng-D-Ge, SODA'19] $\tilde{\Theta}(Nd/\mathrm{poly}(\epsilon))$.
  - ❖ [Depersin-Lecue, Arxiv-June 2019] $\tilde{\Theta}(Nd)$.
  - ❖ [Dong-Hopkins-Li, June 2019] $\tilde{\Theta}(Nd)$.

- How about more general estimation tasks?
  - ❖ Robust Covariance Estimation [Cheng-D-Ge-Woodruff, COLT'19]
  - ❖ Robust *Sparse* Estimation?
  - ❖ List-Decodable Learning?

Broader Research Directions

General Algorithmic Theory of Robustness

How can we robustly learn rich representations of data, based on natural hypotheses about the structure in data?

Can we robustly *test* our hypotheses about structure in data before learning?

Broader Challenges:

- Richer Families of Problems and Models
- Connections to Non-convex Optimization, Adversarial Examples, GANs, …
- Relation to Related Notions of Algorithmic Stability
  (Differential Privacy, Adaptive Data Analysis)
- Further Applications (ML Security, Computer Vision, …)
- Other notions of robustness?

Thank you!
Questions?