Robust Sparse Statistics

Daniel M. Kane

Departments of CS/Math
University of California, San Diego
dakane@ucsd.edu

June 23rd, 2019
Overview

- Sparse Estimation
- Robust Version
- Convex Relaxation
- Further Directions
Sparse Mean Estimation

- Given $X \sim N(\mu, I) \subset \mathbb{R}^d$ it takes $O(d/\epsilon^2)$ samples to learn $\mu$ to error $\epsilon$.
- What if extra information is known about $\mu$? Can we do better?
  - In particular, what if $\mu$ is known to be sparse?

\[
|\hat{\mu}|_0 \leq k,
\]

then with $O\left(\frac{k \log(d)}{\epsilon^2}\right)$ samples suffices:

- Sample mean learns each coordinate to error $\frac{\epsilon}{\sqrt{k}}$.
- Truncating to $k$ largest coordinates ($\hat{\mu}_k$) gives error $\epsilon$.
- For $k \ll d$, this is a substantial improvement.
Sparse Mean Estimation

- Given $X \sim N(\mu, I) \subset \mathbb{R}^d$ it takes $O(d/\epsilon^2)$ samples to learn $\mu$ to error $\epsilon$.

- What if extra information is known about $\mu$? Can we do better? 
  - In particular, what if $\mu$ is known to be sparse?

If $|\mu|_0 \leq k$, then with $O(k \log(d)/\epsilon^2)$ samples suffices:
- Sample mean learns each coordinate to error $\epsilon/2\sqrt{k}$.
- Truncating to $k$ largest coordinates ($\hat{\mu}_k$) gives error $\epsilon$.
- For $k \ll d$, this is a substantial improvement.
Robust Sparse Mean Estimation

What if we want to do this robustly? Can we learn $\mu$ up to error $\tilde{O}(\epsilon)$ in the presence of adversarial errors with $o(d)$ samples?
Robust Sparse Mean Estimation

What if we want to do this robustly? Can we learn $\mu$ up to error $\tilde{O}(\epsilon)$ in the presence of adversarial errors with $o(d)$ samples?

First considered by [Balakrishnan–Du–Li–Singh ’17].
Basic Algorithm

Non-Sparse Robust Mean Estimation:

- \( \hat{\mu} \approx \mu \) unless there is a \( v \) with \( |v|_2 = 1 \) and \( v \cdot (\hat{\mu} - \mu) \) large.
- If such \( v \) exists, \( \text{Var}(v \cdot X) \) large.
- Determine if there is a \( v \) with \( |v|_2 = 1 \) and \( v^T \text{Cov}(X) v \) large.
  - If not, return \( \hat{\mu} \)
  - If so, filter on \( v \cdot X \) and repeat
Basic Algorithm

Sparse Robust Mean Estimation:

- $\hat{\mu} \approx \mu$ unless there is a $\nu$ with $|\nu|_2 = 1$ and $\nu \cdot (\hat{\mu} - \mu)$ large.
- If such $\nu$ exists, $\text{Var}(\nu \cdot X)$ large.
- Determine if there is a $\nu$ with $|\nu|_2 = 1$ and $\nu^T \text{Cov}(X) \nu$ large.
  - If not, return $\hat{\mu}$
  - If so, filter on $\nu \cdot X$ and repeat
Basic Algorithm

Sparse Robust Mean Estimation:

- \( \hat{\mu}_k \approx \mu \) unless there is a \( \nu \) with \( \|\nu\|_2 = 1 \) and \( \nu \cdot (\hat{\mu} - \mu) \) large.
- If such \( \nu \) exists, \( \text{Var}(\nu \cdot X) \) large.
- Determine if there is a \( \nu \) with \( \|\nu\|_2 = 1 \) and \( \nu^T \text{Cov}(X) \nu \) large.
  - If not, return \( \hat{\mu} \)
  - If so, filter on \( \nu \cdot X \) and repeat
Basic Algorithm

Sparse Robust Mean Estimation:

- \( \hat{\mu}_k \approx \mu \) unless there is a 2k-sparse \( |\nu|_2 = 1 \) with \( \nu \cdot (\hat{\mu} - \mu) \) large.
- If such \( \nu \) exists, \( \text{Var}(\nu \cdot X) \) large.
- Determine if there is a \( \nu \) with \( |\nu|_2 = 1 \) and \( \nu^T \text{Cov}(X) \nu \) large.
  - If not, return \( \hat{\mu} \)
  - If so, filter on \( \nu \cdot X \) and repeat
Basic Algorithm

Sparse Robust Mean Estimation:

- $\hat{\mu}_k \approx \mu$ unless there is a 2k-sparse $|\nu|_2 = 1$ with $\nu \cdot (\hat{\mu} - \mu)$ large.
- If such $\nu$ exists, $\text{Var}(\nu \cdot X)$ large.
- Determine if there is a 2k-sparse $\nu$ with $|\nu|_2 = 1$ and $\nu^T \text{Cov}(X) \nu$ large.
  - If not, return $\hat{\mu}$
  - If so, filter on $\nu \cdot X$ and repeat
Basic Algorithm

Sparse Robust Mean Estimation:

- \( \hat{\mu}_k \approx \mu \) unless there is a 2\( k \)-sparse \(|v|_2 = 1 \) with \( v \cdot (\hat{\mu} - \mu) \) large.
- If such \( v \) exists, \( \text{Var}(v \cdot X) \) large.
- Determine if there is a 2\( k \)-sparse \( v \) with \(|v|_2 = 1 \) and \( v^T \text{Cov}(X) v \) large.
  - If not, return \( \hat{\mu}_k \)
  - If so, filter on \( v \cdot X \) and repeat
Sample Complexity

We need our good set of points to have:

- $\nu \cdot (\hat{\mu} - \mu)$ small for $\nu$ 2k-sparse.
- $\text{Var}(\nu \cdot X) \approx 1$ for $\nu$ 2k-sparse.
- $\nu \cdot X$ to have appropriate tails for $\nu$ 2k-sparse.
Sample Complexity

We need our good set of points to have:

- $v \cdot (\hat{\mu} - \mu)$ small for $v$ $2k$-sparse.
- $\text{Var}(v \cdot X) \approx 1$ for $v$ $2k$-sparse.
- $v \cdot X$ to have appropriate tails for $v$ $2k$-sparse.

Can cover $2k$-sparse vectors with cover of size $\binom{d}{2k}2^{O(k)}$. Need $O(k \log(d)/\epsilon^2)$ samples.
Problem

To find directions of large variance need to solve:

$$\sup_{\|v\|_2 \leq 1, \|v\|_0 \leq 2k} v^T M v$$

with $M = \text{Cov}(X)$. 
Problem

To find directions of large variance need to solve:

\[
\sup_{\|v\|_2 \leq 1, \|v\|_0 \leq 2k} v^T M v
\]

with \( M = \text{Cov}(X) \).

This is NP-Hard in general!
Convex Relaxation

Instead solve a relaxation.

- If $\nu$ is $2k$-sparse, $|\nu|_1 \leq \sqrt{2k}$.
- $|\nu \nu^T|_1 \leq 2k$ and $\nu \nu^T \cdot \text{Cov}(X)$ large.
Convex Relaxation

Instead solve a relaxation.

- If $v$ is $2k$-sparse, $|v|_1 \leq \sqrt{2k}$.
- $|vv^T|_1 \leq 2k$ and $vv^T \cdot \text{Cov}(X)$ large.

Solve

$$\sup_{H \geq 0, |H|_1 \leq 2k, \text{tr}(H) = 1} H \cdot \text{Cov}(X).$$  \hspace{1cm} (1)$$

If solution is small, $\hat{\mu}_k \approx \mu$.

If not, filter?
Convex Relaxation

Instead solve a relaxation.

- If $v$ is $2k$-sparse, $|v|_1 \leq \sqrt{2k}$.
- $|vv^T|_1 \leq 2k$ and $vv^T \cdot \text{Cov}(X)$ large.

Solve

$$\sup_{H \geq 0, |H|_1 \leq 2k, \text{tr}(H) = 1} H \cdot \text{Cov}(X).$$  \hspace{1cm} (1)

- If solution is small, $\hat{\mu}_k \approx \mu$.
- If not, filter?
Good Samples

Assuming that we took $\Omega(k^2 \log(d)/\epsilon^2)$ samples, with high probability each entry of $\hat{\Sigma} - \Sigma$ is $O(\epsilon/k)$.
Good Samples

Assuming that we took $\Omega(k^2 \log(d)/\epsilon^2)$ samples, with high probability each entry of $\hat{\Sigma} - \Sigma$ is $O(\epsilon/k)$.

If so, for any $H$ with $|H|_1 \leq 2k$, and $\text{tr}(H) = 1$

$$H \cdot \hat{\Sigma} = H \cdot \Sigma + O(\epsilon) = 1 + O(\epsilon).$$
Assuming that we took $\Omega(k^2 \log(d)/\epsilon^2)$ samples, with high probability each entry of $\hat{\Sigma} - \Sigma$ is $O(\epsilon/k)$.

If so, for any $H$ with $|H|_1 \leq 2k$, and $\text{tr}(H) = 1$

$$H \cdot \hat{\Sigma} = H \cdot \Sigma + O(\epsilon) = 1 + O(\epsilon).$$

- If $H \cdot \hat{\Sigma}$ is much larger, discrepancy due to bad samples.
- Filter entries where $(x - \hat{\mu})H(x - \hat{\mu})$ is large (or add to convex program).
Upshot

Have an algorithm where if $\mu$ is known to be $k$-sparse, learn $\mu$ to error $\tilde{O}(\epsilon)$ with $\epsilon$ adversarial error with $O(k^2 \log(d)/\epsilon^2)$ samples in polynomial time.
Further Extensions

[BDLS] Also give robust sparse estimation algorithms for:

- Estimating $\Sigma = I + \Omega$ when $|\Omega|_0 \leq k$.
- Estimating $\Sigma = I + \rho \nu \nu^T$ when $\nu$ is $k$-sparse.
- Linear regressions $y \approx x \cdot \beta$ when $\beta$ is $k$-sparse.
Further Extensions

[BDLS] Also give robust sparse estimation algorithms for:

- Estimating $\Sigma = I + \Omega$ when $|\Omega|_0 \leq k$.
- Estimating $\Sigma = I + \rho vv^T$ when $v$ is $k$-sparse.
- Linear regressions $y \approx x \cdot \beta$ when $\beta$ is $k$-sparse.

Recent work by [Diakonikolas–Kane–Karmalkar–Price] does much of this using spectral techniques instead of convex programs.