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 μ to error ϵ.
- What if extra information is known about μ? Can we do better?
  - In particular, what if μ is known to be sparse?

Sparse Mean Estimation

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

Robust Sparse Mean Estimation

What if we want to do this robustly? Can we learn μ 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 μ 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 $|v|_2 = 1$ with $v \cdot (\hat{\mu} - \mu)$ large.
- If such v exists, $\mathrm{Var}(v \cdot X)$ large.
- Determine if there is a v with $|v|_2 = 1$ and $v^T \mathrm{Cov}(X) v$ 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 $|v|_2 = 1$ with $v \cdot (\hat{\mu} - \mu)$ large.
- If such v exists, $\mathrm{Var}(v \cdot X)$ large.
- Determine if there is a v with $|v|_2 = 1$ and $v^T \mathrm{Cov}(X) v$ large.
  - If not, return $\hat{\mu}$
  - If so, filter on $v \cdot X$ and repeat

Basic Algorithm

Sparse Robust Mean Estimation:

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

Basic Algorithm

Sparse Robust Mean Estimation:

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

Basic Algorithm

Sparse Robust Mean Estimation:

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

Basic Algorithm

Sparse Robust Mean Estimation:

- $\hat{\mu}_k \approx \mu$ unless there is a 2k-sparse $|v|_2 = 1$ with $v \cdot (\hat{\mu} - \mu)$ large.
- If such v exists, $\mathrm{Var}(v \cdot X)$ large.
- Determine if there is a 2k-sparse v with $|v|_2 = 1$ and $v^T \mathrm{Cov}(X) v$ 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:

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

Sample Complexity

We need our good set of points to have:

- $v \cdot (\hat{\mu} - \mu)$ small for v $2k$-sparse.
- $\mathrm{Var}(v \cdot X) \approx 1$ 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 $M = \mathrm{Cov}(X)$.

Problem

To find directions of large variance need to solve:

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

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

This is NP-Hard in general!

Convex Relaxation

Instead solve a relaxation.

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

Convex Relaxation

Instead solve a relaxation.

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

Solve

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

Convex Relaxation

Instead solve a relaxation.

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

Solve

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

- 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 $\mathrm{tr}(H) = 1$

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

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 $\mathrm{tr}(H) = 1$

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

- 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 μ is known to be k-sparse, learn μ to error $\tilde{O}(\epsilon)$ with ϵ 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 $\Sigma = I + \rho v v^T$ when v is k-sparse.
- Linear regressions $y \approx x \cdot \beta$ when β is k-sparse.

Further Extensions

[BDLS] Also give robust sparse estimation algorithms for:

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

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