Being Robust (in High Dimensions) Can Be Practical

Paper by Diakonikolas et. al.
ICML 2017

Anant Gupta
agupta225@wisc.edu

Surya Teja Chavali
chavali2@wisc.edu

Muni Sreenivas Pydi
pydi@wisc.edu

Shantanu Gupta
sgupta226@wisc.edu

Introduction

Problem Definition: Robust Mean and Covariance estimation

Given a polynomial number of samples from a high-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$, where an adversary has arbitrarily corrupted an ε-fraction, find a set of parameters $\mathcal{N}'(\widehat{\mu}, \widehat{\Sigma})$ that satisfy $d_{TV}(\mathcal{N}, \mathcal{N}') \leq \widetilde{O}(\varepsilon)$.

Recap: Low Dimension with noise

For low dimensions, median is robust, efficient, computationally tractable.

As we saw in class, median has:

1. Minimax asymptotic bias $O(\varepsilon)$
2. Asymptotic variance $O(\frac{1}{n}) \implies$ sample complexity $\frac{1}{\varepsilon^2}$
3. Computational complexity $O(n)$

Recap: High Dimension with no noise

Sample mean is asymptotically normal:

1. Asymptotic variance $O\left(\frac{I}{n}\right) \implies$ sample complexity $O(d)$
2. Computational complexity $O(nd)$ (Polynomial in n and d)

Error guarantee increases with dimension as $\sqrt{d}$

Goal: High Dimension with noise

We want an estimator that is:

1. Robust: Error bound $\tilde{O}(\varepsilon)$
2. Sample efficient: Sample complexity $\tilde{O}(\frac{d}{\varepsilon^2})$

The Tukey median (1960) achieves these goals.

Goal: High Dimension with noise

We want an estimator that is:

1. Robust: Error bound $\tilde{O}(\varepsilon)$
2. Sample efficient: Sample complexity $\tilde{O}(\frac{d}{\varepsilon^2})$

The Tukey median (1960) achieves these goals.

But it has computational complexity $O\left(n^{d-1} + n \log n\right)$ - **exponential** in d...

Other approaches that don't work well in high dimensions

Generalizations of the median to higher dimensions:

1. Coordinate-wise median
2. Geometric median

Both of these have error bounds $O(\varepsilon\sqrt{\mathbf{d}})$.

Other approaches that don't work well in high dimensions

Generalizations of the median to higher dimensions:

1. Coordinate-wise median
2. Geometric median

Both of these have error bounds $O(\varepsilon\sqrt{\mathbf{d}})$. (*curse of dimensionality*)

Related Work

Robustly learn μ^* given ε-corrupted from $\mathcal{N}(\mu^*, I)$: Error vs computational complexity trade-off:

Algorithm	Error Guarantee	Poly-Time?
Tukey Median	$O(\varepsilon)$	No
Tournament	$O(\varepsilon)$	No
Geometric Median	$O(\varepsilon\sqrt{d})$	Yes
Pruning	$O(\varepsilon\sqrt{d})$	Yes
LRV'16	$O(\varepsilon\sqrt{\log d})$	Yes

Related Work

Robustly learn μ^* given ε-corrupted from $\mathcal{N}(\mu^*, I)$: Error vs computational complexity trade-off:

Algorithm	Error Guarantee	Poly-Time?
Tukey Median	$O(\varepsilon)$	No
Tournament	$O(\varepsilon)$	No
Geometric Median	$O(\varepsilon\sqrt{d})$	Yes
Pruning	$O(\varepsilon\sqrt{d})$	Yes
LRV'16	$O(\varepsilon\sqrt{\log d})$	Yes
Filter	$O(\varepsilon\sqrt{\log(1/\varepsilon)})$	Yes

All these algorithms are sample efficient.

Contamination Model

The paper considers the following contamination model:

$$X_1, X_2, \ldots, X_m \overset{\text{iid}}{\sim} D, D \in \mathcal{D}$$

Contamination Model

The paper considers the following contamination model:

$$X_1, X_2, ..., X_m \overset{\text{iid}}{\sim} D, D \in \mathcal{D}$$

Adversary changes arbitrarily an ε-fraction

$$\downarrow$$

$$Y_1, Y_2, ..., Y_m$$

Contamination Model: contd

Generalization of Huber's model

1. Subsumes Huber

Contamination Model: contd

Generalization of Huber's model

1. Subsumes Huber
2. Allows both insertions and deletions

Contamination Model: contd

Generalization of Huber's model

1. Subsumes Huber
2. Allows both insertions and deletions
3. Adversary allowed to inspect data, i.e. corrupted data is not i.i.d.

Main Result: Mean estimation for Sub-Gaussian Distribution

Theorem (3.1)

If

- *G Sub-Gaussian on $\mathbb{R}^d$, $\nu = \Theta(1)$, mean μ^G, covariance I*
- *S is an ε-corrupted set of samples with $|S| = \tilde{\Omega}(d/\varepsilon^2)$*

Then there exists an efficient algorithm that outputs $\widehat{\mu}$ with prob. $1-\tau$ s.t.

$$\|\widehat{\mu} - \mu^G\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)}).$$

Main Result: Mean estimation for Bounded Second Moment

Theorem (3.2)

If

- P *distribution on* $\mathbb{R}^d$, *mean* μ^P, *covariance* $\Sigma_P \preceq \sigma^2 I$
- S *is an* ε*-corrupted set of samples with* $|S| = \tilde{\Theta}(d/\varepsilon)$

Then there exists an efficient algorithm that outputs $\widehat{\mu}$ *with prob.* $1 - \tau$ *s.t.*

$$\|\widehat{\mu} - \mu^P\|_2 \leq O(\sqrt{\varepsilon}\sigma).$$

Main Result: Covariance Estimation

Theorem (3.3)

If

- $G \sim \mathcal{N}(0, \Sigma)$ *in* d *dimensions*
- S *is an* ε*-corrupted set of samples with* $|S| = \tilde{\Omega}(d^2/\varepsilon^2)$

Then there exists an efficient algorithm that outputs $\widehat{\Sigma}$ *with prob.* $1 - \tau$ *s.t.*

$$\|I - \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}\|_F = O(\varepsilon \log(1/\varepsilon)).$$

A Summary of the Results

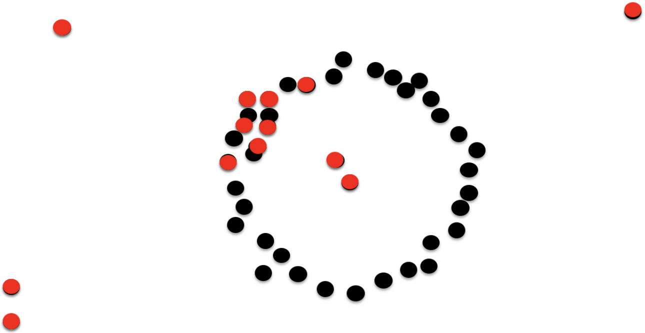
All results with probability $1-\tau$:

	Theorem 3.1	Theorem 3.2	Theorem 3.3
Distribution	Sub-Gaussian Known Cov.	Bounded Covariance	Gaussian
Target	μ_G	μ_P	Σ
Error	$\lVert\widehat{\mu}-\mu^G\rVert_2$	$\lVert\widehat{\mu}-\mu^P\rVert_2$	$\lVert I-\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}\rVert_F$
Error Bound	$O(\varepsilon\sqrt{\log(1/\varepsilon)})$	$O(\sqrt{\varepsilon}\sigma)$	$O\left(\varepsilon\log(1/\varepsilon)\right)$
#(samples)	$\tilde{\Omega}(d/\varepsilon^2)$	$\tilde{\Theta}(d/\varepsilon)$	$\tilde{\Omega}(d^2/\varepsilon^2)$

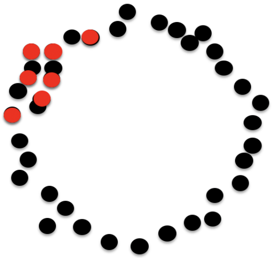
Table: Summarizing the error bounds and sample complexity of the three proposed algorithms

Algorithm

Corrupted Data



Phase-1 Algorithm: Naive Pruning



NaivePrune: Getting a (Nearly) Good Set

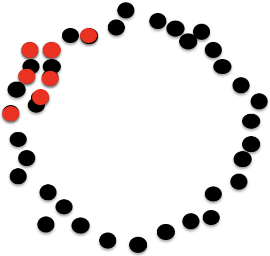
Algorithm 1 Naive Pruning

1: **function** $\textsc{NaivePrune}(X_1, \ldots, X_N)$
2: For $i, j = 1, \ldots, N$, define $\delta_{i,j} = \|X_i - X_j\|_2$.
3: **for** $i = 1, \ldots, j$ **do**
4: Let $A_i = \{j \in [N] : \delta_{i,j} > \Omega(\sqrt{d \log(N/\tau)})\}$
5: **if** $|A_i| > 2\varepsilon N$ **then**
6: Remove X_i from the set.
7: **return** the pruned set of samples.

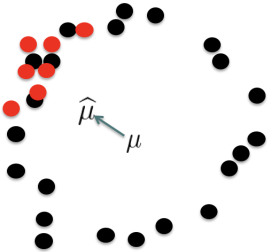
Fact

With high probability, $\textsc{NaivePrune}$ *removes no uncorrupted points, and for all* X_i *that remain,* $\|X_i - \mu\|_2 \leq O\left(\sqrt{d \log(N/\tau)}\right)$.

We're still not in good shape!



We're still not in good shape!



The FILTER (Meta-) Algorithm

Algorithm 2 Filter-based mean estimation

1: **Input:** ε-corrupted sample set S, $\mathrm{Thres}(\varepsilon)$, $\mathrm{Tail}(T, d, \varepsilon, \delta, \tau)$, $\delta(\varepsilon, s)$
2: Compute the sample mean $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$
3: Compute the sample covariance matrix Σ
4: Compute approximations for the largest absolute eigenvalue of Σ, $\lambda^* := \|\Sigma\|_2$, and the associated unit eigenvector v^*.
5: **if** $\|\Sigma\|_2 \leq \mathrm{Thres}(\varepsilon)$ **then**
6: **return** $\mu^{S'}$.
7: Let $\delta = \delta(\varepsilon, \|\Sigma\|_2)$.
8: Find $T > 0$ such that

$$\Pr_{X \in_u S'}\left[|v^* \cdot (X - \mu^{S'})| > T + \delta\right] > \mathrm{Tail}(T, d, \varepsilon, \delta, \tau).$$

9: **return** $\{x \in S' : |v^* \cdot (x - \mu^{S'})| \leq T + \delta\}$.

The Sub-Gaussian Case: Good Sets

Definition (Good Sets)

If G is a Sub-Gaussian distribution on $\mathbb{R}^d$ with parameter $\nu = \Theta(1)$, covariance I, and S is a sample drawn from G, then S is said to be a "good" set, if

(i) $\|x - \mu^G\|_2 \leq O(\sqrt{d \log(|S|/\tau)})$ for all $x \in S$.

(ii) $\forall v, T$, such that $\|v\|_2 = 1$ and $T \in \mathbb{R}$,

$$\left| \Pr_{X \in_u S}[v \cdot (x - \mu^G) \geq T] - \Pr_{X \sim G}[v \cdot (x - \mu^G) \geq T] \right| \leq 8 \exp(-T^2/2\nu) + 8 \frac{\varepsilon}{\tilde{T}^2} .$$

(iii) $\|\mu^S - \mu^G\|_2 \leq \varepsilon$.

(iv) $\|M_S - I\|_2 \leq \varepsilon$.

Instantiating the Sub-Gaussian Case

1. **Sub-Gaussian(ν) Distribution, $\Sigma = I$.**
   - $\mathrm{Thres}(\varepsilon) = O(\varepsilon \log 1/\varepsilon)$
     - Comes from deleted points

Instantiating the Sub-Gaussian Case

1. **Sub-Gaussian(ν) Distribution, $\Sigma = I$.**
   - $\mathrm{Thres}(\varepsilon) = O(\varepsilon \log 1/\varepsilon)$
     - Comes from deleted points
   - $\mathrm{Tail}(T, d, \varepsilon, \delta, \tau) = 8\exp(-T^2/2\nu) + 8\frac{\varepsilon}{\tilde{T}^2}$
     1. Sub-Gaussian

Instantiating the Sub-Gaussian Case

1. **Sub-Gaussian(ν) Distribution, $\Sigma = I$.**
   - $\mathrm{Thres}(\varepsilon) = O(\varepsilon \log 1/\varepsilon)$
     - Comes from deleted points
   - $\mathrm{Tail}(T, d, \varepsilon, \delta, \tau) = 8 \exp(-T^2/2\nu) + 8 \frac{\varepsilon}{\tilde{T}^2}$
     1. Sub-Gaussian
     2. Translate bound from true distribution to empirical

Instantiating the Sub-Gaussian Case

1. **Sub-Gaussian(ν) Distribution, $\Sigma = I$.**
   - $\mathrm{Thres}(\varepsilon) = O(\varepsilon \log 1/\varepsilon)$
     - Comes from deleted points
   - $\mathrm{Tail}(T, d, \varepsilon, \delta, \tau) = 8\exp(-T^2/2\nu) + 8\frac{\varepsilon}{\tilde{T}^2}$
     1. Sub-Gaussian
     2. Translate bound from true distribution to empirical
   - $\delta(\varepsilon, s) = 3\sqrt{\varepsilon(s-1)}$
     - Captures the error in sample mean: $\left\langle v^*, \mu^G - \mu^{S'} \right\rangle$

Instantiations

1. **Sub-Gaussian(ν) Distribution, $\Sigma = I$.**
   - $\mathrm{Thres}(\varepsilon) = O(\varepsilon \log 1/\varepsilon)$
   - $\mathrm{Tail}(T, d, \varepsilon, \delta, \tau) = 8\exp(-T^2/2\nu) + 8\frac{\varepsilon}{\tilde{T}^2}$
   - $\delta(\varepsilon, s) = 3\sqrt{\varepsilon(s-1)}$
2. **Bounded Covariance: $\Sigma \preceq I$.**
   - $\mathrm{Thres}(\varepsilon) = \Theta(1)$
   - $\mathrm{Tail} = Z \max\{|v^* \cdot x - \mu^S| : x \in S\}$
     - Random Threshold will throw out *more* bad points than good.
     - $Z \in [0, 1]$ with PDF $p_Z(z) = 2z$.
   - $\delta = 0$

The Algorithm: Putting it all together

- Firstly, apply NAIVEPRUNE to ensure that the set becomes '2ε-close to good' - outliers that are too far away are removed.
- Iteratively filter out bad points using FILTER to reach a good set with high probability.
- Return the sample mean of the good set.

Proof: The Gory Details

The End Goal

Theorem (A.3)

If

- *G Sub-Gaussian on $\mathbb{R}^d$, $\nu = \Theta(1)$, mean μ^G, covariance I*
- $|S| = \Omega((d/\varepsilon^2)\,\mathrm{poly}\log(d/\varepsilon\tau))$

Then with prob. $1 - \tau$,

$$\|\widehat{\mu} - \mu^G\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)}).$$

Definitions

1. Sub-Gaussian: (tails decay faster than Gaussian)
   - $\Pr_{X\sim P}\left[|v\cdot(X-\mu)|\geq T\right]\leq \exp(-T^2/2\nu)$, $(\|v\|_2=1)$
2. Let S be the set given as input, and S' be the output.
3. Define $\Delta(S,S')=\frac{|S\setminus S'|+|S'\setminus S|}{|S|}$
4. By definition, $\exists$ sets E('entering') and L('leaving') such that $S'=(S\setminus L)\cup E$
5. Note that $\Delta(S,S')=\frac{|E|+|L|}{|S|}$

Definitions

- $\mu^S = \frac{1}{|S|} \sum_{X \in S} X$ Clean Sample mean
- $\mu^{S'} = \frac{1}{|S'|} \sum_{X \in S'} X$ Sample mean
- $\Sigma = \frac{1}{|S'|} \sum_{X \in S'} (X - \mu^{S'})(X - \mu^{S'})^T$ Sample covariance
- $M_{S'} = \frac{1}{|S'|} \sum_{X \in S'} (X - \mu^G)(X - \mu^G)^T$ Modified sample covariance

Definitions

- $\mu^S = \frac{1}{|S|} \sum_{X \in S} X$
- $\mu^L = \frac{1}{|S|} \sum_{X \in L} X$
- $\mu^E = \frac{1}{|S|} \sum_{X \in E} X$
- $M_S = \frac{1}{|S|} \sum_{X \in S'} [(X - \mu^G)(X - \mu^G)^T]$,
- $M_L = \frac{1}{|L|} \sum_{X \in L} [(X - \mu^G)(X - \mu^G)^T]$,
- $M_E = \frac{1}{|E|} \sum_{X \in E} [(X - \mu^G)(X - \mu^G)^T]$.

The Good Set

Lemma (A.5)

If

- G *sub-gaussian on* $\mathbb{R}^d$, $\nu = \Theta(1)$, *covariance* I
- $|S| = \Omega((d/\varepsilon^2) \operatorname{poly}\log(d/\varepsilon\tau))$

Then with prob. $1-\tau$, S *is a "good" set, i.e.*

(i) $\|x - \mu^G\|_2 \leq O(\sqrt{d \log(|S|/\tau)})$ *for all* $x \in S$.

(ii) $\left| \Pr_{X \in_u S}[v \cdot (x - \mu^G) \geq T] - \Pr_{X \sim G}[v \cdot (x - \mu^G) \geq T] \right| \leq \tilde{O}(\varepsilon)$.

(iii) $\|\mu^S - \mu^G\|_2 \leq \varepsilon$.

(iv) $\|M_S - I\|_2 \leq \varepsilon$.

If Algorithm Filter works, then Theorem (A.3) is true

Proposition (A.7)

If

- *G sub-gaussian on $\mathbb{R}^d$, $\nu = \Theta(1)$, covariance I*
- *S is (ε, τ)-good set; $\Delta(S, S') \leq 2\varepsilon$*
  S is uncorrupted, S' is ε-uncorrupted
- *For any $x, y \in S'$, $\|x - y\|_2 \leq O(\sqrt{d \log(d/\varepsilon\tau)})$*
  Consequence of NaivePrune

Then, the algorithm Filter returns one of these:

(i) *A mean vector $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu^G\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$.*

(ii) *A multiset $S'' \subseteq S'$ such that $\Delta(S, S'') \leq \Delta(S, S') - \varepsilon/\alpha$, where $\alpha \stackrel{\text{def}}{=} d \log(d/\varepsilon\tau) \log(d \log(\frac{d}{\varepsilon\tau}))$.*

Algorithm Filter

Algorithm 3 Filter-Sub-Gaussian-Unknown-Mean (S', ε, τ)

1: **Input:** S' such that there exists (ε, τ)-good S with $\Delta(S, S') \leq 2\varepsilon$

2: **Output:** S'' or $\widehat{\mu}$ satisfying Proposition (A.7)

3: Compute $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$ and $\Sigma = \mathbb{E}_{X \in_u S'}\left[(X - \mu^{S'})(X - \mu^{S'})^T\right]$

4: Compute the largest absolute eigenvalue of $\Sigma - I$, $\lambda^* := \|\Sigma - I\|_2$, and the associated unit eigenvector v^*.

5: **if** $\|\Sigma - I\|_2 \leq O(\varepsilon \log(1/\varepsilon))$, **then return** $\mu^{S'}$.

6: Let $\delta := 3\sqrt{\varepsilon \|\Sigma - I\|_2}$. Find $T > 0$ such that

$$\Pr_{X \in_u S'}\left[|v^* \cdot (X - \mu^{S'})| > T + \delta\right] > 8\exp(-T^2/2\nu) + 8\frac{\varepsilon}{T^2 \log\left(d \log(\frac{d}{\varepsilon\tau})\right)}.$$

7: **return** the multiset $S'' = \{x \in S' : |v^* \cdot (x - \mu^{S'})| \leq T + \delta\}$.

Proof of Correctness of Algorithm FILTER

Need to prove:

1. Small spectral norm: If $\|\Sigma - I\|_2 \leq O\left(\varepsilon \log\left(\frac{1}{\varepsilon}\right)\right)$, then

$$\|\mu^{S'} - \mu^{G}\|_2 = O\left(\varepsilon\sqrt{\log\left(\frac{1}{\varepsilon}\right)}\right).$$

2. Large spectral norm: If $\|\Sigma - I\|_2 > \Omega\left(\varepsilon \log\left(\frac{1}{\varepsilon}\right)\right)$, then
   - $\exists$ a threshold T that is used for filtering, such that

$$\Pr_{X \in_u S'}\left[|v^* \cdot (X - \mu^{S'})| > T + \delta\right] > 8e^{\frac{-T^2}{2\nu}} + 8\frac{\varepsilon}{\tilde{T}^2}.$$

   - The algorithm makes progress, i.e. S'' satisfies

$$\Delta(S, S'') \leq \Delta(S, S') - \varepsilon/\alpha$$

Proof of Correctness of Algorithm FILTER

Key Result:

$$\Sigma - I \approx (|E|/|S'|)M_E$$

Intuition: The errors approximately align in the direction of leading eigenvector of $\Sigma - I$.

Proof: By definition,

$$\Sigma - I = \underbrace{(M_{S'} - I)}_{\substack{\approx \frac{|E|}{|S'|} M_E \\ (A)}} - \underbrace{\underbrace{(\mu^{S'} - \mu^{G})}_{\substack{\approx \frac{|E|}{|S'|}(\mu^E - \mu^G) \\ (B)}} (\mu^{S'} - \mu^{G})^T}_{\substack{\|\mu^E - \mu^G\|_2^2 \leq \|M_E\|_2 \\ \text{(Matrix-norm identity)}}}$$

$$\therefore \Sigma - I = (|E|/|S'|)M_E + O(\varepsilon \log(1/\varepsilon)) + O(\varepsilon^2 \|M_E\|_2)$$

Proof of Correctness of Algorithm FILTER

Proof of matrix-norm identity: $\|M_E\|_2 \geq \|\mu^E - \mu^G\|_2^2$

$$\sum (x - \mu^G)(x - \mu^G)^T$$
$$= \sum (x - \mu^E)(x - \mu^E)^T + \sum (\mu^E - \mu^G)(\mu^E - \mu^G)^T$$

Proof of Correctness of Algorithm FILTER

Proof of matrix-norm identity: $\|M_E\|_2 \geq \|\mu^E - \mu^G\|_2^2$

$$\sum (x - \mu^G)(x - \mu^G)^T$$
$$= \sum (x - \mu^E)(x - \mu^E)^T + \sum (\mu^E - \mu^G)(\mu^E - \mu^G)^T$$

$$M_E = \underbrace{\Sigma_E}_{\succeq 0} + \underbrace{(\mu^E - \mu^G)(\mu^E - \mu^G)^T}_{\succeq 0}$$

$$\|M_E\|_2 \geq \|\mu^E - \mu^G\|_2^2$$

Proof of Correctness of Algorithm FILTER

Proof of (A): $M_{S'} - I = \frac{|E|}{|S'|} M_E + O\left(\varepsilon \log \frac{1}{\varepsilon}\right)$

Proof of Correctness of Algorithm FILTER

Proof of (A): $M_{S'} - I = \frac{|E|}{|S'|} M_E + O\left(\varepsilon \log \frac{1}{\varepsilon}\right)$
Let $\tilde{x} = x - \mu^G$. Recall that $S' = (S \backslash L) \cup E$.

Proof of Correctness of Algorithm FILTER

Proof of (A): $M_{S'} - I = \frac{|E|}{|S'|} M_E + O\left(\varepsilon \log \frac{1}{\varepsilon}\right)$

Let $\tilde{x} = x - \mu^G$. Recall that $S' = (S \backslash L) \cup E$.

$$\sum_{X \in S'} \tilde{x}\tilde{x}^T = \sum_{X \in S} \tilde{x}\tilde{x}^T - \sum_{X \in L} \tilde{x}\tilde{x}^T + \sum_{X \in E} \tilde{x}\tilde{x}^T$$

$$|S'| M_{S'} = |S| M_S - |L| M_L + |E| M_E$$

$$M_{S'} = \underbrace{\frac{|S|}{|S'|}}_{1} \underbrace{M_S}_{I + O(\varepsilon)} - \underbrace{\frac{|L|}{|S'|}}_{\varepsilon} \underbrace{M_L}_{O(\log \frac{1}{\varepsilon})} + \frac{|E|}{|S'|} M_E$$

$$\therefore M_{S'} - I = \varepsilon M_E + O\left(\varepsilon \log \frac{1}{\varepsilon}\right)$$

Proof of Correctness of Algorithm FILTER

Proof of (B): $(\mu^{S'} - \mu^{G}) = \frac{|E|}{|S'|}(\mu^{E} - \mu^{G}) + O\left(\varepsilon \log \frac{1}{\varepsilon}\right)$

Proof of Correctness of Algorithm FILTER

Proof of (B): $(\mu^{S'} - \mu^{G}) = \frac{|E|}{|S'|}(\mu^{E} - \mu^{G}) + O\left(\varepsilon \log \frac{1}{\varepsilon}\right)$

Recall that $\tilde{x} = x - \mu^{G}$ and $S' = (S \backslash L) \cup E$.

Proof of Correctness of Algorithm FILTER

Proof of (B): $(\mu^{S'} - \mu^{G}) = \frac{|E|}{|S'|}(\mu^{E} - \mu^{G}) + O\left(\varepsilon \log \frac{1}{\varepsilon}\right)$
Recall that $\tilde{x} = x - \mu^{G}$ and $S' = (S \backslash L) \cup E$.

$$S' = (S \backslash L) \cup E$$

$$\sum_{X \in S'} \tilde{x} = \sum_{X \in S} \tilde{x} - \sum_{X \in L} \tilde{x} + \sum_{X \in E} \tilde{x}$$

$$|S'|(\mu^{S'} - \mu^{G}) = |S|(\mu^{S} - \mu^{G}) - |L|(\mu^{L} - \mu^{G}) + |E|(\mu^{E} - \mu^{G})$$

$$(\mu^{S'} - \mu^{G}) = \underbrace{\frac{|S|}{|S'|}}_{1} \underbrace{(\mu^{S} - \mu^{G})}_{O(\varepsilon)} - \underbrace{\frac{|L|}{|S'|}}_{\varepsilon} \underbrace{(\mu^{L} - \mu^{G})}_{\|\mu^{L} - \mu^{G}\|_2^2 \leq \|M_L\|_2} + \frac{|E|}{|S'|}(\mu^{E} - \mu^{G})$$

$$\therefore (\mu^{S'} - \mu^{G}) = \varepsilon(\mu^{E} - \mu^{G}) + O\left(\varepsilon \sqrt{\log(1/\varepsilon)}\right)$$

Proof of Correctness of Algorithm Filter

Combining (A) and (B), we get:

$$\Sigma - I \quad = \varepsilon M_E + O(\varepsilon \log(1/\varepsilon))$$

(Key Result)

Proof of Correctness of Alg. FILTER: Small Spectral Norm

Step 5. **if** $\|\Sigma - I\|_2 \leq O(\varepsilon \log(1/\varepsilon))$, **return** $\mu^{S'}$.

Proof of Correctness of Alg. FILTER: Small Spectral Norm

Step 5. **if** $\|\Sigma - I\|_2 \leq O(\varepsilon \log(1/\varepsilon))$, **return** $\mu^{S'}$.

To Prove: $\|\mu^{S'} - \mu^G\|_2 \leq O\left(\varepsilon\sqrt{\log(1/\varepsilon)}\right)$

Proof of Correctness of Alg. FILTER: Small Spectral Norm

Step 5. **if** $\|\Sigma - I\|_2 \leq O(\varepsilon \log(1/\varepsilon))$, **return** $\mu^{S'}$.

To Prove: $\|\mu^{S'} - \mu^{G}\|_2 \leq O\left(\varepsilon\sqrt{\log\left(1/\varepsilon\right)}\right)$

Proof:

$$\|\mu^{S'} - \mu^{G}\|_2$$

$$\leq \varepsilon\|\mu^{E} - \mu^{G}\|_2 + O\left(\varepsilon\sqrt{\log\left(1/\varepsilon\right)}\right) \text{ Follows from (B)}$$

$$\leq \varepsilon\sqrt{\|M_E\|_2} + O\left(\varepsilon\sqrt{\log\left(1/\varepsilon\right)}\right), \text{ from matrix-norm identity}$$

$$= \sqrt{\varepsilon\|\Sigma - I\|_2} + O\left(\varepsilon\sqrt{\log\left(1/\varepsilon\right)}\right), \text{ from key result}$$

$$= O\left(\varepsilon\sqrt{\log\left(1/\varepsilon\right)}\right)$$

Proof of Correctness of Alg. FILTER: Large Spectral Norm

Step 6. **if** $\|\Sigma - I\|_2 \geq \Omega(\varepsilon \log(\frac{1}{\varepsilon}))$, **then**:
Find $T > 0$ such that following tail bound is violated

$$\Pr_{X \in_u S'} \left[|v^* \cdot (X - \mu^{S'})| > T + \delta \right] > \mathrm{Tail}(T, d, \varepsilon, \tau).$$

(i.e. at least $\mathrm{Tail}(T, d, \varepsilon, \tau)$-fraction of points fall outside threshold)

Step 7. **reject** $x \in S'$ that fall outside threshold

Proof of Correctness of Alg. FILTER: Large Spectral Norm

if $\|\Sigma - I\|_2 \geq \Omega(\varepsilon \log(\frac{1}{\varepsilon}))$, **return** S''.

To Prove:

Proof of Correctness of Alg. FILTER: Large Spectral Norm

if $\|\Sigma - I\|_2 \geq \Omega(\varepsilon \log(\frac{1}{\varepsilon}))$, **return** S''.
To Prove:

1. A "violation" threshold T for Step 6 exists
$\implies$ allows "many" total points ($\Omega(\varepsilon/\alpha)$) to be rejected

Proof of Correctness of Alg. FILTER: Large Spectral Norm

if $\|\Sigma - I\|_2 \geq \Omega(\varepsilon \log(\frac{1}{\varepsilon}))$, **return** S''.
To Prove:

1. A "violation" threshold T for Step 6 exists
   $\implies$ allows "many" total points ($\Omega(\varepsilon/\alpha)$) to be rejected
2. Using T, Filter rejects "few" good points $O(\varepsilon/\alpha)$

Proof of Correctness of Alg. FILTER: Large Spectral Norm

if $\|\Sigma - I\|_2 \geq \Omega(\varepsilon \log(\frac{1}{\varepsilon}))$, **return** S''.
To Prove:

1. A "violation" threshold T for Step 6 exists
   $\implies$ allows "many" total points ($\Omega(\varepsilon/\alpha)$) to be rejected
2. Using T, Filter rejects "few" good points $O(\varepsilon/\alpha)$
3. Filter rejects more bad points than good and makes progress:

$$\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/\alpha$$

where $S'' = S' \setminus \{\text{rejected points}\}$,
$\alpha = d \log(d/\varepsilon\tau) \log(d \log(\frac{d}{\varepsilon\tau}))$

Proof of Correctness of Alg. FILTER: Large Spectral Norm

1 A threshold T exists

Proof ouline:

1. Suppose not.

Proof of Correctness of Alg. FILTER: Large Spectral Norm

1 A threshold T exists

Proof ouline:

1. Suppose not.
2. Then S' satisfies the sub-gaussian tail bound:

$$\Pr_{X \in_u S'} \left[|v^* \cdot (X - \mu^{S'})| > T + \delta/2 \right] \leq 8 \exp(-T^2/2\nu) + 8 \frac{\varepsilon}{\tilde{T}^2} \,.$$

Proof of Correctness of Alg. Filter: Large Spectral Norm

1 A threshold T exists

Proof ouline:

1. Suppose not.
2. Then S' satisfies the sub-gaussian tail bound:

$$\Pr_{X \in_u S'} \left[|v^* \cdot (X - \mu^{S'})| > T + \delta/2 \right] \leq 8 \exp(-T^2/2\nu) + 8 \frac{\varepsilon}{\tilde{T}^2} .$$

3. Then $\|\Sigma - I\|_2 \leq O(\varepsilon \log(\frac{1}{\varepsilon}))$. Contradiction.

Proof of Correctness of Alg. FILTER: Large Spectral Norm

1 A threshold T exists

Proof ouline:

1. Suppose not.
2. Then S' satisfies the sub-gaussian tail bound:

$$\Pr_{X \in_u S'} \left[|v^* \cdot (X - \mu^{S'})| > T + \delta/2 \right] \leq 8 \exp(-T^2/2\nu) + 8 \frac{\varepsilon}{\tilde{T}^2} .$$

3. Then $\|\Sigma - I\|_2 \leq O(\varepsilon \log(\frac{1}{\varepsilon}))$. Contradiction.
   Therefore T exists and:

Proof of Correctness of Alg. FILTER: Large Spectral Norm

1 A threshold T exists

Proof ouline:

1. Suppose not.
2. Then S' satisfies the sub-gaussian tail bound:

$$\Pr_{X \in_u S'}\left[|v^* \cdot (X - \mu^{S'})| > T + \delta/2\right] \le 8\exp(-T^2/2\nu) + 8\frac{\varepsilon}{\tilde{T}^2}\,.$$

3. Then $\|\Sigma - I\|_2 \le O(\varepsilon \log(\frac{1}{\varepsilon}))$. Contradiction. Therefore T exists and:

$$\underbrace{|E \setminus E'|}_{\#\text{bad points rejected}} + \underbrace{|L' \setminus L|}_{\#\text{good points rejected}} \ge 8\varepsilon|S'|/\tilde{T}^2$$

$$\ge 4\varepsilon|S|/\tilde{T}^2$$

Proof of Correctness of Alg. Filter: Large Spectral Norm

2 Filter rejects at most $\left(2\exp(-T^2/2\nu)+\varepsilon/\tilde{T}^2\right)$-fraction from $S \cap S'$

Proof of Correctness of Alg. Filter: Large Spectral Norm

2 Filter rejects at most $\left(2\exp(-T^2/2\nu) + \varepsilon/\tilde{T}^2\right)$-fraction from $S \cap S'$

Proof:

Points in S satisfy the "goodness" tail bound:

$$\Pr_{X \in_u S}[|w \cdot (X - \mu^{S'})| > T + \|\mu^{S'} - \mu^G\|_2] \leq 2\exp(-T^2/2\nu) + \frac{\varepsilon}{\tilde{T}^2}.$$

Proof of Correctness of Alg. Filter: Large Spectral Norm

2 Filter rejects at most $\left(2\exp(-T^2/2\nu)+\varepsilon/\tilde{T}^2\right)$-fraction from $S \cap S'$

Proof:
Points in S satisfy the "goodness" tail bound:

$$\Pr_{X \in_u S}[|w \cdot (X - \mu^{S'})| > T + \|\mu^{S'} - \mu^G\|_2] \leq 2\exp(-T^2/2\nu) + \frac{\varepsilon}{\tilde{T}^2}.$$

Therefore, $\underbrace{|L' \setminus L|}_{\#\text{good points rejected}} < \varepsilon|S|/\tilde{T}^2$

Proof of Correctness of Alg. Filter: Large Spectral Norm

3 Filter rejects more bad points than good and makes progress:

$$\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/\alpha$$

Proof of Correctness of Alg. Filter: Large Spectral Norm

3 Filter rejects more bad points than good and makes progress:

$$\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/\alpha$$

Proof:

$$\Delta(S, S') - \Delta(S, S'') = \frac{\overbrace{|E \setminus E'|}^{\#\text{bad points rejected}} - \overbrace{|L' \setminus L|}^{\#\text{good points rejected}}}{|S|}$$
$$= \frac{(|E \setminus E'| + |L' \setminus L|) - 2|L' \setminus L|}{|S|}$$
$$\geq 2\varepsilon/\tilde{T}^2 \text{ (Follows from above)}$$
$$\geq 2\varepsilon/\alpha$$

Proof of Correctness of Alg. Filter: Large Spectral Norm

3 Filter rejects more bad points than good and makes progress:

$$\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/\alpha$$

Proof:

$$\Delta(S, S') - \Delta(S, S'') = \frac{\overbrace{|E \setminus E'|}^{\#\text{bad points rejected}} - \overbrace{|L' \setminus L|}^{\#\text{good points rejected}}}{|S|}$$

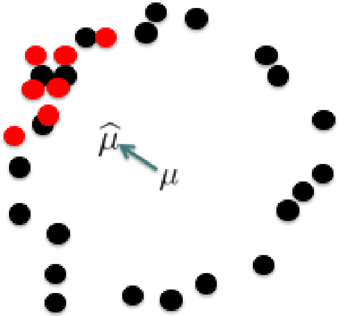
$$= \frac{(|E \setminus E'| + |L' \setminus L|) - 2|L' \setminus L|}{|S|}$$

$$\geq 2\varepsilon/\tilde{T}^2 \text{ (Follows from above)}$$

$$\geq 2\varepsilon/\alpha$$

$T = O(\sqrt{d \log(d/\varepsilon\tau)})$, since all points in S' satisfy $\|x - \mu^{S'}\|_2 \leq O(\sqrt{d \log(d/\varepsilon\tau)})$ (Consequence of NaivePrune).

A sketch of what just happened



1. If norm of covariance is small, μ is close to $\widehat{\mu}$. Algorithm terminates.
2. If not, direction of largest eigenvalue gives a discriminatory tail bound.
3. Threshold ensures that we reject more bad points than good and make progress at certain rate.
4. Algorithm terminates before or when all bad points are rejected.

Experiments

Estimating a $\mathcal{N}(\mu, \mathbf{I})$ Gaussian (known covariance)

- Set $\varepsilon = 0.1$
- $\#$ dimensions (d) from 100 to 400, in steps of 50
- (1 - ε)-fraction samples $\sim \mathcal{N}(\mu, \mathbf{I})$
  - μ is the all-ones vector
- ε-fraction from noise distribution: described in next slide
- $n = \frac{10d}{\varepsilon^2}$ samples $= \tilde{O}\left(\frac{d}{\varepsilon^2}\right)$

Estimating a $\mathcal{N}(\mu, \mathbf{I})$ Gaussian: Noise Model

$$N = \frac{1}{2}\Pi_1 + \frac{1}{2}\Pi_2$$

where

- Π_1: every coordinate is 0 or 1 with probability $1/2$
- Π_2: product distribution of:
  - First coordinate 0 or 12 with probability $1/2$
  - Second coordinate -2 or 0 with probability $1/2$
  - All other coordinates 0

Mean Estimation under Identity Covariance

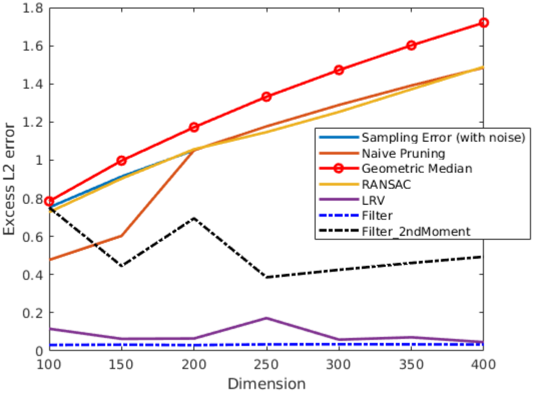


Figure: Mean estimation error with identity covariance matrix

Mean Estimation under (scaled) Identity Covariance

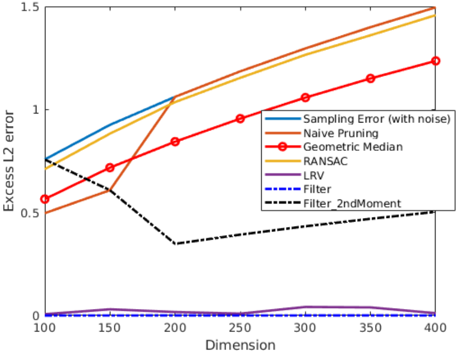


Figure: $C = 0.5\mathbf{I}$

Mean Estimation under (scaled) Identity Covariance

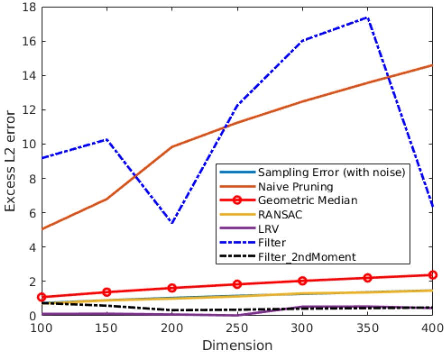


Figure: $C = 2\mathbf{I}$

Trying non-identity covariances: Diagonal covariance

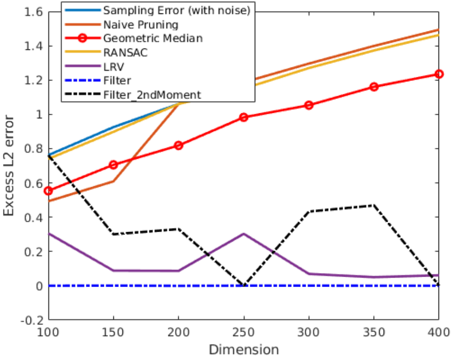


Figure: $C = \texttt{diag(rand(d,1))}$

Trying non-identity covariances: Diagonal covariance

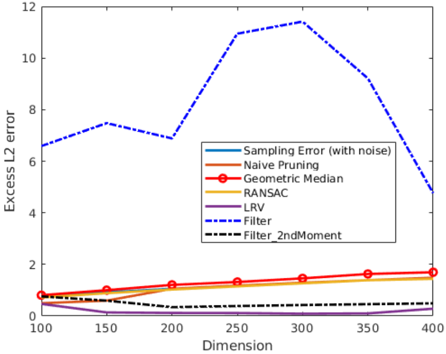


Figure: C $=$ `2 * diag(rand(d,1))`

Trying non-identity covariances: Rotated covariance

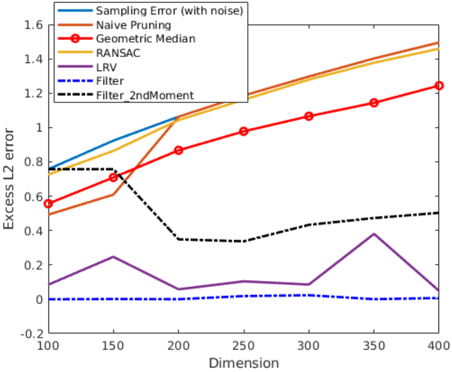


Figure: $[Q,R]=qr(rand(d,d));$ $C=Q*diag(rand(d,1))*Q';$

Trying non-identity covariances: Rotated covariance

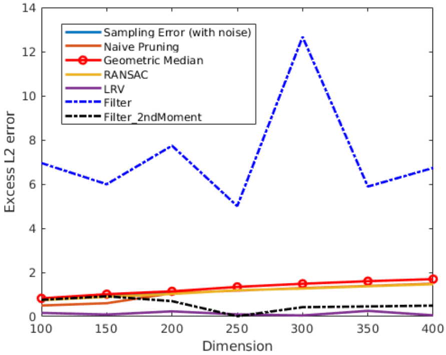


Figure: $$[Q,R]=qr(rand(d,d));$$
$$C=Q*2*diag(rand(d,1))*Q';$$

Additional experiments

Estimating the covariance matrix:

- Synthetic data
    - Isotropic: $\mathcal{N}(0, \mathbf{I})$
    - Spiked: $\mathcal{N}(0, \mathbf{I} + 100 e_1 e_1^T)$

Additional experiments

Estimating the covariance matrix:

- Synthetic data
  - Isotropic: $\mathcal{N}(0, \mathbf{I})$
  - Spiked: $\mathcal{N}(0, \mathbf{I} + 100 e_1 e_1^T)$
- (Noisy) 20-dimensional projection of data from "Genes mirror geography within Europe" (*Nature*, 2008)
  - 2D PCA projection should recover map of Europe, as in original work

Additional experiments

Estimating the covariance matrix:

- Synthetic data
  - Isotropic: $\mathcal{N}(0, \mathbf{I})$
  - Spiked: $\mathcal{N}(0, \mathbf{I} + 100 e_1 e_1^T)$
- (Noisy) 20-dimensional projection of data from "Genes mirror geography within Europe" (*Nature*, 2008)
  - 2D PCA projection should recover map of Europe, as in original work

Not enough time to cover these, unfortunately...

But pretty pictures!

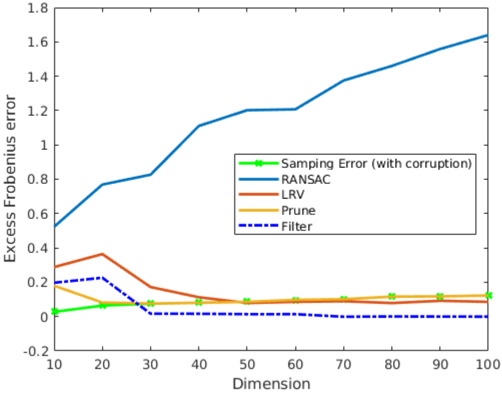
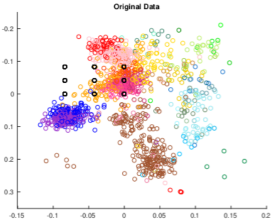


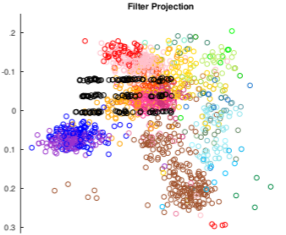
Figure: Covariance estimation error assuming isotropic covariance

Extension to robust PCA (images from the paper)



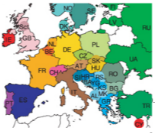
The data projected onto the top two directions of the original data set without noise

(a)



The data projected onto the top two directions returned by the filter

(b)



(c)

Figure: Recovering geographic structure of 2D projection in the presence of noise

Discussion

Some Questions

1. Paper considers two kinds of covariance matrices: fully known ($\mathbf{I}$) and bounded-second-moment ($\preceq \mathbf{I}$)
   - Room for exploring more restricted families of covariance matrices? *Tridiagonal, perhaps?*
   - The main goal appears to be to show $\Sigma - C \approx \varepsilon M_E$, combined with a nice tail bound to guarantee not throwing away inliers.
   - Low-rank case ("Robust PCA?") covered by a coming paper in the presentation schedule. . .

Some Questions

1. Paper considers two kinds of covariance matrices: fully known ($\mathbf{I}$) and bounded-second-moment ($\preceq \mathbf{I}$)
   - Room for exploring more restricted families of covariance matrices? *Tridiagonal, perhaps?*
   - The main goal appears to be to show $\Sigma - C \approx \varepsilon M_E$, combined with a nice tail bound to guarantee not throwing away inliers.
   - Low-rank case ("Robust PCA?") covered by a coming paper in the presentation schedule. . .
2. Theorem 3.1 algorithm seems very sensitive to the scale parameter. Can we make it work with a bounded σ assumption, or better, unknown σ?

Contamination Model Space

The adversary can be thought of as a stochastic process

$$C\left(y_1, y_2, \ldots y_m | x_1, x_2, \ldots x_m\right)$$

Contamination Model Space

The adversary can be thought of as a stochastic process

$$C\left(y_1, y_2, \ldots y_m | x_1, x_2, \ldots x_m\right)$$

which gives rise to a marginal distribution

$$M\left(y_1, y_2, \ldots y_m\right)$$

Contamination Model Space

The adversary can be thought of as a stochastic process

$$C\left(y_1, y_2, \ldots y_m | x_1, x_2, \ldots x_m\right)$$

which gives rise to a marginal distribution

$$M\left(y_1, y_2, \ldots y_m\right)$$

Questions/Conjectures:

1. If the original data is an i.i.d. sample from P, $d_{TV}\left(M\left(y_1, y_2, \ldots y_m\right), \prod_{i=1}^{m} P(y_i)\right) \leq \tilde{O}((d+m)\varepsilon)$?

Contamination Model Space

The adversary can be thought of as a stochastic process

$$C\left(y_1, y_2, \ldots y_m | x_1, x_2, \ldots x_m\right)$$

which gives rise to a marginal distribution

$$M\left(y_1, y_2, \ldots y_m\right)$$

Questions/Conjectures:

1. If the original data is an i.i.d. sample from P, $d_{TV}\left(M\left(y_1, y_2, \ldots y_m\right), \prod_{i=1}^{m} P(y_i)\right) \leq \tilde{O}((d+m)\varepsilon)$?
2. $\tilde{O}(\varepsilon)$?

Contamination Model Space

The adversary can be thought of as a stochastic process

$$C\left(y_1, y_2, \ldots y_m | x_1, x_2, \ldots x_m\right)$$

which gives rise to a marginal distribution

$$M\left(y_1, y_2, \ldots y_m\right)$$

Questions/Conjectures:

1. If the original data is an i.i.d. sample from P, $d_{TV}\left(M\left(y_1, y_2, \ldots y_m\right), \prod_{i=1}^{m} P(y_i)\right) \leq \tilde{O}((d+m)\varepsilon)$?
2. $\tilde{O}(\varepsilon)$?
3. If the marginal of each y_i is Q_i, then $d_{TV}\left(P(y_i), Q_i(y_i)\right) \leq \tilde{O}(\varepsilon)$

Cases not covered by DKK et. al.

- Extend algorithm for Covariance Estimation to Sub-Gaussian from Gaussian.

Cases not covered by DKK et. al.

- Extend algorithm for Covariance Estimation to Sub-Gaussian from Gaussian.
- Can estimate parameters of Gaussian with unknown mean and covariance. *What happens in the Sub-Gaussian case?*
  - Need to estimate covariance using the above, and adjust tail bounds for error in estimation of covariance - specifically the δ term
  - Potentially a worse error bound.

Thank You!