Solving Corrupted Quadratic Equations, Provably

Yuejie Chi

Electrical and Computer Engineering

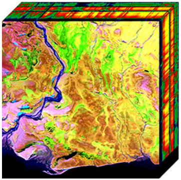
The Ohio State University

University of Michigan
November 2016

Data science

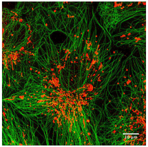
New imaging/sensing modalities allow us to probe the nature in unprecedented manners:

healthcare



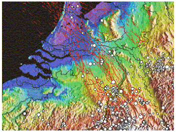
hyperspectral

Internet traffic



microscopy

Radio astronomy

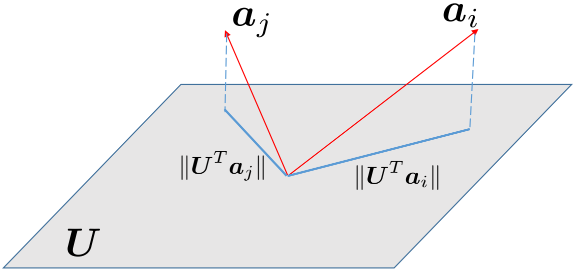


seismic imaging

but also with a lot of new (and exciting) challenges due to the unconventional manner these data are obtained.

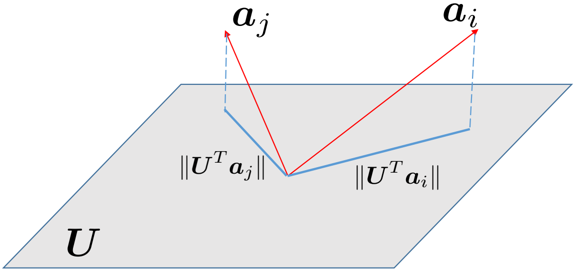
Subspace retrieval using intensity measurements only

- We wish to estimate a subspace $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ by interrogating it with vectors $\{\boldsymbol{a}_i\}_{i=1}^m$ and forming backprojections;



Subspace retrieval using intensity measurements only

- We wish to estimate a subspace $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ by interrogating it with vectors $\{\boldsymbol{a}_i\}_{i=1}^m$ and forming backprojections;



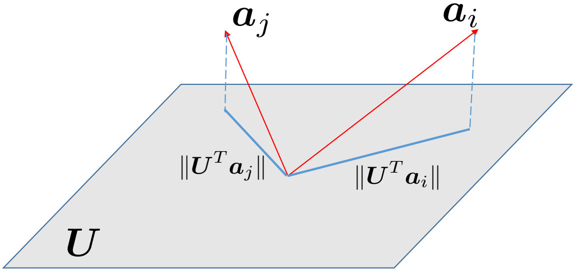
- We only observe the intensity of the backprojections, namely,

$$y_i = \|\boldsymbol{U}^T\boldsymbol{a}_i\|_2^2 = \boldsymbol{a}_i^T(\boldsymbol{U}\boldsymbol{U}^T)\boldsymbol{a}_i, \quad i = 1,\ldots,m.$$

They are quadratic with respect to $\boldsymbol{U}$.

Subspace retrieval using intensity measurements only

- We wish to estimate a subspace $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ by interrogating it with vectors $\{\boldsymbol{a}_i\}_{i=1}^m$ and forming backprojections;



- We only observe the intensity of the backprojections, namely,

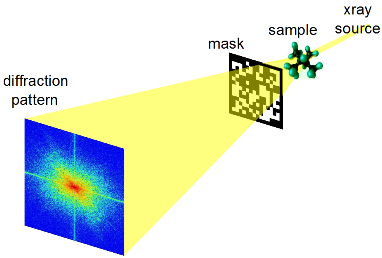
$$y_i = \|\boldsymbol{U}^T\boldsymbol{a}_i\|_2^2 = \boldsymbol{a}_i^T(\boldsymbol{U}\boldsymbol{U}^T)\boldsymbol{a}_i, \quad i = 1, \ldots, m.$$

  They are quadratic with respect to $\boldsymbol{U}$.

- Intensity measurements are much easier to implement by an energy detector for high-frequency and wide-band (THz) applications.

Phase retrieval

How to recover structure of a sample from its diffraction pattern?



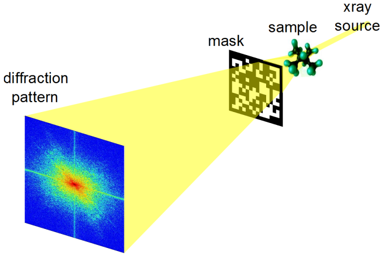
- In the important special case of $r = 1$, it becomes equivalent to **phase retrieval***, namely, recover $\boldsymbol{x} \in \mathbb{R}^n / \mathbb{C}^n$ from

$$y_i = |\mathcal{F}\{\boldsymbol{x}\}|^2, \quad \text{where } \mathcal{F} \text{ is Fourier transform,}$$

*Image credit: E. J. Candès, Y. C. Eldar, T. Strohmer and V. Voroninski, "Phase retrieval via matrix completion," SIAM J. on Imaging Sciences.

Phase retrieval

How to recover structure of a sample from its diffraction pattern?



- In the important special case of $r = 1$, it becomes equivalent to **phase retrieval***, namely, recover $\boldsymbol{x} \in \mathbb{R}^n/\mathbb{C}^n$ from

$$y_i = |\mathcal{F}\{\boldsymbol{x}\}|^2, \quad \text{where } \mathcal{F} \text{ is Fourier transform},$$

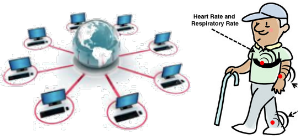
This has wide applications in X-ray crystallography, electron microscopy and coherent diffractive imaging, and leads to winning of Nobel prize (e.g. discovery of double helix structure).

*Image credit: E. J. Candès, Y. C. Eldar, T. Strohmer and V. Voroninski, "Phase retrieval via matrix completion," SIAM J. on Imaging Sciences.

Covariance sketching for streaming data

Multivariate streaming data: a new data snapshot $\boldsymbol{x}_t \in \mathbb{C}^n/\mathbb{R}^n$ is generated by the sensor platform at each time t;

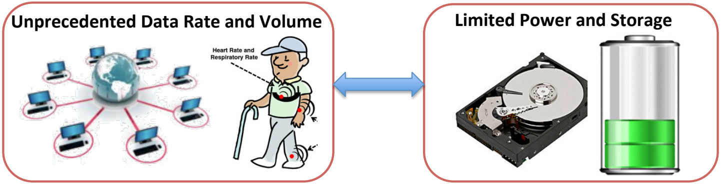
Unprecedented Data Rate and Volume



Limited Power and Storage

Covariance sketching for streaming data

Multivariate streaming data: a new data snapshot $\boldsymbol{x}_t \in \mathbb{C}^n/\mathbb{R}^n$ is generated by the sensor platform at each time t;



- **high-dimensional:** the number of variables, n, is large;
- **real-time:** data processed "on the fly";
- **decentralized:** data collected at decentralized locations;
- **resource-constrained:** cannot store and transmit all data;

Covariance sketching

Observation: Fortunately, inference requires only statistics of the data stream, not the stream itself; we can "sketch"/compress the data at the hope of directly recovering its statistics!

Covariance sketching

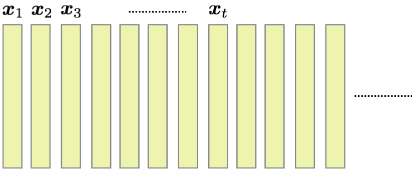
Observation: Fortunately, inference requires only statistics of the data stream, not the stream itself; we can "sketch"/compress the data at the hope of directly recovering its statistics!

Approach: distributed data sketching and aggregation to recover the covariance structure or principal components.

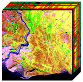
- access each data sample via quadratic (energy) sketches;
- aggregate the sketches into linear observations of the covariance matrix.

Quadratic sampling

How to sketch a high-dimensional data stream in order to recover its covariance matrix?



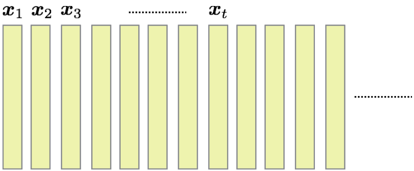
network traffic



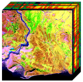
hyperspectral imagery

Quadratic sampling

How to sketch a high-dimensional data stream in order to recover its covariance matrix?



network traffic



hyperspectral imagery

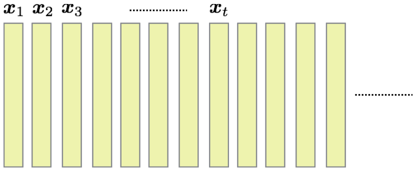
- To meet resource constraints, we would like to sample in a single pass on the fly: a single *quadratic* sketch of $\boldsymbol{x}_t$:

$$z_t = |\langle \boldsymbol{a}_t, \boldsymbol{x}_t \rangle|^2,$$

  which reduces the dim. of each $\boldsymbol{x}_t$ to merely a scalar.
- sketching complexity is linear in length of $\boldsymbol{x}_t$;

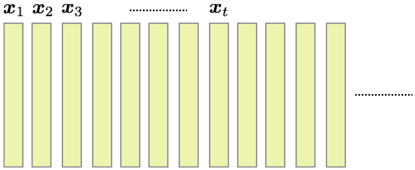
Quadratic sampling for covariance sketching

- Consider a data stream possible distributively observed at m sensors, each with a sketching vector $\boldsymbol{a}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$:



Quadratic sampling for covariance sketching

- Consider a data stream possible distributively observed at m sensors, each with a sketching vector $\boldsymbol{a}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$:



- Sketch a substream indexed by $\{\ell_t^i\}_{t=1}^T$ with $|\langle \boldsymbol{a}_i, \boldsymbol{x}_{\ell_t^i} \rangle|^2$ and compute the average:

$$y_{i,T} = \frac{1}{T} \sum_{t=1}^{T} \left| \left\langle \boldsymbol{a}_i, \boldsymbol{x}_{\ell_t^i} \right\rangle \right|^2 = \boldsymbol{a}_i^T \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_{\ell_t^i} \boldsymbol{x}_{\ell_t^i}^T \right) \boldsymbol{a}_i \xrightarrow{T \to \infty} \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i,$$

where $\boldsymbol{X} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$ is the covariance matrix.

Low-rank covariance estimation

- More generally, quadratic samplers produce the following:

$$y_i = \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i + \eta_i, \quad i = 1, \ldots, m;$$

  where $\boldsymbol{\eta}$ is an additive noise.
  - **linear in the covariance matrix $\boldsymbol{X}$!**

Low-rank covariance estimation

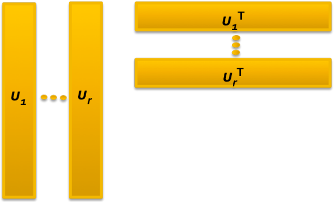
- More generally, quadratic samplers produce the following:

$$y_i = \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i + \eta_i, \quad i = 1, \ldots, m;$$

  where $\boldsymbol{\eta}$ is an additive noise.
  - **linear in the covariance matrix $\boldsymbol{X}$!**

- Low-rank covariance matrix: Many high-dimensional data lie in a low-dimensional subspace, when a small number of components accounts for most of the variability in the data.

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T =$$



- This yields the *subspace retrieval* problem.

Reconstruction?

Two sides of the same coin: We can recover

- either $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T \in \mathbb{R}^{n\times n}$ (when r is possibly unknown) or
- the subspace $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ (when r is known);

	$\boldsymbol{X}$	$\boldsymbol{U}$
measurements	$y_i = \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i$	$y_i = \\|\boldsymbol{U}^T \boldsymbol{a}_i\\|_2^2$
loss	linear	quadratic
prior	$\boldsymbol{X}$ is low-rank	-
dim. of unknowns	n^2	nr
optimization	convex	nonconvex

Reconstruction?

Two sides of the same coin: We can recover

- either $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T \in \mathbb{R}^{n \times n}$ (when r is possibly unknown) or
- the subspace $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ (when r is known);

	$\boldsymbol{X}$	$\boldsymbol{U}$
measurements	$y_i = \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i$	$y_i = \|\boldsymbol{U}^T \boldsymbol{a}_i\|_2^2$
loss	linear	quadratic
prior	$\boldsymbol{X}$ is low-rank	-
dim. of unknowns	n^2	nr
optimization	convex	nonconvex

We will discuss both convex (for reconstructing $\boldsymbol{X}$) and nonconvex methods (for reconstructing $\boldsymbol{U}$), possibly with additional corruptions *in the measurements.*

Low-rank covariance estimation via convex relaxation

- We would like to seek the covariance matrix satisfying the observations with the minimal rank:

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{M} \succeq 0} \ \mathsf{rank}(\boldsymbol{M}) \quad \text{s.t.} \quad y_i = \boldsymbol{a}_i^T \boldsymbol{M} \boldsymbol{a}_i, \ i = 1, \ldots, m.$$

Low-rank covariance estimation via convex relaxation

- We would like to seek the covariance matrix satisfying the observations with the minimal rank:

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{M} \succeq 0} \ \mathsf{rank}(\boldsymbol{M}) \quad \text{s.t.} \quad y_i = \boldsymbol{a}_i^T \boldsymbol{M} \boldsymbol{a}_i, \ i = 1, \ldots, m.$$

- However this is non-convex and NP-hard. Therefore, we replace it by a convex relaxation which is the trace minimization, over all PSD matrices compatible with the measurements:

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{M} \succeq 0} \ \mathsf{Tr}(\boldsymbol{M}) \quad \text{s.t.} \quad y_i = \boldsymbol{a}_i^T \boldsymbol{M} \boldsymbol{a}_i, \ i = 1, \ldots, m.$$

Low-rank covariance estimation via convex relaxation

- We would like to seek the covariance matrix satisfying the observations with the minimal rank:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{M} \succeq 0}{\operatorname{argmin}} \ \mathsf{rank}(\boldsymbol{M}) \quad \text{s.t.} \quad y_i = \boldsymbol{a}_i^T \boldsymbol{M} \boldsymbol{a}_i, \ i = 1, \ldots, m.$$

- However this is non-convex and NP-hard. Therefore, we replace it by a convex relaxation which is the trace minimization, over all PSD matrices compatible with the measurements:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{M} \succeq 0}{\operatorname{argmin}} \ \mathsf{Tr}(\boldsymbol{M}) \quad \text{s.t.} \quad y_i = \boldsymbol{a}_i^T \boldsymbol{M} \boldsymbol{a}_i, \ i = 1, \ldots, m.$$

- Additionally, if $\boldsymbol{X}$ is *Toeplitz*, solve:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{M} \succeq 0, \mathsf{Toeplitz}}{\operatorname{argmin}} \ \mathsf{Tr}(\boldsymbol{M}) \quad \text{s.t.} \quad y_i = \boldsymbol{a}_i^T \boldsymbol{M} \boldsymbol{a}_i, \ i = 1, \ldots, m.$$

Near-optimal recovery via convex programming

Theorem (Chen, C. and Goldsmith)

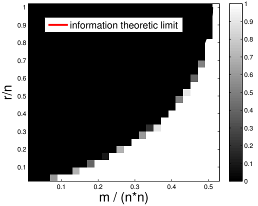
Assuming $\boldsymbol{a}_i$'s are composed of i.i.d. Gaussian entries, with high probability, the solution $\hat{\boldsymbol{X}}$ exactly recovers all rank-r matrices $\boldsymbol{X}$, provided that

$$m \gtrsim nr.$$

If there exists additional Toeplitz constraint, then similar guarantee holds provided

$$m \gtrsim r \text{polylog} n.$$

- **Exact recovery** with $m = O(nr)$;
- **Robust** against approximate low-rankness and bounded noise.
- **Under Toeplitz constraint:**

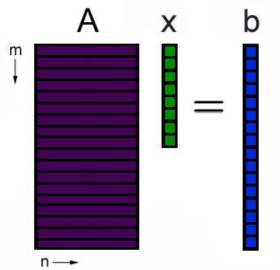


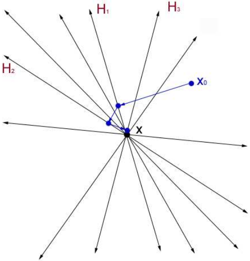
Kaczmarz method for solving quadratic equations

- Goal: reduce the memory and computational cost by directly estimating $\boldsymbol{U} \in \mathbb{R}^{n \times r}$.

Kaczmarz method for solving quadratic equations

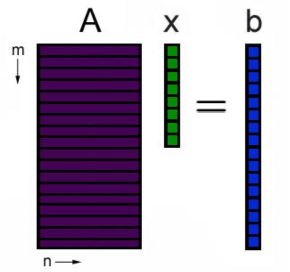
- Goal: reduce the memory and computational cost by directly estimating $\boldsymbol{U} \in \mathbb{R}^{n\times r}$.
- The **Kaczmarz method** is a fast iterative algorithm for solving overdetermined linear system.

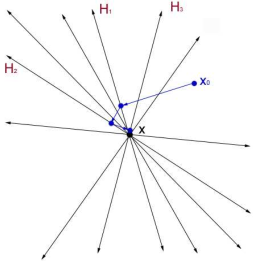




Kaczmarz method for solving quadratic equations

- Goal: reduce the memory and computational cost by directly estimating $\boldsymbol{U} \in \mathbb{R}^{n \times r}$.
- The **Kaczmarz method** is a fast iterative algorithm for solving overdetermined linear system.





- Its randomized version [Strohmer and Vershynin] obtains linear rate in expectation.

Kaczmarz method for solving quadratic equations

- Extend Kaczmarz method by, at each iteration, project the current estimate to the closest signal that satisfies a (quadratic) constraint:[†]

$$\boldsymbol{U}_k = \underset{\boldsymbol{V}:\left\|\boldsymbol{V}^T \boldsymbol{a}_{\ell(k)}\right\|_2^2 = y_{\ell(k)}}{\operatorname{argmin}} \left\|\boldsymbol{U}_{k-1} - \boldsymbol{V}\right\|_{\mathrm{F}}^2,$$

[†] Y. Chi and Y. M. Lu, Kaczmarz Method for Solving Quadratic Equations, IEEE SPL 2016.

Kaczmarz method for solving quadratic equations

- Extend Kaczmarz method by, at each iteration, project the current estimate to the closest signal that satisfies a (quadratic) constraint:[†]

$$\boldsymbol{U}_k = \underset{\boldsymbol{V}:\left\|\boldsymbol{V}^T \boldsymbol{a}_{\ell(k)}\right\|_2^2 = y_{\ell(k)}}{\operatorname{argmin}} \|\boldsymbol{U}_{k-1} - \boldsymbol{V}\|_{\mathrm{F}}^2 ,$$

which can be solved in **closed form** via a rank-one update:

$$\boldsymbol{U}_k = \left[\boldsymbol{I} - \left(\frac{\|\boldsymbol{U}_{k-1}^T \mathbf{a}_{\ell(k)}\|_2 - \sqrt{y_{\ell(k)}}}{\|\boldsymbol{U}_{k-1}^T \mathbf{a}_{\ell(k)}\|_2}\right) \frac{\mathbf{a}_{\ell(k)} \mathbf{a}_{\ell(k)}^T}{\|\boldsymbol{a}_{\ell(k)}\|_2^2}\right] \boldsymbol{U}_{k-1}.$$

[†] Y. Chi and Y. M. Lu, Kaczmarz Method for Solving Quadratic Equations, IEEE SPL 2016.

Kaczmarz method for solving quadratic equations

- Extend Kaczmarz method by, at each iteration, project the current estimate to the closest signal that satisfies a (quadratic) constraint:[†]

$$\boldsymbol{U}_k = \underset{\boldsymbol{V}:\left\|\boldsymbol{V}^T \boldsymbol{a}_{\ell(k)}\right\|_2^2 = y_{\ell(k)}}{\operatorname{argmin}} \left\|\boldsymbol{U}_{k-1} - \boldsymbol{V}\right\|_{\mathrm{F}}^2,$$

which can be solved in **closed form** via a rank-one update:

$$\boldsymbol{U}_k = \left[\boldsymbol{I} - \left(\frac{\|\boldsymbol{U}_{k-1}^T \mathbf{a}_{\ell(k)}\|_2 - \sqrt{y_{\ell(k)}}}{\|\boldsymbol{U}_{k-1}^T \mathbf{a}_{\ell(k)}\|_2}\right) \frac{\mathbf{a}_{\ell(k)} \mathbf{a}_{\ell(k)}^T}{\|\boldsymbol{a}_{\ell(k)}\|_2^2}\right] \boldsymbol{U}_{k-1}.$$

- The solution is equivalent to

$$\min_{\boldsymbol{s}:\|\boldsymbol{s}\|_2=1} \underset{\boldsymbol{V}:\left\|\boldsymbol{V}^T \boldsymbol{a}_{\ell(k)}\right\|_2 = \boldsymbol{s}\sqrt{y_{\ell(k)}}}{\operatorname{argmin}} \left\|\boldsymbol{U}_{k-1} - \boldsymbol{V}\right\|_{\mathrm{F}}^2$$

which corresponds to projecting the current estimate to the hyperplane with the phase that minimizes the projection.

[†]Y. Chi and Y. M. Lu, Kaczmarz Method for Solving Quadratic Equations, IEEE SPL 2016.

Performance Guarantee of Kaczmarz Method

Consider the phase retrieval case.

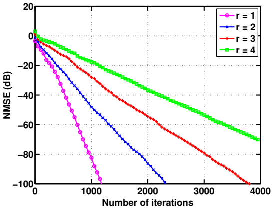
Theorem (Zhang, C., Liang)

Assume $\boldsymbol{a}_i$'s are generated with i.i.d. Gaussian entries, there exist some universal constants $\rho > 0$ such that if $m \gtrsim n$, then with high probability, randomized Kaczmarz update rule yields

$$\mathbb{E}_{i_t}\left[\mathrm{dist}^2(\boldsymbol{z}^{(t+1)}, \boldsymbol{x})\right] \leq \left(1 - \frac{\rho}{n}\right) \mathrm{dist}^2(\boldsymbol{z}^{(t)}, \boldsymbol{x})$$

where $\boldsymbol{z}^{(0)}$ is initialized via the spectral method.

- This establishes linear convergence rate *in expectation*, despite the nonlinearity!
- We can obtain similar guarantees for the block Kaczmarz method which is further accelerated.



What about outliers?

- Outliers happen with
  - sensor failures, malicious attacks, ...
  - For covariance sketching, insufficiently aggregated sketches can be regarded as an outlier;

What about outliers?

- Outliers happen with
  - sensor failures, malicious attacks, ...
  - For covariance sketching, insufficiently aggregated sketches can be regarded as an outlier;

- We're interested when the measurements are corrupted by both *sparse outliers* and *bounded noise*:

$$y_i = \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i + \eta_i + w_i, \quad i = 1, \ldots, m,$$

where $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$, $\|\boldsymbol{\eta}\|_0 \leq sm$ and $\boldsymbol{w}$ is a dense bounded noise.

What about outliers?

- Outliers happen with
  - sensor failures, malicious attacks, ...
  - For covariance sketching, insufficiently aggregated sketches can be regarded as an outlier;

- We're interested when the measurements are corrupted by both *sparse outliers* and *bounded noise*:

$$y_i = \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i + \eta_i + w_i, \quad i = 1, \ldots, m,$$

  where $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$, $\|\boldsymbol{\eta}\|_0 \leq sm$ and $\boldsymbol{w}$ is a dense bounded noise.

- Goal: develop algorithms that are *oblivious* to outliers, and statistically and computationally efficient.
  - small sample size: hopefully m is linear in n;
  - large fraction of outliers: hopefully s is a small constant;
  - low computational complexity and easy to implement.

Outlier-robust recovery by convex programming

- To motivate, ideally one would like to look for low-rank matrices that maintain outlier sparsity:

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{M} \succeq 0} \textsf{cardinality}(\textsf{outliers}), \quad \textsf{s.t.} \quad \textsf{rank}(\boldsymbol{M}) = r$$

Outlier-robust recovery by convex programming

- To motivate, ideally one would like to look for low-rank matrices that maintain outlier sparsity:

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{M} \succeq 0} \text{cardinality}(\text{outliers}), \quad \text{s.t.} \quad \text{rank}(\boldsymbol{M}) = r$$

- By *relaxing* the objective function to the ℓ_1-norm minimization, and *dropping* the rank constraint, we propose to solve

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{M} \succeq 0} \sum_{i=1}^{m} \left| y_i - \boldsymbol{a}_i^T \boldsymbol{M} \boldsymbol{a}_i \right|$$

Outlier-robust recovery by convex programming

- To motivate, ideally one would like to look for low-rank matrices that maintain outlier sparsity:

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{M} \succeq 0} \text{cardinality}(\text{outliers}), \quad \text{s.t.} \quad \text{rank}(\boldsymbol{M}) = r$$

- By *relaxing* the objective function to the ℓ_1-norm minimization, and *dropping* the rank constraint, we propose to solve

$$\hat{\boldsymbol{X}} = \operatorname*{argmin}_{\boldsymbol{M} \succeq 0} \sum_{i=1}^{m} \left| y_i - \boldsymbol{a}_i^T \boldsymbol{M} \boldsymbol{a}_i \right|$$

  - Parameter-free formulation without trace minimization or tuning parameters;
  - No prior information is required for the matrix rank, corruption level or bounded noise level.

Performance guarantee of convex programming

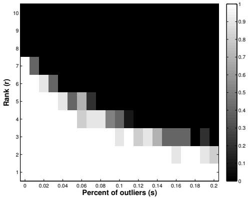
Theorem (Li, Sun and C., 2016)

Suppose that $\|\boldsymbol{w}\|_1 \leq \epsilon$. Assume the support of $\boldsymbol{\eta}$ is selected uniformly at random with the signs of $\boldsymbol{\eta}$ are generated from a symmetric Bernoulli distribution. Then as long as $m \gtrsim nr^2, s \lesssim 1/r$, the solution to the proposed algorithm satisfies

$$\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}\right\|_{\mathrm{F}} \lesssim \frac{r\epsilon}{m}$$

with high probability.

- Exact recovery when $\boldsymbol{w} = 0$ as long as $m \gtrsim nr^2$ and $s \lesssim 1/r$.
- When $r = 1$ recovers a previous result for the phase retrieval case[‡];
- RHS is phase transition for m vs r with 5% corruptions.

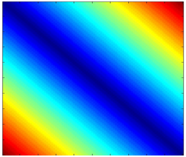


[‡]P. Hand, "Phaselift is robust to a constant fraction of arbitrary errors".

Robust recovery of Toeplitz PSD Matrices

If $\boldsymbol{X}$ is additionally Toeplitz, this can be incorporated:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{M}\succeq 0,\ \text{Toeplitz}}{\operatorname{argmin}} \sum_{i=1}^{m} \left| y_i - \boldsymbol{a}_i^T \boldsymbol{M} \boldsymbol{a}_i \right|.$$



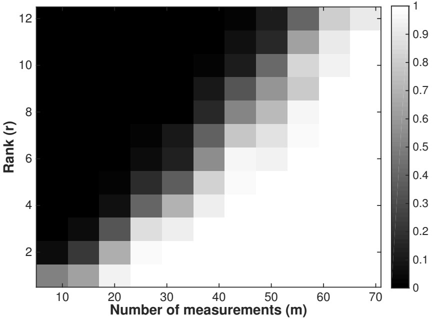


Figure : Phase transitions of low-rank Toeplitz PSD matrix recovery w.r.t. the number of measurements and the rank with 5% of measurements corrupted by standard Gaussian variables, when $n = 64$.

Non-convex approach based on factored model

Can we reduce the computational complexity?

- Recall $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$ where $\boldsymbol{U} \in \mathbb{R}^{n \times r}$, one can directly recover $\boldsymbol{U}$ by attempting:

$$\hat{\boldsymbol{U}} = \underset{\boldsymbol{U} \in \mathbb{R}^{n \times r}}{\operatorname{argmin}} \, \ell(\boldsymbol{U}) := \underset{\boldsymbol{U} \in \mathbb{R}^{n \times r}}{\operatorname{argmin}} \, \frac{1}{m} \sum_{i=1}^{m} \ell(y_i; \boldsymbol{U})$$

Non-convex approach based on factored model

Can we reduce the computational complexity?

- Recall $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$ where $\boldsymbol{U} \in \mathbb{R}^{n \times r}$, one can directly recover $\boldsymbol{U}$ by attempting:

$$\hat{\boldsymbol{U}} = \operatorname*{argmin}_{\boldsymbol{U} \in \mathbb{R}^{n \times r}} \ell(\boldsymbol{U}) := \operatorname*{argmin}_{\boldsymbol{U} \in \mathbb{R}^{n \times r}} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i; \boldsymbol{U})$$

  for some loss function $\ell(y_i, \boldsymbol{U})$:

  - quadratic loss of power: $\ell(\boldsymbol{U}; y_i) = \left(y_i - \left\|\boldsymbol{U}^T \boldsymbol{a}_i\right\|_2^2\right)^2$
  - quadratic loss of amplitude: $\ell(\boldsymbol{U}; y_i) = \left(\sqrt{y_i} - \left\|\boldsymbol{U}^T \boldsymbol{a}_i\right\|_2\right)^2$
  - Poisson loss: $\ell(\boldsymbol{U}; y_i) = \|\boldsymbol{U}^T \boldsymbol{a}_i\|_2^2 - y_i \log \|\boldsymbol{U}^T \boldsymbol{a}_i\|_2^2$

Non-convex approach based on factored model

Can we reduce the computational complexity?

- Recall $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$ where $\boldsymbol{U} \in \mathbb{R}^{n \times r}$, one can directly recover $\boldsymbol{U}$ by attempting:

$$\hat{\boldsymbol{U}} = \underset{\boldsymbol{U} \in \mathbb{R}^{n \times r}}{\operatorname{argmin}} \ell(\boldsymbol{U}) := \underset{\boldsymbol{U} \in \mathbb{R}^{n \times r}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i; \boldsymbol{U})$$

  for some loss function $\ell(y_i, \boldsymbol{U})$:

  - quadratic loss of power: $\ell(\boldsymbol{U}; y_i) = \left(y_i - \left\|\boldsymbol{U}^T \boldsymbol{a}_i\right\|_2^2\right)^2$
  - quadratic loss of amplitude: $\ell(\boldsymbol{U}; y_i) = \left(\sqrt{y_i} - \left\|\boldsymbol{U}^T \boldsymbol{a}_i\right\|_2\right)^2$
  - Poisson loss: $\ell(\boldsymbol{U}; y_i) = \|\boldsymbol{U}^T \boldsymbol{a}_i\|_2^2 - y_i \log \|\boldsymbol{U}^T \boldsymbol{a}_i\|_2^2$

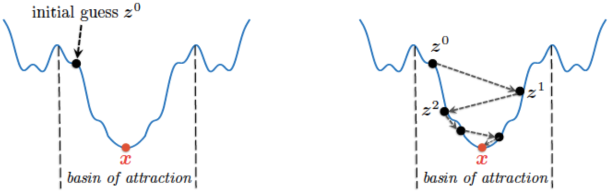
- What are the challenges?
  - $\ell(\boldsymbol{U})$ can be non-convex and non-smooth.
  - With outliers, we want the loss to sum over only clean samples.

Non-convex phase retrieval

Exciting developments (without outliers) – all following the same recipe (for the phase retrieval or rank-1 case):

$$\hat{\boldsymbol{z}} = \underset{\boldsymbol{z}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \ell(y_i;\boldsymbol{z})$$

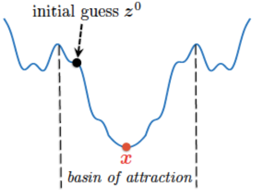
- Initialize $\boldsymbol{z}^{(0)}$ via the (truncated) spectral method to land in the neighborhood of the ground truth;
- Iterative update using (truncated) gradient descent;

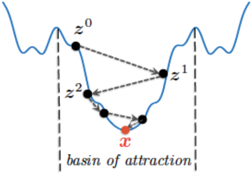


§

§Figure credit: Yuxin Chen.

Non-convex phase retrieval

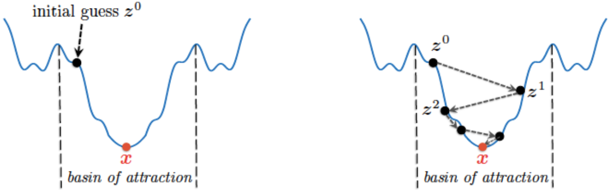




Provable near-optimal performance for Gaussian measurement model:

- Statistically: $m = O(n)$ near-optimal sample complexity
- Computationally: linear convergence with near-linear run time

Non-convex phase retrieval



Provable near-optimal performance for Gaussian measurement model:

- Statistically: $m = O(n)$ near-optimal sample complexity
- Computationally: linear convergence with near-linear run time

Examples: Wirtinger Flow (WF) (Candès et.al. 2014), Truncated Wirtinger Flow (TWF) (Chen and Candès 2015), Reshaped Wirtinger Flow (Zhang and Liang 2016), Truncated Amplitude Flow (Wang, Giannakis and Eldar, 2016)

Non-convex phase retrieval with outliers

In the presence of *arbitrary outliers*, **existing approaches fail:**

- Spectral initialization would fail: the eigenvector of $\boldsymbol{Y}$ can be arbitrarily perturbed

$$\underbrace{\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T}_{\text{WF}} \quad \text{or} \quad \underbrace{\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T \mathbb{1}_{\{|y_i| \le \alpha_y \cdot \mathrm{mean}(\{y_i\})\}}}_{\text{TWF}} .$$

¶with some details hiding

Non-convex phase retrieval with outliers

In the presence of *arbitrary outliers*, **existing approaches fail:**

- Spectral initialization would fail: the eigenvector of $\boldsymbol{Y}$ can be arbitrarily perturbed

$$\underbrace{\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T}_{\text{WF}} \quad \text{or} \quad \underbrace{\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T \mathbb{1}_{\{|y_i| \le \alpha_y \cdot \mathsf{mean}(\{y_i\})\}}}_{\text{TWF}}.$$

- Gradient descent would fail: the search direction can be arbitrarily perturbed

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{\mu}{\|\boldsymbol{z}^{(0)}\|^2} \sum_{i \in \mathcal{T}_t} \nabla \ell(\boldsymbol{z}^{(t)}; y_i)$$

where $\mathcal{T}_t = \{1, \ldots, m\}$ for WF and

$$\mathcal{T}_t = \left\{ i : |y_i - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2| \le \alpha_h \cdot \mathsf{mean}(\{|y_i - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2|\}) \right\}^{\P}$$

for TWF.

¶with some details hiding

Robust phase retrieval via median-truncation

Need better strategy to eliminate outliers!

Key approach: "median-truncation"

- well-known in robust statistics to be outlier-resilient;
- little appearance in high-dimensional estimation;

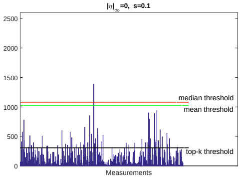
Robust phase retrieval via median-truncation

Need better strategy to eliminate outliers!

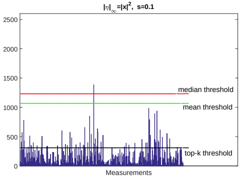
Key approach: "median-truncation"

- well-known in robust statistics to be outlier-resilient;
- little appearance in high-dimensional estimation;

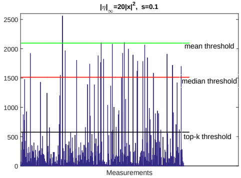
Median is more stable than mean and top-k truncation (which truncates a fixed amount of samples) for various levels of outliers.



no outliers



small outlier magnitudes



large outlier magnitudes

Median-Truncated Wirtinger Flow (median-TWF)

We adopt the Poisson loss function (other loss functions work too) and the Gaussian measurement model.

- **Median-truncated spectral initialization:** Set $\boldsymbol{z}^{(0)} := \lambda_0 \tilde{\boldsymbol{z}}$ where
  - Direction estimation: $\tilde{\boldsymbol{z}}$ is the leading eigenvector of

$$\boldsymbol{Y} = \frac{1}{m} \sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T \mathbb{1}_{\{|y_i| \le 9/0.455 \cdot \text{median}(\{y_i\})\}}.$$

  - Norm estimation: $\lambda_0 = \sqrt{\text{median}(\{y_i\})/0.455}$

$$y_i = |\boldsymbol{a}_i^T \boldsymbol{x}|^2 \sim \chi_1^2 \quad \text{and} \quad \mathbb{E}[\text{median}(\chi_1^2)] = 0.455$$

Median-Truncated Wirtinger Flow (median-TWF)

We adopt the Poisson loss function (other loss functions work too) and the Gaussian measurement model.

- **Median-truncated spectral initialization:** Set $\boldsymbol{z}^{(0)} := \lambda_0 \tilde{\boldsymbol{z}}$ where
  - Direction estimation: $\tilde{\boldsymbol{z}}$ is the leading eigenvector of

$$\boldsymbol{Y} = \frac{1}{m} \sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T \mathbb{1}_{\{|y_i| \leq 9/0.455 \cdot \text{median}(\{y_i\})\}}.$$

  - Norm estimation: $\lambda_0 = \sqrt{\text{median}(\{y_i\})/0.455}$

$$y_i = |\boldsymbol{a}_i^T \boldsymbol{x}|^2 \sim \chi_1^2 \quad \text{and} \quad \mathbb{E}[\text{median}(\chi_1^2)] = 0.455$$

- As long as $m = O(n \log n)$ and $s = O(1)$, the initialization is provably close to the ground truth:

$$\text{dist}(\boldsymbol{z}^{(0)}, \boldsymbol{x}) \leq \frac{1}{10} \|\boldsymbol{x}\|,$$

  where $\text{dist}(\boldsymbol{z}^{(0)}, \boldsymbol{x}) = \min\{\|\boldsymbol{z}^{(0)} + \boldsymbol{x}\|, \|\boldsymbol{z}^{(0)} - \boldsymbol{x}\|\}$.

Median-Truncated Wirtinger Flow (median-TWF)

- **Median-truncated gradient descent:**

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{2\mu}{m} \underbrace{\sum_{i \in \mathcal{E}_1 \cap \mathcal{E}_2} \frac{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2 - y_i}{\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}} \boldsymbol{a}_i}_{\nabla \ell_{tr}(\boldsymbol{z})},$$

where

$$\mathcal{E}_1 = \left\{ i : \ 0.3 \leq \frac{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|}{\|\boldsymbol{z}^{(t)}\|} \leq 5 \right\}, \mathcal{E}_2 = \left\{ i : \ r_i^{(t)} \leq 12 \frac{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|}{\|\boldsymbol{z}^{(t)}\|} \cdot \text{median}(\{r_i^{(t)}\}) \right\},$$

with $r_i^{(t)} = |y_i - (\boldsymbol{a}_i^T \boldsymbol{z}^{(t)})^2|$.

Median-Truncated Wirtinger Flow (median-TWF)

- **Median-truncated gradient descent:**

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{2\mu}{m} \underbrace{\sum_{i \in \mathcal{E}_1 \cap \mathcal{E}_2} \frac{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2 - y_i}{\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}} \boldsymbol{a}_i}_{\nabla \ell_{tr}(\boldsymbol{z})},$$

where

$$\mathcal{E}_1 = \left\{ i : \; 0.3 \le \frac{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|}{\|\boldsymbol{z}^{(t)}\|} \le 5 \right\}, \mathcal{E}_2 = \left\{ i : \; r_i^{(t)} \le 12 \frac{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|}{\|\boldsymbol{z}^{(t)}\|} \cdot \mathsf{median}(\{r_i^{(t)}\}) \right\},$$

with $r_i^{(t)} = |y_i - (\boldsymbol{a}_i^T \boldsymbol{z}^{(t)})^2|$.

- As long as $m = O(n \log n)$ and $s = O(1)$, $\nabla \ell_{tr}(\boldsymbol{z})$ satisfies the *Regularity Condition* $\mathrm{RC}(\mu, \lambda)$ for all $\boldsymbol{z}$, $\boldsymbol{h} = \boldsymbol{z} - \boldsymbol{x}$:

$$-\left\langle \frac{1}{m} \nabla \ell_{tr}(\boldsymbol{z}), \boldsymbol{h} \right\rangle \ge \mu \left\| \frac{1}{m} \nabla \ell_{tr}(\boldsymbol{z}) \right\|^2 + \lambda \|\boldsymbol{h}\|^2, \quad \|\boldsymbol{h}\| \le \frac{1}{10} \|\boldsymbol{z}\|.$$

which guarantees $\mathrm{dist}(\boldsymbol{z}^{(t+1)}, \boldsymbol{x}) \le (1 - \mu\lambda) \mathrm{dist}(\boldsymbol{z}^{(t)}, \boldsymbol{x})$.

Performance guarantee of median-TWF

Theorem (Zhang, C. and Liang, 2016)

Assume $\|\boldsymbol{w}\|_\infty \le c_1\|\boldsymbol{x}\|^2$. *Assume* $\boldsymbol{a}_i$*'s are generated with i.i.d. Gaussian entries. If* $m \gtrsim n \log n$ *and* $s \lesssim s_0$, *then with high probability, median-TWF yields*

$$\mathrm{dist}(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{x}\|} + (1-\rho)^t\|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N}$$

simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ *for some* $0 < \rho < 1$.

- **Exact recovery** when $\|\boldsymbol{w}\| = 0$ with slight more samples ($m = O(n \log n)$) but a constant fraction of outliers $s = O(1)$.
- **Stable recovery** with additional bounded noise;
- Resist outliers **obliviously**: no prior knowledge of outliers.
- *First* non-asymptotic robust recovery guarantee using median: much more involved due to the nonlinearity of median.

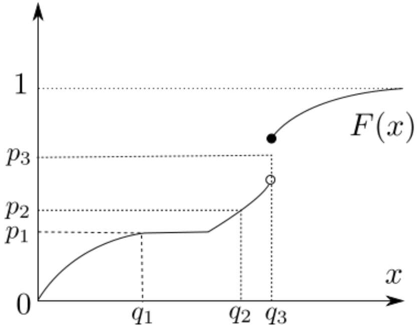
Proof sketch - preparation

Definition (Generalized quantile function)

Let $0 < p < 1$. If F is a CDF, the generalized quantile function is

$$F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

Denote $\theta_p(F) := F^{-1}(p)$ and $\theta_p(\{X_i\}) := \theta_p(\hat{F})$, where $\hat{F}$ is the empirical distribution of the samples $\{X_i\}_{i=1}^{m}$.



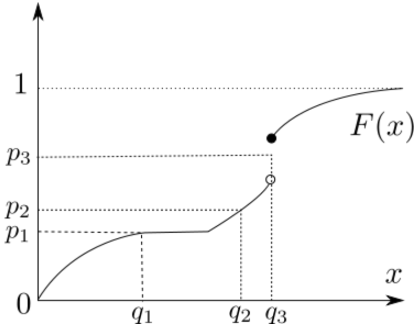
Proof sketch - preparation

Definition (Generalized quantile function)

Let $0 < p < 1$. If F is a CDF, the generalized quantile function is

$$F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

Denote $\theta_p(F) := F^{-1}(p)$ and $\theta_p(\{X_i\}) := \theta_p(\hat{F})$, where $\hat{F}$ is the empirical distribution of the samples $\{X_i\}_{i=1}^{m}$.



Proof sketch

Lemma (Concentration of sample quantile)

Assume $\{X_i\}_{i=1}^m$ are i.i.d. drawn from some distribution F. Under some minor assumptions, w.h.p.

$$|\theta_p(\{X_i\}_{i=1}^m) - \theta_p(F)| < \epsilon$$

Lemma (Sandwich median by quantiles of clean samples)

Consider clean samples $\{\tilde{X}_i\}_{i=1}^m$ and contaminated samples $\{X_i\}_{i=1}^m$. Then

$$\theta_{\frac{1}{2}-s}(\{\tilde{X}_i\}) \le \theta_{\frac{1}{2}}(\{X_i\}) \le \theta_{\frac{1}{2}+s}(\{\tilde{X}_i\}).$$

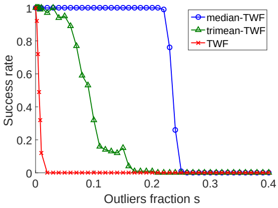
Lemma (Concentration of median)

If $m > c_0 n \log n$, then with probability at least $1 - c_1 \exp(-c_2 m)$, there exist constants β and β' such that

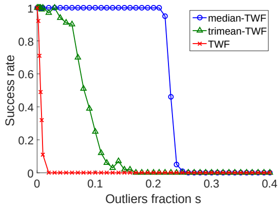
$$\beta\|\boldsymbol{z}\|\|\boldsymbol{h}\| \le \mathsf{median}(\{\big||\boldsymbol{a}_i^T\boldsymbol{x}|^2 - |\boldsymbol{a}_i^T\boldsymbol{z}|^2\big|\}_{i=1}^m) \le \beta'\|\boldsymbol{z}\|\|\boldsymbol{h}\|,$$

holds for all $\boldsymbol{z}, \boldsymbol{h} := \boldsymbol{z} - \boldsymbol{x}$ satisfying $\|\boldsymbol{h}\| < 1/11\|\boldsymbol{z}\|$.

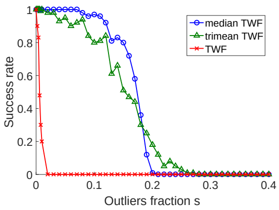
Numerical experiments with median-TWF



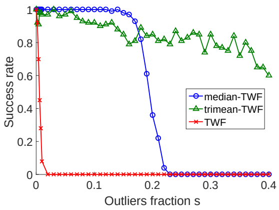
(a) $\|\boldsymbol{\eta}\|_\infty = 0.1\|\boldsymbol{x}\|^2$



(b) $\|\boldsymbol{\eta}\|_\infty = \|\boldsymbol{x}\|^2$



(c) $\|\boldsymbol{\eta}\|_\infty = 10\|\boldsymbol{x}\|^2$



(d) $\|\boldsymbol{\eta}\|_\infty = 100\|\boldsymbol{x}\|^2$

Figure : Success rate of **exact recovery** with outliers for median-TWF, trimean-TWF, and TWF at different levels of outlier magnitudes.

Numerical experiments with median-TWF

Recovery with both dense noise and sparse outliers:

- With outliers, median-TWF achieve better accuracy than TWF.
- Moreover, median-TWF with outliers achieves almost the same accuracy of TWF without outliers.

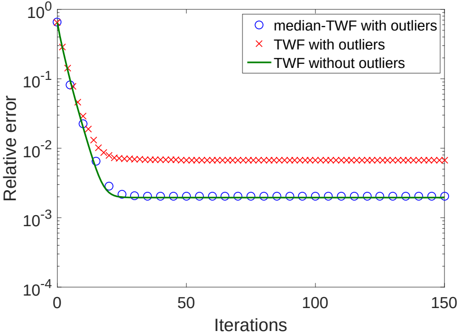


Figure : Relative error of median-TWF vs. TWF w.r.t. iteration when $s = 0.1$, $\|\boldsymbol{w}\|_\infty = 0.01\|\boldsymbol{x}\|^2$, and $\|\boldsymbol{\eta}\|_\infty = \|\boldsymbol{w}\|$.

Conclusions

We have discussed how to solve random quadratic systems of equations, possibly corrupted by a constant fraction of outliers, in a provable manner.

	$\boldsymbol{X}$	$\boldsymbol{U}$
measurements	$y_i = \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i$	$y_i = \Vert \boldsymbol{U}^T \boldsymbol{a}_i \Vert_2^2$
loss	linear/cvx	quadratic/ncvx
without outliers	Semidefinite Prog.	Kaczmarz/SGD
with outliers	Semidefinite Prog.	median-TWF

- **The class of convex methods** are based on convex relaxation for low-rank matrix completion and sparse recovery. It is easier to design but the computational cost is high;
- **The class of non-convex methods** are based on iterative updates with careful initializations. The computational cost is low but the design is a bit of an art.

References

1. Exact and Stable Covariance Estimation from Quadratic Sampling via Convex Programming, IEEE TIT 2015.
2. Low-Rank Positive Semidefinite Matrix Recovery from Corrupted Rank-One Measurements, IEEE TSP 2016.
3. Provable Non-convex Phase Retrieval with Outliers: Median Truncated Wirtinger Flow, ICML 2016.
4. Kaczmarz Method for Solving Quadratic Equations, IEEE SPL 2016.
5. Incremental Reshaped Wirtinger Flow and Its Connection to Kaczmarz Method, NIPS 2016 Workshop on Nonconvex Optimization.

http://www.ece.osu.edu/~chi/

Acknowledgement

- My collaborators: Yuxin Chen (Stanford), Andrea Goldsmith (Stanford), Yuanxin Li (OSU), Huishuai Zhang (Syracuse), Yingbin Liang (Syracuse) and Yue M. Lu (Harvard).

- Research supported by NSF, AFOSR and ONR.