Solving Corrupted Quadratic Equations, Provably

Yuejie Chi

The Ohio State University

London Workshop on Sparse Signal Processing
September 2016

Acknowledgement

- Joint work with Yuanxin Li (OSU), Huishuai Zhuang (Syracuse) and Yingbin Liang (Syracuse).

- Research supported by NSF, AFOSR and ONR.

Estimation of Low-rank PSD Matrices

- Consider estimation of a low-rank positive-semidefinite (PSD) matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ from symmetric rank-one measurements:

$$y_i = \langle \boldsymbol{a}_i \boldsymbol{a}_i^T, \boldsymbol{X} \rangle = \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i, \quad i = 1, \ldots, m.$$

- The measurements are nonnegative since $\boldsymbol{X} \succeq 0$.

- If $\mathrm{rank}(\boldsymbol{X}) = r$, decompose $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$, where $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, then the measurements are *quadratic* in $\boldsymbol{U}$:

$$y_i = \boldsymbol{a}_i^T \boldsymbol{U}\boldsymbol{U}^T \boldsymbol{a}_i = \|\boldsymbol{U}^T \boldsymbol{a}_i\|_2^2.$$

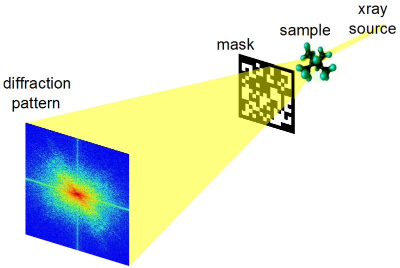
  The rank r may be unknown.

- Goal: recover $\boldsymbol{X}$ or $\boldsymbol{U}$ from as a small number of measurements.
- Related to low-rank matrix recovery but more structured/restricted.

Application - phase retrieval

Quadratic measurements arise in optical applications such as phase retrieval*, namely, recover $\boldsymbol{x} \in \mathbb{R}^n/\mathbb{C}^n$ from

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2 = \boldsymbol{a}_i^*(\boldsymbol{x}\boldsymbol{x}^*)\boldsymbol{a}_i, \quad i = 1, \ldots, m.$$

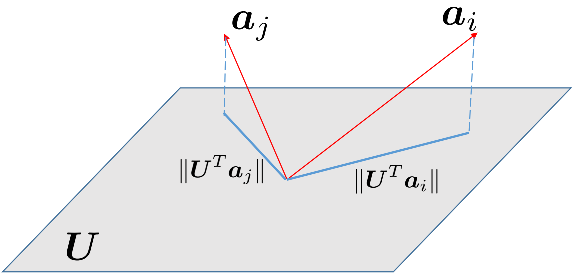


*E. J. Candès, Y. C. Eldar, T. Strohmer and V. Voroninski, "Phase retrieval via matrix completion," SIAM J. on Imaging Sciences.

Application - projection retrieval

Quadratic measurements arise in the problem of projection (or subspace) retrieval via energy measurements[†], namely, recover a subspace $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ from

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



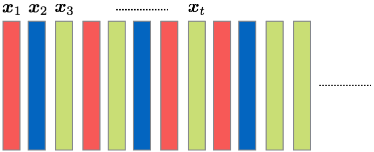
Useful in SAR imaging.

[†]M. Fickus and D. Mixon, "Projection Retrieval: Theory and Algorithms", SAMPTA 2015.

Application - covariance sketching

Question: how to sketch a high-dimensional data stream in order to recover its covariance matrix?‡

- Consider a data stream possible distributively observed at m sensors:



- **Quadratic Sketching:** For each sketching vector $\boldsymbol{a}_i \in \mathbb{R}^n$ with i.i.d. sub-Gaussian entries, $i = 1, \ldots, m$: 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 approximately low-rank.

‡Y. Chen, Y. Chi, and A. J. Goldsmith. "Exact and stable covariance estimation from quadratic sampling via convex programming." IEEE Trans. on IT, 2015.

Near-optimal recovery via convex programming

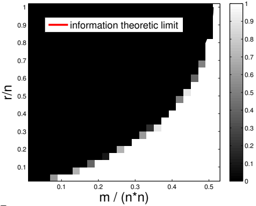
When $\boldsymbol{a}_i$'s are composed of i.i.d. Gaussian entries[§], we aim to recover $\boldsymbol{X}$ using the following trace minimization algorithm:

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

Theorem (Chen, Chi and Goldsmith, 2013)

With probability exceeding $1 - c_1 \exp(-c_2 m)$, the solution $\hat{\boldsymbol{X}}$ exactly recovers all rank-r matrices $\boldsymbol{X}$, provided that $m > c_0 nr$, where c_0, c_1, c_2 are universal constants.

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



[§]similar guarantees also hold for the sub-Gaussian case

What about outliers?

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

- In this talk, we're interested when the measurements are corrupted by both *sparse outliers* and *bounded noise*:

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

  or equivalently

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}) + \boldsymbol{\eta} + \boldsymbol{w},$$

  where $\boldsymbol{\eta}$ is a sparse vector with $\|\boldsymbol{\eta}\|_0 \leq sm$ and $\boldsymbol{w}$ is a dense bounded noise.

Pursuit of outlier-robust algorithms

Previous approaches are sensitive to outliers.

Goal: algorithms that are *oblivious* to outliers, i.e. perform equally well with or without outliers, and *without* any special treatments of outliers. And also 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.

We will outline two approaches, based on convex and non-convex optimization respectively.

Outlier-robust recovery by convex programming

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

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{M} \succeq 0}{\operatorname{argmin}} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_0 , \quad \text{s.t.} \quad \text{rank}(\boldsymbol{M}) = r$$

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

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{M} \succeq 0}{\operatorname{argmin}} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_1$$

  We call this algorithm ℓ_1*-regularized Phaselift*, or Phaselift-ℓ_1.

  - 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 Phaselift-ℓ_1

Theorem (Li, Sun and Chi, 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 for a fixed rank-r PSD matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, there exist some absolute constants $C_1 > 0$ and $0 < s_0 < 1$ such that as long as $m > C_1 n r^2, s \leq s_0/r,$ the solution to the proposed algorithm satisfies

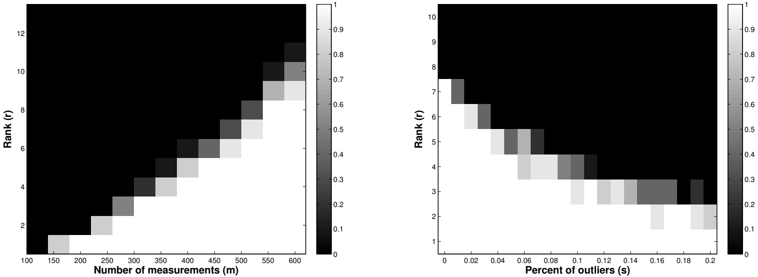
$$\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}\right\|_{\mathrm{F}} \leq C_2 \frac{r\epsilon}{m},$$

with probability exceeding $1 - e^{-\gamma m/r^2}$ for some constants C_2 and γ.

- Proof by dual certificate construction.
- Exact recovery when $\boldsymbol{w} = 0$ as long as $m \gtrsim nr^2$ and $s \lesssim 1/r$.
- When $r = 1$ we obtain near-optimal guarantee, which recovers the result by Hand for the phase retrieval case[¶];

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

Numerical Performance: Outlier robustness



(a) Fix s, r vs m

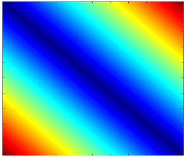
(b) Fix m, r vs s

Figure: Phase transitions of PSD matrix recovery with respect to (a) the number of measurements and the rank, with 5% of measurements corrupted by arbitrary standard Gaussian variables; (b) the percent of outliers and the rank, when the number of measurements is $m = 400$, where $n = 40$.

Robust recovery of Toeplitz PSD Matrices

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

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{M}\succeq 0,\ \text{Toeplitz}}{\operatorname{argmin}} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{M})\|_1.$$



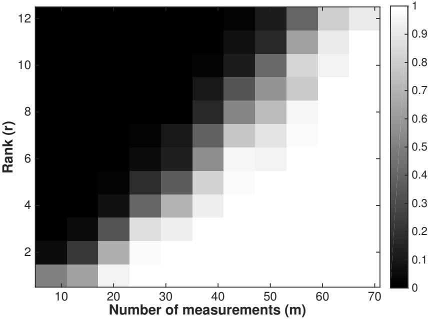


Figure: Phase transitions of low-rank Toeplitz PSD matrix recovery w.r.t. the number of measurements and the rank with 5% of measurements corrupted by standard Gaussian variables, when $n = 64$.

Non-convex approach based on factored model

Can we reduce the computational complexity?

- If rank($\boldsymbol{X}$) is known a priori as r, using the Cholesky factorization $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$ where $\boldsymbol{U} \in \mathbb{R}^{n \times r}$, one can directly recover $\boldsymbol{U}$:

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

  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

Rank-1 case (phase retrieval):

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2 + \eta_i + w_i, \quad i = 1, \ldots, m$$

where $\|\boldsymbol{\eta}\|_0 = s \cdot m$ is the outliers, $\boldsymbol{w}$ is the additive noise.

Exciting developments (without outliers) – all following the same recipe:

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

Provable near-optimal performance for Gaussian measurement model:

- Statistically: $m = O(n)$ near-optimal sample complexity
- Computationally: geometric 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

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

- 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

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

for TWF.

‖with some details hiding

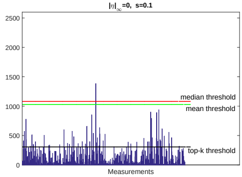
Robust phase retrieval via median-truncation

Need better strategy to eliminate outliers!

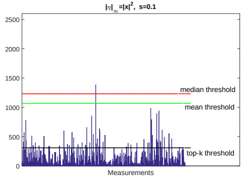
Key approach: "median-truncation"

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

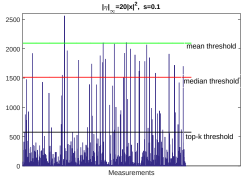
Median is more stable than mean and top-k truncation (which truncates a fixed amount of samples) for various levels of outliers.



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

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

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

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

  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:**

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

where

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

with $r_i^{(t)} = |y_i - (a_i^T z^{(t)})^2|$.

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

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

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

Performance guarantee of median-TWF

Theorem (Zhang, Chi and Liang, 2016)

Consider the model $y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2 + w_i + \eta_i$, *where* $\|\boldsymbol{w}\|_\infty \le c_1 \|\boldsymbol{x}\|^2$ *and* $\|\boldsymbol{\eta}\|_0 \le sm$. *If* $m \ge c_0 n \log n$ *and* $s < s_0$, *then with prob.* $1 - c_1 \exp(-c_2 m)$, *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 $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ *and some constants* $c_0, c_1, c_2 > 0$ *and* $0 < \rho < 1$.

- **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

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 $\{X_i\}_{i=1}^m$.

- *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.

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

- *Bound median by quantiles of clean samples:* Consider *clean samples* $\{\tilde{X}_i\}_{i=1}^m$ and *contaminated samples* $\{X_i\}_{i=1}^m$. Then

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

Proof sketch

Lemma (Concentration of median)

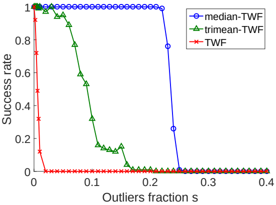
If $m > c_0 n \log n$*, then with probability at least* $1 - c_1 \exp(-c_2 m)$*, there exist constants* β *and* β' *such that*

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

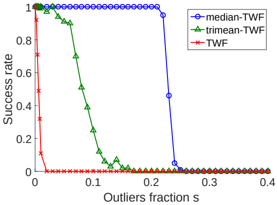
holds for all $\boldsymbol{z}, \boldsymbol{h} := \boldsymbol{z} - \boldsymbol{x}$ *satisfying* $\|\boldsymbol{h}\| < 1/11 \|\boldsymbol{z}\|$.

- A similar property is established for the *mean* when $m = O(n)$;
- here we lose a factor of $\log n$ due to working with the median.

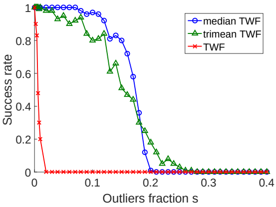
Numerical experiments with median-TWF



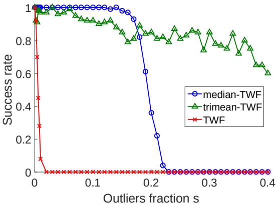
(a) $\|\boldsymbol{\eta}\|_\infty = 0.1\|\boldsymbol{x}\|^2$



(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.

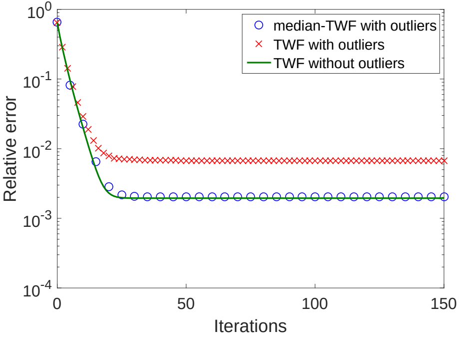


Figure: Relative error of median-TWF vs. TWF w.r.t. iteration when $s = 0.1$, $\|\boldsymbol{w}\|_\infty = 0.01\|\boldsymbol{x}\|^2$, and $\|\boldsymbol{\eta}\|_\infty = \|\boldsymbol{w}\|$.

Conclusion

We have discussed two approaches for combating outliers:

- **Convex** optimization based on PhaseLift-ℓ_1:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{X} \succeq 0}{\operatorname{argmin}} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_1$$

- **Non-convex** optimization based on median-TWF:

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{\mu}{m} \sum_{i=1}^{m} \nabla \ell(y_i, \boldsymbol{z}^{(t)}) \mathbb{1}_{\mathcal{E}_i}$$

  - **No prior knowledge** of outliers are required: we can run these algorithms as if outliers do not exist;
  - **Exact recovery** guarantees for Gaussian measurement model are obtained, even with a constant proportion of arbitrary outliers;
  - **Stability** against additional bounded noise.

Work in progress: extending median-TWF to the low-rank setting.

References

1. Outlier-Robust Recovery of Low-rank Positive Semidefinite Matrices From Magnitude Measurements, ICASSP 2016
2. Provable Non-convex Phase Retrieval with Outliers: Median Truncated Wirtinger Flow, ICML 2016

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