Residual-based Sampling for Online Outlier Robust PCA

Tianhao Zhu Jie Shen

Stevens Institute of Technology

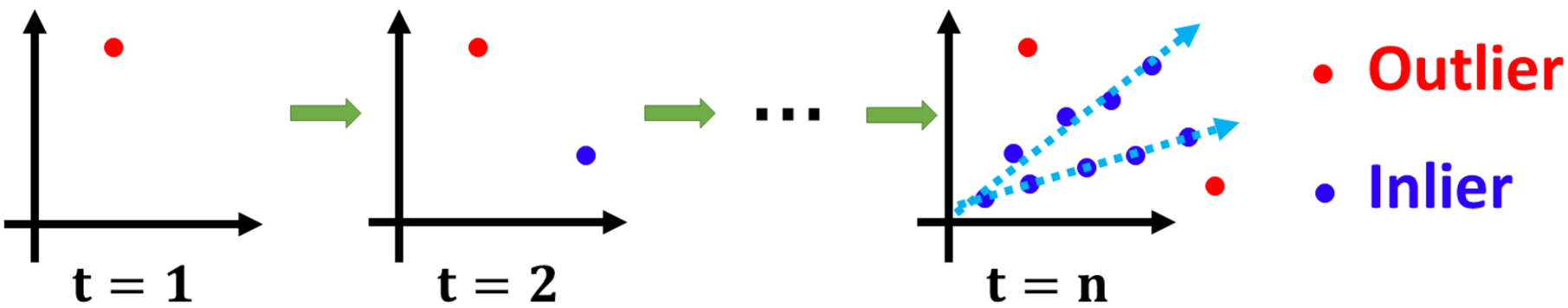
July 19, 2022

Problem Setup

This Work: The **Online Outlier Robust** PCA for data with z outliers:

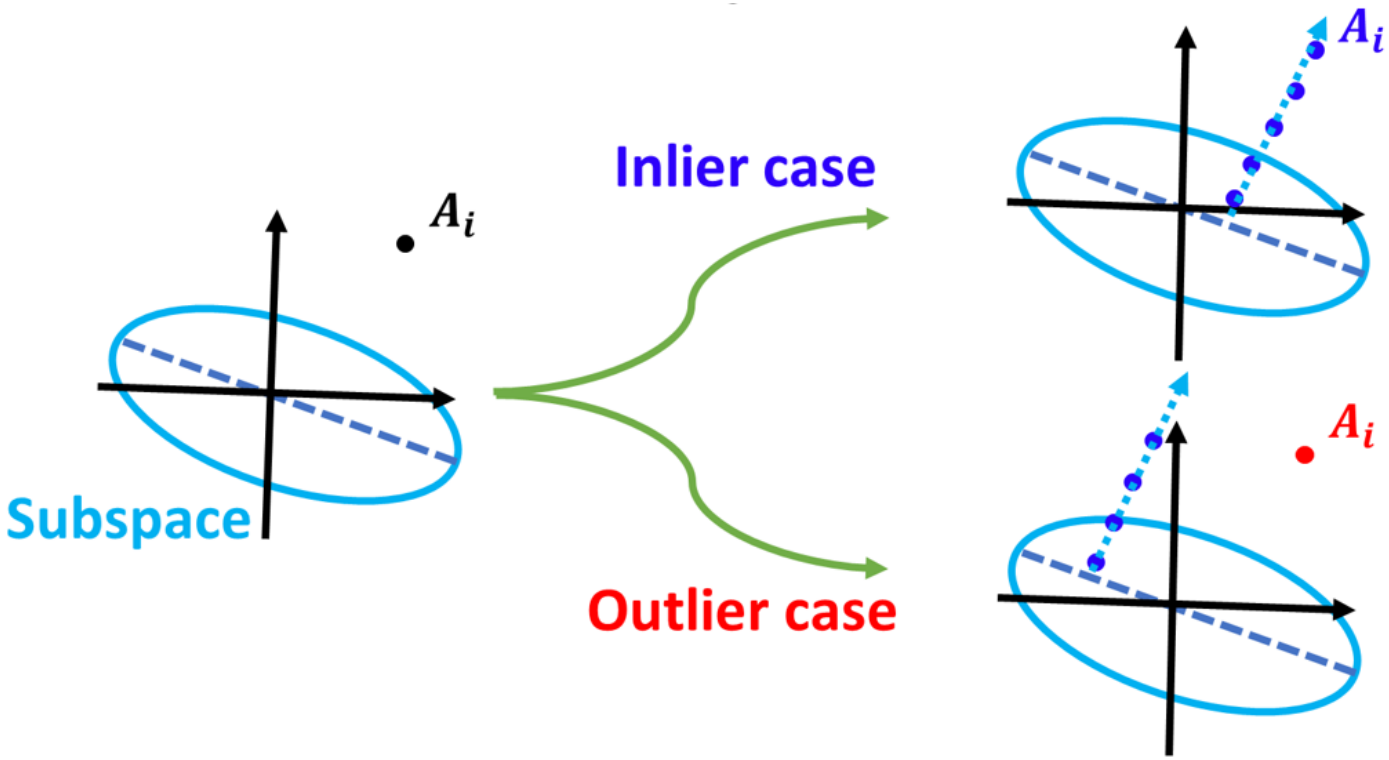
$$A = A_{\text{in}} \cup A_{\text{out}}, \quad k \ll z \ll n$$

$$\text{OPT}_k = \min \|A_{\text{in}} - UV\|_F^2, \text{s.t. } |A_{\text{out}}| \leq z.$$



Challenge

When we encounter a point far from the current subspace, we are not sure if it is **inlier** or **outlier**.



Algorithm

Residual-based Sampling:

Residual norm: $\left\|\prod_U^{\perp} A_i\right\| = \left\|A_i - UU^T A_i\right\|$.

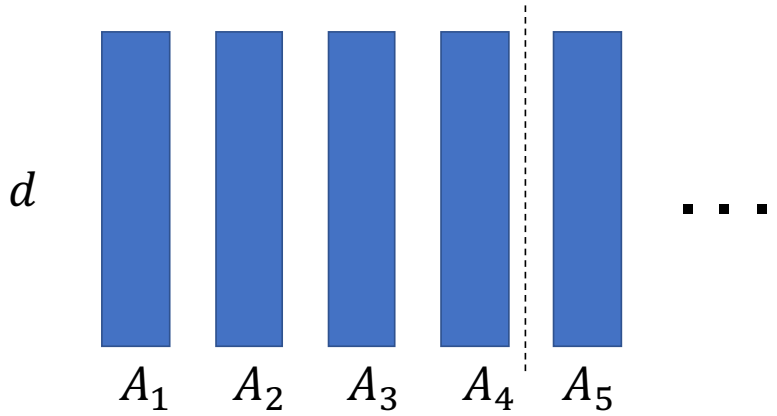
– target error $\xi \geq \mathrm{OPT}_k$.

Key: Residual norm is tied to objective contribution.

– Residual norm $\geq \frac{\xi}{k}$ ==> A new direction is formed.

– Residual norm $< \frac{\xi}{k}$ ==> Contribution to objective is small.

Framework



Algorithm

Threshold based Outliers Removal:

Problem: Distant inliers and outliers are indistinguishable.

Key: Each dimension of the subspace contains at least z/k inliers.
– Residual norm $\geq \frac{\xi}{z}$ ==> Marked outlier.
– Residual norm $< \frac{\xi}{z}$ ==> Marked inlier.

Three categories:

– Residual norm $\leq \frac{\xi}{z}$ ==> **Non-informative inlier**.
Linearly combine the point.

– Residual norm between $\frac{\xi}{z}$ and $\frac{\xi}{k}$ ==> **Non-informative outlier**.
Linearly combine the point with probability $O(k \left\| \prod_U^{\perp} A_i \right\|^2 / \xi)$**.**

– Residual norm $\geq \frac{\xi}{k}$ ==> **Informative outlier**.
Add it to the subspace with probability k/z**.**

Contributions

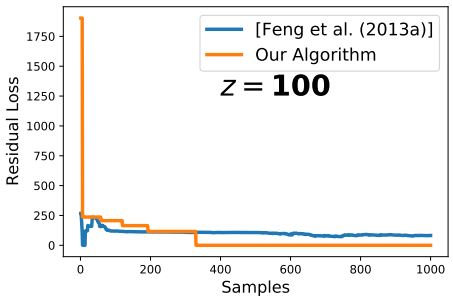
The first to provide the **provable algorithm** using sampling method.

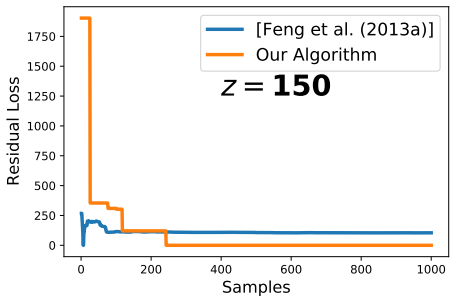
Work	Online?	Outlier-robust?	Error
[Xu et al., 2012]	✗	✓	OPT_k
[Bhaskara et al., 2019]	✓	✗	$O((\log \frac{\|A_{\text{in}}\|_F^2}{\xi})^2 \cdot \xi)$
[Feng et al., 2013]	✓	✓	unknown
Algorithm 1	✓	✓	$O((\log \frac{\|A_{\text{in}}\|_F^2}{\xi})^2 \cdot \xi)$
Algorithm 2	$\approx$	✓	$\mathrm{OPT}_k + \epsilon\xi$

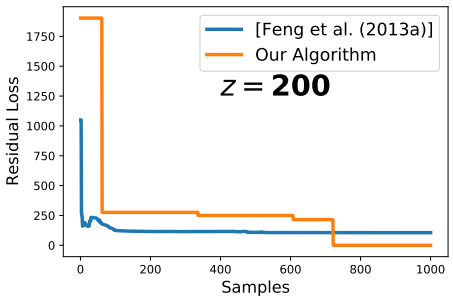
Experiments

Benchmark model [Feng et al. 2013]: Updating subspace using **SVD**.

$n = 1000$, $d = 500$, $k = 5$







Algorithm	Embedding Dimension	Marked Outliers ($z = 100$)	Execution Time (s)
Our algorithm	12	132	7.37
[Feng et al., 2013]	5	29	41.52

Conclusion

Summary

Our Online ORPCA is fast, outlier robust and returns non-asymptotic results.

Thank you!

See you in the Poster Session 1, Hall E #1108.

Reference


Bhaskara, A., Lattanzi, S., Vassilvitskii, S., and Zadimoghaddam, M. (2019).
Residual based sampling for online low rank approximation.
In *Proceedings of the 60th IEEE Annual Symposium on Foundations of Computer Science*, pages 1596–1614.


Feng, J., Xu, H., Mannor, S., and Yan, S. (2013).
Online PCA for contaminated data.
In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 764–772.


Xu, H., Caramanis, C., and Sanghavi, S. (2012).
Robust PCA via outlier pursuit.
IEEE Transactions on Information Theory, 58(5):3047–3064.