Robust Streaming PCA

Daniel Bienstock[1], Minchan Jeong[2], Apurv Shukla[1], Se-Young Yun[2]
(Authors are ordered alphabetically)

[1] IEOR Department, Columbia University, United States of America
[2] Kim Jaechul Graduate School of AI, KAIST, Republic of Korea

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

KAIST AI
Graduate School of AI

O/i Optimization and Statistical Inference LAB

NEURAL INFORMATION PROCESSING SYSTEMS

TL; DR

Streaming principal component analysis when the stochastic data -generating model is subject to perturbations.

Motivation

Principal component analysis (PCA) is one of the most extensively studied methods for obtaining the low-dimensional representation of observed data. Streaming PCA focuses on the online PCA algorithms with data-generating model.

Most algorithms assume that all the observations belong to the same low-dimensional space. However, this situation is unlikely when the unknown/unexplored alterations corrupt a system's observations. For instance:

- Typical data attacks on power grids can significantly change the estimated covariance matrix of the data observed from sensors.
- PCA can be used to explain stock returns in terms of macroeconomic factors, which varies with the time.

In all these scenarios, **the underlying data-generating model changes with time**, and the decisions are based on identifying the changed model.

Non-Stationary Environment

Time-Variant Spiked Covariance Model: We consider the time-dependent environment:

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \mathbf{A}_t\mathbf{A}_t^\top + \sigma^2\mathbf{I}_{p\times p})$$

where $\mathbf{A}_t \in \mathbb{R}^{p\times k}$ can vary with time. Standard spiked covariance model[1] indicates the case $\mathbf{A}_t = \mathbf{A}$.

Task: Algorithm ϕ should recover top-k principal components of covariance matrix at the final time step T.

Temporal Uncertainty Set: We only allow the sequence of matrices $\mathbf{A}_t\mathbf{A}_t^\top$ that lie in an temporal uncertainty set defined as:

$$\mathrm{Tu}(\Gamma,\delta) := \left\{ (\mathbf{A}_t)_{t=1}^T \;:\; s_k(\mathbf{A}_t\mathbf{A}_t^\top) \geq \delta\,,\, \|\mathbf{A}_t\mathbf{A}_t^\top - \mathbf{A}_{t-1}\mathbf{A}_{t-1}^\top\| \leq \Gamma \right\}$$

Metric and Algorithm Optimality

Estimation Error: For streaming algorithm ϕ and the sampled data stream $\mathcal{X} = (\mathbf{x}_t)_{t=1}^T \sim \mathscr{A} = (\mathbf{A}_t)_{t=1}^T \in \mathrm{Tu}(\delta,\Gamma)$, we consider the metric $d(\mathrm{ran}(\mathbf{A}_T), \phi_{\mathcal{X}})$, where d is the matrix 2-norm between projectors.

Performance of Algorithm: For each streaming algorithm ϕ, the maximum expected error of ϕ is defined as $\mathscr{R}^\phi := \sup_{\mathscr{A}\in\mathrm{Tu}(\delta,\Gamma)} \mathbb{E}_{\mathcal{X}\sim\mathscr{A}}\left[d\left(\mathrm{ran}(\mathbf{A}_T), \phi_{\mathcal{X}}\right)\right]$.

Fundamental Lower Bound: Fundamental minimax lower bound is the infimum over maximum expected error $\mathscr{R}^* := \inf_{\phi\in\Phi} \mathscr{R}^\phi$.

Rate Optimal Algorithm: Streaming algorithm ϕ is rate optimal if $\mathscr{R}^\phi \leq C\cdot\mathscr{R}^*$, where C is a constant independent with T, δ, p, k, and Γ.

Contributions

On the **non-stationary streaming PCA environment**, we provide :

1. **Fundamental minimax lower bound**
   - For $T = \mathcal{O}(\Gamma^{-2/3})$, the minimax error decreases as $\mathcal{O}(p^{1/2}T^{-1/2})$.
   - On the other hand, for $T = \Omega(\Gamma^{-2/3})$, the error stagnates to $\mathcal{O}(p^{1/3}\Gamma^{1/3})$, and does not decrease upon collecting more observations.
2. **Analysis for two streaming PCA algorithms**
   - There exists regime for the best learning parameters.
   - Noisy power method is rate optimal under mild conditions.
   - We validate some findings via numerical experiments.

1. Minimax Lower Bound

When $\mathscr{A} = (\mathbf{A}_t)_{t=1}^T$ belongs to the temporal uncertainty set $\mathrm{Tu}(\delta,\Gamma)$, an algorithm designed to recover the principal components of $\mathbf{A}_T$ from the observations cannot guarantee converges-to-zero estimation error.

Theorem 1. Assume $\delta > \Gamma \geq 0$ and $p > 2k+1$. Then:

$$\mathscr{R}^* \geq \Theta\left(\min\left\{1\,,\frac{1}{\sqrt{T}}\left(\frac{p\sigma^2(\sigma^2+\delta)}{\delta^2}\right)^{1/2} + \left(\frac{\Gamma}{\delta}\right)^{1/3}\left(\frac{p\sigma^2(\sigma^2+\delta)}{\delta^2}\right)^{1/3}\right\}\right)$$

For standard streaming PCA problem ($\Gamma = 0$), the fundamental limit is expected $\Theta(1/\sqrt{T})$ dependence [2,3].

On the other hand, for the case ($\Gamma > 0$), only the last T_c observations are essential for estimation since the information quickly becomes stale in a dynamic environment. ($T_c = (\Gamma/\delta)^{-2/3}\left(p\sigma^2(\sigma^2+\delta)/\delta^2\right)^{1/3}$)

2. Algorithm Analysis

In this section, $\mathscr{M} = 2(k\delta + p\sigma^2)(1+\Theta(\log(pT^2)/T))$ **and** $\mathscr{V} = 2\mathscr{M}(\delta+\sigma^2)$.

Update Rule of Noisy Power Method [4]:

$$\hat{\mathbf{U}}(\ell) \leftarrow \text{Gram-Schmidt}\left(\frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\mathbf{x}_t\mathbf{x}_t^\top\cdot\hat{\mathbf{U}}(\ell-1)\right)$$

Since only the last observations are essential, it becomes imperative to find the block size B that can be used to recover the principal components.

Theorem 2. Assume that $\delta \geq 0.71\sigma^2$and $\Gamma = \mathcal{O}(\delta^3/(\mathscr{V}\log(2pT^2)))$.
For $B = \Theta(\mathscr{V}^{1/3}\log(2pT^2)^{1/3}\Gamma^{-2/3})$, we have:

$$\mathscr{R}^{\mathrm{NPM}} = \tilde{\mathcal{O}}\left(\left(\frac{\Gamma}{\delta}\right)^{1/3}\left(\frac{(p\sigma^2+k\delta)(\sigma^2+\delta)}{\delta^2}\right)^{1/3}\right)$$

If $T = \Omega(\max(T_c, \delta(p\sigma^2)^{-1}))$, $\Gamma = \Omega((c^{\Omega(p-k+1)} + e^{-\Omega(p)})\delta^2(p\sigma^2)^{-1})$, and $s_1(\mathbf{A}_t\mathbf{A}_t^\top) = \Theta(\delta)$.

Noisy power method [4] becomes order-wise identical to the fundamental limit established in the Theorem 1, when $p\sigma^2$ dominates $k\delta$. This regime is the case of noisy practical situations.

2. Algorithm Analysis (Continues)

Update Rule of Oja's algorithm [5]:

$$\hat{\mathbf{U}}(t) \leftarrow \text{Gram-Schmidt}\left((\mathbf{I} + \zeta\mathbf{x}_t\mathbf{x}_t^\top)\cdot\hat{\mathbf{U}}(t-1)\right)$$

We establish similar analysis for the Oja's algorithm using virtual block size $B_\zeta = \lceil\zeta^{-1}\rceil$. The regime for optimal inverse learning rate ζ^{-1} becomes:

$$\zeta^{-1} = \Theta(\mathscr{M}^{2/3}\log(pT^2)^{1/3}\Gamma^{-2/3}).$$

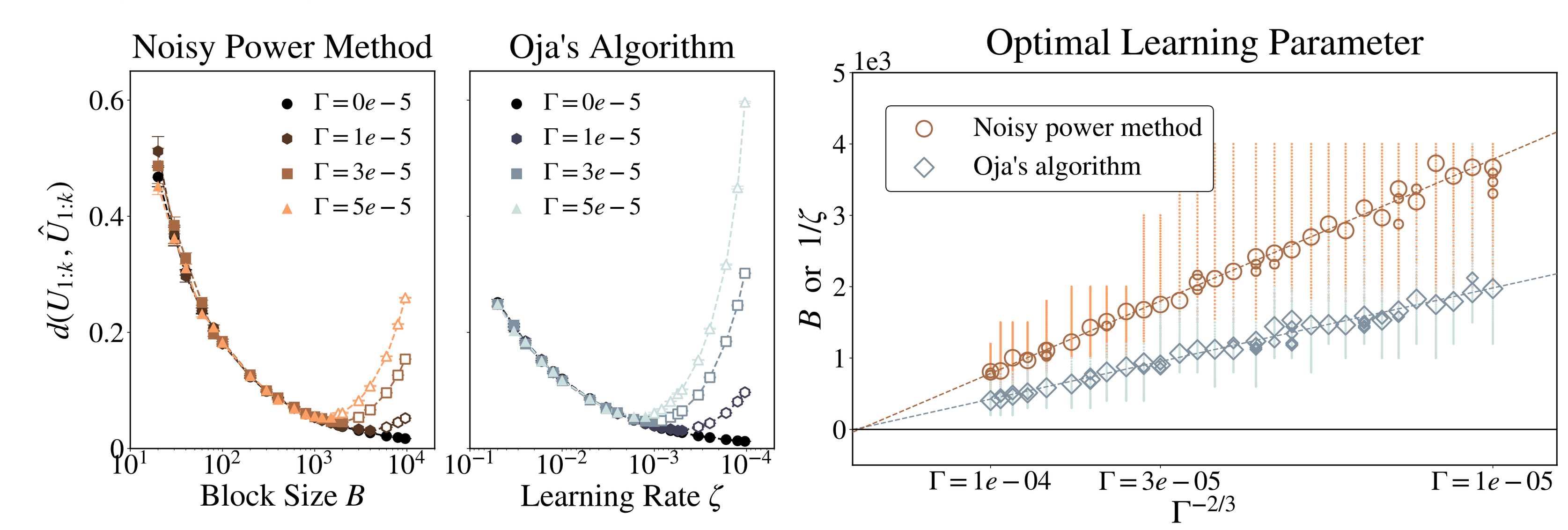
Unlike the noisy power method, the upper bound for Oja's algorithm $\mathcal{O}(p^{2/3})$ is not rate optimal (See Theorem 3 of the paper for details). This theoretical gap occurs because the proof uses different (multiplicative) matrix concentration inequalities [6], different from the matrix Bernstein inequality used for noisy power method analysis.

Experiments

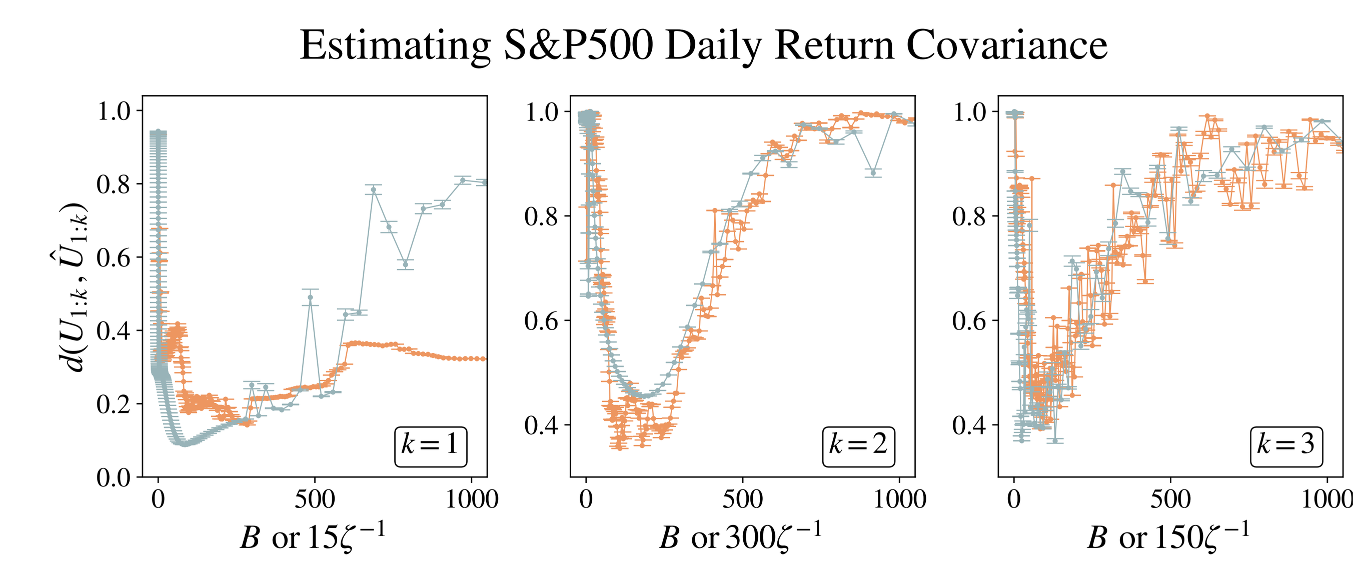
We verify the below findings via experiments:

- Existence of the optimal regime for block size B and the learning rate ζ .
- $\Gamma^{-2/3}$ dependencies of that optimal learning parameters B and ζ^{-1}.

- **Synthetic Experiments**



- **S&P500 Return Covariance Analysis**



References

[1] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295 – 327, 2001.
[2] Tony Cai, Zongming Ma, and Yihong Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3):781–815, Apr 2015.
[3] Vincent Q. Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905 – 2947, 2013.
[4] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. *In Advances in Neural Information Processing Systems*, pages 2861–2869, 2014
[5] Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.
[6] De Huang, Jonathan Niles-Weed, Joel A Tropp, and Rachel Ward. Matrix concentration for products. *arXiv preprint arXiv*:2003.05437, 2020