On Distributed Averaging for Stochastic k-PCA

Aditya Bhaskara, Maheshakya Wijewardena

University of Utah

October 26, 2019

Stochastic PCA

PCA

Given a collection of data points $x1, x2, \ldots, xn$, find a subspace U of dimension k such that captures the most mass of the points.

Find $U \in \mathbb{R}^{d \times k}$ with orthonormal columns such that $||\Sigma U||_F$ is maximized where $\Sigma = \sum_i x_i x_i^T$ is the covariance matrix.

Stochastic PCA

$x_1, x_2, \ldots$ are sampled from an unknown distribution $\mathcal{D}$ over points in $\mathbb{R}^d$, and the goal is to find the top k directions $U \in \mathbb{R}^{d \times k}$ that maximizes $||\Sigma U||_F$ where $\Sigma = \mathbb{E}_{x \sim \mathcal{D}} x x^T$ is the the distribution covariance matrix (or the second moment matrix).

Distributed Averaging

Framework:

- There are m machines each of which receives n samples.
- Each machine performs some computation and sends an $O(k)$-size summary of the local dataset to the central server.
- Central server performs an aggregation and computes the desired quantity.

Goal: achieve the same effect as the server computing using mn samples by itself.

Distributed Averaging

Previous results:

- Garber et. al.([2]) showed that appropriately signed average of the leading eigenvectors sent by each machine gives a good approximation to the leading eigenvector of the distribution covariance matrix.

$$1 - \left(\frac{\tilde{u_1}^T u_1}{||\tilde{u_1}||}\right)^2 = O\left(\frac{1}{mn} + \frac{1}{n^2}\right), n \geq O\left[\frac{1}{(\lambda_1 - \lambda_2)^2}\right]$$

- Fan et. al.([1]) show that the average of the covariance matrices generated by top k subspaces sent by each machine give a good approximation for the top k subspace of which the majority of mass lies in the distribution.

$$||\tilde{U_k}\tilde{U_k}^T - U_k U_k||_F^2 = O\left(\frac{1}{mn} + \frac{1}{n^2}\right), n \geq O\left[\frac{1}{(\lambda_k - \lambda_{k+1})^2}\right]$$

Our Contributions

- Assuming $n \geq O\left(\frac{1}{(\lambda_k - \lambda_{k+1})^2}\right)$, our algorithm can compute top k leading eigenvalues and eigenvectors with a good approximation guarantee.

$$1 - (\tilde{u}_i{}^T u_i)^2 \leq \frac{1}{\delta_i^2} \cdot O\left(\frac{1}{n^2} + \frac{1}{mn}\right)$$

$$|\widetilde{\lambda}_i - \lambda_i| \leq \frac{1}{\delta_i} \cdot O\left(\frac{1}{n} + \frac{1}{\sqrt{mn}}\right)$$

  where $\delta_i := \min(\lambda_{i-1} - \lambda_i, \lambda_i - \lambda_{i+1})$

- If there exists a $k \in (k_0, k_1)$ such that $\lambda_k - \lambda_{k+1}$ is large enough, then using a single round of communicating k_1 vectors by each machine (i.e. space $O(k_1 d)$ per machine), we show an efficient way to find k.

Algorithm

Let the covariance matrix in machine j be $\hat{A}^{(j)} = \hat{U}\hat{\Sigma}\hat{U}^T$.
Let $\hat{V}^{(j)} = \hat{U}_k \hat{\Sigma}_k^{1/2}$.

Algorithm

1. **Local:** On each machine, compute the rank-k PCA of the empirical covariance matrix $\hat{A}^{(j)}$, and send $\hat{V}^{(j)}$ to the server.
2. **Server:** On the central server, compute $\tilde{A}_k = \frac{1}{m}\sum_{j=1}^{m} \hat{V}^{(j)}(\hat{V}^{(j)})^T$. Then output the top k eigenvalues and the corresponding eigenvectors of $\tilde{A}_k$.

Analysis of the algorithm

Let $A^* = \mathbb{E}[\hat{A}]$. Now,

$$||A_k - \tilde{A}_k|_F \leq ||A_k - A^*||_F + ||A^* - \tilde{A}_k||_F$$

$||A_k - A^*||_F$ is the bias and $||A^* - \tilde{A}_k||_F$ is the variance.

We prove that:

$$||A_k - A^*||_F \leq O(1/n)$$
$$||A^* - \tilde{A}_k||_F \leq O(1/\sqrt{mn})$$

The statement about eigenvalues follows from Weyl's inequality while the statement about eigenvectors follow from David-Kahan $\sin\theta$-theorem.

Empirical Evaluation

Synthetic dataset:

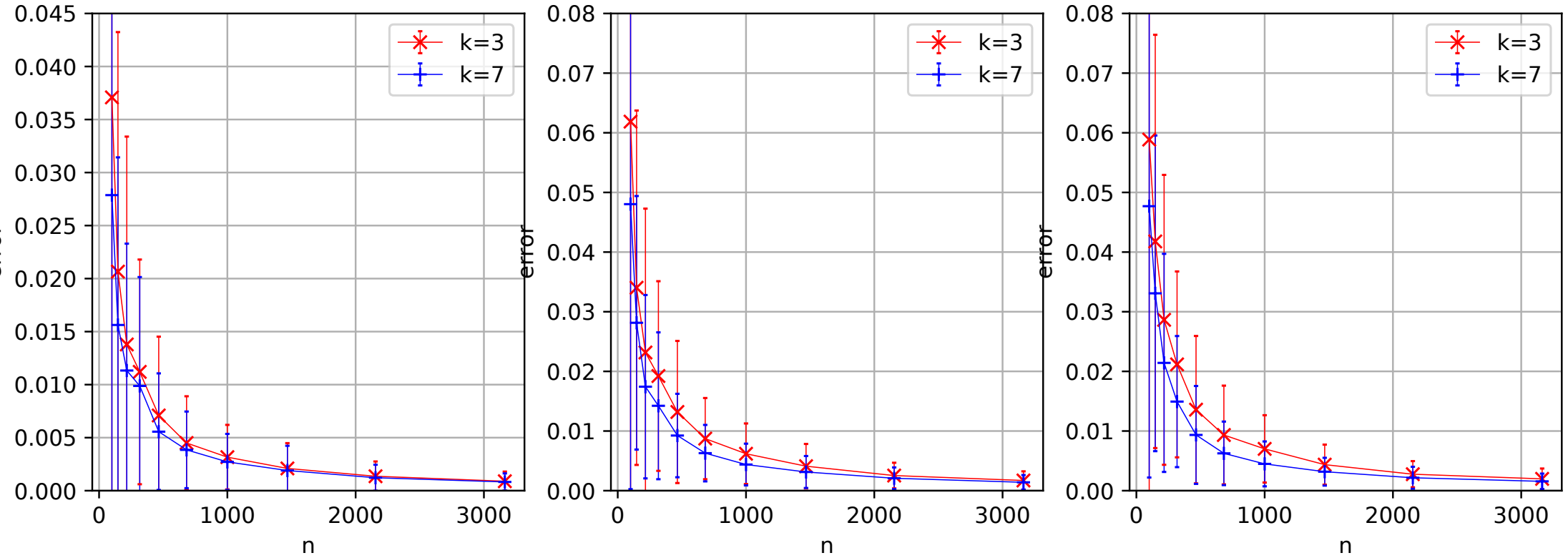

Figure: Estimation errors of first eigenvector (left), second eigenvector (middle), and third eigenvector (right) for $k = 3$ and $k = 7$ vs. samples size n per machine.

Empirical Evaluation

Real datasets:

Dataset	N	d	r	t
MNIST-small	20000	196	5	15
NIPS-papers	11463	150	5	15
FMA-music	21314	518	10	70

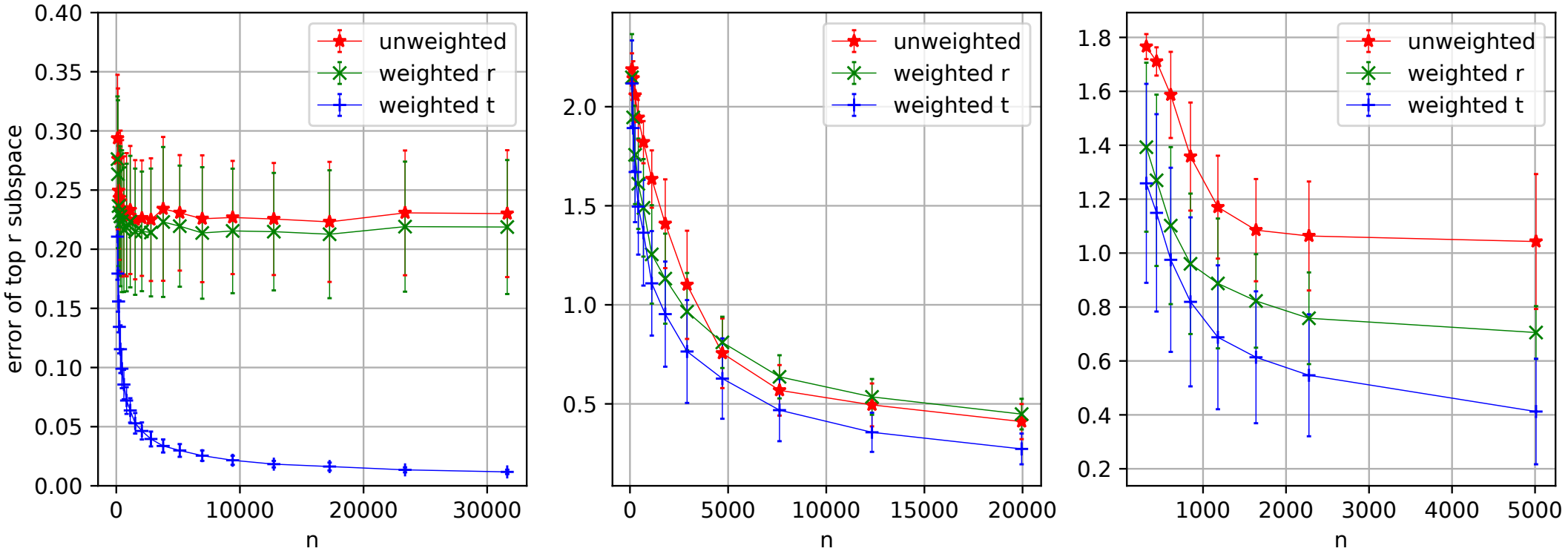

Figure: Estimation errors of top r subspace of MNIST-small dataset (left), NIPS-papers dataset (middle), FMA-music dataset (right) vs. unweighted, weighted r, weighted t averaging.

References

- Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. arXiv preprint arXiv:1702.06488, 2017.
- Communication-efficient algorithms for dis- tributed stochastic principal component analysis. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, pages 1203–1212. JMLR. org, 2017.