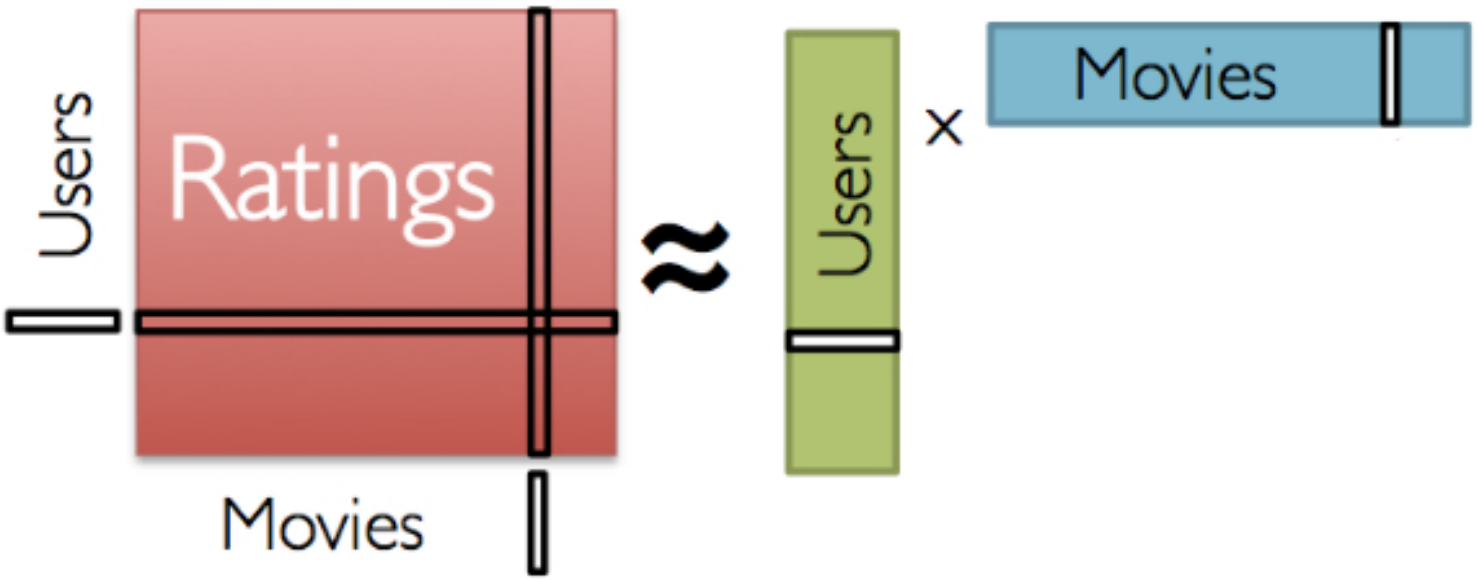
Estimating shared subspace with misalignment: the power and limitation of multiple data matrices

Yuepeng Yang

Mar 21, 2025

based on joint work with Cong Ma

Many data can be approximated as a low-rank matrix



$\mathbb{R}^{n \times d}$ $\mathbb{R}^{n \times r}$ $\mathbb{R}^{r \times d}$

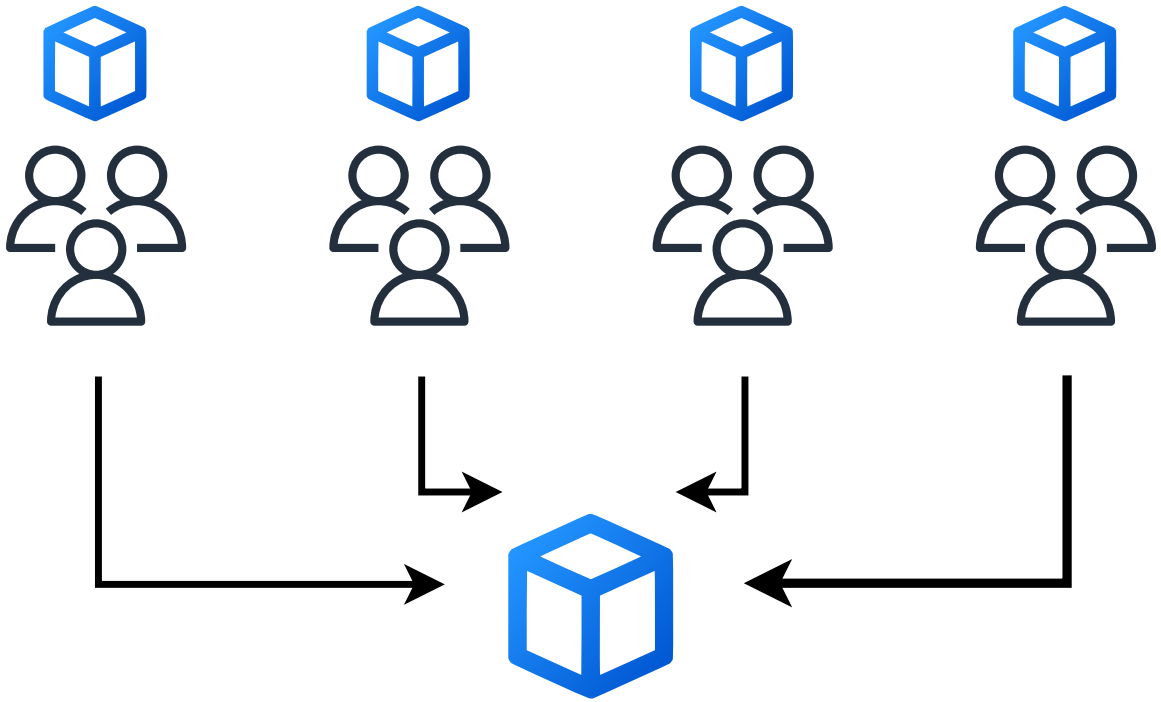
- Let $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ be a rank-r matrix, then it can be written as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^\top$$

  for some $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ such that $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_r$ and $\boldsymbol{V} \in \mathbb{R}^{d \times r}$

- $\boldsymbol{U}$ is unique up to rotation and represent $\mathrm{col}(\boldsymbol{A})$
- Given a noisy observation of $\boldsymbol{A}$, our goal is to estimate the column space $\boldsymbol{U}$

Learning shared data structure



Learning a shared structure using multiple data sources

Subspace estimation with shared component

- Consider K data matrices $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_K$ with shared column space

$$\boldsymbol{A}_k = \boldsymbol{U}\boldsymbol{V}_k^\top + \boldsymbol{E}_k$$

- $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ is the shared column space
- $\boldsymbol{V}_k \in \mathbb{R}^{d_k \times r}$ are the loading matrices
- $\boldsymbol{E}_k \in \mathbb{R}^{n \times d_k}$ is entry-wise i.i.d. Gaussian noise

Can we estimate $\boldsymbol{U}$ from $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_K$ statistically efficiently?

Stack-SVD

1. Stack all data matrices into a single rank-r matrix

$$\boldsymbol{A} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_K] = \boldsymbol{U}[\boldsymbol{V}_1^\top \ldots \boldsymbol{V}_K^\top] + [\boldsymbol{E}_1 \ldots \boldsymbol{E}_K]$$

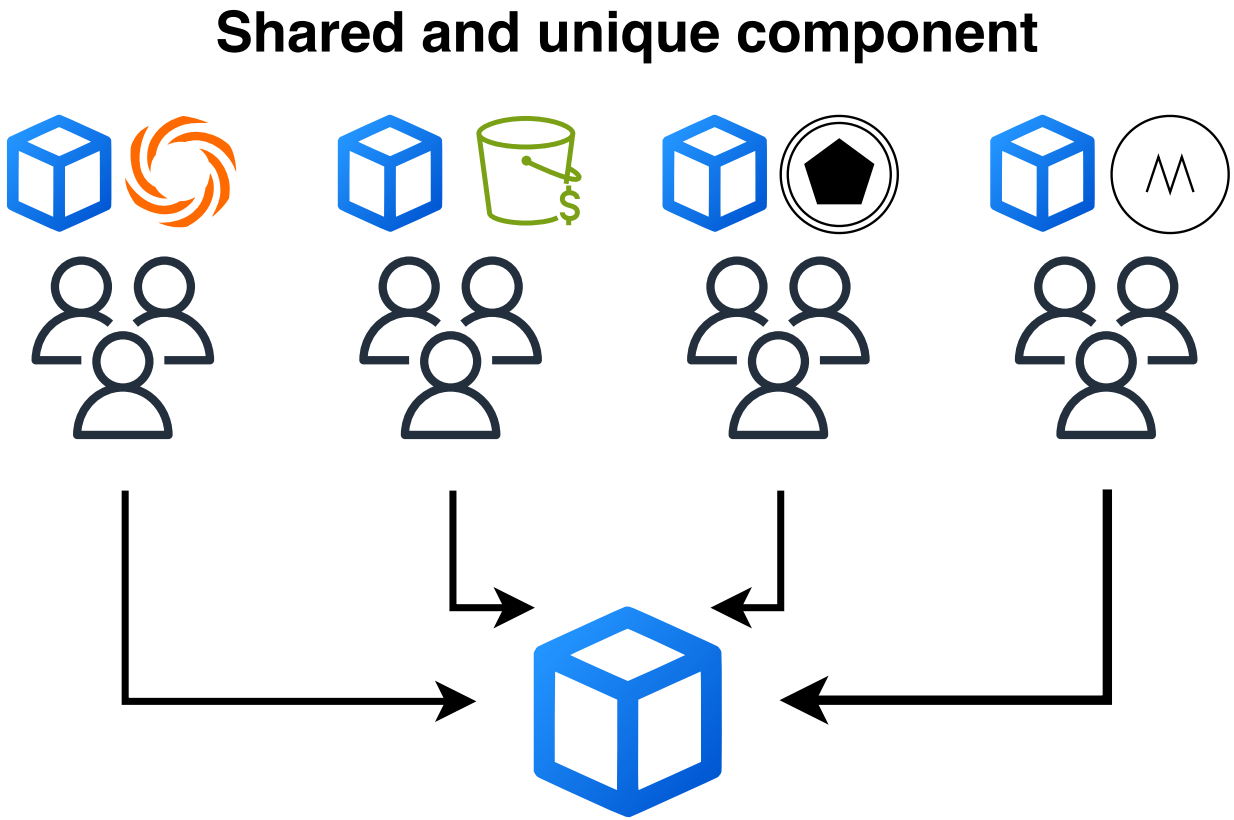
2. Compute $\widehat{\boldsymbol{U}}$, the top-r left singular vectors of $\boldsymbol{A}$

Let $\boldsymbol{E}_k \sim \mathcal{N}(0, \sigma^2)$ and for simplicity take $n = d_1 = \ldots = d_K$

Stack-SVD is minimax optimal:

$$\|\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\| \lesssim \sigma\sqrt{\frac{n}{K}}$$

In the setting where the column space is completely shared, we can use multiple data matrices to better estimate $\boldsymbol{U}$



Learning a shared structure using multiple data sources

Assume each data matrix is the sum of a component with shared column space and a component with unique column space

$$\boldsymbol{A}_k = \underbrace{\boldsymbol{U}\boldsymbol{V}_k^\top}_{\substack{\text{rank}-r \\ \text{shared component}}} + \underbrace{\boldsymbol{U}_k\boldsymbol{W}_k^\top}_{\substack{\text{rank}-r_k \\ \text{unique component}}} + \underbrace{\boldsymbol{E}_k}_{\text{Gaussian noise}}$$

- $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ is the shared column space
- $\boldsymbol{U}_k \in \mathbb{R}^{n \times r_k}$ are the unique column spaces, $\boldsymbol{U}_k \perp \boldsymbol{U}$
- $\boldsymbol{V}_k \in \mathbb{R}^{d_k \times r}$ and $\boldsymbol{W}_k \in \mathbb{R}^{d_k \times r_k}$ are the loading matrices
- $\boldsymbol{E}_k \in \mathbb{R}^{n \times d_k}$ is entry-wise i.i.d. Gaussian noise

Can we estimate $\boldsymbol{U}$ given $\boldsymbol{A}_k$ in presence of the unique components?

- Can we simply use Stack-SVD?
- Is the estimation of $\boldsymbol{U}$ fundamentally different when unique components exist?

Can we estimate $\boldsymbol{U}$ given $\boldsymbol{A}_k$ in presence of the unique components?

- Can we simply use Stack-SVD?
  No. It fails even in the noiseless case.
- Is the estimation of $\boldsymbol{U}$ fundamentally different when unique components exist?
  Yes.

Angle-based Joint and Individual Variation Explained (AJIVE) Feng et al. (2018)

AJIVE is a simple, non-iterative two-step spectral algorithm:

1. Estimate the shared + unique column space each $\boldsymbol{A}_k$:
   Let $\widetilde{\boldsymbol{U}}_k$ the top-$(r + r_k)$ left singular vectors of $\boldsymbol{A}_k$
2. Estimate the shared column space:
   Let $\widehat{\boldsymbol{U}}$ be the top-r eigenvectors of $\sum_{k=1}^{K} \widetilde{\boldsymbol{U}}_k \widetilde{\boldsymbol{U}}_k^\top$

- Angle-based: $\widehat{\boldsymbol{U}}$ minimizes its angles from $\{\mathrm{col}(\boldsymbol{A}_k)\}_{k=1}^{K}$
- AJIVE exactly recovers $\boldsymbol{U}$ in noiseless setting

Performance guarantee of AJIVE:

- Assumptions on the misalignment of the unique components
- The power and limitation of using multiple data matrices
- The role of the unique component

JIVE is not yet well-defined

$$\boldsymbol{A}_k = \underbrace{\boldsymbol{U}\boldsymbol{V}_k^\top}_{\substack{\text{rank}-r \\ \text{shared component}}} + \underbrace{\boldsymbol{U}_k\boldsymbol{W}_k^\top}_{\substack{\text{rank}-r_k \\ \text{unique component}}}$$

- Faithfulness: $\text{col}(\boldsymbol{U}) \subset \cap_{k=1}^K \text{col}(\boldsymbol{A}_k)$

  equiv. $\text{rank}(\boldsymbol{A}_k) = r + r_k$

- Exhaustiveness: $\cap_{k=1}^K \text{col}(\boldsymbol{A}_k) \subset \text{col}(\boldsymbol{U})$

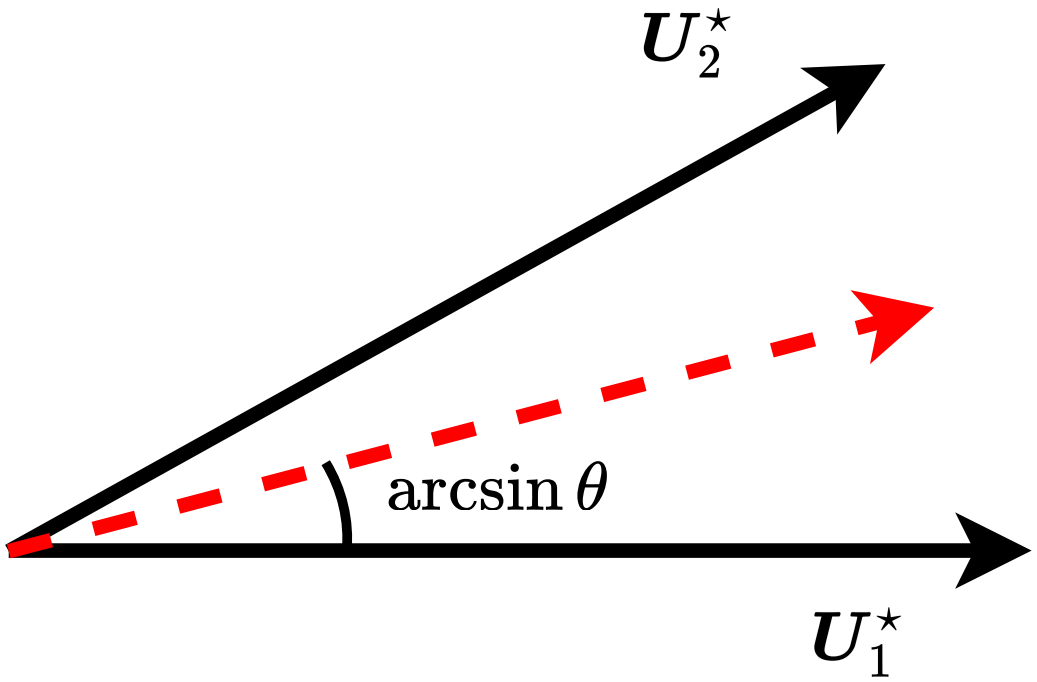
  equiv.1: $\underbrace{\left\| \frac{1}{K}\sum_{k=1}^K \boldsymbol{U}_k\boldsymbol{U}_k^\top \right\|}_{\text{Avg. Proj. Mat}} < 1$ equiv.2: $\cap_{k=1}^K \text{col}(\boldsymbol{U}_k) = \varnothing$

The misalignment parameter θ

Misalignment: $\| \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{U}_k \boldsymbol{U}_k^\top \| = 1 - \theta$

θ measures

- how hard it is to represent $\{\mathrm{col}(\boldsymbol{U}_k)\}_{k=1}^K$ with one direction
- how far away the unique column spaces are from being shared



The power of AJIVE is not well-understood theoretically

Naive approach (Wedin / Davis-Kahan) yields a suboptimal rate:

$$\|\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top - \boldsymbol{U}\boldsymbol{U}^\top\| \lesssim \frac{\sigma\sqrt{n}}{\theta}$$

We want to give a statistical analysis of AJIVE that

- shows the power of using multiple data matrices
- understand the role of the misalignment parameter θ

For simplicity suppose $n = d_1 = \cdots = d_K$, $r = r_1 = \cdots = r_K$,

Theorem 1: Upper bound (simplified for $K\theta > 1$)

Assuming $\sigma_{r+r_k}(\boldsymbol{A}_k) \geq 1$, $\sigma\sqrt{n} \leq \sqrt{\theta}$,

$$\left\|\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\right\| \lesssim \sigma\sqrt{\frac{n}{K} + \frac{r}{K\theta}} + \frac{\sigma^2 n}{\theta}$$

- $\sigma\sqrt{\frac{n}{K} + \frac{r}{K\theta}}$: first-order error that scales with $\frac{1}{\sqrt{K}}$
- $\frac{\sigma^2 n}{\theta}$: non-diminishing second-order error

AJIVE is first-order minimax optimal

Theorem 2: Minimax lower bound

Assuming $\sigma_{r+r_k}(\boldsymbol{A}_k) \geq 1$,

$$\inf_{\widetilde{\boldsymbol{U}}} \sup_{\{\boldsymbol{A}_k\}} \left\| \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top \right\| \gtrsim \sigma \sqrt{\frac{n}{K} + \frac{r}{K\theta}}$$

- $\sigma\sqrt{\frac{n}{K}}$: expected error without unique component
- $\sigma\sqrt{\frac{r}{K\theta}}$: additional mild error due to misalignment

AJIVE is first-order minimax optimal

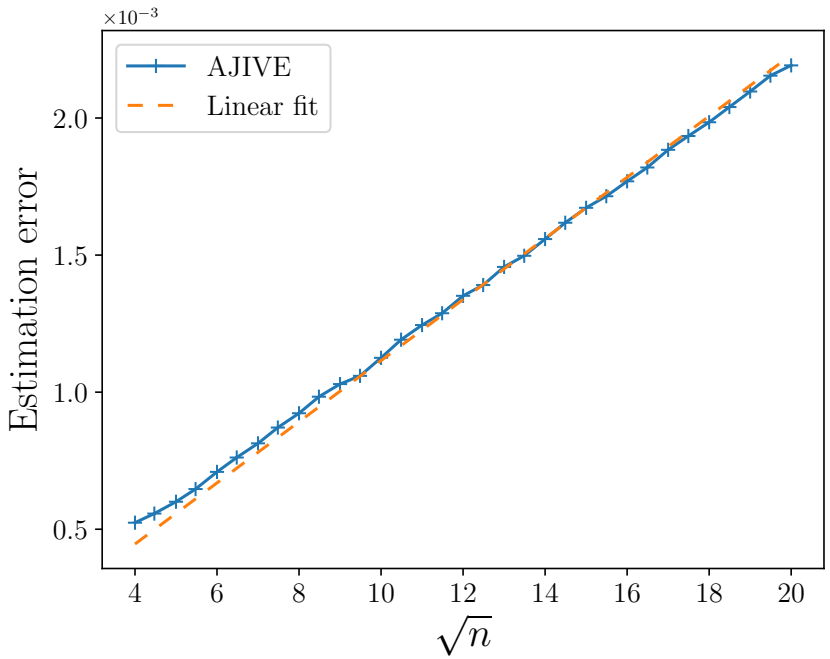
Theorem 3: Minimax lower bound

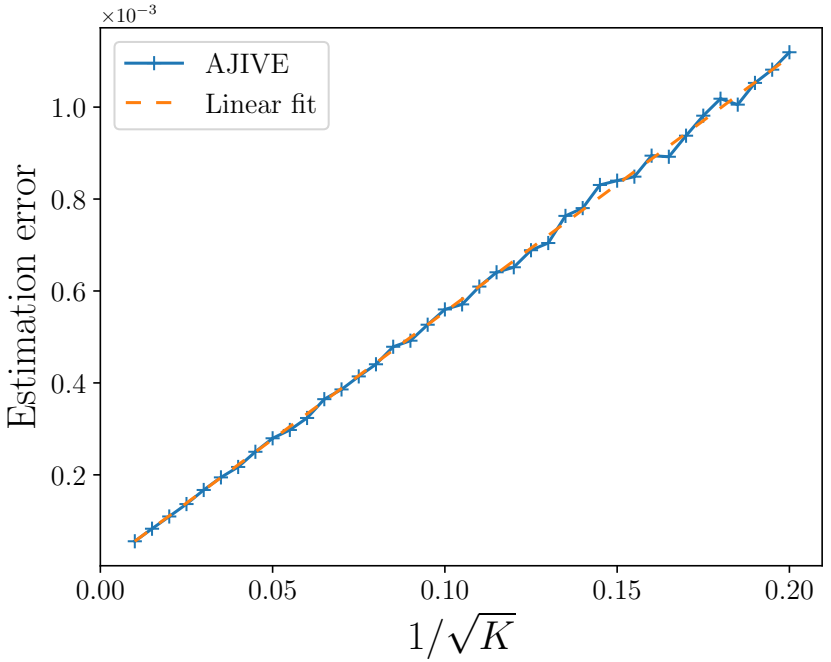
Assuming $\sigma_{r+r_k}(\boldsymbol{A}_k) \geq 1$,

$$\inf_{\widetilde{\boldsymbol{U}}} \sup_{\{\boldsymbol{A}_k\}} \left\| \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top \right\| \gtrsim \sigma \sqrt{\frac{n}{K} + \frac{r}{K\theta}}$$

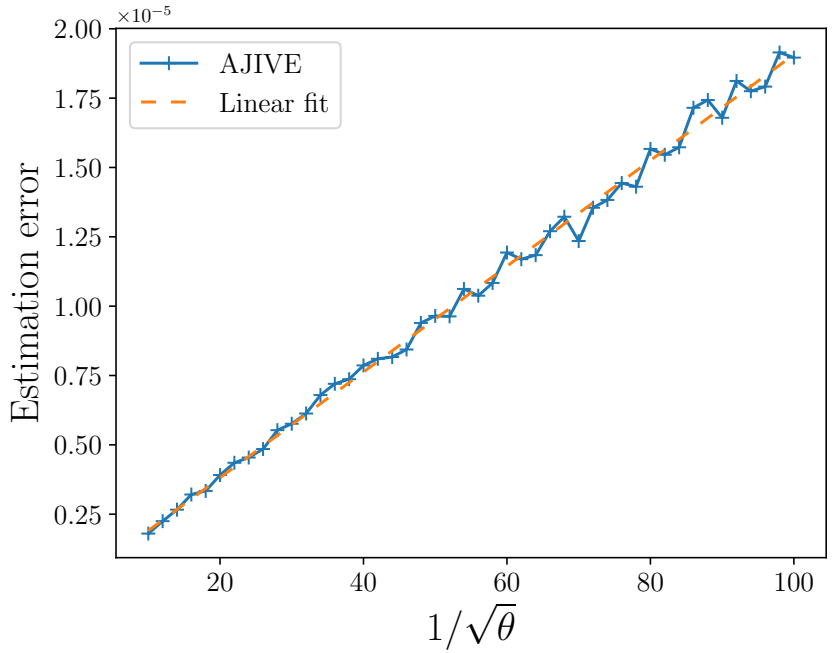
- $\sigma\sqrt{\frac{n}{K}}$: expected error without unique component
- $\sigma\sqrt{\frac{r}{K\theta}}$: additional mild error due to misalignment

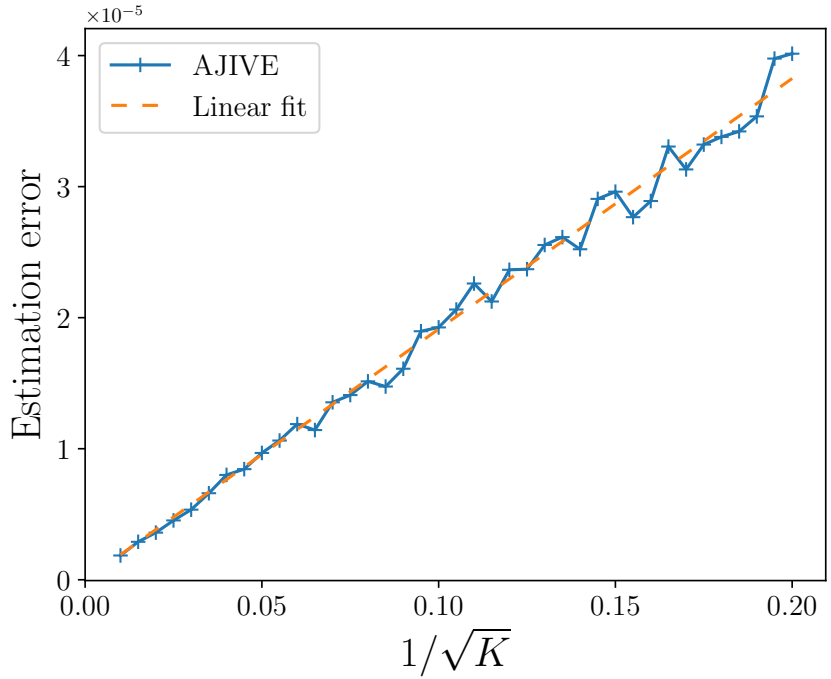
Large θ (0.5): $\|\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\| \propto \sqrt{n/K}$



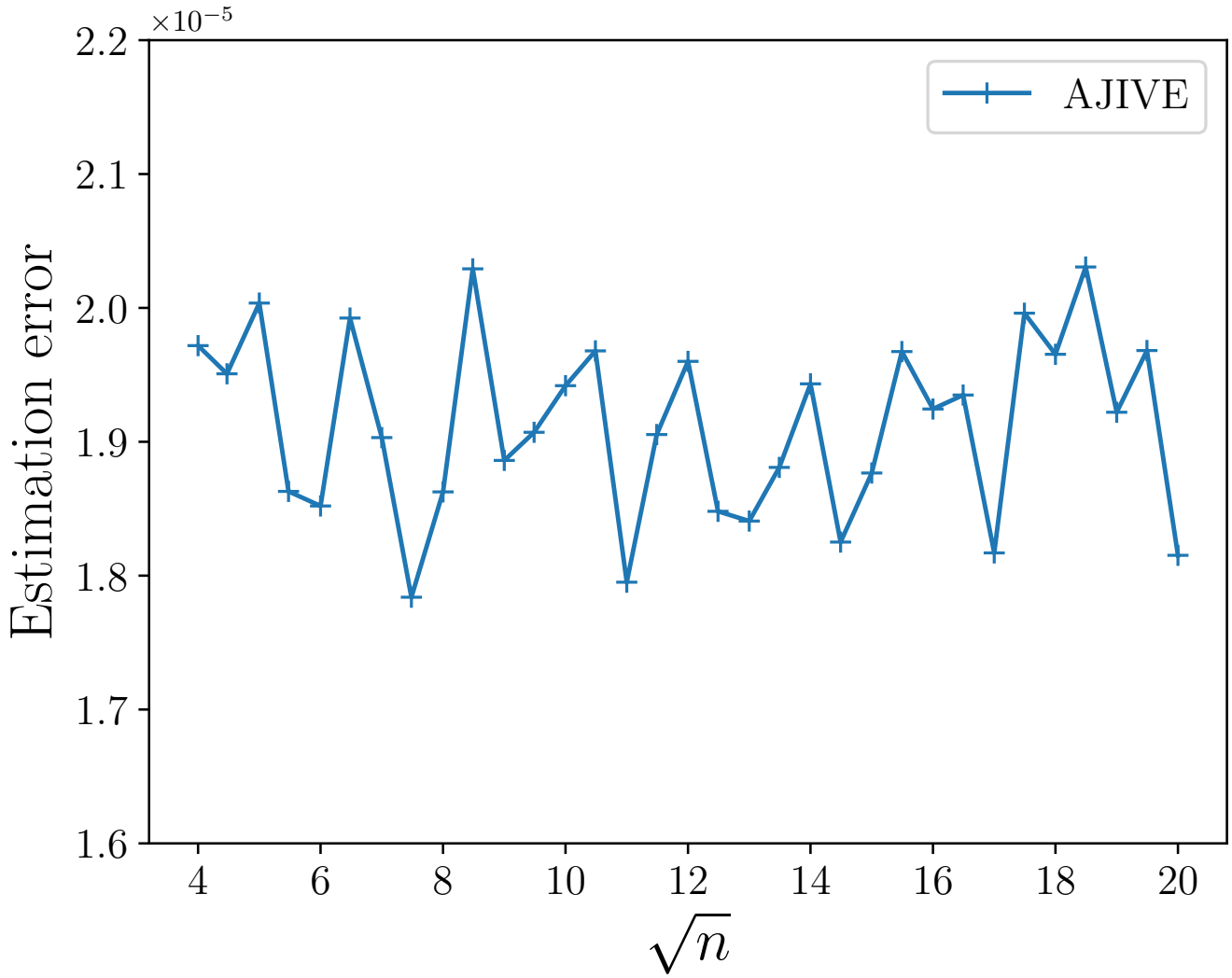


Small θ (0.0001 to 0.01): $\|\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\| \propto \sqrt{1/K\theta}$





Empirical results: small θ (0.0001)



Take-aways for the performance guarantee of AJIVE

- Using multiple data matrices is be powerful even in presence of the misaligned unique components ($\propto 1/\sqrt{K}$)
- AJIVE is first-order minimax optimal
- The extra error induced by unique components scales with intrinsic dimension r

Non-diminishing term in the Performance guarantee of AJIVE

Revisiting the upper bound of AJIVE:

Assuming $\sigma_{r+r_k}(\boldsymbol{A}_k) \geq 1$, $\sigma\sqrt{n} \leq \sqrt{\theta}$, and $K\theta \geq 1$

$$\|\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\| \lesssim \sigma\sqrt{\frac{n}{K} + \frac{r}{K\theta}} + \underbrace{\frac{\sigma^2 n}{\theta}}_{\substack{\text{non-diminishing} \\ \text{second-order error}}}$$

Non-diminishing term in the Performance guarantee of AJIVE

Revisiting the upper bound of AJIVE:

Assuming $\sigma_{r+r_k}(\boldsymbol{A}_k) \geq 1$, $\sigma\sqrt{n} \leq \sqrt{\theta}$, and $K\theta \geq 1$

$$\|\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\| \lesssim \sigma\sqrt{\frac{n}{K} + \frac{r}{K\theta}} + \underbrace{\frac{\sigma^2 n}{\theta}}_{\substack{\text{non-diminishing} \\ \text{second-order error}}}$$

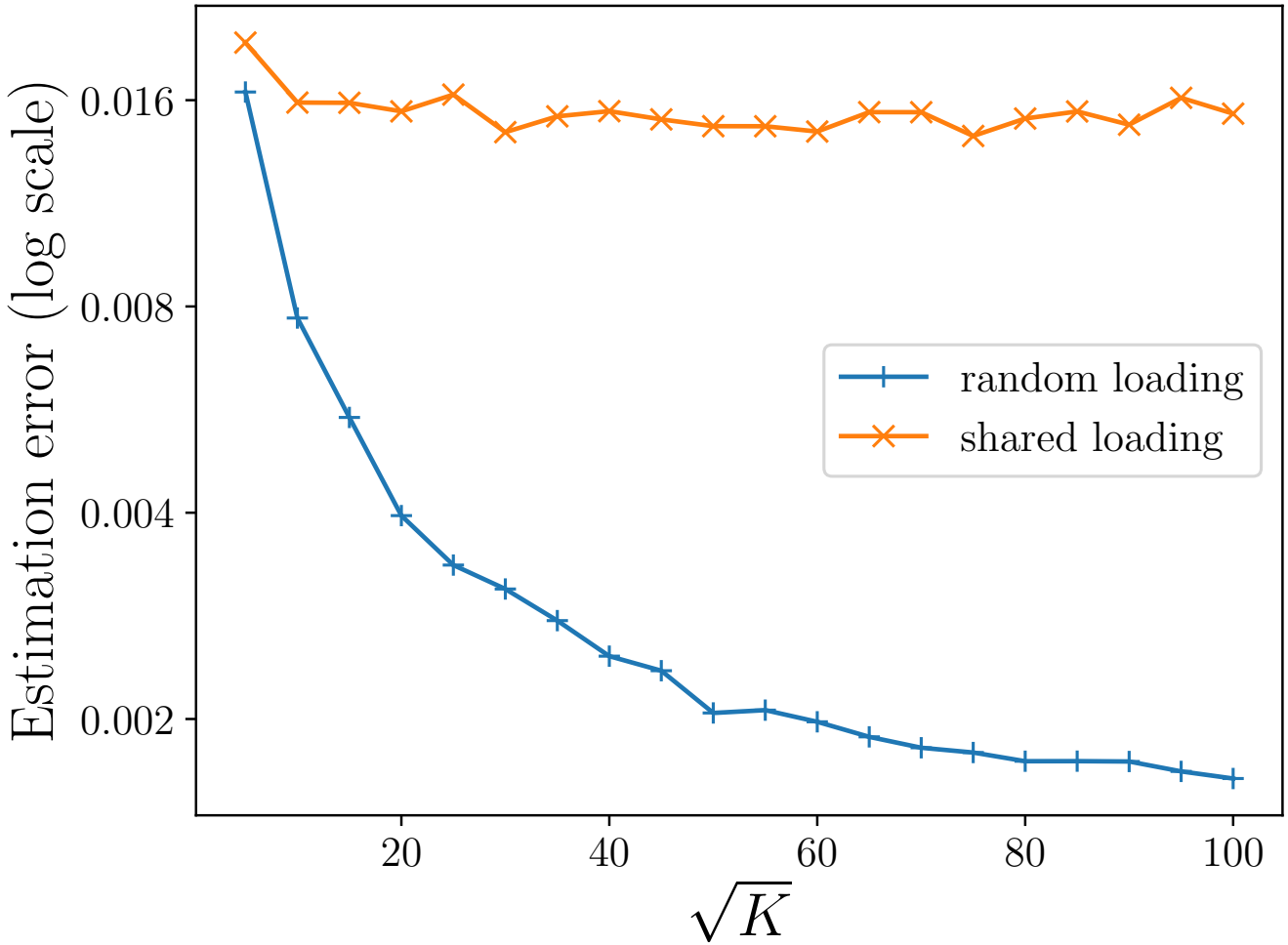
Is this term a technical artifact or a fundamental limitation?

Two experiment settings

$$\boldsymbol{A}_k = \underbrace{\boldsymbol{U}\boldsymbol{V}_k^\top}_{\substack{\text{rank}-r \\ \text{shared component}}} + \underbrace{\boldsymbol{U}_k\boldsymbol{W}_k^\top}_{\substack{\text{rank}-r_k \\ \text{unique component}}} + \underbrace{\boldsymbol{E}_k}_{\text{Gaussian noise}}$$

- Random loading: Let $\boldsymbol{V}_k$ and $\boldsymbol{W}_k$ be i.i.d. random orthonomral matrices
- Shared loading: Let $\boldsymbol{V}$ and $\boldsymbol{W}$ be random orthonomral matrices. Let $\boldsymbol{V}_k = \boldsymbol{V}$ and $\boldsymbol{W}_k = \boldsymbol{W}$ for all k

Empirical results: non-diminishing error is real



Shared vs random loading on $\|\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\|$ vs K.

AJIVE

1. Let $\widetilde{\boldsymbol{U}}_k$ be the top-$(r + r_k)$ left singular vectors of $\boldsymbol{A}_k$.
2. Let $\widehat{\boldsymbol{U}}$ be the top-r eigenvectors of $\sum_{k=1}^{K} \widetilde{\boldsymbol{U}}_k \widetilde{\boldsymbol{U}}_k^\top$.

- SVD is biased: $\mathbb{E}[\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^\top] \neq \boldsymbol{U}\boldsymbol{U}^\top + \boldsymbol{U}_k\boldsymbol{U}_k^\top$
- Such bias can be aligned (e.g. with shared loading), inducing a non-diminishing error in the second step of AJIVE

AJIVE

1. Let $\widetilde{\boldsymbol{U}}_k$ be the top-$(r + r_k)$ left singular vectors of $\boldsymbol{A}_k$.
2. Let $\widehat{\boldsymbol{U}}$ be the top-r eigenvectors of $\sum_{k=1}^{K} \widetilde{\boldsymbol{U}}_k \widetilde{\boldsymbol{U}}_k^\top$.

- SVD is biased: $\mathbb{E}[\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^\top] \neq \boldsymbol{U}\boldsymbol{U}^\top + \boldsymbol{U}_k\boldsymbol{U}_k^\top$
- Such bias can be aligned (e.g. with shared loading), inducing a non-diminishing error in the second step of AJIVE

The non-diminishing error does exist for AJIVE, but is it fundamental to the shared subspace estimation problem?

Non-convex loss (neg. log lik. under Gaussian noise)

$$\ell(\widetilde{\boldsymbol{U}}, \{\widetilde{\boldsymbol{U}}_k\}, \{\widetilde{\boldsymbol{V}}_k\}, \{\widetilde{\boldsymbol{W}}_k\}) = \sum_{k=1}^{K} \|\boldsymbol{A}_k - \boldsymbol{U}\boldsymbol{V}_k^\top - \boldsymbol{U}_k\boldsymbol{W}_k^\top\|_{\mathrm{F}}^2$$

- Let $\widehat{\boldsymbol{U}}$ be one-step alternating minimization of the non-convex loss starting from the ground truth $\boldsymbol{U}$
- Non-diminishing error still provably exists:

$$\sup_{\boldsymbol{U}, \{\boldsymbol{U}_k\}, \{\boldsymbol{V}_k\}, \{\boldsymbol{W}_k\}} \|\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \boldsymbol{U}\boldsymbol{U}^\top\| \gtrsim \sigma^4 n^2$$

In the Neyman and Scott's problem, the law of a sequence of ind. observations $(\boldsymbol{A}_k)$ is governed by two sets of parameters:

- structural parameters, which appears in the law of every observation $(\boldsymbol{U})$
- incidental parameters, which only appears in each individual observation $(\{\boldsymbol{U}_k\}, \{\boldsymbol{V}_k\}, \{\boldsymbol{W}_k\})$

MLE can be inconsistent for the structural parameters

Summary

- The JIVE framework and AJIVE algorithm
- Power: AJIVE is first-order optimal
- Limitation: Non-diminishing error appears to be fundamental

Summary

- The JIVE framework and AJIVE algorithm
- Power: AJIVE is first-order optimal
- Limitation: Non-diminishing error appears to be fundamental

Future directions:

- Fully settling the non-diminishing error
- Going beyond the JIVE framework

Reference

Feng, Q., Jiang, M., Hannig, J., and Marron, J. (2018). Angle-based joint and individual variation explained. *Journal of multivariate analysis*, 166:241–265.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523.