Can matrix coherence be efficiently and accurately estimated?

Ameet Talwalkar
UC Berkeley

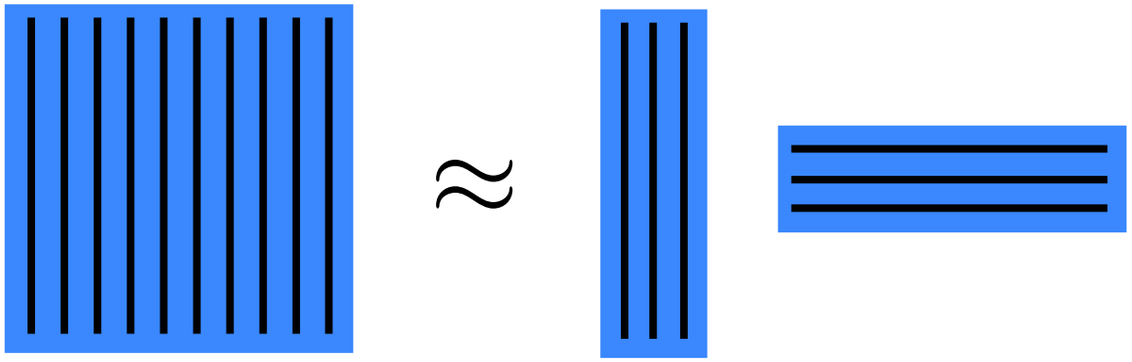
joint with Mehryar Mohri (NYU, Google)

Motivation

- Many modern datasets are large, high-dimensional and can be represented by large matrices
  - e.g., videos, images, documents on the web

- Low-rank approximations are often appropriate
  - e.g., dimensionality reduction, collaborative filtering

$\mathbf{X} =$



$n \times m \qquad n \times r \qquad r \times m$

$r \ll n, m$

Motivation

- Many modern datasets are large, high-dimensional and can be represented by large matrices
  - e.g., videos, images, documents on the web

- Low-rank approximations are often appropriate
  - e.g., dimensionality reduction, collaborative filtering

- Sampling-based methods work with subset of columns
  - tractable when SVD is not
  - interpretability

Key Assumptions

1. *Good low-rank structure, i.e.,* $\mathbf{X} \approx \mathbf{X}_r$

2. *Finding a good subset of columns is possible*

$$\mathbf{X} = \begin{bmatrix} | & & | & | & & | \\ \mathbf{e}_1 & \dots & \mathbf{e}_r & \mathbf{0} & \dots & \mathbf{0} \\ | & & | & | & & | \end{bmatrix}$$

- non-uniform sampling, e.g., [Drineas et al., '05]
- uniform sampling often works well in practice [Kumar et al., '09]
  - *incoherence* ⟶ uniform sampling [Talwalkar & Rostamizadeh, '10]
  - but, computing matrix coherence is expensive
  - this work: can we estimate matrix coherence?

Outline

- ESTIMATE-COHERENCE algorithm
- Analysis in low-rank setting
- Experiments

Matrix Coherence

Definition: Matrix Coherence, $\mu_0(\cdot)$

- $\mathbf{U}_r$: top left singular vectors of $\mathbf{X} \in \mathbb{R}^{n \times m}$
- $\mathbf{P}_r = \mathbf{U}_r \mathbf{U}_r^\top$: orthogonal projection matrix
- $\mu_0(\mathbf{U}_r) = \dfrac{n}{r} \max_i \|\mathbf{P}_r \mathbf{e}_i\|^2$

- Degree to which $\mathbf{U}_r$ corresponds to canonical basis
  - min: 1 (e.g., all matrix entries identical)
  - max: n/r (e.g., canonical basis)

$$\mathbf{X} = \begin{bmatrix} | & & | & | & & | \\ \mathbf{e}_1 & \dots & \mathbf{e}_r & \mathbf{0} & \dots & \mathbf{0} \\ | & & | & | & & | \end{bmatrix}$$

- incoherence $\longrightarrow$ uniform sampling

Matrix Coherence

Definition: Matrix Coherence, $\mu_0(\cdot)$

- $\mathbf{U}_r$: top left singular vectors of $\mathbf{X} \in \mathbb{R}^{n \times m}$
- $\mathbf{P}_r = \mathbf{U}_r \mathbf{U}_r^\top$: orthogonal projection matrix
- $\mu_0(\mathbf{U}_r) = \frac{n}{r} \max_i \|\mathbf{P}_r \mathbf{e}_i\|^2$

- Coherence assumptions used in other lines of work
  - e.g., matrix completion, robust PCA

[Candes and Recht, '09; Candes et al., '09; Keshavan et al. '09]

- We study related quantity:

$$0 < \gamma(\mathbf{U}_r) = \frac{r}{n} \mu_0(\mathbf{U}_r) \leq 1$$

Proposed Algorithm

- $\mathbf{X}_1$: $n \times l$ matrix containing $l \ll m$ columns of $\mathbf{X}$ sampled uniformly at random
- Idea: $\gamma(\mathbf{X}) \approx \gamma(\mathbf{X}_1)$
- If $\mathbf{X}$ is low rank:

$\textsc{Estimate-Coherence}(\mathbf{X}_1)$

1 $\mathbf{U}_{X_1} \leftarrow \textsc{CompactSVD}(\mathbf{X}_1)$
2 $\gamma(\mathbf{X}_1) \leftarrow \textsc{Calculate-Gamma}(\mathbf{U}_{X_1})$
3 **return** $\gamma(\mathbf{X}_1)$

Proposed Algorithm

- $\mathbf{X}_1$: $n \times l$ matrix containing $l \ll m$ columns of $\mathbf{X}$ sampled uniformly at random
- Idea: $\gamma(\mathbf{X}) \approx \gamma(\mathbf{X}_1)$
- For rank r approximation of arbitrary $\mathbf{X}$:

$\textsc{Estimate-Coherence}(\mathbf{X}_1, r)$

1 $\mathbf{U}_{X_1} \leftarrow \textsc{CompactSVD}(\mathbf{X}_1, r)$
2 $\gamma(\mathbf{X}_1) \leftarrow \textsc{Calculate-Gamma}(\mathbf{U}_{X_1})$
3 **return** $\gamma(\mathbf{X}_1)$

Outline

- ESTIMATE-COHERENCE algorithm
- Analysis in low-rank setting
- Experiments

Low-rank Analysis

Observation 1:

- $\gamma(\mathbf{X}_1) \leq \gamma(\mathbf{X})$ *and is monotonically increasing as a function of* l
- $\gamma(\mathbf{X}_1) = \gamma(\mathbf{X})$ *when* $l \geq \tilde{O}(r^2 \mu_0(\mathbf{U}_r) 1/\delta)$ *with probability* $1 - \delta$

- Quality of estimation depends on coherence itself

Low-rank Analysis

Observation 1:

- $\gamma(\mathbf{X}_1) \leq \gamma(\mathbf{X})$ *and is monotonically increasing as a function of* l
- $\gamma(\mathbf{X}_1) = \gamma(\mathbf{X})$ *when* $l \geq \tilde{O}(r^2 \mu_0(\mathbf{U}_r) 1/\delta)$ *with probability* $1 - \delta$

Proof Sketch:

- Relate projection matrices associated with column spaces of $\mathbf{X}$ and $\mathbf{X}_1$
- Utilize coherence analysis of sampling-based approximations

[Talwalkar & Rostamizadeh, '10]

Low-rank Analysis

Observation 2: Fix r, n, m and let $\hat{\gamma} \ll 1$ be a constant. For any γ such that $\hat{\gamma} < \gamma \leq 1$, there exists an $\mathbf{X}$ with $\gamma(\mathbf{X}) = \gamma$ such that

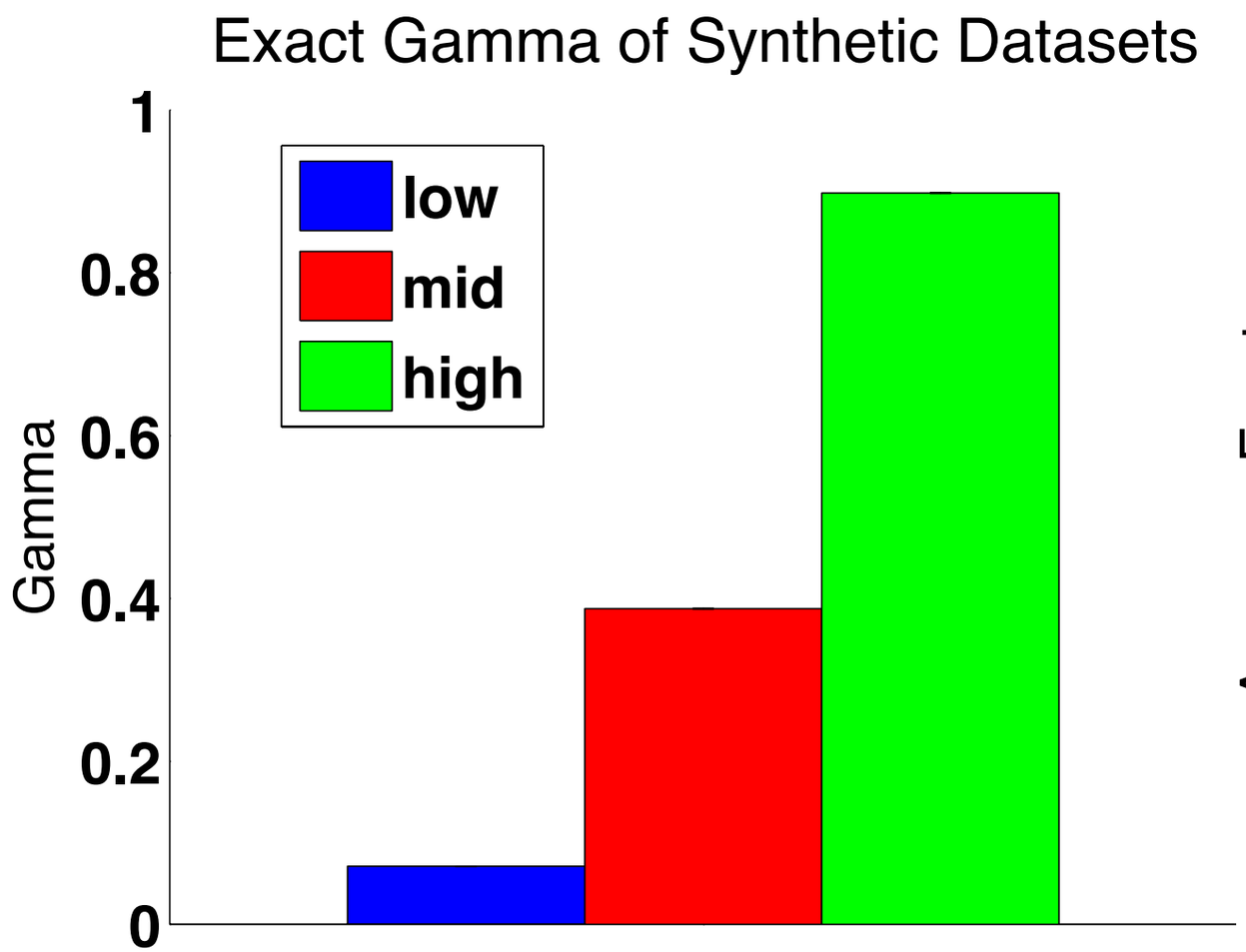
- $\gamma(\mathbf{X}_1) \leq \hat{\gamma}$, *if* $\mathbf{X}_1$ *does not include* $\mathbf{X}^{(1)}$
- $\gamma(\mathbf{X}_1) = \gamma$, *otherwise*

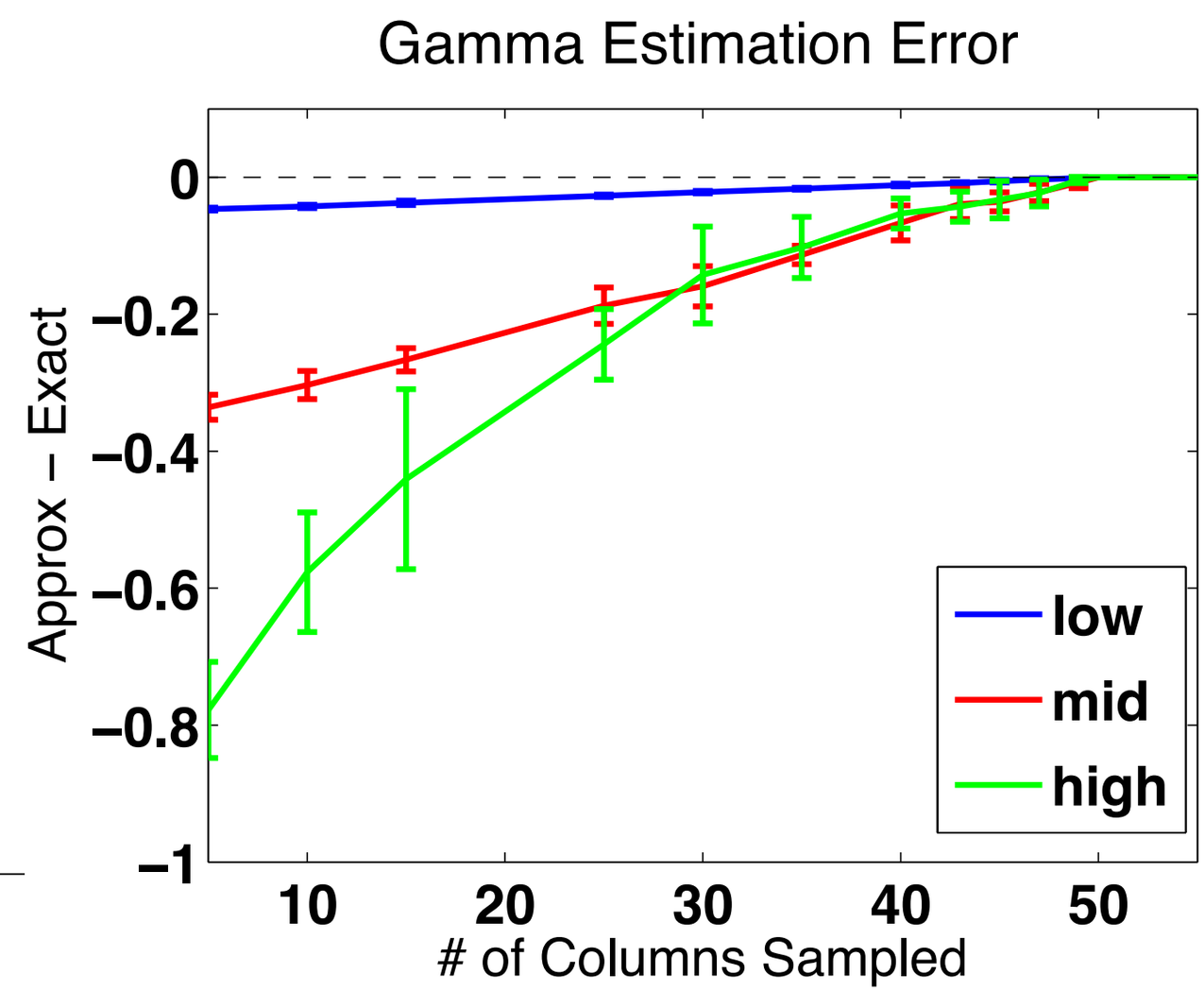
- Gap ($\gamma - \hat{\gamma}$) is proportional to γ
- Simple construction that makes use of coherence properties of 'random orthogonal model' [Candes and Recht, '09]

Outline

- ESTIMATE-COHERENCE algorithm
- Analysis in low-rank setting
- Experiments

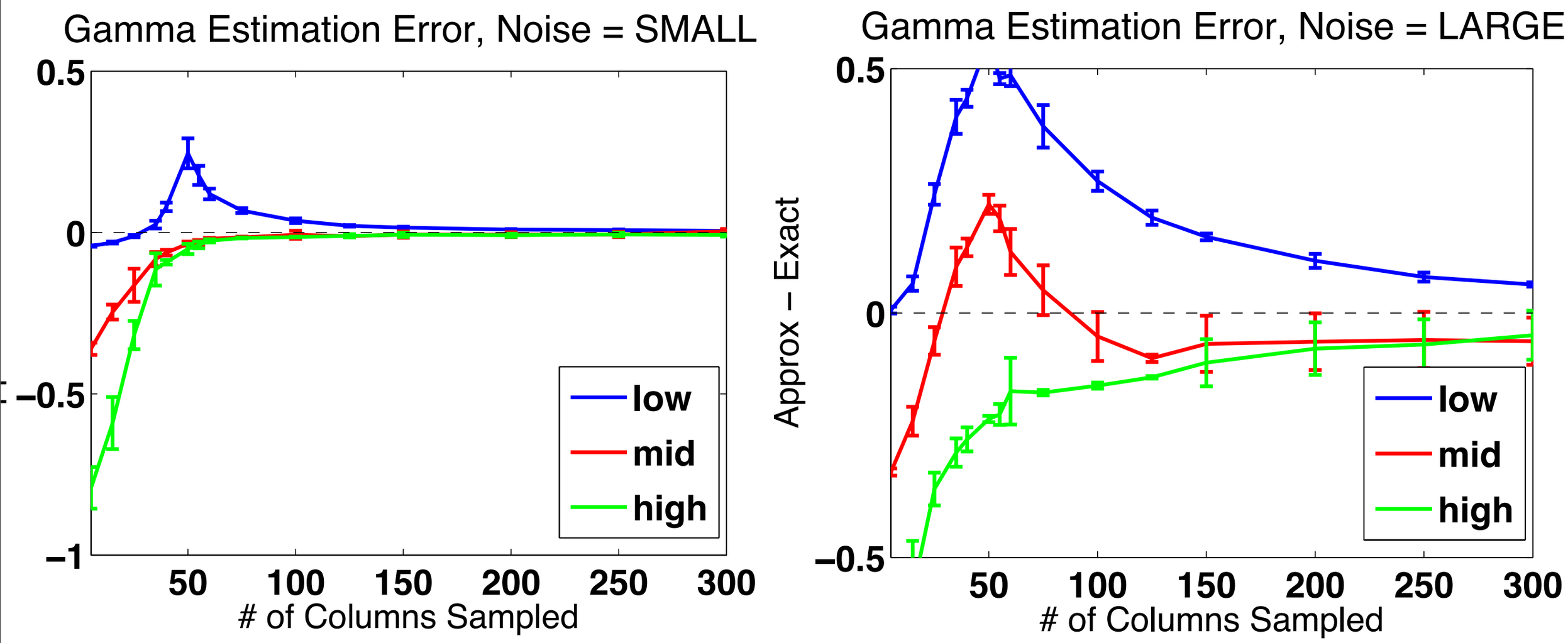
Low-rank Synthetic Data





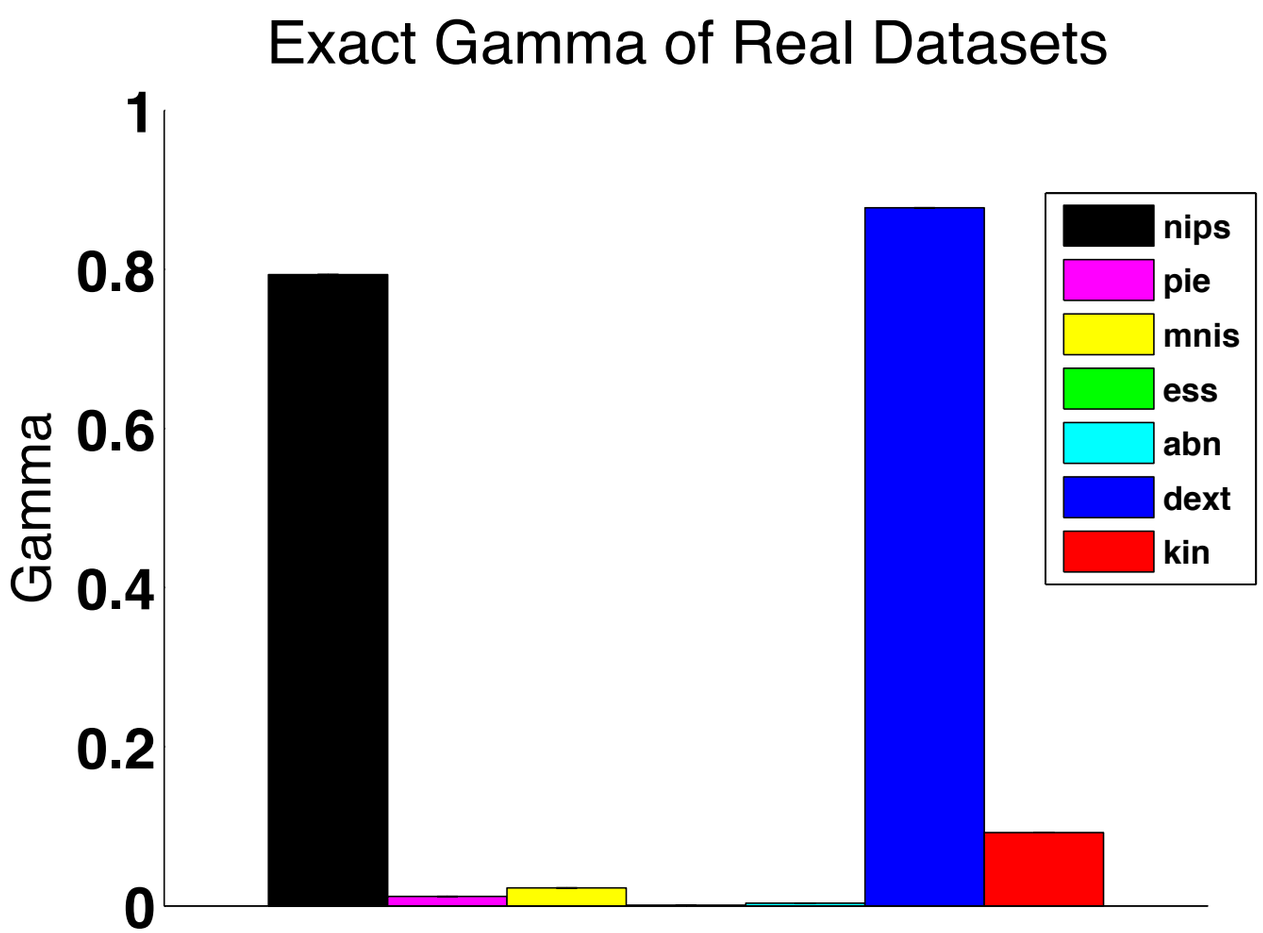
- $n = m = 1000; r = 50$
- True coherence recovered when $l \geq r$

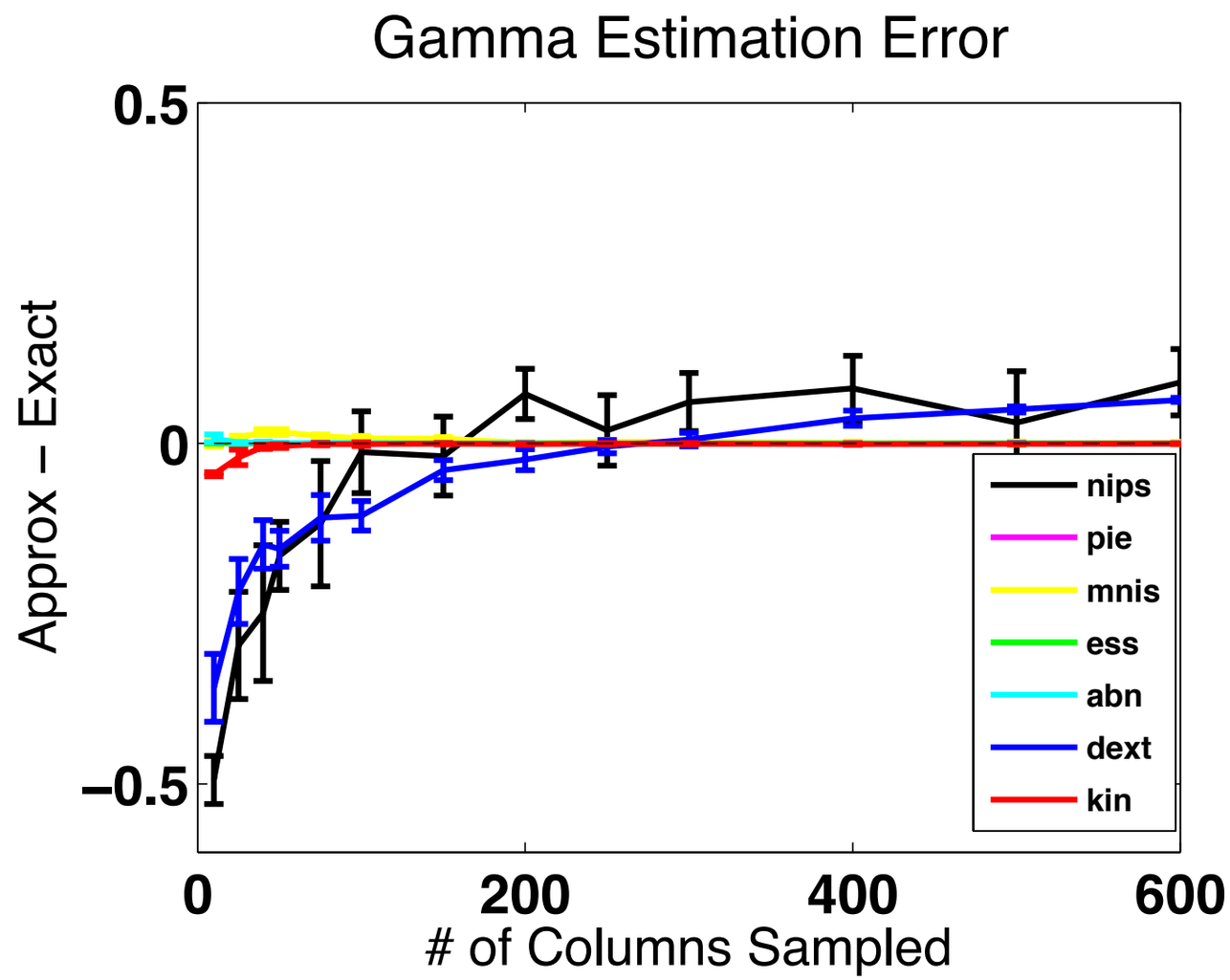
Full-rank Synthetic Data



- Set remaining $n - r$ singular values equal to $\epsilon \cdot \sigma_r$
- Good recovery when l is a small multiple of r

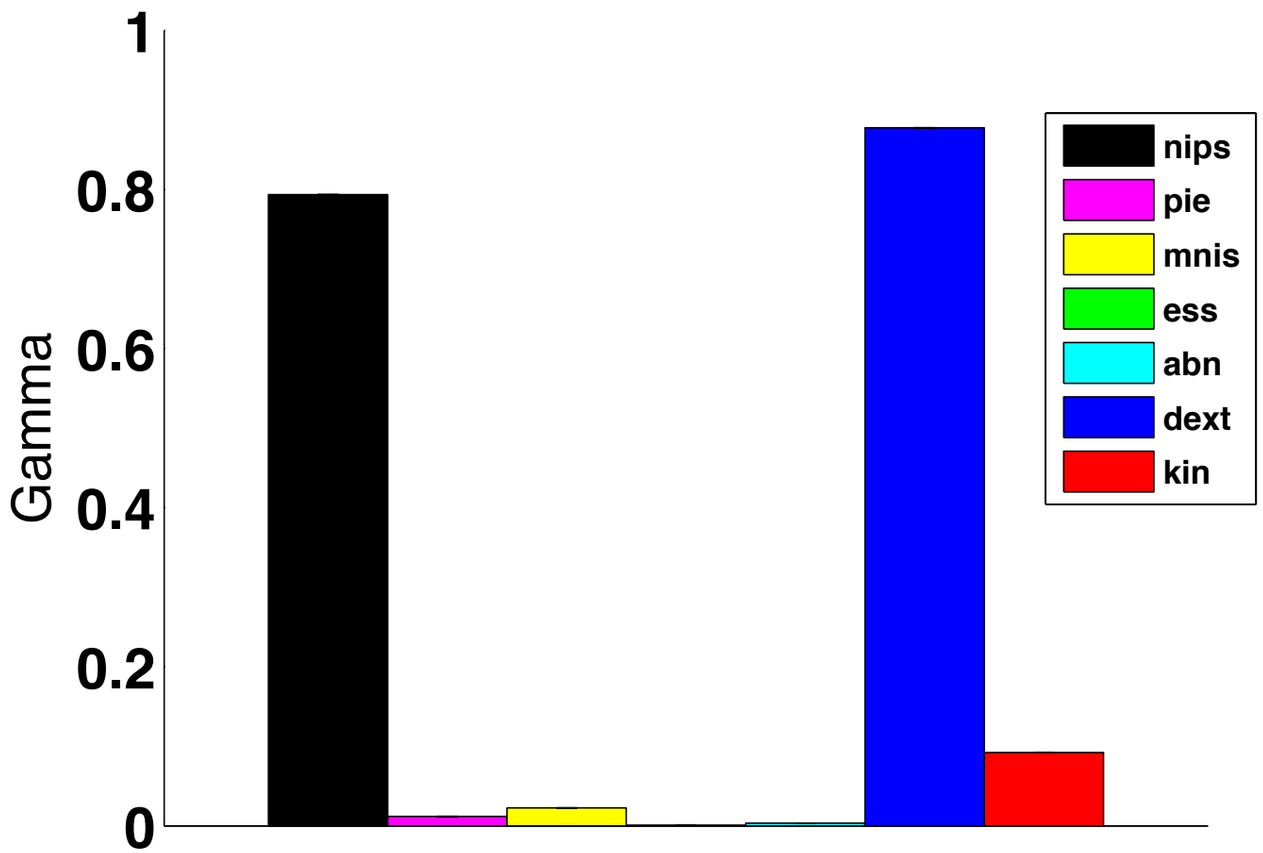
Real Data

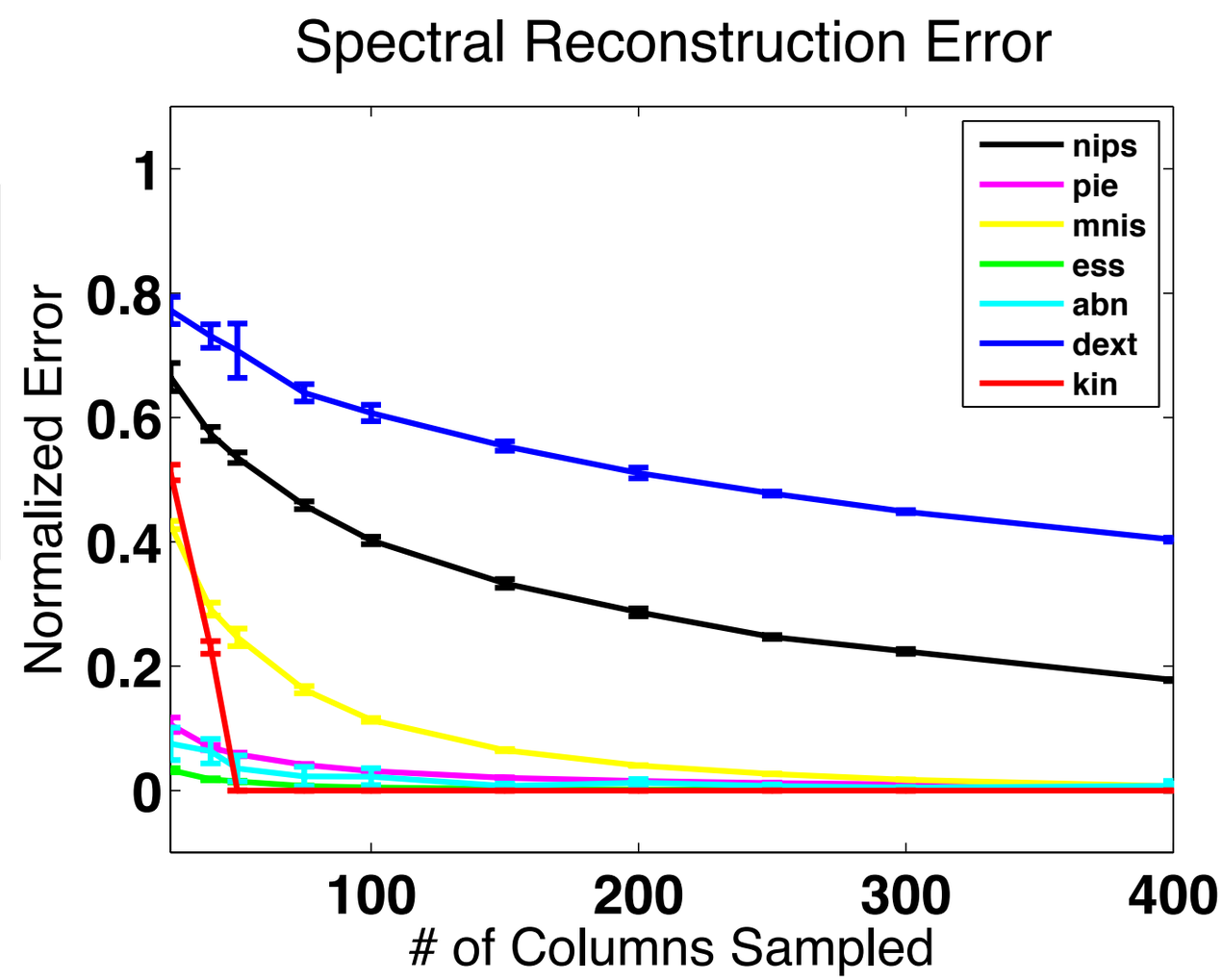




- Good estimates for after sampling ~100 columns
- High coherence ⟶ slower convergence, higher variance

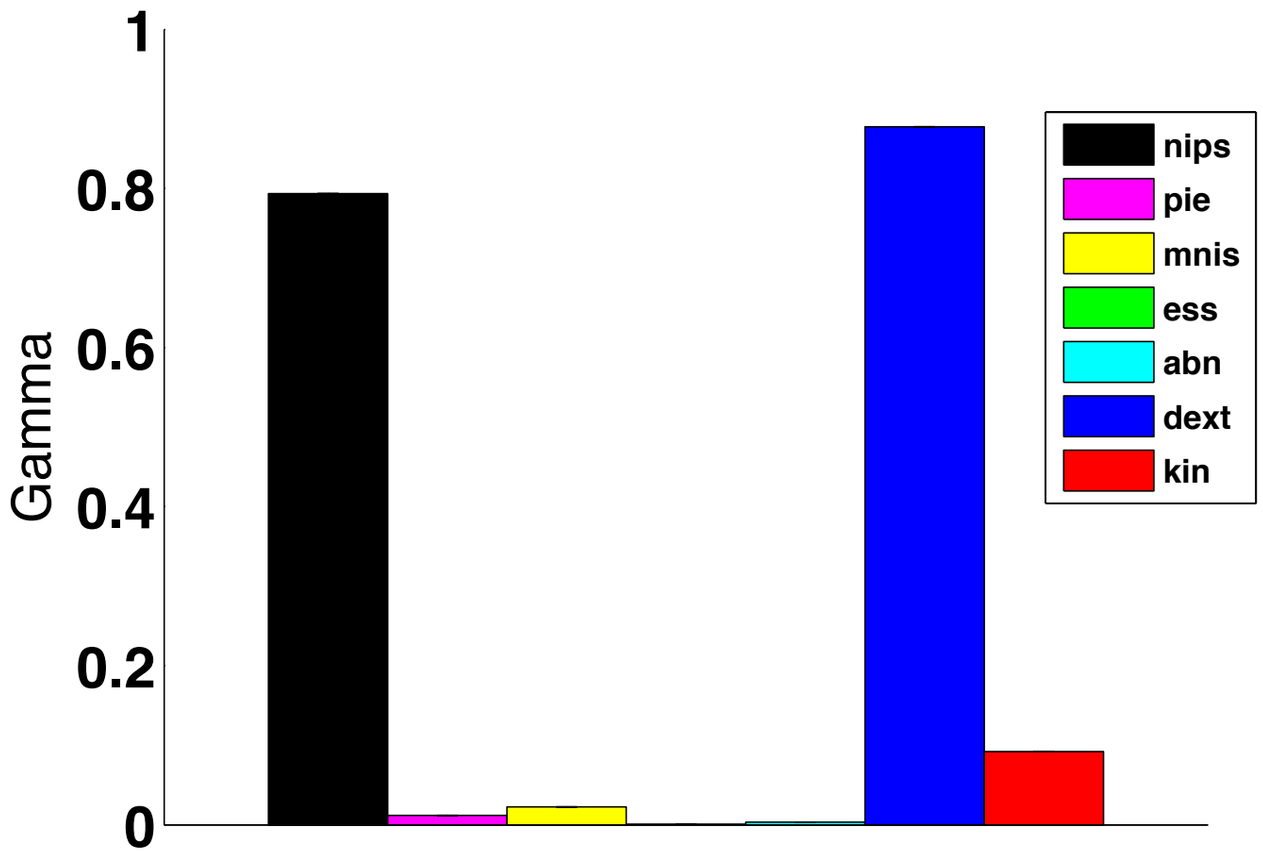
Coherence + Low-Rank

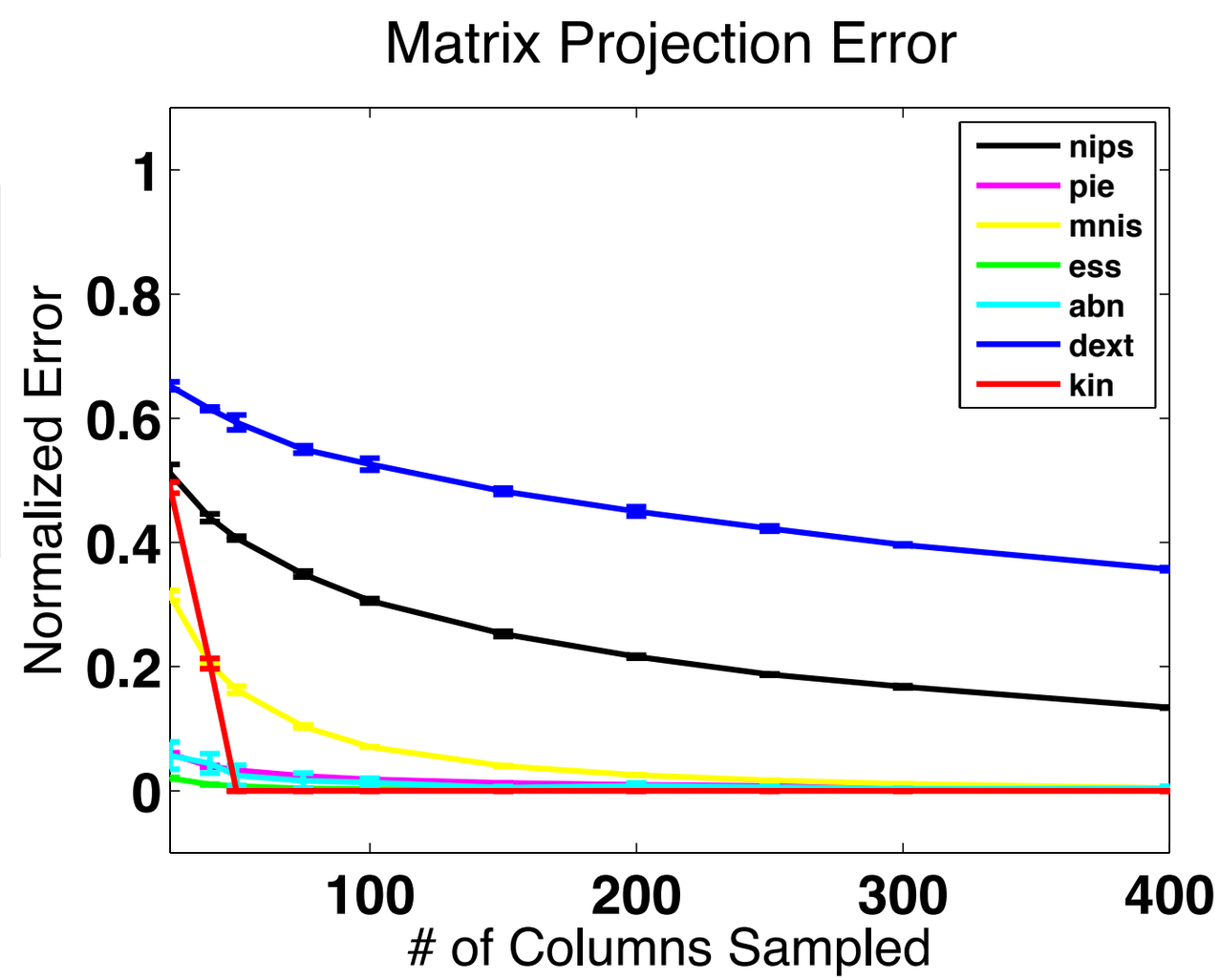




- Normalized Error: $\|\mathbf{X} - \widetilde{\mathbf{X}}\|_F / \|\mathbf{X}\|_F$
- Gamma is an excellent predictor of quality of sampling based low-rank approximation

Coherence + Low-Rank





- Normalized Error: $\|\mathbf{X} - \widetilde{\mathbf{X}}\|_F / \|\mathbf{X}\|_F$
- Gamma is an excellent predictor of quality of sampling based low-rank approximation

Conclusion

- Novel algorithm to estimate matrix coherence
  - theory: estimate depends on coherence itself
  - practice: accurate estimates across coherence
  - predicts effectiveness of low-rank approximation

- Future work:
  - Analyze estimates in full-rank setting
  - Combine several estimates for better approximation