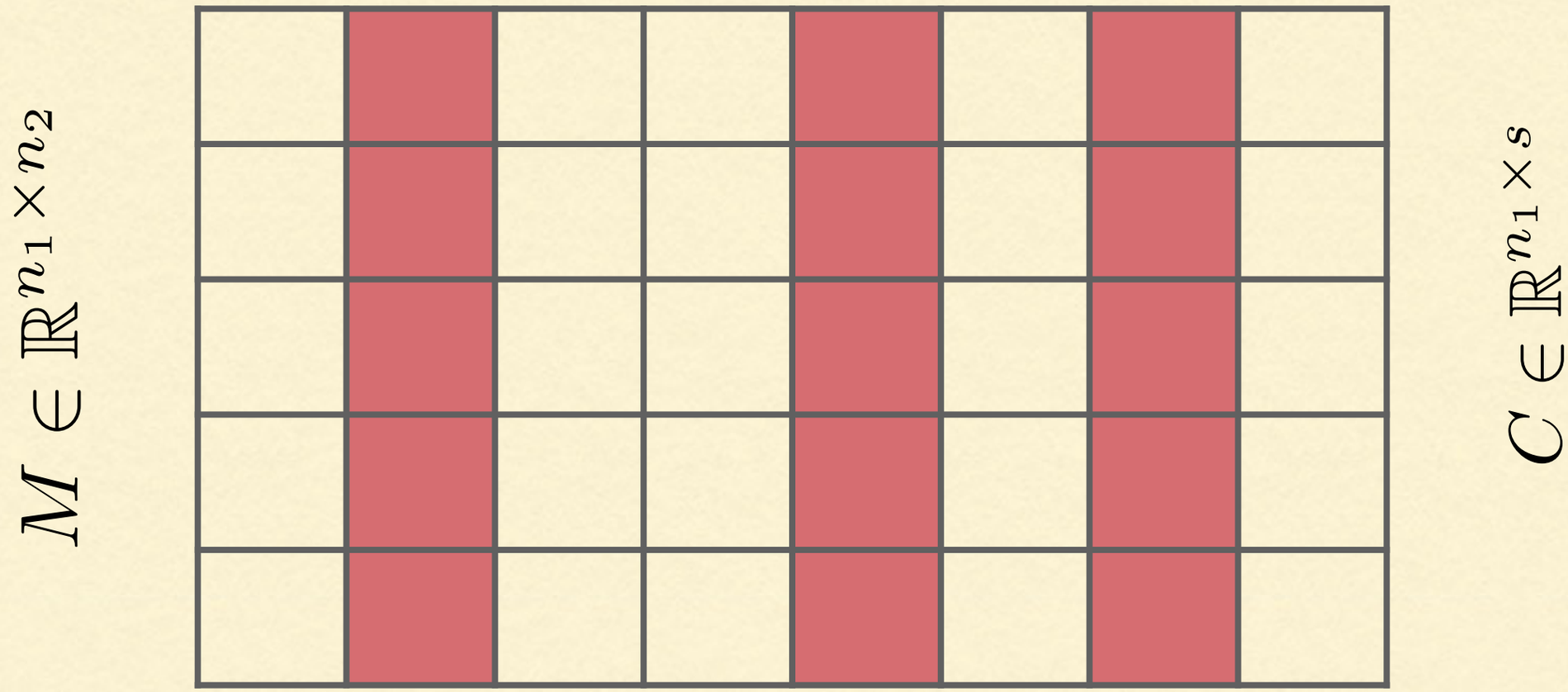
EMPIRICAL COMPARISON OF COLUMN SUBSET SELECTION ALGORITHMS

Yining Wang, Aarti Singh

Machine Learning Department, Carnegie Mellon University

COLUMN SUBSET SELECTION



$$\min_{|C| \leq s} \|M - CC^{\dagger} M\|_F$$

COLUMN SUBSET SELECTION

- Interpretable low-rank approximation (compared to PCA)
- Applications:
  - Unsupervised feature selection
  - Image compression
  - Genetic analysis: target SNP selection, etc.
- Challenges:
  - Exact column subset selection is NP-hard

ALGORITHMS

- Deterministic Algorithms
  - *Rank-revealing QR (RRQR)* [Chan, 87]
    - Most accurate, but expensive: $O(n^3)$
- Sampling based algorithms,slightly inaccurate, but cheap: $O(n^2k)$
  - Norm sampling [Frieze et. al., 04]
  - Leverage score sampling [Drineas et. al., 08]
  - Iterative norm sampling (approximate volume sampling) [Deshpand & Vempala, 06]

NORM SAMPLING

- The algorithm:

  1. Compute column norms $\|M_{(i)}\|_2$
  2. Sample each column with probability $p_i \propto \|M_{(i)}\|_2^2$

- Time complexity: $O(n^2)$

- Error analysis:

"Additive error"

$$\|M - CC^\dagger M\|_F^2 \leq \|M - M_k\|_F^2 + O(k/s) \cdot \|M\|_F^2$$

LEVERAGE SCORE SAMPLING

- The algorithm:

  1. Top-k truncated SVD: $M = U_k \Sigma_k V_k^\top + U_{-k} \Sigma_{-k} V_{-k}^\top$
  2. Leverage score sampling: $p_i \propto \|U_k e_i\|_2^2$

- Time complexity: $O(n^2 k)$

- Error analysis: assuming $s = \Omega(k^2/\epsilon^2)$ "Relative error"

$$\|M - CC^\dagger M\|_F \leq (1+\epsilon)\|M - M_k\|_F$$

ITERATIVE NORM SAMPLING

- Initialize C=0. Repeat until s columns are selected:

  1. Compute residue: $r_i = M_{(i)} - CC^\dagger M_{(i)}$
  2. Residue norm sampling: $p_i \propto \|r_i\|_2^2$

- Time complexity: $O(n^2 s)$

- Error analysis:

"Multiplicative error"

$$\mathbb{E}_c \left[\|M - CC^\dagger M\|_F^2 \right] \leq (k+1)! \|M - M_k\|_F^2$$

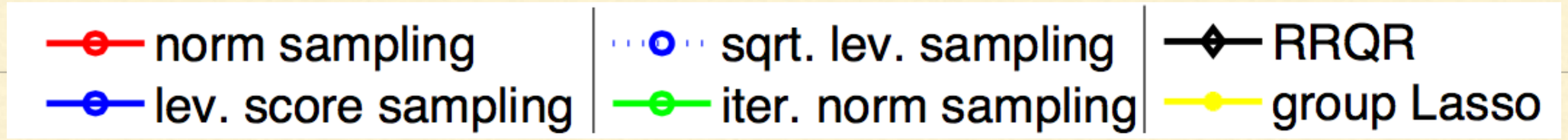
QUESTION

- Three different algorithms
  - Norm sampling: $\|M - CC^{\dagger}M\|_F^2 \leq \|M - M_k\|_F^2 + \epsilon\|M\|_F^2$
    - Leverage score sampling: $\|M - CC^{\dagger}M\|_F^2 \leq (1+\epsilon)\|M - M_k\|_F^2$
  - Iterative norm sampling: $\|M - CC^{\dagger}M\|_F^2 \leq (k+1)!\|M - M_k\|_F^2$
- Which one works best in practice?

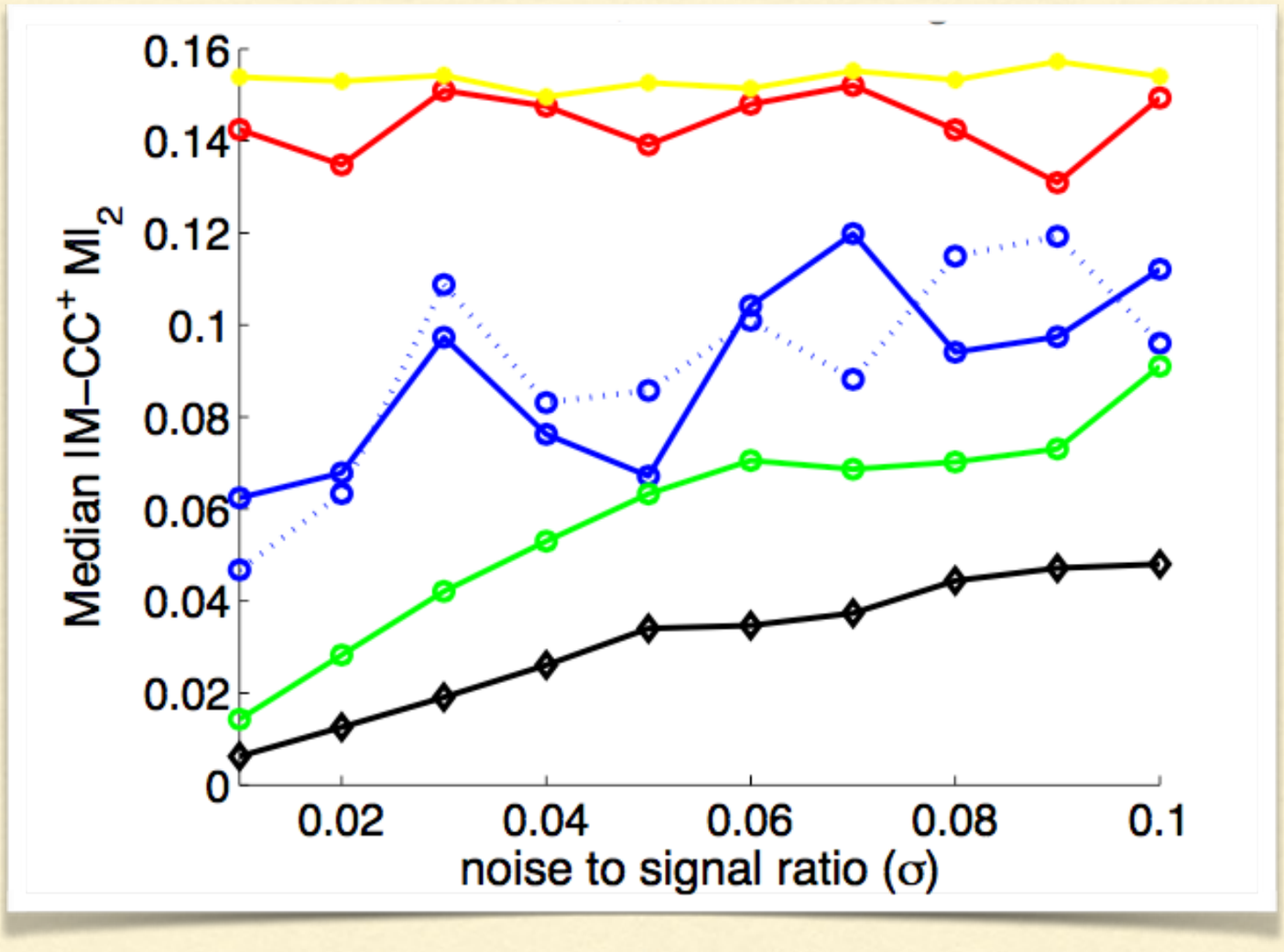
EXPERIMENTS

- Synthetic data:
  - Generate an n x k random Gaussian matrix A
  - Set $M = AA^T$, then normalize so that M has unit F norm
  - *Coherent design:* pick a random column in M, enlarge its norm by 10 times and repeat the same column five times.
  - Noise corruption: impose entrywise zero-mean noise on the normalized matrix M.

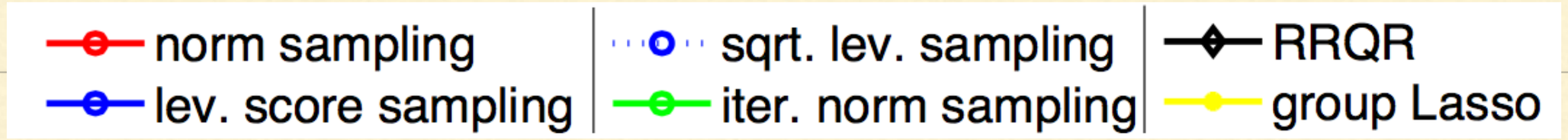
EXPERIMENTS



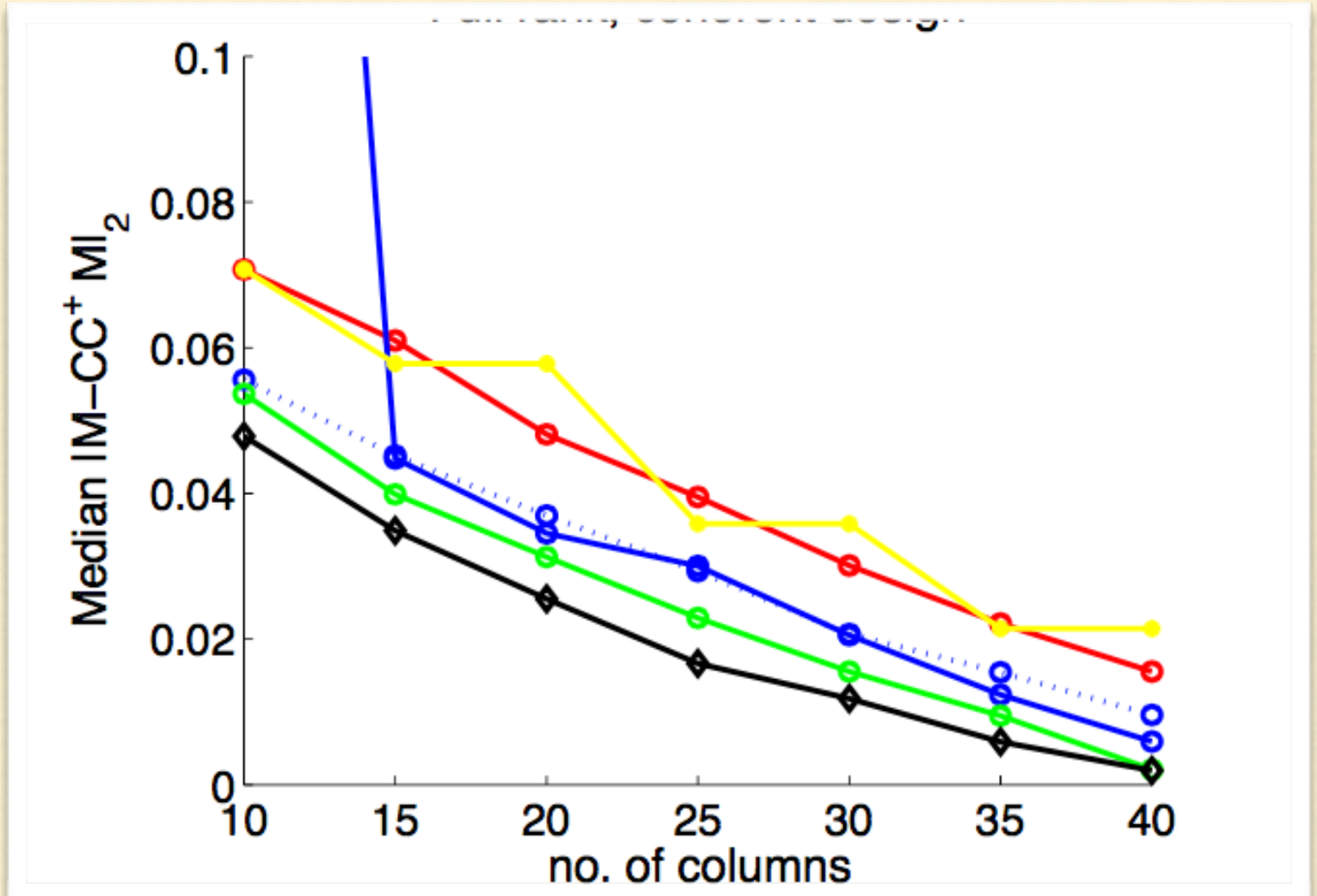
- Low-rank input, coherent design



EXPERIMENTS



- Full-rank input, coherent design



EXPERIMENTS

- Computational efficiency

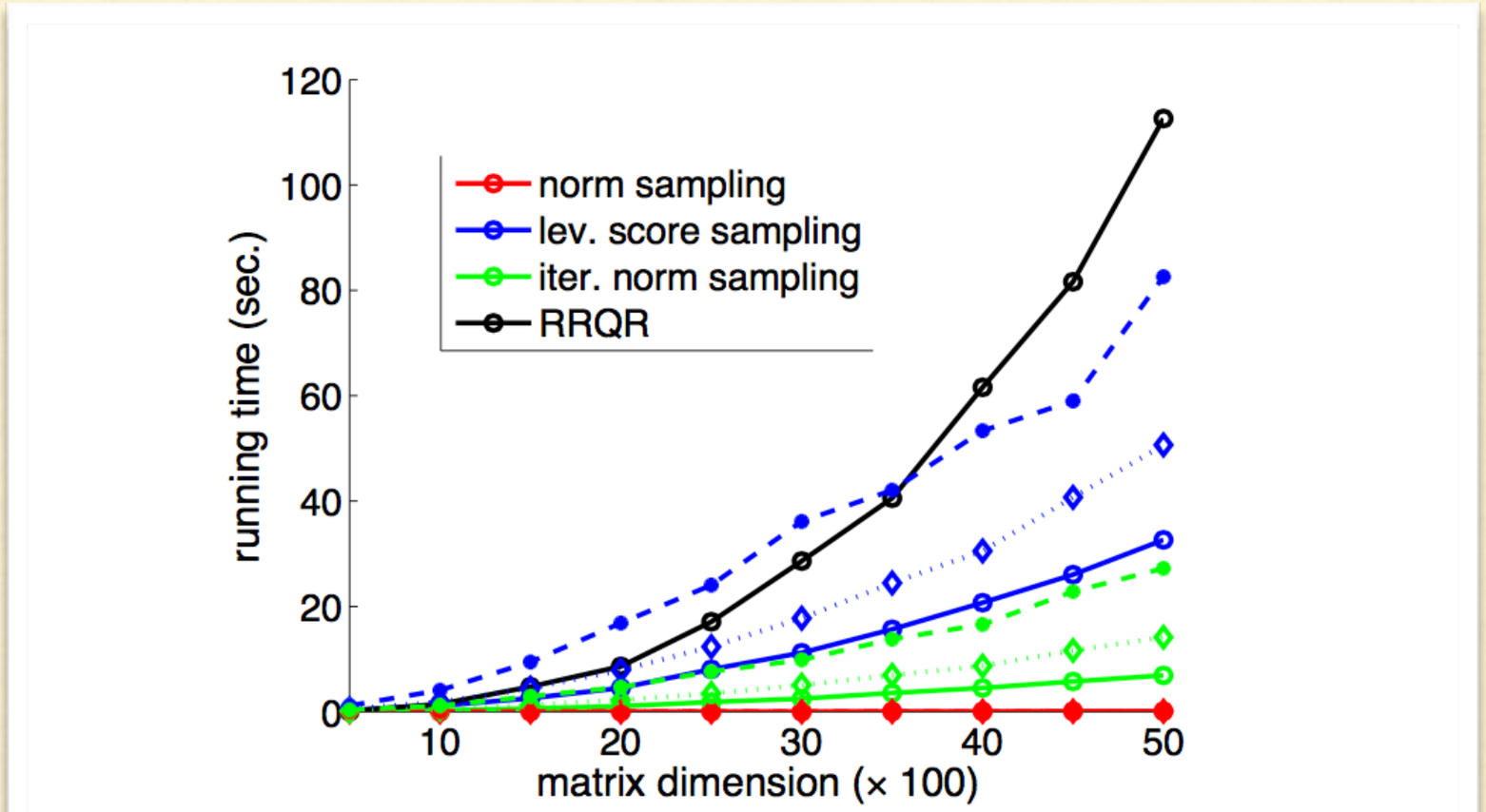
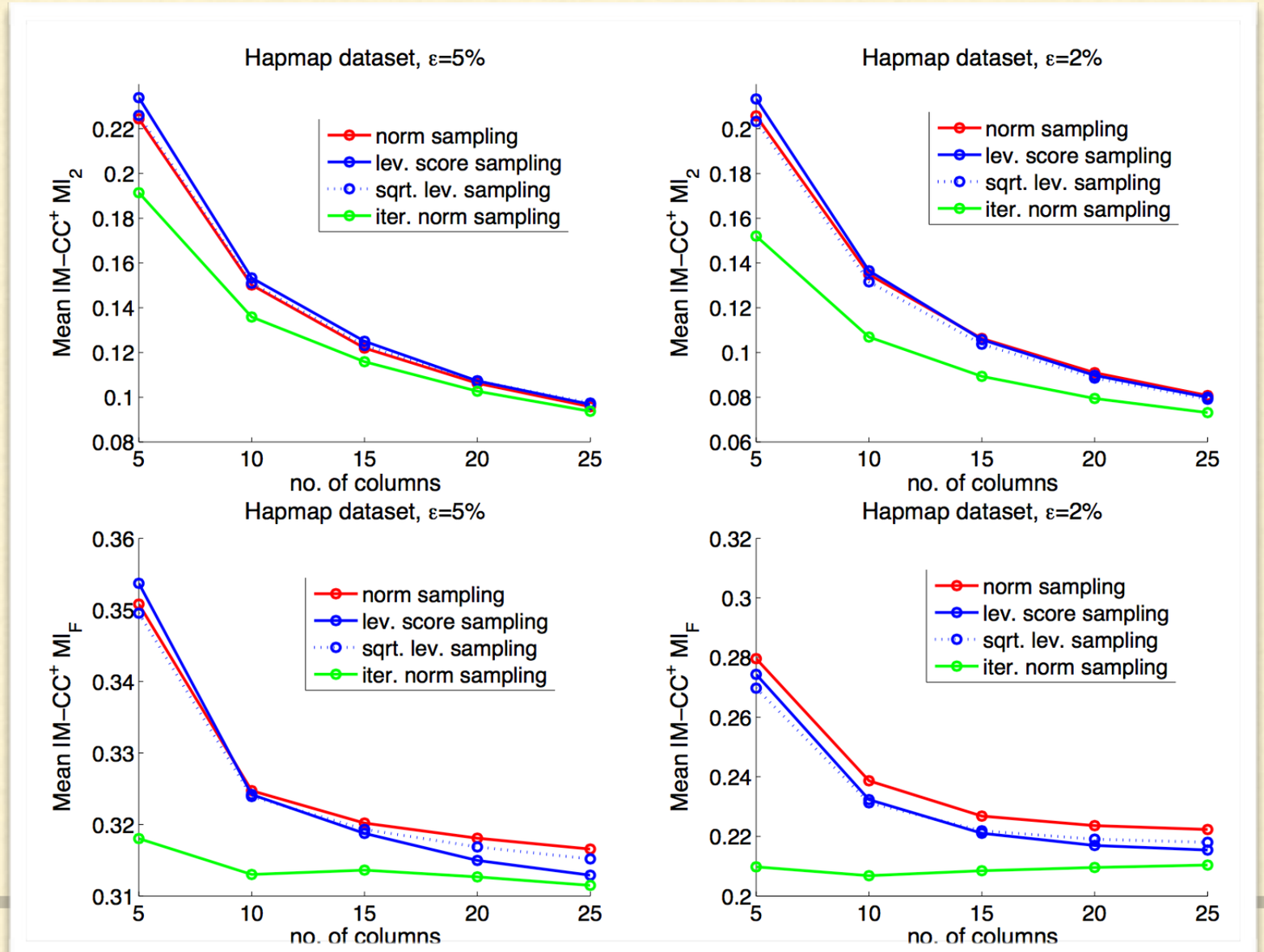


Fig. 2. Running time (seconds) of norm sampling, leverage score sampling, iterative norm sampling and RRQR factorization with respect to different matrix sizes. Solid line: $k = 25$; dotted line: $k = 50$; dashed line: $k = 100$.

EXPERIMENTS

- Human genetic data: *Hapmap Phase II*



CONCLUSION

- **Iterative norm sampling** performs much better than **leverage score sampling** in practice, which is not predicted by existing theoretical results.

- **Iterative norm sampling** is also computationally cheaper then **leverage score sampling**, which requires truncated SVD.

- Calls for improved analysis of iterative norm sampling!

REFERENCES

- T.F. Chan, "Rank Revealing QR Factorizations," *Linear Algebra and Its Applications*, vol. 88, pp. 67-82, 1987.
- A. Frieze, R. Kannan and S. Vempala, "Fast Monte-Carlo Algorithms for Finding Low-rank Approximations," *Journal of the ACM*, vol. 51, no. 6, pp. 1025-1041, 2004.
- P. Drineas, M.W. Mahoney and S. Muthukrishnan, "Relative-error CUR Matrix Decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 844-881, 2008.
- A. Deshpande and S. Vempala, "Adaptive Sampling and Fast Low-rank Matrix Approximation," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, 2006, pp. 292-303.