Spectrum-Revealing Matrix Factorizations Theory and Algorithms

Ming Gu

Department of Mathematics
University of California, Berkeley

April 5, 2016

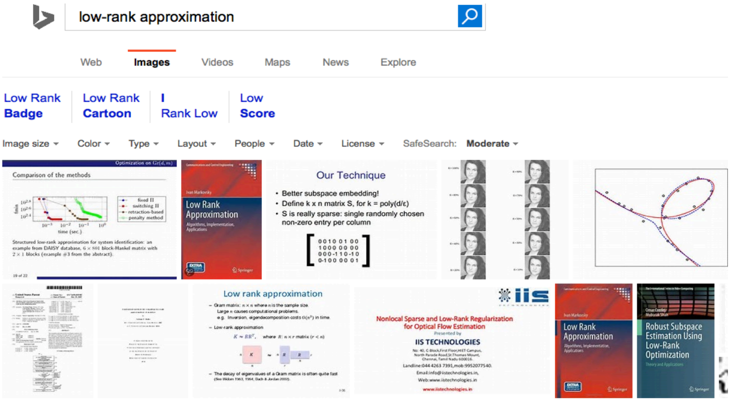
Joint work with D. Anderson, J. Deursch, C. Melgaard, J. Xiao

Content

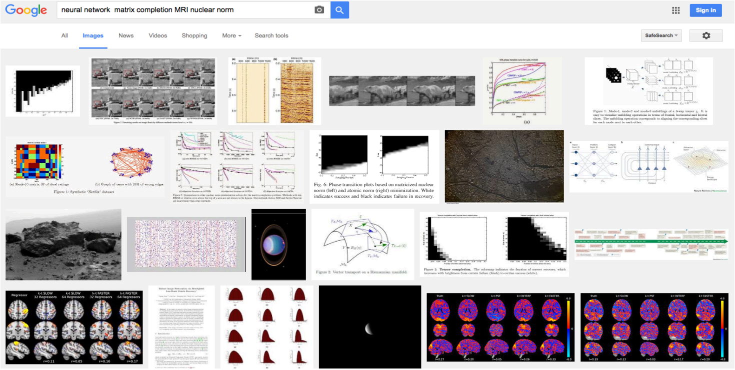
- **Low-rank matrix approximations**
- Rank-revealing matrix factorizations vs. Spectrum-revealing matrix factorizations.
- Related new algorithms
- Numerical Experiments
- Future work

Background–Applications: Bing Search

- Scientific Computing/Data Analysis
- Computer Vision
- Machine learning; kernel learning.



Google Search on Low-rank Approximation



Goals in Low-rank Matrix Approximations

- Highly efficient
- Minimum communication
- As accurate/reliable as truncated SVD
- Beyond SVD: preserve sparsity, structure, interpretability

Background: Low-Rank Approximation Methods

- Deterministic low-rank approximations
  - Interpolative decomposition (ID) (Cheng et. al. 2005, Liberty et. al 2007)
  - Deterministic CUR with various column selection algorithms
- Randomized low-rank approximations
  - Using the subsampled fft technique, for $\ell \gg k$: "optimal" complexity is
    $$O\left(mn\log(k) + \ell^2(m+n)\right)$$
  - Several approximation algorithms (Halko et. al. 2010)
  - Some of the first randomized algorithms. (Rokhlin et. al. 2008)
  - Randomized LU decomposition for low-rank approximation (Shabat et. al. 2015)

Spectrum-revealing factorizations stronger in theory and efficiency

Background: Other Data Analysis Methods

Data Analysis: Low-rank matrix approximations that respect data sparsity, structure, and interpretability; more informative than truncated SVD.

- CX decomposition: for medical data, text data, etc

  $$A \approx CX,$$

  where C consists of well-chosen columns of A.
- CUR decomposition: for recommendation systems, text data analysis

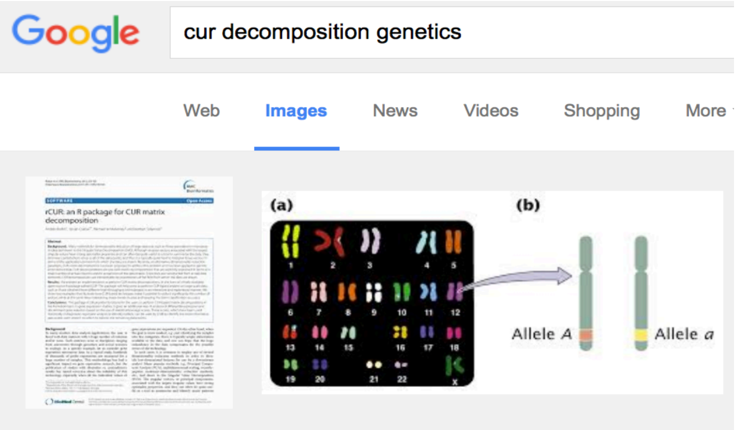
  $$A \approx CUR,$$

  where C and R consist of well-chosen columns and rows of A.
- LHL^T decomposition: Approximate kernel methods, independent component analysis

  $$A \approx LHL^T,$$

  where L consists of well-chosen columns of SPD matrix A.

Google Search on CUR Decomposition



Golden Standard: Truncated SVD

Given matrix $M \in \mathbf{R}^{m \times n}$ with $m \le n$, its SVD is

$$M = U \Sigma V^T = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix} \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}^T,$$

where σ_j is the j^{th} largest singular value of M. The rank-k truncated SVD is

$$M_k = U_k \Sigma_k V_k^T = \begin{pmatrix} u_1 & \cdots & u_k \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix} \begin{pmatrix} v_1 & \cdots & v_k \end{pmatrix}^T.$$

Theorem: (Eckart & Young, 1936)

$$\mathbf{Min}_{\text{rank}(H) \le k} \|M - H\|_2 = \|M - M_k\|_2 = \sigma_{k+1}(M).$$

Review: Rank-revealing Matrix factorizations: QR

- **Given:** Matrix $M \in \mathbb{C}^{m \times n}$ and $1 \leq \ell \leq \mathbf{min}(m, n) = n$,
  Existence: QR factorization with column permutation Π

$$M\Pi = Q \begin{pmatrix} A_\ell & B_\ell \\ 0 & C_\ell \end{pmatrix} \quad \text{with}$$

$$\begin{aligned} \frac{\sigma_j(M)}{\sqrt{1 + \ell(n-\ell)}} &\leq \sigma_j(A_\ell) \leq \sigma_j(M), \quad j = 1, \cdots, \ell, \\ \sigma_{\ell+1}(M) &\leq \|C_\ell\|_2 \leq \sigma_{\ell+1}(M)\sqrt{1 + \ell(n-\ell)}, \end{aligned}$$

where $\sigma_j(\cdot)$ is the j^{th} largest singular value.

Rank-revealing: A_ℓ reveals numerical rank of M.

Example: Kahan Matrix

- For $s, c > 0$ and $s^2 + c^2 \leq 1$, Kahan Matrix $K_n = S_n C_n$:
$$S_n = \mathbf{diag}(1, s, \cdots, s^{n-1}), C_n = \begin{pmatrix} 1 & -c & -c & \cdots & -c \\ & 1 & -c & \cdots & -c \\ & & 1 & \cdots & -c \\ & & & \ddots & \vdots \\ & & & & 1 \end{pmatrix}$$
  - Kahan Matrix not in rank revealed form:
$$\sigma_{\min}(K_n) = O\left(\left(\frac{s}{1+c}\right)^n\right) \ll s^{n-1}$$
- Little consensus on definition of "reveal."

New: Spectrum-revealing factorizations (I): QR/CX

- **Given:** Matrix $M \in \mathbb{C}^{m \times n}$ and $1 \le k \le \ell \le \mathbf{min}(m,n) = n$,
  Existence: QR factorization with column permutation Π

$$M\Pi = Q \begin{pmatrix} A_\ell & B_\ell \\ 0 & C_\ell \end{pmatrix} \quad \text{with} \quad \tau_j = \frac{\sigma_{\ell+1}(M)}{\sigma_j(M)} \quad j = 1, \cdots, k+1,$$

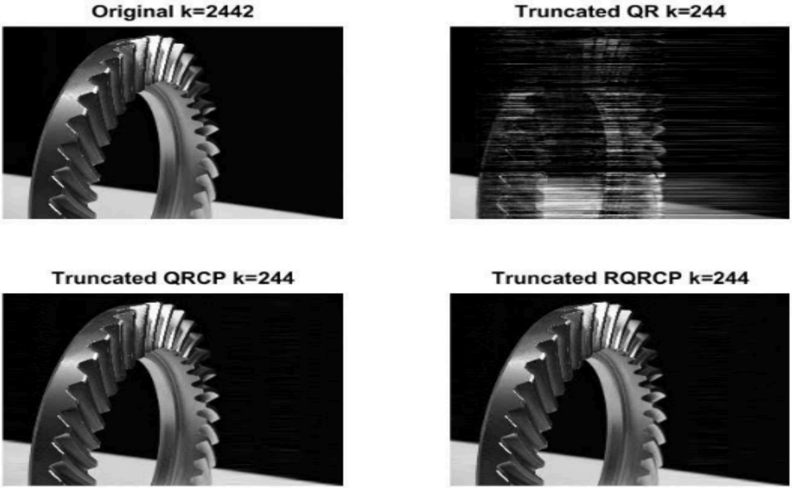
$$\frac{\sigma_j(M)}{\sqrt{1 + O\left(\tau_j^2\right)}} \le \sigma_j\left(\begin{pmatrix} A_\ell & B_\ell \end{pmatrix}\right) \le \sigma_j(M)$$

$$\sigma_{k+1}(M) \le \|M\Pi - QM_k\| \le \sigma_{k+1}(M)\sqrt{1 + O\left(\tau_{k+1}^2\right)}$$

$$\left(\text{cf. } \frac{\sigma_j(M)}{\sqrt{1 + \ell(n-\ell)}} \le \sigma_j(A_\ell), \quad \|C_\ell\|_2 \le \sigma_{\ell+1}(M)\sqrt{1 + \ell(n-\ell)}.\right)$$

M_k = truncated SVD of $\begin{pmatrix} A_\ell & B_\ell \\ & 0 \end{pmatrix}$; $\sigma_j(\cdot) = j^{\text{th}}$ largest singular value.

QM_k well approximates $M\Pi$ and reveals its leading singular values.



Original k=2442
Truncated QR k=244
Truncated QRCP k=244
Truncated RQRCP k=244

New: Spectrum-revealing factorizations (II) Cholesky

- **Given:** SPD matrix $M \in \mathbb{R}^{n \times n}$ and $1 \leq k \leq \ell \leq n$,
  Existence: Cholesky factorization with diagonal permutation

$$\Pi M \Pi^T = L L^T, \quad L = \begin{pmatrix} A_\ell & 0 \\ B_\ell & C_\ell \end{pmatrix} \quad \text{with} \quad \tau_j = \frac{\sigma_{\ell+1}(M)}{\sigma_j(M)}$$

$$\frac{\sigma_j(M)}{1 + O(\tau_j)} \leq \sigma_j^2 \begin{pmatrix} A_\ell \\ B_\ell \end{pmatrix} \leq \sigma_j(M), \quad j = 1, \cdots, k,$$

$$\sigma_{k+1}(M) \leq \|\Pi M \Pi^T - L_k L_k^T\| \leq \sigma_{k+1}(M)\left(1 + O(\tau_{k+1})\right)$$

$L_k =$ truncated SVD of $\begin{pmatrix} A_\ell \\ B_\ell \end{pmatrix}$.

New: Spectrum-revealing factorizations (II) LHL^T

- **Given:** SPD matrix $M \in \mathbb{R}^{n\times n}$ and $1 \le k \le \ell \le n$,
- **Existence:** LHL^T factorization with diagonal permutation

$$\Pi M \Pi^T = LL^T, \quad L = \begin{pmatrix} A_\ell & 0 \\ B_\ell & C_\ell \end{pmatrix},$$

$$\widehat{M} = \Pi^T \begin{pmatrix} A_\ell \\ B_\ell \end{pmatrix} \begin{pmatrix} A_\ell \\ B_\ell \end{pmatrix}^{\dagger} \left(\Pi M \Pi^T\right) \begin{pmatrix} A_\ell \\ B_\ell \end{pmatrix} \begin{pmatrix} A_\ell \\ B_\ell \end{pmatrix}^{\dagger} \Pi$$

$$\frac{\sigma_j(M)}{\sqrt{1 + O\left(\tau_j^2\right)}} \le \sigma_j\left(\widehat{M}\right) \le \sigma_j(M), \quad j = 1, \cdots, k$$

$$\sigma_{k+1}(M) \le \|M - \widehat{M}_k\| \le \sigma_{k+1}(M)\sqrt{1 + O\left(\tau_{k+1}^2\right)}$$

$\widehat{M}_k$ = truncated SVD of $\widehat{M}$; $\quad \tau_j = \frac{\sigma_{\ell+1}(M)}{\sigma_j(M)}$

New: Spectrum-revealing factorizations (III) LU

- **Given:** Matrix $M \in \mathbb{C}^{m \times n}$ and $1 \leq k \leq \ell \leq \mathbf{min}(m, n) = n$,
  Existence: LU with row, column permutations Π_l, Π_r

$$\Pi_l M \Pi_r = LU, \quad L = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}, \quad U = \begin{pmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{pmatrix}$$

$$\text{with} \quad \tau_j = \frac{\sigma_{\ell+1}(M)}{\sigma_j(M)}, \quad j = 1, \cdots, \ell$$

$$\begin{aligned} \frac{\sigma_j(M)}{1 + O(\tau_j)} &\leq \sigma_j\left(\begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix} \cdot \begin{pmatrix} U_{11} & U_{12} \end{pmatrix}\right) \leq \sigma_j(M), \\ \sigma_{k+1}(M) &\leq \|M - \widehat{M}_k\| \leq \sigma_{k+1}(M)\left(1 + O(\tau_{k+1})\right) \end{aligned}$$

$$\widehat{M}_k = \text{truncated SVD of } \begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix} \cdot \begin{pmatrix} U_{11} & U_{12} \end{pmatrix}.$$

New: Spectrum-revealing factorizations (III) CUR

- **Given:** Matrix $M \in \mathbb{C}^{m \times n}$ and $1 \leq k \leq \ell \leq \mathbf{min}(m, n) = n$,
  Existence: CUR with row, column permutations Π_l, Π_r

$$\Pi_l M \Pi_r = LU, \quad L = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}, \quad U = \begin{pmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{pmatrix}$$

$$\Pi_l \widehat{M} \Pi_r = \begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix} \begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix}^{\dagger} (\Pi_l M \Pi_r) \begin{pmatrix} U_{11} & U_{12} \end{pmatrix}^{\dagger} \begin{pmatrix} U_{11} & U_{12} \end{pmatrix}.$$

$$\frac{\sigma_j\left(M\right)}{\sqrt{1 + O\left(\tau_j^2\right)}} \leq \sigma_j\left(\widehat{M}\right) \quad \leq \quad \sigma_j\left(M\right), \quad j = 1, \cdots, k$$

$$\sigma_{k+1}\left(M\right) \leq \|M - \widehat{M}_k\| \quad \leq \quad \sigma_{k+1}\left(M\right)\sqrt{1 + O\left(\tau_{k+1}^2\right)}$$

$\widehat{M}_k$ = truncated SVD of $\widehat{M}$; $\quad \tau_j = \frac{\sigma_{\ell+1}(M)}{\sigma_j(M)}$

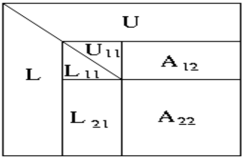
Talks: Spectrum-Revealing Matrix Factorizations and Randomized Algorithms

New Algorithms (Ia)

- Solving linear systems of equations $Ax = b$:
  - Standard method: Gaussian Elimination with Partial Pivoting (GEPP): $\Pi\, A = LU$: most efficient, but can be unstable.
  - New method: randomized Gaussian Elimination with Complete Pivoting (RGECP): $\Pi_l A \Pi_r = LU$: almost as efficient, but is always stable. (joint work with C. Melgaard.)
  - Communication-avoiding version would be numerically stable, in contrast to CA-LU with partial pivoting.

New Algorithms (Ib)

- Column pivots random: **Successive Schur Sampling.**
- RGECP is $O(rn^2)$ more flops than GEPP:



- **Theorem:** Given $\delta \in (0,1)$. If

$$r > 32 \ln\left(\frac{n(n+1)}{2\delta}\right),$$

then the pivot growth factor of RGECP satisfies

$$\rho(A) \leq 3\sqrt{e(n+1)}\, n^{2+\frac{1}{2}\ln(n)}$$

with probability greater than $1-\delta$.

New Algorithms (IIa)

- Low-rank Matrix Approximations
  - Randomized algorithm for computing a Spectrum-revealing LU Decomposition (joint work with D. Anderson): First of its kind. Computes

$$\Pi_l A \Pi_r \approx LU, \quad L = \begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix}, \quad U = \begin{pmatrix} U_{11} & U_{12} \end{pmatrix}$$

    without ever computing the Schur complement. Likely the most efficient for low-rank matrix approximation.
  - Non-classical communication patterns:
    - Computing/Updating random projection.
    - Both row and column swaps.
    - No Schur updating.

New Algorithms (IIb)

MAIN ALGORITHM: Truncated Randomized LUCP

Input: matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, target rank k, oversampling parameter p
Output: $\mathbf{L} \in \mathbb{R}^{m \times k}$, $\mathbf{U} \in \mathbb{R}^{k \times n}$, with $\mathbf{LU} \approx \mathbf{A}$

b	$<$	p	$<$	k	$\leq$	ℓ	$\ll$	m, n
block size		our oversamp. param.		target rank		prev. work oversamp. param		dim.

Initialization: compress matrix with random projection $\Omega \in \mathbb{R}^{p \times m}$:
$\mathbf{R} = \Omega\mathbf{A}$ Iterate:

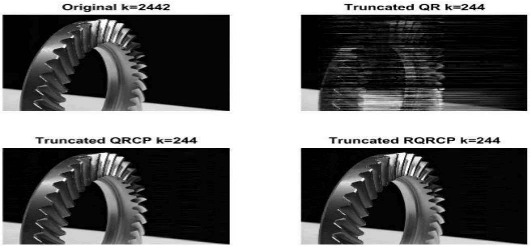
1. Perform column selection (QR or LU with partial column pivoting) on $\mathbf{R}$ to determine column pivots
2. Update block column of $\mathbf{L}$
3. Perform iteration of block LU with partial row pivots on $\mathbf{A}$
4. Update block row of $\mathbf{U}$
5. Update projection so that $\mathbf{R}$ is a random projection of the Schur complement

New Algorithms (III)

- Low-rank Matrix Approximations
  - Randomized algorithm for computing a Spectrum-revealing QR factorization (J. Duersch, Nov. 2015)

$$A\Pi \approx Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix}$$

Never need to update trailing submatrix in R. Up to twice as fast as **truncated QR.**

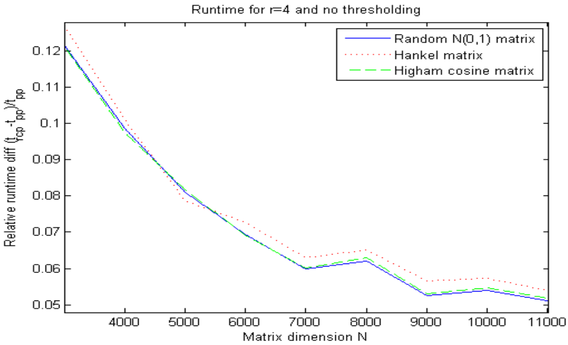


New Algorithms (IV)

- Low-rank Matrix Approximations
  - Randomized algorithm for computing a Spectrum-revealing Cholesky factorization
    Randomized block left looking Cholesky. Much promise in quality and efficiency.

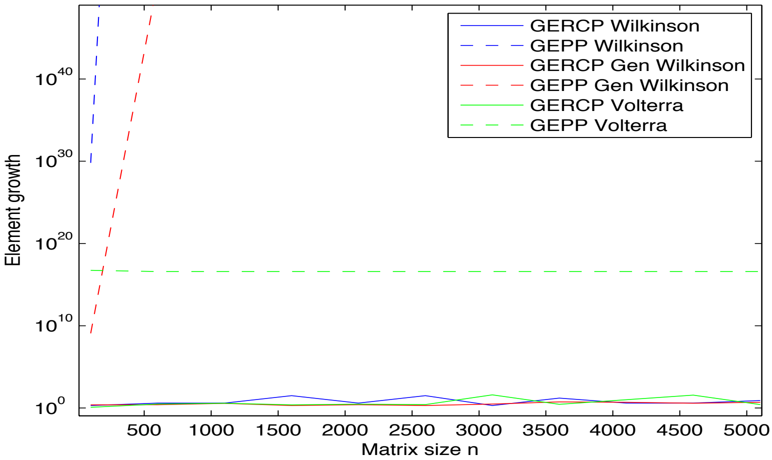
Numerical Experiments for RGECP (I)

- Fortran implementation, optimized BLAS.
- Randomized GECP: Execution Time Relative to GEPP



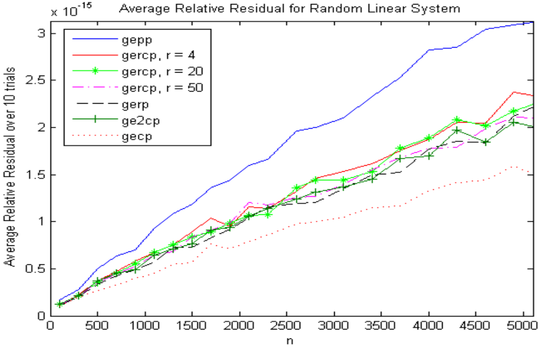
Numerical Experiments for RGECP (II)

- GEPP and Randomized GECP Element Growth



Numerical Experiments for RGECP (III)

- GEPP and Randomized GECP Residuals for random linear systems.



Numerical Experiments for Spectrum-revealing LU (I)

- SRLU vs. truncated LU, execution times.

Experiments: Benchmarking

Time comparison (seconds) with truncated DGETRF (truncated rank=300)
(rand orth)*(decaying diag)*(rand orth), numeric rank≈340

TR-LUCP (recursive column selection)

		Cols 2000	4000	6000	8000
Rows	2000	0.090	0.163	0.236	0.309
	4000	0.153	0.253	0.357	0.460
	6000	0.215	0.344	0.477	0.605
	8000	0.271	0.429	0.589	0.754

DGETRF_TRUNC

		Cols 2000	4000	6000	8000
Rows	2000	0.105	0.215	0.326	0.435
	4000	0.228	0.436	0.659	0.879
	6000	0.325	0.656	0.989	1.328
	8000	0.442	0.890	1.339	1.788

TIME RATIO

		Cols 2000	4000	6000	8000
Rows	2000	86%	75%	73%	71%
	4000	67%	58%	54%	52%
	6000	66%	52%	48%	46%
	8000	61%	48%	44%	42%

Edison, single core, MKL

Numerical Experiments for Spectrum-revealing LU (II)

- SRLU vs. truncated LU, Complexity comparison.

Experiments: Scaling

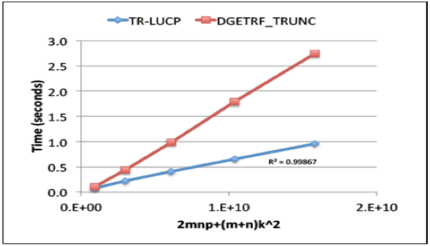
Time comparison (seconds) with truncated DGETRF (truncated rank=300)
(rand orth)*(decaying diag)*(rand orth), numeric rank≈340

TR-LUCP (recursive column selection)

Rows \ Cols	2000	4000	6000	8000	10000
2000	**0.076**	0.145	0.219	0.289	
4000	0.130	**0.218**	0.316	0.412	
6000	0.170	0.290	**0.413**	0.557	
8000	0.214	0.358	0.507	**0.657**	
10000					**0.958**

DGETRF_TRUNC

Rows \ Cols	2000	4000	6000	8000	10000
2000	**0.105**	0.215	0.326	0.435	
4000	0.228	**0.436**	0.659	0.879	
6000	0.325	0.656	**0.989**	1.328	
8000	0.442	0.890	1.339	**1.788**	
10000					**2.738**

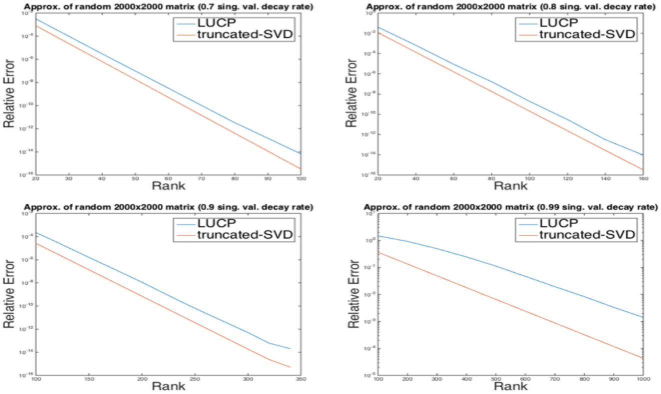


Edison, single core, MKL

Numerical Experiments for Spectrum-revealing LU (III)

- SRLU vs. truncated SVD.

Experiments: SVD-like Accuracy



Numerical Experiments for Spectrum-revealing QR (I)

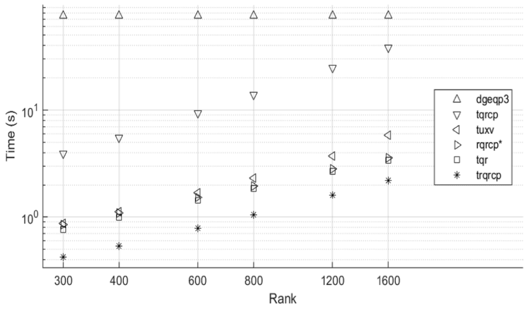
- Edison (12 core, shared memory machine).
- Matrix A is 12000×12000.
- `dgeqp2` (full decomposition), 75.84s
- `dgeqrf` (full decomposition), 5.91s
- truncated k=1200 randomized qrcp, 1.59s
- truncated k=600 randomized qrcp, 0.76s

Numerical Experiments for Spectrum-revealing QR (II)

True BLAS-3 QRCP with Random Sampling and Low Rank Approxi | Low-rank approximations

Performance of TRQRCP, rank scaling

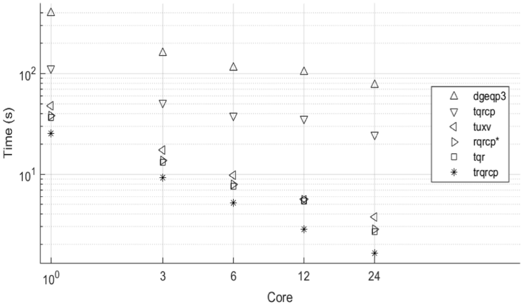
Random 12000 × 12000 rank-r factorizations on 24-cores:



Numerical Experiments for Spectrum-revealing QR (III)

Performance of TRQRCP, parallel scaling

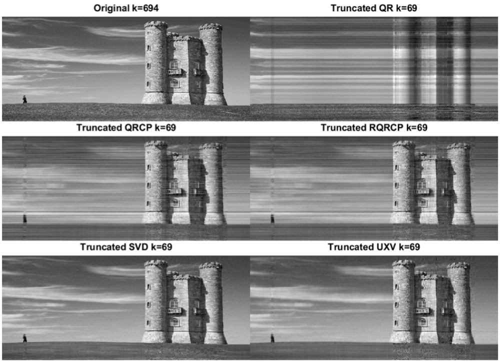
Random 12000 × 12000 rank-1200 factorizations on p-cores:



Numerical Experiments for Spectrum-revealing QR (IV)

True BLAS-3 QRCP with Random Sampling and Low Rank Approxi | Approximate SVD

Tower



Conclusions

- A new suite of randomized algorithms for numerical linear algebra and low-rank matrix approximations, along with theoretical analysis.
- Codes in different stages of development.
- Communication-avoiding variants needed.

Thank you