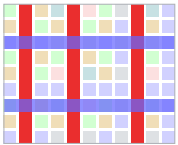
Basics of pivoting and column selection

Low-rank approximation from selected rows and columns



Alice Cortinovis
MoRAT Summer School

Road map

0. Randomized SVD and randomized Nyström for SPSD matrices

Explaining the equivalence between these.

1. Building low-rank approximations from (only?) some rows and columns

Cross approximation, CUR approximation, column subset selection, Nyström approximation.

2. Pivoting for cross approximation and Nyström

Maximum volume, complete pivoting, and randomly pivoted Cholesky.

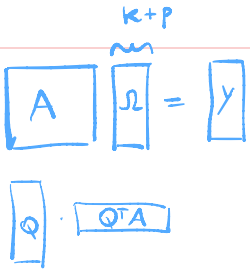
3. Column subset selection

Frobenius-norm existence results, QR with column pivoting, volume sampling, DPPs, leverage scores, and adaptive randomized pivoting.

Yuji's lecture: more on CUR and friends!

Randomized SVD and randomized Nyström

Recap of standard randomized SVD



Draw $\Omega \in \mathbb{R}^{n\times(k+p)}$ with independent Gaussian entries and form

$$Y = A\Omega, \qquad Q = \mathrm{Orth}(Y), \qquad \widehat{A} = QQ^TA.$$

For $p \geq 2$, the expected projection error satisfies the standard bound

$$\boxed{\mathbb{E}\|(I - QQ^T)A\|_F^2 \leq \left(1 + \frac{k}{p-1}\right)\|A - A_k\|_F^2.}$$

Standard randomized rangefinder result; cf. slide 46 of `intro3.pdf` and [HMT11].

The SPSD case: Randomized Nyström

Let $A \succeq 0$, draw $\Omega \in \mathbb{R}^{n\times(k+p)}$, and form $Y = A\Omega$.

Randomized Nyström approximation

$$\widehat{A} = A\Omega(\Omega^T A\Omega)^{\dagger}(A\Omega)^T.$$

It uses only $Y = A\Omega$ and small Gram matrices: one pass and streaming-compatible.

Good implementation

1. Draw $\Omega \in \mathbb{R}^{n\times\ell}$; form $Y = A\Omega$.
2. Set $\nu = \sqrt{n}\,\varepsilon\,\|Y\|$ and $Y_\nu = Y + \nu\Omega$.
3. $C = \mathrm{chol}(\Omega^T Y_\nu)$; $B = Y_\nu C^{-1}$.
4. Compute $[U, \Sigma, \sim] = \mathrm{svd}(B)$.
5. $\Lambda = \max\{0, \Sigma^2 - \nu I\}$; retain k terms.
6. Return $U\Lambda U^T$ in factorized form.

How do you write orthogonal projections?

Let $B \in \mathbb{R}^{n \times \ell}$ be tall and skinny ($n > \ell$).

How do we write the matrix that represents the orthogonal projection onto the range of B?

If B has orthonormal columns:

$$P_B = BB^T$$

If B has full column rank:

$$P_B = B\left(B^TB\right)^{-1}B^T$$

The Nyström residual

Let $Z = A^{1/2}\Omega$ and let's write the orthogonal projection onto the range of Z:

$$\Pi_Z = A^{1/2}\Omega \left(\Omega^T A^{1/2} A^{1/2} \Omega\right)^\dagger \Omega^T A^{1/2}$$
$$= A^{1/2}\Omega \left(\Omega^T A \Omega\right)^\dagger \Omega^T A^{1/2}$$

Let's write the error of the randomized Nyström approximation:

$$\text{error} = A - A\Omega(\Omega^T A\Omega)^\dagger \Omega^T A = A^{1/2}\left(I - \Pi_Z\right)A^{1/2}$$

Let's relate it to the error of randomized SVD:

$$\|\text{error}\|_* = \text{tr}\left(\text{error}\right) = \text{tr}\left(A^{1/2}(I-\Pi_Z)^2 A^{1/2}\right)$$
$$= \|(I-\Pi_Z)A^{1/2}\|_F^2$$
$$= \|(I - \Pi_{A^{1/2}\Omega})A^{1/2}\|_F^2$$

Nyström analysis from randomized SVD analysis

Gram correspondence

Nyström residual for A $\Leftrightarrow$ randSVD residual for $A^{1/2}$

Nyström nuclear-norm error for A $\Leftrightarrow$ $\|\cdot\|_F^2$ error of randSVD applied to $A^{1/2}$

Algorithms for SPSD matrix A $\Leftrightarrow$ Algorithms for column subset selection for any B s.t. $B^T B = A$

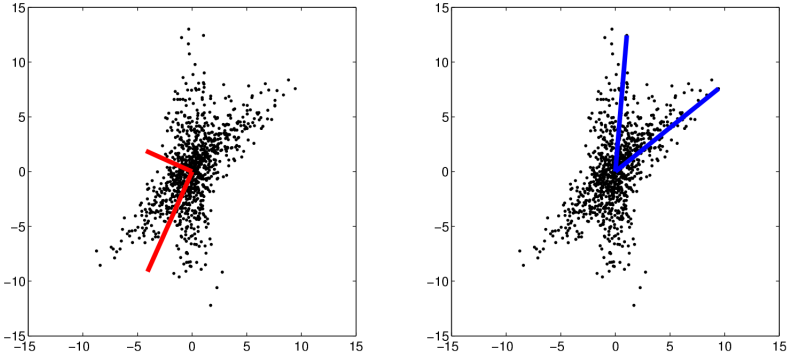
For Gaussian Ω and oversampling $p \geq 2$, the standard rangefinder bound therefore yields

$$\mathbb{E}\left\|A - \widehat{A}\right\|_* \leq \left(1 + \frac{k}{p-1}\right) \sum_{j>k} \lambda_j(A) = \left(1 + \frac{k}{p-1}\right) \|A - A_k\|_* .$$

See Ethan Epperly's blog entry on the Gram correspondence.

Rows, columns, and crosses

Sometimes, (truncated) SVD does not look great...

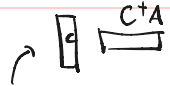


Singular vectors (left) and representatives that look closer to the data (right).

Image from Sorensen–Embree (2018), *A DEIM-induced CUR factorization.*

Column subset selection

Choose $J \subset \{1, \ldots, n\}$, $|J| = k$, and set $C = A(:, J)$.



The best approximation of the form "$C \cdot$ something" is $A \approx CC^{\dagger}A$.

Geometrical interpretation

CC^{+} = orth. projection onto Range(C)

Interpolation property

$CC^{+}A$ coincides with A in the chosen columns

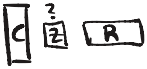
Objective

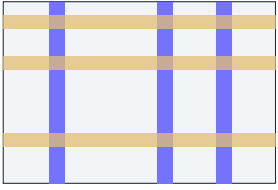
$A - CC^{+}A$ small (in some norm)

CUR approximation by orthogonal projection

Choose columns $C = A(:, J)$ and rows $R = A(I, :)$. What is the best matrix Z such that

$A \approx CZR$?



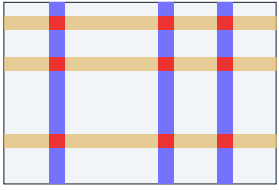


CUR approximation

$$CC^{+}AR^{+}R$$

Cross approximation

Choose I, J, with $|I| = |J| = k$: $A \approx A(:, J)A(I, J)^{-1}A(I, :)$.



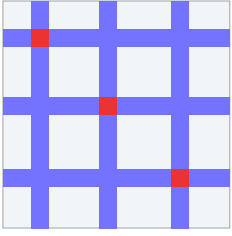
Interpolation property

$A(:,J)\,A(I,J)^{-1}\,A(I,:)$ coincides with A in the columns J and the rows I

SPSD matrices: Nyström approximation

For $A = A^T \succeq 0$, take the same indices for rows and columns:

$$A \approx A(:, J)A(J, J)^{\dagger}A(J, :).$$



Properties

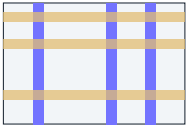
same interpolation property as before

Four ways to use selected rows and columns

Column subset selection

$P_C A$

CUR

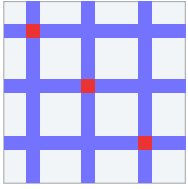


$P_C A P_R$

Cross approximation

$CA(I, J)^{-1}R$

Nyström



$A(:, J)A(J, J)^{\dagger}A(J, :)$

How should we choose the rows and/or columns?

Cross approximation and pivoting

A bad pivot can be arbitrarily bad

Consider

$$A = \begin{bmatrix} 1 & 1 \\ 1 & \varepsilon \end{bmatrix}, \qquad 0 < \varepsilon \ll 1.$$

Choose the pivot $A_{11} = 1$

$$\begin{bmatrix} 1 & 1 \\ 1 & \varepsilon \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon - 1 \end{bmatrix}$$

Choose the pivot $A_{22} = \varepsilon$

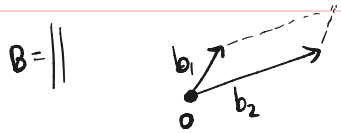
$$\begin{bmatrix} 1 & 1 \\ 1 & \varepsilon \end{bmatrix} - \begin{bmatrix} 1/\varepsilon & 1 \\ 1 & \varepsilon \end{bmatrix} = \begin{bmatrix} 1 - 1/\varepsilon & 0 \\ 0 & 0 \end{bmatrix}$$

Lesson learned:

Be careful about the choice of the indices!!

Local maximum volume is good

For a full-rank $B \in \mathbb{R}^{p \times q}$, $p \geq q$, define



$$\mathrm{Vol}(B) = \prod_{j=1}^{q} \sigma_j(B) = \sqrt{\det(B^T B)}.$$

Theorem (Goreinov–Tyrtyshnikov–Zamarashkin (1997))

Let I, J select a maximum-volume $k \times k$ submatrix of A. Then

$$\left\| A - A(:, J)A(I, J)^{-1}A(I, :) \right\|_{\max} \leq (k+1)\sigma_{k+1}(A).$$

The same is true if $A(I, J)$ has local maximum-volume in A, that is, you cannot increase its volume by changing one row index and/or one column index.

Thm: $\exists$ I, J of cardinality k s.t.
$\|A - A(:,J)A(I,J)^{-1}A(I,:)\|_F \leq (k+1)\|A - A_k\|_F$

Existence is not yet an algorithm...

Global maximum volume

NP-hard

Local maximum volume

Can be done in polynomial time (but still expensive)

Gaussian elimination with complete pivoting

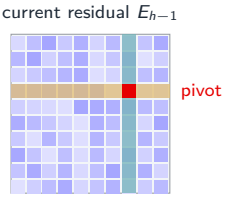
Set $E_0 = A$. At step h:

1. find the largest residual entry

$$(i_h, j_h) \in \arg\max_{i,j} |(E_{h-1})_{ij}|;$$

2. subtract the rank-one cross

$$E_h = E_{h-1} - \frac{E_{h-1}(:, j_h) E_{h-1}(i_h, :)}{E_{h-1}(i_h, j_h)}.$$



After k steps

The sum of the rank-1 factors that we subtract is

$$A(:, J)\, A(I, J)^{-1}\, A(I, :)$$

Complete pivoting is greedy volume maximization

Suppose I_{h-1}, J_{h-1} have already been selected. A bordered determinant identity gives

$$\det A(I_{h-1} \cup \{i\}, J_{h-1} \cup \{j\}) = \det A(I_{h-1}, J_{h-1})\,(E_{h-1})_{ij}.$$

Choosing the largest residual entry produces the largest possible one-step increase in volume.

Gaussian elimination with complete pivoting = greedy maxvol

Theorem

With k steps of this strategy,

$$\|A - A(:, J)A(I, J)^{-1}A(I, :)\|_{\max} \leq 4^k g_k \sigma_{k+1}(A),$$

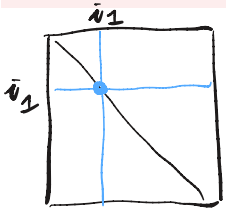
where $g_k \leq 2\sqrt{k+1}(k+1)^{\log(k+1)/4}$ is the growth factor of Gaussian elimination.

Complete-pivoting bound: [CKM20].

Cross approximation for SPSD matrices

Question

Let $A \succeq 0$. Where can you find its entry of maximum absolute value? on the diagonal!



$$A - A(:, i_1)\, A(i_1, :) \cdot \frac{1}{A(i_1, i_1)}$$

$\rightsquigarrow O(nk^2)$ time

Pivoted Cholesky = greedy Nyström

Starting from $E_0 = A \succeq 0$, choose $i_h \in \arg\max_i (E_{h-1})_{ii}$ and update

$$E_h = E_{h-1} - \frac{E_{h-1}(:, i_h) E_{h-1}(i_h, :)}{E_{h-1}(i_h, i_h)}.$$

After k steps,

$$A - E_k = A(:, J) A(J, J)^{-1} A(J, :) = F F^T.$$

Advantage over the general case

Randomly pivoted Cholesky

Replace the greedy rule by sampling from the residual diagonal:

$$\Pr(i_h = i \mid E_{h-1}) = \frac{(E_{h-1})_{ii}}{\operatorname{tr}(E_{h-1})}.$$

Then apply the same rank-one Schur-complement update.

Adaptivity

The sampling distribution changes after each step

Why randomize?

To hope to avoid the ugly 4^k factor...

Randomly pivoted Cholesky: [Che+24].

What happens after the first step (in expectation)

Residual after one step when choosing pivot i:

Average:

A multi-step guarantee for RPCholesky

Let A_r be the best rank-r approximation of $A \succeq 0$. A new analysis shows that

$$\boxed{\mathbb{E}\,\mathrm{tr}(A - \widehat{A}_k) \leq (1+\varepsilon)\,\mathrm{tr}(A - A_r)}$$

as soon as

$$k \geq \frac{r}{\varepsilon} + 2r\sqrt{\log r} + r\log\frac{1}{\varepsilon} + 2.3r.$$

Improvement over deterministic algorithm

At the cost of some oversampling, we get sth very close to the best rank-r approximation.

Corollary 1.3 of [Epp26].

Column subset selection

Projection instead of interpolation

For $C = A(:, J)$, column subset selection returns

$$P_C A = CC^\dagger A.$$

Questions we will try to answer

1) How close can we get to best rank-k approximation?
2) How do we EFFICIENTLY select columns?

The Frobenius-norm existence theorem

Theorem (Deshpande–Rademacher–Vempala–Wang)

For every $A \in \mathbb{R}^{m\times n}$ and $k < \operatorname{rank}(A)$, there exists J, $|J| = k$, such that

$$\|A - P_J A\|_F^2 \le (k+1)\|A - A_k\|_F^2 = (k+1)\sum_{i>k}\sigma_i(A)^2.$$

The proof does more: it supplies a probability distribution whose expected error satisfies this bound.

Column subset selection by volume sampling: [Des+06].

Volume sampling

For $|J| = k$, volume sampling chooses $\{1, \dots, n\}^k$

$$\mathbb{P}(J) = \frac{\det(A(:,J)^T A(:,J))}{\sum_{|S|=k} \det(A(:,S)^T A(:,S))}.$$

This is exactly a fixed-size determinantal point process (k-DPP) with kernel $L = A^T A$:

$$\det L(J,J) = \det(A(:,J)^T A(:,J)) = \mathrm{Vol}(A(:,J))^2.$$

If J is a random subset of columns according to volume sampling:

$$\mathbb{E}\left[\|A - A(:,J)A(:,J)^+ A\|_F^2\right] \le (k+1)\,\|A - A_k\|_F^2$$

$\Rightarrow$ existence follows

Volume sampling (Sampling a DPP): feasible, but usually too costly

$A \quad n \times n \quad \rightarrow \quad O(n^3)$

QR with column pivoting

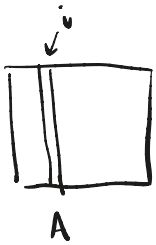
At step h, choose the column farthest from the span already selected:

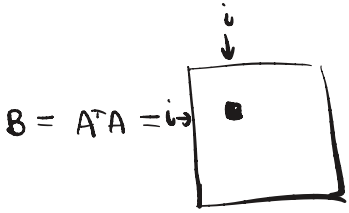
$$j_h \in \arg\max_j \left\| (I - Q_{h-1} Q_{h-1}^T) A(:, j) \right\|_2^2 .$$

Theoretical guarantee Let J be selected columns in the first k steps of CPQR. Then

$$\|A - P_J A\|_2 \leq 2^k \sqrt{n-k}\, \sigma_{k+1}(A)$$

One-to-one correspondence between CPQR and pivoted Cholesky





Randomly pivoted QR

Replace the maximum by adaptive sampling from the residual column energies:

$$\mathbb{P}(j_h = j \mid Q_{h-1}) = \frac{\left\|(I - Q_{h-1}Q_{h-1}^T)A(:,j)\right\|_2^2}{\left\|(I - Q_{h-1}Q_{h-1}^T)A\right\|_F^2}.$$

After selecting j_h, orthogonalize that column and update every residual norm.

Comparing to randomly pivoted Cholesky, are they *really* the same thing in practice?

SPSD case is "cheaper" because we don't ever need to look at the whole matrix!

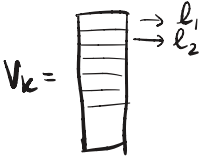
Randomly pivoted QR / adaptive sampling: [Des+06; Epp26].

Randomized column subset selection

Leverage scores

Let $A_k = U_k \Sigma_k V_k^T$. The rank-k column leverage scores are

$$\ell_j = \|V_k(j,:)\|_2^2, \qquad \sum_{j=1}^{n} \ell_j = k.$$



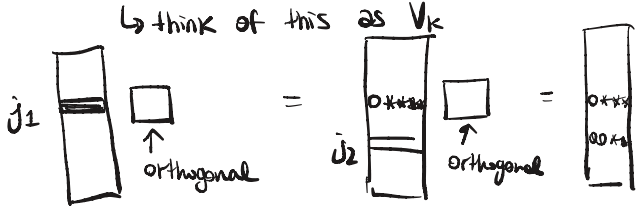
They quantify how strongly column j participates in the dominant right singular subspace.

Leverage score sampling

$O(k \log k)$ samples (iid)

Adaptive Randomized Pivoting (ARP)

Input: matrix $V \in \mathbb{R}^{n \times k}$ with orthonormal columns such that $A \approx AVV^T$.



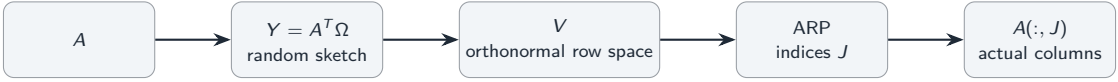
Leverage scores and adaptivity

1st sample: leverage score. Then we UPDATE the probability distribution by "eliminating" that row of V with a step of QR.
Equivalently: random CPQR on V^T

Guarantee

$$\mathbb{E}\left[\|A - A(:,J)\,A(:,J)^{+}A\|_F^2\right] \leq (k+1)\,\|A - AVV^T\|_F^2$$
$$\leq \mathbb{E}\left[\|A - A(:,J)\,V(J,:)^{-T}V^T\|_F^2\right]$$

Randomized SVD + ARP: fast overall algorithm



Conditional ARP guarantee

For the selected J,

$$\mathbb{E}_J \|A - P_J A\|_F^2 \leq (k+1) \left\|A - AVV^T\right\|_F^2 .$$

Final guarantee

Taking a Gaussian sketching matrix $\Omega \in \mathbb{R}^{n\times(k+2)}$ we have

$$\mathbb{E}_{J,\Omega} \|A - P_J A\|_F^2 \leq (k+1)(k+3) \|A - A_k\|_F^2 .$$

Some more thoughts on ARP

CPQR on V^T where V has orthonormal columns:

Back to volume sampling:

Deterministic algorithms:

References

[Che+24] Y. Chen et al. **"Randomly Pivoted Cholesky: Practical Approximation of a Kernel Matrix with Few Entry Evaluations".** In: *Communications on Pure and Applied Mathematics* (2024). DOI: 10.1002/cpa.22234.

[CK25] A. Cortinovis and D. Kressner. **"Adaptive Randomized Pivoting for Column Subset Selection, DEIM, and Low-Rank Approximation".** In: *SIAM Journal on Matrix Analysis and Applications* (2025). DOI: 10.1137/24M1719189.

[CKM20] A. Cortinovis, D. Kressner, and S. Massei. **"On Maximum Volume Submatrices and Cross Approximation for Symmetric Semidefinite and Diagonally Dominant Matrices".** In: *Linear Algebra and its Applications* 593 (2020), pp. 251–268. DOI: 10.1016/j.laa.2020.02.010.

References ii

[Des+06] A. Deshpande et al. **"Matrix Approximation and Projective Clustering via Volume Sampling".** In: *Theory of Computing* 2 (2006), pp. 225–247. DOI: 10.4086/toc.2006.v002a012.

[DM21] M. Derezinski and M. W. Mahoney. **"Determinantal Point Processes in Randomized Numerical Linear Algebra".** In: *Notices of the American Mathematical Society* 68.1 (2021), pp. 34–45.

[Epp26] E. N. W. Epperly. ***A New Analysis of the Randomly Pivoted Cholesky Algorithm.*** 2026. arXiv: 2608.20633 [math.NA].

[GE96] M. Gu and S. C. Eisenstat. **"Efficient Algorithms for Computing a Strong Rank-Revealing QR Factorization".** In: *SIAM Journal on Scientific Computing* 17.4 (1996), pp. 848–869. DOI: 10.1137/0917055.

References iii

[HMT11] N. Halko, P.-G. Martinsson, and J. A. Tropp. **"Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions".** In: *SIAM Review* 53.2 (2011), pp. 217–288. DOI: 10.1137/090771806.

[Osi24] A. I. Osinsky. **"Volume-Based Subset Selection".** In: *Numerical Linear Algebra with Applications* 31.1 (2024), e2525. DOI: 10.1002/nla.2525.