On Column-Row Matrix Approximations

Keaton Hamm
University of Arizona, USA
Email: hamm@math.arizona.edu

Longxiu Huang
Vanderbilt University, USA
Email: longxiu.huang@vanderbilt.edu

***Abstract*—This article discusses column-row factorizations of low-rank matrices, and extensions of these to approximations of full rank matrices. We give perturbation estimates for CUR decompositions, and provide some numerical illustrations of the practical aspects of these approximations.**

I. Introduction

In modern times, low rank approximations have become extremely important both in theory and application, particularly as a dimensionality reduction or compression tool. Oftentimes, data matrices are well-approximated by a low rank matrix, i.e. are of the form $\tilde{A} = A + E$, where A is low rank and E is some noise (deterministic or random). There are many classical *factorizations* of low-rank matrices which give rise to natural low-rank *approximations* of matrices. While the truncated Singular Value Decomposition (SVD) provides the best approximation (in the spectral or Frobenius norm) to a given matrix, it has an issue of interpretability and computational feasibility [9]. That is, if a matrix consists of data vectors whose interpretations are clear to a domain scientist (such as gene expression data), then using the SVD for dimension reduction could come at the cost of a clear interpretation, e.g. what is an eigengene or an eigenpatient? Additionally, computing even the truncated SVD of a huge matrix can be somewhat costly; the naïve algorithm takes $O(k^2 \min\{m, n\})$ operations for an $m \times n$ matrix, whereas finding the full SVD of the matrix has complexity $O(\min\{mn^2, m^2n\})$. Finally, we note that storing a full SVD requires storing $O(m^2 + n^2 + k)$ entries.

One of many alternative low-rank approximations is the so-called *CUR decomposition*. Here, one seeks to express a given low-rank matrix in the form $A = CUR$, where C and R are actual column and row submatrices of A, and U is suitably chosen (more on how to do this later). At minimum this amounts to storage savings in that one may store a rank k matrix via this decomposition by keeping only $O(k(m+n+k))$ values, which is notably smaller than the full SVD, but the same as the truncated SVD. Moreover, if one knows appropriate columns and rows to choose to form C and R, then the complexity of forming a suitable U can be $O(k^3)$. Many works consider random sampling of columns and rows [3], [4], [5], [9] and such methods can provide good error estimates, although there are some sensitivity issues in the presence of noise [8]; however, recent works have considered fast methods for deterministically selecting columns and rows [12], [14].

With this setup in mind, there are two natural questions to be asked:

- How should one choose columns and rows to form C and R, and how many should one choose?
- What is the right matrix U?

Let us begin by considering the second question above. Suppose we are given $A \in \mathbb{R}^{m\times n}$ and suitable column and row submatrices C and R, respectively (suitable here meaning that at least $\operatorname{rank}(C) = \operatorname{rank}(R) = \operatorname{rank}(A)$ so that there is hope of obtaining a factorization of A). Then we desire a function f satisfying the following properties:

- $A = Cf(A)R$, and
- f is stable under small perturbations, e.g. $\|f(A + E) - f(A)\|$ is small if $\|E\|$ is small.

II. Exact CUR Decomposition

We begin by illustrating a particular choice for the function f above which satisfies the factorization condition. This is the original (and most general) formulation of what is today called the CUR decomposition. It's history is not easy to pin down, but it appears implicitly at least as far back as a paper of Penrose [10]. For notation here, $[n]$ is the set $\{1, \dots, n\}$; given index sets $I \subset [m]$ and $J \subset [n]$, $A(I, J)$ is the $|I| \times |J|$ submatrix of A whose entries are indexed by $I \times J$, and $A(I, :)$ is a $|I| \times n$ row submatrix of A while $A(:, J)$ is a $m \times |J|$ column submatrix. Given a matrix A, $A^\dagger$ represents its Moore–Penrose pseudoinverse. Note that throughout this paper, $\|\cdot\|$ will represent the spectral norm of a matrix, and $\|\cdot\|_F$ is the Frobenius norm.

Theorem II.1 ([6]). *Let $A \in \mathbb{R}^{m\times n}$ have rank r, and let $I \subset [m]$ and $J \subset [n]$ with $|I| \geq r$ and $|J| \geq r$. Let $C = A(:, J)$, $R = A(I, :)$, and $U = A(I, J)$. If $\operatorname{rank}(U) = \operatorname{rank}(A)$, then*

$$A = CU^\dagger R.$$

For a very simple proof of Theorem II.1, consult [1]. This theorem suggests that one reasonable choice of f is to set $f(A) = A(I, J)$ if good index sets I and J are given.

III. Column Selection and CUR Approximations

Another way to find a CUR decomposition of A given C and R would be to find the minimizer of $\|A - CZR\|$. This minimizer is known in the Frobenius norm as the following shows:

Proposition III.1 ([13]). *Let $A \in \mathbb{R}^{m\times n}$ (not necessarily low rank) and C and R be column and row submatrices of A, respectively. Then the following holds:*

$$\underset{Z}{\operatorname{argmin}} \|A - CZR\|_F = C^\dagger A R^\dagger.$$

It should be noted that Proposition III.1 is not true for spectral norm. Given the result of Proposition III.1, much of the literature surrounding the CUR decomposition takes $A \approx CC^\dagger AR^\dagger R$ to be the CUR decomposition of A even in the case that A is not low rank.

As a first note, these two manners of performing a CUR approximation of A are the same in the exact decomposition case:

Proposition III.2. *If $A = CU^\dagger R$ is an exact CUR decomposition of A, then $U^\dagger = C^\dagger A R^\dagger$.*

Sketch of Proof. It suffices to note that $C^\dagger C$ and $U^\dagger U$ are orthogonal projections onto the same subspace, and likewise so are $RR^\dagger$ and $UU^\dagger$. Then

$$\begin{aligned} C^\dagger A R^\dagger &= C^\dagger C U^\dagger R R^\dagger \\ &= U^\dagger U U^\dagger U U^\dagger \\ &= U^\dagger, \end{aligned}$$

where the final step follows from basic properties of the Moore–Penrose pseudoinverse. □

Note also that by a simple computation noting that $CC^\dagger$ is an orthogonal projection, it follows that

$$\|A - CC^\dagger AR^\dagger R\| \le \|A - CC^\dagger A\| + \|A - AR^\dagger R\|.$$

Here, $CC^\dagger A$ is the projection of the columns of A onto the span of the columns in C, and $AR^\dagger R$ is the projection of the rows of A onto the span of the rows in R. Going back to the question of choosing good columns and rows, we see that if we are able to find the *best* column submatrix of A on which to project and similarly the best row submatrix, then we will have a good CUR approximation of A. This problem is termed the *Column Subset Selection Problem* in the theoretical computer science literature [2], and was recently shown to be NP–hard [11].

Thus far we have shown that two common choices for the function f given columns C and R are $f_1(A) = A(I, J)$ and $f_2(A) = C^\dagger A R^\dagger$. These both satisfy the first property listed in the introduction. In general, $f_1(A)$ requires much less computation than $f_2(A)$, but at the expense of not giving the minimal distance from A as in Proposition III.1. Demonstrating stability under small perturbations is difficult for general matrices given the necessity of estimating the norms of pseudoinverses of submatrices; this issue is left to future work.

IV. Perturbations of CUR Decompositions

Armed with some notion of how to form CUR decompositions of low-rank matrices, let us discuss what happens under small perturbations. We will consider matrices of the form $\tilde{A} = A + E$ where A has low rank k, and E is a (typically) full rank noise matrix. In our numerical experiments we will take E to be a random matrix, but here we do not make any assumption on its entries. We give upper bounds on a CUR approximation of $\tilde{A}$ in terms of the underlying CUR decomposition of A and the magnitude of the noise.

If $\tilde{A} = A + E$, and we consider $\tilde{C} = \tilde{A}(:, J)$, $\tilde{R} = \tilde{A}(I, :)$, and $\tilde{U} = \tilde{A}(I, J)$ for some index sets I and J, then we write

$$\tilde{C} = C + E(:, J), \quad \tilde{R} = R + E(I, :), \quad \tilde{U} = U + E(I, J) \tag{1}$$

where $C := A(:, J)$, $R := A(I, :)$, and $U := A(I, J)$. Thus if we choose columns and rows, $\tilde{C}$ and $\tilde{R}$ of $\tilde{A}$, we would like to determine how this compares to the underlying approximation of the low rank matrix A by its columns and rows, C and R.

For ease of notation, we will use the conventions that $E_I := E(I, :)$, $E_J := E(:, J)$, and $E_{I,J} := E(I, J)$. The following proposition gives a first estimate of the performance of the CUR approximation suggested by the exact CUR decomposition of Theorem II.1 in terms of the underlying CUR decomposition of A (proofs of these and further results are in a forthcoming manuscript [7]).

Proposition IV.1. *Suppose that A, C, U, and R are as in Theorem II.1 but with no assumption on the rank of U, and let $\tilde{A} = A + E$ for some fixed but arbitrary $E \in \mathbb{R}^{m\times n}$. Let $\tilde{C}, \tilde{R}$, and $\tilde{U}$ be as in* (1)*. Then the following holds:*

$$\begin{aligned} \|A - \tilde{C}\tilde{U}^\dagger \tilde{R}\| \le \|A - CU^\dagger R\| &+ \|C\tilde{U}^\dagger\| \|E_I\| \\ &+ \|\tilde{U}^\dagger \tilde{R}\| \|E_J\| \\ &+ \|CU^\dagger\| \|U^\dagger R\| \|E_{I,J}\| (3 + \|\tilde{U}^\dagger\| \|E_{I,J}\|). \end{aligned}$$

If columns and rows are chosen such that a valid CUR decomposition of the underlying low-rank matrix A is obtained (i.e. $A = CU^\dagger R$), then the error term on the right-hand side above will be dominated by the noise E as long as the other terms are not too large. However, these may be difficult to assess for general matrices. Thus Proposition IV.1 gives only a preliminary estimate, but is also flexible since it allows the use of any submultiplicative norm. While the decomposition considered here is the direct analogue of that in Theorem II.1, there is one key difference due to the presence of noise: namely that the rank of $\tilde{U}$ is typically larger than the rank of A provided more than $\operatorname{rank}(A)$ columns or rows are chosen. Therefore, $\tilde{C}\tilde{U}^\dagger \tilde{R}$ is an approximation of A that has larger rank; in fact we shall see experimentally that this leads to much worse approximation (cf. Figure 1). It is natural then to consider what happens if the target rank is enforced. By modifying the proof of Proposition IV.1, one has the following.

Proposition IV.2. *With the notations and assumptions of Proposition IV.1, suppose $\operatorname{rank}(A) = \operatorname{rank}(U) = k$, and let $\tilde{U}_k$ be the best rank k approximation of $\tilde{U}$. Then*

$$\begin{aligned} \|A - \tilde{C}\tilde{U}_k^\dagger \tilde{R}\| \le\ & \|C\tilde{U}_k^\dagger\| \|E_J\| + \|\tilde{U}_k^\dagger R\| \|E_I\| \\ & + \|CU^\dagger\| \|U^\dagger R\| (3\|U - \tilde{U}_k\| \\ & + \|\tilde{U}_k^\dagger\| \|U - \tilde{U}_k\|^2). \end{aligned}$$

Note that if we are able to select U such that $\|E_{I,J}\|$ is small compared to $\sigma_k(U)$, then $\|U^\dagger - \tilde{U}_k^\dagger\|$ can be well estimated.

Doing so will give estimations of $\|A-\tilde{C}\tilde{U}_k^\dagger\tilde{R}\|$ which depend primarily on A itself rather than $\|U^\dagger\|$ which can be arbitrarily large. The following gives a preliminary estimate.

Theorem IV.3. *With the notations and assumptions of Proposition IV.1, if additionally* $\sigma_k(U) > 2\mu\,\|E_{I,J}\|$, *then*

$$\begin{aligned}\|A-\tilde{C}\tilde{U}_k^\dagger\tilde{R}\| \leq \frac{\|U^\dagger\|}{1-2\mu\|A^\dagger\|_2\|E_{I,J}\|} \\ \big\{2\|E_{I,J}\|\left[\|E_J\|\|AR^\dagger\|+\|E_I\|\|C^\dagger A\|+\right. \\ \left.+2\|E_{I,J}\|\|AR^\dagger\|\|C^\dagger A\|\right]+\|E_I\|\|E_J\|\big\} \\ +\|AR^\dagger\|\|E_J\|+\|C^\dagger A\|\|E_I\|+6\|AR^\dagger\|\|C^\dagger A\|\|E_{I,J}\|.\end{aligned}$$

As noted in [7], the norms of $C^\dagger U$ and $UR^\dagger$ are equal to row and column submatrices of the left and right singular vector matrices of A, respectively. In the case that the rows and columns are selected to give the maximal volume submatrices of the singular vectors, then all estimates in Theorem IV.3 can be made dependent upon A and the noise E, and are of the form

$$\|A-\tilde{C}\tilde{U}_k^\dagger\tilde{R}\| = O(\|E\|+\|A^\dagger\|\|E\|^2).$$

V. Numerical simulations

Here, we illustrate the performance of some of the basic CUR approximations mentioned previously on matrices of the form $\tilde{A} = A + E$, where A is low-rank and E is a small perturbation matrix. In the experiments, we will take the entries of E to be i.i.d. Gaussian with mean zero and a prescribed variance σ^2.

Experiment 1. In this simulation, we test how enforcing the low-rank constraint on $\tilde{U}_r$ will influence the relative error $\|A-\tilde{C}\tilde{U}_r^\dagger\tilde{R}\|_2/\|A\|_2$, where r varies from $\mathrm{rank}\,(A)$ to $\mathrm{rank}\,(\tilde{U})$. We consider a 500×500 matrix of rank 10 perturbed by Gaussian noise with standard deviation 10^{-4}. We randomly choose 60 columns and rows, and for each fixed r, we repeat this process 100 times and compute the average error. For illustration, we use three common sampling methods: Leverage Scores, column lengths, and uniform; each corresponds to sampling columns with replacement with probabilities given by

$$p_i^{\text{lev}} := \frac{1}{k}\|V_k(i,:)\|_2^2, \quad p_i^{\text{col}} := \frac{\|A(:,i)\|_2^2}{\|A\|_F^2}, \quad p_i^{\text{unif}} := \frac{1}{n}.$$

The row sampling probabilities are defined analogously. Figure 1 shows that if r is closer to $\mathrm{rank}\,(A)$, the relative error is smaller as one might expect, while as the rank increases the error is saturated by the noise.

Experiment 2. In this simulation, the set-up is the same as in Experiment 1. We compared the relative error $\|A-\tilde{C}\tilde{U}_r^\dagger\tilde{R}\|_2/\|A\|_2$ with the error $\|A-\tilde{C}\tilde{C}^\dagger\tilde{A}\tilde{R}^\dagger\tilde{R}\|_2/\|A\|_2$, when r varies from rank(A) to rank(U). The result is reported in Figure 2. The figure shows that if we could set $r = rank(A)$, then $\tilde{C}\tilde{U}_r^\dagger\tilde{R}$ is almost the best CUR estimation of A.

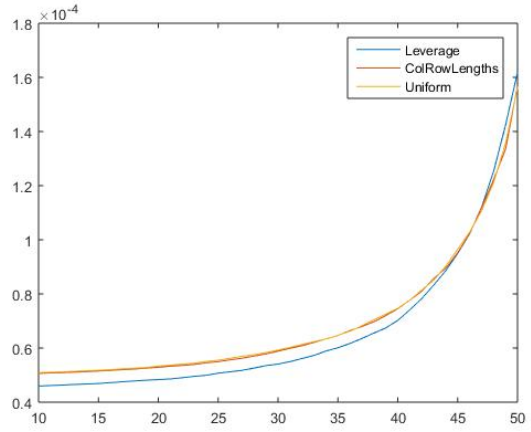

Fig. 1: Averaged errors $\|A-\tilde{C}\tilde{U}_r^\dagger\tilde{R}\|_2/\|A\|_2$ vs. r over 100 trials of sampling columns/rows.

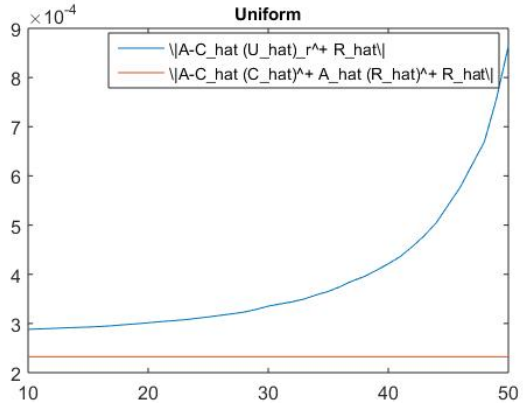

Fig. 2: The blue line represents $\|A-\tilde{C}\tilde{U}_r^\dagger\tilde{R}\|_2/\|A\|_2$ vs. r which varies from $\mathrm{rank}\,(A)$ to 50 for averaged errors over 100 trials of sampling columns and rows uniformly, and the red line stands for the averaged error $\|A-\tilde{C}\tilde{C}^\dagger\tilde{A}\tilde{R}^\dagger\tilde{R}\|_2/\|A\|_2$ over 100 trials of sampling columns/rows under the uniform probability.

Acknowledgements

KH was partially supported through the NSF TRIPODS project under grant CCF-1423411.

References

[1] Akram Aldroubi, Keaton Hamm, Ahmet Bugra Koku, and Ali Sekmen. CUR decompositions, similarity matrices, and subspace clustering. *Frontiers in Applied Mathematics and Statistics*, 4:65, 2019.

[2] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

[3] Jiawei Chiu and Laurent Demanet. Sublinear randomized algorithms for skeleton decompositions. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1361–1383, 2013.

[4] Petros Drineas and Michael W Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(Dec):2153–2175, 2005.

[5] Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

[6] S. A. Goreĭnov, N. L. Zamarashkin, and E. E. Tyrtyshnikov. Pseudo-skeleton approximations of matrices. *Dokl. Akad. Nauk*, 343(2):151–152, 1995.

[7] Keaton Hamm and Longxiu Huang. Cur decompositions, approximations, and perturbations. *arXiv preprint arXiv:1903.09698*, 2019.

[8] John T Holodnak, Ilse CF Ipsen, and Thomas Wentworth. Conditioning of leverage scores and computation by qr decomposition. *SIAM Journal on Matrix Analysis and Applications*, 36(3):1143–1163, 2015.

[9] Michael W Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[10] R. Penrose. On best approximate solutions of linear matrix equations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 52(1):1719, 1956.

[11] Yaroslav Shitov. Column subset selection is NP-complete. *arXiv preprint arXiv:1701.02764*, 2017.

[12] Danny C Sorensen and Mark Embree. A DEIM induced CUR factorization. *SIAM Journal on Scientific Computing*, 38(3):A1454–A1482, 2016.

[13] GW Stewart. Four algorithms for the the efficient computation of truncated pivoted qr approximations to a sparse matrix. *Numerische Mathematik*, 83(2):313–323, 1999.

[14] Sergey Voronin and Per-Gunnar Martinsson. Efficient algorithms for CUR and interpolative matrix decompositions. *Advances in Computational Mathematics*, 43(3):495–516, 2017.