

Efficient Algorithms for CUR and Interpolative Matrix Decompositions

Sergey Voronin¹ and Per-Gunnar Martinsson²

¹Department of Mathematics, Tufts University, Medford, MA 02155, USA

²Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, USA

October 20, 2016

Abstract

The manuscript describes efficient algorithms for the computation of the CUR and ID decompositions. The methods used are based on simple modifications to the classical truncated pivoted QR decomposition, which means that highly optimized library codes can be utilized for implementation. For certain applications, further acceleration can be attained by incorporating techniques based on randomized projections. Numerical experiments demonstrate advantageous performance compared to existing techniques for computing CUR factorizations.

1 Introduction

In many applications, it is useful to approximate a matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ by a factorization of rank $k < \min(m, n)$. When the singular values of \mathbf{A} decay sufficiently fast so that an accurate approximation can be obtained for a rank k that is substantially smaller than either m or n , great savings can be obtained both in terms of storage requirements, and in terms of speed of any computations involving \mathbf{A} . A low rank approximation that is in many ways optimal is the truncated *singular value decomposition* (SVD) of rank k , which approximates \mathbf{A} via the product

$$\begin{matrix} \mathbf{A} & \approx & \mathbf{U}_k & \mathbf{\Sigma}_k & \mathbf{V}_k^* \\ m \times n & & m \times k & k \times k & k \times n \end{matrix} \quad (1.1)$$

where the columns of the orthonormal matrices \mathbf{U}_k and \mathbf{V}_k are the left and right singular vectors of \mathbf{A} , and where $\mathbf{\Sigma}_k$ is a diagonal matrix holding the singular values of \mathbf{A} . However, a disadvantage of the low rank SVD is its storage requirements. Even if \mathbf{A} is a sparse matrix, \mathbf{U}_k and \mathbf{V}_k are usually dense. This means that if \mathbf{A} is large and very sparse, compression via the SVD is only efficient when the rank k is *much* smaller than $\min(m, n)$.

As an alternative to the SVD, the so called *CUR-factorization* [8, 19, 13] has recently received much attention [15, 21]. The CUR-factorization approximates an $m \times n$ matrix \mathbf{A} as a product

$$\begin{matrix} \mathbf{A} & \approx & \mathbf{C} & \mathbf{U} & \mathbf{R}, \\ m \times n & & m \times k & k \times k & k \times n \end{matrix} \quad (1.2)$$

where \mathbf{C} contains a subset of the columns of \mathbf{A} and \mathbf{R} contains a subset of the rows of \mathbf{A} . The key advantage of the CUR is that the factors \mathbf{C} and \mathbf{R} (which are typically much larger than \mathbf{U}) inherit properties such as *sparsity* or *non-negativity* from \mathbf{A} . Also, the index sets that point out which columns and rows of \mathbf{A} to include in \mathbf{C} and \mathbf{R} often assist in *data interpretation*. Numerous algorithms for computing the CUR factorization have been proposed (see e.g. [5, 21]), with some of the most recent and popular approaches relying on a method known as leverage scores [5, 13], a notion originating from statistics [11].

A third factorization which is closely related to the CUR is the so called *interpolative decomposition* (ID), which decomposes \mathbf{A} as

$$\begin{matrix} \mathbf{A} & \approx & \mathbf{C} & \mathbf{V}^*, \\ m \times n & & m \times k & k \times n \end{matrix}, \quad (1.3)$$

where again \mathbf{C} consists of k columns of \mathbf{A} . The matrix \mathbf{V} contains a $k \times k$ identity matrix as a submatrix and can be constructed so that $\max_{i,j} |\mathbf{V}(i,j)| \leq 1$, making \mathbf{V} fairly well-conditioned. Of course, one could equally well express \mathbf{A} as

$$\begin{matrix} \mathbf{A} & \approx & \mathbf{W} & \mathbf{R}, \\ m \times n & & m \times k & k \times n \end{matrix}, \quad (1.4)$$

where \mathbf{R} holds k rows of \mathbf{A} , and the properties of \mathbf{W} are analogous to those of \mathbf{V} . A third variation of this idea is the *two-sided interpolative decomposition* (tsID), which decomposes \mathbf{A} as the product

$$\begin{matrix} \mathbf{A} & \approx & \mathbf{W} & \mathbf{A}_{\text{skel}} & \mathbf{V}^*, \\ m \times n & & m \times k & k \times k & k \times n \end{matrix}, \quad (1.5)$$

where \mathbf{A}_{skel} consists of a $k \times k$ submatrix of \mathbf{A} . The two sided ID allows for data interpretation in a manner entirely analogous to the CUR, but has an advantage over the CUR in that it is inherently better conditioned, cf. Remark 2.3. On the other hand, the factors \mathbf{W} and \mathbf{V} do not inherit properties such as sparsity or non-negativity. This makes the two-sided ID only marginally better than the SVD in terms of storage requirements for sparse matrices.

In this manuscript, we describe a set of efficient algorithms for computing approximate ID and CUR factorizations. The algorithms are obtained via slight variations on the classical “rank-revealing QR” factorizations [4] and are easy to implement—the most expensive parts of the computation can be executed using highly optimized standard libraries such as, e.g., LAPACK [1]. We also demonstrate how the computations can be accelerated by using randomized algorithms [10]. For instance, randomization allows us to improve the asymptotic complexity of computing the CUR decomposition from $O(mnk)$ to $O(mn \log(k) + (m+n)k^2)$. Section 6 illustrates via several numerical examples that the techniques described here for computing the CUR factorization compare favorably in terms of both speed and accuracy with recently proposed CUR implementations. All the ID and CUR factorization algorithms discussed in this article are efficiently implemented as part of the open source RSVDPACK package [20].

2 Preliminaries

In this section we review some existing matrix decompositions, notably the pivoted QR, ID, and CUR decompositions [10]. We follow the notation of [7] (the so called “Matlab style notation”): given any matrix \mathbf{A} and (ordered) subindex sets I and J , $\mathbf{A}(I, J)$ denotes the submatrix of \mathbf{A} obtained by extracting the rows and columns of \mathbf{A} indexed by I and J , respectively; and $\mathbf{A}(:, J)$ denotes the submatrix of \mathbf{A} obtained by extracting the columns of \mathbf{A} indexed by J . For any positive

integer k , $1 : k$ denotes the ordered index set $(1, \dots, k)$. We take $\|\cdot\|$ to be the spectral or operator norm (largest singular value) and $\|\cdot\|_F$ the Frobenius norm: $\|x\|_F = \left(\sum_{k=1}^n |x_k|^2\right)^{\frac{1}{2}}$.

2.1 The singular value decomposition (SVD)

The SVD was introduced briefly in the introduction. Here we define it again, with some more detail added. Let \mathbf{A} denote an $m \times n$ matrix, and set $r = \min(m, n)$. Then \mathbf{A} admits a factorization

$$\begin{array}{ccccc} \mathbf{A} & = & \mathbf{U} & \mathbf{\Sigma} & \mathbf{V}^*, \\ m \times n & & m \times r & r \times r & r \times n \end{array}, \quad (2.1)$$

where the matrices \mathbf{U} and \mathbf{V} are orthonormal, and $\mathbf{\Sigma}$ is diagonal. We let $\{\mathbf{u}_i\}_{i=1}^r$ and $\{\mathbf{v}_i\}_{i=1}^r$ denote the columns of \mathbf{U} and \mathbf{V} , respectively. These vectors are the left and right singular vectors of \mathbf{A} . As in the introduction, the diagonal elements $\{\sigma_j\}_{j=1}^r$ of $\mathbf{\Sigma}$ are the singular values of \mathbf{A} . We order these so that $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$. We let \mathbf{A}_k denote the truncation of the SVD to its first k terms, $\mathbf{A}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^*$. It is easily verified that

$$\|\mathbf{A} - \mathbf{A}_k\| = \sigma_{k+1}, \quad \text{and that} \quad \|\mathbf{A} - \mathbf{A}_k\|_F = \left(\sum_{j=k+1}^{\min(m,n)} \sigma_j^2 \right)^{1/2}. \quad (2.2)$$

Moreover, the Eckart-Young theorem [6] states that these errors are the smallest possible errors that can be incurred when approximating \mathbf{A} by a matrix of rank k .

2.2 Pivoted QR factorizations

Let \mathbf{A} be an $m \times n$ matrix with real or complex entries, and set $r = \min(m, n)$. The (compact) QR-factorization of \mathbf{A} then takes the form

$$\begin{array}{ccccc} \mathbf{A} & \mathbf{P} & = & \mathbf{Q} & \mathbf{S}, \\ m \times n & n \times n & & m \times r & r \times n \end{array}, \quad (2.3)$$

where \mathbf{P} is a permutation matrix, \mathbf{Q} has orthonormal columns, and \mathbf{S} is upper triangular (the matrix we call “ \mathbf{S} ” is customarily labeled “ \mathbf{R} ”, but we use that letter for one of the factors in the CUR-decomposition). The permutation matrix \mathbf{P} can more efficiently be represented via a vector $J \in \mathbb{Z}_+^n$ of indices such that $\mathbf{P} = \mathbf{I}(:, J)$ where \mathbf{I} is the $n \times n$ identity matrix. The factorization (2.3) can then be written

$$\begin{array}{ccccc} \mathbf{A}(:, J) & = & \mathbf{Q} & \mathbf{S}. \\ m \times n & & m \times r & r \times n \end{array} \quad (2.4)$$

The QR-factorization is often computed via column pivoting combined with either the Gram-Schmidt process, Householder reflectors [7], or Givens rotations [4]. The resulting factor \mathbf{S} then satisfies various decay conditions [7], such as:

$$\mathbf{S}(j, j) \geq \|\mathbf{S}(j : m, \ell)\|_2 \quad \text{for all } j < \ell.$$

The QR-factorization (2.4) expresses \mathbf{A} as a sum of r rank-one matrices

$$\mathbf{A}(:, J) \approx \sum_{j=1}^r \mathbf{Q}(:, j) \mathbf{S}(j, :).$$

The QR-factorization is often built incrementally via a greedy algorithm such as column pivoted Gram-Schmidt. This opens up the possibility of stopping after the first k terms have been computed and settling for a “partial QR-factorization of \mathbf{A} ”. We can express the error term by splitting the factors in (2.4) as follows:

$$\mathbf{A}(:, J) = \begin{matrix} & k & r-k \\ m & \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \end{matrix} \times \begin{matrix} k \\ r-k \end{matrix} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix} = \mathbf{Q}_1 \mathbf{S}_1 + \mathbf{Q}_2 \mathbf{S}_2. \quad (2.5)$$

Observe that since the SVD is optimal, it is always the case that

$$\sigma_{k+1}(\mathbf{A}) \leq \|\mathbf{Q}_2 \mathbf{S}_2\| = \|\mathbf{S}_2\|.$$

We say that a factorization is a “rank-revealing QR-factorization (RRQR)” if the ratio $\frac{\|\mathbf{S}_2\|}{\sigma_{k+1}(\mathbf{A})}$ is guaranteed to be bounded [9]. (Some authors require additionally that $\sigma_j(\mathbf{S}_1) \approx \sigma_j(\mathbf{A})$ for $1 \leq j \leq k$). Classical column pivoted Gram-Schmidt *typically* results in an RRQR, but there are counter-examples. More sophisticated versions such as [9] provably compute an RRQR, but are substantially harder to code, and the gain compared to standard methods is typically modest.

2.3 Low rank interpolative decomposition

An approximate rank k interpolative decomposition (ID) of a matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ is the approximate factorization:

$$\begin{matrix} \mathbf{A} \\ m \times n \end{matrix} \approx \begin{matrix} \mathbf{C} \\ m \times k \end{matrix} \begin{matrix} \mathbf{V}^* \\ k \times n \end{matrix}, \quad (2.6)$$

where the partial column skeleton $\mathbf{C} \in \mathbb{C}^{m \times k}$ is given by a subset of the columns of \mathbf{A} and \mathbf{V} is well-conditioned in a sense that we will make precise shortly. The interpolative decomposition approximates \mathbf{A} using only some of its columns, and one of the advantages of doing so is that the more compact description of the range of \mathbf{A} given by its skeleton preserves some of the properties of the original matrix \mathbf{A} such as sparsity and non-negativity. In this section we show one way of obtaining a low rank interpolative decomposition, via the truncated QR with column pivoting.

From (2.5), we see that as long as $\|\mathbf{S}_2\|_2$ is small, we can approximate $\mathbf{A}(:, J)$ by $\mathbf{Q}_1 \mathbf{S}_1$. We show that the approximation term $\mathbf{Q}_1 \mathbf{S}_1$ provides a rank k ID to the matrix \mathbf{A} . In fact, the approximation term $\mathbf{Q}_1 \mathbf{S}_1$ is the image of a skeleton of \mathbf{A} , i.e., the range of $\mathbf{Q}_1 \mathbf{S}_1$ is contained in the span of k columns of \mathbf{A} . Splitting the columns of \mathbf{S}_1 and \mathbf{S}_2 as follows:

$$\mathbf{S}_1 = \begin{matrix} k & n-k \\ k & \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \end{bmatrix} \end{matrix} \quad \text{and} \quad \mathbf{S}_2 = \begin{matrix} k & n-k \\ r-k & \begin{bmatrix} \mathbf{0} & \mathbf{S}_{22} \end{bmatrix} \end{matrix}, \quad (\text{i.e., } \mathbf{S} = \begin{matrix} k & n-k \\ r-k & \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix} \end{matrix},) \quad (2.7)$$

it is immediate that

$$\mathbf{A}(:, J) = \mathbf{Q}_1 \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \end{bmatrix} + \mathbf{Q}_2 \begin{bmatrix} \mathbf{0} & \mathbf{S}_{22} \end{bmatrix} = \begin{matrix} k & n-k \\ m & \begin{bmatrix} \mathbf{Q}_1 \mathbf{S}_{11} & \mathbf{Q}_1 \mathbf{S}_{12} + \mathbf{Q}_2 \mathbf{S}_{22} \end{bmatrix} \end{matrix}.$$

In other words, we see that the matrix $\mathbf{Q}_1 \mathbf{S}_{11}$ equals the first k columns of $\mathbf{A}(:, J)$. We now define the factor \mathbf{C} in (2.6) via

$$\mathbf{C} := \mathbf{A}(:, J(1:k)) = \mathbf{Q}_1 \mathbf{S}_{11}.$$

Then the dominant term $\mathbf{Q}_1 \mathbf{S}_1$ in (2.5) can be written

$$\mathbf{Q}_1 \mathbf{S}_1 = \begin{bmatrix} \mathbf{Q}_1 \mathbf{S}_{11} & \mathbf{Q}_1 \mathbf{S}_{12} \end{bmatrix} = \mathbf{Q}_1 \mathbf{S}_{11} \begin{bmatrix} \mathbf{I}_k & \mathbf{T}_l \end{bmatrix} = \mathbf{C} \begin{bmatrix} \mathbf{I}_k & \mathbf{T}_l \end{bmatrix},$$

where \mathbf{T}_l is a solution to the matrix equation

$$\mathbf{S}_{11} \mathbf{T}_l = \mathbf{S}_{12}. \quad (2.8)$$

The equation (2.8) obviously has a solution whenever \mathbf{S}_{11} is non-singular. If \mathbf{S}_{11} is singular, then one can show that \mathbf{A} must necessarily have rank k' less than k , and the bottom $k - k'$ rows in (2.8) consist of all zeros, so there exists a solution in this case as well. We now recover the factorization (2.6) upon setting

$$\mathbf{V}^* = [\mathbf{I}_k \quad \mathbf{T}_l] \mathbf{P}^*. \quad (2.9)$$

The approximation error of the ID obtained via truncated QR with pivoting is the same as that of the truncated QR:

$$\mathbf{A} - \mathbf{C}\mathbf{V}^* = \mathbf{Q}_2 \mathbf{S}_{22} \quad (2.10)$$

Remark 2.1 *This section describes a technique for converting a QR decomposition of \mathbf{A} into the interpolative decomposition (1.3). By applying an analogous procedure to the adjoint \mathbf{A}^* of \mathbf{A} , we obtain the sibling factorization (1.4) that uses a sub-selection of rows of \mathbf{A} to span the row space. In other words, to find the column skeleton, we perform Gram-Schmidt on the columns on \mathbf{A} , and in order to find the row skeleton, we perform Gram-Schmidt to the rows of \mathbf{A} .*

2.4 Two sided interpolative decomposition

A two sided ID approximation for matrices, is constructed via two successive one sided IDs. Assume that we have performed the one sided decomposition to obtain (2.9). Then perform an ID of the adjoint of \mathbf{C} to determine a matrix \mathbf{W} and an index vector I such that

$$\begin{array}{ccc} \mathbf{C}^* & = & \mathbf{C}(I(1:k), :)^* \mathbf{W}^*. \\ k \times m & & k \times k \quad k \times m \end{array} \quad (2.11)$$

In other words, the index vector I is obtained by performing a pivoted Gram-Schmidt process on the rows of \mathbf{C} . Observe that the factorization (2.11) is exact since it is a *full* (as opposed to *partial*) QR factorization. We next insert (2.11) into (2.6), using that $\mathbf{C}(I(1:k), :) = \mathbf{A}(I(1:k), J(1:k))$, and obtain

$$\mathbf{A} \approx \mathbf{C}\mathbf{V}^* = \mathbf{W}\mathbf{A}(I(1:k), J(1:k))\mathbf{V}^*. \quad (2.12)$$

We observe that the conversion of the single-sided ID (2.9) into the two-sided ID (2.12) is *exact* in the sense that no additional approximation error is incurred:

$$\mathbf{A} - \mathbf{C}\mathbf{V}^* = \mathbf{A} - \mathbf{W}\mathbf{A}(I(1:k), J(1:k))\mathbf{V}^* = \mathbf{Q}_2 \mathbf{S}_2.$$

Remark 2.2 *The index vector I and the basis matrix \mathbf{W} computed using the approach described in this section form an approximate row-ID for \mathbf{A} in the sense that $\mathbf{A} \approx \mathbf{W}\mathbf{A}(I, :)$. However, the resulting error tends to be slightly higher than the error incurred if Gram-Schmidt is performed directly on the rows of \mathbf{A} (rather than on the rows of \mathbf{C}), cf. Lemma 3.2.*

2.5 The CUR Decomposition

A rank k CUR factorization of a matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ is given by

$$\begin{array}{ccccc} \mathbf{A} & \approx & \mathbf{C} & \mathbf{U} & \mathbf{R}, \\ m \times n & & m \times k & k \times k & k \times n \end{array}$$

where \mathbf{C} consists of k columns of \mathbf{A} , and \mathbf{R} consists of k rows of \mathbf{A} . The decomposition is typically obtained in three steps [15]. First, some scheme is used to assign a weight or the so called leverage score (of importance) to each column and row in the matrix. This is typically done either using the ℓ_2 norms of the columns and rows or by using the leading singular vectors of \mathbf{A} [5]. Next, the matrices \mathbf{C} and \mathbf{R} are constructed via a randomized sampling procedure, using the leverage scores to assign a sampling probability to each column and row. Finally, the \mathbf{U} matrix is computed via:

$$\mathbf{U} \approx \mathbf{C}^\dagger \mathbf{A} \mathbf{R}^\dagger, \quad (2.13)$$

with \mathbf{C}^\dagger and \mathbf{R}^\dagger being the pseudoinverses of \mathbf{C} and \mathbf{R} .

Many techniques for computing CUR factorizations have been proposed. In particular, we mention the recent work of Sorensen and Embree [18] on the DEIM-CUR method. A number of standard CUR algorithms is implemented in the software package rCUR [2] which we use for our numerical comparisons. The methods in the rCUR package utilize eigenvectors to assign weights to columns and rows of \mathbf{A} . Computing the eigenvectors exactly amounts to doing the SVD which is very expensive. However, instead of the full SVD, when a CUR of rank k is required, we can utilize instead the randomized SVD algorithm [10] to compute an approximate SVD of rank k at substantially lower cost.

Remark 2.3 (Conditioning of CUR) *For matrices whose singular value experience substantial decay, the accuracy of the CUR factorization can deteriorate due to effects of ill-conditioning. To simplify slightly, one would normally expect the leading k singular values of \mathbf{C} and \mathbf{R} to be of roughly the same order of magnitude as the leading k singular values of \mathbf{A} . Since low-rank factorizations are most useful when applied to matrices whose singular values decay reasonably rapidly, we would typically expect \mathbf{C} and \mathbf{R} to be highly ill-conditioned, with condition numbers roughly on the order of $\sigma_1(\mathbf{A})/\sigma_k(\mathbf{A})$. Hence, in the typical case, evaluation of the formula (2.13) can be expected to result in substantial loss of accuracy due to accumulation of round-off errors. Observe that the ID does not suffer from this problem; in (1.5), the matrix \mathbf{A}_{skel} tends to be ill-conditioned, but it does not need to be inverted. (The matrices \mathbf{W} and \mathbf{V} are well-conditioned.)*

3 The CUR-ID algorithm

In this section, we demonstrate that the CUR decomposition can easily be constructed from the basic two-sided ID (which in turn, recall, can be built from a column pivoted QR factorization), via a procedure we call “CUR-ID”. The difference between recently popularized algorithms for CUR computation and CUR-ID is in the choice of columns and rows of \mathbf{A} for forming \mathbf{C} and \mathbf{R} . In the CUR-ID algorithm, the columns and rows are chosen via the two sided ID. The idea behind the use of ID for obtaining the CUR factorization is that the matrix \mathbf{C} in the CUR factorization is immediately available from the ID (see (2.9)), and the matrix $\mathbf{V} \in \mathbb{C}^{n \times k}$ not only captures a rough row space description of \mathbf{A} but also is of rank at most k . A rank k ID on \mathbf{C} , being an exact factorization of \mathbf{C} which is of rank at most k , could hint on the relevant rows of \mathbf{A} that approximate the entire row space of \mathbf{A} itself. Specifically, similar to (2.9) where approximating $\text{range}(\mathbf{A})$ using \mathbf{C} incurs an error term $[\mathbf{0} \quad \mathbf{Q}_2 S_{22}]$, we can estimate the error of approximating $\text{range}(\mathbf{A}^*)$ using $\mathbf{A}(I(1:k), :)$; see Lemma 3.2 below.

The CUR-ID algorithm is based on the two sided ID factorization, and as a starting point, we assume the factorization (2.12) has been computed using the procedures described in Section 2. In other words, we assume that the index vectors I and J , and the basis matrices \mathbf{V} and \mathbf{W} , are all available. We then define

$$\mathbf{C} = \mathbf{A}(:, J(1:k)) \quad \text{and} \quad \mathbf{R} = \mathbf{A}(I(1:k), :). \quad (3.1)$$

Consequently, \mathbf{C} and \mathbf{R} are respectively subsets of columns and of rows of \mathbf{A} , with J and I determined by the pivoted QR factorizations. Next we construct a $k \times k$ matrix \mathbf{U} such that $\mathbf{A} \approx \mathbf{CUR}$. We know that

$$\mathbf{A} \approx \mathbf{C}\mathbf{V}^*, \quad (3.2)$$

and we seek a factor \mathbf{U} such that

$$\mathbf{A} \approx \mathbf{CUR}. \quad (3.3)$$

By inspecting (3.2) and (3.3), we find that we would achieve our objective if we could determine a matrix \mathbf{U} such that

$$\begin{array}{ccc} \mathbf{U} & \mathbf{R} & = \mathbf{V}^* \\ k \times k & k \times m & k \times m \end{array} \quad (3.4)$$

Unfortunately, (3.4) is an over-determined system, but at least intuitively, it seems plausible that it should have a fairly accurate solution, given that the rows of \mathbf{R} and the rows of \mathbf{V}^* should, by construction, span roughly the same space (namely, the space spanned by the k leading right singular vectors of \mathbf{A}). Solving (3.4) in the least-square sense, we arrive at our definition of \mathbf{U} :

$$\mathbf{U} := \mathbf{V}^* \mathbf{R}^\dagger. \quad (3.5)$$

The construction of \mathbf{C} , \mathbf{U} , and \mathbf{R} in the previous paragraph was based on heuristics. We next demonstrate that the approximation error is comparable to the error resulting from the original QR-factorization. First, let us define \mathbf{E} and $\tilde{\mathbf{E}}$ as the errors in the column and row IDs of \mathbf{A} , respectively,

$$\mathbf{A} = \mathbf{C}\mathbf{V}^* + \mathbf{E}, \quad (3.6)$$

$$\mathbf{A} = \mathbf{W}\mathbf{R} + \tilde{\mathbf{E}}. \quad (3.7)$$

Recall that \mathbf{E} is a quantity we can control by continuing the original QR factorization until $\|\mathbf{E}\|$ is smaller than some given threshold. We will next prove two lemmas. The first states that the error in the CUR decomposition is bounded by $\|\mathbf{E}\| + \|\tilde{\mathbf{E}}\|$. The second states that $\|\tilde{\mathbf{E}}\|$ is small whenever $\|\mathbf{E}\|$ is small (and again, $\|\mathbf{E}\|$ we can control).

Lemma 3.1 *Let \mathbf{A} be an $m \times n$ matrix that satisfies the approximate factorizations (3.6) and (3.7). Suppose further that \mathbf{R} is full rank, and that the $k \times k$ matrix \mathbf{U} is defined by (3.5). Then*

$$\|\mathbf{A} - \mathbf{CUR}\| \leq \|\mathbf{E}\| + \|\tilde{\mathbf{E}}\|. \quad (3.8)$$

Proof. Using first (3.5) and then (3.6), we find

$$\mathbf{A} - \mathbf{CUR} = \mathbf{A} - \mathbf{C}\mathbf{V}^* \mathbf{R}^\dagger \mathbf{R} = \mathbf{A} - (\mathbf{A} - \mathbf{E}) \mathbf{R}^\dagger \mathbf{R} = (\mathbf{A} - \mathbf{A} \mathbf{R}^\dagger \mathbf{R}) + \mathbf{E} \mathbf{R}^\dagger \mathbf{R}. \quad (3.9)$$

To bound the term $\mathbf{A} - \mathbf{A} \mathbf{R}^\dagger \mathbf{R}$ we use (3.7) and the fact that $\mathbf{R} \mathbf{R}^\dagger \mathbf{R} = \mathbf{R}$ to achieve

$$\mathbf{A} - \mathbf{A} \mathbf{R}^\dagger \mathbf{R} = \mathbf{A} - (\mathbf{W}\mathbf{R} + \tilde{\mathbf{E}}) \mathbf{R}^\dagger \mathbf{R} = \mathbf{A} - \mathbf{W}\mathbf{R} - \tilde{\mathbf{E}} \mathbf{R}^\dagger \mathbf{R} = \tilde{\mathbf{E}} - \tilde{\mathbf{E}} \mathbf{R}^\dagger \mathbf{R} = \tilde{\mathbf{E}} (\mathbf{I} - \mathbf{R}^\dagger \mathbf{R}). \quad (3.10)$$

Inserting (3.10) into (3.9) and taking the norms of the result, we get

$$\|\mathbf{A} - \mathbf{CUR}\| = \|\tilde{\mathbf{E}} (\mathbf{I} - \mathbf{R}^\dagger \mathbf{R}) + \mathbf{E} \mathbf{R}^\dagger \mathbf{R}\| \leq \|\tilde{\mathbf{E}} (\mathbf{I} - \mathbf{R}^\dagger \mathbf{R})\| + \|\mathbf{E} \mathbf{R}^\dagger \mathbf{R}\| \leq \|\tilde{\mathbf{E}}\| + \|\mathbf{E}\|,$$

where in the last step we used that $\mathbf{R} \mathbf{R}^\dagger$ and $\mathbf{I} - \mathbf{R} \mathbf{R}^\dagger$ are both orthonormal projections. \square

Lemma 3.2 Let \mathbf{A} be an $m \times n$ matrix that admits the factorization (3.6), with error term \mathbf{E} . Suppose further that $I = [I_{\text{skel}}, I_{\text{res}}]$ and \mathbf{T} form the output of the ID of the matrix \mathbf{C} , so that

$$\mathbf{C} = \mathbf{W}\mathbf{C}(I_{\text{skel}}, :), \quad \text{where} \quad \mathbf{W} = \mathbf{P} \begin{bmatrix} \mathbf{I} \\ \mathbf{T}^* \end{bmatrix}, \quad (3.11)$$

and where \mathbf{P} is the permutation matrix for which $\mathbf{P}\mathbf{A}(I, :) = \mathbf{A}$. Now define the matrix \mathbf{R} via

$$\mathbf{R} = \mathbf{A}(I_{\text{skel}}, :). \quad (3.12)$$

Observe that \mathbf{R} consists of the k rows of \mathbf{A} selected in the skeletonization of \mathbf{C} . Finally, set

$$\mathbf{F} = \begin{bmatrix} -\mathbf{T}^* & \mathbf{I} \end{bmatrix} \mathbf{P}^*. \quad (3.13)$$

Then the product $\mathbf{W}\mathbf{R}$ approximates \mathbf{A} , with a residual error

$$\tilde{\mathbf{E}} = \mathbf{A} - \mathbf{W}\mathbf{R} = \mathbf{P} \begin{bmatrix} \mathbf{0} \\ \mathbf{F}\mathbf{E} \end{bmatrix}. \quad (3.14)$$

Proof. From the definitions of \mathbf{W} in (3.11) and \mathbf{R} in (3.12) we find

$$\begin{aligned} \mathbf{A} - \mathbf{W}\mathbf{R} &= \mathbf{P}\mathbf{A}(I, :) - \mathbf{W}\mathbf{R} = \mathbf{P} \begin{bmatrix} \mathbf{A}(I_{\text{skel}}, :) \\ \mathbf{A}(I_{\text{res}}, :) \end{bmatrix} - \mathbf{P} \begin{bmatrix} \mathbf{I} \\ \mathbf{T}^* \end{bmatrix} \mathbf{A}(I_{\text{skel}}, :) \\ &= \mathbf{P} \begin{bmatrix} \mathbf{0} \\ \mathbf{A}(I_{\text{res}}, :) - \mathbf{T}^* \mathbf{A}(I_{\text{skel}}, :) \end{bmatrix} = \mathbf{P} \begin{bmatrix} \mathbf{0} \\ \mathbf{F}\mathbf{A} \end{bmatrix}. \end{aligned} \quad (3.15)$$

To bound the term $\mathbf{F}\mathbf{A}$ in (3.15), we invoke (3.6) to obtain

$$\mathbf{F}\mathbf{A} = \mathbf{F}\mathbf{C}\mathbf{V}^* + \mathbf{F}\mathbf{E} = \{\text{Insert (3.11)}\} = \mathbf{F}\mathbf{W}\mathbf{C}(I_{\text{skel}}, :)\mathbf{V}^* + \mathbf{F}\mathbf{E} = \mathbf{F}\mathbf{E}, \quad (3.16)$$

since $\mathbf{F}\mathbf{W} = \mathbf{0}$ due to (3.11) and (3.13). Finally, insert (3.16) into (3.15) to obtain (3.14). \square

Equation (3.14) allows us to bound the norm of the error $\tilde{\mathbf{E}}$ in (3.7). Simply observe that the definition of \mathbf{F} in (3.13) implies that for any matrix \mathbf{X} we have:

$$\mathbf{F}\mathbf{X} = \begin{bmatrix} -\mathbf{T}^* & \mathbf{I} \end{bmatrix} \mathbf{P}^* \mathbf{X} = \begin{bmatrix} -\mathbf{T}^* & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{X}(I_{\text{skel}}, :) \\ \mathbf{X}(I_{\text{res}}, :) \end{bmatrix} = -\mathbf{T}^* \mathbf{X}(I_{\text{skel}}, :) + \mathbf{X}(I_{\text{res}}, :),$$

so that:

$$\|\mathbf{F}\mathbf{X}\| = \|\mathbf{X}(I_{\text{res}}, :) - \mathbf{T}^* \mathbf{X}(I_{\text{skel}}, :)\| \leq \|\mathbf{X}(I_{\text{res}}, :)\| + \|\mathbf{T}\| \|\mathbf{X}(I_{\text{skel}}, :)\| \leq (1 + \|\mathbf{T}\|) \|\mathbf{X}\|. \quad (3.17)$$

This leads us to the following Corollary to Lemma 3.2:

Corollary 3.3 *Under the same assumptions as in Lemma 3.2, we have*

$$\|\tilde{\mathbf{E}}\| \leq (1 + \|\mathbf{T}\|) \|\mathbf{E}\|. \quad (3.18)$$

Further, assuming additionally that the conditions of Lemma 3.1 are satisfied,

$$\|\mathbf{A} - \mathbf{CUR}\| \leq (2 + \|\mathbf{T}\|) \|\mathbf{E}\|. \quad (3.19)$$

Proof. To show (3.18), we use (3.14) and (3.17):

$$\|\tilde{\mathbf{E}}\| = \left\| \mathbf{P} \begin{bmatrix} \mathbf{0} \\ \mathbf{FE} \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} \mathbf{0} \\ \mathbf{FE} \end{bmatrix} \right\| \leq (1 + \|\mathbf{T}\|) \|\mathbf{E}\|.$$

For (3.19), we use (3.8) and (3.18):

$$\|\mathbf{A} - \mathbf{CUR}\| \leq \|\mathbf{E}\| + \|\tilde{\mathbf{E}}\| \leq (2 + \|\mathbf{T}\|) \|\mathbf{E}\|.$$

□

Now recall that the matrix \mathbf{T} contains the expansion coefficients in the interpolative decomposition of \mathbf{C} . These can be guaranteed [12] to all be bounded by $1 + \nu$ in magnitude for any positive number ν . The cost increases as $\nu \rightarrow 0$, but for, e.g., $\nu = 1$, the cost is very modest. Consequently, we find that for either the spectral or the Frobenius norm, we can easily guarantee $\|\mathbf{T}\| \leq (1 + \nu)\sqrt{k(n - k)}$, with practical norm often far smaller.

4 Efficient deterministic algorithms

Sections 2 and 3 describe how to obtain the ID, two-sided ID, and the CUR decompositions from the output of the column pivoted rank k QR algorithm. In this section, we discuss implementation details, and computational costs for each of the three algorithms.

4.1 The one-sided interpolative decomposition

We start discussing the algorithm for computing an ID decomposition which returns an index vector J and a matrix \mathbf{V} such that $\mathbf{A} \approx \mathbf{A}(:, J(1 : k))\mathbf{V}^*$, and is summarized as Algorithm 1. The only computational complication here is how to evaluate $\mathbf{T} = \mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ on Line 4 of the algorithm. Observe that \mathbf{S}_{11} is upper triangular, so as long as \mathbf{S}_{11} is not too ill-conditioned, a simple backwards solve will compute \mathbf{T} very efficiently. When highly accurate factorizations are sought, however, \mathbf{S}_{11} will typically be sufficiently ill-conditioned that it is better to view \mathbf{T} as the solution to a least squares system:

$$\mathbf{T} = \arg \min_{\mathbf{U}} \|\mathbf{S}_{11}\mathbf{U} - \mathbf{S}_{12}\|. \quad (4.1)$$

This equation can be solved using stabilized methods. For instance, we can form a stabilized pseudo-inverse of \mathbf{S}_{11} by first computing its SVD $\mathbf{S}_{11} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^*$. Dropping all terms involving singular values smaller than some specified threshold, we obtain a truncated decomposition $\mathbf{S}_{11} \approx \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}^*$. Then set $\mathbf{T} = \hat{\mathbf{V}}\hat{\mathbf{D}}^{-1}\hat{\mathbf{U}}^*\mathbf{S}_{12}$. We can also amend (4.1) with a regularization term (i.e. $\lambda\|\mathbf{U}\|$), turning the minimization into a Tikhonov type problem, solvable by an application of the conjugate gradient scheme.

There exists a variation of Algorithm 1 that results in an interpolation matrix \mathbf{V} whose entries are *assured* to be of moderate magnitude. The idea is to replace the column pivoted QR on

Line 1 by the so called “strongly rank revealing QR factorization” algorithm described by Gu and Eisenstat in [9]. They prove that for any $\epsilon > 0$, one can construct matrices \mathbf{S}_{11} and \mathbf{S}_{12} such that the equation $\mathbf{S}_{11}\mathbf{T} = \mathbf{S}_{12}$ has a solution for which $|\mathbf{T}(i, j)| \leq 1 + \epsilon$ for every i and j . The cost of the algorithm increases as $\epsilon \rightarrow 0$, but remains reasonable as long as ϵ is not too close to 0. While such a provably robust algorithm has strong appeal, we have found that in practice, standard column pivoted QR works so well that the additional cost and coding effort required to implement the method of [9] is not worthwhile.

With respect to storage cost, if \mathbf{A} is $m \times n$, to store the ID representation of \mathbf{A} , we require $mk + k(n - k)$ units (since \mathbf{V} contains within it an identity matrix).

4.2 The two-sided interpolative decomposition

Next, we consider the two-sided ID described in Section 2.4, and summarized here as Algorithm 2. The main observation is that \mathbf{C}^* is a matrix of rank at most k . Hence, a rank k QR decomposition would reconstruct it exactly so that the steps in Algorithm 1 produce an exact decomposition. Typically, if the dimensions are not too large, the QR decomposition for step 2 can be performed using standard software packages, such as, e.g., LAPACK. For the two sided ID, the storage requirement for an $m \times n$ matrix is $k(m - k) + k^2 + k(n - k)$, which is the same as for the one sided ID above.

4.3 The CUR decomposition

As demonstrated in Section 3, it is simple to convert Algorithm 2 for computing a two-sided ID into an algorithm for constructing the CUR decomposition. We summarize the procedure as Algorithm 3. The only complication here concerns solving the least squares problem

$$\begin{array}{ccc} \mathbf{U} & \mathbf{R} & = \mathbf{V}^* \\ k \times k & k \times n & k \times n \end{array} \quad (4.2)$$

for \mathbf{U} . In applications like data-mining, where n might be very large, and modest accuracy is sought, one may simply form the normal equations and solve those. For higher accuracy, stabilized techniques based on a truncated QR or SVD decomposition of \mathbf{R} is preferable.

If feasible, one may also consider some adjustment to (4.2) based on the error introduced by the truncated QR factorization. Including the error term from (2.10), we may write:

$$\mathbf{A} = \mathbf{C}\mathbf{V}^* + \mathbf{E} = \mathbf{C}\mathbf{U}\mathbf{R},$$

from which we obtain the modified system:

$$\mathbf{U}\mathbf{R} = \mathbf{V}^* + \mathbf{C}^\dagger \mathbf{E}, \quad (4.3)$$

where \mathbf{E} can be obtained from $\mathbf{E} = \mathbf{A} - \mathbf{Q}\mathbf{R}$ once the partial rank k QR factorization has been performed. One can then obtain matrix \mathbf{U} from a least squares problem corresponding to (4.3). For CUR, the storage requirement for an $m \times n$ matrix is $mk + kn + k^2$, noting that the $k \times k$ matrix \mathbf{U} is not a diagonal.

4.4 Computational and storage costs

All the algorithms discussed in this section have asymptotic cost $O(mnk)$. The dominant part of the computation is almost always the initial rank- k QR factorization. All subsequent computations involve only matrices of sizes $m \times k$ or $k \times n$, and have cost $O((m + n)k^2)$. In terms of memory

storage, when the matrix \mathbf{A} is dense, the two ID decompositions of \mathbf{A} require the least space, followed by the SVD, and then the CUR. However, if \mathbf{A} is a sparse matrix and sparse storage format is used for the factor matrices, the ID and CUR decompositions can be stored more efficiently. Note that the factors \mathbf{C} and \mathbf{R} will be sparse if \mathbf{A} is sparse and so in the sparse case, the CUR storage will in general be minimal amongst all the factorizations.

Algorithm 1: A rank k ID decomposition

Input : $\mathbf{A} \in \mathbb{C}^{m \times n}$ and parameter $k < \min(m, n)$.

Output: A column index set J and a matrix $\mathbf{V} \in \mathbb{C}^{n \times k}$ such that $\mathbf{A} \approx \mathbf{A}(:, J(1:k))\mathbf{V}^*$.

- 1 Perform a rank k column pivoted QR factorization to get $\mathbf{A}\mathbf{P} = \mathbf{Q}_1\mathbf{S}_1$;
 - 2 define the ordered index set J via $\mathbf{I}(:, J) = \mathbf{P}$;
 - 3 partition \mathbf{S}_1 : $\mathbf{S}_{11} = \mathbf{S}_1(:, 1:k)$, $\mathbf{S}_{12} = \mathbf{S}_1(:, k+1:n)$;
 - 4 $\mathbf{V} = \mathbf{P} \begin{bmatrix} \mathbf{I}_k & \mathbf{S}_{11}^{-1}\mathbf{S}_{12} \end{bmatrix}^*$;
-

Algorithm 2: A rank k two sided ID decomposition

Input : $\mathbf{A} \in \mathbb{C}^{m \times n}$ and parameter $k < \min(m, n)$.

Output: A column index set J , a row index set I and a matrices $\mathbf{V} \in \mathbb{C}^{n \times k}$ and $\mathbf{W} \in \mathbb{C}^{m \times k}$ such that $\mathbf{A} \approx \mathbf{W}\mathbf{A}(I(1:k), J(1:k))\mathbf{V}^*$.

- 1 Perform a one sided rank k ID of \mathbf{A} so that $\mathbf{A} \approx \mathbf{C}\mathbf{V}^*$ where $\mathbf{C} = \mathbf{A}(:, J(1:k))$;
 - 2 Perform a full rank ID on \mathbf{C}^* so that $\mathbf{C}^* = \mathbf{C}^*(:, I(1:k))\mathbf{W}^*$;
-

Algorithm 3: A rank k CUR-ID algorithm

Input : $\mathbf{A} \in \mathbb{C}^{m \times n}$ and parameter $k < \min(m, n)$.

Output: Matrices $\mathbf{C} \in \mathbb{C}^{m \times k}$, $\mathbf{R} \in \mathbb{C}^{k \times n}$, and $\mathbf{U} \in \mathbb{C}^{k \times k}$ (such that $\mathbf{A} \approx \mathbf{C}\mathbf{U}\mathbf{R}$).

- 1 Construct a rank k two sided ID of \mathbf{A} so that $\mathbf{A} \approx \mathbf{W}\mathbf{A}(I(1:k), J(1:k))\mathbf{V}^*$;
 - 2 Construct matrices $\mathbf{C} = \mathbf{A}(:, J(1:k))$ and $\mathbf{R} = \mathbf{A}(I(1:k), :)$;
 - 3 Construct matrix \mathbf{U} via $\mathbf{U} = \mathbf{V}^*\mathbf{R}^\dagger$;
-

5 Efficient randomized algorithms

The computational costs of the algorithms described in Section 4 tend to be dominated by the cost of performing the initial k steps of a column pivoted QR-decomposition of \mathbf{A} (at least when the rank k is substantially smaller than the dimensions m and n of the matrix). This initial step can often be accelerated substantially by exploiting techniques based on randomized projections. These ideas were originally proposed in [14, 17], and further developed in [16, 22, 12, 10].

Observe that in order to compute the column ID of a matrix, all we need is to know the linear dependencies among the columns of \mathbf{A} . When the singular values of \mathbf{A} decay reasonably rapidly, we can determine these linear dependencies by processing a matrix \mathbf{Y} of size $\ell \times n$, where ℓ can be much smaller than n . The rows of \mathbf{Y} consist of random linear combinations of the rows of \mathbf{A} , and as long as the number of samples ℓ is a “little bit” larger than the rank k , highly accurate approximations result. In this section, we provide a brief description of how randomization can be used to accelerate the ID and the CUR factorizations, for details and a rigorous analysis of sampling errors, see [10].

The techniques in this section are all designed to compute a one-sided ID. Once this factorization is available, either a two-sided ID, or a CUR decomposition can easily be obtained using the techniques outlined in Section 3.

5.1 A basic randomized algorithm

Suppose that we are given an $m \times n$ matrix \mathbf{A} and seek to compute a column ID, a two-sided ID, or a CUR decomposition. As we saw in Section 4, we can perform this task as long as we can identify an index vector $J = [J_{\text{skel}}, J_{\text{res}}]$ and a basis matrix $\mathbf{V} \in \mathbb{C}^{n \times k}$ such that

$$\begin{array}{ccccc} \mathbf{A} & = & \mathbf{A}(:, J_{\text{skel}}) & \mathbf{V}^* & + & \mathbf{E} \\ m \times n & & m \times k & k \times n & & m \times n \end{array}$$

where \mathbf{E} is small. In Section 4, we found J and \mathbf{V} by performing a column pivoted QR factorization of \mathbf{A} . In order to do this via randomized sampling, we first fix a small over-sampling parameter p , say $p = 10$ for now (see Remark 5.1 for details). Then draw a $(k + p) \times m$ random matrix $\mathbf{\Omega}$ whose entries are i.i.d. standardized Gaussian random variables, and form the *sampling matrix*

$$\begin{array}{ccccc} \mathbf{Y} & = & \mathbf{\Omega} & \mathbf{A}. \\ (k + p) \times n & & (k + p) \times m & m \times n \end{array} \quad (5.1)$$

One can prove that with high probability, the space spanned by the rows of \mathbf{Y} contains the dominant k right singular vectors of \mathbf{A} to high accuracy. This is precisely the property we need in order to find both the vector J and the basis matrix \mathbf{V} . All we need to do is to perform k steps of a column pivoted QR factorization of the sample matrix to form a partial QR factorization

$$\begin{array}{ccccc} \mathbf{Y}(:, J) & \approx & \mathbf{Q} & \mathbf{S}. \\ (k + p) \times n & & (k + p) \times k & k \times n \end{array}$$

Then compute the matrix of expansion coefficients via $\mathbf{T} = \mathbf{S}(1 : k, 1 : k)^{-1} \mathbf{S}(1 : k, (k + 1) : n)$, or a stabilized version, as described in Section 4.1. The matrix \mathbf{V} is formed from \mathbf{T} as before, resulting in Algorithm 4. The asymptotic cost of Algorithm 4 is $O(mnk)$, just like the algorithms described in Section 4. However, substantial practical gain is achieved due to the fact that the matrix-matrix multiplication is much faster than a column-pivoted QR factorization. This effect gets particularly pronounced when a matrix is very large and is stored either out-of-core, or on a distributed memory machine.

Remark 5.1 *Careful mathematical analysis is available to guide the choice of the over-sampling parameter p [10]. However, in practical applications, choosing $p = 10$ is almost always more than sufficient. If a very close to optimal skeleton is desired, one could increase the parameter up to $p = 2k$, but this is generally far higher than needed.*

Algorithm 4: A randomized rank k ID Decomposition

Input : $\mathbf{A} \in \mathbb{C}^{m \times n}$, a rank parameter $k < \min(m, n)$, and an oversampling parameter p .

Output: A column index set J and a matrix $\mathbf{V} \in \mathbb{C}^{n \times k}$ (such that $\mathbf{A} \approx \mathbf{A}(:, J(1:k))\mathbf{V}^*$).

- 1 Construct a random matrix $\mathbf{\Omega} \in \mathbb{R}^{(k+p) \times m}$ with i.i.d. Gaussian entries;
 - 2 Construct the sample matrix $\mathbf{Y} = \mathbf{\Omega}\mathbf{A}$;
 - 3 Perform full pivoted QR factorization on \mathbf{Y} to get: $\mathbf{Y}\mathbf{P} = \mathbf{Q}\mathbf{S}$;
 - 4 Remove p columns of \mathbf{Q} and p rows of \mathbf{S} to construct \mathbf{Q}_1 and \mathbf{S}_1 ;
 - 5 Define the ordered index set J via $\mathbf{I}(:, J) = \mathbf{P}$;
 - 6 Partition \mathbf{S}_1 : $\mathbf{S}_{11} = \mathbf{S}_1(:, 1:k)$, $\mathbf{S}_{12} = \mathbf{S}_1(:, k+1:n)$;
 - 7 $\mathbf{V} = \mathbf{P} \begin{bmatrix} \mathbf{I}_k & \mathbf{S}_{11}^{-1}\mathbf{S}_{12} \end{bmatrix}^*$;
-

5.2 An accelerated randomized scheme

At this point, all algorithms described have asymptotic complexity $O(mnk)$. Using the randomized projection techniques, we can reduce this to $O(mn \log(k) + k^2(m+n))$. The idea is to replace the Gaussian randomized matrix $\mathbf{\Omega}$ we used in Section 5.1 by a random matrix that has enough structure that the matrix-matrix multiplication (5.1) can be executed in $O(mn \log(k))$ operations. For instance, one can use a *subsampled random Fourier transform (SRFT)*, which takes the form

$$\begin{matrix} \mathbf{\Omega} & = & \sqrt{\frac{m}{\ell}} & \mathbf{R} & \mathbf{F} & \mathbf{D} \\ \ell \times m & & & \ell \times m & m \times m & m \times m \end{matrix} \quad (5.2)$$

where \mathbf{D} is an $m \times m$ diagonal matrix whose entries are independent random variables uniformly distributed on the complex unit circle; where \mathbf{F} is the $m \times m$ unitary discrete Fourier transform, whose entries take the values $\mathbf{F}(p, q) = m^{-1/2} e^{-2\pi i(p-1)(q-1)/m}$ for $p, q = 1, 2, \dots, m$; and where \mathbf{R} is an $\ell \times m$ matrix that samples ℓ coordinates from m uniformly at random (i.e., its ℓ rows are drawn randomly without replacement from the rows of the $m \times m$ identity matrix).

When using an SRFT, a larger number of samples is sometimes required to attain similar accuracy. In practice $\ell = 2k$ is almost always sufficient, see [10, Sec. 4.6].

Replacing lines 1 and 2 in Algorithm 4 by the SRFT (5.2) reduces the cost of executing these lines to $O(mn \log(k))$, assuming $\ell = 2k$. The remaining operations have complexity $O(k^2(m+n))$.

5.3 An accuracy enhanced scheme

The randomized sampling schemes described in Sections 5.1 and 5.2 are roughly speaking as accurate as the techniques based on a column pivoted QR factorization described in Section 4 as long as the singular values of \mathbf{A} exhibit reasonable decay. For the case where the singular values decay slowly (as often happens in data mining and analysis of statistical data, for instance), the accuracy deteriorates. However, high accuracy can easily be restored by slightly modifying the construction of the sampling matrix \mathbf{Y} . The idea of the power sampling scheme is roughly to choose a small integer q (say $q = 1$ or $q = 2$), and then form the sampling matrix via

$$\mathbf{Y} = \mathbf{\Omega} \mathbf{A} (\mathbf{A}^* \mathbf{A})^q. \quad (5.3)$$

The point here is that if \mathbf{A} has singular values $\{\sigma_j\}_{j=1}^{\min(m,n)}$, then the singular values of $\mathbf{A}(\mathbf{A}\mathbf{A}^*)^q$ are $\{\sigma_j^{2q+1}\}_{j=1}^{\min(m,n)}$, which means that the larger singular values are weighted much more heavily versus the lower ones.

For computational efficiency, note that the evaluation of (5.3) should be done by successive multiplications of \mathbf{A} and \mathbf{A}^* , so that line 2 in Algorithm 4 gets replaced by:

```

(2a)   $\mathbf{Y} = \Omega\mathbf{A}$ 
(2b)  for  $i = 1 : q$ 
(2c)       $\mathbf{Y} \leftarrow \mathbf{Y}\mathbf{A}^*$ 
(2d)       $\mathbf{Y} \leftarrow \mathbf{Y}\mathbf{A}$ 
(2e)  end

```

In cases where very high computational precision is required (higher than $\epsilon_{\text{mach}}^{1/(2q+1)}$, where ϵ_{mach} is the machine precision), one typically needs to orthonormalize the sampling matrix in between multiplications, resulting in:

```

(2a)   $\mathbf{Y} = \Omega\mathbf{A}$ 
(2b)  for  $i = 1 : q$ 
(2c)       $\mathbf{Y} \leftarrow \text{orth}(\mathbf{Y})\mathbf{A}^*$ 
(2d)       $\mathbf{Y} \leftarrow \text{orth}(\mathbf{Y})\mathbf{A}$ 
(2e)  end

```

where `orth` refers to orthonormalization of the *rows*, without pivoting. In other words, if $\mathbf{Q} = \text{orth}(\mathbf{Y})$, then \mathbf{Q} is a matrix whose rows form an orthonormal basis for the rows of \mathbf{Y} .

The asymptotic cost of the algorithm described in this section is $O((2q+1)mnk + k^2(m+n))$.

Remark 5.2 *It is to the best of our knowledge not possible to accelerate the accuracy enhanced technique described in this section to $O(mn \log(k))$ complexity.*

6 Numerics

In this section, we present numerical comparisons between the proposed CUR-ID algorithm, and previously proposed schemes, specifically those implemented in the rCUR package [2] and the algorithm from [18].

We first compare the proposed method for computing the CUR decomposition (Algorithm 3) against four existing CUR algorithms, one based on the newly proposed DEIM-CUR method as described in [18] and three algorithms as implemented in the rCUR package. We first use the full SVD with each algorithm:

- CUR-H The full SVD is computed and provided to rCUR, and then the “highest ranks” option is chosen. This generally offers good performance and reasonable runtime in our experiments.
- CUR-1 The full SVD is computed and provided to rCUR, and then the “orthogonal top scores” option is chosen. This is an expensive scheme that we believe gives the best performance in rCUR for many matrix types. However, when the decay of singular values of the input matrix is very rapid or abrupt (as in the example in Figure 3 below), the scheme performs poorly. This scheme is also considerably slower than the others.
- CUR-2 The full SVD is computed and provided to DEIM-CUR. This generally offers good performance and reasonable runtime in our experiments.
- CUR-3 The full SVD is computed and provided to rCUR, and then the “top scores” option is chosen. This procedure reflects a common way that “leverage scores” are used. It has slightly worse performance than CUR-1 and CUR-H in our experiments but better runtime.

Our first set of test matrices (“Set 1”) involves matrices \mathbf{A} of size 1000×3000 , of the form $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^*$ where \mathbf{U} and \mathbf{V} are random orthonormal matrices, and \mathbf{D} is a diagonal matrix with entries that are logspaced between 1 and 10^b , for $b = -2, -4, -6$. The second set (“Set 2”) are simply the transposes of the matrices in Set 1 (so these are matrices of size 3000×1000). Figure 1 plots the median relative errors in the spectral norm between the matrix \mathbf{A} and the corresponding factorization (with the error defined as $E = \frac{\|\hat{\mathbf{A}}_k - \mathbf{A}\|}{\|\mathbf{A}\|}$ where $\hat{\mathbf{A}}_k = \mathbf{CUR}$ is the corresponding approximation of given rank). We plot median quantities collected over 5 trials. In addition to the four CUR algorithms, we also include plots for the two sided ID and the SVD of given rank (providing the optimal approximation). Based on the plots, we make three conjectures for matrices conditioned similar to those used in this example (note that CUR-1 performs poorly in some of our other experiments):

- The accuracies of CUR-ID, CUR-1, and CUR-2, are all very similar. CUR-H offers slightly worse approximations.
- The accuracy of CUR-3 is worse than all other algorithms tested.
- The two-sided ID is in every case more accurate than the CUR-factorizations.

Next, in Figure 2, we compare the performance and runtimes of CUR-H, CUR-1, and CUR-2 algorithms with the randomized SVD [10] (which gives close results to the true SVD of given rank but at substantially less cost) and the CUR-ID algorithm using the randomized ID, as described in this text (using $q = 2$ in the power sampling scheme (5.3)). This comparison allows us to test algorithms which can be used in practice on large matrices, since they involve randomization. We again use random matrices constructed as above whose singular values are logspaced, ranging from 10^0 to 10^{-3} , but of larger size: 2000×4000 . We notice that the performance with all schemes is similar but the runtime with the randomized CUR-ID algorithm is substantially lower than with the other schemes. The runtime of CUR-1 is substantially greater than of the other schemes. The plotted quantities are again medians over 5 trials.

In Figure 3, we repeat the experiment using the randomized SVD with the two matrices \mathbf{A}_1 and \mathbf{A}_2 defined in the preprint [18]. The matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{300,000 \times 300}$ are constructed as follows:

$$\mathbf{A}_1 = \sum_{j=1}^{10} \frac{2}{j} x_j y_j^T + \sum_{j=11}^{300} \frac{1}{j} x_j y_j^T \quad \text{and} \quad \mathbf{A}_2 = \sum_{j=1}^{10} \frac{1000}{j} x_j y_j^T + \sum_{j=11}^{300} \frac{1}{j} x_j y_j^T,$$

where x and y are sparse vectors with random non-negative entries. One problem with using traditional CUR algorithms for these matrices stems from the fact that the singular values of \mathbf{A}_1 and \mathbf{A}_2 decay rapidly. Due to this, the performance of CUR-1 (and of CUR-3, which we do not show) for these examples is poor. It appears that this is because for these schemes, the rapid decay of the singular values of the input matrix translates into the inversion of ill-conditioned matrices, which adversely effects performance. On the other hand, CUR-ID and CUR-2 offer similar performance, close to the approximate SVD results. In Figure 3, we show the medians of relative errors versus k over 5 trials.

In Figure 4, we show comparison between absolute errors given by our non-randomized and randomized CUR-ID algorithms and the truncated SVD and QR factorizations in terms of the square of the Frobenius norm and the spectral norm. We use 600×600 test matrices, with varying singular value decay, as before. In particular, we check here if the optimistic bound:

$$\|\mathbf{A} - \mathbf{CUR}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2 \quad \text{with} \quad \mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^* \quad (6.1)$$

from [3] holds with $1 < \epsilon < 2$ for the non-randomized CUR-ID scheme. For $\epsilon \approx 2$ and $k \ll \min(m, n)$ the bound sometimes holds, but it does not hold for all k . Despite this, we may also observe from the bottom row of Figure 4 that for matrices with rapid singular value decay, the CUR-ID error in the spectral norm is sometimes lower even than that of the truncated QR.

In Figure 5, we have an image compression experiment, using CUR-ID and CUR-1, CUR-2, and CUR-H with the full SVD. We take two black and white images (of size 350×507 and 350×526) and transform the matrix using four levels of the 2D CDF 97 wavelet transform. We then threshold the result, leaving a sparse $m \times n$ matrix \mathbf{M} with about 30% nonzeros (with same dimensions as the original image). Then we go on to construct a low rank CUR approximation of this wavelet thresholded matrix (with $k = \min(m, n)/15$) to further compress the image data. Storing the three matrices \mathbf{C} , \mathbf{U} , and \mathbf{R} corresponds to storing about 8 time less nonzeros vs storing \mathbf{M} . To reconstruct the image from this compressed form, we perform the inverse CDF 97 WT transform on the matrix product \mathbf{CUR} , which approximates the wavelet thresholded matrix. From the plots, we see that CUR-ID produces a \mathbf{U} which has less rapid singular value decay than the \mathbf{U} matrix obtained with the CUR-1 and CUR-H algorithms. In particular, the reconstructions obtained with CUR-1 are very poor and the \mathbf{U} obtained from this scheme has rapidly decaying singular values, comparable to those of \mathbf{M} .

Thus, in each case, we observe comparable or even better performance with CUR-ID than with existing CUR algorithms. For large matrices, existing CUR algorithms that rely on the singular vectors must be used in conjunction with an accelerated scheme for computing approximate singular vectors, such as, e.g., the randomized method of [10], or to use CUR-ID with the randomized ID. We find that for random matrices the performance is similar, but CUR-ID is easier to implement and is generally more efficient. Also, as in the case of the imaging example we present, existing CUR algorithms suffer from a badly conditioned \mathbf{U} matrix when the original matrix is not well conditioned. The \mathbf{U} matrix returned by the CUR-ID algorithm tends to be better conditioned.

Finally, we again remark that optimized codes for the algorithms we propose are available as part of the RSVDPACK software package [20].

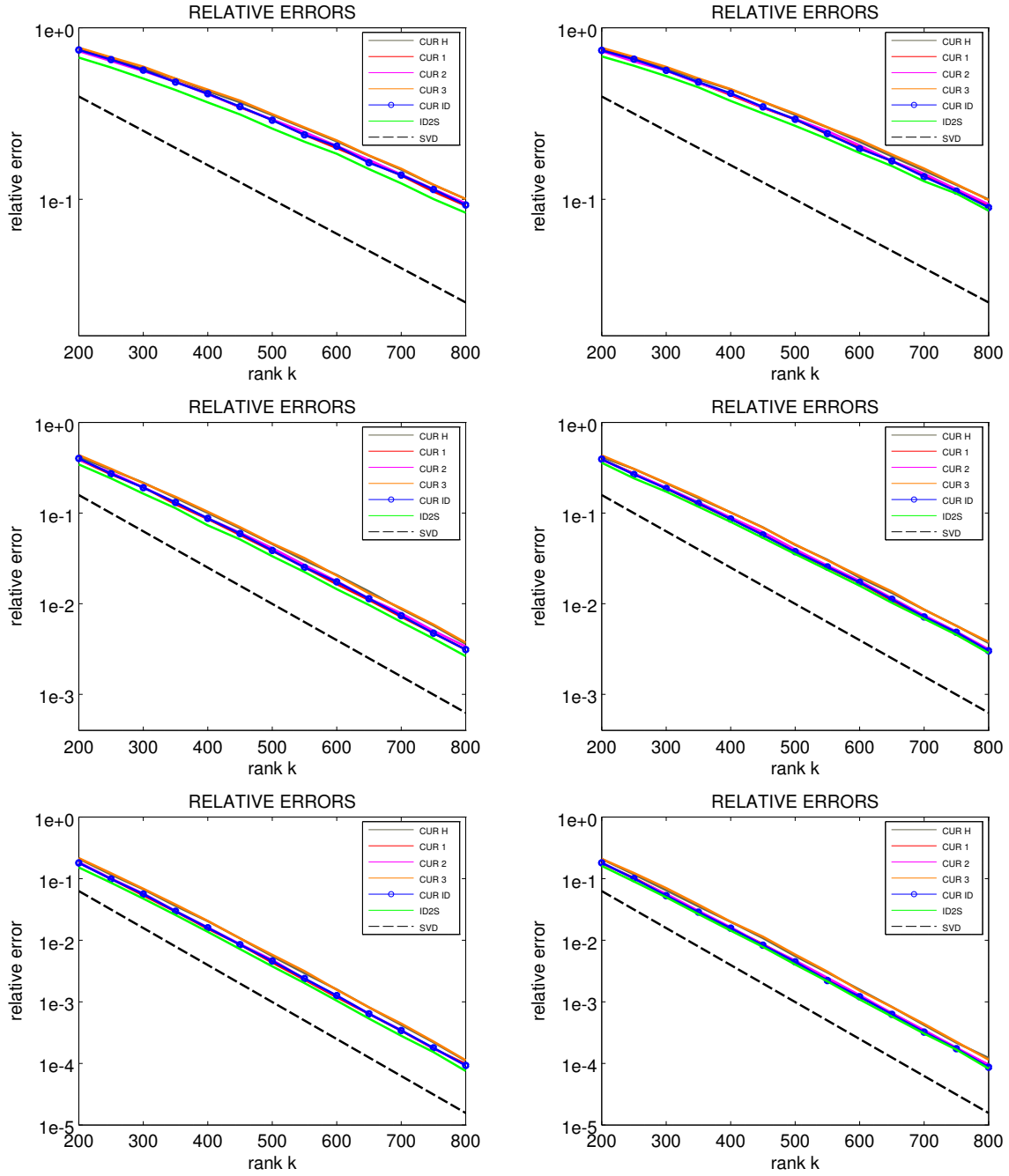


Figure 1: Relative errors for differently conditioned matrices approximated with various algorithms. Left: fat matrices (1000×3000), right: thin matrices (3000×1000). Top to bottom: faster drop off of logspaced singular values.

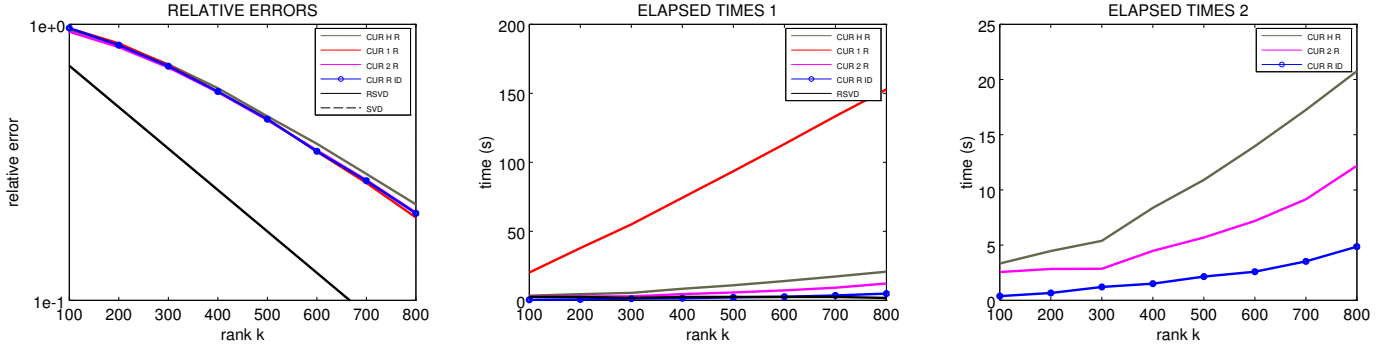


Figure 2: Relative errors and elapsed times for CUR-H,CUR-1,CUR-2 with randomized SVD and CUR-ID with the randomized ID using larger matrices of size 2000×4000 . First time plot shows runtimes for all algorithms. Second time plot shows runtimes of CUR-H, CUR-2, and CUR-ID.

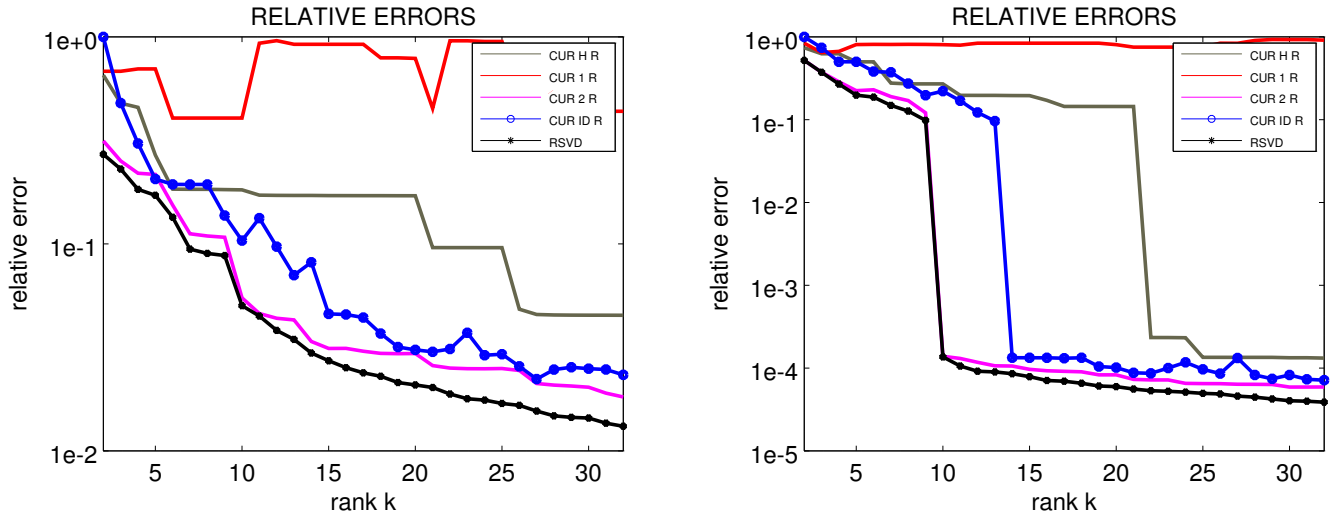


Figure 3: Relative errors versus k for matrices \mathbf{A}_1 (left) and \mathbf{A}_2 (right) from [18] approximated using CUR-H,CUR-1,CUR-2 with randomized SVD and CUR-ID with the randomized ID.

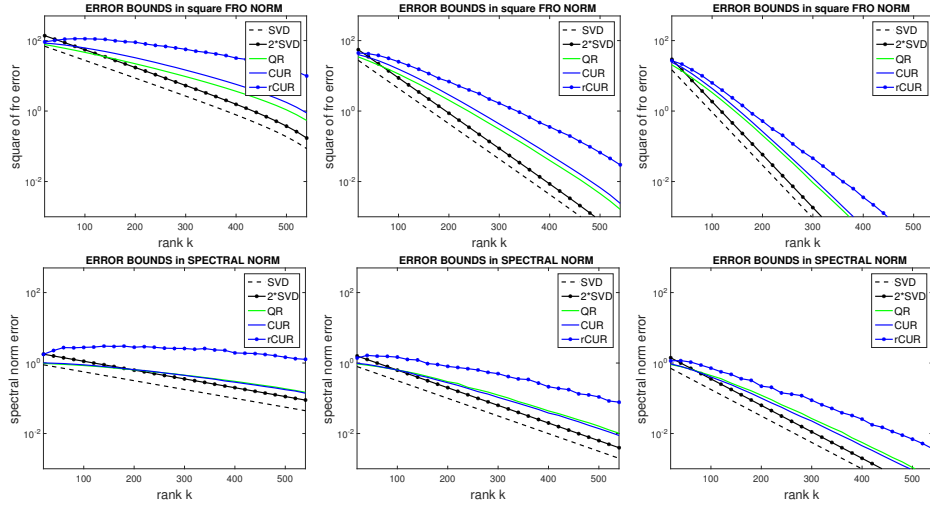


Figure 4: Comparison of absolute error bounds for rank k CUR-ID and CUR-ID with randomization in comparison to truncated rank k SVD and truncated QR decompositions in terms of square Frobenius norm (top) and spectral norm (bottom) for matrices with singular values distributed on a logarithmic scale between 1 and 10^{-b} with $b = 1.5, 3, 4.5$.

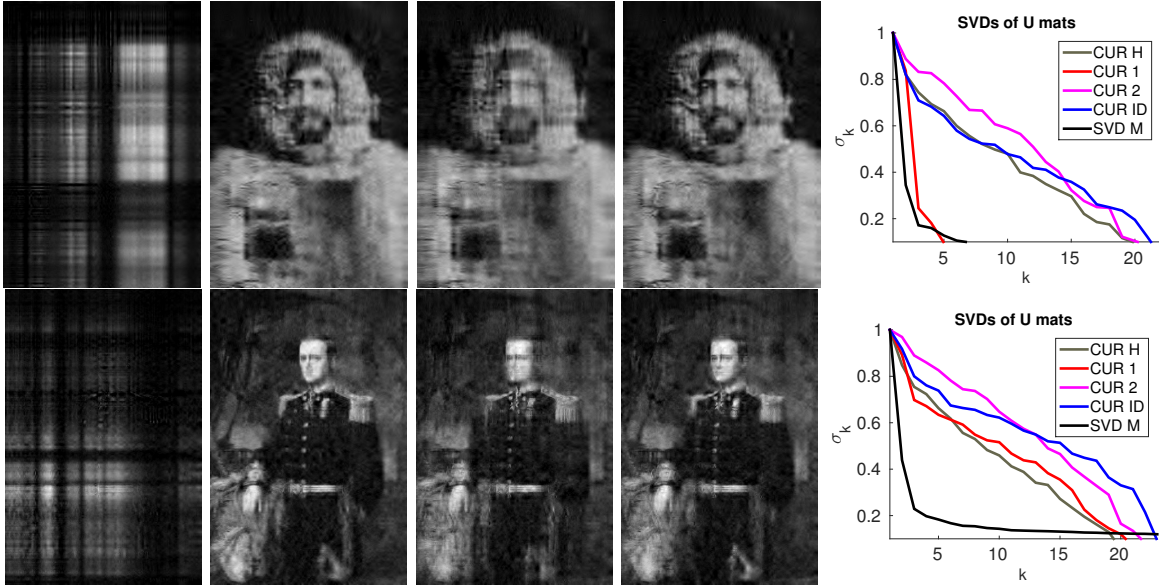


Figure 5: Reconstructed images with CUR compression of the wavelet transformed image. Images resulting from applying Inverse Wavelet transform to matrix product **CUR** obtained with CUR-1 in column 1, CUR-H in column 2, CUR-2 in column 3, and with CUR-ID in column 4. Column 5 plots: singular value distributions of output **U** matrices with the different algorithms compared.

7 Conclusions

This paper presents efficient algorithms for computing ID and CUR decompositions. The algorithms are obtained by very minor modifications to the classical pivoted QR factorization. As a result, the new CUR-ID algorithm provides a direct and efficient way to compute the CUR factorization using standard library functions, as provided in, e.g., BLAS and LAPACK.

Numerical tests illustrate that the new algorithm CUR-ID leads to substantially smaller approximation errors than methods that select the rows and columns based on leverage scores only. The accuracy of the new scheme is comparable to existing schemes that rely on additional information in the leading singular vectors, such as, e.g., the DEIM-CUR [18] of Sorensen and Embree, or the “orthogonal top scores” technique in the package rCUR. However, we argue that CUR-ID has a distinct advantage in that it can easily be coded up using existing software packages, and our numerical experiments indicate an advantage in terms of computational speed.

The paper also demonstrates that the two-sided ID is superior to the CUR-decomposition in terms of both approximation errors and conditioning of the factorization. The ID offers the same benefits as the CUR decomposition in terms of data interpretation. However, for very large and very sparse matrices, the CUR decomposition can be more memory efficient than the ID.

Finally, the paper demonstrates that randomization can be used to very substantially accelerate algorithms for computing the ID and CUR-decompositions, including techniques based on leverage scores, the DEIM-CUR algorithm, and the newly proposed CUR-ID. Moreover, randomization can be used to reduce the overall complexity of the CUR-ID-algorithm from $O(mnk)$ to $O(k^2m + k^2n + mn \log k)$.

Acknowledgement The research reported was supported by the Defense Advanced Projects Research Agency under the contract N66001-13-1-4050, and by the National Science Foundation under contracts 1320652 and 0748488.

References

- [1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. LAPACK Users’ Guide. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999. 2
- [2] András Bodor, István Csabai, Michael Mahoney, and Norbert Solymosi. rCUR: an R package for CUR matrix decomposition. BMC Bioinformatics, 13(1), 2012. 6, 15
- [3] Christos Boutsidis and David P Woodruff. Optimal cur matrix decompositions. In Proceedings of the 46th Annual ACM Symposium on Theory of Computing, pages 353–362. ACM, 2014. 16
- [4] Tony F. Chan. Rank revealing QR factorizations. Linear Algebra Appl., 88/89:67–82, 1987. 2, 3
- [5] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. SIAM J. Matrix Anal. Appl., 30(2):844–881, 2008. 2, 6
- [6] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. Psychometrika, 1(3):211–218, 1936. 3
- [7] Gene H. Golub and Charles F. Van Loan. Matrix computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013. 2, 3

- [8] Sergei A Goreinov, Eugene E Tyrtyshnikov, and Nikolai L Zamarashkin. A theory of pseudoskeleton approximations. Linear Algebra and its Applications, 261(1):1–21, 1997. [1](#)
- [9] Ming Gu and Stanley C. Eisenstat. Efficient algorithms for computing a strong rank-revealing qr factorization. SIAM J. Sci. Comput., 17(4):848–869, July 1996. [4](#), [10](#)
- [10] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev., 53(2):217–288, 2011. [2](#), [6](#), [11](#), [12](#), [13](#), [15](#), [16](#)
- [11] David C. Hoaglin and Roy E. Welsch. The Hat matrix in regression and ANOVA. The American Statistician, 32(1):17–22, 1978. [2](#)
- [12] Edo Liberty, Franco Woolfe, Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. Randomized algorithms for the low-rank approximation of matrices. Proceedings of the National Academy of Sciences, 104(51):20167–20172, 2007. [9](#), [11](#)
- [13] Michael W. Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. Proc. Natl. Acad. Sci. USA, 106(3):697–702, 2009. With supplementary material available online. [1](#), [2](#)
- [14] Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. A randomized algorithm for the approximation of matrices. Technical Report Yale CS research report YALEU/DCS/RR-1361, Yale University, Computer Science Department, 2006. [11](#)
- [15] Nikola Mitrovic, Muhammad Tayyab Asif, Umer Rasheed, Justin Dauwels, and Patrick Jaillet. CUR decomposition for compression and compressed sensing of large-scale traffic data. Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems, 2013. [1](#), [6](#)
- [16] Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A randomized algorithm for principal component analysis. SIAM Journal on Matrix Analysis and Applications, 31(3):1100–1124, 2009. [11](#)
- [17] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06), pages 143–152. IEEE, 2006. [11](#)
- [18] D. C. Sorensen and M. Embree. A DEIM Induced CUR Factorization. ArXiv e-prints, July 2014. [6](#), [15](#), [16](#), [18](#), [20](#)
- [19] Eugene Tyrtyshnikov. Incomplete cross approximation in the mosaic-skeleton method. Computing, 64(4):367–380, 2000. [1](#)
- [20] Sergey Voronin and Per-Gunnar Martinsson. Rsvdpack: Subroutines for computing partial singular value decompositions via randomized sampling on single core, multi core, and gpu architectures. arXiv preprint arXiv:1502.05366, 2015. [2](#), [16](#)
- [21] Shusen Wang and Zhihua Zhang. Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling. J. Mach. Learn. Res., 14:2729–2769, 2013. [1](#), [2](#)
- [22] Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert. A fast randomized algorithm for the approximation of matrices. Applied and Computational Harmonic Analysis, 25(3):335–366, 2008. [11](#)