IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 February 2023, accepted 26 February 2023, date of publication 2 March 2023, date of current version 8 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3251744

RESEARCH ARTICLE

Energy-Based Adaptive CUR Matrix Decomposition

LIWEN XU, XUEJIAO ZHAO, AND YONGXIA ZHANG
College of Science, North China University of Technology, Beijing 100144, China
Corresponding authors: Liwen Xu (xulw163@163.com) and Yongxia Zhang (zhangyx@ncut.edu.cn)

This work was supported in part by the National Social Science Foundation of China under Grant 20BTJ046.

ABSTRACT CUR decompositions are interpretable data analysis tools that express a data matrix in terms of a small number of actual columns and/or actual rows of the data matrix. One bottleneck of existing relative-error CUR algorithms lies on high computational complexity for computing important sampling probabilities. In this paper, we provide a simple yet effective framework that considers energy-based sampling algorithm. On one hand, we provide an intuitive and fast relative-error sampling algorithm for column selection problem. On the other hand, by combining the relative-error sampling algorithm with adaptive sampling algorithm we provide a novel CUR matrix approximation algorithms which is referred to as energy-based adaptive sampling algorithm. The sampling algorithm is the first adaptive relative-error CUR decomposition in the coherent sense. Specially, in each stage of our algorithm, we sample columns or rows from data matrix using sampling probabilities that are directly proportional to Euclidean norms of the columns or rows of the original data and residual matrix, respectively. Our empirical results exactly indicate that the new adaptive sampling algorithm typically achieves a good balance between computational complexity and approximate accuracy.

INDEX TERMS Randomized algorithm, CUR decomposition, subsampling, energy sampling.

I. INTRODUCTION

In modern data analysis applications, dealing with and approximating large data matrices are common. Examples of such data matrices include web images, web documents and gene responses. The truncated singular value decomposition (SVD) plays a fundamental role that provides the best low-rank approximations of the original data matrices, with respect to any unitarily invariant norm. In view of the difficulties of interpretation, there exists a great concern to find and study matrix approximations that are explicitly expressed in terms of a small number of actual rows or columns of the original data matrices.

In this context, CUR decompositions are particularly interesting, as they directly sample actual rows or columns of matrices to form the random sample and preserve the interpretability of the original data [1]. Therefore, the CUR matrix decompositions have been extensively discussed in the theoretical computer science, the machine learning, and the numerical linear algebra community [2], [3], [4], [5], [6], [7]. The more applications of CUR decomposition include modeling large-scale traffic networks, large-scale retrieval, and compression sensing, and so on. Moreover, the related studies have been extended to tensor CUR which can be thought of as multi-dimensional generalizations of CUR decompositions [8], [9], [10]. The computational complexity of sampling probabilities and approximate accuracy are two fundamental problems.

Given a data matrix $A \in \mathbb{R}^{m\times n}$, employing simple nonuniform sampling probabilities that depend on the Euclidean norms of rows and/or columns of A itself, Drineas et al. [11], [12], [13] proposed a CUR algorithm with additive-error bound. An additive-error bound may be expressed as $\|A - CUR\|_F \leq \|A - A_k\|_F + \epsilon\|A\|_F$, where $0 < \epsilon < 1$, A_k denotes the best rank-k approximation to A, and $\|A\|_F$ denotes the Frobenius norm of A (cf. Section II). Mitrovic et al. [14] referred the algorithm as the energy sampling algorithm since the square of Euclidean norm of a signal vector

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Huang.

represents the total energy contained in the signal. Although these methods also can be used to approximate matrix multiplication or compute low-rank approximations, none of them addresses relative-error approximations aspects of this approach. A relative-error bound may be expressed as $\|A - CUR\|_F \leq (1+\epsilon)\|A - A_k\|_F = \|A - A_k\|_F + \epsilon\|A - A_k\|_F$. Noting form $\|A - A_k\|_F \leq \|A\|_F$ that in general the scale of the additive-error is somewhat larger, Drineas et al. [15] devised a randomized CUR algorithm called the subspace sampling algorithm which has relative-error bound by using the statistical leverage scores, so the computational cost of this algorithm is at least equal to the cost of the truncated SVD of A.

The CUR matrix decomposition problem has a close connection with the column selection problem. As pointed out in Wang and Zhang [16], CUR is a harder problem than column selection because ''one can get good columns or rows separately'' does not mean that ''one can get good columns and rows together''. Motivated by this problem, Wang and Zhang [16] developed an adaptive sampling algorithm for improving existing CUR. Combining the near-optimal column selection algorithm of Boutsidis et al. [17] and the adaptive sampling algorithm for solving the CUR problem, Wang and Zhang [18] provided an algorithm with a much tighter theoretical bound than existing algorithms. However, under common scenarios, the algorithm becomes less efficient when the column number c and row number r are large [18]. On the other hand, U matrix in CUR decomposition can be computed in different ways after constructing C and R. Wang et al. [19] provided a technique for computing the U matrix more efficiently in CUR matrix decomposition. Boutsidis and Woodruff [20] developed relative-error CUR algorithms selecting the optimal number of columns and rows, together with a matrix U with optimal rank.

The above studies lead us to the following questions: (1) Can simple energy sampling algorithm generate relative-error approximation? (2) Can we achieve a good balance between the computational complexity of sampling probabilities and approximate accuracy?

In this paper we focus on how to efficiently construct high-quality C and R matrices for CUR. We develop a simple yet effective sampling algorithm which is the first adaptive sampling in the coherent sense. Specifically, in each stage of the energy adaptive sampling algorithm, we sample columns or rows from data matrix with the same type of probabilities that are directly proportional to Euclidean norms of the columns or rows of the original data and residual matrix, respectively. Meanwhile, we provide a new relative-error bound theory for the adaptive CUR decomposition. Based on the theoretical results, our main empirical contribution is to provide an evaluation of the running time and approximation errors of our adaptive sampling algorithm on several real data sets. The empirical results indicate that the simple energy adaptive sampling typically achieves low computational cost with comparable accuracy.

The rest of this paper is organized as follows. Section III introduces several existing column selection and CUR algorithms. Section IV describes and analyzes our novel CUR algorithm. Section V empirically compares our proposed algorithm with two widely known algorithms.

II. NOTATION

For a vector $a \in \mathbb{R}^m$, let $\|a\|_2$ denote the Euclidean norm of a. For a matrix $A \in \mathbb{R}^{m\times n}$, let $A^{(j)}, j = 1, \ldots, n$, denote the jth column of A as a column vector, $A_{(i)}, i = 1, \ldots, m$, denote the ith row of A as a row vector, and A_{ij} denote the (i, j)th element of A. For any orthogonal matrix $U \in \mathbb{R}^{m\times l}$, let $U^{\perp} \in \mathbb{R}^{m\times(m-l)}$ denote an orthogonal matrix whose columns are an orthonormal basis spanning the subspace of $\mathbb{R}^m$ that is orthogonal to the column space of U. Let the rank of A be $\rho \leq \min\{m, n\}$. The SVD of A is denoted by

$$\begin{aligned} A &= \sum_{i=1}^{\rho} \sigma_i u_i v_i^T = U_A \Sigma_A V_A^T \\ &= U_{A_k} \Sigma_{A_k} V_{A_k}^T + U_{A_k}^{\perp} \Sigma_{A_k\perp} V_{A_k}^{\perp T}, \end{aligned} \tag{1}$$

where $\sigma_i = \sigma_i(A)$ denotes the ith singular value, $i = 1, \ldots, \rho$, $U_{A_k} \in \mathbb{R}^{m\times k}$, $\Sigma_{A_k} \in \mathbb{R}^{k\times k}$, $V_{A_k} \in \mathbb{R}^{n\times k}$ correspond to the top k singular values, $\Sigma_{A_k\perp} \in \mathbb{R}^{(\rho-k)\times(\rho-k)}$ is the diagonal matrix containing the bottom $\rho - k$ nonzero singular values of A. $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T = U_{A_k} \Sigma_k V_{A_k}^T$ denotes the best rank-k approximation to A, and $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ denote the maximum and minimum singular values of A, respectively. The condition number of A is $\kappa(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$. The Moore-Penrose generalized inverse of A may be expressed as $A^+ = V_A \Sigma_A^{-1} U_A^T$. The Frobenius norm of A is defined by $\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2\right)^{1/2}$.

Based on SVD, the statistical leverage scores of the columns of A relative to the best rank-k approximation to A are defined as

$$lev_j = \|(V_{A_k}^T)^{(j)}\|_2^2, \quad j = 1, \ldots, n. \tag{2}$$

The corresponding subspace sampling probabilities are defined as

$$p_j = lev_j / \|V_{A_k}\|_F^2 = lev_j / k, \quad j = 1, \ldots, n. \tag{3}$$

The simple energy-based sampling probabilities satisfy

$$p_j = \|A^{(j)}\|_2^2 / \|A\|_F^2, \quad j = 1, \ldots, n. \tag{4}$$

III. RELATED WORK

In Section III-A, we describe the subspace sampling algorithms for column selection and CUR of Drineas et al. [15] which will be used as a benchmark for comparison. In Section III-B, we present an adaptive sampling algorithm for CUR and its relative-error bound established by Wang and Zhang [18]. This algorithm is a building block of some more powerful algorithms, and our novel CUR algorithm also relies on this algorithm.

A. THE SUBSPACE SAMPLING ALGORITHM

1) COLUMN SELECTION VIA SUBSPACE SAMPLING ALGORITHM

Given a matrix $A \in R^{m\times n}$, column selection is a problem of selecting c columns of A to construct $C \in R^{m\times c}$ to minimize $\|A-CC^{+}A\|_F$. There are a variety of column selection algorithms achieving relative-error bounds in the literature [15], [17], [21], [22]. We only present some results related to this work.

*Lemma 1 (The Subspace Sampling for Column Selection [15]):*Given a matrix $A \in \mathbb{R}^{m\times n}$, a target rank $k \ll \min\{m, n\}$, and the probabilities in (3), the subspace sampling algorithm selects $c = O(k^2\epsilon^{-2}\log(1/\delta))$ with replacement to construct $C \in \mathbb{R}^{m\times c}$. Then

$$\|A - CC^{+}A\|_F \leq (1+\epsilon)\|A - A_k\|_F$$

holds with probability at least $1-\delta$.

2) THE SUBSPACE SAMPLING ALGORITHM FOR CUR

Drineas et al. [15] proposed a two-stage randomized CUR algorithm which has a relative-error bound with high probability (w.h.p.). In the first stage the algorithm samples c columns of A to construct C, and in the second stage it samples r rows from A and C simultaneously to construct R and W and let $U = W^{+}$. The sampling probabilities in the two stages are proportional to the leverage scores of A and C, respectively. Here we show the main results of the subspace sampling algorithm in the following lemma.

Lemma 2 (The Subspace Sampling for CUR [15]): Given a matrix $A \in \mathbb{R}^{m\times n}$ and a target rank $k \ll \min\{m, n\}$, the subspace sampling algorithm selects $c = O(k^2\epsilon^{-2}\log(1/\delta))$ columns and $r = O(c^2\epsilon^{-2}\log(1/\delta))$ rows with replacement to construct $C \in \mathbb{R}^{m\times c}$ and $R \in \mathbb{R}^{r\times n}$. Then

$$\|A - CUR\|_F = \|A - CW^{+}R\|_F \leq (1+\epsilon)\|A - A_k\|_F$$

holds with probability at least $1-\delta$, where W contains the rows of C with scaling.

B. THE ADAPTIVE SAMPLING ALGORITHM

The relative-error adaptive sampling algorithm is originally established in Theorem 2.1 of Deshpande et al. [21]. The algorithm is based on the following idea: after selecting a proportion of columns from A to form C_1 by an arbitrary algorithm, the algorithm randomly samples additional c_2 columns according to the residual $A - C_1C_1^{+}A$. Wang and Zhang [18] proved the following more general error bound for the same adaptive sampling algorithm.

Lemma 3 (The Adaptive Sampling Algorithm [18]): Given a matrix $A \in \mathbb{R}^{m\times n}$ and a matrix $C \in \mathbb{R}^{m\times c}$ such that $\text{rank}(C) = \text{rank}(CC^{+}A) = \rho(\rho \leq c \leq n)$. We let $R_1 \in \mathbb{R}^{r_1\times n}$ consist of r_1 rows of A, and define the residual $B = A - AR_1^{+}R_1$. Additionally, for $i = 1, \ldots, m$, we define $p_i = \|B_{(i)}\|_2^2/\|B\|_F^2$. We further sample r_2 rows i.i.d. from A, in each trial of which the i-th row is chosen with the probabilities p_i. Let $R_2 \in \mathbb{R}^{r_2\times n}$ contain the r_2 sampled rows and let $R = [R_1^T, R_2^T]^T \in \mathbb{R}^{(r_1+r_2)\times n}$. Then we have

$$\mathbb{E}\|A - CC^{+}AR^{+}R\|_F^2 \leq \|A - CC^{+}A\|_F^2 + \|A - AR_1^{+}R_1\|_F^2,$$

where the expectation is taken w.r.t. R_2.

Guaranteed by Lemma 3, any column selection algorithm with relative-error bound can be applied to CUR approximation. We show the result in the following lemma.

Lemma 4 (The Adaptive Sampling for CUR [18]): Given a matrix $A \in \mathbb{R}^{m\times n}$, a target rank $k(\ll m, n)$, and a column selection algorithm $\mathcal{A}_{col}$ which achieves relative-error upper bound by selecting $c \geq f(k, \epsilon)$ columns. By selecting $c \geq f(k, \epsilon)$ columns of A to construct C and $r_1 = c$ rows to construct R_1, both using the algorithm $\mathcal{A}_{col}$, followed by selecting additional $r_2 = c/\epsilon$ rows using the adaptive sampling algorithm to construct R_2, the CUR matrix decomposition achieves relative-error upper bound in expectation:

$$\mathbb{E}\|A - CUR\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2,$$

where $R = (R_1^T, R_2^T)^T$ and $U = C^{+}AR^{+}$.

According to Lemma 4, Wang and Zhang [18] combined the near-optimal column selection algorithm of Boutsidis et al. [17] and the adaptive sampling algorithm for solving the CUR problem, giving rise to an algorithm with a much tighter theoretical bound than previous algorithms. The analysis of this algorithm is given in Lemma 5.

Lemma 5 (A Special Adaptive Sampling for CUR [18]): Given a matrix $A \in \mathbb{R}^{m\times n}$, a target rank $k(\ll m, n)$, the CUR algorithm described in Algorithm 2 of Wang and Zhang [18] randomly selects $c = \frac{2k}{\epsilon}(1+o(1))$ columns of A to construct $C \in \mathbb{R}^{m\times c}$, and then selects $r = \frac{c}{\epsilon}(1+\epsilon)$ rows of A to construct $R \in \mathbb{R}^{r\times n}$. Then we have

$$\mathbb{E}\|A - CUR\|_F^2 = \mathbb{E}\|A - C(C^{+}AR^{+})R\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2.$$

Algorithm 1 Column Selection Algorithm via Energy-Based Sampling

1: **Input:** a real matrix $A \in \mathbb{R}^{m\times n}$, target rank k, error parameter $\epsilon \in (0, 1]$, confidence parameter $1-\delta \in [0, 1)$, target column number $c = O(\tau^2\epsilon^{-2}\log(1/\delta))$, where τ^2 is given in Theorem 8 and $m \geq n$ (without losing generality);
2: Compute energy sampling probabilities in (4);
3: Sampling c columns from A with the probabilities in (4) to construct C;
4: **return** C.

IV. MAIN RESULT

We first establish a novel relative-error bound theory for energy sampling column selection algorithm in Section IV-A. We then combine the energy sampling column selection algorithm and the adaptive sampling algorithm for solving the CUR problem in Section IV-B, giving rise to a new energy sampling CUR algorithm with a low time complexity and comparable accuracy.

Algorithm 2 Adaptive CUR via Energy-Based Sampling Algorithm

1: **Input:** a real matrix $A \in \mathbb{R}^{m\times n}$, target rank k, error parameter $\epsilon \in (0, 1]$, confidence parameter $1-\delta \in [0, 1)$, target column number $c = O(\tau^2\epsilon^{-2}\log(1/\delta))$, target row number $r = \frac{c}{\epsilon}(1+\epsilon)$;
2: *Stage 1*: Select c columns of A to construct $C \in \mathbb{R}^{m\times c}$ using Algorithm 1;
3: *Stage 2*: Select r rows of A to construct $R \in \mathbb{R}^{r\times n}$ using the following procedures:
4: Select $r_1 = c$ rows of A to construct $R_1 \in \mathbb{R}^{r_1\times n}$ using Algorithm 1,
5: Residual matrix $B \leftarrow A - AR_1^+R_1$,
6: Adaptively sample $r_2 = c/\epsilon$ rows from A to construct $R_2 \in \mathbb{R}^{r_2\times n}$ according to the sampling probabilities $p_i = \|B_{(i)}\|_2^2/\|B\|_F^2$, $i = 1, \ldots, m$;
7: **return** C, $R = (R_1^T, R_2^T)^T$, and $U = C^+AR^+$.

A. COLUMN SELECTION VIA ENERGY SAMPLING ALGORITHM

In this section, we first consider any probabilities that satisfy the following conditions:

$$p_j = \|A_k^{(j)}\|_2^2/\|A_k\|_F^2, \quad j = 1, \ldots, n, \tag{5}$$

where A_k is the best rank-k approximation to A.

*Theorem 6: (Approximate Energy Sampling for Column Selection):*Given a matrix $A \in \mathbb{R}^{m\times n}$ and a target rank $k \ll \min\{m, n\}$, and the probabilities in (5), the energy sampling algorithm selects $c = O(\tau^2\epsilon^{-2}\log(1/\delta))$ with replacement to construct $C \in \mathbb{R}^{m\times c}$. Then

$$\|A - CC^+A\|_F \le (1+\epsilon)\|A - A_k\|_F$$

holds with probability at least $1-\delta$, where $\tau^2 = \tau^2(A_k) = (\sum_{i=1}^k \sigma_i^2)(\sum_{i=1}^k \sigma_i^{-2})$.

Our main tools in the proofs of Theorem 6 are the use of a novel matrix inequality and similar techniques as in Drineas et al. [15]. We provide the complete proofs of Theorem 6 in Appendix.

Remark 7: In Theorem 6, if $\kappa(A_k) = 1$, then $\tau^2 = \left(\sum_{i=1}^k \sigma_i^2\right)\left(\sum_{i=1}^k \sigma_i^{-2}\right) = k^2$, and hence we may obtain the same condition on the chosen column number c as that in the subspace sampling algorithm (cf. Lemma 1). It follows from the submultiplicativity of Frobenius norm that $k^2 = \|I_k\|_F^2 = \|\Sigma_{A_k}\Sigma_{A_k}^{-1}\|_F^2 \le \|\Sigma_{A_k}\|_F^2\|\Sigma_{A_k}^{-1}\|_F^2 = \tau^2$. Thus, in general energy sampling algorithm requires more columns than subspace sampling algorithm to be chosen for the worst-case bounds. Note that the energy sampling algorithm has the same type of sampling probabilities as that in the adaptive sampling stage. We will demonstrate in Section V that the following energy-based adaptive CUR can provide very good Frobenius norm reconstruction in real data analysis by sampling a number of columns and/or rows that equals a small constant, e.g., 2 or 3, times the rank parameter k.

It is easy to see that the sampling probabilities in (5) require the calculation of SVD of A. The requirement may be either unrealistic or inefficient. Fortunately, we have a natural approach to deal with this problem. In fact, we may use the energy sampling probabilities in (4) and obtain the following theorem as a corollary of Theorem 6 in which $k = \rho$ is set.

Theorem 8 (Energy Sampling for Column Selection): Given a matrix $A \in \mathbb{R}^{m\times n}$, $\rho = \text{rank}(A)$, $\epsilon \in (0, 1]$, and the sampling probabilities in (4), the energy sampling algorithm selects $c = O(\tau^2(A)\epsilon^{-2}\log(1/\delta))$ with replacement to construct $C \in \mathbb{R}^{m\times c}$. Then

$$A = CC^+A$$

holds with probability at least $1-\delta$, where $\tau^2(A) = \tau^2(A_\rho) = (\sum_{i=1}^\rho \sigma_i^2)(\sum_{i=1}^\rho \sigma_i^{-2})$.

B. ADAPTIVE CUR VIA ENERGY SAMPLING ALGORITHM

In this section, we combine the energy sampling column selection algorithm (Theorems 6 and 8) with the adaptive sampling algorithm (Lemma 4) to the CUR problem, obtaining effective and efficient CUR algorithm.

*Theorem 9: ([Adaptive CUR With Approximate Energy Sampling):*Given a matrix $A \in \mathbb{R}^{m\times n}$, a target rank $k(\ll m, n)$. By selecting $c \ge O(\tau^2(A_k)\epsilon^{-2}\log(1/\delta))$ columns of A to construct C and $r_1 = c$ rows to construct R_1, both using Algorithm 1, followed by selecting additional $r_2 = c/\epsilon$ rows using the adaptive sampling algorithm (Algorithm 2) to construct R_2, the CUR matrix decomposition achieves relative-error upper bound:

$$\|A - CUR\|_F \le (1+\epsilon)\|A - A_k\|_F,$$

holds with probability at least $1-\delta$, where $R = (R_1^T, R_2^T)^T$ and $U = C^+AR^+$.

Similar to Theorem 8, we may use Algorithm 2 and obtain the following theorem as a corollary of Theorem 9 in which $k = \rho$ is set.

Theorem 10 (Adaptive CUR With Energy Sampling): Let a matrix $A \in \mathbb{R}^{m\times n}$, $\rho = \text{rank}(A)$, $\epsilon \in (0, 1]$. By selecting $c \ge O(\tau^2(A)\epsilon^{-2}\log(1/\delta))$ columns of A to construct C and $r_1 = c$ rows to construct R_1, both using Algorithm 1, followed by selecting additional $r_2 = c/\epsilon$ rows using the adaptive sampling algorithm (Algorithm 2) to construct R_2, we have

$$A = CUR$$

holds with probability at least $1-\delta$, where $R = (R_1^T, R_2^T)^T$ and $U = C^+AR^+$.

V. EMPIRICAL COMPARISON

In Section V-A, we conduct empirical comparisons among our energy sampling based adaptive CUR algorithm (Algorithm 2) with the other two CUR algorithms introduced in Section IV. We report the error ratio and the running time of each algorithm on each data set. The error ratio is defined by

$$\text{Error Ratio} = \frac{\|A - \widetilde{A}\|_F}{\|A - A_k\|_F},$$

where $\widetilde{A} = CUR$ for the CUR matrix decomposition, A_k is the best rank-k approximation and can be used as a relative-error. Note that we only consider the running time of each algorithm for constructing C and R, since the three CUR algorithms used two different procedures for the computation of matrix

TABLE 1. A summary of the data sets.

Dataset	Type	Size	Source
Macaw	natural image	2560×1600	http://abc.2008php.com/tuku/2011/0119/900
Owl	natural image	2560×1600	http://abc.2008php.com/tuku/2011/0119/900
Arcene	biology	10000×900	http://archive.ics.uci.edu/ml/datasets/Arcene

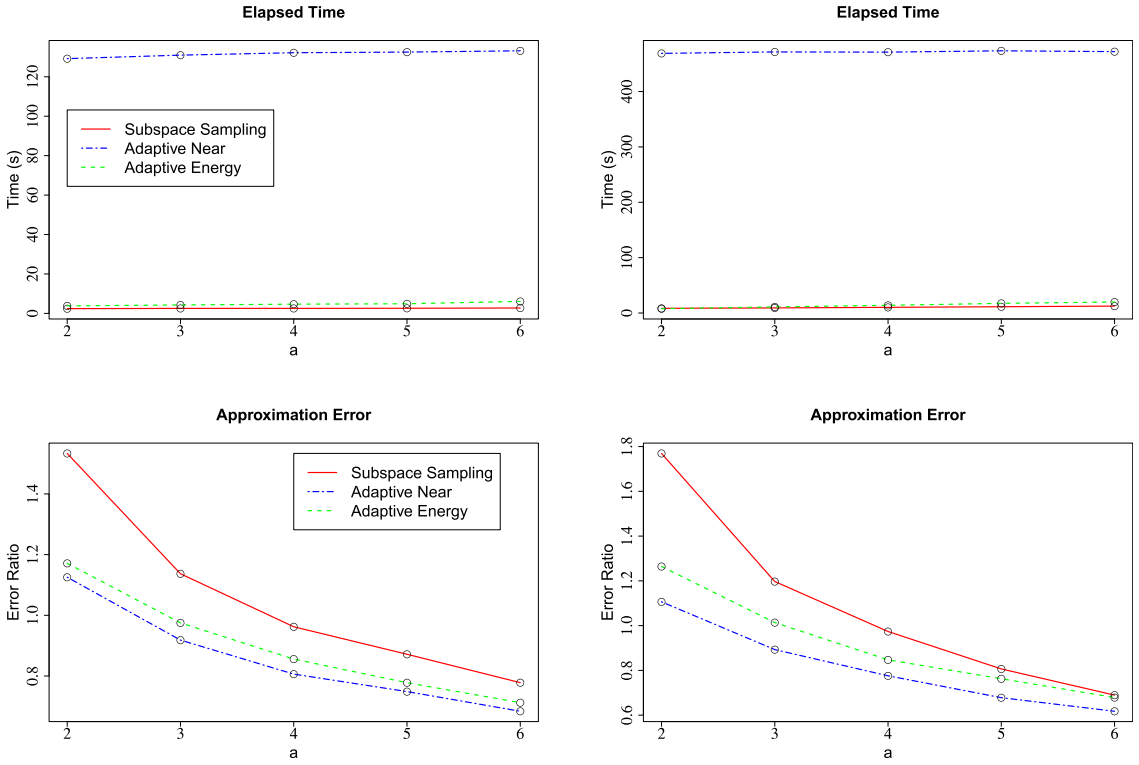


FIGURE 1. Empirical results on the Macaw data set (dense matrix); left panels: $k = 10$, $c = ak$, and $r = ac$; right panels: $k = 50$, $c = ak$, and $r = ac$.

U after constructing C and R. In Section V-B we intuitively demonstrate the true effectiveness of our method.

A. COMPARISON OF CUR ALGORITHM

In this section, we empirically compare our adaptive CUR algorithm (Algorithm 2) with the adaptive CUR algorithm of Wang and Zhang [18] and the subspace sampling algorithm of Drineas et al. [15]. For the last two sampling algorithms, we compute the sampling probabilities exactly via the truncated SVD for the sake of saving time.

We conduct experiments on three datasets obtained from the Internet, including two 2560×1600 natural images with dense matrix structure and a set of biology data with sparse matrix structure. Table 1 briefly summarizes some information of the datasets. Arcene is from the UCI datasets [23]. Arcene is a biology dataset with 900 instances and 10000 attributes. Each dataset is actually represented as a data matrix, upon which we apply the CUR algorithms. We conduct the experiments on an iMac with Intel Core i5 3.1GHz CPUs, 8GB RAM, and macOS Sierra 10.12.6 system. We implement the algorithms in R environment, and use the R Spectra package function 'svds' for truncated SVD.

The parametric setup adopted here was essentially the same as Wang and Zhang [18]. For each data set and each algorithm, we set $k = 10$ or 50, and $c = ak$, $r = ac$, where a ranges in each set of experiments. We repeat each of the three randomized algorithms 10 times, and report the minimum error ratio and the total elapsed time for constructing C and R of the 10 rounds. We depict the error ratios and the elapsed time of the three CUR matrix decomposition algorithms in Figures 1, 2, and 3.

First of all, let us compare the three CUR algorithms via examining their error ratio and running time under dense matrix structure. The results in Figures 1-2 show that the two adaptive CUR algorithms have much lower error ratio than the subspace sampling algorithm in all cases. Although our energy-based adaptive CUR algorithm has slightly larger error ratio than the near-optimal column selection based adaptive CUR algorithm, the former is much more efficient for the approximation to such dense image matrices. The experimental results show our adaptive CUR algorithm exactly achieves a good balance between computational complexity and approximate accuracy.

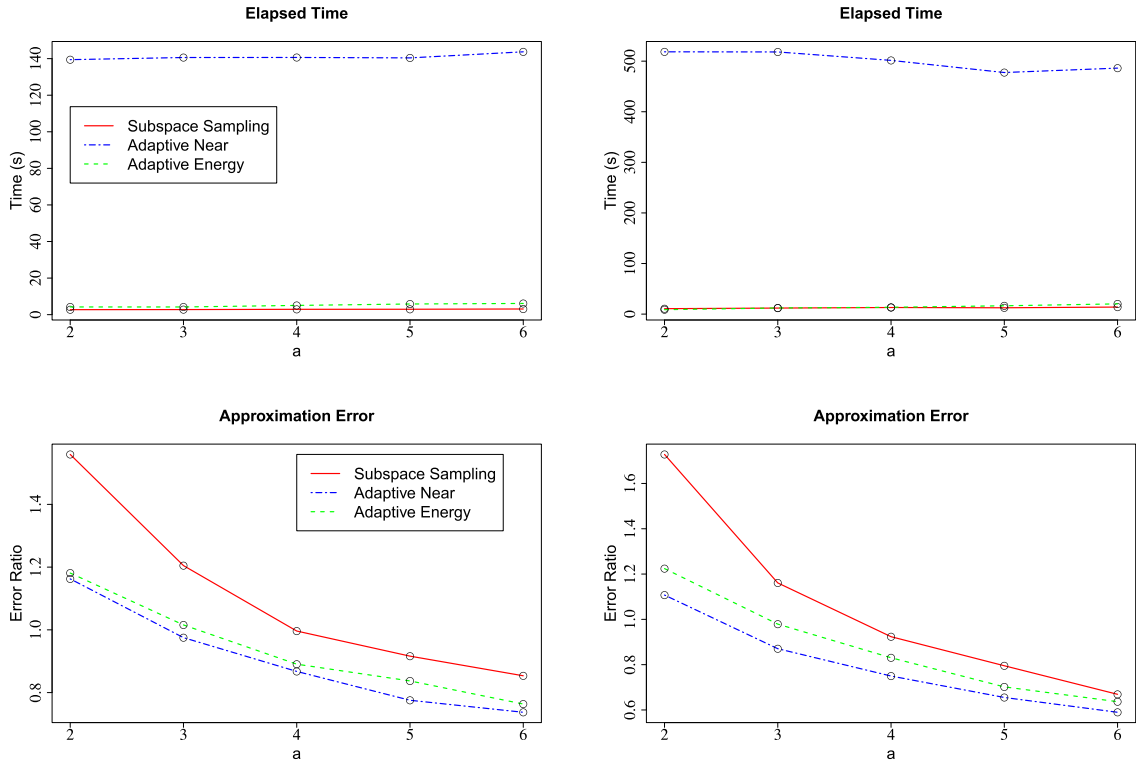


FIGURE 2. Empirical results on the Owl data set (dense matrix); left panels: $k = 10$, $c = ak$, and $r = ac$; right panels: $k = 50$, $c = ak$, and $r = ac$.

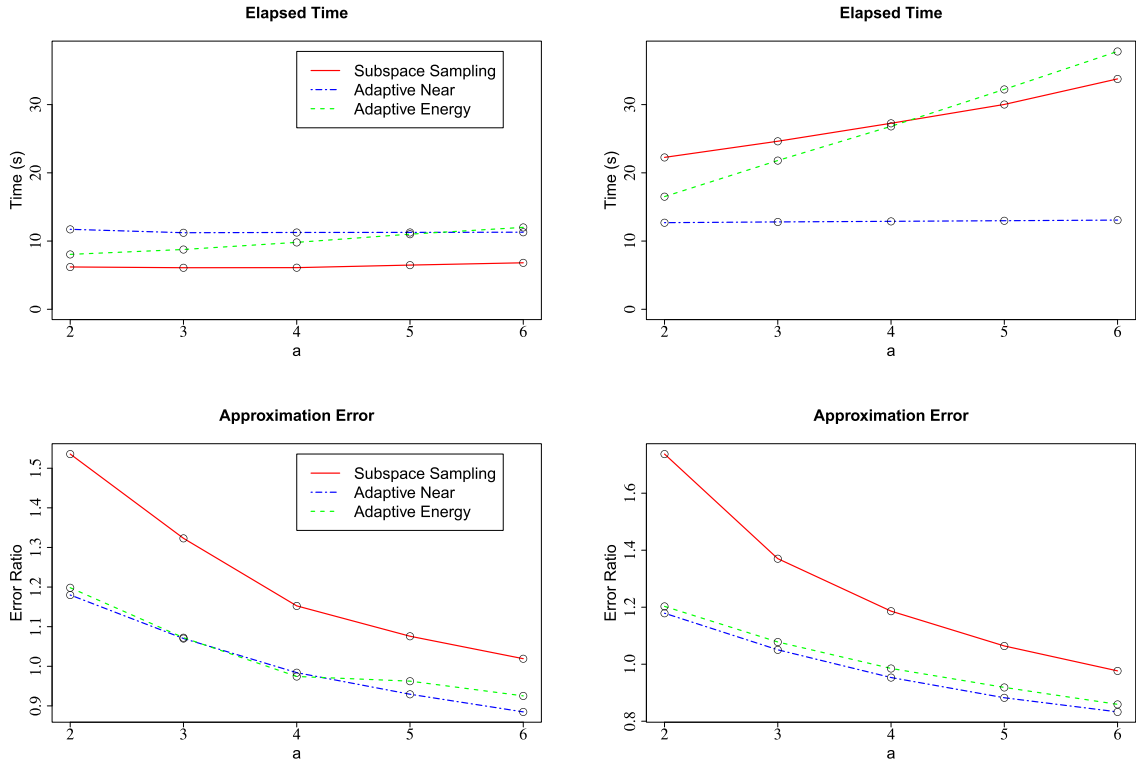
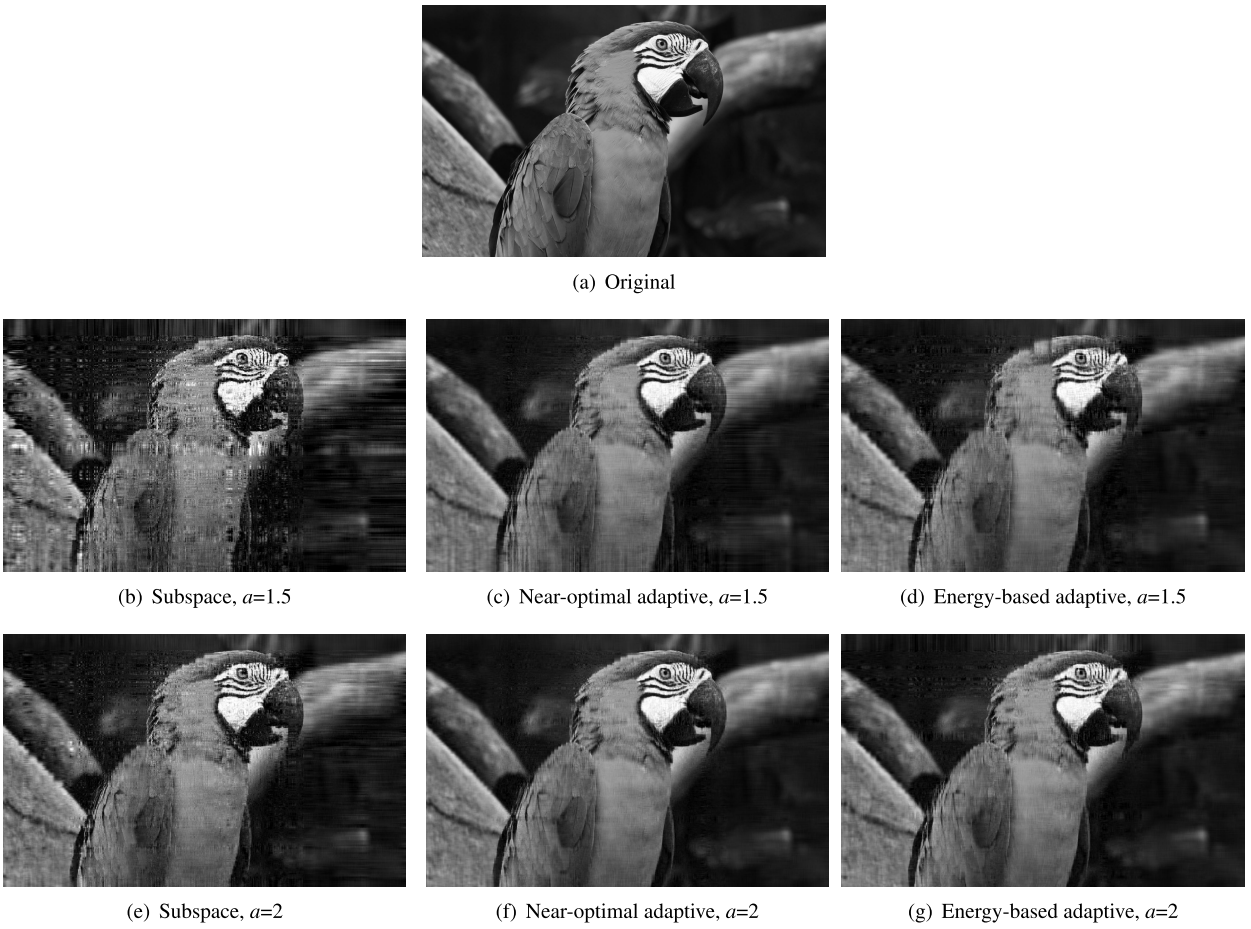


FIGURE 3. Empirical results on the Arcene data set (sparse matrix); left panels: $k = 10$, $c = ak$, and $r = ac$; right panels: $k = 50$, $c = ak$, and $r = ac$.

In the following, let us compare the three CUR algorithms via examining their error ratio and running time under sparse matrix structure. It is seen from Figure 3 that the two adaptive CUR algorithms again have much lower error ratio than the



(a) Original

(b) Subspace, a=1.5 (c) Near-optimal adaptive, a=1.5 (d) Energy-based adaptive, a=1.5

(e) Subspace, a=2 (f) Near-optimal adaptive, a=2 (g) Energy-based adaptive, a=2

FIGURE 4. (a): the original image. (b) to (d): the different CUR decompositions with $a = 1.5$. (e) to (g): the different CUR decompositions with $a = 2$.

subspace sampling algorithm. As for the running time, the other two CUR algorithms are slightly more efficient than our adaptive CUR algorithm in some cases. As pointed out in Wang and Zhang [18], this is reasonable since each of the other two CUR algorithms gains from sparsity of the biology data matrix. Note that our adaptive CUR algorithm grows linearly and is slightly more efficient than the other two CUR algorithm under certain cases. Moreover, for the sparse biology data matrix all the three CUR algorithm is more efficient than the dense image data matrix.

B. COMPARISON OF IMAGE QUALITY

To intuitively demonstrate the effectiveness of our method, we conduct a simple experiment on the 2560×1600 Macaw image discussed in Section V-A. We set target rank parameter $k = 50$ and sample $c = ak$ columns to form C and $r = ac$ rows to form R according to each of the three CUR algorithms by varying a. We show the image $\widetilde{A} = CUR$ in Figure 4.

Figure 4(b), (c) and (d) are obtained by the subspace sampling algorithm [15], the adaptive CUR algorithm of Wang and Zhang [18] and our adaptive CUR algorithm in Section IV, where $a = 1.5$. Obviously, the approximation quality of the two adaptive sampling is much better than the subspace sampling algorithm in this setting.

Figure 4(e), (f) and (g) are obtained by taking $a = 2$ and using the three CUR algorithms respectively. The approximation quality is significantly improved. Moreover the approximation quality of the two adaptive sampling is still slightly better than the subspace sampling algorithm, and with low computaional complexity, our adaptive CUR approximation quality is nearly as good as that of Wang and Zhang [18].

VI. CONCLUSION

In this paper we have built a new relative-error bound for the energy-based adaptive sampling algorithm. Accordingly, we have devised novel CUR matrix decomposition and approximation algorithms which possesses an elegant balance between computational complexity and approximate accuracy. We have shown that our adaptive sampling algorithm achieves relative-error upper bound by using very simple sampling probabilities and the corresponding adaptive sampling algorithm. Our proposed CUR algorithm is scalable provided that matrix multiplication can be highly efficiently executed. Finally, the empirical

comparisons have also demonstrated the benefits of our algorithm.

APPENDIX
PROOF

We provide the proofs of Theorem 6 in this paper in this section. Our main tools in the proofs are the use of a matrix inequality and several lemmas of general interest.

Lemma 11 (A Matrix Inequality): Let $A \in \mathbb{R}^{\rho\times\rho}$ be a symmetric matrix, and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_\rho) \in \mathbb{R}^{\rho\times\rho}$ a diagonal invertible matrix. Then

$$\|A\|_F \leq \|\Sigma A \Sigma^{-1}\|_F, \tag{6}$$

and the equality holds if and only if

$$|\sigma_1| = \cdots = |\sigma_\rho|.$$

Proof: Recall that for any matrix $D \in \mathbb{R}^{m\times n}$,

$$\|D\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij}^2,$$

and note that

$$\Sigma A \Sigma^{-1} = \begin{pmatrix} \sigma_1\sigma_1^{-1}A_{11} & \cdots & \sigma_1\sigma_\rho^{-1}A_{1\rho} \\ \vdots & \ddots & \vdots \\ \sigma_\rho\sigma_1^{-1}A_{\rho 1} & \cdots & \sigma_\rho\sigma_\rho^{-1}A_{\rho\rho} \end{pmatrix}.$$

Thus we have from the symmetry of A that

$$\begin{aligned}
\|\Sigma A \Sigma^{-1}\|_F^2 &= \sum_{i=1}^{\rho}\sum_{j=1}^{\rho} \sigma_i^2\sigma_j^{-2}A_{ij}^2 \\
&= \sum_{i=1}^{\rho} A_{ii}^2 + \sum_{i\neq j} \sigma_i^2\sigma_j^{-2}A_{ij}^2 \\
&= \sum_{i=1}^{\rho} A_{ii}^2 + \sum_{i<j} \left(\frac{\sigma_i^2}{\sigma_j^2} + \frac{\sigma_j^2}{\sigma_i^2}\right) A_{ij}^2 \\
&\geq \sum_{i=1}^{\rho} A_{ii}^2 + \sum_{i<j} 2A_{ij}^2 \\
&= \sum_{i=1}^{\rho} A_{ii}^2 + \sum_{i\neq j} A_{ij}^2 \\
&= \|A\|_F^2,
\end{aligned}$$

and the lemma follows.

Since our main algorithms will involve sampling columns and/or rows from input matrices, we provide a brief review of a sampling matrix formalism that was introduced in Drineas et al. [15] and with respect to which our sampling matrix operations may be conveniently expressed. First, assume that c columns of A are chosen in c i.i.d. trials by randomly sampling according to a probability distribution $\{p_i\}_{i=1}^n$, and assume that the i_tth column of A is chosen in the tth (for $t = 1, \ldots, c$) independent random trial. Then, define a block sampling matrix $S \in \mathbb{R}^{n\times c}$ to be a zero-one matrix where $S_{i_t t} = 1$ and $S_{ij} = 0$ otherwise, and define the rescaling matrix $D \in \mathbb{R}^{c\times c}$ to be the diagonal matrix with $D_{tt} = 1/\sqrt{cp_{i_t}}$, where p_{i_t} is the probability of choosing the i_tth column. Clearly, $C = ASD$ is an $m \times c$ matrix consisting of sampled and rescaled copies of the columns of A, and $R = (SD)^T A = D^T S^T A$ is a $c \times n$ matrix consisting of sampled and rescaled copies of the rows of A. In certain cases, we subscript S and D with C or R (e.g., $C = AS_C D_C$ and $R = D_R^T S_R^T A$) to make explicit that the corresponding sampling and rescaling matrices are operating on the columns or rows, respectively, of A.

For simplicity of notation in the proofs of the next several lemmas, we let $\mathcal{S} = DS^T$ denote the rescaled row sampling matrix. Let the rank of the matrix $\mathcal{S}U_A = DS^T U_A$ be $\widetilde{\rho}$, and let its SVD be

$$\mathcal{S}U_A = U_{\mathcal{S}U_A}\Sigma_{\mathcal{S}U_A}V_{\mathcal{S}U_A}^T,$$

where $\widetilde{\rho} \leq \rho$.

Lemma 12: Let $\epsilon \in (0, 1]$, e be the Euler's number, approximately equal to 2.71828, $\tau^2(A) = \left(\sum_{i=1}^{\rho}\sigma_i^2(A)\right)\left(\sum_{i=1}^{\rho}\sigma_i^{-2}(A)\right)$, and define $\Upsilon = (\mathcal{S}U_A)^+ - (\mathcal{S}U_A)^T$. If the sampling probabilities satisfy (4) and if $r \geq 36\tau\text{e}^2/(\alpha\epsilon^2)$, then with probability at least $1 - 1/3\text{e}$:

$$\widetilde{\rho} = \rho \; i.e., \;\; \text{rank}(\mathcal{S}U_A) = \text{rank}(U_A) = \text{rank}(A), \tag{7}$$

$$(\mathcal{S}A)^+ = V_A\Sigma_A^{-1}(\mathcal{S}U_A)^+, \tag{8}$$

$$\|\Upsilon\|_2 = \|\Sigma_{\mathcal{S}U_A} - \Sigma_{\mathcal{S}U_A}^{-1}\|_2 \leq \epsilon/\sqrt{2}. \tag{9}$$

Proof: To prove (7), note that for all $i \in \{1, \ldots, \rho\}$

$$\begin{aligned}
&|1 - \sigma_i^2(\mathcal{S}U_X)| \\
&\quad = |\sigma_i(U_X^T U_X) - \sigma_i^2(U_X^T\mathcal{S}^T\mathcal{S}U_X)| \\
&\quad \leq \|U_X^T U_X - U_X^T\mathcal{S}^T\mathcal{S}U_X\|_2 && (10)\\
&\quad \leq \|U_X^T U_X - U_X^T\mathcal{S}^T\mathcal{S}U_X\|_F && (11)\\
&\quad \leq \|\Sigma_X\left(U_X^T U_X - U_X^T\mathcal{S}^T\mathcal{S}U_X\right)\Sigma_X^{-1}\|_F && (12)\\
&\quad = \|V_X\Sigma_X\left(U_X^T U_X - U_X^T\mathcal{S}^T\mathcal{S}U_X\right)\Sigma_X^{-1}\|_F && (13)\\
&\quad = \|X^T U_X\Sigma_X^{-1} - X^T\mathcal{S}^T\mathcal{S}U_X\Sigma_X^{-1}\|_F. && (14)
\end{aligned}$$

Note that (10) follows from Corollary 8.1.6 of Golub and Van Loan [24], (11) follows since $\|\cdot\|_2 \leq \|\cdot\|_F$, (12) follows from Lemma 19, and (13) follows since V_X is a matrix with orthogonal columns. To bound the error of approximating $\|X^T U_X\Sigma_X^{-1}\|_F$ by $\|X^T\mathcal{S}^T\mathcal{S}U_X\Sigma_X^{-1}\|_F$, we apply Theorem 6 in Drineas et al. [11]. Since the sampling probabilities p_i satisfy (4), it follows from Theorem 1 in Drineas et al. [11]. and by applying Markov's inequality that with probability at least $1 - 1/3\text{e}$:

$$\begin{aligned}
&\|X^T U_X\Sigma_X^{-1} - X^T\mathcal{S}^T\mathcal{S}U_X\Sigma_X^{-1}\|_F \\
&\quad \leq 3\text{e}\mathbf{E}\left[\|X^T U_X\Sigma_X^{-1} - X^T\mathcal{S}^T\mathcal{S}U_X\Sigma_X^{-1}\|_F\right] \\
&\quad \leq \frac{3\text{e}}{\sqrt{\alpha b}}\|X^T\|_F\|U_X\Sigma_X^{-1}\|_F. && (15)
\end{aligned}$$

By combining (14) and (15), recalling that $\tau = \|\Sigma_X\|_F\|\Sigma_X^{-1}\|_F = \|X^T\|_F\|U_X\Sigma_X^{-1}\|_F$, and using the

assumed choice of b, it follows that

$$|1-\sigma_i^2(\mathcal{S}U_X)| \le \epsilon/2 \le 1/2$$

since $\epsilon \le 1$. This implies that all singular values of $\mathcal{S}U_X$ are strictly positive, and thus that $\text{rank}(\mathcal{S}U_X) = \text{rank}(U_X) = \text{rank}(X)$, which establishes the claim (7).

To prove the second claim, note that

$$\begin{aligned}(\mathcal{S}X)^+ &= (\mathcal{S}U_X\Sigma_X V_X^T)^+ \\ &= (U_{\mathcal{S}U_X}\Sigma_{\mathcal{S}U_X}V_{\mathcal{S}U_X}^T\Sigma_X V_X^T)^+ \\ &= V_X(\Sigma_{\mathcal{S}U_X}V_{\mathcal{S}U_X}^T\Sigma_X)^+U_{\mathcal{S}U_X}^T. \end{aligned} \tag{16}$$

Notice that since $\rho = \widetilde{\rho}$ with probability at least $1-1/3\text{e}$, all three matrices $\Sigma_{\mathcal{S}U_X}$, $V_{\mathcal{S}U_X}^T$, and Σ_X are square $\rho\times\rho$ matrices with full rank, and thus are invertible. In this case,

$$\begin{aligned}(\Sigma_{\mathcal{S}U_X}V_{\mathcal{S}U_X}^T\Sigma_X)^+ &= (\Sigma_{\mathcal{S}U_X}V_{\mathcal{S}U_X}^T\Sigma_X)^{-1} \\ &= \Sigma_X^{-1}V_{\mathcal{S}U_X}\Sigma_{\mathcal{S}U_X}^{-1}. \end{aligned} \tag{17}$$

By combining (16) and (17) we obtain that

$$\begin{aligned}(\mathcal{S}X)^+ &= V_X\Sigma_X^{-1}V_{\mathcal{S}U_X}\Sigma_{\mathcal{S}U_X}^{-1}U_{\mathcal{S}U_X}^T \\ &= V_X\Sigma_X^{-1}(\mathcal{S}U_X)^+, \end{aligned}$$

which establishes the second claim (8).

To prove the third claim (9), we use the SVD of $\mathcal{S}U_X$ and note that

$$\begin{aligned}&\|\Upsilon\|_2 \\ &= \|(\mathcal{S}U_X)^+ - (\mathcal{S}U_X)^T\|_2 \\ &= \|(U_{\mathcal{S}U_X}\Sigma_{\mathcal{S}U_X}V_{\mathcal{S}U_X}^T)^+ - (U_{\mathcal{S}U_X}\Sigma_{\mathcal{S}U_X}V_{\mathcal{S}U_X}^T)^T\|_2 \\ &= \|V_{\mathcal{S}U_X}\left(\Sigma_{\mathcal{S}U_X}^{-1} - \Sigma_{\mathcal{S}U_X}\right)U_{\mathcal{S}U_X}^T\|_2 \\ &= \|\Sigma_{\mathcal{S}U_X}^{-1} - \Sigma_{\mathcal{S}U_X}\|_2, \end{aligned}$$

which establishes the equality in (9). The last equality holds since $V_{\mathcal{S}U_X}$ and $U_{\mathcal{S}U_X}$ are matrices with orthogonal columns. Finally, to prove the inequality in (9), recall that under the assumptions of the lemma $\rho = \widetilde{\rho}$ with probability at least $1-1/3\text{e}$, and thus $\sigma_i(\mathcal{S}U_X) > 0$ for all $i \in \{1,\dots,\rho\}$. Note that

$$\begin{aligned}&\|\Sigma_{\mathcal{S}U_X}^{-1} - \Sigma_{\mathcal{S}U_X}\|_2 \\ &= \max_{i\in\{1,\dots,\rho\}}\left|\sigma_i(\mathcal{S}U_X) - \frac{1}{\sigma_i(\mathcal{S}U_X)}\right| \\ &= \max_{i\in\{1,\dots,\rho\}}\left|\frac{\sigma_i^2(\mathcal{S}U_X)-1}{\sigma_i(\mathcal{S}U_X)}\right| \end{aligned} \tag{18}$$

Using the fact that, by (10), for all $i \in \{1,\dots,\rho\}$,

$$|1-\sigma_i^2(\mathcal{S}U_X)| \le \|U_X^TU_X - U_X^T\mathcal{S}^T\mathcal{S}U_X\|_2,$$

it follows that for all $i \in \{1,\dots,\rho\}$

$$\frac{1}{\sigma_i(\mathcal{S}U_X)} \le \frac{1}{\sqrt{1-\|U_X^TU_X - U_X^T\mathcal{S}^T\mathcal{S}U_X\|_2}}.$$

When these are combined with (18) it follows that

$$\begin{aligned}&\|\Sigma_{\mathcal{S}U_X}^{-1} - \Sigma_{\mathcal{S}U_X}\|_2 \\ &\le \frac{\|U_X^TU_X - U_X^T\mathcal{S}^T\mathcal{S}U_X\|_2}{\sqrt{1-\|U_X^TU_X - U_X^T\mathcal{S}^T\mathcal{S}U_X\|_2}}. \end{aligned}$$

Combining this with the Frobenius norm bound of (15), and noticing that our choice for b guarantees that $1-\|U_X^TU_X - U_X^T\mathcal{S}^T\mathcal{S}U_X\|_2 \ge 1/2$, concludes the proof of the inequality in (9). This concludes the proof of the lemma.

The next lemma provides an approximate matrix multiplication bound that is useful in the proof of Lemma 15.

Lemma 13: Let $\epsilon \in (0,1]$, and τ be defined as in Lemma 12. If the sampling probabilities satisfy (4) and if $b \ge 36\tau\text{e}^2/(\alpha\epsilon^2)$, then with probability at least $1-1/3\text{e}$:

$$\|U_X^T\mathcal{S}^T\mathcal{S}U_X^\perp U_X^{\perp T}Y\|_F \le \frac{\epsilon}{2}\|U_X^\perp U_X^{\perp T}Y\|_F.$$

Proof: First, note that since U_X is an orthogonal matrix and since $U_X^TU_X^\perp = 0$, we have that

$$\begin{aligned}&\|U_X^T\mathcal{S}^T\mathcal{S}U_X^\perp U_X^{\perp T}Y\|_F \\ &= \|\Sigma_X^{-1}\Sigma_X U_X^T\mathcal{S}^T\mathcal{S}U_X^\perp U_X^{\perp T}Y\|_F \\ &\le \|\Sigma_X^{-1}\|_F\|\Sigma_X U_X^T\mathcal{S}^T\mathcal{S}U_X^\perp U_X^{\perp T}Y\|_F \\ &= \|\Sigma_X^{-1}\|_F\|\Sigma_X U_X^TU_X^\perp U_X^{\perp T}Y - \Sigma_X U_X^T\mathcal{S}^T\mathcal{S}U_X^\perp U_X^{\perp T}Y\|_F \\ &= \|\Sigma_X^{-1}\|_F\|X^TU_X^\perp U_X^{\perp T}Y - X^T\mathcal{S}^T\mathcal{S}U_X^\perp U_X^{\perp T}Y\|_F \end{aligned} \tag{19}$$

Since the sampling probabilities satisfy (4) and thus are appropriate for bounding the right-hand side of (19). Thus, it follows from Markov's inequality and Theorem 6 in Drineas et al. [11] that with probability at least $1-1/3\text{e}$:

$$\begin{aligned}&\|U_X^T\mathcal{S}^T\mathcal{S}U_X^\perp U_X^{\perp T}Y\|_F \\ &\le 3\text{e}\mathbf{E}\|U_X^T\mathcal{S}^T\mathcal{S}U_X^\perp U_X^{\perp T}Y\|_F \\ &\le \frac{3\text{e}}{\sqrt{\alpha b}}\|\Sigma_X^{-1}\|_F\|X^T\|_F\|U_X^\perp U_X^{\perp T}Y\|_F \\ &= \frac{3\tau\text{e}}{\sqrt{\alpha b}}\|U_X^\perp U_X^{\perp T}Y\|_F. \end{aligned}$$

The lemma follows by the choice of b.

Lemma 14: With probability at least $1-1/3\text{e}$:

$$\|\mathcal{S}U_X^\perp U_X^{\perp T}Y\|_F \le \sqrt{3\text{e}}\|U_X^\perp U_X^{\perp T}Y\|_F.$$

Proof: Let $Q = U_X^\perp U_X^{\perp T}Y$, and let $i_1, i_2, \dots, i_b$ be the b rows of Q that were included in $\mathcal{S}Q = DS^TQ$. Clearly,

$$\begin{aligned}\mathbf{E}\left[\|\mathcal{S}U_X^\perp U_X^{\perp T}Y\|_F^2\right] &= \mathbf{E}\left[\|DS^TQ\|_F^2\right] \\ &= \sum_{t=1}^{b}\sum_{i=1}^{M}p_i\frac{\|Q_{[i]}\|_F^2}{bp_i} = \|Q\|_F^2. \end{aligned}$$

The lemma follows by applying Markov's inequality and taking the square root of both sides of the resulting inequality.

Lemma 15: Suppose $X \in \mathbb{R}^{n\times p}$ has rank ρ, $Y \in \mathbb{R}^{n\times q}$. Let

$$\mathcal{Z} = \min_{\Theta\in\mathbb{R}^{p\times q}}\|Y - X\Theta\|_F = \|Y - X\widehat{\Theta}\|_F,$$

where $\widehat{\Theta} = X^+Y$, let us run Algorithm 1 with the sampling probabilities $\{p_i\}_{i=1}^N$ of the form (4) and assume the algorithm returns as output a $p\times q$ vector $\widetilde{\Theta} = (\mathcal{S}X)^+\mathcal{S}Y$, let

$\epsilon \in (0, 1]$, and let $\tau = \left(\sum_{i=1}^{\rho} \sigma_i^2(X)\right)\left(\sum_{i=1}^{\rho} \sigma_i^{-2}(X)\right)$. If $b = O\left(\tau\epsilon^{-2}\ln(1/\delta)\right)$ rows are chosen with Algorithm 1, then with probability at least $1-\delta$:

$$\|Y - X\widetilde{\Theta}\|_F \leq (1+\epsilon)\mathcal{Z}. \tag{20}$$

Proof: We provide a bound for $\|Y - X\widetilde{\Theta}\|_F$ in terms of $\mathcal{Z}$, thus proving (20). First, we prove the claim (a): if $b = 324\tau e^2/(\Theta\epsilon^2)$ rows are chosen with the algorithm, then equation (20) holds with probability at least $1 - 1/e$.

For the moment, let us assume that $b = 36\tau e^2/(\Theta\epsilon^2)$, in which case the assumption on b is satisfied for each of Lemma 12, Lemma 13, and Lemma 14. Thus, the claims of all three lemmas hold simultaneously with probability at least $1 - 3(1/3e) = 1 - 1/e$, and so let us condition on this event.

First, we have that

$$\begin{aligned} Y - X\widetilde{\Theta} &= Y - X(\mathcal{S}X)^+\mathcal{S}Y \\ &= Y - U_X(\mathcal{S}U_X)^+\mathcal{S}Y \qquad (21) \\ &= Y - U_X(\mathcal{S}U_X)^+\mathcal{S}U_X U_X^T Y \\ &\quad - U_X(\mathcal{S}U_X)^+\mathcal{S}U_X^\perp U_X^{\perp T} Y \qquad (22) \\ &= U_X^\perp U_X^{\perp T} Y - U_X(\mathcal{S}U_X)^+\mathcal{S}U_X^\perp U_X^{\perp T} Y. \qquad (23) \end{aligned}$$

Equation (21) follows from (8) of Lemma 12, (22) follows by inserting $U_X U_X^T + U_X^\perp U_X^{\perp T} = I_n$, and (23) follows since $(\mathcal{S}U_X)^+\mathcal{S}U_X = I_\rho$ by Lemma 12.

By taking the Frobenius norm of both sides of (23), by using the triangle inequality, and recalling that $\Upsilon = (\mathcal{S}U_X)^+ - (\mathcal{S}U_X)^T$, we have that

$$\begin{aligned} &\|Y - X\widetilde{\Theta}\|_F \\ &\leq \|U_X^\perp U_X^{\perp T} Y\|_F + \|U_X(\mathcal{S}U_X)^T\mathcal{S}U_X^\perp U_X^{\perp T} Y\|_F \\ &\quad + \|U_X\Omega\mathcal{S}U_X^\perp U_X^{\perp T} Y\|_F \\ &\leq \|U_X^\perp U_X^{\perp T} Y\|_F + \|U_X^T\mathcal{S}^T\mathcal{S}U_X^\perp U_X^{\perp T} Y\|_F \\ &\quad + \|\Upsilon\|_2\|\mathcal{S}U_X^\perp U_X^{\perp T} Y\|_F \qquad (24) \end{aligned}$$

where (24) follows by submultiplicativity and since U_X has orthogonal columns. By combining (24) with the bounds provided by Lemmas 12, 13, and 14, it follows that

$$\begin{aligned} \|Y - X\widetilde{\Theta}\|_F &\leq (1 + \epsilon/2 + \sqrt{3e}\epsilon/\sqrt{2})\mathcal{Z} \\ &\leq (1 + 3\epsilon)\mathcal{Z}. \end{aligned}$$

Equation (20) follows with probability at least $1 - 1/e$ by setting $\epsilon' = \epsilon/9$ and using the value of $b = 324\tau e^2/(\alpha\epsilon^2)$. The claim (a) now can be boosted to hold with probability at least $1-\delta$ using standard methods. In particular, consider the following: run the algorithm with $b = 324\tau e^2/(\alpha\epsilon^2)$ independently $\ln(1/\delta)$ times, and return a $\widetilde{\mathcal{Z}}$ such that the $\widetilde{\mathcal{Z}}$ is smallest. Then, since in each trial the claim (a) fails with probability less than $1/e$, the claim (a) will fail for every trial with probability less than $(1/e)^{\ln(1/\delta)} = \delta$. This establishes equation (20).

The Proofs of Theorem 6.

Since for every set of columns $C = AS_C D_C$, $\widehat{\Theta} = C^+A$ is the matrix that minimizes $\|A - C\Theta\|_F$, it follows that

$$\begin{aligned} &\|A - CC^+A\|_F \\ &\quad = \|A - (AS_C D_C)(AS_C D_C)^+A\|_F \\ &\quad \leq \|A - (AS_C D_C)(A_k S_C D_C)^+A_k\|_F. \qquad (25) \end{aligned}$$

To bound (25), consider the problem of approximating the solution to $\min_{\Theta\in\mathbb{R}^{m\times m}} \|A - \Theta A_k\|_F$ by randomly sampling columns of A_k and of A. It follows as a corollary of Lemma 15 that

$$\begin{aligned} &\|A - (AS_C D_C)(A_k S_C D_C)^+A_k\|_F \\ &\quad \leq (1+\epsilon)\|A - AA_k^+A_k\|_F \\ &\quad \leq (1+\epsilon)\|A - A_k\|_F \end{aligned}$$

which establishes the theorem by combining (25).

ACKNOWLEDGMENT

The authors are grateful to the anonymous referees for many valuable comments and constructive suggestions which resulted in the present version.

REFERENCES

[1] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 3, pp. 697–702, Jan. 2009.
[2] D. P. Woodruff, "Sketching as a tool for numerical linear algebra," *Found. Trends Theor. Comput. Sci.*, vol. 10, nos. 1–2, pp. 1–157, 2014.
[3] M. W. M. Boyd, "Randomized algorithms for matrices and data," *Found. Trends Mach. Learn.*, vol. 3, no. 2, pp. 123–224, 2010.
[4] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, Jan. 2011.
[5] C. Chen, M. Gu, Z. Zhang, W. Zhang, and Y. Yu, "Efficient spectrum-revealing CUR matrix decomposition," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 766–775.
[6] Y. Dong and P.-G. Martinsson, "Simpler is better: A comparative study of randomized algorithms for computing the CUR decomposition," 2021, *arXiv:2104.05877*.
[7] H. Cai, K. Hamm, L. Huang, and D. Needell, "Robust CUR decomposition: Theory and imaging applications," *SIAM J. Imag. Sci.*, vol. 14, no. 4, pp. 1472–1503, Jan. 2021.
[8] H. Q. Cai, K. Hamm, L. Huang, and D. Needell, "Mode-wise tensor decompositions: Multi-dimensional generalizations of CUR decompositions," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 8321–8356, 2021.
[9] M. Che, J. Chen, and Y. Wei, "Perturbations of the TCUR decomposition for tensor valued data in the tucker format," *J. Optim. Theory Appl.*, vol. 194, no. 3, pp. 852–877, Sep. 2022.
[10] J. Chen, Y. Wei, and Y. Xu, "Tensor CUR decomposition under T-product and its perturbation," *Numer. Funct. Anal. Optim.*, vol. 43, no. 6, pp. 698–722, Apr. 2022.
[11] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication," *SIAM J. Comput.*, vol. 36, no. 1, pp. 132–157, Jan. 2006.
[12] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix," *SIAM J. Comput.*, vol. 36, no. 1, pp. 158–183, Jan. 2006.
[13] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition," *SIAM J. Comput.*, vol. 36, no. 1, pp. 184–206, Jan. 2006.
[14] N. Mitrovic, M. T. Asif, U. Rasheed, J. Dauwels, and P. Jaillet, "CUR decomposition for compression and compressed sensing of large-scale traffic data," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2013, pp. 1475–1480.

[15] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error *CUR* matrix decompositions," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 844–881, Jan. 2008.
[16] S. Wang and Z. Zhang, "A scalable cur matrix decomposition algorithm: Lower time complexity and tighter bound," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–13.
[17] C. Boutsidis, P. Drineas, and M. Magdon-Ismail, "Near-optimal column-based matrix reconstruction," *SIAM J. Comput.*, vol. 43, no. 2, pp. 687–717, Jan. 2014.
[18] S. Wang and Z. Zhang, "Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling," *J. Mach. Learn. Res.*, vol. 14, pp. 2729–2769, Jan. 2013.
[19] S. Wang, Z. Zhang, and T. Zhang, "Towards more efficient SPSD matrix approximation and CUR matrix decomposition," *J. Mach. Learn. Res.*, vol. 17, no. 210, pp. 1–49, 2016.
[20] C. Boutsidis and D. P. Woodruff, "Optimal CUR matrix decompositions," *SIAM J. Comput.*, vol. 46, no. 2, pp. 543–589, Jan. 2017.
[21] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang, "Matrix approximation and projective clustering via volume sampling," *Theory Comput.*, vol. 2, no. 2006, pp. 225–247, 2006.
[22] A. Deshpande and L. Rademacher, "Efficient volume sampling for row/column subset selection," in *Proc. IEEE 51st Annu. Symp. Found. Comput. Sci.*, Oct. 2010, pp. 329–338.
[23] A. Frank and A. Asuncion, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2010. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Arcene
[24] G. H. Golub and C. Van Loan, *Matrix Computations*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996.

LIWEN XU was born in Anhui, China, in 1977. He received the B.S. degree in mathematics from Changsha Institute of Electric Power, Changsha, in 2000, the M.S. degree in applied mathematics from Hunan University, Changsha, in 2003, and the Ph.D. degree in probability and mathematical statistics from the Beijing University of Technology, Beijing, in 2006. From 2006 to 2008, he was a Postdoctoral Researcher with the Department of Mathematical Science, Tsinghua University. Since 2008, he has been an Assistant Professor with the Department of Statistics, North China University of Technology, where he was a Full Professor of statistics, in 2015. He is the author of three books and more than 50 articles. His research interests include decentralized learning over multitask networks, deep learning, subsampling, and smoothing spline.

XUEJIAO ZHAO was born in Hebei, China, in 2002. She is currently pursuing the B.S. degree in statistics with the College of Science, North China University of Technology. Her research interests include object detection, deep learning, and machine learning.

YONGXIA ZHANG was born in Gansu, China, in 1989. She received the B.S. degree in automation from Wuhan University, Wuhan, in 2010, the M.S. degree in computer application technology from the University of Chinese Academy of Sciences, Beijing, in 2013, and the Ph.D. degree in actuarial science from the Renmin University of China, Beijing, in 2018. From 2018 to 2021, she was a Postdoctoral Researcher with the School of Statistics, Renmin University of China. Since 2021, she has been an Assistant Professor with the Department of Statistics, North China University of Technology. She is the author of seven articles. Her research interests include quantile regression, latent factor model, Bayesian statistics, and deep learning.