Deterministic Sparse Column Based Matrix Reconstruction via Greedy Approximation of SVD

Ali Çivril and Malik Magdon-Ismail

Computer Science Department, RPI, 110 8th Street, Troy, NY 12180
{civria,magdon}@cs.rpi.edu

Abstract. Given a matrix $A \in \mathbb{R}^{m\times n}$ of rank r, and an integer $k < r$, the top k singular vectors provide the best rank-k approximation to A. When the columns of A have specific meaning, it is desirable to find (provably) "good" approximations to A_k which use only a small number of columns in A. Proposed solutions to this problem have thus far focused on randomized algorithms. Our main result is a simple greedy deterministic algorithm with guarantees on the performance and the number of columns chosen. Specifically, our greedy algorithm chooses c columns from A with $c = O\left(\frac{k^2 \log k}{\epsilon^2}\mu^2(A) \ln\left(\frac{\sqrt{k}\|A_k\|_F}{\epsilon\|A-A_k\|_F}\right)\right)$ such that

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

where C_{gr} is the matrix composed of the c columns, C_{gr}^{+} is the pseudo-inverse of C_{gr} ($C_{gr}C_{gr}^{+}A$ is the best reconstruction of A from C_{gr}), and $\mu(A)$ is a measure of the *coherence* in the normalized columns of A. The running time of the algorithm is $O(SVD(A_k)+mnc)$ where $SVD(A_k)$ is the running time complexity of computing the first k singular vectors of A. To the best of our knowledge, this is the first deterministic algorithm with performance guarantees on the number of columns and a $(1+\epsilon)$ approximation ratio in Frobenius norm. The algorithm is quite simple and intuitive and is obtained by combining a generalization of the well known *sparse approximation problem* from information theory with an *existence* result on the possibility of sparse approximation. Tightening the analysis along either of these two dimensions would yield improved results.

1 Introduction

Most data can be represented as an $m \times n$ matrix where the columns are objects and the rows are the features associated with them. Hence, given a matrix $A \in \mathbb{R}^{m\times n}$, one might be interested in obtaining the "important" spectral information of A by using some compressed representation. The usual approach to this problem is to take the best rank k ($k \ll \min\{m,n\}$) approximation A_k, which minimizes the error with respect to any unitarily invariant norm. A_k can be constructed from the top k singular vectors in $O(\min\{mn^2, m^2n\})$ time. The

first k singular vectors required to construct A_k can be computed efficiently using Lanczos methods. The problem with this general approach, which was also pointed out by [10] is that the singular vector representation might not be suitable to make inferences about the actual underlying data, because they are generally combinations of all the columns of the raw information in A. An example of this is the microarray data where the combinations of the column vectors have no sensible interpretation [16]. Hence, it is of practical importance to represent the approximation to A by a small number of columns of A.

1.1 Our Contributions

We give a deterministic greedy algorithm for low rank matrix reconstruction which is based on the sparse approximation of the SVD of A. We first generalize the sparse approximation problem of approximating vector [18] to one of approximating a subspace, using a small number of columns from A. We analyse a greedy algorithm which generalizes the analysis in [18]; in order to correct a minor technical error in the proof therein, we introduce a *coherence* parameter for a matrix, the *rank coherence parameter* which can be thought of as a more general and robust version of the coherence parameters defined in [21].

Our algorithm first computes the top k left singular vectors of A, and then selects columns of A in a greedy fashion so as to "fit" the space spanned by the singular vectors, appropriately scaled according to the singular values. The performance charasteristics of the algorithm depend on how well the greedy algorithm approximates the optimal choice of such columns from A, and on how good the optimal columns themselves are. We give an existence result on the quality of the optimal columns, and the necessary analysis of the greedy algorithm to arrive at the following result:

Theorem 1 *The greedy algorithm chooses a column submatrix $C_{gr} \subseteq A$ with $c = O\left(\frac{k^2 \log k}{\epsilon^2}\mu^2(A) \ln\left(\frac{\sqrt{k}\|A_k\|_F}{\epsilon\|A-A_k\|_F}\right)\right)$ columns such that*

$$\|A - C_{gr}C_{gr}^{+}A\|_F \leq (1+\epsilon)\,\|A - A_k\|_F.$$

The term $\frac{k \log k}{\epsilon^2}$ arises from an upper bound on the number of columns the optimal solution would choose (the existence result), and the remaining terms are contributed by the analysis of the greedy algorithm. The coherence parameter, $\mu(A)$ restricts the class of matrices for which the algorithm is useful. To the best of our knowledge, this is the first deterministic algorithm with $(1+\epsilon)$ approximation. Note that, in order to achieve this approximation ratio, we choose more than k columns. When $\mu = O(1)$, setting $\epsilon = \sqrt{k \log k}$ and ignoring logarithmic factors, we have a $1 + \sqrt{k \log k}$ approximation ratio with $O(k)$ columns.

We believe that a result without the coherence parameter should be possible, however have not been able to construct one. In any case, improving either the upper bound on the optimal reconstruction of the singular vectors, or improving the analysis of the greedy algorithm would yield a tighter result. The running time of the algorithm is governed by the computation of the top k singular vectors, which is $O(SVD(A_k))$ and the greedy selection phase, which is $O(mnc)$.

1.2 Comparison to Related Work

With the advent of massive data sets, much work in theoretical computer science has been spent on finding algorithms for matrix reconstruction by considering a careful choice of a subset of the columns of the data matrix. The seminal paper by Frieze, Kannan and Vempala [12] gives a randomized algorithm that chooses a subset of columns $C \in \mathbb{R}^{m\times c}$ of A such that $\|A - \Pi_C A\|_F \leq \|A - A_k\|_F + \epsilon\|A\|_F$, where Π_C is a projection matrix obtained by the SVD of C and $c = poly(k, 1/\epsilon, 1/\delta)$, where δ is the failure probability of the algorithm. Subsequent work [8, 7, 20] introduced several improvements on the dependence of c on $k, 1/\epsilon$ and $1/\delta$ also extending the analysis to the spectral norm. Recently, the effort has been towards eliminating the additive term in the inequality thereby yielding a relative approximation in the form $\|A - \Pi_C A\|_F \leq (1+\epsilon)\|A - A_k\|_F$. Along these lines, Deshpande et al. [5] first shows the existence of such approximations introducing a sampling technique related to the volume of the simplex defined by the column subsets of size k, without giving a polynomial time algorithm. Specifically, they show that there exists k columns with which one can get a $\sqrt{k+1}$ relative error approximation in Frobenius norm, which is tight. Later, Deshpande and Vempala [6] provides an algorithm with two steps which yields a relative approximation in expectation: first, approximate the "volume sampling" introduced in [5] by successively choosing one column at each step with carefully chosen probabilities; then, choose $O(k/\epsilon + k^2 \log k)$ columns in $O(k \log k)$ rounds in a similar fashion. The complexity of their algorithm is $O(M(k/\epsilon + k^2 \log k) + (m+n)poly(k,\epsilon))$, where M is the number of non-zero elements in A.

Recent result of Drineas et al. [10] provides two randomized algorithms fir relative error approximation in Frobenius norm using "subspace sampling", i.e. selecting columns proportional to the row-norms of the matrix of top k right singular vectors. One of the algorithms chooses exactly $c = O(k^2 \log(1/\delta)/\epsilon^2)$ columns; the other chooses $c = O(k \log k \log(1/\delta)/\epsilon^2)$ columns in expectation and both of them runs in $O(SVD(A_k))$ time, i.e. the time required to compute A_k, where δ is the failure probability. All of these algorithms exploit the power of randomization and they introduce a trade-off between the the number of columns chosen, the error parameter and the failure probability of the algorithm. The proof techniques presented in these papers break when the random sampling approach is sacrificed and a deterministic column selection procedure is used.

When it comes to deterministic reconstruction, no $(1+\epsilon)$ approximation algorithms are known. The linear algebra community has developed deterministic algorithms in the framework of *rank revealing QR (RRQR) factorizations* [1] which yield some approximation guarantees in spectral norm. Given a matrix $A \in \mathbb{R}^{n\times n}$, consider the QR factorization of the form

$$A\Pi = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix} \tag{1}$$

where $R_{11} \in \mathbb{R}^{k\times k}$ and $\Pi \in \mathbb{R}^{n\times n}$ is a permutation matrix. By the interlacing property of singular values (see [13]), $\sigma_k(R_{11}) \leq \sigma_k(A)$ and $\sigma_1(R_{22}) \geq \sigma_{k+1}(A)$. If the numerical rank of A is k, i.e. $\sigma_k(A) \gg \sigma_{k+1}(A)$, then one would like

to find a permutation Π for which $\sigma_k(R_{11})$ is sufficiently large and $\sigma_1(R_{22})$ is sufficiently small. A QR factorization is said to be a rank revealing QR (RRQR) factorization if $\sigma_k(R_{11}) \geq \sigma_k(A)/p(k,n)$ and $\sigma_1(R_{22}) \leq \sigma_{k+1}(A)p(k,n)$, where $p(k,n)$ is a low degree polynomial in k and n.

Much research on finding RRQR factorizations has yielded improved results for $p(k,n)$ [1, 2, 4, 14, 15, 19]. These algorithms make use of the *local maximum volume* concept and are generally complicated. Tight bounds for $p(k,n)$ can be used to give deterministic low rank matrix reconstruction with respect to the spectral norm, via the following simple fact.

Theorem 2 *Let Π_k be the matrix of first k columns of Π in (1). Then,*

$$\|A - (A\Pi_k)(A\Pi_k)^+ A\|_2 \leq p(k,n)\|A - A_k\|_2.$$

The best $p(k,n)$ was proposed by Gu and Eisenstat [14]. The authors show that there exists a permutation Π for which $p(k,n) = \sqrt{1+k(n-k)}$. It is not known whether such a permutation can be computed in polynomial time. Instead, algorithms with $p(k,n) = \sqrt{1+f^2k(n-k)}$ were given which run in $O((m+n\log_f n)n^2)$ time for $f > 1$ [14]. Hence, for constant f, the approximation ratio depends on n and the running time is $O(mn^2 + n^3 \log n)$. Note that, these algorithms consider choosing exactly k columns and the results are not directly comparable to ours as they provide bounds on the spectral norm. It is not clear whether these algorithmic results can be extended to give non-trivial bounds in Frobenius norm or to choose more than k columns so as to yield $(1+\epsilon)$ approximation.

Our results rely on a generalization of the sparse approximation problem which was formally proposed by Natarajan [18]: given $A \in \mathbb{R}^{m\times n}$, a vector $b \in \mathbb{R}^m$, and $\epsilon > 0$, find a vector $x \in \mathbb{R}^n$ satisfying $\|Ax - b\|_2 \leq \epsilon$ such that x has the fewest non-zero entries over all such vectors. This problem was also considered by Tropp [21]. Natarajan [18] proves that the problem is NP-hard and gives a greedy algorithm based on choosing the column vector from A with largest projection on b at each step. After correcting a minor technical error in his proof, his result gives that the greedy algorithm chooses at most $\lceil 18\, Opt(\epsilon/2)\mu^2(A)\ln(\|b\|_2/\epsilon)\rceil$ columns, $\mu(A)$ is a parameter defining the coherence between the normalized columns of A and $Opt(\epsilon/2)$ is the optimal number of vectors at error $\epsilon/2$. More recently, from an information theoretic point of view, Tropp [21] analyzed some previously known algorithms (e.g. *Matching Pursuit (MP)* [11, 17], *Basis Pursuit (BP)* [3]) for the sparse approximation problem, showing that these algorithms perform well for dictionaries (matrices) which are close to orthonormal. A formalization of this notion is represented by the *coherence parameter* [17], which is the maximum absolute inner product betweeen two distinct column vectors. Tropp gives a natural generalization of this concept, the *cumulative coherence parameter*, which is the maximum coherence between a fixed column vector and a collection of other column vectors. Intuitively, these parameters measure how "close" the column vectors of a matrix are and smaller values indicate an *incoherent* (almost orthonormal) matrix.

1.3 Notation and Preliminaries

From now on $A \in \mathbb{R}^{m\times n}$ is the matrix we wish to reconstruct. $A_{(i)}$ denotes the i^{th} row of A for $1 \leq i \leq m$, and $A^{(j)}$, the j^{th} column of A for $1 \leq j \leq n$. A_{ij} is the element at i^{th} row and the j^{th} column. Typically, we use C to denote a subset of columns of A, written $C \subset A$, i.e. C is a column submatrix of A. $span(C)$ denotes the subspace spanned by the column vectors in C. The Singular Value Decomposition of $A \in \mathbb{R}^{m\times n}$ of rank r is denoted by $A = U\Sigma V^T$ where $U \in \mathbb{R}^{m\times m}$ is the matrix of left singular vectors, $\Sigma \in \mathbb{R}^{m\times r}$ is the diagonal matrix containing the singular values of A in order, i.e. $\Sigma = (\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, and $V \in \mathbb{R}^{n\times n}$ is the matrix of right singular vectors. The "best" rank k approximation to A is $A_k = U_k\Sigma_k V_k$ where U_k, Σ_k, and V_k are the first k columns of the corresponding matrices in the full SVD of A. The pseudo-inverse of A is denoted by $A^+ = V\Sigma^+U^T$, where $\Sigma^+ = \left(\frac{1}{\sigma_1}, \ldots \frac{1}{\sigma_r}, 0, \ldots, 0\right)$. The Frobenius norm of A is $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2}$, and the spectral norm of A is $\|A\|_2 = \sigma_1(A)$. We also define the maximum column norm of a matrix A, $\|A\|_{col} = \max_{i=1}^{n}\{\|A^{(i)}\|_2\}$. $S^{\perp}$ is the space orthogonal to the space spanned by the vectors in S.

1.4 Organization of the paper

The rest of the paper is organized as follows. In Section 2, we define a generalized version of the sparse approximation problem which asks for a small set of columns that approximates the subspace spanned by a given set of target vectors. We give a greedy algorithm along with its analysis. Section 3 gives our column based rank matrix reconstruction algorithm, which can be viewed as a special case of the generalized sparse approximation problem, where the target vectors are the left singular vectors of A.

2 Generalized Sparse Approximation

Instead of seeking sparse approximation to a single vector [18], we propose the following generalization: given matrices $A \in \mathbb{R}^{m\times n}$, a set of vectors $B \in \mathbb{R}^{m\times k}$, and $\epsilon > 0$, find a matrix $X \in \mathbb{R}^{n\times k}$ satisfying

$$\|AX - B\|_F \leq \epsilon \tag{2}$$

such that $\sum_{i=1}^{n} \nu_i(X)$ is minimum over all possible choices of X, where $\nu_i(X) = 1$ if the row $X_{(i)}$ contains non-zero entries, $\nu_i(X) = 0$ if $X_{(i)} = \overrightarrow{0}$. Intuitively, the problem asks for a minimum number of set of column vectors of A whose span is close to those of B.

2.1 The Algorithm

A greedy strategy for solving this problem is to choose the column v from A at each iteration, for which $\|B^T v\|_2$ is maximum, and project the column vectors

of B and the other column vectors of A onto the space orthogonal to the chosen column. The algorithm proceeds greedily on these residual matrices until the norm of the residual B drops below the required threshold ϵ. Naturally, if the error ϵ cannot be attained, the algorithm will fail after selecting a maximal independent set of columns.

Greedy(A, B, ϵ)

1: normalize each column of A to have norm 1.
2: $l \leftarrow 0$, $\Lambda \leftarrow \emptyset$, $A_0 \leftarrow A$, $B_0 \leftarrow B$.
3: **while** $\|B_l\|_F > \epsilon$ **do**
4: choose $i \in \{1, \ldots, n\} - \Lambda$ such that $\|B_l^T A_l^{(i)}\|_2$ is maximum
5: $B_{l+1}^{(j)} \leftarrow B_l^{(j)} - \left({B_l^{(j)}}^T A_l^{(i)}\right) A_l^{(i)}$ for $i = 1, \ldots, k$, i.e. project $B_l^{(j)}$'s onto $\{A_l^{(i)}\}^\perp$.
6: $\Lambda \leftarrow \Lambda \cup \{i\}$.
7: $A_{l+1}^{(j)} \leftarrow A_l^{(j)} - \left({A_l^{(j)}}^T A_l^{(i)}\right) A_l^{(i)}$ for $j \in \{1, \ldots, n\} - \Lambda$, i.e. project $A_l^{(j)}$'s onto $\{A_l^{(i)}\}^\perp$.
8: normalize $A_{l+1}^{(j)}$ for $j \in \{1, \ldots, n\} - \Lambda$.
9: $l \leftarrow l + 1$.
10: **end while**
11: return $C = \Lambda(A)$, the selected columns.

Fig. 1. A greedy algorithm for Generalized Sparse Approximation

We first define the coherence of a matrix.

Definition 3 (Coherence) *The rank coherence of A, $\mu(A)$ is the inverse of the least singular value over all non-singular normalized column submatrices of A. Namely,*

$$\mu(A) = \max_{\substack{C \subseteq A \\ rank(C) = rank(A)}} \frac{1}{\sigma_r(C)}. \tag{3}$$

Remark 4 $1 \leq \mu(A) < \infty$. *Small values of $\mu(A)$ indicate a matrix with near orthonormal columns.*

Theorem 5 *The number of columns chosen by Greedy is at most*

$$O\left(Opt(\epsilon/2)\mu^2(A) \ln\left(\frac{\|B\|_F}{\epsilon}\right)\right)$$

where $Opt(\epsilon/2)$ is the optimal number of columns at error $\epsilon/2$.

We will establish Theorem 5 through a sequence of lemmas. The proof follows similar reasoning to the proof in [18]. Let t be the total number of iterations of

Greedy. At the beginning of the l^{th} iteration of the algorithm, for $0 \leq l < t$, let U_l be an optimal solution to the generalized sparse approximation problem with error parameter $\epsilon/2$, i.e. U_l minimizes $\sum_{i=1}^{n} \nu_i(X)$ over $X \in \mathbb{R}^{n \times k}$ such that $\|A_l U_l - B_l\|_F \leq \epsilon/2$, where $\nu_i(X) = 1$ if the row $X_{(i)}$ contains non-zero entries, $\nu_i(X) = 0$ if $X_{(i)} = \overrightarrow{0}$. Let $N_l = \sum_{i=1}^{n} \nu_i(U_l)$ and $Q_l = A_l U_l$. Define

$$\lambda = 4 \max_{0 \leq l < t} \frac{N_l \|U_l\|_F^2}{\|B_l\|_F^2}. \tag{4}$$

The proofs of the following lemmas which essentially bound the number of iterations of the algorithm, are given in the appendix. Assuming that the Greedy has not terminated, the first lemma states that the next step makes significant progress.

Lemma 6 *For the l^{th} iteration of Greedy, $\|B_l^T A_l\|_{col} \geq \frac{\|B_l\|_F^2}{2\sqrt{N_l}\|U_l\|_F}$.*

Thus, there exists a column in the residual A_l which will reduce the residual B_l significantly, because B_l has a large projection onto this column. Therefore, since every step of Greedy makes significant progress, there cannot be too many steps, which is the content of the next lemma.

Lemma 7 *$t \leq \left\lceil 2\lambda \ln\left(\frac{\|B\|_F}{\epsilon}\right)\right\rceil$, where t is the number of Greedy iterations.*

What remains is to bound λ. First, we will bound $\|U_l\|_F$ in terms of $\|B_l\|_F$ both of which appear in the expression for λ. Let $\sigma_l = \{i | U_{l(i)} \neq \overrightarrow{0}\}$ be the indices of rows of U_l which are not all zero. Recall that these indices denote which columns are chosen by the optimal solution for A_l. Let $\tau_l = \{i_1, i_2, \ldots, i_l\}$ be the indices of the first l columns picked by the algorithm. Given an index set γ, let the set of column vectors $\{A^{(i)} | i \in \gamma\}$ be denoted by $\gamma(A)$. The proofs of the following lemmas are also in the appendix.

Lemma 8 *$\sigma_l(A) \cup \tau_l(A)$ is a linearly independent set for all $l \geq 0$.*

Lemma 9 *For $0 \leq l < t$, $\|U_l\|_F \leq \frac{3}{2}\mu(A)\|B_l\|_F$.*

Proof of Theorem 5: First, we note that the number of non-zero rows in the optimal solution is non-increasing as the algorithm proceeds, that is $N_l \geq N_{l+1}$ for $l > 0$, which follows from an argument identical to the proof of Lemma 3 in [18]. Since $Opt(\epsilon/2) = N_0$, we have

$$\lambda \leq 4 \max_{0 \leq l < t} \frac{N_0 \|U_l\|_F^2}{\|B_l\|_F^2} \leq 9 Opt(\epsilon/2)\mu^2(A)$$

where the last inequality is due to the result of Lemma 9. Combining this with Lemma 7, we have that the number of iterations of the algorithm is bounded by

$$t \leq \left\lceil 18 Opt(\epsilon/2)\mu^2(A) \ln\left(\frac{\|B\|_F}{\epsilon}\right)\right\rceil$$

3 Deterministic Low-Rank Matrix Reconstruction

In this section, we give a deterministic algorithm for low rank matrix reconstruction based on the greedy approach that we have introduced and analyzed for the generalized sparse approximation problem:

LowRankApproximation(A, k)
1: compute U_k and Σ_k of A
2: return Greedy(A, $U_k\Sigma_k$, $\epsilon\|A - A_k\|_F$)

Fig. 2. The low-rank approximation algorithm

The algorithm first computes U_k, the top k left singular vectors of A and Σ_k the first k singular values of A, which can be performed by standard methods like Lanczos. The columns of A are then selected in a greedy fashion so as to "fit" them to the subspace spanned by the columns of $U_k\Sigma_k$. Intuitively, we select columns of A which are close to the columns of $U_k\Sigma_k$ and the analysis shows that the submatrix C of A we obtain is provably close to the "best" rank-k approximation to A. The error parameter which is given as an input to the greedy algorithm is $\epsilon\|A - A_k\|_F$. The following result provides an upper bound on the number of columns of the optimal solution at error $\epsilon\|A - A_k\|_F/2$.

Lemma 10 *There exists a column submatrix C of A with $c = O(k \log k/\epsilon^2)$ columns such that $\|U_k\Sigma_k - CC^+U_k\Sigma_k\|_F \leq \epsilon\|A - A_k\|_F/2$.*

Proof. The proof is given in the appendix due to space limitations.

We now, give the proof of Theorem 1.

Proof of Theorem 1: By the algorithm, we have

$$U_k\Sigma_k = C_{gr}C_{gr}^+U_k\Sigma_k + E.$$

for some generic error matrix E satisfying $\|E\|_F \leq \epsilon\|A - A_k\|_F$. Multiplying both sides by V_k^T, we get

$$A_k = C_{gr}C_{gr}^+A_k + EV_k^T,$$

Hence, $A - C_{gr}C_{gr}^+A_k = A - A_k + EV_k^T$. Taking norms of both sides, and noting that $\|V_k\|_F = \sqrt{k}$, and C_{gr}^+A is the minimizer of $\|A - C_{gr}X\|_F$, we obtain

$$\begin{aligned}\|A - C_{gr}C_{gr}^+A\|_F &\leq \|A - C_{gr}C_{gr}^+A_k\|_F \\ &\leq \|A - A_k\|_F + \epsilon\sqrt{k}\|A - A_k\|_F \\ &= (1 + \epsilon\sqrt{k})\|A - A_k\|_F\end{aligned}$$

Choosing an error parameter $\epsilon' = \epsilon/\sqrt{k}$ and combining Theorem 5 and Lemma 10 gives the desired result.

Note that, the number of columns chosen by the algorithm depends on $\mu(A)$, i.e. the structure of A. To get an idea of what this result implies when the number of columns chosen is of order k, we give the following corollary, which immediately follows upon a careful choice of error parameter.

Corollary 11 *The greedy algorithm chooses a submatrix C of $\tilde{O}(k)$ columns of A for which $\|A - CC^+A\|_F \leq \mu(A)\sqrt{k \log k}\|A - A_k\|_F$.*

Acknowledgments: We would like to thank Petros Drineas for helpful discussions.

References

1. T. F. Chan. Rank revealing QR factorizations. *Linear Algebra Appl.*, (88/89):67–82, 1987.
2. S. Chandrasekaran and I. C. F. Ipsen. On rank-revealing factorizations. *SIAM J. Matrix Anal. Appl.*, 15:592–622, 1994.
3. S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
4. F. R. de Hoog and R. M. M. Mattheijb. Subset selection for matrices. *Linear Algebra and its Applications*, (422):349–359, 2007.
5. A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *SODA '06*, pages 1117–1126. ACM Press, 2006.
6. A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. In *RANDOM'06*, pages 292–303. Springer, 2006.
7. P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *SODA '99: Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, pages 291–299. SIAM, 1999.
8. P. Drineas, R. Kannan, and M. W. Mahoney. Fast monte carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36(1):158–183, 2006.
9. P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-based methods. In *APPROX-RANDOM*, pages 316–326, 2006.
10. P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: column-row-based methods. In *ESA'06: Proceedings of the 14th conference on Annual European Symposium*, pages 304–314. Springer-Verlag, 2006.
11. J. H. Friedman and W. Stuetzle. Projection pursuit regressions. *J. Amer. Statist. Soc.*, 76:817–823, 1981.
12. A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the Association for Computing Machinery*, 51(6):1025–1041, 2004.
13. G. H. Golub and C. V. Loan. *Matrix Computations*. Johns Hopkins U. Press, 1996.

14. M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.
15. Y. P. Hong and C. T. Pan. Rank-revealing QR factorizations and the singular value decomposition. *Mathematics of Computation*, 58:213–232, 1992.
16. F. G. Kuruvilla, P. J. Park, and S. L. Schreiber. Vector algebra in the analysis of genome-wide expression data. *Genome Biology*, (3), 2002.
17. S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.
18. B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.
19. C. T. Pan and P. T. P. Tang. Bounds on singular values revealed by QR factorizations. *BIT Numerical Mathematics*, 39:740–756, 1999.
20. M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4), 2007.
21. J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

Appendix

Proof of Lemma 6 : Let $E \in \mathbb{R}^{m\times k}$ be a generic error matrix such that $\|E\|_F \leq \epsilon/2$, and Let $\|E^{(j)}\|_2 = \epsilon_j/2$ for $i = 1, \ldots, k$. Hence, $\sum_{i=1}^{k} \epsilon_j^2 \leq \epsilon^2$. Now, we can write $B_l^{(j)} = \left(\sum_{i=1}^{n} A_l^{(i)} U_{lij}\right) + E^{(j)}$ for $j = 1, \ldots, k$. Then,

$$\|B_l\|_F^2 = \sum_{j=1}^{k} {B_l^{(j)}}^T B_l^{(j)} = \sum_{j=1}^{k}\sum_{i=1}^{n} U_{lij} {B_l^{(j)}}^T A_l^{(i)} + \sum_{j=1}^{k} {B_l^{(j)}}^T E^{(j)} \tag{5}$$

We will first bound the double summation in the above expression.

$$\begin{aligned}
\sum_{j=1}^{k}\sum_{i=1}^{n} U_{lij} {B_l^{(j)}}^T A_l^{(i)} &\leq \sum_{i=1}^{n} \left(\left(\sum_{j=1}^{k} U_{lij}^2 \right)^{1/2} \left(\sum_{j=1}^{k} \left({B_l^{(j)}}^T A_l^{(i)} \right)^2 \right)^{1/2} \right) \\
&\leq \max_{1\leq i\leq n} \left\{ \left(\sum_{j=1}^{k} \left({B_l^{(j)}}^T A_l^{(i)} \right)^2 \right)^{1/2} \right\} \sum_{i=1}^{n} \left(\sum_{j=1}^{k} U_{lij}^2 \right)^{1/2} \\
&\leq \|B_l^T A_l\|_{col} \sqrt{N_l} \|U_l\|_F
\end{aligned}$$

The first line is due to Cauchy-Schwartz inequality. The last inequality bounds the double summation in the second line as follows. Define n dimensional vectors a and b such that $a_i = \left(\sum_{j=1}^{k} U_{lij}^2\right)^{1/2}$ and $b_i = 1$ if there exists a non-zero entry in the i^{th} row of U_l, $b_i = 0$ if all the elements in the i^{th} row of U_l are zero, for $i = 1, \ldots, n$. Then, applying Cauchy-Schwartz inequality to a and b, we obtain $\sum_{i=1}^{n} \left(\sum_{j=1}^{k} U_{lij}^2\right)^{1/2} = \sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \left(\sum_{i=1}^{n} b_i^2\right)^{1/2}$. Since $\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n}\sum_{j=1}^{k} U_{lij}^2 = {\|U_l\|_F}^2$, and $\sum_{i=1}^{n} b_i^2 = N_l$, we have that $\sum_{i=1}^{n} \left(\sum_{j=1}^{k} U_{lij}^2\right)^{\frac{1}{2}} \leq \sqrt{N_l}\|U_l\|_F$.

We will now bound the second term in (5).

$$\begin{aligned}
\sum_{j=1}^{k} {B_l^{(j)}}^T E^{(j)} &\leq \sum_{j=1}^{k} \|{B_l^{(j)}}^T\|_2 \|E^{(j)}\|_2 \quad (Cauchy - Schwartz) \\
&= \frac{1}{2} \sum_{j=1}^{k} \epsilon_j \|{B_l^{(j)}}^T\|_2 \\
&\leq \frac{1}{2} \left(\sum_{j=1}^{k} \epsilon_j^2 \right)^{1/2} \left(\sum_{j=1}^{k} \|{B_l^{(j)}}^T\|_2^2 \right)^{1/2} \qquad (Cauchy - Scwartz) \\
&\leq \frac{1}{2} \epsilon \|B_l\|_F \\
&\leq \frac{1}{2} \|B_l\|_F^2
\end{aligned}$$

where the last inequality is due to the fact that $\|B_l\|_F > \epsilon$, i.e. the algorithm is still running.

Combining these bounds in (5), we have $\|B_l\|_F^2 \leq \|B_l^T A_l\|_{col} \sqrt{N_l} \|U_l\|_F + 1/2 \|B_l\|_F^2$, which gives $\|B_l\|_F^2 \leq 2 \|B_l^T A_l\|_{col} \sqrt{N_l} \|U_l\|_F$. The lemma then immediately follows.

Proof of Lemma 7 : Let i be the index of the chosen column at step l and let j be a column index of B. Then, by the execution of the algorithm, $B_{l+1}^{(j)} = B_l^{(j)} - \left({B_l^{(j)}}^T A_l^{(i)} \right) A_l^{(i)}$. Since $B_{l+1}^{(j)}$ is orthogonal to $A_l^{(i)}$ and $\|A_l^{(i)}\|_2 = 1$, we can write $\|B_{l+1}^{(j)}\|_2^2 = \|B_l^{(j)}\|_2^2 - |{B_l^{(j)}}^T A_l^{(i)}|^2$. Summing over all column indices of B_{l+1}, we obtain

$$\begin{aligned}
\|B_{l+1}\|_F^2 = \sum_{j=1}^{k} \|B_{l+1}^{(j)}\|_2^2 &= \sum_{j=1}^{k} \|B_l^{(j)}\|_2^2 - \sum_{j=1}^{k} |{B_l^{(j)}}^T A_l^{(i)}|^2 \\
&= \|B_l\|_F^2 - \|B_l^T A_l^{(i)}\|_2^2 \\
&= \|B_l\|_F^2 - \|B_l^T A_l\|_{col}^2 \\
&\leq \|B_l\|_F^2 - \frac{\|B_l\|_F^4}{4 N_l \|U_l\|_F^2} \quad (Lemma\,6) \\
&= \|B_l\|_F^2 \left(1 - \frac{1}{\lambda} \right) \quad (Equation\,(4))
\end{aligned}$$

where the third line follows since the algorithm chooses i to maximize $\|B_l^T A_l^{(i)}\|_2$. Hence, $\|B_l\|_F^2 \leq (1 - 1/\lambda) \|B_0\|_F^2$. Since the algorithm stops when $\|B_t\|_F^2 \leq \epsilon^2$, it suffices for t to satisy $(1 - 1/\lambda)^t \|B_0\|_F^2 \leq \epsilon^2$. Rearranging, and taking logarithms

we obtain $t\ln(1-1/\lambda) \leq \ln(\epsilon^2/\|B_0\|_F^2)$. Since $\ln(1-1/\lambda) \leq -1/\lambda$, we get that $t \geq \lambda \ln(\|B\|_F^2/\epsilon^2) = 2\lambda \ln(\|B\|_F/\epsilon)$ iterations are enough for Greedy to terminate.

Proof of Lemma 8 : Note that for $l = 0$, we only have $\sigma_0(A)$ and by the definition of the optimality of U_0, this set should be linearly independent. For $l \geq 1$, we will argue by contradiction. Assume that the given set, $\sigma_l(A) \cup \tau_l(A)$ is not a linearly independent set. Hence, some linear combination of some vectors from the set sum to 0. Since, by the execution of the algorithm, $\tau_l(A)$ is a linearly independent set, at least one of these vectors should be from $\sigma_l(A)$, and this vector u can be written as a linear combination of some other vectors in $\sigma_l(A) \cup \tau_l(A)$. To this end, recall that σ_l denotes the indices of columns of A_l chosen by the optimal solution U_l, and $\sigma_l(A)$ is the set of columns of A with these indices. Consider a column vector v in $\sigma_l(A)$. According to the algorithm, at the end of the l^{th} iteration, the residual vector v_l (which is in $\sigma_l(A_l)$) is precisely the projection of v onto the space orthogonal to the vectors chosen by the algorithm, namely $\tau_l(A)$. Since this is the case for all possible v's, we have that $\sigma_l(A_l)$ is the projection of $\sigma_l(A)$ onto the space orthogonal to $\tau_l(A)$. Hence, according to our last assumption, u_l which is the projection of u onto the space orthogonal to $\tau_l(A)$ can be expressed as a linear combination of some other vectors in $\sigma_l(A_l)$ since no vector from $\tau_l(A)$ can contribute in the expansion of u_l. This contradicts the optimality of U_l, i.e. that the number of columns it "selects" from A_l is the fewest among all possible choices.

Proof of Lemma 9 Consider the column indices $\{i_1, i_2, \ldots, i_l\}$ of the first l vectors chosen by the algorithm. Specifically, let $\tau_l(A_l) = \{A_l^{(i_1)}, A_l^{(i_2)}, \ldots, A_l^{(i_l)}\}$ be the columns in A_l chosen by the algorithm in the order selected. Note that these vectors are orthogonal. At the end of the l^{th} iteration of the algorithm, for $i \in \sigma_l$, we can write

$$A_l^{(i)} = \frac{A_{l-1}^{(i)} - v_l^{(i)}}{\sqrt{1 - \|v_l^{(i)}\|_2^2}} \tag{6}$$

where $v_l^{(i)}$ is in the span of $A_l^{(i_l)}$. Similarly, we can express $A_{l-1}^{(i)}$ in terms of $A_{l-2}^{(i)}$, i.e.

$$A_{l-1}^{(i)} = \frac{A_{l-2}^{(i)} - v_{l-1}^{(i)}}{\sqrt{1 - \|v_{l-1}^{(i)}\|_2^2}}$$

where $v_{l-1}^{(i)}$ is in the span of $A_l^{(i_{l-1})}$. Note that, since the vectors in $\tau_l(A_l)$ are orthogonal, we have $\|v_l^{(i)} + v_{l-1}^{(i)}\|_2^2 = \|v_l^{(i)}\|_2^2 + \|v_{l-1}^{(i)}\|_2^2$. Using this, we can recursively express $A_l^{(i)}$ in (6) as

$$A_l^{(i)} = \frac{A^{(i)} - v^{(i)}}{\sqrt{1 - \|v^{(i)}\|_2^2}} \tag{7}$$

for some $v^{(i)} \in span(\tau_l(A))$. (Note that $span(\tau_l(A_l) = span(\tau_l(A_0) = span(\tau_l(A)$ and the columns of A are normalized). Thus, noting that $Q_l^{(j)} = \sum_{i \in \sigma_l} A_l^{(i)} U_{lij}$, and $v^{(i)}$ can be expressed as a linear combination of the column vectors of $\tau_l(A)$, we have

$$Q_l^{(j)} = \sum_{i \in \sigma_l} U_{lij} \frac{A^{(i)} - v^{(i)}}{\sqrt{1 - \|v^{(i)}\|_2^2}} = \sum_{i \in \sigma_l} \frac{U_{lij}}{\sqrt{1 - \|v^{(i)}\|_2^2}} A^{(i)} + \sum_{i \in \tau_l} \delta_i A^{(i)} \tag{8}$$

where δ_i's are appropriate coefficients in the expansion of $v^{(i)}$. Now, let S_l be the matrix of the columns from $\sigma_l(A) \cup \tau_l(A)$. Note that, S_l is a column submatrix of A which has full rank by Lemma 8. Since S_l is a linearly independent set, Q_l has a unique expansion in the basis S_l given by $W_l = S_l^+ Q_l$. Specifically, for $i \in \sigma_l$, $W_{lij} = U_{lij} / \sqrt{1 - \|v^{(i)}\|_2^2}$, and for $i \in \tau_l$, $W_{lij} = \delta_i$. Since $\sqrt{1 - \|v^{(i)}\|_2^2} < 1$, $|U_{lij}| \leq |W_{lij}|$ for $i \in \sigma_l$. For $i \in \tau_l$, we have $U_{lij} = 0$ and hence trivially $|U_{lij}| \leq |W_{lij}|$. Applying this inequality to the j^{th} column of U_l, we obtain $\|U_l^{(j)}\|_2 \leq \|W_l{}^{(j)}\|_2 \leq \|S_l^+\|_2 \|Q_l^{(j)}\|_2$. The last inequality is due to sub-multiplicativity of the spectral norm. Noting that $Q_l^{(j)} = B_l^{(j)} + E^{(j)}$, where E is a generic error matrix with $\|E\|_F \leq \epsilon/2$, and hence $\sum_{j=1}^k \|E^{(j)}\|_2^2 \leq \epsilon^2/4$, we obtain

$$
\begin{aligned}
\|U_l\|_F^2 &= \sum_{j=1}^{k} \|U_l^{(j)}\|_2^2 \\
&\leq \|S_l^+\|_2^2 \sum_{j=1}^{k} \|Q_l^{(j)}\|_2^2 \\
&\leq \|S_l^+\|_2^2 \sum_{j=1}^{k} \left(\|B_l^{(j)} + E^{(j)}\|_2^2\right) \\
&\leq \|S_l^+\|_2^2 \sum_{j=1}^{k} \left(\|B_l^{(j)}\|_2 + \|E^{(j)}\|_2\right)^2 \quad (Triangle\ Inequality) \\
&= \|S_l^+\|_2^2 \left(\sum_{j=1}^{k} \|B_l^{(j)}\|_2^2 + \sum_{j=1}^{k} \|E^{(j)}\|_2^2 + 2\sum_{j=1}^{k} \|B_l^{(j)}\|_2 \|E^{(j)}\|_2\right) \\
&\leq \|S_l^+\|_2^2 \left(\|B_l\|_F^2 + \frac{\epsilon^2}{4} + 2\sum_{j=1}^{k} \|B_l^{(j)}\|_2 \|E^{(j)}\|_2\right) \\
&\leq \|S_l^+\|_2^2 \left(\frac{5}{4}\|B_l\|_F^2 + 2\sum_{j=1}^{k} \|B_l^{(j)}\|_2 \|E^{(j)}\|_2\right) \quad (\|B_l\|_F > \epsilon)
\end{aligned}
$$

Applying Cauchy-Schwartz inequality to the second term in the parantheses, we obtain

$$
\begin{aligned}
\|U_l\|_F^2 &\leq \|S_l^+\|_2^2 \left(\frac{5}{4}\|B_l\|_F^2 + 2\left(\sum_{j=1}^{k} \|B_l^{(j)}\|_2^2\right)^{1/2} \left(\sum_{j=1}^{k} \|E^{(j)}\|_2^2\right)^{1/2}\right) \\
&= \|S_l^+\|_2^2 \left(\frac{5}{4}\|B_l\|_F^2 + 2\|B_l\|_F \|E\|_F\right) \\
&\leq \|S_l^+\|_2^2 \left(\frac{5}{4}\|B_l\|_F^2 + \epsilon\|B_l\|_F\right) \quad (\|E\|_F \leq \epsilon/2) \\
&\leq \|S_l^+\|_2^2 \left(\frac{5}{4}\|B_l\|_F^2 + \|B_l\|_F^2\right) \quad (\|B_l\|_F > \epsilon) \\
&= \frac{9}{4}\|S_l^+\|_2^2 \|B_l\|_F^2.
\end{aligned}
$$

Hence, we have $\|U_l\|_F \leq \frac{3}{2}\|S_l^+\|_2 \|B_l\|_F$. Now, note that the rank of S_l is less than or equal to r, the rank of A. S_l can be obtained by deleting columns of a full-rank submatrix Z of A, which has exactly r columns. $\|S_l^+\|_2$, which is the inverse of

the least singular value of S_l is smaller than that of such a matrix Z (see [13]). Then, by the definition of $\mu(A)$, we clearly have $\|S_l^+\|_2 \leq \|Z^+\|_2 \leq \mu(A)$ and the lemma follows.

Proof of Lemma 10 : We will make use of the following result which is proved in [9]. They give a randomized algorithm which constructs, with non-zero probability a set of columns with a particular approximation property. This immediately translates to an existence result. For a set of columns $C \in A$, denote the sampling matrix which selects the columns by S so that $C = AS$. Let V_k be the matrix of the first k right singular vectors of A. Let V_{r-k} be the matrix containing the last $r-k$ right singular vectors of A, and let Σ_k and Σ_{r-k} be the diagonal matrices containing the first k and the last $r-k$ singular values of A.

Theorem 12 ([9]) *There exists a set of $c = O(k \log k/\epsilon^2)$ columns from A and corresponding sampling matrix S, with $C = AS$ such that $rank(V_k^T S) = rank(V_k)$, $\|\Sigma_{r-k} V_{r-k}^T S (V_k^T S)^+\|_F \leq \epsilon \|A - A_k\|_F$ where Σ_{r-k} is the diagonal matrix containing the smallest $r-k$ singular values of A, and V_{r-k} is the matrix containing the last $r-k$ right singular vectors of A.*

Let $C = AS$ be the column sub-matrix whose existence is guaranteed by the theorem above. We have

$$\begin{aligned}
\epsilon^2 \|A - A_k\|_F^2 &\geq \|\Sigma_{r-k} V_{r-k}^T S (V_k^T S)^+\|_F^2 \\
&= \|\Sigma_k - \Sigma_k V_k^T S (V_k^T S)^+\|_F^2 + \|\Sigma_{r-k} V_{r-k}^T S (V_k^T S)^+\|_F^2
\end{aligned}$$

where the first term in the last expression is just 0 as $V_k^T S (V_k^T S)^+ = I_k$. Combining the last two terms into one expression, we have

$$\begin{aligned}
\epsilon^2 \|A - A_k\|_F^2 &\geq \left\| \begin{pmatrix} \Sigma_k \\ 0 \end{pmatrix} - \begin{pmatrix} \Sigma_k V_k^T \\ \Sigma_{r-k} V_{r-k}^T \end{pmatrix} S (V_k^T S)^+ \right\|_F^2 \\
&= \left\| \begin{pmatrix} \Sigma_k \\ 0 \end{pmatrix} - \begin{pmatrix} \Sigma_k & 0 \\ 0 & \Sigma_{r-k} \end{pmatrix} \begin{pmatrix} V_k^T \\ V_{r-k}^T \end{pmatrix} S (V_k^T S)^+ \right\|_F^2 \\
&= \left\| \begin{pmatrix} \Sigma_k \\ 0 \end{pmatrix} - (\Sigma V^T S)(\Sigma_k V_k^T S)^+ \Sigma_k \right\|_F^2 \\
&= \left\| \begin{pmatrix} \Sigma_k \\ 0 \end{pmatrix} - (\Sigma V^T S) Y \right\|_F^2
\end{aligned}$$

where $Y = (\Sigma_k V_k^T S)^+ \Sigma_k$. Let A, B be arbitrary matrices. Then, $\min_X \|A - BX\|_F^2 = \|A - BB^+ A\|_F^2$ (see [13]). Hence, we continue as follows,

$$
\begin{aligned}
\left\| \begin{pmatrix} \Sigma_k \\ 0 \end{pmatrix} - (\Sigma V^T S) Y \right\|_F^2 &\geq \min_{X \in \mathbb{R}^{c \times k}} \left\| \begin{pmatrix} \Sigma_k \\ 0 \end{pmatrix} - (\Sigma V^T S) X \right\|_F^2 \\
&= \left\| \begin{pmatrix} \Sigma_k \\ 0 \end{pmatrix} - (\Sigma V^T S)(\Sigma V^T S)^+ \begin{pmatrix} \Sigma_k \\ 0 \end{pmatrix} \right\|_F^2 \\
&= \left\| \begin{pmatrix} I_k \\ 0 \end{pmatrix} \Sigma_k - (\Sigma V^T S)(\Sigma V^T S)^+ \begin{pmatrix} I_k \\ 0 \end{pmatrix} \Sigma_k \right\|_F^2 \\
&= \left\| U \begin{pmatrix} I_k \\ 0 \end{pmatrix} \Sigma_k - (U \Sigma V^T S)(\Sigma V^T S)^+ U^T U_k \Sigma_k \right\|_F^2 \\
&= \left\| U_k \Sigma_k - (U \Sigma V^T S)(U \Sigma V^T S)^+ U_k \Sigma_k \right\|_F^2 \\
&= \left\| U_k \Sigma_k - C C^+ U_k \Sigma_k \right\|_F^2
\end{aligned}
$$

where we have used $U \Sigma V^T = A$ and $C = AS$. Choosing an error parameter $\epsilon' = \epsilon/2$ gives the desired result.