Theoretical Computer Science 421 (2012) 1–14

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Column subset selection via sparse approximation of SVD

A. Çivril [a,*], M. Magdon-Ismail [b]

[a] Meliksah University, Computer Engineering Department, Talas, Kayseri 38280, Turkey
[b] Rensselaer Polytechnic Institute, Computer Science Department, 110 8th Street Troy, NY 12180-3590, USA

ARTICLE INFO

Article history:
Received 7 February 2011
Received in revised form 19 August 2011
Accepted 17 November 2011
Communicated V.Y. Pan

Keywords:
Subset selection
SVD
Sparse approximation

ABSTRACT

Given a real matrix $A \in \mathbb{R}^{m\times n}$ of rank r, and an integer $k < r$, the sum of the outer products of top k singular vectors scaled by the corresponding singular values provide the best rank-k approximation A_k to A. When the columns of A have specific meaning, it might be desirable to find good approximations to A_k which use a small number of columns of A. This paper provides a simple greedy algorithm for this problem in Frobenius norm, with guarantees on the performance and the number of columns chosen. The algorithm selects c columns from A with $c = \tilde{O}\left(\frac{k\log k}{\epsilon^2}\eta^2(A)\right)$ such that

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

where C is the matrix composed of the c columns, Π_C is the matrix projecting the columns of A onto the space spanned by C and $\eta(A)$ is a measure related to the *coherence* in the normalized columns of A. The algorithm is quite intuitive and is obtained by combining a greedy solution to the generalization of the well known *sparse approximation problem* and an *existence* result on the possibility of sparse approximation. We provide empirical results on various specially constructed matrices comparing our algorithm with the previous deterministic approaches based on QR factorizations and a recently proposed randomized algorithm. The results indicate that in practice, the performance of the algorithm can be significantly better than the bounds suggest.

1. Introduction

The usual approach to find a "good" subspace that approximates the column span of a matrix A is to take the best rank k approximation $A_k = \sum_{i=1}^{k}\sigma_i u_i v_i^T$, which minimizes the residual error with respect to any unitarily invariant norm. In some application areas such as statistical data analysis, this approach might be undesirable since the singular vector representation is not suitable to make inferences about the actual underlying data; because they are generally combinations of all the columns of A. An example of this is the micro-array data where the combinations of the column vectors have no sensible interpretation [25]. Hence, it is of practical importance to find an approximation to A_k which is composed of a small number of columns of A. The problem also bears a theoretical importance in the sense that one might want to know how well the column vectors of a matrix can represent its spectrum. This paper considers the problem of finding a small number of columns of a matrix A such that the expression $\|A - \Pi_C A\|_F$ is close to $\|A - A_k\|_F$, for a given number $k < r = rank(A)$ where $\Pi_C = CC^+$ is the matrix projecting the columns of A onto the space spanned by the columns of C.

[*] Corresponding author. Tel.: +90 541 513 6385.
E-mail addresses: acivril@meliksah.edu.tr, alicivril@yahoo.com, alicivril@gmail.com (A. Çivril), magdon@cs.rpi.edu (M. Magdon-Ismail).

0304-3975/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2011.11.019

We give a deterministic greedy algorithm for this problem which is based on the *sparse approximation* of the SVD of A. We first generalize the problem of sparse approximation [11,27] to one of approximating a subspace. This is conceptually the same problem with the one so-called *simultaneous sparse approximation* in signal processing and approximation theory in Hilbert spaces (e.g. [26,34]). We then propose and analyze a greedy algorithm for this problem and derive our main result in the special case where the subspace to be approximated is the space spanned by the first k left singular vectors of A. In words, the 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 characteristics 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 combine an existence result on the quality of the optimal columns with the analysis of the greedy algorithm to arrive at the following result:

Theorem 1.1. *Given a matrix $A \in \mathbb{R}^{m\times n}$, an integer $k < r = rank(A)$ and $\epsilon < \frac{\|A_k\|_F}{\|A-A_k\|_F}$, there exists a polynomial-time algorithm which selects a column sub-matrix $C \in \mathbb{R}^{m\times c}$ of A with $c = O\left(\frac{k\log k}{\epsilon^2}\eta^2(A)\ln\left(\frac{\|A_k\|_F}{\epsilon\|A-A_k\|_F}\right)\right)$ columns such that*

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

where $\eta(A)$ is a measure related to the coherence in the normalized columns of A.

The requirement on ϵ is to make sure that the expression with the natural logarithm is meaningful. The term $\eta(A)$ arises from the analysis of the generalized sparse approximation problem. In our analysis, the possibility of eliminating this parameter or replacing it with a low order polynomial in k and ϵ would yield a much more desirable result. We would like to note that such input-dependent parameters naturally arise in the analysis of sparse approximation problems [34,35]. Yet, considering the special nature of the subspace we wish to approximate, we think an improvement is possible.

1.1. Related work

The theoretical computer science community has investigated the low-rank matrix approximation problem which asks for a k-dimensional subspace that approximates A_k in the spectral and Frobenius norm. The solutions developed thus far have mostly focused on randomized algorithms, and the set of columns chosen by these algorithms have more than k columns which is proven to contain an arbitrarily close approximation to A_k. This approximation has the nice property of having the same dimensionality with that of A_k, but cannot directly be interpreted in terms of the columns of A. The numerical linear algebra community on the other hand, implicitly provides deterministic solutions for approximating A_k in the context of rank revealing QR factorizations, which primarily aim to determine the numerical rank of A. The algorithms developed in this framework usually focus on spectral norm and they select exactly k columns providing approximations as a function of k and n. The algorithm we provide has hybrid features in the sense that it is deterministic, and the error ratio drops with increasing number of selected columns.

The seminal paper by Frieze et al. [19] gives a randomized algorithm that selects a subset of columns $C \in \mathbb{R}^{m\times c}$ of A such that $\|A - \Pi_C A\|_F \le \|A - A_k\|_F + \epsilon\|A\|_F$, where Π_C is a projection matrix obtained by the truncated SVD of C and c is a polynomial in k, $(1/\epsilon)$ and $(1/\delta)$, where δ is the failure probability of the algorithm. Subsequent work [15,16] introduced several improvements on the dependence of c on k, $1/\epsilon$ and $1/\delta$ also extending the analysis to the spectral norm, while Rudelson and Vershynin [29,30], provided results of the form $\|A - \Pi_C A\|_2 \le \|A - A_k\|_2 + \epsilon\sqrt{\|A\|_2\|A\|_F}$. Recently, the effort has been toward eliminating the additive term in the inequality thereby yielding a relative approximation in the form $\|A - \Pi_C A\|_F \le (1+\epsilon)\|A - A_k\|_F$. Along these lines, Deshpande and Vempala [14] and Drineas et al. [17] provided algorithms with different sampling schemes attaining the $(1+\epsilon)$ approximation where the number of columns is a function of k and ϵ. Other recent approaches for the problem we consider includes random projections [31], and sampling which exploits geometric properties of high dimensional spaces [32]. [13] also considers the subspace approximation problem in general l_p norms. Achlioptas and McSherry approaches the problem by zero-ing out and quantizing the individual elements of the matrix randomly [1]. All of these algorithms exploit the power of randomization and they introduce a trade-off between the number of columns chosen, the error parameter and the failure probability of the algorithm.

During the submission of this paper, a deterministic algorithm for matrix reconstruction was proposed by Guruswami and Sinop [22] based on carefully implementing a scheme akin to volume sampling. Their algorithm uses optimal number of columns which does not involve $\eta(A)$, but the running time is $O(m^\omega nr\log m)$ where ω is the exponent in matrix multiplication. Compared to [22], the algorithm we present in this paper is less sophisticated and more intuitive.

The linear algebra community has developed deterministic algorithms in the framework of *rank revealing QR (RRQR) factorizations* [7] 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

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