FAST EXEMPLAR SELECTION ALGORITHM FOR MATRIX APPROXIMATION AND REPRESENTATION: A VARIANT oASIS ALGORITHM

V. Abrol, P. Sharma, A. K. Sao

School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi

ABSTRACT

Extracting inherent patterns from large data using decompositions of data matrix by a sampled subset of exemplars has found many applications in machine learning. We propose a computationally efficient algorithm for adaptive exemplar sampling, called fast exemplar selection (FES). The proposed algorithm can be seen as an efficient variant of the oASIS algorithm [1]. FES iteratively selects incoherent exemplars based on the exemplars that are already sampled. This is done by ensuring that the selected exemplars forms a positive definite Gram matrix which is checked by exploiting its Cholesky factorization in an incremental manner. FES is a deterministic rank revealing algorithm delivering a tighter matrix approximation bound. Further, FES can also be used to exactly represent low rank matrices and signals sampled from a unions of independent subspaces. Experimental results show that FES performs comparable to existing methods for tasks such as matrix approximation, feature selection, outlier detection, and clustering.

***Index Terms*—** Matrix factorization, exemplar selection, low rank approximation, sparse coding.

1. INTRODUCTION

Exemplar selection (ES) aims at finding a subspace (composed of a small number of exemplars) that approximates the column span of a matrix $\mathbf{X} \in \mathbb{R}^{n\times l}$ [2]. From theoretical perspective the aim is to know how well the column vectors of a matrix can represent its spectrum, as this can further help in summarizing and visualizing large datasets of natural scenes, objects, faces, videos, and text [3, 4]. Further, as opposed to a larger dataset one can improve the efficiency, memory requirement and computational time of e.g., classification and clustering algorithms by working on a reduced number of representative exemplars [5]. This paper considers the problem of sampling a small number of columns of a matrix $\mathbf{X}$ such that $\|\mathbf{X} - \mathbf{\Pi X}\|_F$ is close to $\|\mathbf{X} - \mathbf{X}_k\|_F$ i.e., error between the target matrix $\mathbf{X}$ and its rank-k approximation $\mathbf{X}_k$ with respect to any unitarily invariant norm [6]. Here, $\mathbf{\Pi}$ is the projection matrix of the sampled exemplars, and $k < r = rank(\mathbf{X})$. However, this problem is believed to be NP-hard, as one have to search over all possible $\binom{l}{k}$ choices [7]. This paper proposes a scalable and deterministic greedy fast exemplar selection (FES) algorithm for this problem. We then show the application of these selected exemplars for matrix representation and in making inferences about the actual underlying data.

FES iteratively samples an exemplar from $\mathbf{X}$ by estimating how well the previously selected exemplars represents the remaining signals. To achieve this, FES searches for the exemplars which does not lie in the span of the already selected exemplars. In other words, FES extracts a linearly independent subset which captures the full range of the dataset. To this aim, FES ensures that the sampled exemplars have a positive definite (PD) Gram matrix, which guarantees linear independence [8]. However, checking positive definiteness of a matrix requires computing its eigenvalues, which is computational expensive. Hence, this paper proposes a computationally efficient way using the fact that a PD Gram matrix has a unique Cholesky decomposition. To handle larger datasets, FES uses incremental Cholesky decomposition and block matrix inversion algorithms [9]. Our results demonstrate that FES provides a computationally efficient alternative to existing approaches.

The rest of the paper is organized as follows: In Section 2, we briefly review the existing approaches to solve the ES problem. In Section 3, we have mathematically described the motivation behind the proposed column sampling method, in Section 4 we propose an efficient algorithm for the ES problem. Section 5 presents the applications and experimental results. The summary of the paper is given in Section 7.

2. PRIOR WORK

Various methods to solve the ES problem have been studied extensively in both communities of numerical linear algebra and theoretical computer science [6]. Since, $\|\mathbf{X} - \mathbf{\Pi X}\|_F$ is lower bounded by $\|\mathbf{X} - \mathbf{X}_k\|_F$, a large number of approximation algorithms have been proposed to sample columns of $\mathbf{X}$ such that matrix $\mathbf{\Pi}$ satisfies

$$\|\mathbf{X} - \mathbf{X}_k\|_F \leq \|\mathbf{X} - \mathbf{\Pi X}\|_F \leq f(l,k)\|\mathbf{X} - \mathbf{X}_k\|_F \qquad (1)$$

for some function $f(.)$ [7].

In numerical linear algebra, deterministic solutions to the ES problem are obtained in the context of rank revealing factorizations (RRF) [10, 11]. RRF seeks a permutation matrix using which one can select a well-conditioned collection of columns that spans the (numerical) range of the matrix $\mathbf{X}$ [10]. However, these algorithms are not preferred mainly due to their large time complexity which scales to an order of $\mathcal{O}(n^3)$ for square matrices.

In contrast, the theoretical computer science community has investigated the ES problem by constructing a low-rank matrix approximation to $\mathbf{X}$ in the spectral or Frobenius norm sense [2]. The usual solution to this problem is the rank-r matrix $\mathbf{X}_k = \mathbf{X}\mathbf{Y}\mathbf{Y}^T$, where the columns of $\mathbf{Y}$ are the top r right singular vectors of $\mathbf{X}$ obtained using the Singular Value Decomposition (SVD) [7]. When the rank-r is unknown, leverage-based sampling approaches can be used where, columns of $\mathbf{X}$ are sampled based on the so called "statistical leverage scores" [12]. However, since SVD is also used for estimation of leverage scores, for larger dataset the running time might be too large, and in practice singular values can be irrational because of which one can only compute an approximate SVD. In order to obtain approximate and fast solutions, one can use random sampling based approaches, where column are sampled with probabilities proportional to their squared norms [13]. Alternatively, based on trade-off between speed or a better error bound, methods such as sequential er-

ror sampling [14], accelerated sequential incoherence selection (oASIS) [1] and sparse modeling representative selection (SMRS) [15] can also be used. The proposed algorithm is related to the oASIS algorithm, and can be seen as its efficient variant. FES gains improvement on how the gram matrix necessary for the algorithm is stored, updated and used. Specifically, instead of storing and updating the matrix inverse using Nystrom approximation, as suggested in [1], FES suggest storing and updating the Cholesky factorization of the matrix, leading to faster computation. In addition, this paper provides an alternative and simpler derivation of the method underlying the oASIS algorithm.

3. APPROXIMATE SOLUTION TO THE ES PROBLEM

The ES problem attempts to identify the best exemplars as columns of matrix $\mathbf{X}_S$ to represent the entire matrix $\mathbf{X}$ in a geometric sense [4]. This problem can be solved iteratively. Specifically, if any column $\mathbf{x}_i$ from matrix $\mathbf{X}$ has to be added in matrix $\mathbf{X}_S$ (of already selected columns from set S), then the following metric can be used

$$i = \underset{i \notin S}{\operatorname{argmax}} \|\mathbf{x}_i - \mathbf{\Pi}_S \mathbf{x}_i\|_2^2 = \|\mathbf{x}_i - \mathbf{X}_S \mathbf{X}_S^+ \mathbf{x}_i\|_2^2 \tag{2}$$

It computes the distance of vector $\mathbf{x}_i$ to the space spanned by the set $\mathbf{X}_S$. Here, $\mathbf{\Pi}_S$ is the projection matrix, $\mathbf{X}_S\mathbf{X}_S^+\mathbf{x}_i$ is the projection of $\mathbf{x}_i$ on to $\mathbf{X}_S$, and $+$ denotes the pseudo-inverse. One can begin the algorithm with the assumption $\mathbf{X}_S = \emptyset$, and iteratively sample the column using the criteria of (2). The algorithm can stop when the following quantity reaches to a threshold value [3]:

$$\min \|\mathbf{X} - \mathbf{X}_S \mathbf{X}_S^+ \mathbf{X}\|_F \tag{3}$$

Problem in (2) appears in various forms in the existing literature related to column/row/feature selection algorithms [2, 3, 5, 16, 17]. For instance (2) is solved explicitly using pseudo-inverse in SES approach [14].

Taking the term inside l_2-norm and multiplying both sides by $\mathbf{x}_i^T$, the expression in (2) can be rewritten as:

$$\mathbf{x}_i^T \mathbf{x}_i - \mathbf{x}_i^T \mathbf{X}_S \mathbf{X}_S^+ \mathbf{x}_i \tag{4}$$

Assuming the columns already sampled in $\mathbf{X}_S$ are independent, the expression in (4) can be expanded as

$$\mathbf{x}_i^T \mathbf{x}_i - \mathbf{x}_i^T \mathbf{X}_S (\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{X}_S^T \mathbf{x}_i \tag{5}$$

which can be further simplified as

$$\Delta_i = \mathbf{d}_i - \mathbf{a}_i^T (\mathbf{W})^{-1} \mathbf{a}_i \tag{6}$$

where $\mathbf{d}_i = \mathbf{x}_i^T \mathbf{x}_i$, $\mathbf{a}_i = \mathbf{X}_S^T \mathbf{x}_i$ and $\mathbf{W} = \mathbf{X}_S^T \mathbf{X}_S$. This result is employed in different ways to govern the sampling rule in existing approaches [2, 3]. This is preferable as updating a new column in set S is equivalent to rank-1 update in $\mathbf{W}$ and if $\mathbf{W}$ is invertible, the inverse can be computed using block matrix update [9].

4. PROPOSED FES APPROACH

In the proposed FES approach, the matrix $\mathbf{X}_S$ is build by sequentially sampling exemplars from the training dataset. The proposed FES method ensures that the sampled columns of $\mathbf{X}_S$ forms a positive definite (PD) Gram matrix $\mathbf{W}$. This is because a PD matrix is the Gram matrix of linearly independent set of vectors [18]. However, to check if a matrix is PD, one needs to compute the eigenvalues, which is computationally expensive. To address this issue, the proposed approach exploits the fact that a PD matrix has a unique Cholesky decomposition. Assuming the columns of $\mathbf{X}$ are normalized, and $\mathbf{x}_i$ is the new sampled exemplar in an iteration, one can write the updated Gram matrix as [8, 9]:

$$\mathbf{W}_{k+1} = \begin{bmatrix} \mathbf{X}_S^T \mathbf{X}_S & \mathbf{a} \\ \mathbf{a}^T & \mathbf{x}_i^T \mathbf{x}_i \end{bmatrix} = \begin{bmatrix} \mathbf{W}_k & \mathbf{a} \\ \mathbf{a}^T & 1 \end{bmatrix} \tag{7}$$

Assuming $\mathbf{W}_k$ to be SPD matrix such that there exists a unique lower triangular Cholesky decomposition $\mathbf{W}_k = \mathbf{L}_k \mathbf{L}_k^T$, the updated Gram matrix can be expressed via block matrix update as [9]:

$$\begin{bmatrix} \mathbf{W}_k & \mathbf{a} \\ \mathbf{a}^T & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{L}_k & 0 \\ \mathbf{c}^T & d \end{bmatrix} \begin{bmatrix} \mathbf{L}_k^T & \mathbf{c}^T \\ 0 & d \end{bmatrix} = \begin{bmatrix} \mathbf{L}_k \mathbf{L}_k^T & \mathbf{L}_k \mathbf{c} \\ \mathbf{L}_k^T \mathbf{c}^T & \mathbf{c}^T \mathbf{c} + d^2 \end{bmatrix} \tag{8}$$

where, $\mathbf{c}$ is a real vector, and d a positive scalar. Comparing left and right sides of (8), we have

$$\mathbf{a} = \mathbf{L}_k \mathbf{c} \text{ or } \mathbf{c} = \mathbf{L}_k^{-1} \mathbf{a} \text{ and } \quad d \quad = \sqrt{1 - \mathbf{c}^T \mathbf{c}} \tag{9}$$

where $\mathbf{L}_k$ is the Cholesky factor of $\mathbf{W}_k$, and d a positive scalar. In other words, the matrix $\mathbf{W}_{k+1}$ is PD, invertible and has a unique Cholesky decomposition if and only if, $\mathbf{c}^T \mathbf{c} < 1$. Hence, FES proposes to iteratively sample columns using the criteria $\mathbf{c}^T \mathbf{c} < 1$. Further updating $\mathbf{L}_k^{-1}$ using block matrix inversion is preferable, as $\mathbf{L}_k$ is lower triangular. Now, after back substituting the variables in (9) we have

$$d^2 = 1 - \mathbf{a}_i^T (\mathbf{L}_k^{-1T} \mathbf{L}_k^{-1}) \mathbf{a}_i = 1 - \mathbf{a}_i^T (\mathbf{W}_k^{-1}) \mathbf{a}_i \tag{10}$$

It can be verified that (10) reduces to (6) when the data is normalized, and thus the sampling rule in FES is in other way a new formulation of the standard result. Theoretically data normalization doesn't affect the sampling metric, but it certainly affects its estimation numerically. Further, the proposed method has the following advantages: 1) $\mathbf{c}^T \mathbf{c} = \|\mathbf{c}\|_2^2$ i.e., sum of the square of entries of $\mathbf{c}$, which is faster to compute than Δ, 2) normalization saves from the computation of subtraction in (6) for each candidate in each iteration, allowing us to just ensure that $\mathbf{c}^T \mathbf{c} < 1$, 3) instead of $\mathbf{W}^{-1}$, estimating $\mathbf{L}^{-1}$ is efficient as it is a lower triangular matrix. In addi-

Algorithm 1 Fast Exemplar Selection (FES) Algorithm

Inputs: Training signal matrix $\mathbf{X} \in \mathbb{R}^{n \times l}$ containing l signals
Outputs: Matrix $\mathbf{X}_S \in \mathbb{R}^{n \times p}$ with p exemplars
Initialization: s and p

Randomized Step
1: Normalize the data matrix
2: Randomly keep s column indexes in set S and rest in set R.
Deterministic Step
3: Set $\mathbf{W} = \mathbf{X}_S^T \mathbf{X}_S$, $\mathbf{L} = chol(\mathbf{W})$
4: **Until** $p > 1$
$\mathbf{C} = \mathbf{L}^{-1} \mathbf{X}_S^T \mathbf{X}$
$\forall i \subset R$, keep the i-th column as a dictionary atom as:
$\Delta \leftarrow diag(\mathbf{C}^T \mathbf{C}), i \leftarrow \operatorname{argmin}[\Delta(i)]$
If $\forall_i \Delta_i \geq 1$ **Go to** (5);
Else Update
$d \leftarrow \sqrt{1 - \Delta_i}, \mathbf{c} \leftarrow \mathbf{C}(:, i), S \leftarrow S \cup i, R \leftarrow R/i$
$\mathbf{L}^{-1} \leftarrow \begin{bmatrix} \mathbf{L}^{-1} & 0 \\ -(1/d)\mathbf{c}^T \mathbf{L}^{-1} & 1/d \end{bmatrix}, p \leftarrow p - 1$
5: Randomly sample remaining p columns from $\mathbf{X}_R$ as exemplars.

tion, one can speed up this computation via approximating $\mathbf{L}_{k+1}^{-1}$ by performing rank-1 updates to the inverse matrix $\mathbf{L}_k^{-1}$ i.e.,

$$\begin{bmatrix} \mathbf{L}_k & 0 \\ \mathbf{c}^T & d \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{L}_k^{-1} & 0 \\ -(1/d)\mathbf{c}^T\mathbf{L}_k^{-1} & 1/d \end{bmatrix} \tag{11}$$

The pseudo-code of the proposed approach is shown in Algorithm 1.

4.1. Comparison with Existing Works

Similar to FES, various sequential sampling methods have been proposed in the past. The main difference lies in how projection matrix $\mathbf{\Pi}_S$ or equivalently $\mathbf{W}^{-1}$ is computed or approximated after each iteration. Further, note that the updated matrix $\mathbf{W}_{k+1}$ is invertible only if its Schur complement is positive, and it can be verified that Δ is in-fact the Schur complement [18]. For instance, instead of computing $\mathbf{W}$ or its inverse directly, work in [19] proposed an iterative method to obtain a sequence of gradually better approximations, and called the approximation of Δ the *generalized stretch*. Similarly, In [20], authors proposed a recursive formula for the approximation of projection matrix $\mathbf{\Pi}_S$, using the Schur compliment and the block matrix updates. In [17], authors introduced (6) as *dependency margin* for optimal feature selection. In [1], the oASIS algorithm exploits the Nystrom approximation of the matrix $\mathbf{W}$ to compute Δ, in contrast to Cholseky decomposition employed in the proposed FES algorithm[1]. The proposed FES method can be seen as a variant of oASIS algorithm with an alternative and computationally efficient implementation. What differentiates the algorithms is (i) the logic used to motivate and justify the use of the Schur complement and (ii) the method of computing the inverse $\mathbf{W}^{-1}$.

4.2. Computational Complexity

The rate-limiting step of Algorithm 1 is the computation of $\mathbf{L}_k^{-1}$ for l training signals, and (11) allows this to be performed using $\mathbf{L}_{k-1}^{-1}$. The complexity of a single iteration is thus $\mathcal{O}(kl)$. If p columns are sampled in total, then $\sum_{k=1}^{p} kl = p(p+1)l/2$ entries need to be updated. Thus the resulting complexity of the FES approximately scales as $\mathcal{O}(p^2 l)$. However, in general $p < l$ and $\mathbf{L}$ being lower triangular does not require estimating all its entries. This makes the proposed algorithm considerably more efficient than existing methods for which the time complexity scales as $\mathcal{O}(nl^2)$ [15, 21].

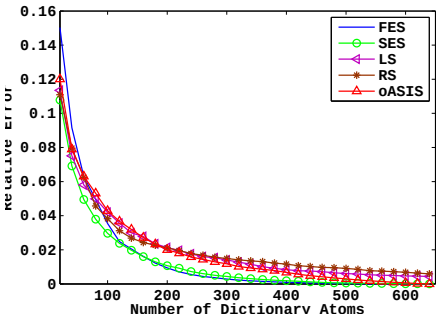
5. APPLICATIONS AND EXPERIMENTAL RESULTS

This section demonstrates the use of FES for matrix approximation, feature selection and clustering. As a prerequisite it is shown that FES fulfills a sufficient condition for all the above mentioned problems i.e., the columns of $\mathbf{X}_S$ captures the full range of the training set or $\mathbf{X} = \mathbf{\Pi}_S\mathbf{X}$.

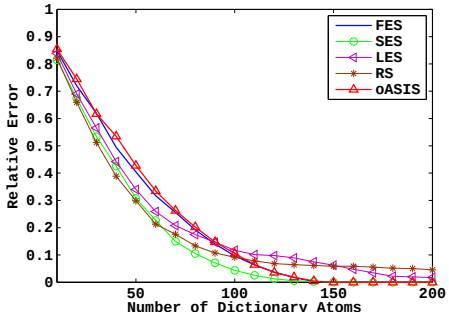
5.1. Exact Matrix Recovery and Low Rank Approximation

The proposed FES approach yields exact matrix recovery, which follows from the fact that the sampled exemplars span the whole space of $\mathbf{X}$. To understand this, let the matrices $\mathbf{X}$ and $\mathbf{X}_S$ are of rank r and $r1$, respectively. Thus, when $r1 = r$ exact recovery of $\mathbf{X}$ is guaranteed, and the proof follows by induction. In each iterations until the condition $\mathbf{c}^T\mathbf{c} < 1$ is violated, any new column will be independent to previously sampled columns, as the Gram matrix will be PD. Similar arguments follows for the low rank approximation problem. The aim is to minimize the objective $\|\mathbf{X} - \mathbf{X}_k\|_F^2$ i.e., the error between the data matrix $\mathbf{X}$ and a low rank approximating matrix $\mathbf{X}_k$. In this case, one needs to terminate the FES algorithm after selecting $k < r$ linearly independent columns of $\mathbf{X}$.

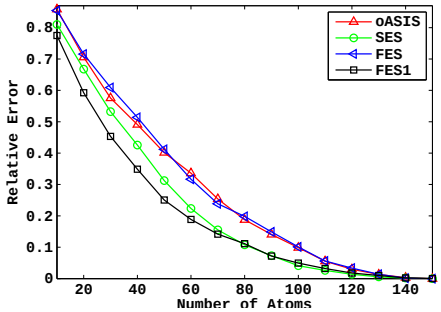
To illustrate the performance of FES for matrix approximation, Fig.1 shows the approximation error as a function of number of dictionary atoms sampled, for a rank-150 matrix taken from the synthetic union-of-subspaces UoS dataset [22] and a rank-631 matrix taken from the Yale-B face dataset [23]. Other sampling based approaches compared in this and following experiments are oASIS [1], uniform random sampling (RS) [6], sequential error sampling (SES) [14] and leverage sampling (LES) [12]. It can be observed that only SES, FES and oASIS achieve exact recovery when the number of



(a)



(b)



(c)

Fig. 1. Approximation error vs number of dictionary atoms for (a) Face, (b) and (c) UoS datasets.

[1]We came across [1] at the time of development of our algorithm and thus we have mentioned the same in the title, to give proper attribution to the oASIS algorithm.

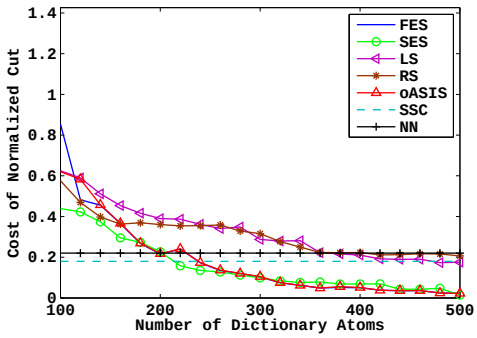


Fig. 2. Normalized cut ratios for clustering task on the face dataset.

sampled exemplars equals the rank. Note that the error curves of oASIS and FES are similar but not exactly same. The reason could be the error introduced in approximating the projection matrix via different methods. One can observe similar decay in the approximation error for both the synthetic and real datasets. To investigate further, Fig.1 (c) shows the error curves for the proposed method when for $\mathbf{c}^T\mathbf{c} < 1$: 1) the candidate with the minimum value is sampled, denoted by FES, 2) the candidate is sampled randomly, denoted by FES1. It can be observed that FES1 with random sampling achieves best error curve with a significant improvement over other methods. This shows that the error curves for FES/oASIS trails behind SES not because $\mathbf{W}^{-1}$ is approximated, but in the way how candidates are selected. Although, this needs further investigation which we defer as future work.

5.2. Optimal Feature Selection in Union of Subspaces

In optimal feature selection (OFS) problem, every signal is represented using only signals from within its own subspace. This requires that at least t linearly independent columns that span each t-dimensional subspace exist in $\mathbf{X}$. When this occurs, $\mathbf{X}$ provides a complete reference set for each subspace present in the data. It has been proved in [24, 25] that whenever $\mathbf{X}_S$ yields exact matrix recovery it is guaranteed that $\mathbf{X}_S$ also provides a complete reference set for a union of subspaces. Thus, using results in [24, 25] and of Section 5.1, produce guarantees that OFS occurs for the decomposition obtained via FES. This can also be verified from Fig. 1(a) which shows exact matrix recovery for FES on UoS dataset.

5.3. Application to Sparse Representation Based Clustering

In recent years, signal representations using a subset of training signals, called *exemplars*, has been widely explored in the field of machine learning [15, 26] and has been successfully applied in the context of classification [27, 28], clustering [21], and low-rank matrix approximation [29]. Here, the objective is to represent each signal in the dataset as a linear combination of a small subset of exemplars [15]. Specifically, one seeks the factorization $\mathbf{A}$ for a collection of signals as columns of matrix $\mathbf{X}$ by minimizing the objective function

$$\begin{aligned} \|\mathbf{X} - \mathbf{X}\mathbf{A}\|_F^2 &= \sum_{i=1}^{l} \|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2^2, \quad \text{or} \\ \|\mathbf{X} - \mathbf{X}_S\mathbf{A}\|_F^2 &= \sum_{i=1}^{l} \|\mathbf{x}_i - \mathbf{X}_S\mathbf{a}_i\|_2^2 \end{aligned} \tag{12}$$

Here, one represent each training signal in terms of other signals in the dataset. Depending on the application, additional constraints such as convexity or sparsity can also be imposed over $\mathbf{A}$, as done in case of approaches such as archetypal analysis [30], sparse subspace clustering (SSC) [21], self-expressive decomposition [22], and sparse modeling by representative selection (SMRS) [15].

Sparse representation based clustering, uses the fact that a signal can be sparsely represented using exemplars from its own class (or subspace) [31]. Thus using FES, one can cluster the data using the sparsity patterns of the obtained decomposition i.e., $\mathbf{A}$. In order to illustrate the performance of FES for clustering, we follow the approach of [22]. Here $\mathbf{A}$ is considered as representing the edges of a bi-partite graph for which the average cost of a normalized cut is computed, for all the classes as a function of number of dictionary atoms sampled. This cost is a measure of how easy it is to cluster the graph into its correct classes [32]. Fig. 2, shows the normalized cut ratios with maximum number of dictionary atoms to sample equal to 30% of the whole data, on a subset of the Yale-B Face dataset consisting of 10 different subjects under various illumination conditions. One can observe that FES, SEED and SES achieve normalized cuts less than nearest-neighbor (NN) and SSC methods. The gap between SSC and FES grows as the number of dictionary atoms are increased. The performance of LS and RS appears to flatline just near the cut ratios for NN and SSC methods.

6. REPRODUCIBLE RESEARCH

For reproducible-research purposes, a GPL Matlab implementation and Creative Commons data related to the presented work are available on request via votrix13[at]ieee[dot]org.

7. SUMMARY

This paper introduced FES, a low complexity rank revealing exemplar selection approach. FES sequentially sample linearly independent exemplars from the dataset. This is done by seeding an exemplar matrix column by column, such that it has a PD Gram matrix. To ensure PD, we exploited the Cholesky decomposition and block matrix inversion properties. Thus, using the rank revealing property of FES approach, one can analyze and discover structures, patterns and other properties of the training signals. We have experimentally demonstrated that FES performs well for various signal processing and machine learning problems, ranging from matrix approximation, to feature selection and clustering.

8. REFERENCES

[1] R. Patel, T. A. Goldstein, E. L. Dyer, A. Mirhoseini, and R. G. Baraniuk, "oASIS: Adaptive Column Sampling for Kernel Matrix Approximation," *ArXiv e-prints*, may 2015.

[2] A. Çivril, "Column subset selection problem is UG-hard," *Journal of Computer and System Sciences*, vol. 80, no. 4, pp. 849 – 859, 2014.

[3] F. de Hoog and R. Mattheij, "Subset selection for matrices," *Linear Algebra and its Applications*, vol. 422, no. 23, pp. 349 – 359, 2007.

[4] A. Çivril and M. Magdon-Ismail, "Column subset selection via sparse approximation of {SVD}," *Theoretical Computer Science*, vol. 421, pp. 1 – 14, 2012.

[5] A. Deshpande and S. Vempala, "Adaptive sampling and fast low-rank matrix approximation," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, ser. APPROX'06/RANDOM'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 292–303.

[6] J. A. Tropp, "Column subset selection, matrix factorization, and eigenvalue optimization," in *ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, 2009, pp. 978–986.

[7] A. Farahat, A. Elgohary, A. Ghodsi, and M. Kamel, "Greedy column subset selection for large-scale data sets," *Knowledge and Information Systems*, pp. 1–34, 2014.

[8] D. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas (Second Edition)*, ser. Princeton reference. Princeton University Press, 2009.

[9] G. Stewart, "On the stability of sequential updates and downdates," *IEEE Transactions on Signal Processing*, vol. 43, no. 11, pp. 2642–2648, November 1995.

[10] T. F. Chan, "Rank revealing QR factorizations," *Linear Algebra and its Applications*, vol. 8889, no. 0, pp. 67 – 82, 1987.

[11] M. Gu and S. C. Eisenstat, "Efficient algorithms for computing a strong rank-revealing QR factorization," *SIAM Journal on Scientific Computing*, vol. 17, no. 4, pp. 848–869, 1996.

[12] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. Woodruff, "Fast approximation of matrix coherence and statistical leverage," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3475–3506, December 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2503308.2503352

[13] A. Frieze, R. Kannan, and S. Vempala, "Fast Monte-carlo algorithms for finding Low-rank approximations," *Journal of ACM*, vol. 51, no. 6, pp. 1025–1041, November 2004.

[14] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang, "Matrix approximation and projective clustering via volume sampling," *Theory of Computing*, vol. 2, no. 12, pp. 225–247, 2006. [Online]. Available: http://www.theoryofcomputing.org/articles/v002a012

[15] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 1600–1607.

[16] A. K. Farahat, A. Ghodsi, and M. S. Kamel, "A novel greedy algorithm for nyström approximation," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 15, April 2011, pp. 278–286.

[17] Y. Liu, F. Tang, and Z. Zeng, "Feature selection based on dependency margin," *IEEE Transactions on Cybernetics*, vol. 45, no. 6, pp. 1209–1221, June 2015.

[18] G. Strang, *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 2003. [Online]. Available: http://math.mit.edu/~gs/linearalgebra/

[19] M. Li, G. L. Miller, and R. Peng, "Iterative row sampling," in *IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, October 2013, pp. 127–136.

[20] A. K. Farahat, A. Ghodsi, and M. S. Kamel, "A fast greedy algorithm for generalized column subset selection," *NIPS Workshop on Greedy Algorithms*, December 2013.

[21] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, November 2013.

[22] E. L. Dyer, T. A. Goldstein, R. Patel, K. P. Kording, and R. G. Baraniuk, "Self-Expressive Decompositions for Matrix Approximation and Clustering," *ArXiv e-prints*, May 2015.

[23] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, June 2001.

[24] E. Elhamifar, G. Sapiro, and R. Vidal, "Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery," in *Advances in Neural Information Processing Systems*, December 2012, pp. 19–27.

[25] E. Dyer, A. Sankaranarayanan, and R. Baraniuk, "Greedy feature selection for subspace clustering," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2487–2517, January 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2567709.2567741

[26] V. Abrol, P. Sharma, and A. Sao, "Greedy double sparse dictionary learning for sparse representation of speech signals," *Speech Communication*, vol. 85, pp. 71 – 82, December 2016.

[27] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, February 2009.

[28] V. Abrol, P. Sharma, and A. Sao, "Greedy dictionary learning for kernel sparse representation based classifier," *Pattern Recognition Letters*, vol. 78, pp. 64 – 69, 2016.

[29] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, January 2013.

[30] M. Mørup and L. K. Hansen, "Archetypal analysis for machine learning and data mining," *Neurocomputing*, vol. 80, pp. 54 – 63, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231211006060

[31] B. Gowreesunker and A. Tewfik, "Learning sparse representation using iterative subspace identification," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3055–3065, June 2010.

[32] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, August 2001, pp. 269–274.