Select to *Better* Learn: Fast and Accurate Deep Learning using Data Selection from Nonlinear Manifolds

Mohsen Joneidi*, Saeed Vahidian†, Ashkan Esmaeili*, Weijia Wang†, Nazanin Rahnavard*, Bill Lin†, and Mubarak Shah#

* University of Central Florida, Department of Electrical and Computer Engineering

† University of California, San Diego, Department of Electrical and Computer Engineering

University of Central Florida, Center for Research in Computer Vision

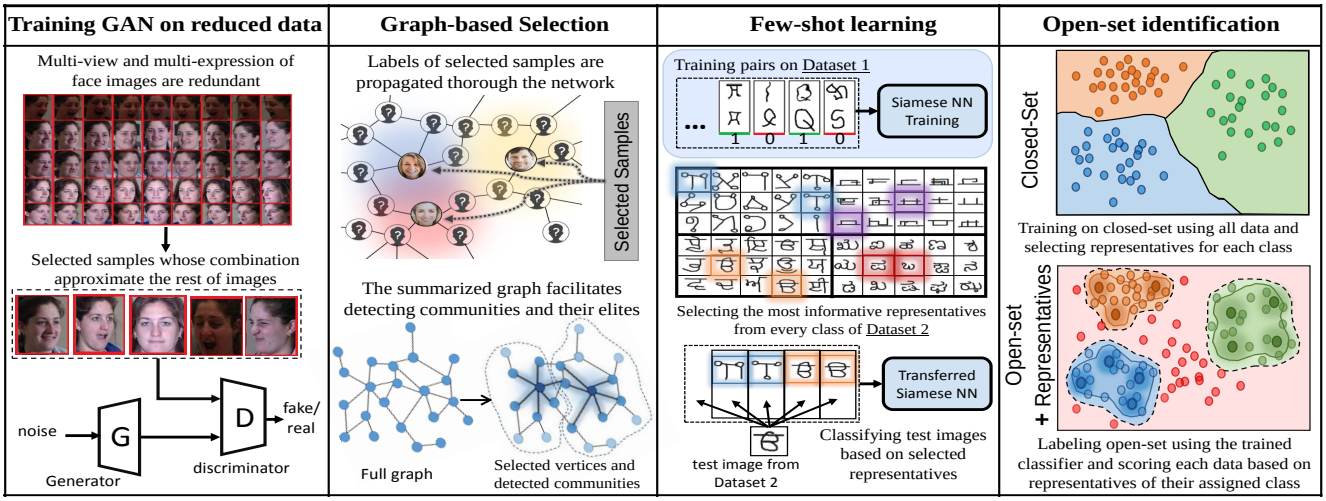


Figure 1: Several deep learning applications of our proposed data selection algorithms discussed in this paper.

Abstract

Finding a small subset of data whose linear combination spans other data points of dataset, also called column subset selection problem (CSSP), is an important open problem in computer science with a wide range of applications in computer vision and deep learning such as the ones shown in Fig. 1. There are some studies that solve CSSP in a polynomial *time complexity w.r.t. the size of the original dataset. A simple and efficient selection algorithm with a linear complexity order, referred to as* spectrum pursuit *(SP), is proposed that pursuits spectral components of the dataset using available sample points. The proposed non-greedy algorithm aims to iteratively find K data samples whose span is close to that of the first K spectral components of entire data. SP has no parameter to be fine tuned and this desirable property makes it problem-independent. The simplicity of SP enables us to extend the underlying linear model to more complex models such as nonlinear manifolds and graph-based models. The nonlinear extension of SP is introduced as the* kernel-SP *(KSP) algorithm. The superiority of the proposed algorithms is demonstrated in many applications including training generative adversarial networks, graph-based label propagation, few shot classification, graph summarization and open-set identification.*

1. Introduction

Processing M data samples, each including N features, is not feasible for of most the systems when M is a very large number. Therefore, it is crucial to select a small subset of $K << M$ data from the entire set such that the selected data can capture the underlying properties or structure of the entire data. This way, complex systems such as deep learning (DL) networks can operate on the informative selected data rather than the redundant entire data. Randomly selecting K out of M data, while computationally simple, is inefficient in many cases, since non-informative or redundant instances may be among the selected ones. On the other hand, the optimal selection of data for a specific task implies solving an NP-hard problem [2]. For example, finding an optimal subset of data to be employed in training a DL network with the best performance requires $\binom{M}{K}$ number of trial and errors, which is not tractable. It is essential to define a versatile objective function and to develop a method that efficiently selects the K samples that optimize the function. Let us assume the M data samples are organized as the columns of a matrix $\boldsymbol{A} \in \mathbb{R}^{N \times M}$. The following is a general purpose cost function for subset selection, known as *column subset selection problem* (CSSP), which is an open problem [3]:

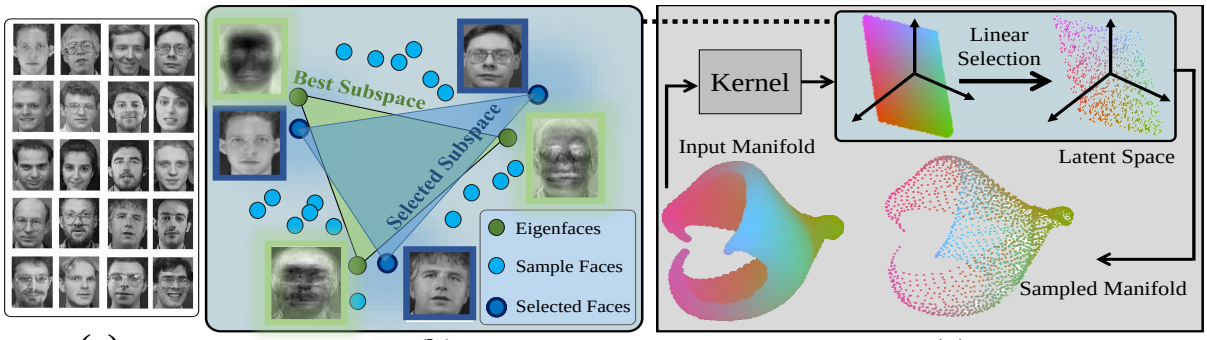


Figure 2: Intuitive illustration of our main contributions in this paper. (a) A dataset including 20 real images from AT&T face database [1] is considered. (b) the images in (a) are represented as blue dots. Three most significant eigenfaces are shown by green dots. However, these eigenfaces are not among data samples. Here we are interested in selecting the best 3 out 20 real images whose span is the closest to the span of the 3 eigenfaces. There are $\binom{20}{3}$ possible combinations from which the best subset must be selected. In this paper, we propose the SP algorithm to select K samples such that their span pursuits the span of the first K singular vectors. (c) Utilizing the proposed linear selection algorithm (SP), a tractable algorithm is developed for selecting from low-dimensional manifolds. First a kernel which is defined by neighborhood transforms the given data on a manifold to a latent space. Next, the linear selection operator is performed.

$$\mathbb{S}^* = \underset{|\mathbb{S}|\leq K}{\operatorname{argmin}} \|\boldsymbol{A} - \pi_{\mathbb{S}}(\boldsymbol{A})\|_F^2, \tag{1}$$

where $\pi_{\mathbb{S}}$ is the linear projection operator on the span of K columns of $\boldsymbol{A}$ indicated by set $\mathbb{S}$. This problem has been shown to be NP hard [2]. Moreover, the cost function is not sub-modular [4] and greedy algorithms are not efficient to tackle Problem (1). Computer scientists and mathematicians during the last 30 years have proposed many tractable selection algorithms that guarantee an upper bound for the projection error $\|\boldsymbol{A} - \pi_{\mathbb{S}}(\boldsymbol{A})\|_F^2$. These works include algorithms based on QR decomposition of matrix $\boldsymbol{A}$ with column pivoting (QRCP) [5, 6, 7], methods based on volume sampling (VS) [8, 9, 10] and matrix subset selection algorithms [3, 11, 12]. However, the guaranteed upper bounds are very loose and the corresponding selection results are far from the actual minimizer of CSSP in practice. Interested readers are referred to [13, 11] and Sec. 2.1 in [14] for detailed discussions. For example, in VS it is shown that the projection error on the span of K selected samples is guaranteed to be less than $K+1$ times of the projection error on the span of the K first left singular vectors (which is too loose for a large K). Recently, it was shown that VS performs even worse than random selection in some scenarios [15]. Moreover, some efforts have been made using convex relaxation and regularization. Fine tuning of these methods is not straightforward. Moreover their cubical complexity is an obstacle to employ these methods for diverse applications.

Recently, a low-complexity approach was proposed to solve CSSP, referred to as iterative projection and matching (IPM) [16]. IPM is a greedy algorithm that selects K consecutive and locally optimum samples, without the option of revisiting the previous selections and escaping local optima. Moreover, IPM samples the data from *linear subspaces*, while in general data points reside in the union of *nonlinear manifolds*.

In this paper, an efficient non-greedy algorithm is proposed to approach Problem (1) with a linear order of complexity. The proposed subspace-based algorithm outperforms the state-of-the-art algorithms in terms of accuracy for CSSP. In addition, the simplicity and accuracy of the proposed algorithm enable us to extend it for efficient sampling from nonlinear manifolds. The intuition behind our work is depicted in Fig. 2. Assume for solving CSSP, we are not restricted to selecting representatives from data samples and we are allowed to generate pseudo-data and select them as representatives. In this scenario, the best K representatives are the first K spectral components of data according to definition of singular value decomposition (SVD) [17]. However, the spectral components do not reside in the dataset. Our proposed algorithm aims to find K data samples such that their span is close to that of the first K spectrum of data. We refer to our proposed algorithm as *spectrum pursuit (SP)*. Fig. 2 (b) shows the intuition behind SP and Fig. 2 (c) shows a straightforward extension of SP for sampling from nonlinear manifolds. We refer to this algorithm as *Kernel Spectrum Pursuit (KSP)*.

Our main contributions can be summarized as:

- We introduce SP, a non-greedy selection algorithm with linear order complexity w.r.t. the number of original data points. SP captures spectral characteristics of dataset using only a small number of samples. To the best of our knowledge, SP is the most accurate solver for CSSP.

- Further, we extend SP to Kernel-SP for manifold-based data selection.

- We provide extensive evaluations to validate our proposed selection schemes. In particular, we evaluate the proposed algorithms on training generative adversarial networks, graph-based label propagation, few shot classification, and open-set identification, as shown in Fig. 1. We demonstrate that our proposed algorithms outperform the state-of-the-art algorithms.

2. Data Selection from Linear Subspaces

In this section, we first introduce the related work on matrix subset selection and then we propose our algorithm for CSSP.

2.1. Related Work

A simple approach to selection is to reduce the entire data and evaluate a criterion *only* for the reduced set, $\boldsymbol{A}_{\mathbb{S}}$. Mathematically speaking, we need to solve the following problem [18, 9]:

$$\mathbb{S}^* = \underset{|\mathbb{S}|\leq K}{\operatorname{argmin}} \ \phi\left((\boldsymbol{A}_{\mathbb{S}}^T \boldsymbol{A}_{\mathbb{S}})^{-1}\right). \tag{2}$$

Here, $\phi(.)$ is a function of matrix eigenvalues, such as the determinant or trace function. This is an NP hard and non-convex problem that can be solved via convex relaxation of ℓ_0 norm with time complexity of $O(M^3)$ [19, 18]. There are several other efforts in this area for designing function ϕ [9, 20, 21]. Inspired by D-optimal design, VS [10] considers a selection probability for each subset of data, which is proportional to the determinant (volume) of the reduced matrix [9, 22, 23]. To the best of our knowledge the tightest bound for CSSP is introduced in NIPS 2015 paper as follows for selecting K columns [11]:

$$\|\boldsymbol{A} - \pi_{\mathbb{S}^*}(\boldsymbol{A})\|_F^2 \leq 3\ \|\boldsymbol{A} - \boldsymbol{A}_K\|_F^2,$$

where $\boldsymbol{A}_K$ is the best rank-K approximation of $\boldsymbol{A}$. Moreover, VS guarantees a projection error up to $K+1$ times worse than the first K singular vectors [10]. A set of diverse samples optimizes cost function (2) and algorithms such as VS assign a higher probability for them to be chosen. However, selecting some diverse samples that are solely different from each other probably does not provide good representative for all (un-selected) data.

Ensuring that selected samples are able to reconstruct unselected samples is a more robust approach than selecting a diverse subset. The exact solution of Problem (1) aims to find such a subset. An equivalent problem to the original problem (1) is proposed in [24]. Their suggested equivalent problem exploits the mixed norm, $\|.\|_{2,0}$, which is not a convex function and they propose to employ ℓ_1 regularization to relax it [24]. There is no guarantee that convex relaxation provides the best approximation for an NP-hard problem. Furthermore, such methods which approach the problem using convex programming are usually computationally intensive for large datasets [24, 25, 26, 27]. In this paper, we present another reformulation of Problem (1) and propose a fast and accurate algorithm for addressing CSSP.

2.2. Spectrum Pursuit (SP)

Projection of all data onto the subspace spanned by K columns of $\boldsymbol{A}$, indexed by $\mathbb{S}$, i.e., $\pi_{\mathbb{S}}(\boldsymbol{A})$, can be expressed by a rank-K factorization, $\boldsymbol{U}\boldsymbol{V}^T$. In this factorization, $\boldsymbol{U} \in \mathbb{R}^{N\times K}$, $\boldsymbol{V} \in \mathbb{R}^{M\times K}$, and $\boldsymbol{U}$ includes a set of K normalized columns of $\boldsymbol{A}$, indexed by $\mathbb{S}$. Therefore, the optimization problem (1) can be restated as [16]:

$$\underset{\boldsymbol{U},\boldsymbol{V}}{\operatorname{argmin}} \|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{V}^T\|_F^2 \text{ s.t. } \boldsymbol{u}_k \in \mathbb{A}, \tag{3}$$

where, $\mathbb{A} = \{\tilde{\boldsymbol{a}_1}, \tilde{\boldsymbol{a}_2}, \ldots, \tilde{\boldsymbol{a}}_M\}$, $\tilde{\boldsymbol{a}}_m = \boldsymbol{a}_m/\|\boldsymbol{a}_m\|_2$, and $\boldsymbol{u}_k$ is the k^{th} column of $\boldsymbol{U}$. It should be noted that $\boldsymbol{U}$ is restricted to be a collection of K normalized columns of $\boldsymbol{A}$, while there is no constraint on $\boldsymbol{V}$. As mentioned before, this is an NP hard problem. Recently, IPM [16], a fast sub-optimal approach to tackle (3), was proposed. In IPM, samples are selected in a greedy manner until K samples are collected. Although sequential selection is desirable in certain applications such as active learning and online data selection, it may result in a local optimum. In this paper, we propose a new selection algorithm, referred to as Spectrum Pursuit (SP), which can select a better solution for Problem (3). The time complexity of both IPM and SP are linear with respect to the number of samples and the samples' dimension, which is desirable for selection from very large datasets. The idea behind SP is that instead of making K consecutive locally optimal selections, we start with a set of K samples, selected either randomly or for example by IPM. We then iteratively update this set by removing one sample and adding a new one such that the new set of K samples minimizes (3). This facilitates *revising* our selection and escaping from local optima. In SP, we modify (3) into two sub-problems. The first one is built upon the assumption that we have already selected $K-1$ data points and the goal is to select the next best data. However, it relaxes the constraint $\boldsymbol{u}_k \in \mathbb{A}$ in (3) to a moderate constraint $\|\boldsymbol{u_k}\| = 1$. This relaxation makes finding the solution tractable at the expense of coming up with a solution that may not belong to our data points. To fix this, we introduce a second sub-problem that reimposes the underlying constraint and selects the datapoint that has the highest correlation with the point selected in the first sub-problem. These sub-problems are formulated as

$$(\boldsymbol{u}_k, \boldsymbol{v}_k) = \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmin}} \|\boldsymbol{A} - \boldsymbol{U}_{\overline{k}}\boldsymbol{V}_{\overline{k}}^T - \boldsymbol{u}\boldsymbol{v}^T\|_F^2 \text{ s.t. } \|\boldsymbol{u}\|_2 = 1, \tag{4a}$$

$$\mathbb{S}_k = \underset{m}{\operatorname{argmax}} |\boldsymbol{u}_k^T \tilde{\boldsymbol{a}}_m|. \tag{4b}$$

Here $\mathbb{S}_k$ is a singleton that contains the index of the selected data point. Matrices $\boldsymbol{U}_{\overline{k}}$ and $\boldsymbol{V}_{\overline{k}}$ are obtained by removing the k^{th} column of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Subproblem (4a) is equivalent to finding the first left singular vector (LSV) of $\boldsymbol{E}_k \triangleq \boldsymbol{A} - \boldsymbol{U}_{\overline{k}}\boldsymbol{V}_{\overline{k}}^T$. The constraint $\|\boldsymbol{u}\| = 1$ keeps $\boldsymbol{u}$ on the unit sphere to remove scale ambiguity between $\boldsymbol{u}$ and $\boldsymbol{v}$. Moreover, the unit sphere is a superset for $\mathbb{A}$ and keeps the modified problem close to the recast problem (3). After solving for $\boldsymbol{u}_k$, we find the data point that matches it the most in (4b). The steps of the SP algorithm are elaborated in Algorithm 1. Fig. 3 illustrates Problem (4) pictorially. SP is a low-complexity algorithm with no parameters to be tuned. The complexity order of computing the first singular component of an $M \times N$ matrix is $O(MN)$ [28]. As the proposed algorithm only needs the first singular component for each selection, the computational complexity of SP is $O(NM)$ per iteration which is much faster

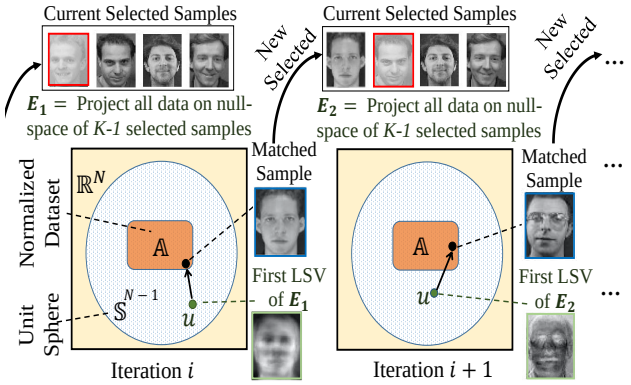


Figure 3: Two consecutive iterations of SP algorithm. The first LSV of the residual matrix is a vector on $\mathbb{S}^{N-1}$ and the goal is to find K samples which pursuit the spectral characteristics of dataset over iterations.

than convex relaxation-based algorithms with complexity $O(M^3)$ [18]. Moreover, SP performs faster than K-medoids algorithm and volume sampling, whose complexity is of order $O(KN(M-K)^2)$ and $O(MKN\log N)$, respectively [29, 30]. The stopping criterion can be convergence of set $\mathbb{S}$ or reaching a pre-defined maximum number of iterations.

Simplicity and accuracy of SP facilitate its extension to nonlinear manifold sampling with a wide range of applications. We will refer to this extended version as kernel-SP (KSP) which is discussed next in Section 3.

Algorithm 1 Spectrum Pursuit Algorithm

Require: $\boldsymbol{A}$ and K
Output: $\boldsymbol{A}_{\mathbb{S}}$
Initialization:
$\mathbb{S} \leftarrow$A random subset of $\{1,\dots,M\}$ with $|\mathbb{S}| = K$
$\{\mathbb{S}_k\}_{k=1}^{K} \leftarrow$Partition $\mathbb{S}$ into K subsets, each containing one element.
iter$= 0$
while a stopping criterion is not met
  $k =$mod(iter,K)+1
  $\boldsymbol{U}_{\overline{k}} =$ normalize column$(\boldsymbol{A}_{\mathbb{S}\setminus\mathbb{S}_k})$
  $\boldsymbol{V}_{\overline{k}} = \boldsymbol{A}^T\boldsymbol{U}_{\overline{k}}(\boldsymbol{U}_{\overline{k}}^T\boldsymbol{U}_{\overline{k}})^{-1}$
  $\boldsymbol{E}_{\overline{k}} = \boldsymbol{A} - \boldsymbol{U}_{\overline{k}}\boldsymbol{V}_{\overline{k}}^T$ (null-space projection)
  $\boldsymbol{u}_k =$ find the first left singular vector of $\boldsymbol{E}_{\overline{k}}$ by solving (4a)
  $\mathbb{S}_k \leftarrow$ index of the most correlated data with $\boldsymbol{u}_k$(4b)
  $\mathbb{S} \leftarrow \bigcup_{k'=1}^{K} \mathbb{S}_{k'}$
  iter=iter+1
end while

3. Kernel SP: Selection based on a Locally Linear Model

The goal of CSSP introduced in (1) is to select a subset of data whose *linear subspace* spans all data. Obviously, this model is not proper for general data types that mostly lie on nonlinear manifolds. Accordingly, we generalize (1) and propose the following selection problem in order to efficiently sample from a union of manifolds

$$\underset{|\mathbb{S}|\le K}{\operatorname{argmin}} \sum_{m=1}^{M} \|\boldsymbol{a}_m - \pi_{\mathbb{S}_m}(\boldsymbol{a}_m)\|_F^2 \quad \text{s.t. } \mathbb{S}_m \subseteq \mathbb{S} \cap \Omega_m, \tag{5}$$

where Ω_m indicates the indices of local neighbors of $\boldsymbol{a}_m$ based on an assumed distance metric. This problem is simplified to CSSP in Problem (1) if Ω_m is assumed to be equal to $\{1,\cdots,M\}$. Problem (2) is written for each column of $\boldsymbol{A}$ separately in order to engage neighborhood for each data. This problem facilitates fitting a customized and locally linear subspace for each data sample in terms of its neighbors.

Similar to Section 2, where we introduced SP as a low-complexity algorithm to tackle the NP-hard Problem (1), here we propose an extention of SP, referred to as kernel SP (KSP), to tackle the combinatorial search Problem (2). Manifold-based dimension reduction techniques and clustering algorithms do not provide prototypes suitable for data selection. However, inspired by spectral clustering of manifolds [31], main tool for nonlinear data analysis that partitions data into nonlinear clusters based on spectral components of the corresponding normalized similarity matrix, we formulate KSP as

$$\mathbb{S} = \underset{|\mathbb{S}|\le K}{\operatorname{argmin}} \|\boldsymbol{L} - \pi_{\mathbb{S}}(\boldsymbol{L})\|_F^2, \tag{6}$$

where $\boldsymbol{L} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}$, is the normalized similarity matrix of the data, $\boldsymbol{S} = [s_{ij}] \in \mathbb{R}^{M\times M}$ is defined as the similarity matrix of data and $\boldsymbol{D}$ is a diagonal matrix and $d_{ii} = \sum_{j\neq i} s_{ij}$. Note that since the SP algorithm models data with a low-rank matrix, identity matrix is not subtracted in order to keep $\boldsymbol{L}$ low-rank. The defined $\boldsymbol{L}$ differs Laplacian matrix only by an identity matrix. Note that problem (6) is the same as problem (1), where $\boldsymbol{A}$ is replaced by $\boldsymbol{L}$. The steps of the KSP algorithm are summarized in Algorithm 2.

Algorithm 2 Kernel Spectrum Pursuit

Require: $\boldsymbol{A}$, α, and K
Output: $\mathbb{S}$
$\boldsymbol{S} \leftarrow$Similarity Matrix: $s_{ij} = e^{-\alpha\|\boldsymbol{a}_i - \boldsymbol{a}_j\|}$
Form diagonal matrix $\boldsymbol{D}$ where $d_{ii} = \sum_{i\neq j} \boldsymbol{s}_{ij}$
$\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$.
$\mathbb{S} \leftarrow$ Apply SP on $\boldsymbol{L}$ with K (Alg. 1)

4. Empirical Results and Some Applications of SP/KSP

To evaluate the performance of our proposed selection algorithms we considered several applications and ran extensive experiments. For all the experiments, we compared our results with other state-of-the-art selection algorithms. The selected applications in this paper are (i) fast GAN training using reduced dataset, (ii) semi-supervised learning on graph-based datasets, (iii) large graph summarization, (iv) few-shot learning, and (v) open-set identification.

4.1. Training GAN

Here, we present our experimental results on CMU Multi-PIE Face Database [32] for representative selection. We use 249 subjects from the first session with 13 poses, 20 illuminations, and two expressions. Thus, there are $13 \times 20 \times 2$ images per subject. Fig. 4 (top) depicts 10

selected images from 520 images of a subject based on different selection methods: SP (our proposed) is compared with DS3 [33], VS [30], and K-medoids [34] as three well-known selection algorithms. As it can be seen, SP selects from more diverese angles. Fig. 4 (bottom) compares the performance of different state-of-the-art selection algorithms in terms of normalized projection error of CSSP, which is defined as the cost function in (1) for a give selection method normalized by the projection error of the random selection. As shown, SP outperforms all other methods. There is also a considerable performance gap between SP and IPM [16], the second best algorithm.

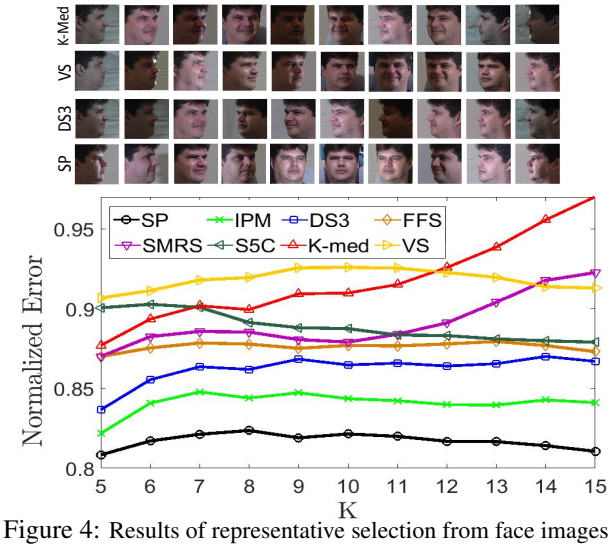


Figure 4: Results of representative selection from face images of Multi-pie dataset. (Top) Visualization of selection 10 images from 520 images of a subject. (Bottom) Averaged projection error for different number of representatives from 249 subjects. The projection error is normalized by projection error of random selection for all methods and the ratio is reported. The proposed SP algorithm is compared with IPM [16], DS3 [33], FFS [35], SMRS [24], S5C [36], K-medoids [34] and volume sampling [30].

Next, to investigate the effectiveness of selection in a real application, we use the selected samples to train a generative adversarial network (GAN) to generate multi-view images from a single-view input. For that, the GAN architecture proposed in [37] is employed. The experimental setup and the implementation details in [37] are considered where the first 200 subjects are used for training and the rest for testing. We select only 9 images from each subject and train the network with the selected images for 300 epochs using the batch size of 36. Table 1 shows the normalized ℓ_2 distances between features of the real and generated images, indicated as identity dissimilarities, averaged over all the images in the testing set. Features are extracted using a ResNet18 trained on MS-Celeb-1M dataset [38, 39]. As can be seen, SP and KSP outperform other selection methods. Moreover, KSP performs better than SP due to the selection from a nonlinear manifold.

Table 1: Identity dissimilarities between real and GAN-generated images for different selection methods. For each method, GAN is trained based on the selected data points.

SMRS	S5C	FFS	DS3	K-Med	VS	IPM	SP	KSP
0.631	0.617	0.608	0.602	0.599	0.583	0.553	0.550	**0.546**
Trained GAN using All Data 0.5364								

4.2. Graph-based Semi-supervised Learning

To evaluate the performance of our proposed selection algorithm on more complicated scenarios, we consider the graph convolutional neural network (GCN) proposed in [40] that serves as a semi-supervised classifier on graph-based datasets. Indeed, a GCN takes a feature matrix and an adjacency matrix as inputs and for every vertex of the graph produces a vector, whose elements correspond to the score of belonging to different classes. Moreover, every row of the feature matrix defines the feature of a vertex in the graph. The semi-supervised task here considers the case where only a *selected* subset of nodes are labeled in the training set and the loss is computed based on the output vectors of these labeled nodes to perform back-propagation. Moreover, we inherit the same two-layer network architecture from [40] and we follow their pre-processing techniques. To be more specific, an identity matrix is added to the original adjacency matrix so that every node is assigned with a self-connection. Further, we normalize the summation of two matrices using the kernel discussed in lines 2 and 3 of Algorithm 2 while the adjacency matrix serves as the similarity matrix S.

Our proposed KSP algorithm, together with other baselines, is tested on Cora dataset which is a real citation network dataset with $2,708$ nodes and $5,429$ number of edges as well as a random cluster-based graph datasets. The neural network is trained based on semi-supervised learning, i.e., the network is fed with the feature and adjacency matrices of the entire graph while the loss is only computed on the labeled vertices. Here the labeled vertices is the subset of vertices that is selected by performing our proposed algorithm (KSP) on the normalized adjacency matrix. We train both datasets for a maximum of 100 epochs using Adam [41] with a learning rate of 0.01 and early stopping with a window size of 10, i.e. we stop training if the validation loss does not decrease for 10 consecutive epochs. The results are summarized in Figure 5. Due to the inherent randomness of training neural networks using gradient descent based optimizers, some ripples appear in the curves. However, it can still be identified that, as expected, the test accuracy tends to increase as more labeled points are utilized for training. Further, as can be seen from the figure our proposed KSP algorithm significantly outperforms other algorithms for almost the whole range of selected points. This implies the superior performance of KSP in selecting the subset of data that comprises the most representative points of clusters. Lastly, because of the existence of outliers in a random graph, the accuracy of the proposed algorithm starts to improve slowly at about 70%, whereas other competitors saturate at about 60%. However, we note that the model is trained with only

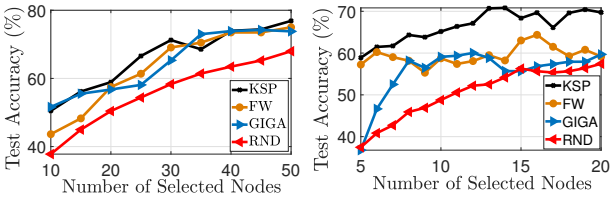


Figure 5: Semi-supervised classification accuracy of GCN on (Left) the Cora dataset [42] and (Right) a random cluster-based graph dataset. Only the selected nodes are labeled and the subset selection is performed using the proposed KSP algorithm, in comparison with GIGA [43], FW [43], and random selection (RND).

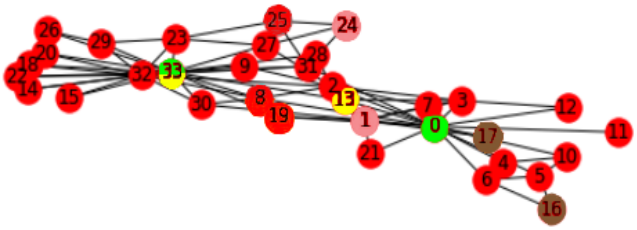


Figure 6: Zachary's Karate Club is a small social network where a conflict arises between the admin and the instructor in the club [52]. Each node of the club network represents a member of the karate club and a link between members indicate that they interact outside the club. The admin and the instructor which are the two nodes of this graph are $\{0, 33\}$, respectively. We apply KSP and two other algorithms to choose two of the main vertices. GIGA, MP and FW select ●, IS selects ●, VS selects ●, and KSP, FFS and DS3 select ●.

10% of data, so this also implicitly suggests that our algorithm successfully picks out the most informative nodes.

4.3. Graph Summarization

Clusters (also known as communities) in a graph are those groups of vertices that share common properties. Identification of communities is a crucial task in graph-based systems. Instances include protein-protein interaction networks in biology [53], recommendation systems [54] in computer science, social media networks, etc. In the following, we design an experiment to find the vertices with a central position on several types of graphs, produced both by real datasets such as [47] and also synthetic graph which contains the aggregated network of some users' Facebook friends. In the former dataset, vertices represent individuals on Facebook, and edges between two users mean they are Facebook friends.

Various community detection based algorithms such as betweenness centrality (BC) has been proposed to measure the importance of a user in the network [46] by looking at how many shortest paths pass through that user (vertex) for connecting each pair of other users (vertices). The more shortest paths that pass through the user, the more central the user is in the Facebook social network. Now assuming that a graph $\mathcal{G}$ or a similarity matrix is given, the aim is to first implement our method on the graph to approximate it with a subset of the vertices and then the exploit the measure of shortest path to evaluate the accuracy. We report the following performance measures: instead of computing the average shortest path between each vertex of the graph and all the other vertices which is really expensive (use of Dijkstra's algorithm n^2 times where n is the number of vertices), we compute the average shortest path between all the vertices and the selected vertices by KSP. The latter can be computed by using Dijkstra's algorithm only kn times, where k is the number of selected vertices.

Further, in this experiment we evaluate the performance of KSP compared with several state-of-the-art algorithms for data selection and coreset construction. The results of these experiments are shown in Table 2 where 10 vertices from each graph are selected (except for Karate Club sketched in Fig. 6 from which we select 2 vertices) by different data selection algorithms. As can be seen our proposed method provides significant improvements in shortest path error over the state-of-the-art.

4.4. Few Shot Learning

Training on Sampled Pairs: The previous sections display the exceptional performance of our selection algorithms on cluster-based graphs. Next, we would like to further evaluate the performance of SP on a more common data such as images and features. This analysis is motivated by the work in [55], as we employ their proposed neural network architecture named Siamese neural network. Moreover, we adopt the Omniglot dataset and split it into three subsets for training, validation, and test, each of which consists of totally different classes. For training and validation process two images are randomly sampled from their own corresponding data and are fed as the input to the Siamese neural network and a binary label is assigned to each pair according to the classes that they are sampled from. The network trained on these pairs achieves 90%+ accuracy in distinguishing inter-class and intra-class pairs.

Classification with Few-Shot Learning: After being fully trained on the sampled pairs, the model is further developed for few-shot classification. In other words, if the model is accurate enough to distinguish the identity of classes from which the pairs come from, given few representatives of a specific class, a trained Siamese network could serve as a binary classifier that verifies if the test instance belongs to this class. Therefore, the problem reduces to selecting the best representatives of every class to be paired with any test images. The class that produces the pairings with the highest average score is then identified as the classification result. This subset selection problem can be addressed by our SP algorithm. The test set of Omniglot after splitting comprises 352 different classes, each of which is composed of around 20 images. We sequentially deploy our algorithm on every one of the 352 classes to choose the most informative subset of the 20 images. The classifier made from the Siamese network and the selected 352 representative groups are then evaluated on all the 7,000+ images in the test set. An example of selected groups and the few-shot learning results are illustrated in Figure 7.

It can be observed in Figure 7 that images selected by the evaluated algorithms are generally more standard and more identifiable than the others. Among all these competitor algorithms, SP makes the best selection for this character. Specifically, both GIGA [43] and FW [43] pick the

Table 2: Performance of different state-of-the-art coreset construction algorithms for Graph summarization (central vertex selection) on various types of graphs. Practically all major social networks provide social clusters for instance, 'circles' on Google+, and 'lists' on Facebook and Twitter. For example, concerning Facebook ego graph, with SP algorithm we define the task of identifying users' social clusters on a user's ego-network by exploiting the network structure. The table shows that our proposed algorithm outperforms other algorithms on all graphs except Florentine.

Graph/Algorithm	RND	IS [44]	VS [30]	FFS [35]	MP [45]	DS3 [25]	IPM [16]	FW [43]	BC [46]	GIGA [43]	KSP
Facebook Ego [47]	0.2960	0.1250	0.2210	0.0142	0.0250	0.0147	0.0140	0.0190	0.0149	0.0145	**0.0130**
Powerlaw Cluster [48]	0.2739	0.2735	0.2732	**0.0167**	0.2701	0.0275	**0.0167**	0.2730	0.0358	0.0296	**0.0167**
Barabasi [49]	0.1630	0.1625	0.0142	0.0184	0.1625	0.0154	0.0156	0.1628	0.0378	0.0169	**0.0122**
Geo [50]	0.0685	0.0674	0.0683	0.0424	0.0493	0.0411	0.0299	0.0673	0.0014	0.0017	**0.0012**
Florentine [51]	0.0026	0.0006	0.0007	**0.0003**	0.001	0.0019	**0.0003**	0.0009	**0.0003**	**0.0003**	0.0004
Karate Club [52]	0.1388	0.0158	0.0326	**0.0117**	0.0146	**0.0117**	**0.0117**	0.0146	**0.0117**	0.0146	**0.0117**
Synthesized Graph	0.1421	0.1430	0.0115	0.0120	0.0143	0.0127	0.0122	0.0143	0.0143	0.0124	**0.0106**

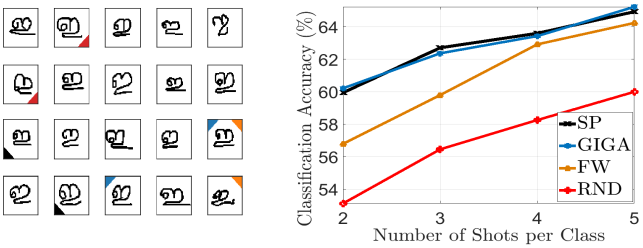


Figure 7: Learning of Omniglot's dataset on Siamese Neural Network using few shots. (Left) Visualization of the 2-image selection from the first class of Omniglot's test set. Images selected by an algorithm are marked in corners with the same color used in the right plot. (Right) Classification accuracy with few-shot learning.

last image of the first row that does not show a clear spiral and the last character of the second row chosen by FW is written huddled. Due to the fact that the classification accuracy is evaluated based on the 352 test classes while they do not appear in the training set, around 60% of correct classification is considerably acceptable. In particular, SP achieves accuracies of 59.84%, 62.70%, 63.55%, and 64.88% for 2-shot, 3-shot, 4-shot, and 5-shot classifications, respectively, which is comparable to the GIGA results of 60.21%, 62.36%, 63.42%, and 65.22% while outperforming other baseline algorithms. This is while SP needs less memory requirement and its computational complexity is less than its peers.

4.5. Open-Set Identification

In this experiment, the open-set identification problem is addressed based on selection which results in significant accuracy improvement compared to the state-of-the-art. In open-set identification, test data of a classification problem may come from unknown classes other than the provided classes at the time of training, and the goal is to identify such samples belong to open-set and not the known labeled classes [56]. Interested readers are referred to [57, 58, 59, 60, 61] to know about the state-of-the-art milestone of approaches towards open-set problem.

Employing the entire closed-set data during the training procedure leads to inclusion of untrustworthy samples of the closed-set. Even regularized or underfitting models may suffer from slightly memorizing the behavior of such samples which exacerbate the separation between open and closed-set by adding ambiguity to the decision boundary between the closed and open-set classes. To resolve this issue, we utilize our proposed selection method, KSP, which selects the core representatives. Therefore, selected representatives provide distinctive open-set determiner as they are more robust in rejecting open-set test samples which do not fit well to the core representatives. We pictorially illustrate the proposed scheme for open-set identification in Fig. 1 on the rightmost panel and the proposed algorithm which is referred to as selection-based open-set identification scheme (SOSIS) hereunder (Algorithm 3).

Algorithm 3 Selection-based open-set identification (SOSIS)

Require: $\boldsymbol{A}^X$ (closed-set training data), and $\boldsymbol{A}^Y=\{\boldsymbol{a}_p^Y\}_{p=1}^P$ (test set)
1: Train a classifier on $\boldsymbol{A}^X$ on H classes
2: $\mathbb{S}_h \leftarrow$ set of K selected samples for class $\#h$ in $\boldsymbol{A}^X$
3: $\ell(p) \leftarrow$ label $\boldsymbol{a}_p^Y$ using trained classifier in Step 1 $(\forall p)$
4: $\text{err}(p) = \|\boldsymbol{a}_p^Y - \pi_{\mathbb{S}_{\ell(p)}}(\boldsymbol{a}_p^Y)\|_2$ $(\forall p)$
5: $c_1,\ c_2 \leftarrow$ perform kmeans on **err** with 2 centroids
Output: open-set$\leftarrow \{\boldsymbol{a}_p^Y \mid \textbf{err}(p) \geq \frac{c_1+c_2}{2}\}$

Experiment Set-up: We use MNIST dataset as the closed-set with samples from Omniglot as the open-set. The ratio of Omniglot to MNIST test dataset is set to $1:1$ (10,000 from each) same as the simulation scenario in [59]. A classifier with ResNet-164 architecture [62] is trained on MNIST as for step 1 in Alg. 3. Results of macro-averaged F1-score [63] for SOSIS method with different selection methods and different number of samples are listed in Table 3 as well as the sate-of-the-art in [59]. The best achieved F1-score is 0.964 belonging to SOSIS with KSP selection using 50 representatives. The second best performance goes to SOSIS with SP selection again using 50 representatives. Performance downgrade is observed for both scenarios of choosing too few representatives such as 5 or fewer and obsessively choosing all data. The gap between the error values resulted from projection of open and closed-set onto selected samples computed in step 4 of Alg. 3 differs significantly compared to that of the projection onto the entire dataset (due to overfitting and memorization effect). We call this splitting property as reflected in Fig. 8 (a) (entire dataset) vs. 8 (b) (selected samples) at the testing phase. For a better visualization, projection errors are sorted separately for closed-set and open-set data at the testing phase. As observed, fewer number of representatives results in higher projection error. However, at the same time *closed-set and*

Table 3: Comparing F1-score of the proposed SOSIS algorithm with state-of-the-art methods for open-set identification. SP, KSP and FFS [35] are employed as the core of SOSIS.

method/K	5	20	50	100	500	All Data
SOSIS (based on FFS)	0.876	0.913	0.944	0.952	0.841	0.792
SOSIS (based on SP)	0.904	0.945	0.958	0.952	0.824	
SOSIS (based on KSP)	0.928	0.959	**0.964**	0.959	0.827	

Supervised only [59]	0.680
LadderNet [59]	0.764
DHRNet [59]	0.793

open-set test data are better split as also observed in Fig. 8. It is worth noting that the threshold also can be assigned by validation if all values in **err** are not available at testing time. In online applications, due to the splitting property one can set the threshold according to error values on closed set test samples benefiting form the splitting property.

Fig. 9 contains the macro-averaged F1-score vs. threshold for different selected representatives using SP data selection. Fine-tuning the open-set identifier by selecting best representatives enhances the accuracy significantly as observed in Fig. 9. As the number of representatives decreases, the performance sensitivity to the threshold adjustment increases which means there is a trade-off between accuracy using selection-based scheme and the stability of performance w.r.t the designed threshold range. Fig. 9 also shows that between 50-100 samples from each training class (each containing about 6000) leads to optimal F1-score.

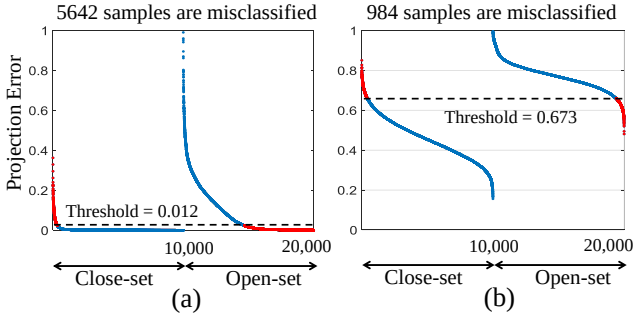


Figure 8: Sorted values of **err** in Step 4 of Alg. 3 for 20,000 test samples (10,000 per each closed/open set). (a) all data are selected as representatives. (b) only 20 representatives are selected. For both (a) and (b), a projection error above/below the threshold leads to classifying a sample as open-set/closed-set. Blue and red points correspond to the correctly-classified and missclassified samples, respectively. As shown, implementing SOSIS enabled by KSP has significantly reduced the number of misclassified samples, from 5642 to 984.

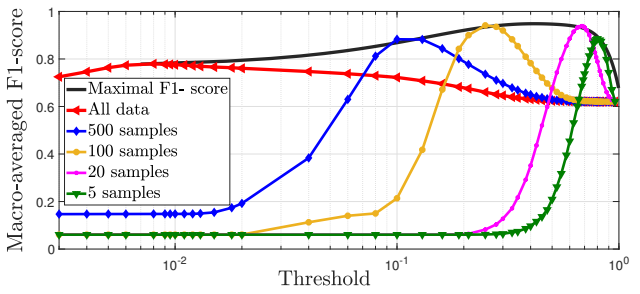


Figure 9: F1-score vs. threshold for different number of selected representatives (Accuracy-Sensitivity Trade-off)

5. Conclusion

A novel approach to data selection from linear subspaces is proposed and its extension for selection from nonlinear manifolds is presented. The proposed SP algorithm demonstrates an accurate solution for CSSP. Moreover, SP and KSP have shown superior performance in many applications. The investigated fast and efficient deep learning frameworks, empowered by our selection methods, have shown that dealing with selected representatives is not only fast but can also be more effective. This manuscript is allocated mostly for algorithm designs and applications of data selection. Theoretical results and more buttressing experiments can be found in the supplementary document.

References

[1] UK. ATT Laboratories, Cambridge. The orl database of faces (now att the database of faces. In *Available online: https://gitdisl.github.io/GTDLBench/datasets/att$_f$ace$_d$ataset/*, 1994.

[2] Ali Çivril. Column subset selection problem is ug-hard. *Journal of Computer and System Sciences*, 80(4):849–859, 2014.

[3] Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 968–977. SIAM, 2009.

[4] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions - I. *Math. Program.*, 14(1):265–294, 1978.

[5] Tony F Chan and Per Christian Hansen. Some applications of the rank revealing qr factorization. *SIAM Journal on Scientific and Statistical Computing*, 13(3):727–741, 1992.

[6] Tony F Chan. Rank revealing qr factorizations. *Linear algebra and its applications*, 88:67–82, 1987.

[7] Jianwei Xiao, Ming Gu, and Julien Langou. Fast parallel randomized qr with column pivoting algorithms for reliable low-rank matrix approximations. In *2017 IEEE 24th International Conference on High Performance Computing (HiPC)*, pages 233–242. IEEE, 2017.

[8] Venkatesan Guruswami and Ali Kemal Sinop. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1207–1214. SIAM, 2012.

[9] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Polynomial time algorithms for dual volume sampling. In *Advances in Neural Information Processing Systems*, pages 5038–5047, 2017.

[10] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1117–1126. Society for Industrial and Applied Mathematics, 2006.

[11] Saurabh Paul, Malik Magdon-Ismail, and Petros Drineas. Column selection via adaptive sampling. In *Advances in neural information processing systems*, pages 406–414, 2015.

[12] Yining Wang and Aarti Singh. Provably correct algorithms for matrix column subset selection with selectively sampled data. *Journal of Machine Learning Research*, 18:1–42, 2018.

[13] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

[14] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[15] Michal Derezinski, Manfred K Warmuth, and Daniel J Hsu. Leveraged volume sampling for linear regression. In *Advances in Neural Information Processing Systems*, pages 2505–2514, 2018.

[16] Alireza Zaeemzadeh, Mohsen Joneidi, Nazanin Rahnavard, and Mubarak Shah. Iterative Projection and Matching: Finding Structure-Preserving Representatives and Its Application to Computer Vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5414–5423, 2019.

[17] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[18] Siddharth Joshi and Stephen Boyd. Sensor Selection via Convex Optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2 2009.

[19] Ali Çivril and Malik Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47-49):4801–4811, 2009.

[20] Zelda E Mariet and Suvrit Sra. Elementary symmetric polynomials for optimal experimental design. In *Advances in Neural Information Processing Systems*, pages 2139–2148, 2017.

[21] Haim Avron and Christos Boutsidis. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.

[22] Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat. Proportional volume sampling and approximation algorithms for a-optimal design. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1369–1386. SIAM, 2019.

[23] Michal Derezinski and Manfred Warmuth. Subsampling for Ridge Regression via Regularized Volume Sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 716–725, 2018.

[24] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1607. IEEE, 2012.

[25] Ehsan Elhamifar, Guillermo Sapiro, and S. Shankar Sastry. Dissimilarity based sparse subset selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2182–2197, 2016.

[26] Jingjing Meng, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan. From Keyframes to Key Objects: Video Summarization by Representative Object Proposal Selection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1039–1048. IEEE, 6 2016.

[27] Huaping Liu, Yunhui Liu, and Fuchun Sun. Robust Exemplar Extraction Using Structured Sparse Coding. *IEEE Transactions on Neural Networks and Learning Systems*, 26(8):1816–1821, 8 2015.

[28] P Comon and G H Golub. Tracking a few extreme singular values and vectors in signal processing. *Proceedings of the IEEE*, 78(8):1327–1343, 1990.

[29] P A Vijaya, M Narasimha Murty, and D K Subramanian. Leaders–Subleaders: An efficient hierarchical clustering algorithm for large data sets. *Pattern Recognition Letters*, 25(4):505–513, 2004.

[30] Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 329–338. IEEE, 2010.

[31] Yong Wang, Yuan Jiang, Yi Wu, and Zhi-Hua Zhou. Spectral clustering on multiple manifolds. *IEEE Transactions on Neural Networks*, 22(7):1149–1161, 2011.

[32] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 5 2010.

[33] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.

[34] P A Vijaya, M Narasimha Murty, and D K Subramanian. Leaders–Subleaders: An efficient hierarchical clustering algorithm for large data sets. *Pattern Recognition Letters*, 25(4):505–513, 2004.

[35] Chong You, Chi Li, Daniel P Robinson, and René Vidal. Scalable exemplar-based subspace clustering on class-imbalanced data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018.

[36] Shin Matsushima and Maria Brbic. Selective sampling-based scalable sparse subspace clustering. In *Advances in Neural Information Processing Systems*, 2019.

[37] Yu Tian, Xi Peng, Long Zhao, Shaoting Zhang, and Dimitris N. Metaxas. CR-GAN: Learning Complete Representations for Multi-view Generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 942–948, California, 7 2018. International Joint Conferences on Artificial Intelligence Organization.

[38] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy. Pose-Robust Face Recognition via Deep Residual Equivariant Mapping. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World. *Electronic Imaging*, 2016(11):1–6, 2 2016.

[40] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016.

[41] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[42] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. Technical report, 2008.

[43] Trevor Campbell and Tamara Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In *International Conference on Machine Learning*, 2018.

[44] Andreas Krause Olivier Bachem, Mario Lucic. Practical coreset constructions for machine learning. *Thesis at Department of Computer Science, ETH Zurich*, 2017.

[45] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 2115–2123, 2011.

[46] Santo Fortunato. Community detection in graphs. *CoRR*, abs/0906.0612, 2009.

[47] Jure Leskovec and Julian J. Mcauley. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems 25*, pages 539–547. 2012.

[48] William Aiello, Fan Chung Graham, and Linyuan Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10(1):53–66, 2001.

[49] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[50] Aric Hagberg Milan Bradonjic and Allon George Percus. Giant component and connectivity in geographical threshold graphs. *in Algorithms and Models for the Web-Graph (WAW), Antony Bonato and Fan Chung (Eds)*, pages 209–216, 9 2007.

[51] Ronald L. Breiger and Philippa E. Pattison. Cumulated social roles: The duality of persons and their algebras. *Social Networks*, 8, 9 1986.

[52] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[53] Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

[54] Wang-Cheng Kang, Eric Kim, Jure Leskovec, Charles Rosenberg, and Julian J. McAuley. Complete the look: Scene-based complementary product recommendation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10532–10541, 2019.

[55] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. 2015.

[56] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

[57] Qianyu Feng, Guoliang Kang, Hehe Fan, and Yi Yang. Attract or distract: Exploit the margin of open set. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7990–7999, 2019.

[58] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8362–8371, 2019.

[59] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4016–4025, 2019.

[60] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8688–8696, 2018.

[61] Trung Pham, Vijay BG Kumar, Thanh-Toan Do, Gustavo Carneiro, and Ian Reid. Bayesian semantic instance segmentation in open set world. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[63] Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maximize f1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–239. Springer, 2014.

[64] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Supplementary Material

The supplementary material provided in this document is organized as follows. In Section A, we present some theoretical results on performance guarantees of SP and KSP. Then, in Section B, further experiments are provided to investigate the performance of the proposed approaches on several different real datasets.

A. Theoretical Results

Theorem 1 expresses that if dataset is grouped into P clusters. SP selects a number of samples according to the importance of each group (subspace). Importance refers to the rank of each subspace. Theorem 2 provides a tight upper bound for CSSP. Finally, Lemma 2 shows that the proposed locally linear selection problem in Equation (5) of the main paper is equivalent to the conventional CSSP where the selection is performed on a similarity matrix instead of raw data.

Theorem 1 *Let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ be collection of M samples in N dimensional space. Assume columns of $\boldsymbol{A}$ can be grouped into P clusters and each cluster forms a k_p-dimensional subspace in which $\sum_{p=1}^{P} k_p = K \leq N$. Selection of K samples using SP provides exactly k_p samples from each cluster.*

Proof of Theorem 1: Let $\mathcal{S}$ denote the span of the selected data and $\mathcal{N}$ be the null space of $\mathcal{S}$. If k_p samples are selected from the p^{th} cluster (subspace) using SP, the projection of all the samples of the corresponding cluster onto $\mathcal{N}$ is $\mathbf{0}$, and SP does not select further samples from the p^{th} cluster anymore. Suppose SP selects n_p samples from the p^{th} cluster. Thus, the number of selected samples from each cluster cannot exceed the dimension of the cluster subspace, i.e., $n_p \leq k_p$. If SP totally selects K samples from the entire data, and the inequality $n_p < k_p$ holds for a cluster, then there exists another cluster, i, for which $n_i > k_i$ that is in contradiction to the previously stated result, $n_p \leq k_p$. Thus, n_p must be equal to k_p for $p = 1, \cdots, P$. ■

Theorem 2 *If the columns of matrix $\boldsymbol{A}$ contain M zero-mean samples in N dimensional space and $\boldsymbol{a}_i$ is the first selected sample using SP, then,*

$$\|\boldsymbol{A} - \pi_i(\boldsymbol{A})\|_F^2 \leq \big(1 + \mathcal{R}_{\boldsymbol{A}}^2(1+\mathcal{R}_{\boldsymbol{A}})(1-\mathcal{R}_{\boldsymbol{A}})\big)\|\boldsymbol{A} - \boldsymbol{A}_1\|_F^2,$$

where $\|\boldsymbol{A} - \pi_i(\boldsymbol{A})\|_F^2$ is the projection error on the span of the selected sample and $\boldsymbol{A}_1$ is the best rank-one approximation.

Obviously, $\|\boldsymbol{A} - \boldsymbol{A}_1\|_F^2$ is the lower bound for projection error based on the definition of SVD. However, this theorem states that the upper bound is a scale (≥ 1) of the lower bound and the scale is $1 + \mathcal{R}_{\boldsymbol{A}}^2(1+\mathcal{R}_{\boldsymbol{A}})(1-\mathcal{R}_{\boldsymbol{A}})$.

Proposition 1 *Assume $\boldsymbol{a}_i$ is the first selected sample using SP. Then,*

$$\|\boldsymbol{A} - \pi_i(\boldsymbol{A})\|_F^2 \leq 1.25\|\boldsymbol{A} - \boldsymbol{A}_1\|_F^2.$$

When $\mathcal{R}_{\boldsymbol{A}} = 1$, the upper bound of projection error is equal to its lower bound since the dataset is rank-one. Thus, any selection (even random selection) provides the same subspace which is equal to the subspace of rank-one approximation. On the other hand, when $\mathcal{R}_{\boldsymbol{A}}$ is too small,[1] distribution of points in the dataset is symmetric. Thus, a specific data point does not have a priority to be selected. Therefore, for such datasets even random selection of a sample provides a close projection error in comparison to the best projection error. In other words, for very low-rank and very high-rank datasets, selection is not challenging and there are trivial solutions. The most challenging scenario for selection of a new sample occurs when $\mathcal{R}_{\boldsymbol{A}} = \sqrt{2}/2$ and the gap between the lower bound and the upper bound is maximized. In this case, the role of selection algorithm is more critical because the dataset is neither highly structured nor symmetrically-spread in the space. Fig. 10 shows the ratio of projection error obtained by selection to projection error using the best rank-1 approximation for several datasets.

[1]Please note that $\mathcal{R}_{\boldsymbol{A}}$ is greater than $1/\sqrt{N}$.

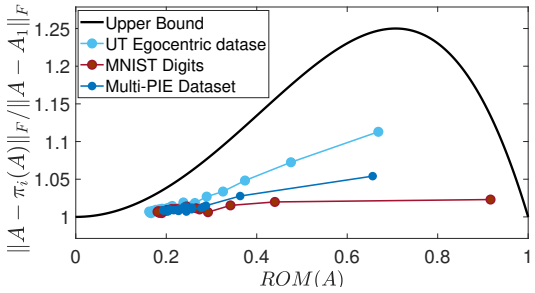


Figure 10: Trajectory of the normalized projection error over iterations of SP algorithm in terms of the rank-oneness measure. The projection error is normalized by projection error of the best rank-1 approximation. This ratio is surely greater than 1. However, we have shown it is less than 1.25 for our algorithm. In each iteration, matrix $\boldsymbol{A}$ is assumed to be the projection of entire data onto null space of previously selected data and one sample is selected.

Before jumping to proof of Theorem 2, we need to borrow the following definition and lemma from [16].

Definition 1 *[16] Rank-oneness measure (ROM) for matrix $\boldsymbol{A}$ with singular values $\sigma_1, \sigma_2, \ldots, \sigma_R$ is defined as $\mathcal{R}_{\boldsymbol{A}} = \sqrt{\frac{\sigma_1^2}{\sum_{r=1}^{R}\sigma_r^2}} = \frac{\sigma_1}{\|\boldsymbol{A}\|_F}$.*

Lemma 1 *[16] Let $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M \in \mathbb{R}^N$ be M given data points of dimension N organized as columns of $\boldsymbol{A}$. Let σ_1, $\boldsymbol{u}$ and $\boldsymbol{v}$ denote the first singular value, the corresponding left and right singular vectors of $\boldsymbol{A}$, respectively. Then, there exists at least one data point such that the absolute value of its inner product with $\boldsymbol{u}$ is greater than or equal to $\frac{\sigma_1}{\sqrt{M}}$. Hence, $\max_m |\boldsymbol{a}_m^T\boldsymbol{u}| \geq \frac{\sigma_1}{\sqrt{M}}$.*

Proof of Theorem 2: Matrix $\boldsymbol{A}$ can be written in terms of its singular components as follows,

$$\boldsymbol{A} = \boldsymbol{u}_1\boldsymbol{u}_1^T\boldsymbol{A} + \cdots + \boldsymbol{u}_N\boldsymbol{u}_N^T\boldsymbol{A}.$$

The scaled version of selected data is decomposed in terms of the first LSV and a vector which is orthogonal to it, denoted by $\boldsymbol{u}_\perp$. Mathematically,

$$\tilde{\boldsymbol{a}}_i = \frac{\boldsymbol{u}_1 + \alpha\boldsymbol{u}_\perp}{\sqrt{1+\alpha^2}}. \tag{7}$$

Projection of $\boldsymbol{A}$ on the space of the selected data can be cast as follows,

$$\begin{aligned}\pi_i(\boldsymbol{A}) =& \tilde{\boldsymbol{a}}_i\tilde{\boldsymbol{a}}_i^T\boldsymbol{A}/\tilde{\boldsymbol{a}}_i^T\tilde{\boldsymbol{a}}_i \\ =& \frac{1}{(1+\alpha^2)}\boldsymbol{u}_1\boldsymbol{u}_1^T\boldsymbol{A} + \frac{\alpha^2}{(1+\alpha^2)}\boldsymbol{u}_\perp\boldsymbol{u}_\perp^T\boldsymbol{A}.\end{aligned} \tag{8}$$

Note that $\boldsymbol{u}_\perp$ is a normalized vector perpendicular to $\boldsymbol{u}_1$. Consequently, the projection error can be presented in terms of singular components as follows.

$$\|\boldsymbol{A}-\pi_i(\boldsymbol{A})\|_F^2 = \tag{9}$$
$$\|\boldsymbol{u}_1\boldsymbol{u}_1^T\boldsymbol{A}+...+\boldsymbol{u}_N\boldsymbol{u}_N^T\boldsymbol{A}-\frac{1}{(1+\alpha^2)}\big(\boldsymbol{u}_1\boldsymbol{u}_1^T\boldsymbol{A}+\alpha^2\boldsymbol{u}_\perp\boldsymbol{u}_\perp^T\boldsymbol{A}\big)\|_F^2.$$

It is straightforward to show that $\|\boldsymbol{A}-\pi_i(\boldsymbol{A})\|_F^2$ is minimized if $\boldsymbol{u}_\perp = \boldsymbol{u}_N$. Let $\mathrm{Res}_{\min}$ denote the minimum value of $\|\boldsymbol{A}-\pi_i(\boldsymbol{A})\|_F^2$. Thus, $\mathrm{Res}_{\min}$ is equal to

$$\|\boldsymbol{u}_1\boldsymbol{u}_1^T\boldsymbol{A}+...+\boldsymbol{u}_N\boldsymbol{u}_N^T\boldsymbol{A}-\frac{1}{(1+\alpha^2)}\big(\boldsymbol{u}_1\boldsymbol{u}_1^T\boldsymbol{A}+\alpha^2\boldsymbol{u}_N\boldsymbol{u}_N^T\boldsymbol{A}\big)\|_F^2 \tag{10}$$

It is clear that $\|\boldsymbol{A}-\boldsymbol{A}_1\|_F^2$ is a lower bound for $\|\boldsymbol{A}-\pi_i(\boldsymbol{A})\|_F^2$. Therefore,

$$\|\boldsymbol{A}-\boldsymbol{A}_1\|_F^2 \le \mathrm{Res}_{\min}.$$

On the other hand, there exists a value ϵ such that $\mathrm{Res}_{\min}$ is upper bounded by a factor of its lower bound as follows,

$$\mathrm{Res}_{\min} \le (1+\epsilon)\|\boldsymbol{A}-\boldsymbol{A}_1\|_{\mathrm{F}}^2.$$

Thus, substituting the expansion of $\mathrm{Res}_{\min}$ from (10) we are looking for an ϵ that satisfies the following inequality,

$$\|\boldsymbol{u}_1\boldsymbol{u}_1^T\boldsymbol{A}+...+\boldsymbol{u}_N\boldsymbol{u}_N^T\boldsymbol{A}-\frac{1}{(1+\alpha^2)}\big(\boldsymbol{u}_1\boldsymbol{u}_1^T\boldsymbol{A}+\alpha^2\boldsymbol{u}_N\boldsymbol{u}_N^T\boldsymbol{A}\big)\|_F^2 \le (1+\epsilon)\|\boldsymbol{A}-\boldsymbol{A}_1\|_F^2,$$

which is equivalent to

$$\frac{\alpha^4}{(1+\alpha^2)^2}\sigma_1^2+\sigma_2^2+\sigma_3^2+\cdots+\sigma_{N-1}^2+\frac{1}{(1+\alpha^2)^2}\sigma_N^2 \le (1+\epsilon)(\sigma_2^2+\cdots+\sigma_N^2). \tag{11}$$

Since the data is pre-processed to be zero-mean, $\sigma_N = 0$ and (11) can be simplified as

$$\frac{\alpha^4}{(1+\alpha^2)^2}\sigma_1^2 \le \epsilon(\|\boldsymbol{A}\|_F^2-\sigma_1^2) \tag{12}$$

By dividing both sides to σ_1^2, the right side can be cast in terms of ROM($\boldsymbol{A}$) as

$$\frac{\alpha^4}{(1+\alpha^2)^2} \le \epsilon(\frac{1-\mathcal{R}_A^2}{\mathcal{R}_A^2}).$$

Now let us write the right side in terms of correlation of the first left singular vector and i^{th} data which is selected as stated in (7) (their correlation is indicated by c).

$$\frac{(c^2-1)^2/c^4}{1/c^4} \le \epsilon(\frac{1-\mathcal{R}_A^2}{\mathcal{R}_A^2}).$$

According to Lemma 1 the correlation is lower-bounded by $\mathcal{R}_A$. Thus,

$$(1-\mathcal{R}_A)^2(\mathcal{R}_A+1)^2 \le \epsilon(\frac{1-\mathcal{R}_A^2}{\mathcal{R}_A^2}).$$

And finally ϵ is upper bounded in terms of ROM of $\boldsymbol{A}$ as follows,

$$\epsilon \ge \mathcal{R}_A^2(1+\mathcal{R}_A)(1-\mathcal{R}_A).$$

Since $\mathcal{R}_A$ is bounded between 0 and 1, each $\epsilon \ge \frac{1}{4}$ which is the maximum of the right side establishes the desired upper bound. ■

Lemma 2 *Consider M data points and the neighborhood for each one are denoted by $\boldsymbol{a}_m$ and Ω_m, respectively. The following problems have the same selection results using the SP algorithm.*

$$P1: \underset{|\mathbb{S}|\le K}{argmin} \sum_{m=1}^{M}\|\boldsymbol{a}_m-\pi_{\mathbb{S}_m}(\boldsymbol{a}_m)\|_F^2 \;\; s.t.\; \mathbb{S}_m \subseteq \mathbb{S}\cap\Omega_m,$$

and,

$$P2: \underset{|\mathbb{S}|\le K}{argmin} \|\boldsymbol{H}-\pi_{\mathbb{S}}(\boldsymbol{H})\|_F^2,$$

where $h_{ij} = [\,|\Omega_i\cap\Omega_j|\boldsymbol{a}_i^T\boldsymbol{a}_j]$.

Proof of Lemma 2: Matrix $\boldsymbol{X}_m \in \mathbb{R}^{N\times M}$ is defined as an all-zero matrix except in rows indexed by Ω_m. The non-zero rows are equal to $\boldsymbol{a}_m^T$ (repeated for all those rows). Matrix $\boldsymbol{X} \in \mathbb{R}^{MN\times M}$ is defined as follows,

$$\boldsymbol{X} = [vec(\boldsymbol{X}_1),\cdots,vec(\boldsymbol{X}_M)].$$

Operator $vec(.)$ reshapes a matrix to a vector. Using the definition of $\boldsymbol{X}$, Problem P1 can be cast in terms of $\boldsymbol{X}$ as follows,

$$\underset{|\mathbb{S}|\le K}{\text{argmin}} \|\boldsymbol{X}-\pi_{\mathbb{S}}(\boldsymbol{X})\|_F^2.$$

It is straightforward to show that the k^{th} left singular vector of $\boldsymbol{X}^T\boldsymbol{X}$ is proportional to $\boldsymbol{X}^T\boldsymbol{u}_k$, where $\boldsymbol{u}_k$ is the k^{th} left singular vector of $\boldsymbol{X}$. Given the singular value decomposition of $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices and $\boldsymbol{\Sigma}$ is the diagonal matrix of singular values, one can write $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T$. Thus, the k-th left eigenvector is aligned with $\boldsymbol{v}_k$ which is the k-th column of $\boldsymbol{V}$. Similarly,

$$\boldsymbol{X}^T\boldsymbol{u}_k = \sum_{i=1}^{rank(\boldsymbol{X})}\sigma_i\boldsymbol{v}_i\boldsymbol{u}_i^T\boldsymbol{u}_k = \sigma_k\boldsymbol{v}_k, \tag{13}$$

where the last equality follows from orthogonality of $\boldsymbol{U}$. Now, it is trivial that the k-th left eigenvector of $\boldsymbol{X}^T\boldsymbol{X}$ is aligned with $\boldsymbol{X}^T\boldsymbol{u}_k$ (both aligned with $\boldsymbol{v}_k$).

As the following step of the proof, we proceed to state that the same data index, m, which maximizes $|\boldsymbol{x}_m^T\boldsymbol{u}_k|$ (as in Eq. 4(b) in the main text) also maximizes $|\boldsymbol{h}_m^T\boldsymbol{X}^T\boldsymbol{u}_k|$, where $\boldsymbol{h}_m$ is the m^{th} column of $\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{X}$. This can be proved as follows: m^* is in fact given as $m^* = \underset{m}{\text{argmax}}\,|\boldsymbol{x}_m^T\boldsymbol{u}_k|$, i.e., the index which picks the largest magnitude in vector $\boldsymbol{X}^T\boldsymbol{u}_k$. Similarly, one can write $\boldsymbol{h}_m^T\boldsymbol{X}^T\boldsymbol{u}_k = (\boldsymbol{X}^T\boldsymbol{X})_m\boldsymbol{X}^T\boldsymbol{u}_k$ and therefore, $m^* = \underset{m}{\text{argmax}}\,|\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{u}_k|$. We also have,

$$\underbrace{\boldsymbol{X}^T\boldsymbol{X}}_{\boldsymbol{H}}\overbrace{\boldsymbol{X}^T\boldsymbol{u}_k}^{k^{th}\text{ left singular vector of }\boldsymbol{H}} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{u}_k = \boldsymbol{V}\boldsymbol{\Sigma}^3\boldsymbol{U}^T\boldsymbol{u}_k = \sigma_k^3\boldsymbol{v}_k. \tag{14}$$

This means both optimization mentioned above on index m result in the same function with a difference in scale which does not affect the solution. Therefore, selection with SP results in the same selection by solving the following problem as solution of *P1*.

$$\underset{|\mathbb{S}|\le K}{\text{argmin}} \|\boldsymbol{H}-\pi_{\mathbb{S}}(\boldsymbol{H})\|_F^2.\blacksquare$$

Matrix $\boldsymbol{H}$ is equal to the weighted replica of auto-correlation matrix of data, $\boldsymbol{A}^T\boldsymbol{A}$. The weights come from the neighborhood information. For example, if data i and data j are not neighbors,

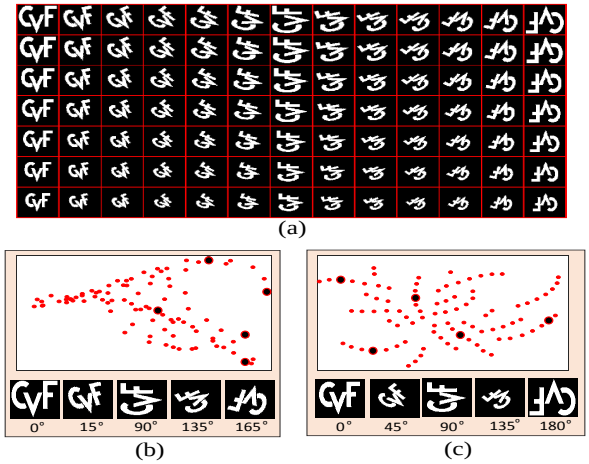


Figure 11: (a) A dataset lies on a two dimensional manifold identified by two parameters, rotation and size. However, the rank of corresponding matrix to this dataset is a large number. (b) Linear embedding using linear PCA and selection using linear SP. (c) nonlinear embedding using tSNE[64] and selection using kernel-SP. Un-selected and selected samples are shown as red and black dots in the embedded space, respectively.

then $h_{ij} = 0$. And if they share P neighbors then $h_{ij} = P\boldsymbol{a}_i^T\boldsymbol{a}_j$. Matrix $\boldsymbol{H}$ is a similarity matrix and any other graph-based similarity matrix is reasonable to substitute $\boldsymbol{H}$. In the main paper, we employ normalized similarity matrix, the definition of which is inspired by Laplacian graph of neighborhood. This choice is a conventional similarity matrix in the context of manifold-based dimension reduction. Moreover, it can be employed easily for graph summarization which is investigated in the main manuscript. The neighborhood and weighting in definition of matrix $\boldsymbol{H}$ is hard, while the normalized similarity matrix based on Gaussian kernel provides a soft neighborhood definition via smooth weighting. Employing the normalized similarity matrix results in Problem (6) in the main paper.

Fig. 11 illustrates the impact of nonlinear modeling on a toy example containing a set of 100×100 images where each image is a rotated and resized version of other images (Fig. 11(a)). Since none of the images lie on the linear subspace spanned by the rest of images, the ensemble of these data do not form a linear subspace. Therefore, this dataset is of high rank and the union of linear subspaces is not a proper underlying model for it. The KSP algorithm is implemented using a Gaussian kernel with parameter α, i.e., $s_{ij} \triangleq e^{-\alpha\|\boldsymbol{a}_i-\boldsymbol{a}_j\|^2}$. As shown in Fig. 11 (c), the nonlinear selection algorithm has been able to discover the intrinsic structure of data and select data from more distinguished angles than that of Fig. 11 (b) in which the plain SP is applied.

B. Supplementary Experiments

Further experiments in this section support experiments of the main paper.

B.1. Convergence of SP

Provably convergent version of SP algorithm needs a slight modification in the algorithm which is out of scope of this material. However, lots of experiments show that the proposed SP algorithm in the main paper converges in less than $5K$ iterations for selecting K samples. Fig. 12 and Fig. 13 show convergence behavior of SP and KSP for selecting from multi-pie face data set and Cora citation dataset within less than $5K$ iterations.

B.2. GAN on Multi-pie Face Dataset

As it is discussed in the main paper, we select only 9 images from each subject (1800 total subjects), and train the network with the reduced dataset for 300 epochs using the batch size of 36. Fig. 14 shows the generated images of a subject in the testing set, using the trained network on the reduced dataset, as well as using the complete dataset. The network trained on samples selected by KSP (fifth row) is able to generate more realistic images, with fewer artifacts, compared to other selection methods (rows 1-4). The parameter of KSP is set as $1e-4$ for constructing the similarity matrix.

B.3. Graph Summarization

The motivation for our experiment here which is in line with the experiment provided in Section 4.3 of the paper is to elaborate upon one of the important applications of KSP algorithm i.e., graph summarization. Herein, we aim at comparing the central vertex selection and community detection capability of KSP with other state-of-the-art algorithms provided in table 2 for the Power-law Cluster graph [48] as in Fig 16.

B.4. Training a Classifier using Reduced Data

The t-SNE visualization [64] of the selected representatives for two randomly selected classes of UCF-101 dataset is shown in Fig. 15. The contours represent the decision boundary of an SVM trained using all or a set of selected samples. This exper-

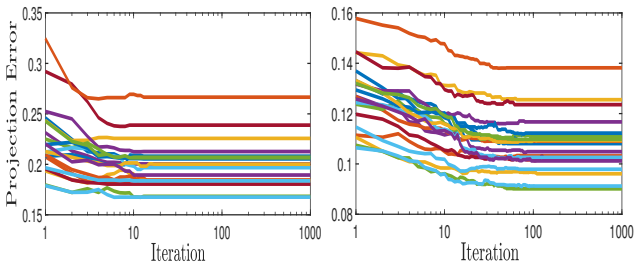


Figure 12: Selecting 5 and 20 representatives from the first 20 classes of Multi-pie dataset. Each class has 520 samples and the error trajectory of each single implementation is depicted in order to show that SP algorithm converges to its solution for each independent selection. (Left) Projection error for selecting 5 samples versus iterations. (Right) Projection error for selecting 20 samples versus iterations. Typically, SP selects K representatives in $5K$ iterations.

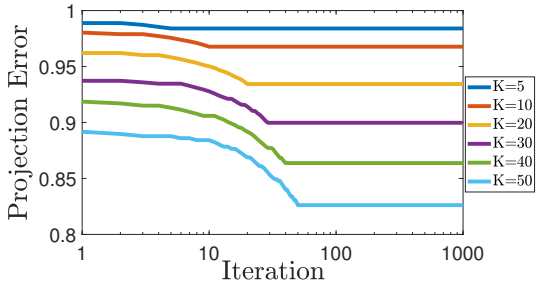


Figure 13: Selecting different number of nodes from Cora dataset which is a graph-based dataset. SP on the similarity matrix of this graph converges in only K iterations which is the minimum number of iterations for updating K selected nodes.

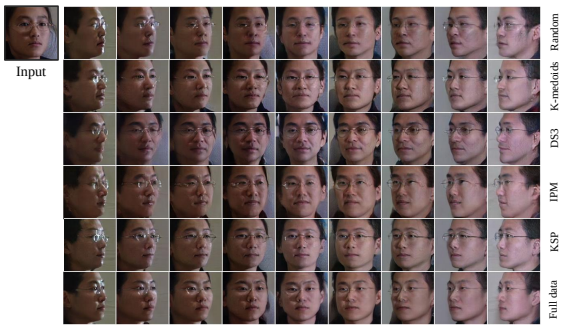
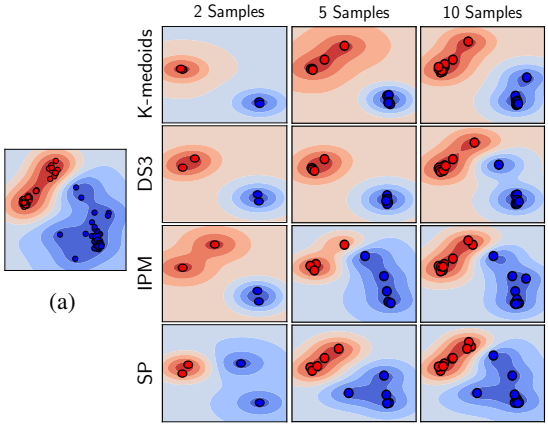


Figure 14: Multi-view face generation results for a sample subject in testing set using CR-GAN [37]. The network is trained on a selected subset of training set (9 images per subject) using random selection (first row), K-medoids (second row), DS3 [25] (third row), IPM (fourth row), and our proposed KSP algorithm. The sixth row shows the results generated by the network trained on all the data (360 images per subject). KSP generates closest results to the complete dataset. In the main paper, a quantitative measure is studied for comparing the generated images and the ground truth from different viwes.



(a)

(b)

Figure 15: t-SNE visualization of two classes of UCF-101 dataset and their representatives selected by different methods. (a) Decision boundary learned using entire data of both classes. (b) Decision boundary obtained using 2, 5, and 10 representatives per class, employing K-medoids (first row), DS3 [25] (second row), IPM [16] (third row), and SP (fourth row).

iment illustrates that SP represent the actual boundary of classes in the t-SNE space more accurately, comparing with other selection methods since its boundary is closer to the boundary which is obtained by entire data. IPM algorithm [16] which is a greedy algorithm has no option to revise the selected samples. Therefore, selecting 10 samples results in the result of selecting 5 samples using IPM plus 5 more selected samples. Similarly, the selected 2 samples using IPM also are present in the selection result for 5 and 10 samples. However, SP optimizes the selection for the given desired K. Thus, selecting 2 samples results an independent selection comparing to selecting 5 and 10 samples using SP. This phenomena is depicted in Fig. 15. As it can be seen, IPM keeps the previously selected samples and choose more greedily. While, SP can result in different selection for different values of K.

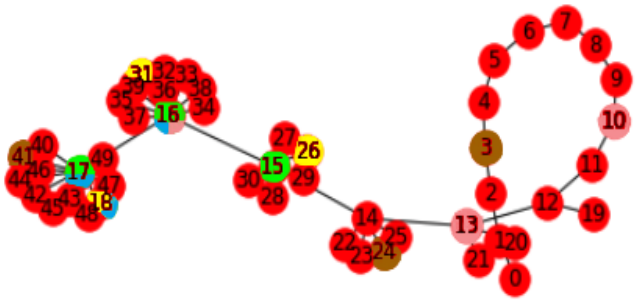


Figure 16: We apply KSP and other algorithms as in table 2, to choose three of the main vertices from another graph, i.e., Powerlaw Cluster graph as was provided in table 2. The results are: GIGA, MP and FW select ●, IS selects ●, VS selects ●, DS3 selects ●, and KSP and FFS select ●. As is evident, KSP and FFS are the only ones that are able to detect the clusters and their corresponding vertices.

B.5. Open-set Identification

It is worth noting that in some contexts, open-set is defined as the set containing both known and unknown classes. In this paper, *we have assumed that open-set is only used for the unknown classes* and the known classes at the time of training are called the closed-set.

Here, we provide a discussion on how to select the threshold in the open-set identification experiment setup. In Fig. 18, a network is trained on the MNIST training data as shown on the left partition. Next, the validation data consisting of data from both the known and unknown classes is used to find the threshold as in algorithm 3 in the main text.

In Fig. 17, the receiver operating characteristic (ROC) of area under the curve (AUC) is plotted for the KSP method in the open-set identification. Different number of selected representatives in the proposed SOSIS algorithm (Alg. 3 in the main text) are considered. Sweeping through the threshold range, the ROC-AUC is achieved for SOSIS algorithm with each desired number of selected samples. As observed and magnified in Fig. 17, the best ROC-AUC performance (higher in plot) is achieved for about $20 - 50$ number of selected representatives.

In Fig. 18, threshold selection is investigated with more scrutiny. At the time of test, a pre-determined threshold is required for deciding on test samples. Our proposed method works based on accessing a set of error values by splitting them and deciding on the threshold. Using one test sample at a time does not lead to a set of error values for splitting at a time. Therefore, one can simply assign the threshold to be a value slightly larger than maximum of error values relating to projecting training samples on selected representatives from each class. Alternatively, if the learning framework is allowed to access validation data, the threshold can be achieved by clustering error values in the balanced validation data into two groups with two centroids, and then taking their average (1:1 sample ratio for Omniglot and MNIST in our case).

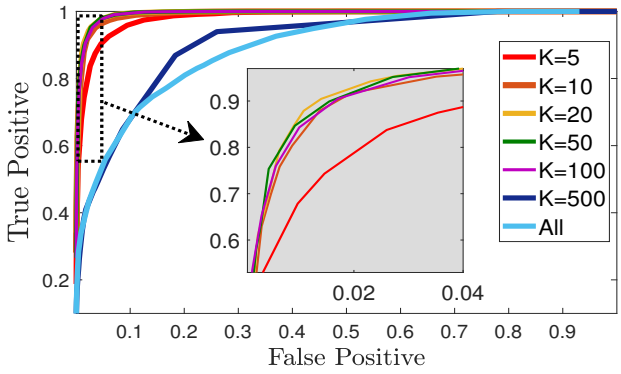


Figure 17: ROC of the proposed selection-based open-set identification employing KSP. The parameter of KSP for constructing the similarity matrix is set to 0.6.

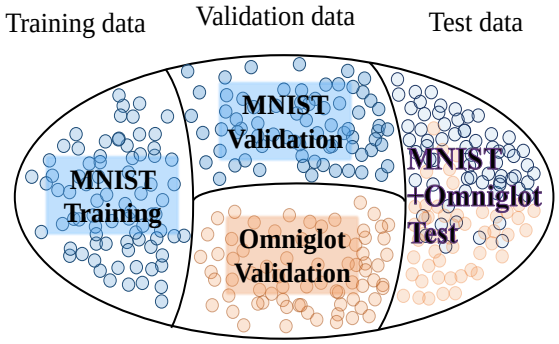


Figure 18: In some scenarios where we access to a validating set, a reliable threshold can be estimated. In this case, validating data from the open-set are not engaged for training the classifier and they are only employed for estimating an optimal threshold.