E-optimal Sensor Selection for Compressive Sensing-based Purposes

Mohsen Joneidi, Alireza Zaeemzadeh, Behzad Shahrasbi, Guo-Jun Qi, and Nazanin Rahnavard

Abstract—Collaborative estimation of a sparse vector $\mathbf{x}$ by M potential measurements is considered. Each measurement is the projection of $\mathbf{x}$ obtained by a regressor, i.e., $y_m = \mathbf{a}_m^T\mathbf{x}$. The problem of selecting K sensor measurements from a set of M potential sensors is studied where $K \ll M$ and K is less than the dimension of $\mathbf{x}$. In other words, we aim to reduce the problem to an under-determined system of equations in which a sparse solution is desired. This paper suggests selecting sensors in a way that the reduced matrix construct a well conditioned measurement matrix. Our criterion is based on E-optimality, which is highly related to the restricted isometry property that provides some guarantees for sparse solution obtained by ℓ_1 minimization. The proposed basic E-optimal selection is vulnerable to outlier and noisy data. The robust version of the algorithm is presented for distributed selection for big data sets. Moreover, an online implementation is proposed that involves partially observed measurements in a sequential manner. Our simulation results show the proposed method outperforms the other criteria for collaborative spectrum sensing in cognitive radio networks (CRNs).

Our suggested selection method is evaluated in machine learning applications. It is used to pick up the most informative features/data. Specifically, the proposed method is exploited for face recognition with partial training data.

Index Terms—Sensor Selection, E-optimality, Restricted Isometry Property (RIP), and Sparse Recovery

1 Introduction

Complex systems containing very large numbers of data-gathering devices, were developed in the last decade. However, dealing with large number of sources of data is challenging. The emerging research area, *big data*, aims to address challenges of such complex systems. Representing the underlying structure of data by a succinct format is a crucial issue in the big data literature. For instance, dimension reduction techniques and different clustering-based approaches aim to extract a concise format of data. Representatives obtained by such methods are often not easy to interpret. Furthermore, obtaining each representative implies processing of all data or a large portion of data. In order to have a straightforward interpretation, it is desired to find the representatives by selection from data. There are some clustering approaches that select the representatives from data such as k-medoids clustering [1]. However these clustering methods assign each data to only one prototype which is the cluster representer, while in the case of highly structured data only one prototype from data does not contain sufficient information to capture the underlying structure of the whole cluster.

An example of big data system is wireless sensor networks, where the processing unit has to deal with an excessively large number of observations acquired by the various sensors. Often there exist some redundancies within the sensed data and they should be pruned. Sensor selection and sensor scheduling aim to address this problem. In many applications the sensor selection task is non-trivial and possibly consists of addressing an NP-hard problem (i.e., there are $\binom{M}{K}$ possibilities of choosing K distinct sensors out of M available ones). This essentially implies that an optimal solution cannot be efficiently computed, in particular when the number of sensors becomes excessively large. A convex relaxation of the original NP-hard problem has been suggested in [2]. The most prominent advantage of this approach over other methods is its practicality, thanks to many well-established computationally-efficient convex optimization techniques. In addition to convex relaxation, a sub-modular cost function as the criterion of sensor selection allows us to take advantage of greedy optimization methods for selecting sensors [3]. The existing studies on sensor selection mostly consider heuristic approaches. For example, in [2] the volume of the reduced bases is considered. This method is called *D-optimality*. In addition, *A-optimality* [4] and *E-optimality* [4] are suggested as some other alternative heuristics already introduced in convex optimization. These heuristics are presented without any specific justification for sensor selection application. In this paper we are going to exploit a criteria more judiciously in favor of compressed sensing (CS) theoretical guarantees.

Roadmap: This paper reviews existing work on sensor selection and matrix subset selection. Relation of these two topics is elaborated. Then, a new method for matrix subset selection is proposed which is equivalent to our proposed sensor selection algorithm. The distributed and robust implementation is also presented. The performance bound of the proposed scheme is derived and its applicability is studied for two practical cases.

Table 1 presents the employed notations throughout this paper. The rest of paper is organized as follows. Section 2 illustrates the motivation of sensor selection inspired by

• *Mohsen Joneidi, Alireza Zaeemzadeh, Behzad Shahrasbi, and Nazanin Rahnavard are with the Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL, 32816.*
E-mails: {joneidi, zaeemzadeh, behzad, and nazanin}@eecs.ucf.edu
This material is based upon work supported by the National Science Foundation under Grant No. CCF-1718195.

TABLE 1: Employed notations in this paper.

Variable Type	Notation
Constant Scalar	X
Vector	$\boldsymbol{x}$
Matrix	$\boldsymbol{X}$
Set	$\mathbb{X}$
Selected Rows of $\boldsymbol{A}$ by set $\mathbb{X}$	$\boldsymbol{A}_{\mathbb{X}}$
Number of non-zero entries of $\boldsymbol{x}$	$\|\boldsymbol{x}\|_0$
Number of non-zero rows of $\boldsymbol{X}$	$\|X\|_{2,0}$
Trace of Matrix $\boldsymbol{X}$	$\mathrm{Tr}(\boldsymbol{X})$
Expectation of $\boldsymbol{X}$	$\mathbb{E}\{X\}$
Singularvalue of $\boldsymbol{X}$	$\sigma(\boldsymbol{X})$

compressed sensing theory. Section 3 states the problem of sensor selection and reviews some existing methods. E-optimal sampling is introduced in Section 4 and a new sensor selection method is proposed. The extension of E-optimal sensor selection to RIP-based sensor selection is presented in Section 5. Section 6 proposes the distributed implementation. Section 7 presents the simulation results and Section 8 concludes the paper.

2 Motivation

Compressed sensing is a technique by which sparse signals can be measured at a rate less than conventional Nyquist sampling theorem [5, 6]. There exist vast applications of CS in signal and image processing [7], channel estimation [8], cognitive radio [9] and spectrum sensing [10]. CS aims to recover a sparse vector, $\mathbf{x}$, using a small number of measurements $\mathbf{y}$. The CS problem can be formulated as,

$$\hat{\mathbf{x}} = \underset{x}{\operatorname{argmin}} \|\mathbf{x}\|_0 \quad \text{s.t.} \;\; \mathbf{y} = \mathbf{\Phi}\mathbf{x}, \tag{1}$$

where, $\|.\|_0$ represents the number of non-zero elements of a vector. $\mathbf{\Phi} \in \mathbb{R}^{K\times N}$ is called measurement matrix that provides us K measurements collected in $\boldsymbol{y}$. These measurements sense from an unknown N dimensional vector. Exact solution of the above optimization problem is through the combinational search among all possible subsets. Due to its high computational burden, this algorithm is impractical for high dimension scenarios. Many sub-optimal algorithms have been proposed such as OMP [11], smoothed ℓ_0 [12] and basis pursuit [13]. Basis pursuit is based on relaxing ℓ_0 to ℓ_1 norm and is popular due to theoretical guarantees and reasonable computational burden [14]. The theoretical guarantees for ℓ_1 minimization arise from several sufficient conditions based on some suggested metrics. These include the mutual coherence [15], null space property [16], spark [17] and restricted isometery property (RIP) [18]. Except for the mutual coherence, none of these measures can be efficiently calculated for an arbitrary given measurement matrix $\mathbf{\Phi}$. For example, the RIP requires enumerating over an exponential number of index sets. RIP is defined as follows.

Definition 1. [18] A measurement matrix is said to satisfy symmetric form RIP of order S with constant δ_S if δ_S is the smallest number that

$$(1-\delta_S)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi}\mathbf{x}\|_2^2 \leq (1+\delta_S)\|\mathbf{x}\|_2^2, \tag{2}$$

holds for every S-sparse $\mathbf{x}$ (i.e. $\mathbf{x}$ contains at most S nonzero entries).

Based on this definition several guarantees are proposed in terms of δ_{2S}, δ_{3S} and δ_{4S} in [19] and [20] in order to guarantee recovering S-sparse vectors. By S-sparse we mean a vector that has S non-zero entries. In [21] an asymmetric form of definition 1 is introduced in order to more precisely quantify the RIP.

Definition 2. [21] For a measurement matrix the asymmetric RIP constants δ_S^L and δ_S^U are defined as,

$$\begin{aligned} \delta_S^L(\mathbf{\Phi}) &= \underset{c\geq 0}{\operatorname{argmin}} \; (1-c)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi}\mathbf{x}\|_2^2, \;\; \forall \mathbf{x} \in \mathcal{X}_S^N, \\ \delta_S^U(\mathbf{\Phi}) &= \underset{c\geq 0}{\operatorname{argmin}} \; (1+c)\|\mathbf{x}\|_2^2 \geq \|\mathbf{\Phi}\mathbf{x}\|_2^2, \;\; \forall \mathbf{x} \in \mathcal{X}_S^N, \end{aligned} \tag{3}$$

where, $\mathcal{X}_S^N$ refers to the set of S-sparse vectors in $\mathbb{R}^N$.

Remark 1. [21] Although both the smallest and largest singular values of $\mathbf{\Phi_S}^T\mathbf{\Phi_S}$ [1] affect the stability of the reconstruction algorithms, the smaller eigenvalue is dominant for compressed sensing in that it allows distinguishing between sparse vectors, $\mathcal{X}_S^N$, given their measurements by $\mathbf{\Phi}$.

This paper suggests to design a sensor selection method inspired by the RIP of a matrix. The goal is to reduce a measurement matrix to only a small fraction of its rows, while optimizing the proposed RIP-based criterion. In other words we aim to reduce number of equations such that the reduced system of equations would be a well-conditioned inverse problem.

For many scenarios, the big data are modeled by matrices and tensors. While conventional numerical algebra has been of interest for decades in many fields of sciences, it has been revisited for analysis of large datasets. For example algebraic tools such as singular value decomposition and subspace clustering are well-known methods for data mining, however their essential considerations for big data analysis are studied recently under the context of big data [22–24]. To this aim, parallel, distributed, scalable, and randomized algorithms are developed based on novel optimization strategies such as ADMM (alternating direction method of multipliers) [25–27]. Selection strategies are helpful for big data analysis and there is a strong connection between matrix subset selection and other analysis methods based on low-rank data expression [28]. A modified matrix subset selection is proposed in Chapter III of [29] in which big data considerations are addressed by a randomized approach. In this paper, a successive and a parallel algorithm are proposed to tackle big data scenarios. The parallel algorithm is designed based on distribution of data on machines. Moreover, theoretical bounds are studied.

The main contributions of the paper are summarized as,

- The link between matrix subset selection, especially volume sampling and sensor selection, is investigated,
- A new criterion for matrix subset selection is proposed, which results in a new sensor selection method,
- The suitability of the E-optimal criterion is discussed, which is equivalent to optimization of an upper

1. $\mathbb{S}$ represents a set with cardinality of S and $\mathbf{\Phi_S}$ represents any combination of columns of $\mathbf{\Phi}$.

bound for RIP coefficients in compressive sensing literature,
- An approximation for RIP coefficients is proposed and utilized to extend E-optimality to an RIP-based criteria, and
- Successive and parallel algorithms are proposed as practical algorithms for selection from large data sets. Their performances are compared with the centralized algorithm.

3 Problem Statement and Related Work

Solving the sensor selection problem by evaluating the performance for each of the possible choices of $\binom{M}{K}$ is impractical unless the sizes are sufficiently small.

Suppose we want to estimate a vector $\mathbf{x} \in \mathbb{R}^N$ from M linear measurements, corrupted by additive noise, given by

$$\mathbf{y} = \boldsymbol{A}\mathbf{x} + \boldsymbol{\nu}, \tag{4}$$

where, $\mathbf{y} \in \mathbb{R}^M$ and $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{\nu}$ is normally distributed with zero mean and σ^2 variance. In other words, we want to only select just K rows of $\boldsymbol{A}$ to have K measurements out of maximum M measurements. The maximum likelihood (ML) estimator is given by [2],

$$\hat{\mathbf{x}}_{ML} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\mathbf{y}. \tag{5}$$

The estimation error $\mathbf{x}-\hat{\mathbf{x}}$ has zero mean and the covariance matrix is equal to

$$\boldsymbol{\Sigma}_{ML} = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}. \tag{6}$$

To involve selection operator in the equations let us first write the ML solution as follows,

$$\hat{\mathbf{x}}_{ML} = (\sum_{m=1}^{M} \mathbf{a}_m\mathbf{a}_m^T)^{-1} \sum_{m=1}^{M} y_m\mathbf{a}_m, \tag{7}$$

where, $\mathbf{a}_m^T$ is the m^{th} row of $\boldsymbol{A}$. The estimation error is distributed in a high dimensional ellipsoid that its center is located at origin and its shape is according to the covariance matrix of error [2]. Minimization of volume of this ellipsoid (D-optimality) is the heuristic used in [2] that results in the following problem:

$$\begin{aligned} \hat{\boldsymbol{w}} &= \underset{\boldsymbol{w}}{\text{argmin}} \log\det(\sum_{m=1}^{M} w_m\mathbf{a}_m\mathbf{a}_m^T)^{-1}, \\ &\text{subject to} \quad \|\boldsymbol{w}\|_0 = K \text{ and } \boldsymbol{w} \in \mathbb{B}^M, \end{aligned} \tag{8}$$

where $\boldsymbol{w}$ determines whether or not each column is involved and $\mathbb{B} = \{0, 1\}$.

The computationally tractable algorithms are divided into two main categories, convex relaxation and greedy selection. The first approach approximates the search space to the nearest convex set and exploits convex optimization methods to solve the problem, while greedy methods gradually select suitable sensors or prune inefficient ones.

3.1 Convex Relaxation

A convex relaxation for (8) is proposed in [2] as given by

$$\begin{aligned} \hat{\boldsymbol{w}} &= \underset{\boldsymbol{w}}{\text{argmin}} \log\det(\sum_{m=1}^{M} w_m\mathbf{a}_m\mathbf{a}_m^T)^{-1}, \\ &\text{subject to} \quad \|\boldsymbol{w}\|_1 = K \text{ and } \boldsymbol{w} \in \mathbb{C}^M, \end{aligned} \tag{9}$$

for which ℓ_0 norm is replaced by ℓ_1 norm and $\mathbb{C}$, the continuous set $[0, 1]$, is used instead of $\mathbb{B}$. Another heuristic (A-optimality) is proposed in [30] based on minimization of MSE$= \mathbb{E}[\|x - \hat{x}\|_2^2] = \sigma^2\text{tr}(\sum_{m=1}^{M} \mathbf{a}_m\mathbf{a}_m^T)^{-1}$ given by,

$$\begin{aligned} \hat{\boldsymbol{w}} &= \underset{\boldsymbol{w}}{\text{argmin}} \|\boldsymbol{w}\|_1, \\ &\text{subject to} \quad \text{tr}(\sum_{m=1}^{M} w_m\mathbf{a}_m\mathbf{a}_m^T)^{-1} \le \eta \text{ and } \boldsymbol{w} \in \mathbb{C}^M, \end{aligned} \tag{10}$$

where, η is a regularization parameter. As η increases, the number of selected sensors would be decreased. There is a performance gap between the best subset and the heuristic solution of the convex relaxation for maximizing the volume. Although simulations show the gap is small in many cases, there is no guarantee that the gap between the performance of the chosen subset and the best performance is always small [2].

3.2 Greedy Algorithms

The greedy algorithms are faster than convex relaxation methods in addition to providing some guarantees for the optimality of the solution in the case of a sub modular condition [31]. For example, it is possible to rewrite (8) as the following sub-modular problem [3],

$$\begin{aligned} \hat{\boldsymbol{w}} &= \underset{\boldsymbol{w}}{\text{argmax}} \log\det(\sum_{m=1}^{M} w_m\mathbf{a}_m\mathbf{a}_m^T), \\ &\text{subject to} \quad \|\boldsymbol{w}\|_0 = K \text{ and } \boldsymbol{w} \in \mathbb{B}^M. \end{aligned} \tag{11}$$

To solve this problem, we can select sensors sequentially. At the step t, a sensor will be selected that maximizes $\log\det\{(\sum_{m=1}^{t-1} \mathbf{a}_{S_m}\mathbf{a}_{S_m}^T) + \mathbf{a}_z\mathbf{a}_z^T\}$ with respect to $\mathbf{a}_z$ in which S_m stacks the indices of the selected sensors in previous iterations and the obtained z is the index of the new selected sensor. Solving the maximization results in $\mathbf{a}_{S_t}$. This procedure will continue till $t = K$.

3.3 Matrix subset selection

The sensor selection problem is highly related to column/row sub-matrix selection, a fundamental problem in applied mathematics. There exists many efforts in this area [32–35]. Generally, they aim at devising a computationally efficient algorithm in which the span of the selected columns/rows cover the columns/rows space as close as possible. Mathematically, a general guarantee can be stated as one of the following forms [33, 36],

$$\begin{aligned} \mathbb{E}\{\|\boldsymbol{A} - \pi_{\mathbb{T}}(\boldsymbol{A})\|_F^2\} &\le (K+1)\|\boldsymbol{A} - \boldsymbol{A}_K\|_F^2, \\ \|\boldsymbol{A} - \pi_{\mathbb{T}}(\boldsymbol{A})\|_F^2 &\le p(K, M, N)\|\boldsymbol{A} - \boldsymbol{A}_K\|_F^2, \end{aligned}$$

in which, $\pi_{\mathbb{T}}(\boldsymbol{A})$ represents projection of rows of $\boldsymbol{A}$ on to the span of selected rows indexed by $\mathbb{T}$ set. $\mathbb{E}$ indicates expectation operator with respect to $\mathbb{T}$, i.e., all the combinatorial selection of K rows of $\boldsymbol{A}$ out of M. Moreover,

$p(K, M, N)$ is a polynomial function of the number of selected elements, the number of columns and the number of rows. $\boldsymbol{A}_K$ is the best rank-K approximation of $\boldsymbol{A}$ that can be obtained by singular value decomposition. The first form suggests the distribution of potential sets for selection and it expresses an upper bound for expected value of error. The second form guarantees existence of a deterministic subset that bounds the error by a polynomial function of the parameters.

Volume sampling is the most well-known approach to achieve the desired selection that satisfies one of the aforementioned bounds. The following theorem expresses the probabilistic form volume sampling.

Theorem 3.1 ([33]). Let $\mathbb{T}$ be a random $K-$subset of rows of a given matrix $\boldsymbol{A}$ chosen with probability

$$Pr(\mathbb{T}) = \frac{\det(\boldsymbol{A}_{\mathbb{T}}\boldsymbol{A}_{\mathbb{T}}^T)}{\sum_{|\mathbb{U}|=K}\det(\boldsymbol{A}_{\mathbb{U}}\boldsymbol{A}_{\mathbb{U}}^T)}$$

Then,

$$\mathbb{E}\{\|\boldsymbol{A} - \pi_{\mathbb{T}}(\boldsymbol{A})\|_F^2\} \leq (K+1)\|\boldsymbol{A} - \boldsymbol{A}_K\|_F^2.$$

Volume sampling considers more probability of selection for those rows whose volume is greater. The volume of a subset of a matrix, $\boldsymbol{A}_{\mathbb{T}}$, is proportional to the determinant of $\boldsymbol{A}_{\mathbb{T}}\boldsymbol{A}_{\mathbb{T}}^T$. Thus, (8) aims to find the most probable subset according to volume sampling.

Volume sampling and D-optimality pursue the same heuristic objective. This heuristic does not promote a well-shaped matrix for compressive sensing purposes based on RIP. However, the analysis of optimization w.r.t the RIP coefficient is not an easy task due to the columns combinatorial behavior in addition to row selection for the basic sensor selection problem. To eliminate the column combinations, we consider all of the columns and consequently we come up with an optimization problem w.r.t the minimum eigenvalue that is known as E-optimality in the optimization literature [4]. Assume a simple selection from rows of $\boldsymbol{A} \in \mathbb{R}^{100\times 3}$. Each row of $\boldsymbol{A}$, associated with a sensor, corresponds to a point in $\mathbb{R}^3$. We are to select 2 sensors out of 100 based on D-optimality and E-optimality. Both solutions are initialized by the same sensor (sensor 1) and the criteria for the next selection varies. The D-optimal solution aims to maximize the surrounded area (gray area in Fig. 1) which is vulnerable to be an ill-shaped area while, E-optimal solution comes up with a well-shaped area due to maximizing the minimum eigenvalue (shaded area in Fig. 1).[2]

The following simple example illustrates the effect of E-optimality. Consider two matrices, $\begin{bmatrix}2 & 0\\ 0 & 0.5\end{bmatrix}$ and $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$. The determinant of both matrices are equal, thus D-optimality does not favor one over the other, however, the second matrix is optimum based on E-optimality.

2. The presented intuition about D-optimality and E-optimality relates to the condition number of a matrix in linear algebra [37]. Diverged eigenvalues results in a large condition number and an ill-conditioned system of equations; accordingly, we refer to the polygon of an ill-conditioned system of equations as ill-shaped where the vertexes of shape are the rows of the matrix. On the other hand, close eigenvalues correspond to a small condition number and a well-conditioned system of equations. The corresponding polygon is referred as well-shaped in Fig 1. Having well-conditioned matrices, is a central concern in CS as evidenced by the role played by the RIP [38].

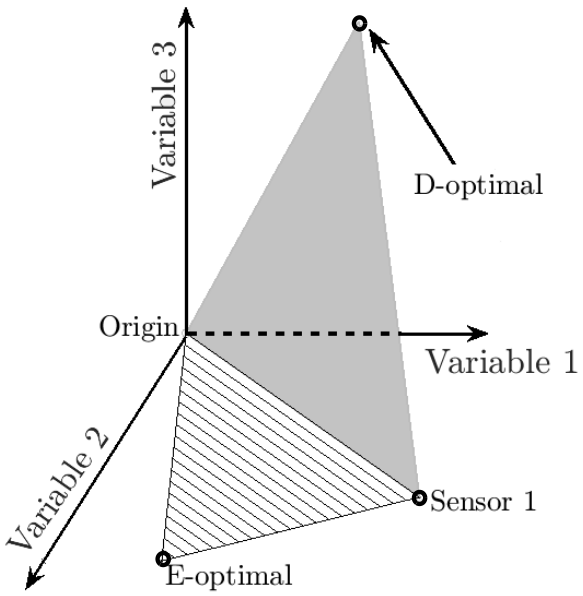


Fig. 1: Comparison of D-optimality and E-optimality for selecting 2 sensors in the 3D space. The gray area is the maximum achievable area by selecting the second sensor based on D-optimality. The shaded area is a well-shaped polygon obtained by E-optimality.

As we will see in the next section, for selection of K rows of $\boldsymbol{A} \in \mathbb{R}^{M\times N}$, the E-optimal criterion is equivalent to optimizing the RIP coefficient of order N, which is an upper bound for any arbitrary order of RIP coefficients. In the next section E-optimality will be exploited to develop a new sampling method for which its performance guarantee is analyzed. E-optimal criterion suggests optimization of an upper bound for any order of RIP. Moreover, in this paper we suggest a method to approximate a specific order of RIP. Based on it, a new RIP-based sensor selection algorithm is proposed.

4 E-OPTIMAL SAMPLING

Remark 1 promotes us to develop a new matrix subset selection method that reduces the matrix to have a well-conditioned sub-matrix in the CS sense. The dominant factor of RIP constant comes from the minimum eigenvalue of the reduced matrix. It suggests to exploit the following optimization problem for sensor selection,

$$\begin{aligned}\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \; &\|\Big(\sum_{m=1}^{M} w_m \mathbf{a}_m \mathbf{a}_m^T\Big)^{-1}\|,\\ \text{subject to} \quad &\|\boldsymbol{w}\|_0 = K \text{ and } \boldsymbol{w} \in \mathbb{B}^M.\end{aligned} \tag{12}$$

In which, $\|.\|$ denotes the spectral norm of a matrix that is defined as its maximum eigenvalue. The following lemma shows that the minimum eigenvalue is an upper bound for δ_S^L.

Lemma 4.1. For any $\boldsymbol{A} \in \mathbb{R}^{M\times N}$, the following inequality holds.

$$1-\sigma^2_{\min}(\boldsymbol{A}) = \delta^L_N(\boldsymbol{A}) \geq \delta^L_{N-1}(\boldsymbol{A}) \geq \cdots \geq \delta^L_2(\boldsymbol{A}).$$

Proof: According to the definition of RIP constant δ_S and considering that the set of at most S-1 non-zero vectors is subset of the set of at most S non-zero vectors, it easily concluded that $\delta_S(\boldsymbol{A}) \geq \delta_{S-1}(\boldsymbol{A})$ for any $S = 2, \cdots, N$.

Lemma 4.1 suggests that E-optimality, i.e., minimization of δ^L_N, actually is equivalent an upper bound for an arbitrary order of RIP coefficient.

Similar to volume sampling, we design a probability of sampling according to their minimum eigenvalue.

Definition 3. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$, *E-optimal sampling* is defined as picking a subset of $\mathbb{T}$ with the following probability,

$$Pr(\mathbb{T}) = \frac{\sigma^2_{\min}(\boldsymbol{A}_{\mathbb{T}})}{\sum_{|\mathbb{U}|=K}\sigma^2_{\min}(\boldsymbol{A}_{\mathbb{U}})}.$$

Definition 4. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$, $\bar{\delta}^L_K$ is defined as one minus the average of minimum of squared-singularvalues of all sub-matrices of A with K columns. Mathematically, it can be expressed as follows,

$$\bar{\delta}^L_K(\boldsymbol{A}) = 1-\bar{\sigma}^2_{\min}(\boldsymbol{A}) = 1-\frac{1}{\binom{M}{K}}\sum_{|\mathbb{U}|=K}\sigma^2_{\min}(\boldsymbol{A}_{\mathbb{U}}),$$

in which $\mathbb{U} \subset [1, \cdots, N]$ indicates a subset of K columns of $\boldsymbol{A}$.

Definition 5. [17] Given a matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$, the spark of $\boldsymbol{A}$ is defined as the smallest number of columns that are linearly dependent. It can be stated as follows,

$$Spark(\boldsymbol{A}) = \min \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \boldsymbol{A}\mathbf{x} = \mathbf{0} \ \text{ and } \ \mathbf{x} \neq \mathbf{0}.$$

The upper bound for spark is the rank of matrix plus 1. However any linear dependencies among some columns of the matrix may decrease the spark. Based on the above definitions we present the following theorem that expresses an upper bound for projection error of E-optimal sampling.

Theorem 4.2. Assume spark of $\boldsymbol{A}^T$ is greater than $K+1$. E-optimal selection of K rows implies

$$\mathbb{E}\{\|\boldsymbol{A}-\pi_{\mathbb{T}}(\boldsymbol{A})\|^2_F\} \leq \frac{M-K}{\gamma} \times \frac{1-\bar{\delta}^L_{K+1}(\boldsymbol{A}^T)}{1-\bar{\delta}^L_K(\boldsymbol{A}^T)},$$

where γ is a positive number a function of the dependencies of rows.

Proof: See appendix.

E-optimal sampling implies an upper bound for the expectation of projection error in a probabilistic manner. However, we need to select some sensors deterministically. To this aim, we propose the following iterative algorithm. Actually, this algorithm is an approximation for the maximum likelihood estimator in which the likelihood comes from the suggested probability in Definition 3.

Table 2 compares computational burden of three well-known selection methods with the proposed method. Convex relaxation is not able to work effectively for big data sets since the complexity of the algorithm grows with M^3 [2]. Complexity of volume sampling also depends on M^2.

TABLE 2: Complexity of different selection strategies.

Algorithm	Complexity
Convex Optimization [2]	$O(M^3)$
Volume sampling [32]	$O(KNM^2\log M)$
Greedy Submodular Selection [3]	$O(MK^3)$
Greedy E-optimal selection (proposed)	$O(MNK^2)$

Likewise, complexity of greedy algorithms which process data one-by-one increase linearly w.r.t size of data. However, in some big data scenarios we still need to decrease computational complexity w.r.t data size. To this aim in Section 6, two remedies are studied based on data partitioning.

Algorithm 1 Greedy E-Optimal Sensor Selection

Require: $\boldsymbol{A}$ and K
Initialization: $\mathbb{S}$ with a random sensor
for $k = 1, \cdots, K$
  for $m = 1, \cdots, M$
    $\mathbb{T} = \mathbb{S}\bigcup\{m\}$
    $p(m) = \sigma_{\min}(\boldsymbol{A}_{\mathbb{T}})$
  end
$s_k = \text{argmax}\ p(m)$
$\mathbb{S} = \mathbb{S}\bigcup\overset{m}{s_k}$
end

5 RIP-Based Sensor Selection

The structure of the reduced measurement matrix plays a critical role in sparse recovery. Several criteria have been suggested to evaluate suitability of a measurement matrix including the mutual coherency and the RIP coefficient. In order to guarantee a well-conditioned matrix to recover a S-sparse vector, the criteria based on RIP depend on the RIP constant of order PS. Different guarantees suggest some bounds in terms of δ_{2S}, δ_{3S} and δ_{4S}, i.e., δ_{PS} for $P = 2, 3, 4$ [19] [20]. As Remark 1 suggests, the lower RIP constant defined in (3) is the dominant factor for compressive sensing. Thus, we employ the lower constant of order PS in (2) denoted by δ^L_{PS} (3) to propose the following problem for sensor selection,

$$\begin{aligned}\hat{\boldsymbol{W}} = &\underset{w_{km}\in\{0,1\}}{\text{argmin}}\ \delta^L_{PS}(\boldsymbol{W}\boldsymbol{A}),\\ &\text{subject to } \|\mathbf{w}_k\|_0 = 1\ \forall k = 1, \cdots, K.\end{aligned} \tag{13}$$

In which $\boldsymbol{W} \in \mathbb{R}^{K\times M}$ reduces the matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ by some selected rows. I other words, matrix $\boldsymbol{W}$ is the selector operator and the goal of sensor selection is to estimate this matrix. P is a constant between 2 and 4 and $\mathbf{w}_k$ is the k^{th} row of $\boldsymbol{W}$. In each row of $\boldsymbol{W}$ there is only one entry 1 and all the other entries are zero, i.e., $\|\mathbf{w}_k\|_0 = 1$. According to the definition of RIP, the above problem can be cast to the following form,

$$\begin{aligned}\hat{\boldsymbol{W}} = &\underset{w_{km}\in\{0,1\}}{\text{argmax}}\ \min_{x}\|\boldsymbol{W}\boldsymbol{A}\boldsymbol{x}\|^2_2,\\ &\text{subject to } \|\mathbf{w}_k\|_0 = 1, \|\boldsymbol{x}\|_2 = 1 \text{ and } \|\boldsymbol{x}\|_0 \leq PS.\end{aligned} \tag{14}$$

This problem is a jointly combinatorial search with respect to both $\boldsymbol{W}$ and $\boldsymbol{x}$. It is shown that finding the solution with

respect to $\boldsymbol{x}$ is NP-hard with a fixed $\boldsymbol{W}$ [39]. On the other hand, with a fixed $\boldsymbol{x}$, it is easy to show that the problem is sub-modular with respect to $\boldsymbol{W}$. The reduction matrix selects the most significant entries of the error $\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}$. In the next section we will propose an optimization algorithm that first approximates the solution w.r.t $\boldsymbol{x}$ and then pursues a greedy method to update $\boldsymbol{W}$. Please note that by ignoring the last constraint, the problem turns into the E-optimal sensor selection.

Although matrix subset selection and sensor selection formulation are highly related to each other, they have their own approaches to the problem. Sensor selection aims to reduce a system of equations which is not specified for a fixed unknown vector. For instance, in Problem (14) we minimize w.r.t $\boldsymbol{x}$ and Problem (8) is derived by minimizing expectation of estimation error of $\boldsymbol{x}$. However, a specific $\boldsymbol{x}$ generates the corresponding values of potential sensors. So far we have assumed that we do not optimize the problem for a specific observed $\boldsymbol{y}$. If we have access to the measurements in a fusion center, we can exploit this information in the selection decision. To consider more valuable measurements, their values are involved in the following problem in which we call it data-aware RIP based sensor selection.

$$\begin{aligned}\hat{\boldsymbol{W}} = \underset{w_{km}\in\{0,1\}}{\operatorname{argmax}} \min_{\boldsymbol{x}} &\|\boldsymbol{W}\boldsymbol{A}\boldsymbol{x}\|_2^2+\lambda\|\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}(\mathbb{O}))\|_2^2,\\ \text{s.t. } &\|\mathbf{w}_k\|_0= 1, \|\mathbf{w}^{(m)}\|_0\leq 1, \|\boldsymbol{x}\|_2= 1 \text{ and } \|\boldsymbol{x}\|_0\leq PS.\end{aligned} \tag{15}$$

This problem promotes the sensor selection to select sensors from areas with high vulnerability to error. In a same time, their corresponding bases construct a well-conditioned matrix based on RIP coefficient. $\mathbf{w}^{(m)}$ denotes the m^{th} column of $\boldsymbol{W}$. The constraint, $\|\mathbf{w}^{(m)}\|_0\leq 1$ avoids repetitive selection of the same sensor. Note that repetitive selection may occur for large values of λ and there is no need for this constraint in (14) because a repetitive column results in a zero eigenvalue while the cost function maximizes the minimum eigenvalue. By considering the model's error, we aim to compensate the error of model by an intelligent sensor selection. Aggregating all sensors' measurements in a fusion center is in contrast with the goal of sensor selection. However, we devise a dynamic framework that needs a partial set of sensors for adapting the sensor selection algorithm with the dynamic of the sensors. These measurements might be derived by a low-frequency sampling from all sensors or set of recent measured sensors. $\mathbb{O}$ denotes the set of observed measurements and $\boldsymbol{x}(\mathbb{O})$ refers to the estimation based on the partial observed data. $\boldsymbol{A}\boldsymbol{x}(\mathbb{O})$ indicates the approximation of the measurements $\boldsymbol{y}$. $\boldsymbol{x}(\mathbb{O})$ is obtained by solving the following regression problem.

$$\boldsymbol{x}(\mathbb{O}) = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{y}_{\mathbb{O}} - \boldsymbol{A}_{\mathbb{O}}\boldsymbol{x}\|_2^2+\lambda_{LASSO}\|\boldsymbol{x}\|_1. \tag{16}$$

Where λ_{LASSO} regularizes sparsity. In order to obtain the error of model, we need to observe all the sensors, while we aim to keep the number of observed sensors limited. The following interpolation in terms of the observed sensors is exploited to derive the error of model for all sensors.

$$y_m = \frac{\sum_{j\in\mathbb{O}}\gamma_{mj}y_j}{\sum_{j\in\mathbb{O}}\gamma_{mj}}, \tag{17}$$

where γ_{mj} is a similarity function between m^{th} and j^{th} sensor. The estimated observation of unobserved sensors help us to evaluate their fidelity to the model. E.g., if the interpolated measurement of the m^{th} sensor, y_m, also satisfies $y_m \approx \boldsymbol{a}_m^T\boldsymbol{x}(\mathbb{O})$, it implies that this sensor can be predicted by some other sensors based on the model. Thus, this sensor is reliable and it does not maximize the cost function (15) significantly. This data-driven approach is inspired by dynamic sensor selection introduced in [40, 41]. For a given model $\mathcal{M}$ on the data, dynamic sensor selection determines set $\mathbb{S}$ such that the estimation error of the rest of sensors, $\mathbb{S}^c$, is minimized. The estimation is obtained based on the model, $\mathcal{M}$, and observed sensors, $\mathbb{S}$ [40]. The assumed model in our proposed approach is indicated in (4).

The parameter λ in (15) regularizes the weight of the energy of error and the RIP coefficient of selected bases. In other words, $\boldsymbol{W}$ reduces the rows of $\boldsymbol{A}$ in an optimal sense and simultaneously, it selects some vulnerable sensors to model's error. In the experimental results we will show the effect of the regularization parameter. According to our simulations, the importance of the main term of objective function is more than the energy of the model's error. Even by $\lambda = 0$ we have a well-spread set of selected sensors corresponding to a well-conditioned system of equations while, by $\lambda \to \infty$ a set of concentrated sensors would be concluded which corresponds to an ill-posed system of equations. Simulations show a relatively wide range of λ could be a good choice.

Finding RIP of a matrix requires solving an NP-hard problem [39]. Thus, for a large-scale problem, it is not feasible to search among all the subsets. A greedy algorithm is proposed to approximate the RIP of a matrix. To this end, let us consider the following problem.

$$\begin{aligned}&\delta_{PS}(\boldsymbol{A}) = \\ &\quad 1 - \min\|\boldsymbol{A}\mathbf{x}\|_2^2 \quad \text{st: } \|\mathbf{x}\|_2= 1 \text{ and } \|\mathbf{x}\|_0\leq PS.\end{aligned} \tag{18}$$

The solution is approximated in (19). The suggested problem neglects the last constraint in (18) and obtains a solution, then projects the obtained solution to the feasible set spanned by the neglected constraint.

$$\begin{aligned}&\tilde{\delta}_{PS}(\boldsymbol{A}) = \\ &\quad 1 - \|\boldsymbol{A}\Omega_{\ell_2}(T_{PS}\{\underset{x}{\operatorname{argmin}}\|\boldsymbol{A}\mathbf{x}\|_2^2 \text{ st: } \|\mathbf{x}\|_2= 1 \})\|_2^2.\end{aligned} \tag{19}$$

In which, $T_{PS} : \mathbb{R}^N \to \mathbb{R}^N$ is the truncate function that keeps only PS most significant entries and makes the rest zero. As the truncated vector no longer satisfies the unit norm constraint, $\Omega_{\ell_2} : \mathbb{R}^N \to \mathbb{R}^N$ normalizes the truncated vector to the unit ℓ_2 ball. The solution of the alternative problem denoted by $\tilde{\delta}_{PS}(\boldsymbol{A})$ can be solved efficiently using singular value decomposition.

$$\begin{aligned}&\tilde{\delta}_{PS}(\boldsymbol{A}) = 1 - \|\boldsymbol{A}\mathbf{x}^*\|_2^2, \quad \mathbf{x}^* = \Omega_{\ell_2}(T_{PS}\{\boldsymbol{U}(:,k)\}),\\ &\boldsymbol{A} = \boldsymbol{V}^T\boldsymbol{\Lambda}\boldsymbol{U},\end{aligned} \tag{20}$$

in which, $\boldsymbol{U}(:,k)$ is the k^{th} column[3] of $\boldsymbol{U}$. In other words, $\mathbf{x}^*$ is obtained by setting it to the normalized and truncated Eigenvector corresponding the minimum Eigenvalue. By exploiting the approximation of δ_{PS} the sensor selection problem can be cast as the following form,

$$\hat{\boldsymbol{W}} = \underset{w_{km}\in\{0,1\}}{\text{argmin}}\ \tilde{\delta}_{PS}(\boldsymbol{W}\boldsymbol{A}) \quad \text{st: } \|\mathbf{w}_k\|_0 = 1,\ \forall k = 1,\cdots,K\ . \tag{21}$$

By using the obtained approximation in (20), we conclude

$$\hat{\boldsymbol{W}} = \underset{w_{km}\in\{0,1\}}{\text{argmax}}\ \|\boldsymbol{W}\boldsymbol{A}\mathbf{x}^*\|_2^2 \quad \text{st: } \|\mathbf{w}_k\|_0 = 1,\ \forall k = 1,\cdots,K\ , \tag{22}$$

in which,

$$\begin{aligned} &\mathbf{x}^* = \Omega_{\ell_2}(T_{PS}\{\boldsymbol{U}(:,K)\}) \\ &\boldsymbol{W}\boldsymbol{A} = \boldsymbol{V}^T\boldsymbol{\Lambda}\boldsymbol{U}. \end{aligned} \tag{23}$$

Algorithm 2 shows the steps of our proposed greedy algorithm to solve the obtained optimization problem. To evaluate each sensor we need to compute the most dominant k eigen components which implies performing singular value decomposition (SVD). However, truncated SVD up to the k^{th} component will be sufficient. A similar algorithm can be used to solve Problem (15). To this aim, Step 6 in Algorithm 2 should be modified to consider the error of m^{th} sensor, i.e., $p(m) = \|\mathbf{x}^*\|_2^2 + \lambda|y_m - \boldsymbol{a}_m^T\boldsymbol{x}(\mathbb{O})|$. However, it is not practical to have all the measurements at the fusion center. An online algorithm is proposed that observes one new measurement sequentially. In each sequence, the observed set of sensors is updated and this set is initialized by the output of Algorithm 2. In other words, the selected sensors in Algorithm 2 are sensed. Our data-aware algorithm needs an approximation of the observed data in terms of the corresponding reduced $\boldsymbol{A}$ using (16).

As mentioned in the last section, the online data-aware framework, Algorithm 3, uses an interpolation as the prediction of unobserved measurements. It will be an enabling step for estimation of model's error in order to adapt the sensor selection to the measurements. The interpolation is based on weighted averaging of observed measurements where the weight is a similarity metric that depends on the underlying application. For example, we consider a simple channel gain between two sensors in CRNs simulations which is an inverse function of distance as the similarity criterion in 17.

The bottleneck of complexity order of Algorithm 2 at the k^{th} iteration is performing a truncated singular value decomposition to obtain the first k eigen components. Thus, the complexity of the algorithm in the k^{th} iteration will be $O(kMN^2\})$ [42]. Therefore, selection of K sensors implies complexity order of $O(K^2N^2M)$.

Algorithm 2 and Algorithm 3 can be implemented in a distributed manner similar to the proposed idea in Section 6. The selection procedure is as same as before in each machine but the number of data are decreased by factor C which is the number of machines. This makes complexity to $O(K^2N^2M')$ where, $M' = M/C$.

3. The k^{th} column is represented by $\boldsymbol{U}(:,k)$ and the k^{th} row is represented by $\boldsymbol{U}(k,:)$ in Algorithms 1 and 2. Moreover, $\boldsymbol{U}(\mathbb{S},:)$ represents the reduced matrix by some selected rows indicated by $\mathbb{S}$ set.

6 Distributed Implementation

Data summarizing is an enabling step for more complicated processing procedures. For example, computational burden for training a recognition system increases tremendously by the size of the training data. However, in some cases even data summarizing is not tractable due to the size of data. A naive approach for data summarizing is randomly sampling from data to make it sufficiently small.

There exist some attempts to design randomized algorithms for matrix subset selection [36]. The idea is based on combining deterministic and randomized methods, using a two-phase algorithm. The first phase selects $O(k\log(k))$ rows of the matrix. Then, deterministic subset selection finds exactly the k most informative rows of the matrix. This randomized algorithm achieves the following bound [36],

$$\|\boldsymbol{A} - \pi_{\mathbb{T}}(\boldsymbol{A})\|_F^2 \leq O(k\log^{\frac{1}{2}}(k))\|\boldsymbol{A} - \boldsymbol{A}_K\|_F^2,$$

This bound suggests us that a judiciously or even randomly set of rows of $\boldsymbol{A}$ can provide us a submatrix with a close subspace to the original matrix. The submatrix might be more convenient to deal with, specially when the data size is big. In this section, data partitioning is studied as an enabling step for successive and parallel processing.

6.1 Successive Processing

In order to make the problem tractable, we can employ a method based on successive processing of partitioned data. Suppose data matrix $\boldsymbol{A}$ is partitioned into C blocks that each block, $\boldsymbol{A}_c$, contains M_c rows of $\boldsymbol{A}$. At the first stage K rows are selected out of M_1 rows of the first partition. The selected rows are forwarded to the next stage in order to perform selection among M_2 data of the second part, as well as the already K selected rows. It means at the second stage there are $M_2 + K$ data and the goal is to select only K rows to feed to the next stage. Alg. 4 shows the steps of successive E-optimal sensor selection algorithm. In the experimental results section the performance of this method will be presented.

In addition to the successive method, there is another solution for scenarios that data can be independently processed over distributed machines in a parallel manner. The successive approach performs a series of selection procedures and all of these procedures can be implemented in a same machine. However, in some scenarios we have access to multiple processing nodes in a network. In this case it is desired to implement a distributed algorithm, which is able

Algorithm 2 The blind RIP-based Sensor Selection

Require: $\boldsymbol{A}$, S and K
1: **Initialization:** $\boldsymbol{W} = \mathbf{0} \in \mathbb{R}^{K\times M}$ and $\mathbb{S} = \emptyset$
2: for $k = 1,\cdots,K$ (Optimization of the k^{th} row of $\boldsymbol{W}$)
3: for $m = 1,\cdots,M$
4: SVD on $\boldsymbol{A}(\mathbb{S}\bigcup m,:)$ to obtain $\boldsymbol{U}$ in (23)
5: $\mathbf{x}^* = \Omega_{\ell_2}(T_{PS}\{U(:,k)\})$
6: $p(m) = \|\boldsymbol{A}\mathbf{x}^*\|_2^2$
7: end
8: $s_k = \underset{m}{\text{argmax}}\ p(m)$
9: $\mathbb{S} = \mathbb{S}\bigcup s_k$ and $\boldsymbol{W}_{k,s_k} = 1$

Algorithm 3 The data-aware RIP-based Sensor Selection

Require: $\boldsymbol{A}$, S, K, λ and λ_{LASSO}
Initialization:$\mathbb{O}=$ Output of Algorithm 1
while $\mathbb{O} \neq \{1,\cdots,M\}$
$\boldsymbol{W}=\boldsymbol{0}\in\mathbb{R}^{K\times M}$, $\mathbb{S}=\emptyset$
Observe 1 new measurement and update $\mathbb{O}$
Interpolate $\mathbf{y}_{\overline{\mathbb{O}}}$ using $\mathbf{y}_{\mathbb{O}}$ using (17)
for $k=1,\cdots,K$
for $\forall m\in\mathbb{S}^c$
SVD on $\boldsymbol{A}(\mathbb{S}\bigcup m,:)$ to obtain $\boldsymbol{U}$ in (23)
$\mathbf{x}^*=\Omega_{\ell_2}(T_{PS}\{U(:,k)\})$
$\boldsymbol{x}(\mathbb{O})=LASSO(\boldsymbol{A},\boldsymbol{y}_{\mathbb{O}},\lambda_{LASSO})$ using (16)
$p(m)=\|\boldsymbol{A}\mathbf{x}^*\|_2^2+\lambda|y_m-\boldsymbol{a}_m^T\boldsymbol{x}(\mathbb{O})|$
end
$s_k=\text{argmax}\ p(m)$
$\mathbb{S}=\mathbb{S}\bigcup^m s_k$ and $\boldsymbol{W}_{k,s_k}=1$
end
$\mathbb{O}=\mathbb{O}\bigcup\mathbb{S}$ and return to 2.

to process different part of data simultaneously. We study two methods for distributing data, random partitioning and designed partitioning.

Algorithm 4 Successive E-optimal row selection

Require: $\boldsymbol{A}$, C, and K
Initialization: $\mathbb{S}$ by $\emptyset$.
Partition $\boldsymbol{A}$ to C parts ($\mathbb{A}_c$ indicates the indices of $\boldsymbol{A}_c$).
for $c=1,\cdots,C$
$\mathbb{Z}=\mathbb{S}\bigcup\mathbb{A}_c$
$\mathbb{S}\leftarrow$ select K rows of $\boldsymbol{A}_{\mathbb{Z}}$ using Alg. 1.
end

6.2 Random Partitioning

In this section, the given matrix, $\boldsymbol{A}$, is randomly broken into $\{\boldsymbol{A}_c\}_{c=1}^C$, in which each submatrix contains M_c rows of the original matrix. In order to ensure that row space of each submatrix is close enough to the row space of the original matrix, we need to derive a lower bound on the number of members of each submatrix.

Assumption 1: The matrix $\boldsymbol{A}$ can be expressed as a union of subspaces, i.e., $\boldsymbol{A}=[\boldsymbol{U}_1\boldsymbol{Q}_1,\cdots,\boldsymbol{U}_L\boldsymbol{Q}_L]^T$. Assume rank of $\boldsymbol{A}$ is R and rank of each subspace is $\frac{R}{L}$, where, $\{\boldsymbol{U}_l\in\mathbb{R}^{N\times\frac{R}{L}}\}_{l=1}^L$ and $\{\boldsymbol{Q}_l\in\mathbb{R}^{\frac{R}{L}}\times M'\}_{l=1}^L$, and $M'=M/L\gg\frac{R}{L}$.

Assumption 1 implies that the original matrix, $\boldsymbol{A}$, is a union of L subspaces in which intrinsic dimension of each subspace is at most R/L. This assumption is reasonable for many scenarios in signal processing and data mining [43, 44]. The following lemma suggests an upper bound for the number of parallel machines in order to ensure that the row space of each portion of data is equal to the original data with a high probability.

Lemma 6.1. Assume $\boldsymbol{A}$ follows Assumption 1. If the rows of $\boldsymbol{A}$ are equally partitioned among C parts and samples of each part are drawn uniformly at random and C satisfies the following inequality,

$$C\leq\frac{M}{L\xi(2+(3/\xi)\log\frac{2L}{\delta})},\tag{24}$$

where,

$$\xi=10\gamma\max(R/L,\log M/L)\log\frac{2R}{\delta},$$

then the row space of each part spans row space of $\boldsymbol{A}$ with probability at least $1-2\delta-2\frac{L^4}{M^3}$.

proof: See Appendix.

Proposition 1. The order of minimum number of samples for each parallel machine is $O(R)$ in order to make sure that the span of selected rows is equal to that of the original matrix in each machine with a high probability.

This proposition is clearly derived by the steps of proof of Lemma 6.1 in the appendix. It suggests that each each machine needs a portion of data such that the required size of each portion is linearly dependent to the rank of the original matrix.

Assume K samples are drawn from each partition and KC samples are selected in the first phase. The second phase aims to select only the K most informative samples among the initial selection. Volume sampling and the proposed sampling method select the corners of data such that the selected points constructs a polygonal in which their vertexes are far from each other. However, the selected point could be outlier data, i.e., data is not concentrated about some selected samples. We need to ensure that each selected point represents a relatively large number of non-selected data. Selection algorithms that work based on relative structure of samples are complicated and they can not be used for the big data regime. To tackle this problem a concentration-based selection is performed in the second phase of selection on the KC selected data. K-medoids clustering is a generalization of K-means in which the data centers are selected from the sample points of data. In the first phase we ensure that all the vertexes of the hull of data are selected and in the second phase K-mediods algorithm shrinks the selected data to only K samples. This two-phase algorithm is the practical application of this paper which can be exploited for big data sets. As we will see in the simulation results, the overall two-phase process is faster than performing selection on the whole data using Alg. 1 and it is much faster than performing k-mediods algorithm for whole data. Alg. 5 shows steps of the proposed two-phase algorithm for selecting from big data. This algorithm is the robust and practical version of Alg. 1 for real scenarios which a huge number of noisy data are given.

Algorithm 5 two-phase selection algorithm

Require: $\boldsymbol{A}$, K, C.
Assign $\boldsymbol{A}^{(c)}$ $\forall c=1,\cdots,C$.
for $c=1,\cdots,C$
$\boldsymbol{U}^{(c)}\leftarrow$Algorithm 1 ($\boldsymbol{A}^{(c)},K$).
End for
$\boldsymbol{U}=[\boldsymbol{U}^{(1)},\cdots,\boldsymbol{U}^{(C)}]$.
K-medoids to select K data from $\boldsymbol{U}$.
END

7 Experimental Results

Our proposed schemes are evaluated in three cases including sensor selection in cognitive radio networks (CRNs), and data selection for supervised learning. The underlying model is $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$. Matrix $\boldsymbol{A}$ in CRNs is an array of channel gains from different locations of the network to locations of sensors. In the case of supervised learning, $\boldsymbol{A}$ is the collection of training data. $\boldsymbol{x}$ and $\boldsymbol{y}$ are specified for a test data. However, it is desired that the trained system works for any test data. In the first case, we are estimating a specific $\boldsymbol{x}$ which corresponds to a specific $\boldsymbol{y}$. While for supervised learning it is desired that the selected data constructs a well-conditioned inverse problem that is averagely appropriate

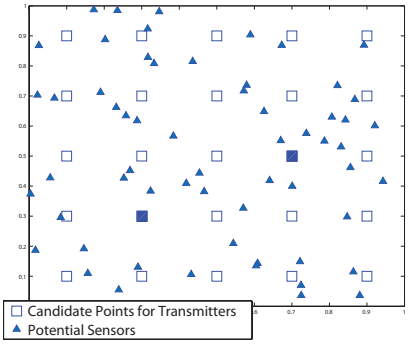


Fig. 2: An example setup with 25 candidate points as transmitters.

for any test data. Thus, we exploit Algorithm 3 only for the first application where we access to the actual measurements of sensors in an online manner.

7.1 Sensor Selection in CRNs

The simulations are performed for collaborative spectrum sensing. Our goal is to estimate vector $\mathbf{x}$ that indicates transmitted spectrum power at some candidate points. We assume a network setup the same as that of [45]. Consider N_s transmitters and M receivers in an area. The receivers receive a superposition of the transmitter signals. Figure 2 shows a setup consist of $N_s = 25$ potential transmitters and 2 active points. The received signals are contaminated by channel gain and additive noise, represented by,

$$\mathbf{y}_m = \boldsymbol{A}_m\mathbf{x} + \sigma_m^2\mathbf{1}, \qquad \forall m = 1 \dots M, \tag{25}$$

where, $\mathbf{1} \in \mathbb{R}^n$, $\mathbf{y}_m \in \mathbb{R}^n$ in which n is the number of frequency samples in each time slot. Moreover, $\boldsymbol{A}_m^T$ contains the corresponding channel gains and σ_m^2 represents noise power at the m^{th} receiver. The following problem aims to estimate $\mathbf{x}$

$$\hat{\mathbf{x}} = \underset{x,\sigma}{\operatorname{argmin}} \|\mathbf{y} - \boldsymbol{A}\mathbf{x} - \boldsymbol{\sigma} \otimes \mathbf{1}\|_2^2 + \gamma\|\mathbf{x}\|_1, \tag{26}$$

in which $\boldsymbol{\sigma} \in \mathbb{R}^M$ indicates the noise level of each sensor. $\mathbf{y}$ and $\boldsymbol{A}$ are concatenation of $\mathbf{y}_m$ and $\boldsymbol{A}_m$ respectively and $\otimes$ denotes kronecker multiplication. Each entry of $\mathbf{x}$ determines the contribution of the s^{th} source on the sensed data. Due to scarce presence of active transmitters and their narrow band communication, $\|\mathbf{x}\|_1$ is exploited which encourages sparsity.

Suppose we have potentially 300 sensors and they are estimating an $\mathbf{x} \in \mathbb{R}^{36}$ that has only 5 active transmitters. Figure 3 shows the performance of different algorithms versus the number of selected sensors. Successful recovery is defined as true estimation of the support of sparse vector using the measurements.

For the first experiment, Problem (26) is solved 200 times by different selected sensor sets for each algorithm. Additive noise is not considered and the iterative re-weighted least square algorithm is employed to obtain the sparse solution [46]. As it can be seen in Figure 3, among the blind methods, sensor selection using RIP coefficient δ_{2S} has the best performance. In the case of known data in a fusion center,

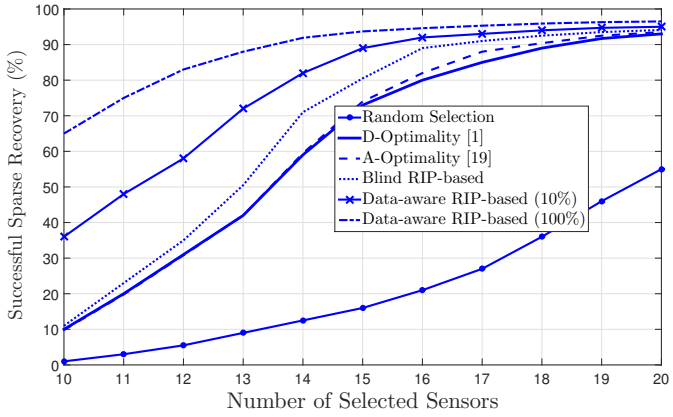


Fig. 3: Performance of different sensor selection algorithms in terms of number of selected sensors

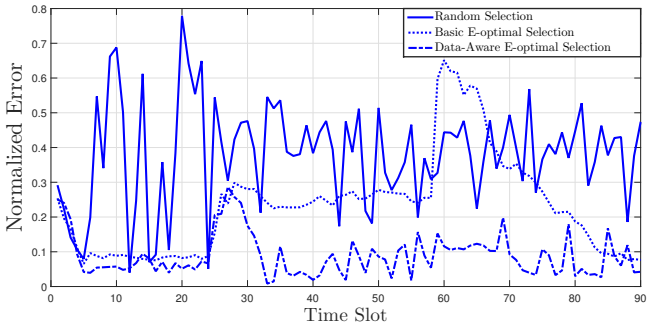


Fig. 4: Performance of the selection algorithm in presence of 0dB AWGN.

the information of sensed data has a great effect on the centralized estimation. The data-aware algorithm observes some sensors in an online manner. [4]

Fig. 4 shows performance of the proposed selection algorithm in a dynamic system where status of the network is changed at time slot 25 and 60. Switching of any propagation point causes a status change. This simulation is performed in presence of 0dB AWGN in addition to 6 tab multipath fading. As it can be seen the blind algorithm performs better than random selection, however, the selected sensors are fixed and independent of the dynamic of system.

Fig. 5 exhibits the effect of involving sensors measurements in the data-aware sensor selection algorithm. Random sensing of only 3% of data (9 sensors within 300 sensors) prior to sensor selection makes an improvement in normalized estimation error; similarly, usage of 15% of data significantly improves the performance to be close to the centralized sensor selection which access to 100% of the data. The normalized error is defined as follows as the criterion for performance,

$$\text{normalized error} = \frac{\|\mathbf{x}^* - \mathbf{x}(\mathbb{O})\|_2}{\|\mathbf{x}^*\|_2}.$$

In which, $\mathbf{x}^*$ is the ground truth solution.

Another simulation is performed to select 20 sensors out of 200 ones to determine the power spectrum in 36

4. The initial sensors can be determined by our blind RIP-based sensor selection and then in each time slot a new sensor will be observed.

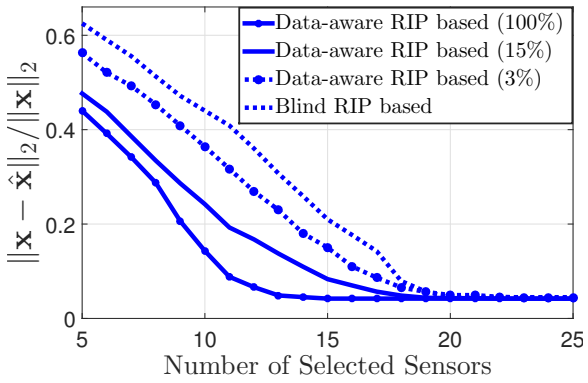
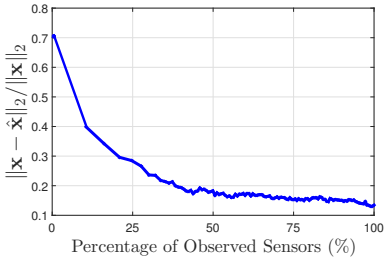
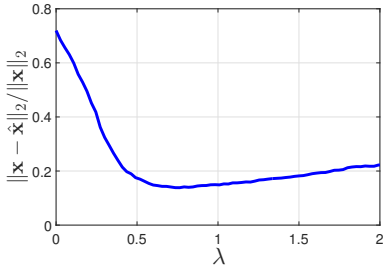


Fig. 5: Performance of blind and data-aware RIP based sensor selection algorithms in terms of number of selected sensors.



(a) Performance of data-aware algorithm versus the percentage of observed measurements. λ is assumed equal to 0.7.



(b) MSE error versus different values of λ for the data-aware algorithm where 100% of measurements are observed.

Fig. 6: Performance of our Data-aware algorithm.

candidate points while 10 of them are active. Figure 6(a) shows the performance of our proposed data-aware method in terms of MSE of the sparse vector estimation while λ is assumed 0.7. The performance improves as the number of observed sensors increases. The performance obtained by observation of 50% of data (100 sensors) is about that of all the sensors because of the redundancy among the sensors. It can be seen in Figure 6(b) that the error of estimation is significantly decreased by setting $\lambda = 0.7$. However, an efficient value of γ depends on the problem setup and should be tuned. Setting $\gamma = 0$ is equivalent to the static E-optimal sensor selection. Simulation shows the proposed reliable sensor selection performs better than the static sensor selection for a relatively wide range of γ, i.e., the problem is not very sensitive to well-tuning of this parameter.

Fig. 7 shows the power spectrum of a network in an area. We have potentially 200 sensors, however we are allowed to use only 8 sensors for collaborative spectrum estimation.

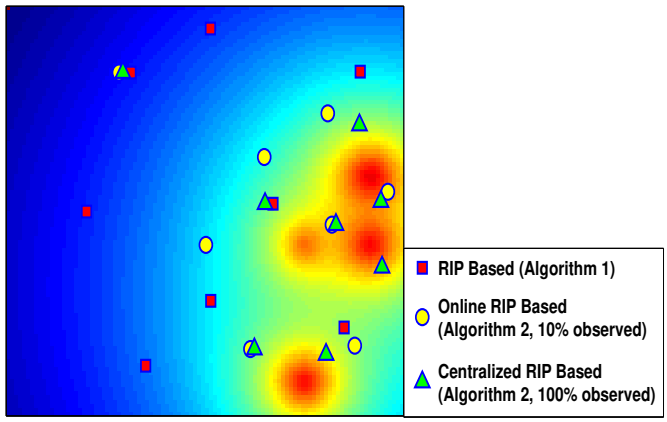


Fig. 7: The true spectrum in the area of interest along the selected sensors obtained by 3 methods in spatial domain

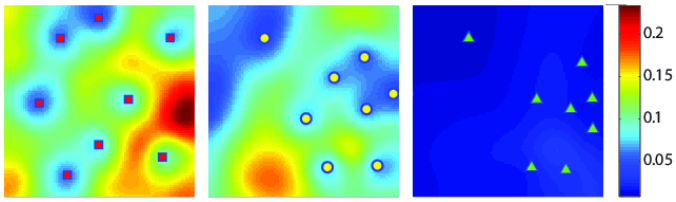


Fig. 8: The error of estimated spectrum in the area of interest corresponding to Fig. 7. (Left) RIP based, Algorithm 1. (Middle) Online RIP based, Algorithm 2 while only 5% of sensors are sensed. (Right) Centralized RIP based, Algorithm 2 while all the sensors are sensed. λ is assumed 0.7

The selected sensors using the blind and data-aware RIP based are marked in this figure. As it can be seen, the selected sensors of the blind RIP based are spread in the area while the selected sensors by the data-aware algorithm have a tendency to move toward the more eventful areas of the network. Figure 8 shows the error of estimated spectrum using different selected sensors in the setup of Fig. 7. To this end, first, the spectrum is estimated in all of sensors and the error is obtained by Euclidean distance of the estimated spectrum and the actual measurements, then a weighted averaging is performed to interpolate the spectrum error in every point.

7.2 Data Selection for Supervised Learning

In this section, the applicability of the proposed selection technique in feature and data selection is studied. This is a challenging problem in computer vision and machine learning [47].

We evaluate the performance of our method as well as other algorithms for finding appropriate representatives for classification. The training data set is reduced to only some selected data for each class. The classifier is then trained solely by the reduced set. We assume that if the representative data are informative enough about the initial data set, the classification performance should be close to the comprehensive classifier. We compare our proposed algorithm with some standard methods for finding representatives. These methods are Kmedoids [48], volume sampling [36], and a simple random selection. Some basic classifiers are utilized for learning and evaluating the test data including

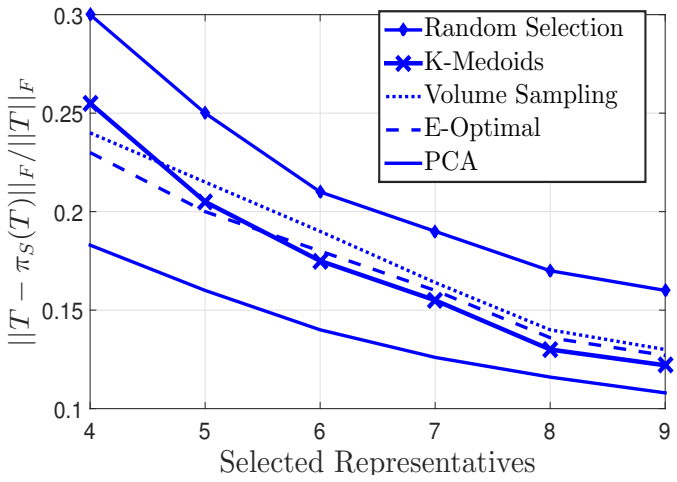


Fig. 9: The projection error of the training data into the subspace spanned by the selected rows.

nearest neighbor (NN), nearest subspace (NS) [49], sparse representation based classifier (SRC) [50], and linear support vector machine (SVM) [51].

Extended Yale-B face images dataset [52] is used to perform the simulations. The dataset consists of 38 subjects in which there exist 56 images for each subject. Data is split into two groups: train set and test set that contain 51 and 5 images, respectively. The selection algorithms aim to pick up a few training images among all 51 ones to train a general classifier which is able to identify the test images.

Fig. 9 shows the normalized error of the projection of training data on the subspace spanned by the representatives. In this figure matrix $\boldsymbol{T}$ is the collection of the training data which representatives are selected from them. It is obvious that PCA indicates the best normalized error [5]. I.e., it can be interpreted as a lower bound for the projection error on any low-dimensional space. However, we aim to indicate the subspace only using few images of the training data set. The performance of random selection, K-medoids, D-optimal, and our suggested E-optimal selections are shown in this figure.

The projection error of test data is depicted in Fig. 10. In this figure matrix $\boldsymbol{T}$ is the collection of the test data which are not seen for selection procedure. Although the error of PCA representatives for training data is much less than the other methods due to over-learning of the bases, in the case of test data the performance of our suggested selection is approximately the same as that of PCA representatives. This means, we could span a generalized subspace by only using few selected images that are able to cover the desired signal space as well as PCA method that uses all of the training data.

Fig. 11 shows 40 images from the third subject of Extended Yale-B data set. As an example we are to select 6 images using K-medoids and our suggested algorithm. The results are shown in Fig. 12. The selected set of images using K-medoids do not contain the shadowing effect from the front side while our selection capture from different point of views.

5. According to the definition of PCA, it spans the best low-rank subspace that minimizes the normalized error defined in Fig. 9 for a set of training data.

TABLE 3: Accuracy of different classifiers using partial data for learning of Extended Yale-B dataset with 5 representatives.

	NN	NS	SRC	SVM
Random	26.8%	45.3%	72.0%	55.7%
Kmedoids	39.0%	61.1%	82.6%	68.2%
Volume sampling	76.3%	71.6%	88.9%	85.3%
E-optimal	**77.9%**	**82.6%**	**94.2%**	**90.0%**
All Data	81.4%	95.8%	97.1%	98.7%

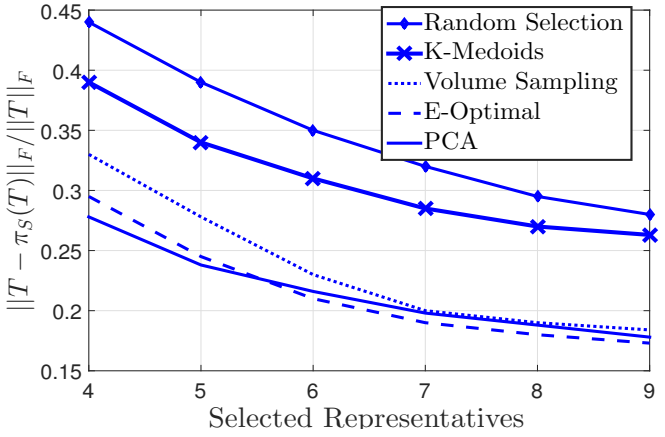


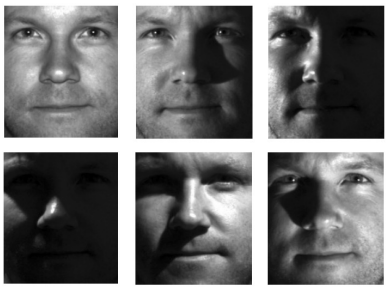
Fig. 10: The projection error of the test data into the subspace spanned by the selected rows.

The effect of data partitioning using the successive E-optimal selection on the performance of selection is studied for a larger data set. MISNT data set is used which contains $60,000$ sample images of handwritten digits [53]. Two criteria are considered, the first one is recognition rate using the learned classifier by reduced data and the second criterion is running time for data selection and data classification. Reducing the number of training data may decrease the performance of a classifier. A proper selection aims to preserve the recognition rate about the one using full data. On the other hand, reduced data make the training algorithm fast. Exploiting full data needs no process for selection but the training process needs a high amount of computations.

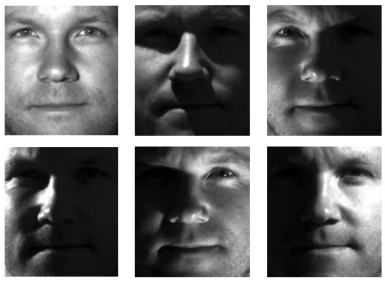
The basic E-optimal criterion is vulnerable to outlier data. It aims to select the most distinguished samples. However, unusual samples are probably different from each other and they satisfy the E-optimal criterion. The proposed two-phase algorithm first selects some candidates for final selection using E-optimal criterion and in the second phase the final selection reduces candidate samples to exact K selection. Fig. 13 shows the effect of two-phase algorithm on selection from 5842 samples of digit 4. The selected samples by E-optimal criterion are exceptional hand-written characters for digit 4. While, the two-phase algorithm selects visually proper representative for this class. Quantitative measures also will be demonstrated.

Sparse subspace classifier is learned by only few selected data. Four criteria are investigated for selection. D-optimal, the proposed E-optimal, K-medoids and the proposed two-phase algorithm are utilized for selection. D-optimal and E-optimal are vulnerable to outlier data as depicted in

Fig. 11: Training data corresponding to the third subject of Extended Yale-B data set. This data set contains different angles of shadowing for each subject.

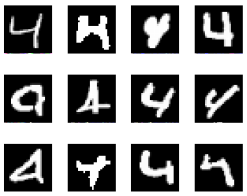


(a) Selected faces by K-Medoids selection.

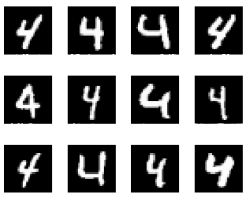


(b) Selected faces by E-optimal selection.

Fig. 12: Comparison of the proposed E-pptimal representatives versus K-mediods selection.



(a) E-optimal criterion on the whole data.



(b) Two phase distributed selection based on E-optimality.

Fig. 13: 12 selected images of digit 4 from 5842 images.

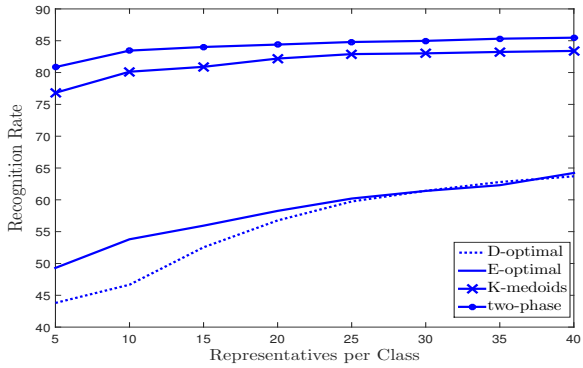


Fig. 14: Performance of nearest subspace classifier learned by few data from each class.

Fig. 13 (a). The k-medoids algorithm performs better than greedy algorithms for selection as it finds some points that data are concentrated around them. However, k-medoids algorithm is not tractable for real-time processing of big data. Our suggested two-phase algorithm outperforms K-medoids in terms successful classification rate. In addition to better representatives, our two-phase selection performs much faster than K-medoids algorithm. The running time of algorithms are shown in Fig. 15. Reducing the number of training signals saves a huge computation burden for training the classifier. In this figure algorithms are performed using an Intel Xeon CPU 3.7 Ghz and 8 GB RAM. A simple one nearest neighbor classifier needs 784 seconds to classify 5000 test images. While by selecting data it decreases to 2.83 seconds.

Deep learning achieves the best results for classification of MNIST data set. In order to compare the the effect of data selection on the state of the art method of classification, a deep neural network is learned with the selected data.

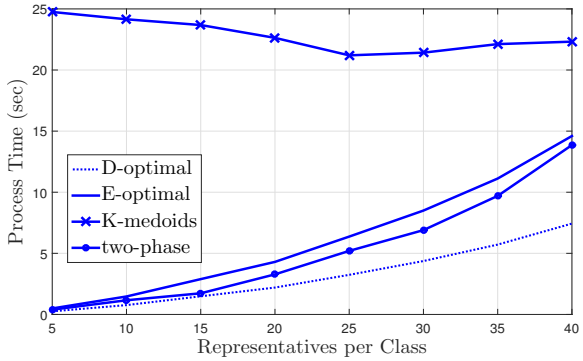


Fig. 15: Running time of selecting few data from each class.

MLP network and Capsules network [54] are employed to perform classification. Table 4 summarizes the accuracy of learned classifiers. The MLP network has three layers and the hidden layer contains 1000 neurons. As it can be seen selected training set improves the classification rate. For example, the learned network using 2,000 random images is working worse than the network which is learned by 1,000 selected images as the training set. However, Capsule net which exhibits one of the best performances for MNIST data set is less sensitive to the input training data and the improvement is less than that of MLP. Table 5 shows processing time for selection from 60,000 images for K-medoids algorithm and our proposed two-phase algorithm. Please note that the proposed algorithm can be implemented parallel which reduces the running time significantly. However, the centralized algorithm is simulated. The effect of data reduction on the speed of learning a deep network is presented in Table 6. The running time for one epoch is reported. MLP needs 20 epochs for convergence and it takes 500 epochs for CapsNet to reach the best performance. Thus, running time of MLP for whole data is about 30 seconds and for CapsNet is about 266 minutes. While, using only 1000 samples the running time for MLP decreases to only 1 second and for CapsNet it takes less than 3 minutes. Deep learning simulations are performed on Chainer framework [55] using 1 GPU of Nvidia TitanX and 12 GB RAM.

8 Conclusion

The problem of sensor selection is considered and its relation to existing work on matrix subset selection is elaborated. We developed a new subset selection method as an extension for the well-known volume sampling. Our criteria is based on E-optimality which is in favor of compressive sensing theory. Moreover the E-optimal criterion is extended to RIP-based sensor selection. Selection is an enabling step for efficient processing of a large amount of data, however for many cases selection from large data also is challenging. To this aim, successive and distributed implementation of the proposed algorithm are developed. Experimental results indicate the performance of our suggested sensor selection algorithm in cognitive radio networks' spectrum sensing as well as supervised learning with partial selected data.

References

[1] X. Jin and J. Han, "K-medoids clustering," in *Encyclopedia of Machine Learning*, pp. 564–565, Springer, 2011.

[2] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *Signal Processing, IEEE Transactions on*, vol. 57, pp. 451–462, Feb 2009.

[3] M. Shamaiah, S. Banerjee, and H. Vikalo, "Greedy sensor selection: Leveraging submodularity," in *Decision and Control (CDC), 2010 49th IEEE Conference on*, pp. 2572–2577, Dec 2010.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[5] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-nyquist sampling of sparse wideband analog signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 375–391, 2010.

[6] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013.

[7] B. Amizic, L. Spinoulas, R. Molina, and A. Katsaggelos, "Compressive blind image deconvolution," *Image Processing, IEEE Transactions on*, vol. 22, pp. 3994–4006, Oct 2013.

[8] X. Guan, Y. Gao, J. Chang, and Z. Zhang, "Advances in theory of compressive sensing and applications in communication," in *Instrumentation, Measurement, Computer, Communication and Control, 2011 First International Conference on*, pp. 662–665, Oct 2011.

[9] Z. Yu, S. Hoyos, and B. M. Sadler, "Mixed-signal parallel compressed sensing and reception for cognitive radio," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3861–3864, IEEE, 2008.

[10] Y. Gwon, H. Kung, and D. Vlah, "Compressive sensing with optimal sparsifying basis and applications in spectrum sensing," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, pp. 5386–5391, Dec 2012.

[11] T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *Information Theory, IEEE Transactions on*, vol. 57, pp. 4680–4688, July 2011.

[12] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed ℓ_0 norm," *Signal Processing, IEEE Transactions on*, vol. 57, pp. 289–301, Jan 2009.

[13] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM JOURNAL ON SCIENTIFIC COMPUTING*, vol. 20, pp. 33–61, 1998.

[14] Q. Geng and J. Wright, "On the local correctness of ℓ_1 minimization for dictionary learning," in *Information Theory (ISIT), 2014 IEEE International Symposium on*, pp. 3180–3184, June 2014.

[15] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, 2001.

[16] A. Cohen, W. Dahmen, and R. Devore, "Compressed sensing and best k-term approximation," *J. Amer. Math. Soc*, pp. 211–231, 2009.

[17] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[18] E. Candes and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, pp. 4203–4215, Dec 2005.

[19] E. Candes, "The restricted isometry property and its implications for compressed sensing," *C. R. Academie des Sciences*, no. 356, pp. 689–592, 2008.

[20] M. Davies and R. Gribonval, "Restricted isometry constants where ℓ_p sparse recovery can fail for $0 < p \leq 1$," *Information Theory, IEEE Transactions on*, vol. 55, pp. 2203–2214, May 2009.

[21] J. D. Blanchard, C. Cartis, and J. Tanner, "Compressed sensing: How sharp is the restricted isometry property?," *SIAM Rev.*, vol. 53, pp. 105–125, Feb. 2011.

[22] X. Peng, H. Tang, L. Zhang, Z. Yi, and S. Xiao, "A unified framework for representation-based subspace clustering of out-of-sample and large-scale data," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 12, pp. 2499–2512, 2016.

[23] T. Sajana, C. S. Rani, and K. Narayana, "A survey on clustering techniques for big data mining," *Indian Journal of Science and Technology*, vol. 9, no. 3, 2016.

[24] W. Song, Z. Deng, L. Wang, B. Du, P. Liu, and K. Lu, "G-ik-svd: parallel ik-svd on gpus for sparse representation of spatial big data," *The Journal of Supercomputing*, pp. 1–18, 2016.

[25] R. Zhang and J. T. Kwok, "Asynchronous distributed admm for consensus optimization.," in *ICML*, pp. 1701–1709, 2014.

[26] S. Scardapane, D. Wang, and M. Panella, "A decentralized training algorithm for echo state networks in distributed

TABLE 4: Performance of selection algorithms in terms percentage of classification rate using a deep neural network learned by partial data. The original data set contains 60,000 training images. (Left) K-medoids, (Middle) two-phase. (Right) Random Selection. Each classifier is learned by only 200, 500, 1000 and 2000 training data out of 60,000.

	200			500			1,000			2,000			60,000
MLP	87.45	**88.26**	80.84	89.81	91.12	88.52	92.13	93.11	91.03	93.79	94.50	92.91	98.25
CapsNet	84.38	81.85	78.61	92.76	**93.10**	91.87	96.11	**96.62**	95.72	**97.91**	97.69	97.59	99.66

TABLE 5: Running time (seconds) of data selection corresponding to Table 4.

	200	500	1000	2000
K-medoids	193.4	218.4	433.2	1328
two-phase	31.61	160.8	191.3	280.6

TABLE 6: Running time (seconds) of neural network learning corresponding to Table 4. Running time per epoch is reported.

	200	500	1000	2000	60,000
MLP	0.029	0.04	0.051	0.078	1.55
CapsNet	0.073	0.18	0.39	0.88	32.18

big data applications," *Neural Networks*, vol. 78, pp. 65–74, 2016.

[27] D. Hajinezhad, T.-H. Chang, X. Wang, Q. Shi, and M. Hong, "Nonnegative matrix factorization using admm: Algorithm and convergence analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 4742–4746, IEEE, 2016.

[28] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

[29] P. Bühlmann, P. Drineas, M. Kane, and M. van der Laan, *Handbook of Big Data*. Chapman and Hall/CRC, 2016.

[30] H. Jamali-Rad, A. Simonetto, and G. Leus, "Sparsity-aware sensor selection: Centralized and distributed algorithms," *Signal Processing Letters, IEEE*, vol. 21, pp. 217–220, Feb 2014.

[31] G. Nemhauser and L. Wolsey, "Best algorithms for approximating the maximum of a submodular set function," CORE Discussion Papers RP 343, Universiti catholique de Louvain, Center for Operations Research and Econometrics (CORE).

[32] A. Deshpande and L. Rademacher, "Efficient volume sampling for row/column subset selection," in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pp. 329–338, IEEE, 2010.

[33] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang, "Matrix approximation and projective clustering via volume sampling," in *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pp. 1117–1126, Society for Industrial and Applied Mathematics, 2006.

[34] 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.

[35] A. K. Farahat, A. Elgohary, A. Ghodsi, and M. S. Kamel, "Greedy column subset selection for large-scale data sets," *Knowledge and Information Systems*, vol. 45, no. 1, pp. 1–34, 2015.

[36] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An improved approximation algorithm for the column subset selection problem," in *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, (Philadelphia, PA, USA), pp. 968–977, Society for Industrial and Applied Mathematics, 2009.

[37] P. Van Dooren, "Numerical linear algebra for signal, systems and control," *Draft notes prepared for the Graduate School in Systems and Control*, vol. 250, 2003.

[38] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[39] A. Tillmann and M. Pfetsch, "The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing," *Information Theory, IEEE Transactions on*, vol. 60, pp. 1248–1259, Feb 2014.

[40] C. C. Aggarwal, Y. Xie, and P. S. Yu, "On dynamic data-driven selection of sensor streams," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1226–1234, ACM, 2011.

[41] G.-J. Qi, C. Aggarwal, D. Turaga, D. Sow, and P. Anno, "State-driven dynamic sensor selection and prediction with state-stacked sparseness," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, (New York, NY, USA), pp. 945–954, ACM, 2015.

[42] M. Holmes, A. Gray, and C. Isbell, "Fast svd for large-scale matrices," in *Workshop on Efficient Machine Learning at NIPS*, vol. 58, pp. 249–252, 2007.

[43] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[44] S. V. Tenneti and P. Vaidyanathan, "A unified theory of union of subspaces representations for period estimation," *IEEE Transactions on Signal Processing*, vol. 64, no. 20, pp. 5217–5231, 2016.

[45] J. Bazerque and G. Giannakis, "Distributed spectrum sensing for cognitive radio networks by exploiting sparsity," *Signal Processing, IEEE Transactions on*, vol. 58, pp. 1847–1862, March 2010.

[46] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3869–3872, March 2008.

[47] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1, pp. 245–271, 1997.

[48] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[49] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Computer vision and pattern recognition, 2003. Proceedings. 2003 IEEE computer society conference on*, vol. 1, pp. I–11, IEEE, 2003.

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

[51] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

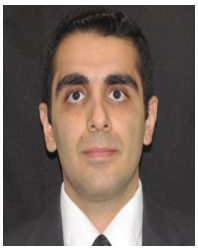
[52] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 5, pp. 684–698, 2005.

[53] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[54] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, pp. 3859–3869, 2017.
[55] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

Mohsen Joneidi received his B.S. and M.S. from Ferdowsi University of Mashhad and Sharif University of Technology in 2009 and 2012, respectively. He joined CWNLAB in Electrical and Computer Engineering Department at University of Central Florida, Orlando, Florida. His research interests include statistical signal processing and compressed sensing.

Alireza Zaeemzadeh (S11) received the B.S. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2014. He is currently working toward the M.Sc. degree in electrical engineering at the University of Central Florida. His current research interests lie in the areas of statistical signal processing and Bayesian data analysis.



Behzad Shahrasbi received the Bachelors degree from Amirkabir University of Technology, Tehran, Iran in 2006, Masters degree from Oklahoma State University, Stillwater, OK in 2011 and the Ph.D. degree from University of Central Florida, Orlando, FL in 2015. His research interests include sparse signal representations, compressed sensing and recovery algorithms, low rank matrix recovery, and approximation.

Guo-Jun Qi Dr. Guo-Jun Qi is an assistant professor in the Department of Computer Science at University of Central Florida. He received the Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Urbana Champaign. His research interests include pattern recognition, machine learning, computer vision and multimedia. He was the co-recipient of the best student paper award in IEEE Conference on Data Mining (2014), and the recipient of the best paper award (2007) and the best paper runner-up (2015) in the ACM International Conference on Multimedia. He has served or will serve as program co-chair of MMM 2016, an area chair of ACM Multimedia (2015, 2016), a senior program committee member of ACM CIKM 2015 and ACM SIGKDD 2016, and program committee members or reviewers for the conferences and journals in the fields of computer vision, pattern recognition, machine learning, and data mining. Dr. Qi has published over 60 academic papers in these areas. He also (co-)edited the two special issues on IEEE transactions on multimedia and IEEE transactions on big data.

Nazanin Rahnavard received her Ph.D. in the School of Electrical and Computer Engineering at the Georgia Institute of Technology, Atlanta, in 2007. She is currently an Associate Professor in the Department of Electrical and Computer Engineering at the University of Central Florida, Orlando, Florida. Dr. Rahnavard is the recipient of NSF CAREER award in 2011. She has interest and expertise in a variety of research topics in the communications, networking, and signal processing areas. She serves on the editorial board of the Elsevier Journal on Computer Networks (COMNET) and on the Technical Program Committee of several prestigious international conferences.