Sparsity-Aware Block Diagonal Representation for Subspace Clustering

Aylin Taştan[†], Michael Muma[‡], Esa Ollila[*], Abdelhak M. Zoubir[†]
[†]*Signal Processing Group, Technische Universität Darmstadt, 64283 Darmstadt, Germany*
[‡]*Robust Data Science Group, Technische Universität Darmstadt, 64283 Darmstadt, Germany*
[*]*Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland*

***Abstract*—A block diagonally structured affinity matrix is an informative prior for subspace clustering which embeds the data points in a union of low-dimensional subspaces. Structuring a block diagonal matrix can be challenging due to the determination of an appropriate sparsity level, especially when outliers and heavy-tailed noise obscure the underlying subspaces. We propose a new sparsity-aware block diagonal representation (SABDR) method that robustly estimates the appropriate sparsity level by leveraging upon the geometrical analysis of the low-dimensional structure in spectral clustering. Specifically, we derive the Euclidean distance between the embeddings of different clusters to develop a computationally efficient density-based clustering algorithm. In this way, the sparsity parameter selection problem is re-formulated as a robust approximation of target between-clusters distances. Comprehensive experiments using real-world data demonstrate the effectiveness of SABDR in different subspace clustering applications.**

***Index Terms*—Block diagonal representation, subspace clustering, spectral clustering, affinity matrix, clustering**

I. Introduction

Determining an embedding so that the data points lie in a union of low-dimensional subspaces is crucial in many real-world problems such as in clustering [1–6], supervised learning [7] and semi-supervised learning [8, 9]. In particular, subspace clustering (SC) has numerous applications e.g. motion segmentation [10, 11], face clustering [1, 3], image segmentation [12] and community clustering in social networks [13]. Motivated by its broad range of applications, SC has been the subject of much research, which can loosely be divided into four main categories, i.e., iterative [14], algebraic [10, 15], statistical [11] and spectral clustering-based methods [1–5, 16]. In recent years, the latter have attracted increasing interest due to their simplicity and promising performance [3, 4].

The first step of spectral clustering-based methods is to compute an affinity matrix (see, e.g., [1–5, 17] for different variations of affinity matrix construction). Block diagonally structured affinity matrices constitute an informative prior, that is frequently used (e.g., [2–4, 17]). A popular strategy to construct a block diagonal (BD) models is to represent the data as a linear combination of feature vectors while regularizing the affinity matrix coefficients, e.g. with an ℓ_1, ℓ_2 or nuclear norm [2–4]. Recent methods apply mixed norms, such as, the elastic net, which have the advantage of providing a tradeoff between sparsity and connectedness [4, 18]. A major challenge for all these approaches is to determine the appropriate level of sparsity which plays a crucial role in SC performance. Different methods building upon supervised learning algorithms [19, 20], similarity coefficients' distribution [21], geometric interpretations [22], connectedness [23] and eigenvalues [17] have been proposed; however, no optimal approach exists, especially in the presence of outliers.

We propose a *Sparsity-Aware Block Diagonal Representation (SABDR)* method to robustly estimate the appropriate level of sparsity for SC. The proposed SABDR approach leverages upon the geometrical analysis of the low-dimensional structure in spectral clustering. In particular, the derived Euclidean distances between the embeddings of different clusters are utilized to construct the BD affinity matrix. Further, we propose a computationally efficient density-based clustering (Con-DBSCAN) algorithm, to obtain a robust estimate of the between-clusters distances that are associated with an available affinity matrix. Unlike the original DBSCAN [24], by leveraging upon a new Theorem that decribes the geometry of the embeddings, Con-DBSCAN determines the neighborhood search radius around given points based on their connectedness, therewith leveraging the derived geometric information. This, in contrast to DBSCAN, enables Con-DBSCAN to efficiently expand clusters with multiple embedding vectors in a single iteration. The proposed modification leads to a considerable speed-up without any performance loss. Building upon our theoretic analysis, we develop a regularization parameter [3] selection by re-formulating the sparsity level selection problem as an approximation of the target between-clusters distances.

Organization: Sec. II presents the motivation and problem formulation while Sec. III is dedicated to the proposed SABDR method. A performance benchmarking against state-of-the-art methods using popular data sets is provided in Sec. IV. Finally, conclusions are drawn in Sec. V.

II. Motivation and Problem Formulation

A. Spectral Analysis of Block Diagonal Affinity Matrix

Let a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathbb{R}^{M\times N}$ with M denoting the dimension and N the number of feature vectors be representable as a graph $G=\{V, E, \mathbf{W}\}$, where V denotes the vertices, E represents the edges, and $\mathbf{W}\in\mathbb{R}^{N\times N}$ is a zero-

The work of A. Taştan is supported by the Republic of Turkey Ministry of National Education.

The work of M. Muma is supported by ERC Starting Grant ScReeningData grant agreement no. 101042407.

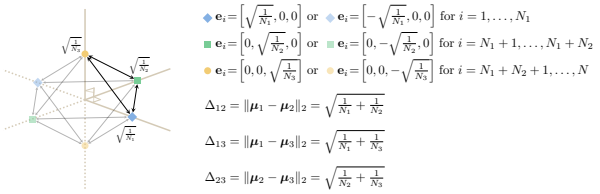
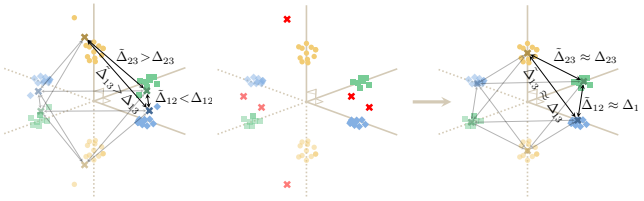


Fig. 1: Spectral embedding according to the eigenvectors of the Laplacian matrix $\mathbf{L}$ when $K = 3$.



(a) Corrupted separation (b) Robust spectral analysis using DBSCAN

Fig. 2: Robustness in spectral analysis.

diagonal symmetric affinity matrix that is computed from $\mathbf{X}$ by choosing an appropriate similarity measure (see, for example, [3]). Assuming that $\mathbf{L} \in \mathbb{R}^{N\times N}$ denotes the nonnegative definite Laplacian matrix, spectral clustering performs K-means clustering on the eigenvectors associated with the K smallest eigenvalues of the eigen-problem

$$\mathbf{L}\mathbf{y}_k = \lambda_k \mathbf{D}\mathbf{y}_k, \quad k = 0, ..., K-1 \tag{1}$$

with associated eigenvalues $0 \leq \lambda_0 \leq \lambda_1 \leq ... \leq \lambda_{N-1}$. Here, $\mathbf{y}_k$ denotes the eigenvector associated with the kth eigenvalue λ_k, $\mathbf{D} \in \mathbb{R}^{N\times N}$ is a diagonal weight matrix with overall edge weights $d_{i,i} = \sum_j w_{i,j}$ on the diagonal and $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

Theorem 1. *Let $\mathbf{L} \in \mathbb{R}^{N\times N}$ be the Laplacian matrix corresponding to a K block zero-diagonal symmetric affinity matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ in which every block $k = 1, ..., K$ is associated to N_k ($N_k \in \mathbb{Z}$) > 1 feature vectors and the affinities outside the blocks are zero-valued. Further, let $\mathbf{Y} = [\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_{K-1}] \in \mathbb{R}^{N\times K}$ be the matrix of eigenvectors associated with the K smallest eigenvalues of $\mathbf{L}$. Finally, let $\mathbf{e}_i$, the ith row vector of $\mathbf{Y}$, denote the embedding vector that represents the M-dimensional ith feature vector $\mathbf{x}_i$ in the reduced K-dimensional space. Assuming that the eigenvectors are orthonormal, the Euclidean distance between any embedding vector pairs $\mathbf{e}_i$ and $\mathbf{e}_j$ associated to distinct blocks k and l is equal to $\|\mathbf{e}_i - \mathbf{e}_j\|_2 = \sqrt{1/N_k + 1/N_l}$ for $k \neq l$ and $i \neq j$.*

Proof. See Appendix A. □

Theorem 1 is illustrated in Fig. 1 for an example consisting of $K = 3$ blocks where $\boldsymbol{\mu}_k \in \mathbb{R}^K$ denotes the cluster centroid corresponding to block $k = 1,...,K$. As can be seen, the Euclidean distance between embeddings of distinct blocks is a function of their block sizes. In the sequel, the geometrical analysis of a BD affinity matrix (Theorem 1) will serve as a target to robustly estimate the distance between clusters.

B. Motivation : DBSCAN for Robust Spectral Analysis

If the affinity matrix of the data is BD, spectral clustering may provide excellent results. Furthermore, according to Theorem 1, a BD affinity matrix will lead to densely connected clusters in the embedding space. Hence, a density-based clustering approach, such as, DBSCAN [24], is a natural approach to achieve a BD structure. However, in real world scenarios the data includes outliers and heavy-tailed noise which may obscure the distance between embeddings of different clusters. Therefore, beyond its computational efficiency that has made DBSCAN very popular, we build upon its intrinsic outlier detection ability to increase robustness for spectral analysis. Fig. 2 illustrates this with an example of $K = 3$ clusters that are hidden in a matrix of corrupted eigenvectors $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_0, \tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2]$. Even though an appropriate level of sparsity provides densely connected clusters, the outliers obscure the distance between different clusters as it is shown in Fig. 2a. By contrast, as illustrated in Fig. 2b, DBSCAN identifies the outliers and robustly estimates the between-clusters distance information.

C. Problem Formulation

Given a dataset of feature vectors $\mathbf{X} \in \mathbb{R}^{M\times N}$, the aim of this work is to robustly and efficiently find a K block zero-diagonal symmetric affinity matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ to obtain a vector of labels $\mathbf{c} \in \mathbb{R}^N$ consisting cluster assignments of $\mathbf{X}$ using the spectral information of BD affinity matrices.

III. PROPOSED METHOD

This section details the proposed SABDR method which estimates a BD affinity matrix in three steps that are detailed in the following sections.

A. Affinity Matrix Construction

Among numerous affinity matrix design methods such as [1–5], this method adapts the BDR method in [3], in which the proposed K-block regularizer promotes a nonnegative symmetric matrix to be K-BD so that the spectral analysis described in Section II-A is directly applicable.

Let $\mathbf{W}^{(m)}$ and $\mathbf{L}^{(m)} \in \mathbb{R}^{N\times N}$, respectively, be the affinity and Laplacian matrix that are computed by using as a regularization parameter pair $\mathbf{p}_m = [p_{m,1}, p_{m,2}]^\top$ from a matrix of candidate regularization parameter pairs $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{N_\mathbf{P}}] \in \mathbb{R}^{2\times N_\mathbf{P}}$. Assuming that for every $\mathbf{p}_m$ there exists a matrix of eigenvectors[1] $\mathbf{Y}^{(m)} \in \mathbb{R}^{N\times K}$, the following sections present the selection of an appropriate $\mathbf{p}_m$ based on robust spectral analysis of $\mathbf{Y}^{(m)}$ with the proposed Con-DBSCAN.

B. Block Size Estimation using Con-DBSCAN

1) Parameter Definition: ϵ: If follows from Theorem 1 that there exists a specific level of sparsity that allows for an embedding, such that, the distance between embeddings of the same cluster is minimal while the distance between embeddings of different clusters is maximal. This important result implies that there exists a minimum neighborhood search radius ϵ that will provide these highly dense clusters. To provide a visual understanding, the geometric definition of a minimum search radius is shown in Fig. 3. Considering a pair of clusters, the two clusters are assumed to have a maximum $N_{\max}$ and a minimum $N_{\min}$ number of samples

[1] If the obtained set is not orthonormal, the Gram-Schmidt algorithm [25] can be used.

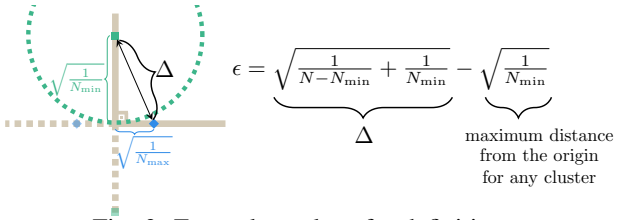


Fig. 3: Examplary plot of ϵ definition.

based on the information that the maximum block size results in the minimum distance from the origin and vice-versa. Clearly, $N_{\max}$ can achieve its greatest value for $N_{\max} = N - N_{\min}$. Then, using Theorem 1, the minimum ball can be simply calculated for the large cluster as $\epsilon = \sqrt{1/(N - N_{\min}) + 1/N_{\min}} - \sqrt{1/N_{\min}}$.

2) Parameter Definition: $N_{\min}$: As the parameter ϵ is a function of the minimum number of points $N_{\min}$, this section discusses finding $N_{\min}$. To begin, DBSCAN assigns a sample to be an outlier if its ϵ-neighborhood does not contain at least $N_{\min}$ neighbors which means that a large value of $N_{\min}$ results in assigning many samples as outlier. On the other hand, a large value of $N_{\min}$ increases the neighborhood search radius which may result in assigning many samples into a big cluster and the remaining samples as outliers. Therefore, $N_{\min}$ must be a reasonably small number which can be easily defined if K is known. In particular, the parameter can be gradually increased as long as the clustering results provide K clusters.

3) Con-DBSCAN Algorithm: For a matrix $\mathbf{E}^{(m)} = (\mathbf{Y}^{(m)})^\top = [\mathbf{e}_1^{(m)}, \mathbf{e}_2^{(m)}, ..., \mathbf{e}_N^{(m)}] \in \mathbb{R}^{K \times N}$ containing embedding vectors associated with $\mathbf{p}_m$, the goal of this section is to assign embedding vectors into mutually exclusive clusters using the proposed Con-DBSCAN. As in the original DBSCAN [24], the method starts with ϵ-neighborhood computation of an unlabelled embedding vector $\mathbf{e}_i^{(m)} \in \mathbf{E}^{(m)}, i = 1, ..., N$ as

$$\mathcal{N}_i^{(m)} = \{\mathbf{e}_j^{(m)} \in \mathbb{R}^K : \|\mathbf{e}_i^{(m)} - \mathbf{e}_j^{(m)}\|_2 < \epsilon\}. \quad (2)$$

Here, $\mathcal{N}_i^{(m)}$ is the ϵ-neighborhood set of $\mathbf{e}_i^{(m)}$ and $\mathbf{e}_j^{(m)}$ is the jth embedding result with $j = 1, ..., N$ and $j \neq i$. Similar to [24], Con-DBSCAN assigns $\mathbf{e}_i^{(m)}$ into a cluster $c^{(m)}$ if its $\mathcal{N}_i^{(m)}$ includes more neighbors than $N_{\min}$. However, there is an important difference in *how* the neighbors are included. While DBSCAN [24] must compute the ϵ-neighborhood for every neighbor and iteratively expand cluster $c^{(m)}$ with embedding vectors, *using Theorem 1, Con-DBSCAN expands the cluster $c^{(m)}$ in a single iteration*, which leads to a considerable speed-up without any loss of performance compared to the original DBSCAN. More specifically, the derived geometric information in Theorem 1 demonstrates that the embeddings associated with connected nodes are concentrated while being far apart from the embeddings of unconnected nodes in the embedding space. Con-DBSCAN exploits this geometric information by comparing the connectedness of a candidate neighbor to that of the least connected embedding in $c^{(m)}$ which are, respectively, defined by

$$\kappa_{\text{cand}}^{(m)} = \sum_{i \in c^{(m)}} w_{\text{cand},i}^{(m)} \ , \quad (3)$$

Algorithm 1: SABDR-based SC

Input: $\mathbf{X} \in \mathbb{R}^{M \times N}$, $\mathbf{P} \in \mathbb{R}^{2 \times N_{\mathbf{P}}}$, $N_{\min}$(See, Sec. III-B2), K (Optional based on BDR method)
for $\mathbf{p}_m = \mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{N_{\mathbf{P}}}$ **do**
  Affinity Matrix Construction
  Obtain an initial affinity matrix $\mathbf{W}^{(m)} \in \mathbb{R}^{N \times N}$, e.g. using [3].
  Compute $\mathbf{Y}^{(m)} \in \mathbb{R}^{N \times K}$ using Eq. (1).
  Block Size Estimation using Con-DBSCAN:
  Calculate ϵ as described in Section III-B1.
  while *there exists an unlabelled embedding vector* **do**
    Select the first unlabelled embedding $\mathbf{e}_i^{(m)}$ and compute the ϵ-neighborhood $\mathcal{N}_i^{(m)}$ via Eq. (2).
    Initialize the cluster label $c^{(m)}$.
    if *the number of ϵ neighbors* $> N_{\min}$ **then**
      Expand cluster $c^{(m)}$ using Eq. (3) and Eq. (4).
    else
      Assign $\mathbf{e}_i^{(m)}$ as outlier.
    end
  end
  Sparsity Level Estimation:
  Calculate $\hat{\mathbf{\Delta}}_{\text{T}}^{(m)}$ and $\hat{\mathbf{\Delta}}^{(m)}$ using Eq. (5) and Eq. (6).
  Update the estimate based on Eq (7).
end
Obtain associated $\hat{\mathbf{Y}}^{(\hat{\mathbf{P}})}$ and perform K-means.
Output: A vector of estimated cluster labels $\hat{\mathbf{c}}$

and

$$\kappa_{\min}^{(m)} = \min\left\{\kappa_i^{(m)} = \sum_{j \in c^{(m)}} w_{i,j}^{(m)}\right\}, \quad (4)$$

where $\kappa_{\text{cand}}^{(m)}$ denotes the connectedness of a candidate neighbor embedding $\mathbf{e}_{\text{cand}}^{(m)} \in \mathcal{N}_i^{(m)}$, $w_{i,j}^{(m)}$ is the i,jth similarity coefficient in $\mathbf{W}^{(m)}$, $\kappa_i^{(m)}$ is the connectedness associated with the ith embedding vector $\mathbf{e}_i^{(m)}$, and $\kappa_{\min}^{(m)}$ is the minimum connectedness in cluster $c^{(m)}$. Based on the embedding idea that connected nodes are embedded closely [26], the method expands clusters using all highly connected neighbors $\forall \kappa_{\text{cand}}^{(m)} \in \mathcal{N}_i^{(m)}$, such that, $\kappa_{\text{cand}}^{(m)} > \kappa_{\min}^{(m)}$. Then, it iterates the ϵ-neighborhood computation on the unlabelled neighbors and repeats the connectedness-based expansion until no new neighbors that can be assigned to $c^{(m)}$ are found.

C. Sparsity Level Estimation

Let $\{\hat{N}_1^{(m)}, \hat{N}_2^{(m)}, ..., \hat{N}_K^{(m)}\} \in \mathbb{Z}^+$ be the block sizes associated with $\mathbf{p}_m$ that have been estimated in Section III-B3. Now, using Theorem 1, the components of the target between-clusters matrix estimate $\hat{\mathbf{\Delta}}_{\text{T}}^{(m)}$ are determined by

$$\hat{\Delta}_{\text{T}_{k,l}}^{(m)} = \begin{cases} \sqrt{\frac{1}{\hat{N}_k} + \frac{1}{\hat{N}_l}}, & \text{if } k \neq l \\ 0, & \text{otherwise} \end{cases} . \quad (5)$$

where $\hat{\Delta}_{\text{T}_{k,l}}^{(m)}$ is the k,lth component of $\hat{\mathbf{\Delta}}_{\text{T}}^{(m)}$ that represents the target Euclidean distance between cluster $k = 1, ..., K$ and $l = 1, ..., K$. Similarly, the components of the between-clusters matrix estimate $\hat{\mathbf{\Delta}}^{(m)}$ are computed by

$$\hat{\Delta}_{k,l}^{(m)} = \begin{cases} \|\hat{\boldsymbol{\mu}}_k^{(m)} - \hat{\boldsymbol{\mu}}_l^{(m)}\|_2, & \text{if } k \neq l \\ 0, & \text{otherwise} \end{cases} , \quad (6)$$

where $\hat{\Delta}_{k,l}^{(m)}$ is the k,lth component of $\hat{\mathbf{\Delta}}^{(m)}$ denoting the Euclidean distance between kth and lth estimated cluster centroids $\hat{\boldsymbol{\mu}}_k^{(m)}$ and $\hat{\boldsymbol{\mu}}_l^{(m)} \in \mathbb{R}^K$, respectively. Since for every

	Subspace Performances on ORL Data Set							Subspace Performances on JAFFE Data Set						
	Min-Max Average Clustering Accuracy for Different Regularization Parameters							Min-Max Average Clustering Accuracy						
K	SSC	EnSC	BD-SSC	RKLRR	IBDLR	BDR-B	**SABDR**	SSC	EnSC	BD-SSC	RKLRR	IBDLR	BDR-B	**SABDR**
2	52.6 - 87.4	53.9 - 60.4	51.3 - 66.6	54.7 - 64.1	x	55.0 - 97.0	95.3	51.4 - 99.2	52.0 - 58.7	50.7 - 68.6	52.0 - 61.8	51.4 - 63.6	52.1 - 99.5	97.5
3	36.7 - 88.1	36.7 - 62.3	36.8 - 59.4	36.7 - 57.5	36.7 - 68.9	36.7 - 92.8	88.5	35.0 - 97.7	35.0 - 68.9	35.3 - 71.1	35.1 - 53.9	35.1 - 63.7	35.1 - 97.8	95.0
5	22.0 - 84.3	22.0 - 67.2	23.3 - 48.1	22.0 - 52.7	22.0 - 64.4	22.0 - 84.8	86.6	21.4 - 97.2	21.4 - 85.3	22.4 - 87.4	21.4 - 70.3	21.4 - 84.2	21.4 - 97.4	96.6
8	13.8 - 83.0	13.8 - 71.5	15.9 - 40	13.8 - 69.7	13.8 - 73.8	13.8 - 82.1	80.2	13.5 - 93.4	13.5 - 82.8	15.1 - 93.1	13.5 - 85.9	13.5 - 87.7	13.5 - 92.5	89.2
10	11.0 - 81.4	11.0 - 70.5	13.2 - 35.3	11.0 - 69.8	11.0 - 73.1	11.0 - 80.6	78.9	10.8 - 85.0	10.8 - 78.9	12.7 - 85.4	10.8 - 84.0	10.8 - 85.4	10.8 - 85.9	76.5

TABLE I: Face clustering performance of different BDR methods on ORL and JAFFE data sets. 'x' denotes the failed results.

	Min-Max Average Clustering Accuracy for Different Regularization Parameters						
K	SSC	EnSC	BD-SSC	RKLRR	IBDLR	BDR-B	**SABDR**
2	52.2 - 93.8	52.4 - 62.8	51.1 - 71.8	52.5 - 60.9	52.2 - 68.8	52.7 - 96.6	95.2
3	35.0 - 88.3	35.0 - 65.0	35.4 - 67.8	35.0 - 52.7	35.0 - 67.4	35.4 - 91.2	83.3
5	21.0 - 85.5	21.0 - 74.8	22.3 - 69.5	21.0 - 53.7	21.0 - 70.6	22.1 - 88.4	83.2
8	13.1 - 79.9	13.1 - 76.7	15.3 - 72.2	13.1 - 66.2	13.1 - 70.6	14.3 - 81.3	73.3
10	10.5 - 76.3	10.5 - 73.5	12.8 - 70.2	10.5 - 65.7	10.5 - 67.9	11.6 - 76.8	70.3

TABLE II: Object clustering performance of different BDR methods on COIL20 data set.

	Min-Max Average Clustering Accuracy for Different Regularization Parameters						
K	SSC	EnSC	BD-SSC	RKLRR	IBDLR	BDR-B	**SABDR**
2	50.8 - 81.9	50.8 - 54.0	50.7 - 94.4	51.0 - 61.8	50.8 - 75.2	50.9 - 89.8	82.0
3	34.0 - 74.5	34.0 - 51.8	34.2 - 87.0	34.0 - 59.7	34.0 - 69.0	34.7 - 78.7	77.1
5	20.4 - 62.7	20.4 - 56.2	22.6 - 79.1	20.4 - 57.8	20.4 - 64.4	21.3 - 62.5	80.3
8	12.8 - 57.8	12.8 - 54.4	19.4 - 70.8	12.8 - 59.1	12.8 - 60.3	13.1 - 56.3	45.6
10	10.2 - 55.8	10.2 - 55.0	18.8 - 66.4	10.2 - 56.6	10.2 - 54.4	10.2 - 61.2	17.4

TABLE III: Handwritten-digit clustering performance of different BDR methods on USPS data set.

$\mathbf{p}_m$ there exists a distance matrix $\hat{\mathbf{\Delta}}^{(m)}$, the appropriate regularization parameter pair controlling the sparsity level can be estimated as follows:

$$\hat{\mathbf{p}} = \underset{\mathbf{p}_m \in \mathbf{P}}{\operatorname{argmin}} \| \hat{\mathbf{\Delta}}_{\mathrm{T}}^{(m)} - \hat{\mathbf{\Delta}}^{(m)} \|_F \tag{7}$$

The proposed SABDR for SC is summarized in Algorithm 1.

IV. Experimental Results

In this section, the SC performance of SABDR is benchmarked against five state-of-the art affinity matrix construction methods, i.e., SSC [1], EnSC [18], BD-SSC [2], BDR-B [3], IBDLR [4] and RKLRR [5] using the real-world data sets of face, object and handwritten digit images. The application details are as follows.

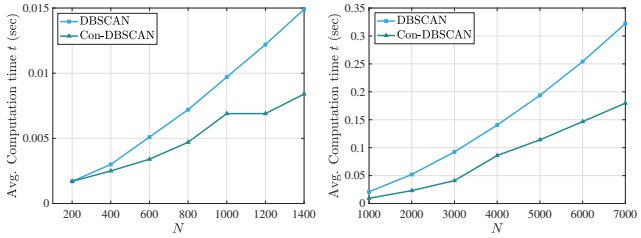
1) *ORL [27]:* As in [4], 400 face images of 40 subjects are resized to 32×32 and $\mathbf{X}$ of size 1024×400 is computed.

2) *JAFFE [28]:* Similarly, 213 images of 10 subjects are resized to 64×64 pixels and $\mathbf{X}$ of size 4096×213 is obtained.

3) *COIL20 [29]:* $\mathbf{X}$ of size 1024×400, whose column vectors contain 32×32 down-sampled images, is generated by selecting 20 images randomly for every object.

4) *USPS [30]:* $\mathbf{X}$ with $M = 256$ and $N = 500$ is generated by randomly selecting 50 handwritten digit images of size 16×16 as feature vectors for every digit.

As in [3], performance analysis of every application is conducted for an increasing value of K, e.g., $K = \{2, 3, 5, 8, 10\}$ using 100 randomly selected subject combinations. To reduce the cost, the feature spaces are, respectively, reduced to 10, 8, 10 and 13 using Principal Component Analysis (PCA), since using a larger feature space did not provide significant improvements. For the competing methods, the regularization parameters are manually tuned on a grid of 50 values. Finally, spectral clustering [31] is performed and the performance is summarized for the average clustering accuracy $\bar{p}_{\mathrm{acc}}$. The performance of SABDR is analyzed for the default parameter choice $N_{\min} = N/(2K)$,



(a) COIL20 (b) USPS

Fig. 4: Numerical results for different $N_{\min}$ parameters.

except for $K = 8$ and $K = 10$ in Table III, where the parameter was increased until it computes K distinct clusters (see Sec. III-B2). A MATLAB implementation of SABDR is available at: https://github.com/A-Tastan/SABDR.

Tables. I, II and III summarize the obtained results on face, object and handwritten digit clustering, respectively. As can be seen, SABDR in nearly all cases reaches a performance close to, or in some cases even better than that of optimally tuned competitors. This demonstrates its excellent sparsity level estimation performance.

In addition to the SC performance analysis, it is interesting to compare the computation time (t) of the proposed Con-DBSCAN method to that of DBSCAN. To analyze t as a function of sample size N, the data sets are generated by randomly selecting an increasing number of samples for every subject in COIL20 [29] and USPS [30], respectively. Then, the original DBSCAN [24] and Con-DBSCAN are run using the eigenvectors associated with the affinity matrices that are obtained by manually tuning BDR-B [3]. As can be seen in Fig. 4, the proposed method is considerably faster than the original DBSCAN [24] which confirms that the derived strategy of expanding the clusters efficiently with multiple highly connected neighbors greatly reduces the number of required ϵ-neighborhood search iterations.

V. Conclusion

The eigenvectors associated with the BD matrix are analyzed to show the significance of the hidden between-clusters distance information in the measure of an appropriate sparsity level for SC. Based on the derived theoretical information, we proposed SABDR which controls the level of sparsity by robustly estimating the regularization parameter/s. To use the available BD structure in the objective function, we proposed an efficient density-based clustering method Con-DBSCAN. SABDR is benchmarked against popular affinity matrix construction methods and it reached similar or higher performance than its optimally tuned competitors.

VI. Appendix : Proof of Theorem 1

Based on the information that the K smallest eigenvalues of the Laplacian matrix associated with

the BD affinity matrix are zero-valued [31], the associated orthonormal set of eigenvectors yields, i.e.

$$\mathbf{y}_0 = [\overbrace{\pm\sqrt{1/N_1},...,\pm\sqrt{1/N_1}}^{N_1}, \overbrace{0,...,0}^{N_2}, \ldots, \overbrace{0,...,0}^{N_K}]^\top$$
$$\mathbf{y}_1 = [0,...,0,\pm\sqrt{1/N_2},...,\pm\sqrt{1/N_2},\ldots,0,...,0]^\top$$
$$\vdots$$
$$\mathbf{y}_{K-1} = [0,...,0,0,...,0,\ldots,\pm\sqrt{1/N_K},...,\pm\sqrt{1/N_K}]^\top$$

where $\mathbf{y}_k \in \mathbb{R}^N$ is the eigenvector associated with the kth zero-valued eigenvalue. The Euclidean distance between any row vector of $\mathbf{Y}=[\mathbf{y}_0,\mathbf{y}_1,...,\mathbf{y}_{K-1}] \in \mathbb{R}^{N\times K}$ associated with different blocks follows as $\sqrt{1/N_k + 1/N_l}$.

REFERENCES

[1] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 2765-2781, 2013.

[2] J. Feng, Z. Lin, H. Xu and S. Yan, "Robust subspace segmentation with block-diagonal prior," in *Proc. IEEE Conf. Comp. Vision Pattern Recognit.*, pp. 3818-3825, 2014.

[3] C. Lu, J. Feng, Z. Lin, T. Mei and S. Yan, "Subspace clustering by block diagonal representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, pp. 487-501, 2018.

[4] X. Xie, X. Guo, G. Liu and J. Wang, "Implicit block diagonal low-rank representation," *IEEE Trans. Image Process.*, vol. 27, pp. 477-489, 2017.

[5] S. Xiao, M. Tan, D. Xu and Z. Y. Dong, "Robust kernel low-rank representation", *IEEE Trans. Neural Networks Learn. Syst.*, vol 27, pp. 2268-2281, 2015.

[6] J. Sui, Z. Liu, L. Liu, A. Jung, T. Liu, B. Peng and X. Li, "Sparse subspace clustering for evolving data streams," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.(ICASSP)*, pp. 7455-7459, 2019.

[7] M. J. Gangeh, A. Ghodsi and M. S. Kamel, "Kernelized supervised dictionary learning," *IEEE Trans. Signal Process.*, vol. 61, pp. 4753-4767, 2013.

[8] C. -G. Li, Z. Lin, H. Zhang and J. Guo, "Learning semi-supervised representation towards a unified optimization framework for semi-supervised learning," in *Proc. IEEE Conf. Comp. Vision*, pp. 2767-2775, 2015.

[9] Y. Qin, H. Wu, X. Zhang and G. Feng, "Semi-Supervised Structured Subspace Learning for Multi-View Clustering," *IEEE Trans. Image Process.*, vol. 31, pp. 1-14, 2021.

[10] M. C. Tsakiris and R. Vidal, "Filtrated algebraic subspace clustering," *SIAM J. Imag. Sci.*, vol. 10, pp. 372-415, 2017.

[11] S. R. Rao, R. Tron, R. Vidal and Y. Ma, "Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 1-8, 2008.

[12] L. Liu, L. Kuang and Y. Ji, "Multimodal MRI brain tumor image segmentation using sparse subspace clustering algorithm," *Comput. Math. Methods Med.*, 2020.

[13] Y. Chen, A. Jalali, S. Sanghavi and H. Xu, "Clustering partially observed graphs via convex optimization," *J. Mach. Learn. Res.*, vol. 15, pp. 2213-2238, 2014.

[14] M. Rahmani and G. K. Atia, "Innovation pursuit: A new approach to subspace clustering," *IEEE Trans. Signal Process.*, vol. 65, pp. 6276-6291, 2017.

[15] M. C. Tsakiris and R. Vidal, "Algebraic clustering of affine subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, pp. 482-489, 2017.

[16] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. Int. Conf. Comp. Vision*, pp. 1615-1622, 2011.

[17] A. Taşkan, M. Muma and A. M. Zoubir, "Eigenvalue-Based Block Diagonal Representation and Application to p-Nearest Neighbor Graphs," in *Proc. 30th European Signal Process. Conf.*, 2022.

[18] C. You, C. -G. Li, D. P. Robinson and R. Vidal, "Oracle based active set algorithm for scalable elastic net subspace clustering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 3928-3937, 2016.

[19] N. García-Pedrajas, J. A. R. Del Castillo and G. Cerruela-García, "A proposal for local k values for k-nearest neighbor rule," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, pp. 470-475, 2015.

[20] S. S. Mullick, S. Datta and S. Das, "Adaptive learning-based k-nearest neighbor classifiers with resilience to class imbalance," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, pp. 5713-5725, 2018.

[21] A. Taştan, M. Muma and A. M. Zoubir, "Sparsity-aware Robust Community Detection," *Signal Process.*, vol. 187, pp. 108147, 2021.

[22] S. Arora, S. Rao and U. Varizani, "Expander flows, geometric embeddings and graph partitioning" *J. ACM*, vol. 56, pp. 1-37, 2009.

[23] B. Nasihatkon and R. Hartley, "Graph connectivity in sparse subspace clustering," in *Proc. CVPR 2011*, pp. 2137-2144, 2011.

[24] M. Ester, H.-P. Kriegel, J. Sander and X. Xiaowei, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*, pp. 226-231, 1996.

[25] W. Cheney, D. Kincaid, *Linear algebra: Theory and applications*, Jones&Bartlett, 2010.

[26] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, vol. 14, 2001.

[27] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. IEEE Int. Workshop Appl. Comput. Vision*, pp. 138-142, 1994.

[28] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, pp. 200-205, 1998.

[29] S. A. Nene, S. K. Nayar and H. Murase, "Columbia object image library (coil-20)," 1995.

[30] J. J Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, pp. 550-554, 1994.

[31] U. Von Luxburg, "A tutorial on spectral clustering,"*Stat. Comput.*, vol. 17, pp. 395-416, 2007.