SUBSPACE CLUSTERING WITH DENSE REPRESENTATIONS

Eva L. Dyer, Christoph Studer, Richard G. Baraniuk

Rice University; e-mail: {e.dyer, studer, richb@rice.edu}

ABSTRACT

Unions of subspaces have recently been shown to provide a powerful nonlinear signal model for collections of high-dimensional data, such as large collections of images or videos. In this paper, we introduce a novel data-driven algorithm for learning unions of subspaces directly from a collection of data; our approach is based upon forming minimum ℓ_2-norm (least-squares) representations of a signal with respect to other signals in the collection. The resulting representations are then used as feature vectors to cluster the data in accordance with each signal's subspace membership. We demonstrate that the proposed least-squares approach leads to improved classification performance when compared to state-of-the-art subspace clustering methods on both synthetic and real-world experiments. This study provides evidence that using least-squares methods to form data-driven representations of collections of data provide significant advantages over current methods that rely upon sparse representations.

Index Terms— Dense representations, subspace clustering, sparsity, pseudo-inverse, orthogonal matching pursuit.

1. INTRODUCTION

Every minute, terabytes of video and image data are uploaded to the internet. Video and image data are expected to account for an estimated 85% of all internet traffic by 2016.[1] To analyze, process, and eventually store such massive collections of high-dimensional data, novel methods that go beyond current compression schemes are essential [1]. One powerful approach used to tackle this problem, is to learn a model that captures the collection's low-dimensional geometric structure, rather than forming independent representations for each signal in the collection.

Despite the power of this type of geometric approach, in large collections of heterogenous data (e.g., images collected under different illumination conditions, viewpoints, etc.), finding a global model that compactly represents the relevant geometric structures across an entire dataset is extremely challenging. Thus, an alternative to learning a *global model* is to instead learn a *hybrid model* or a union of low-dimensional geometric structures (or subspaces) that characterize the structure present in the ensemble.

1.1. Unions of subspaces and subspace clustering

Recently, unions of subspaces have been shown to provide a compact geometric model for a wide range of datasets, and in particular, for collections of visual data acquired from the same scene or object but different viewing/lighting conditions. For example, collections of images of objects under different illumination conditions [2], motion trajectories of point-correspondences arising from different objects [3], and a wide range of structured sparse and block-sparse signals [4–7], can all be well-approximated by a union of low-dimensional subspaces or a union of affine hyperplanes [8].

Unions of subspaces provide a natural extension to single subspace models; however, providing an extension of subspace estimation techniques such as PCA to learn multiple subspaces is extremely challenging. This is due to the fact that *subspace clustering*—or clustering points in accordance with their subspace membership—and *subspace estimation* must be performed jointly. Nevertheless, if one can accurately determine which points lie within the same subspace, then subspace estimation becomes trivial.

1.2. Sparse subspace clustering

Recently, Elhamifar *et al.* [9] introduced a state-of-the-art algorithm for subspace clustering, known as *sparse subspace clustering* (SSC). This method is based upon forming sparse representations of points (or vectors) in the ensemble with respect to the remaining points in the ensemble (see Sec. 2 for the details). The motivation underlying this approach is that the sparse representation of a point under consideration will consist of other points in the same subspace.

After computing a sparse representation for each point in the ensemble, each sparse coefficient vector is placed into a row of a *subspace affinity matrix*. The subspace affinity matrix can be interpreted as a graph, where the (i, j) entry of the

The work of The work of E.L.D was supported by NSF GRFP 0940902 and a Texas Instruments Distinguished Graduate Fellowship. The work of C.S. was supported by the Swiss National Science Foundation (SNSF) under Grant PA00P2-134155. R.G.B. was supported in part by the Grants NSF CCF-0431150, CCF-0728867, CCF-0926127, DARPA/ONR N66001-08-1-2065, N66001-11-1-4090, N66001-11-C-4092, ONR N00014-08-1-1112, N00014-10-1-0989, AFOSR FA9550-09-1-0432, ARO MURIs W911NF-07-1-0185 and W911NF-09-1-0383, and by the Texas Instruments Leadership University Program.

[1]Taken from "Cisco Visual Networking Index: Forecast and Methodology, 2011-2016"

matrix represents the edge between the i^{th} and j^{th} point in the ensemble; the strength of each edge represents the likelihood that two points live in the same subspace. After forming a symmetric affinity matrix, spectral clustering is then performed on the graph Laplacian [10] of the affinity matrix to obtain labels (indicating the subspace membership) for all the points in the ensemble.

1.3. Contributions

While sparse representations result in affinity matrices that contain a small number of edges in the graph linking signals from *different* subspaces, recovering subspace clusters from the affinity matrices obtained via SSC is challenging due to the fact that sparse representations often produce weakly connected components between signals in the *same* subspace.

To circumvent this issue, we propose a novel method for subspace clustering that is based on forming minimum ℓ_2-norm (least-squares) representations[2] from the data instead of sparse representations. We show that the resulting representations tend to be dense (i.e., the energy is spread amongst all nonzero coefficients) and thus, such representations serve to produce affinity matrices that have more tightly connected components than those obtained via SSC. For this reason, spectral clustering algorithms operating on the affinity matrices obtained from least-squares representations can recover the subspace clusters more reliably than SSC.

Our specific contributions are as follows. First, we introduce a novel algorithm for forming data-driven representations that employs a thresholded pseudoinverse operator to obtain least-squares representations of points in an ensemble of data (Sec. 3). Following this, we study the performance of spectral clustering-based approaches to subspace learning for our proposed method, sparse subset selection (SSC) with OMP [11], and nearest neighbor (NN) subset selection from ensembles living on unions of subspaces (Sec. 4). We show that our proposed method outperforms both SSC and NN in the noisy setting and show that our method results in superior classification performance on real data consisting of images of faces under different illumination conditions.

2. BACKGROUND

In this section, we provide background on subspace clustering and introduce the SSC algorithm in [9].

2.1. Sparse subspace clustering

The goal of subspace clustering is to segment a collection of data in accordance with the subspace membership of every point in the dataset. Formally, if we have a collection of d vectors in $\mathbb{R}^n$, contained in the columns of the matrix $\mathbf{Y} \in \mathbb{R}^{n \times d}$, our goal is to determine the number p and dimension k of each of the subspaces present in the ensemble and then label each data point in accordance with the subspace it lives on. In other words, we wish to associate a label $\{1, \ldots, p\}$ for each vector $\mathbf{y}_i \in \mathbb{R}^n$, for $i = 1, \ldots, d$.

Recently, Elhamifar *et al.* [9] proposed a method known as *sparse subspace clustering* (SSC), which achieves state-of-the-art performance on a number of subspace clustering tasks, including motion segmentation and clustering images of faces under different illumination conditions. The authors demonstrate that by forming a sparse representation of a single point with respect to the remaining points in the ensemble, the resulting representation will consist of points that lie in the same subspace.

To be precise, SSC proceeds by first solving the following ℓ_1-norm minimization problem for each point in $\mathbf{Y}$,

$$\mathbf{c}_i^* = \underset{\mathbf{c} \in \mathbb{R}^d}{\operatorname{argmin}} \ \|\mathbf{c}\|_1 \quad \text{subject to } \mathbf{y}_i = \mathbf{Y}_{(i)}\mathbf{c}, \tag{1}$$

where $\mathbf{Y}_{(i)}$ is a matrix containing all vectors in $\mathbf{Y}$ except the i^{th} vector, y_i. After solving (1), each d-dimensional coefficient vector $\mathbf{c}_i^*$ is placed into the i^{th} row of an affinity matrix $\mathbf{C}$. Finally, spectral clustering [10] is performed on the graph Laplacian of $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}^T|$. Subsequent analysis provided conditions for which ℓ_1-norm minimization [12] and orthogonal matching pursuits (OMP) [11] are guaranteed to provide sets of points belonging to the same subspace.

3. DENSE SUBSPACE CLUSTERING

In this section, we propose an alternative to sparsity-based subspace clustering, which relies on forming least squares representations of the data rather than sparse ones.

3.1. Motivation

The motivation underlying SSC is that sparse representations of the data are likely to contain contributions from other points that live in the same subspace. When a representation only contains contributions from other points in the same subspace, then we say that *exact feature selection* (EFS) occurs. EFS is clearly important for providing guarantees for accurate subspace recovery. However, there are many settings where EFS occurs for a significant number of points in the ensemble but graph clustering methods still fail to segment the data appropriately. This is due to the fact that in many cases, affinity matrices formed via SSC produce weakly connected components in addition to outliers that are typically placed into their own cluster.

Instead of forming a sparse representation of the data as in SSC [9], we propose the use of least-squares representation of points to form a subspace affinity matrix for the ensemble. The main motivation behind using least-squares representations is that they are known to provide *dense* representations and thus, vectors that live close to one another in the same subspace will all use one another in their least-squares representations. In contrast, sparse representations by definition

[2] We replace the ℓ_1 norm in (1) with the ℓ_2 norm to obtain the least-squares representation of the data.

Algorithm 1 : Dense subspace clustering (DSC)

Input: Set of d vectors $\mathbf{Y} \in \mathbb{R}^{n \times d}$, number of subspace clusters p, and singular value threshold τ.

Output: A set of labels $\mathcal{L} = \{\ell_1, \ldots, \ell_d\}$ for each point in the dataset, where $\ell_i \in \{1, \ldots, p\}, \forall i$.

1: For each vector $\mathbf{y}_i \in \mathcal{Y}$, compute the dense representation $\bar{\mathbf{c}}_i$ according to $\bar{\mathbf{c}}_i = \mathbf{Y}_{(i)}^{-\tau}\mathbf{y}_i$, where $\mathbf{Y}_{(i)}^{-\tau}$ is the thresholded pseudoinverse of $\mathbf{Y}_{(i)}$ defined in (2).

2: Stack the dense representations $\bar{\mathbf{c}}_i$ into the columns of the transpose of the affinity matrix $\mathbf{C}^T$ and perform spectral clustering on the graph Laplacian of $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}^T|$ to produce a set of d labels $\mathcal{L}$.

would select only one of these points to be included in the representation. By effectively increasing the density of the representation within a particular subspace cluster, this serves to create a more tightly connected cluster in the subspace affinity matrix that is easier to segment with standard graph clustering methods. We show that forming least-squares representations of the data not only provides a higher density of edges within a cluster but also results in high rates of EFS (small number of edges between points in different subspaces). Thus, least-squares representations provide an appropriate balance between the density of the representation and produce sample sets that contain exact features; these two properties result in a more robust subset selection strategy for subspace clustering.

3.2. Dense subspace clustering algorithm

We now detail our proposed method, which we refer to as *dense subspace clustering* (DSC) and provide a summary in Alg. 1. For a set of d signals $\{\mathbf{y}_1, \ldots, \mathbf{y}_d\}$, each of dimension n, the minimum ℓ_2-norm representation of $\mathbf{y}_i$ is given by

$$\bar{\mathbf{c}}_i = \underset{\mathbf{c} \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{c}\|_2 \quad \text{subject to} \quad \mathbf{y}_i = \mathbf{Y}_{(i)}\mathbf{c}.$$

In stark contrast to minimum ℓ_1-norm representations computed using (SSC), which necessitate sophisticated sparse-signal recovery algorithms, the ℓ_2-norm representations can be calculated by standard least-squares methods.

In order to improve the conditioning of the pseudoinverse and consequently, the performance of our algorithm, we make use of the *thresholded pseudoinverse* instead of the full pseudoinverse of the data. We define the thresholded pseudoinverse of $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ as

$$\mathbf{A}^{-\tau} = \mathbf{V}\mathbf{S}^{-\tau}\mathbf{U}^T, \quad [\mathbf{S}^{-\tau}]_{i,i} = \begin{cases} [\mathbf{S}]_{i,i}^{-1}, & \text{if } [|\mathbf{S}|]_{i,i} \geq \tau \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

We compute dense representations for $\mathbf{y}_i$ according to $\bar{\mathbf{c}}_i = \mathbf{Y}_{(i)}^{-\tau}\mathbf{y}_i$ where $\tau > 0$ is a suitable thresholding parameter. In practice, we choose the thresholding parameter τ to preserve most of the energy in the spectrum (singular values in S).

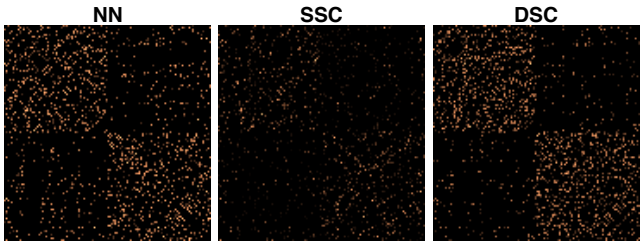

Fig. 1. Subset selection from overlapping subspaces in noise (60% overlap, SNR = 20dB); from left to right: NN, SSC, and the proposed DSC.

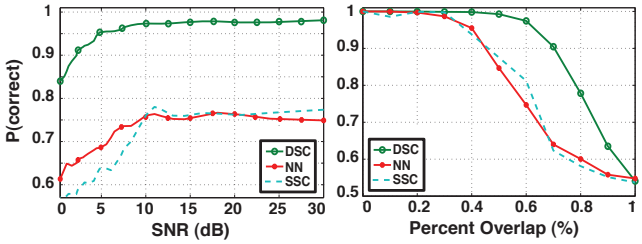

Fig. 2. Average classification rates from overlapping subspaces in noise. On the left, we fix the overlap to 0.6 and vary the SNR. On the right, we fix the SNR to 20dB and vary the SNR.

The algorithm then performs spectral clustering on the graph Laplacian of $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}^T|$ using normalized graph cuts [10], where the matrix $\mathbf{C}$ contains the coefficient vectors $\bar{\mathbf{c}}_i$ in its rows. Alternatively, we can also threshold $\mathbf{W}$ such that each point has only k nonzero elements per row and thus the resulting graph is of degree k. The resulting output of the graph cuts algorithm is a set of labels $\mathcal{L} = \{\ell_1, \ldots, \ell_d\}$ for each point in the dataset, where $\ell_i \in \{1, \ldots, p\}, \forall i$.

In addition to providing advantages in terms of the numerical stability of the algorithm, the thresholded pseudoinverse also provides a natural way to form least-squares representations of the data in noise. In particular, we can set the threshold parameter τ based upon the amount of noise in the data (or via cross-validation) in order to remove erroneous dimensions from the data that the noise alone occupies, i.e., dimensions where the subspaces do not occupy.

4. RESULTS

In this section, we demonstrate the efficacy of the proposed DSC method on both synthetic and real data. First, we study the probability of correct classification for varying levels of signal-to-noise ratio (SNR) and for different amounts of overlap between subspaces (dimension of intersection between subspaces). Following this, we showcase the classification performance of DSC for segmenting images of faces under different illumination conditions.

4.1. Generative model for synthetic data

To study the performance of DSC, we start with a synthetic example where the goal is to separate data living on a union of

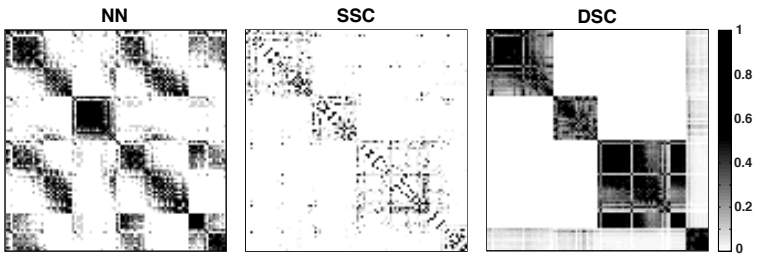

Fig. 3. Affinity matrices from face illumination subspaces; from left to right: NN, SSC, and DSC.

two subspaces. To generate a pair of k-dimensional subspaces with a q-dimensional intersection, we begin by drawing two disjoint index sets Ω_1 and Ω_2 of $k-q$ columns from an orthonormal basis (ONB) $\mathbf{D} \in \mathbb{R}^{n \times n}$, where $\Omega_1 \cap \Omega_2 = \emptyset$. Then, we select another set Ω_c of q distinct and orthogonal columns from $\mathbf{D}$ that both subspaces will share. We refer to the ratio of the intersecting dimensions to the subspace dimension q/k as the 'overlap' between the pair of subspaces.

Let $\mathbf{D}_1 = [\mathbf{D}_{\Omega_1}, \mathbf{D}_{\Omega_c}]$ denote the set of columns from $\mathbf{D}$ that we will use to synthesize points from the first subspace $\mathcal{W}_1$, where $\text{span}(\mathbf{D}_1) = \mathcal{W}_1$. Similarly, let $\mathbf{D}_2 = [\mathbf{D}_{\Omega_2}, \mathbf{D}_{\Omega_c}]$ denote the set of columns from $\mathbf{D}$ that we will use to synthesize points from the second subspace $\mathcal{W}_2$ with $\text{span}(\mathbf{D}_2) = \mathcal{W}_2$. After choosing the subspaces, we synthesize points from $\mathcal{W}_i$ as $\mathbf{Y}_i = \mathbf{D}_i \mathbf{A}_i$, where $\mathbf{A}_i \in \mathbb{R}^{n \times d_i}$ is a matrix of i.i.d. standard Normal coefficients used to mix the vectors in $\mathbf{D}_i$. Finally, we normalize all the points in each dataset $\mathbf{Y}_i$ to unit ℓ_2-norm and stack the resulting vectors into the matrix $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2]$, where $\mathbf{Y} \in \mathbb{R}^{n \times d}$ with $d = d_1 + d_2$.

4.2. Synthetic experiments

Here, we study the classification performance for three methods for computing subspace affinities: (1) nearest neighbor (NN) subset selection, (2) SSC using OMP, and (3) the proposed DSC. We use each of these three methods to form an affinity matrix $\mathbf{C}$ and then apply spectral clustering to the graph Laplacian of $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}^{\mathbf{T}}|$ for each method. In Fig. 1, we show the affinity matrices computed via NN, SSC, and the proposed DSC method, where the number of nonzero elements in each row of $\mathbf{C}$ are at most $k = 20$ and the dimension of the subspaces is ten. For both NN and SSC, two clusters are visible on the block diagonal, but DSC has more of its nonzero edges concentrated in their correct clusters than the NN solution. In contrast, the subspace clusters are barely visible in the affinity matrix obtained from SSC.

To compare the performance of these approaches, in Fig. 2, we show the percent of correctly labeled points (classification error) as we fix the overlap and vary the SNR (left) and fix the SNR and vary the percent overlap between the pair of subspaces (right) . In both experiments shown in Fig. 2, we observe a dramatic improvement in the classification performance of DSC over both NN and SSC, with SSC providing better classification performance in the noisefree setting. Both NN and SSC achieve at most 78% classification performance, while DSC approaches nearly perfect classification performance in the noisefree setting, with classification rates ranging from 84%–97%. In Fig. 2 (bottom), we show the classification performance as we vary the percentage of overlapping blocks between the two subspaces; this simulation also reveals that DSC outperforms NN and SSC by a significant margin.

4.3. Face illumination subspaces

Finally, we compare the performance of DSC with NN and SSC for unions of *illumination subspaces* arising from a collection of images of two different faces under various lighting conditions. By fixing the camera center and position of the persons face and capturing multiple images under varying lighting conditions, the resulting images are well-approximated by a 10-dimensional subspace [2].

In Fig. 3, we show the incidence matrices obtained via NN (left), SSC with OMP (middle), and DSC (right) for a collection of images of four different faces selected at random from the Yale Database B [13], where the number of points in each class equals $(30, 20, 40, 10)$ respectively. Let P_m denote the miss rate (probability that points are not included in the correct subspace cluster) and P_f denote the false alarm rate (probability that points are included in an incorrect subspace cluster). The classification errors for this experiment are NN: $P_m = 34.4\%$, $P_f = 20.9\%$; SSC: $P_m = 8.8\%$, $P_f = 1\%$; DSC: $P_m = 5.6\%$, $P_f = 0.3\%$. This result provides further empirical evidence that DSC yields improved classification performance over SSC and NN on real-world data.

5. CONCLUSIONS

In this paper, we proposed a novel subspace clustering algorithm, referred to as dense subspace clustering (DSC). This method relies on the formation of dense least-squares representations to populate the subspace affinity matrix for the data, rather than sparse representations used by state-of-the-art methods as in [9]. We demonstrated that DSC provides superior performance over sparse subspace clustering (SSC) and nearest neighbor (NN) methods for synthetic as well as real-world data. Future lines of effort include reducing the computational complexity of DCS and providing a theoretical analysis of DSC.

6. REFERENCES

[1] R.G. Baraniuk, "More is less: Signal processing and the data deluge," *Science*, vol. 331, no. 6018, pp. 717–719, 2011.

[2] R. Basri and D.W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 2, pp. 218–233, February 2003.

[3] K. Kanatani, "Motion segmentation by subspace separation and model selection," in *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, 2001.

[4] Y. M. Lu and M. N. Do, "Sampling signals from a union of subspaces," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 41–47, March 2008.

[5] T. Blumensath and M. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 4, pp. 1872–1882, 2009.

[6] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[7] Y.C. Eldar, P. Kuppinger, and H. Bolcskei, "Compressed sensing of block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Processing*, 2009.

[8] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, "Hybrid linear modeling via local best-fit flats," *arXiv:1010.3460v1 [cs.CV]*, October 2010.

[9] E. Elhamifar and R. Vidal, "Clustering disjoint subspaces via sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, March 2010, pp. 1926–1929.

[10] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 888–905, August 2000.

[11] E.L. Dyer, A.C. Sankaranarayanan, and R.G. Baraniuk, "Greedy feature selection for subspace clustering," *Preprint*, April 2012.

[12] M. Soltanolkotabi and E. J. Candès, "A geometric analysis of subspace clustering with outliers," *arXiv:1112.4258v2 [cs.IT]*, January 2012.

[13] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 6, pp. 643–660, 2001.