Probabilistic Subspace Clustering via Sparse Representations

Amir Adler, Michael Elad, and Yacov Hel-Or,

Abstract

We present a probabilistic subspace clustering approach that is capable of rapidly clustering very large signal collections. Each signal is represented by a sparse combination of basis elements (atoms), which form the columns of a dictionary matrix. The set of sparse representations is utilized to derive the co-occurrences matrix of atoms and signals, which is modeled as emerging from a mixture model. The components of the mixture model are obtained via a non-negative matrix factorization (NNMF) of the co-occurrences matrix, and the subspace of each signal is estimated according to a maximum-likelihood (ML) criterion. Performance evaluation demonstrate comparable clustering accuracies to state-of-the-art at a fraction of the computational load.

Index Terms

subspace clustering, sparse representation, dictionary, aspect model, non-negative matrix factorization.

I. Introduction

Subspace clustering is the unsupervised learning problem of clustering a collection of signals drawn from a union of subspaces, according to their spanning subspaces. Subspace clustering algorithms can be divided into four approaches: statistical, algebraic, iterative and spectral clustering-based; see [1] for a review. State-of-the-art approaches such as Sparse Subspace Clustering (SSC) [2], Low-Rank Representation (LRR) [3] and closed form solutions of LRR (LRR-CFS) [4] are spectral-clustering based and provide excellent performance for face clustering and video motion segmentation tasks. However, their complexity practically limits the data sets to be of moderate size.

In this paper[1] we address the problem of applying subspace clustering to data collections of up to millions of signals. This problem is important due to the following reasons: 1) Existing subspace clustering tasks are required

A. Adler and M. Elad are with the CS Department, Technion, Haifa, Israel. Y. Hel-Or is with the CS Department, Interdisciplinary Center, Herzlia, Israel. A. Adler is the recipient of the 2011 Google Europe Fellowship in Multimedia. This research was partly supported by this Google Fellowship and partly supported by HP Labs Innovation Research Program Award.

[1]The content of this paper was presented at an ICML 2012 workshop on Sparsity, Dictionaries and Projections in Machine Learning and Signal Processing.

to handle the ever-increasing amounts of data such as image and video streams. 2) New subspace clustering based solutions could be applied to applications that traditionally could not employ subspace clustering, and require the processing of large data sets. In the following we formulate the subspace clustering problem, explain state-of-the-art algorithms and highlight the main properties of our approach.

Problem Formulation. Let $Y \in \mathbb{R}^{N\times L}$ be a collection of L signals $\{\mathbf{y}_i \in \mathbb{R}^N\}_{i=1}^L$, drawn from a union of $K > 1$ linear subspaces. The bases of the subspaces are $\{B_k \in \mathbb{R}^{N\times d_k}\}_{k=1}^K$ and $\{d_k\}_{k=1}^K$ are their dimensions. The task of subspace clustering is to cluster the signals according to their subspaces. The number of subspaces K is either assumed known or estimated during the clustering process. The difficulty of the problem depends on the following parameters: 1) **Subspaces separation:** the subspaces may be independent[2], disjoint[3] or some of them may intersect, which is considered the most difficult case. 2) **Signal quality:** the collection of signals Y may be corrupted by noise, missing entries or outliers, thus, distorting the true subspaces structure. 3) **Model Accuracy:** the union-of-subspaces model is often only an approximation of a more complex and unknown data model, and the magnitude of the error it induces affects the overall performance.

LRR and SSC are similar algorithms that reveal the relations among signals by finding a self-expressive representation matrix $W \in \mathbb{R}^{L\times L}$, and obtain subspace clustering by applying spectral clustering to the graph induced by W. Both algorithms include two stages: 1) Find W such that $Y \simeq YW$, where $\text{diag}(W) = 0$ for the SSC algorithm. 2) Construct the affinity matrix $B = |W| + |W^T|$ and apply spectral clustering to the graph defined by B. SSC forces W to be sparse by minimizing its l_1 norm whereas LRR forces W to have low-rank by minimizing its *nuclear* norm. SSC outperforms RANSAC [5] and Agglomerative Lossy Compression [6] whereas LRR outperform Local Subspace Affinity [7] and Generalized-PCA [8]. LRR and SSC are restricted to moderate-sized data sets due to the polynomial complexities of their $L \times L$ affinity computation stage and spectral clustering stage (which is $O(L^3)$). LRR-CFS provides closed-form solutions for noisy data and reduces significantly the computational load of LRR. However, the complexity of the spectral clustering stage remains $O(L^3)$. The performance of LRR-CFS was reported in [4] as comparable to SSC and LRR.

In this paper we propose a new approach that is built on sparsely representing the given signals using a compact learned dictionary. This helps in exposing the relations among signals in such a way that leads to a much more efficient subspace-clustering method. The advantages of the proposed approach are as follows: 1) Linear complexity in the collection size L: each signal is represented by a dictionary with M atoms, where $M \ll L$, and the representation is computed by the OMP algorithm [9]. The complexity of solving the representation of all signals is $O(qNML)$, where $q \ll M$ is the average cardinality of the sparse representations. Subspace clustering is obtained by NNMF of the co-occurrences matrix of atoms and signals, a stage with complexity that depends linearly in L.

[2]subspaces are independent if the dimension of their union equals the sum of their dimensions.

[3]subspaces are disjoint if the intersection of each pair of subspaces contains only the null vector. Note that independent subspaces are disjoint, however, disjoint subspaces are not necessarily independent.

2) Immunity to noise: we employ the K-SVD [10] dictionary learning algorithm, which denoises the learned atoms, thus, improving clustering accuracy for noisy signals collections (note that LRR and SSC utilize in such cases the noisy signals as the dictionary).

Paper organization: Section II overviews sparse representations modeling. Section III presents the proposed approach and section IV evaluates its performance.

II. Sparse Representation Modeling of Signals

Sparse representations provide a natural model for signals that live in a union of low dimensional subspaces. This modeling assumes that a signal $\mathbf{y} \in \mathbb{R}^N$ can be described as $\mathbf{y} \simeq D\mathbf{c}$, where $D \in \mathbb{R}^{N \times M}$ is a *dictionary* matrix and $\mathbf{c} \in \mathbb{R}^M$ is sparse. Therefore, $\mathbf{y}$ is represented by a linear combination of *few* columns (atoms) of D. The recovery of $\mathbf{c}$ can be cast as the optimization problem:

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_0 \ \text{ s.t. } \ \|\mathbf{y} - D\mathbf{c}\|_2 \leq \epsilon, \tag{1}$$

for some approximation error threshold ϵ. The l_0 norm $\|\mathbf{c}\|_0$ counts the non-zeros components of $\mathbf{c}$, leading to a NP-hard problem. Therefore, a direct solution of (1) is infeasible. An approximate solution is given by applying the OMP algorithm, which successively approximates the sparsest solution. The recovery of $\mathbf{c}$ can be cast also by an alternative optimization problem that limits the cardinality of $\mathbf{c}$:

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{y} - D\mathbf{c}\|_2 \ \text{ s.t. } \|\mathbf{c}\|_0 \leq T_0, \tag{2}$$

where T_0 is the maximum cardinality. The dictionary D can be either predefined or learned from the given set of signals, see [11] for a review. For example, the K-SVD algorithm learns a dictionary by solving the following optimization problem: $\{D, C\} = \arg\min_{D,C} \|Y - DC\|_F^2$ s.t. $\forall i \, \|\mathbf{c}_i\|_0 \leq T_0$, where $Y \in \mathbb{R}^{N \times L}$ is the signals matrix, containing $\mathbf{y}_i$ in it's *i*-th column. $C \in \mathbb{R}^{M \times L}$ is the sparse representation matrix, containing the sparse representation vector $\mathbf{c}_i$ in it's *i*-th column. Once the dictionary is learned, each one of the signals $\{\mathbf{y}_i\}_{i=1}^L$ is represented by a linear combination of few atoms. Each combination of atoms defines a low dimensional subspace, thus, we will exploit the fact that signals spanned by the same subspace are represented by similar groups of atoms.

III. The Proposed Approach

We propose to interpret the set of sparse representation coefficients in C within a probabilistic framework: by leveraging the *aspect* model [12] to our problem, we associate with each occurrence of an atom $a \in \{a_1, ..., a_M\}$ in the representation of a signal $y \in \{y_1, ..., y_L\}$, a latent variable $s \in \{s_1, ..., s_K\}$ which represents the subspace. We further explain an observed pair (a, y) as follows: we first select a subspace with probability $P(s)$. We further select an atom with probability $P(a|s)$ and finally select a signal with probability $P(y|s)$. The joint probability $P(a_i, y_j, s_k)$ is given by:

$$P(a_i, y_j, s_k) = P(a_i, y_j|s_k)P(s_k) = P(a_i|s_k)P(y_j|s_k)P(s_k), \tag{3}$$

which is based on the assumption that a and y are conditionally independent **given** s (in accordance with the *aspect* model) leading to $p(a, y|s) = p(a|s)p(y|s)$. This assumption is justified for the case of independent subspaces, perfect dictionary, perfect sparse coding and general sampling of the signals within their subspaces (i.e. each signal is represented by all basis elements of its subspace and it is not embedded in a lower dimensional subspace within its subspace). In such cases the atoms that represent a signal are determined only by the subspace that spans the signal. From (3) we can obtain $P(a_i, y_j)$ by marginalization:

$$P(a_i, y_j) = \sum_{k=1}^{K} P(s_k)P(a_i|s_k)P(y_j|s_k). \tag{4}$$

The mixture model (4) can be cast also in matrix form:

$$V^{'} = W^{'}H^{'}, \tag{5}$$

where $V^{'} \in \mathbb{R}^{M\times L}$, $W^{'} \in \mathbb{R}^{M\times K}$ and $H^{'} \in \mathbb{R}^{K\times L}$ are non-negative such that $V^{'}_{ij} = P(a_i, y_j)$, $W^{'}_{ik} = P(s_k)P(a_i|s_k)$ and $H^{'}_{kj} = P(y_j|s_k)$. In the following we discuss how to obtain an estimate of $P(y_j|s_k)$ from the sparse representation coefficients and utilize it for subspace clustering.

NNMF decomposes a non-negative matrix V as the product of two non-negative matrices such that $V \approx WH$. The work of [13] proved that if V is a joint probability matrix that arises from the model (4) then a solution of NNMF that minimizes the KL-divergence [14] is equivalent to an Expectation-Maximization estimation of the mixture components of (4). Therefore, we propose to interpret the co-occurrences matrix of atoms and signals[4] $V = \frac{|C|}{\sum_{ij}|C_{ij}|}$ as emerging from the model (4), apply to it NNMF and recover the conditional probabilities from H. Subspace clustering is obtained by selecting the subspace that maximizes the ML criterion:

$$\hat{k}(y_j) = \arg\max_k \hat{P}(y_j|s_k) = \arg\max_k \bar{H}_{kj}, \tag{6}$$

where $\bar{H}$ equals to H after scaling its rows to unit sum. The proposed approach is summarized in Algorithm 1 and its complexity depends only linearly on L. This complexity is given by $O(qJNML)+O(qNML)+O(TMLK)+O(KL)$, where the first term is K-SVD complexity (with J iterations and assuming $L \gg 1$), the second term is OMP complexity, the third term is NNMF complexity (with T iterations) and the last term is the ML stage complexity.

IV. Performance Evaluation

A. Synthetic Data

Computation time and clustering accuracy of the proposed approach were compared to LRR, SSC and LRR-CFS (using the algorithm of Lemma 1). The experiments were conducted using a computer with Intel *i*7 Quad Core 2.2GHz and 8GB RAM. Experiment 1: MATLAB computation time comparison for clustering L signals in $\mathbb{R}^{128}$ is provided in Fig. 1, indicating linear complexity in L of the proposed approach compared to polynomial

[4]Our choice of V quantifies the contribution of each atom to each signal. The entry $|C|_{ij}$ is the analogy to the "number" of times the *i*-th atom appeared in the representation of the *j*-th signal. By further scaling $|C|$ to unit sum a probabilistic interpretation of this matrix is enabled, in accordance with [13].

Algorithm 1 Probabilistic Sparse Subspace Clustering

Input: signals $Y \in \mathbb{R}^{N \times L}$, # of clusters K, noise σ.

1. **Dictionary Learning:** Employ K-SVD to learn a dictionary $D \in \mathbb{R}^{N \times M}$ from Y.
2. **Sparse Coding:** Find sparse $C \in \mathbb{R}^{M \times L}$, such that $Y \simeq DC$.
3. **Co-occurrences Computation:** $V = \frac{|C|}{\sum_{ij} |C_{ij}|}$.
4. **Conditional Probabilities Estimation:**

   4.a. NNMF: $\min_{W,H} D_{KL}(V||WH)\ s.t.\ W, H \geq 0$.

   4.b. Set $\hat{P}(y_j|s_k) = \bar{H}_{kj}$, where $\bar{H}$ equals to H after scaling its rows to unit sum.
5. **Clustering:** $\hat{k}(y_j) = \arg\max_k \hat{P}(y_j|s_k), j = 1..L$.

Output: cluster labels for all signals $\hat{k}(y_j), j = 1..L$.

complexity of state-of-the-art. The reported durations include a dictionary $D \in \mathbb{R}^{128 \times 128}$ learning stage from the L signals if $L < 2^{16}$ or 2^{16} signals (randomly chosen from the collection) otherwise. Experiment 2: Clustering accuracy was evaluated for signals contaminated by zero mean white Gaussian noise, in the Signal-to-Noise (SNR) range of 5dB to 20dB. Per each experiment we generated a set of L=1000 signals in $\mathbb{R}^{128}$ drawn from a union of 10 subspaces, with equal number of signals per subspace. The bases of all subspaces were chosen as random combinations of the columns of a 128×256 over-complete discrete cosine transform matrix (see section VI.4. in [10]). The coefficients of each signal were randomly sampled from a Gaussian distribution of zero mean and unit variance. Clustering accuracies, averaged over 10 noise realizations, are presented in Fig. 2. The results demonstrate comparable performance of the proposed approach (M=128 learned atoms) to state-of-the-art. Experiment 3: Fig. 3 demonstrates that by increasing the data collection size (hence the dictionary training set), clustering performance improves, with best results for $L/M > 100$. Finally, Fig. 4 depicts an example of the conditional probability matrix $P(y_j|s_k)$ as obtained by NNMF, demonstrating peak probabilities at the same subspace for signals of the same cluster (the signals in the matrix Y were ordered w.l.o.g. according to their subspace association).

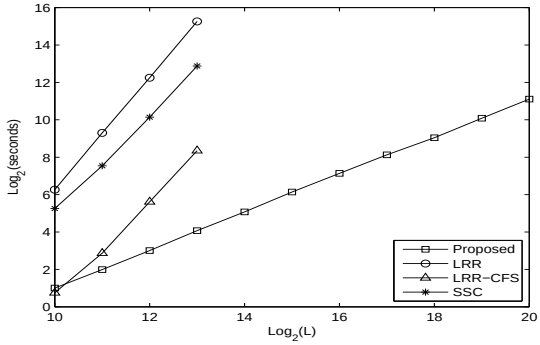

Fig. 1. Computation time vs. the number of signals L, for K=32 subspaces, signals' dimension N=128 and M=128 learned atoms.

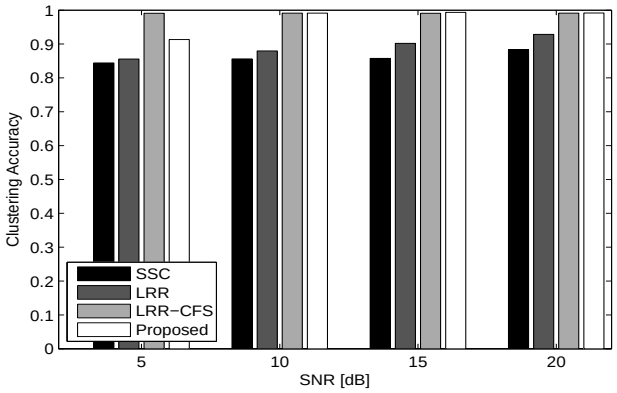

Fig. 2. Clustering accuracy for L=1000 signals in $\mathbb{R}^{128}$ drawn from 10 disjoint subspaces of dimension 10.

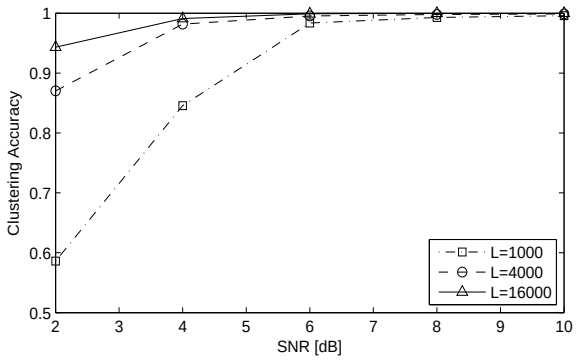

Fig. 3. Clustering accuracy vs. dictionary training set size L: performance improves as L increases (M=128 atoms).

B. Face Clustering

Face clustering is the problem of clustering a collection of facial images according to their human identity. Facial images taken from the same view-point and under varying illumination conditions are well approximated by a subspace of dimension < 10 [15]. Subspace clustering was applied successfully to this problem for example in [3, 16]. Face clustering accuracy of the proposed approach, K-subspaces [16] and state-of-the-art was evaluated using the Extended Yale B database [17], which contains 16128 images of 28 human subjects under 9 view-points and 64 illumination conditions. In our experiments we allocated 10 atoms per human subject (assuming subspaces dimensions < 10), and we found that at least a hundred facial images per subject are required for efficient dictionary training. Therefore, we generated from the complete collection a subset of 1280 images containing the first 10 human subjects, with 128 images per subject, by merging the 4th and 5th view-points which are of very similar view-points as demonstrated in Fig. 5. All images were cropped, resized to 48×42 pixels and column-stacked to vectors in $\mathbb{R}^{2016}$. Clustering accuracy was evaluated for $K = 2..8$ classes, by averaging clustering results over 40 different subsets of human subjects, for each value of K, by choosing 40 different combinations of human subjects out of

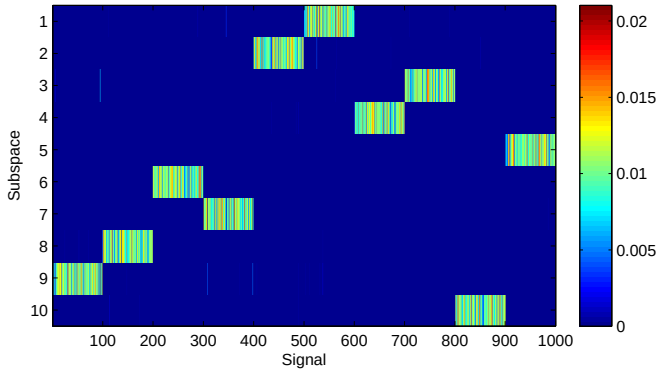

Fig. 4. Example of $\hat{P}(y_j|s_k)$ for L=1000 and K=10 disjoint subspaces (equal size clusters, SNR=10dB and M=128), as obtained by NNMF of V.

the 10 classes. Clustering results, provided in Table I, indicate comparable accuracies of the proposed approach to LRR and LRR-CFS and consistent advantage compared to K-Subspaces. The parameters of each method were optimized for best performance and summarized in Table II (for K-Subspaces d is the dimension of each subspace). In addition, for the proposed approach we employed OMP to approximate the solution of equation (2) and set T_0 to 9.

TABLE I

FACE CLUSTERING ACCURACY (%), AVERAGED OVER 40 DIFFERENT HUMAN SUBJECTS COMBINATIONS PER EACH NUMBER OF CLUSTERS (K).

K =	2	3	4	5	6	7	8
Proposed	92.19	82.07	**78.26**	**68.77**	**61.97**	**56.96**	50.58
LRR	**93.75**	**85.94**	65.47	57.02	51.34	52.86	**54.88**
LRR-CFS	83.17	64.78	57.84	52.08	50.98	51.64	52.56
K-Subspaces	67.60	59.17	50.24	51.34	48.81	48.32	45.35

Fig. 5. Facial images from the Extended Yale B collection: each row presents one of the first five classes in the database, the 8 leftmost columns are from the 4th view point and the 8 rightmost columns are from the 5th view point.

TABLE II

SUBSPACE CLUSTERING ALGORITHMS PARAMETERS SETTINGS.

K =	2	3	4	5	6	7	8
Proposed, $M =$	20	30	40	50	60	70	80
LRR, λ =	0.25	0.25	0.25	0.3	0.3	0.35	0.35
LRR-CFS, τ =	0.35	0.5	0.55	0.65	0.75	0.85	1
K-Subspaces, $d =$	8	8	8	8	8	8	8

V. CONCLUSIONS

This paper presented a probabilistic subspace clustering approach that utilizes a mixture model in conjunction with sparse representations. Performance evaluation with synthetic data and facial images demonstrate comparable accuracies to state-of-the-art, whereas the computation time depends only linearly on the size of the data collection. We further plan to explore the relation between the number of atoms to clustering accuracy, estimation methods for the number of clusters and algorithmic enhancements for data corrupted by missing entries and outliers.

REFERENCES

[1] R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2), 2011.

[2] E. Elhamifar and R. Vidal. Sparse subspace clustering. *CVPR*, 2009.

[3] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 2012.

[4] P. Favaro, R. Vidal, and A. Ravichandran. A closed form solution for robust subspace estimation and clustering. *CVPR*, 2011.

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartgraphy. *Commun. ACM*, 1981.

[6] S. Rao, R. Tron, Y. Ma, and R. Vidal. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. *CVPR*, 2008.

[7] J. Yan and M. Pollefeys. A general framework for motion segmentation: independent, articulated, rigid, non-rigid, degenerate and non-degenarate. *ECCV*, 2006.

[8] Y. Ma, A. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM Review*, 2008.

[9] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. *Asilomar Conf. Signals, Systems, and Computers*, 1993.

[10] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing*, 54(11), 2006.

[11] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proc. of the IEEE*, 98(6), 2010.

[12] T. Hoffman and J. Puzicha. Unsupervised learning from dyadic data. *ICSI Technical Report TR-98-042*, 1998.

[13] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. *SIGIR*, 2005.

[14] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *NIPS*, 2001.

[15] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. Patt. Anal. Mach. Intell.*, 25(2), 2003.

[16] J. Ho, M. Yang, J. Lim, K. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination condtions. *CVPR*, 2003.

[17] 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. Patt. Anal. Mach. Intell.*, 23(6), 2001.