Probabilistic Subspace Clustering via Sparse Representations

Amir Adler ADLERAM@CS.TECHNION.AC.IL
Computer Science Department, Technion, Haifa 32000, Israel

Michael Elad ELAD@CS.TECHNION.AC.IL
Computer Science Department, Technion, Haifa 32000, Israel

Yacov Hel-Or TOKY@GOOGLE.COM
Google Inc., CA , USA

Abstract

We present a probabilistic subspace clustering approach that is capable of rapidly clustering very large signal collections. The signals are modeled as drawn from a union of subspaces and each signal is represented by a sparse combination of basis elements (atoms), which form the columns of a learned dictionary. The set of sparse representations is utilized to derive the co-occurrence matrix of atoms and signals, which is modelled as emerging from a mixture model. The subspace of each signal is chosen as the one that maximizes the conditional probability of the signal given each subspace. This operation is obtained via the non-negative matrix factorization (NNMF) of the co-occurrence matrix, which exposes the conditional probability distribution of all signals. Performance evaluation demonstrate comparable clustering accuracies to state-of-the-art at a fraction of the computational load.

1. 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 (Vidal, 2011) for a review. State-of-the-art approaches such as Sparse Subspace Clustering (SSC) (Elhamifar & Vidal, 2009), Low-Rank Representation (LRR) (Liu et al., 2010) and closed form solutions of LRR (LRR-CFS) (Favaro et al., 2011) are spectral-clustering based and provide excellent performance for face clustering and video motion segmentation tasks. However, their complexity limits the size of the data sets to $\approx 10^4$ signals.

In this paper 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 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[1], disjoint[2] 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.

Submitted to ICML 2012 workshop on Sparsity, Dictionaries and Projections in Machine Learning and Signal Processing.

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

[2] subspaces are disjoint if their intersection contains only the null vector. Note that independent subspaces are disjoint, however, disjoint subspaces are not necessarily independent. Disjoint subspaces are considered more difficult to cluster than independent subspaces.

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 (Fischler & Bolles, 1981) and Agglomerative Lossy Compression (Rao et al., 2008) whereas LRR outperform SSC, Local Subspace Affinity (Yan & Pollefeys, 2006) and Generalized-PCA (Ma et al., 2008). 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 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 (Pati et al., 1993). 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-occurrence matrix of atoms and signals, a stage with complexity that depends linearly in L. 2) Immunity to noise: we employ the K-SVD (Aharon et al., 2006) dictionary learning algorithm, which denoises the 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.

2. 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 an 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 dictionary D can be either predefined or learned from the given set of signals, see (Rubinstein et al., 2010) for a review. For example, the K-SVD algorithm learns a dictionary by solving the following optimization problem:

$$\{D, C\} = \arg\min_{D,C} \|Y - DC\|_F^2 \text{ s.t. } \forall i \, \|\mathbf{c}_i\|_0 \leq T_0, \tag{2}$$

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, and T_0 is a maximal sparsity threshold. 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.

3. The Proposed Approach

3.1. From Sparse Representations to Mixture Models

We propose to interpret the set of sparse representation coefficients in C within a probabilistic framework: by leveraging the *aspect* model (Hoffman & Puzicha, 1998) 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 we select a signal with probability $P(y|s)$. The joint probability $P(a_i, y_j, s_k)$ is given by[3]:

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

from which we can obtain $P(a_i, y_j)$ by marginalization:

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

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

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

[3] We assume here that a and y are conditionally independent **given** s, which is in accordance with the *aspect* model, leading to $p(y, a|s) = p(y|s)p(a|s)$.

```
Algorithm 1 Probabilistic Sparse Subspace Clustering
  Input: signals $Y \in \mathbb{R}^{N \times L}$, # of clusters $K$, noise $\sigma$.
    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-occurrence Computation: $V = \frac{|C|}{\sum_{ij} |C_{ij}|}$.
    4. NNMF: $\{W, H\} = \arg\min_{W,H} D_{KL}(V||WH) \Rightarrow P(y_j|s_k) = \bar{H}_{kj}$, where $\bar{H}$ equals to $H$ after scaling its rows to unit sum.
    5. Clustering: $\hat{k}(y_j) = \arg\max_k P(y_j|s_k), j = 1..L$.
  Output: cluster labels for all signals $\hat{k}(y_j), j = 1..L$.
```

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 recover $P(y_j|s_k)$ from the sparse representation coefficients and utilize it for subspace clustering.

3.2. Subspace Clustering via NNMF

NNMF decomposes a non-negative matrix V as the product of two non-negative matrices such that $V \approx WH$. The work of (Gaussier & Goutte, 2005) 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 $D_{KL}(V||WH)$ is equivalent to an Expectation-Maximization solution for the mixture components of (4). Therefore, we propose to treat the co-occurrence matrix of atoms and signals $V = \frac{|C|}{\sum_{ij} |C_{ij}|}$ as emerging from the model (4), apply to it NNMF and recover the conditional probabilities from H. Subspace clustering is obtained by maximizing the conditional probability per signal:

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

where $\bar{H}$ equals to H after scaling its rows to unit sum. Algorithm 1 summarizes the proposed approach.

4. Performance Evaluation

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 $i7$ 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 complexity of state-of-the-art. The reported durations include a dictionary $D^{128\times128}$ learning stage from the L signals if $L < 2^{16}$ or 2^{16} signals 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 (non-overlapping for disjoint subspaces and overlapping for intersecting subspaces) of the columns of a 128×256 over-complete discrete cosine transform matrix (Aharon et al., 2006). 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).

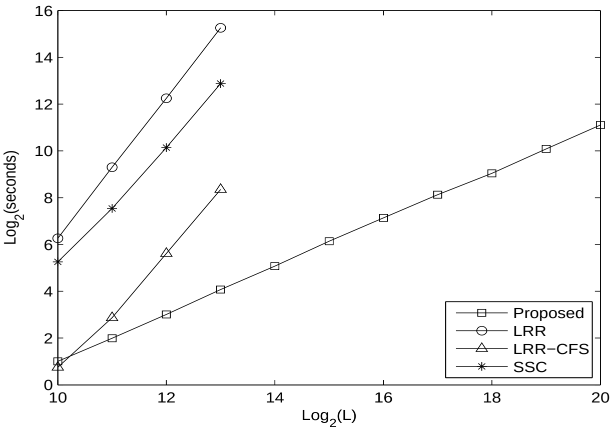
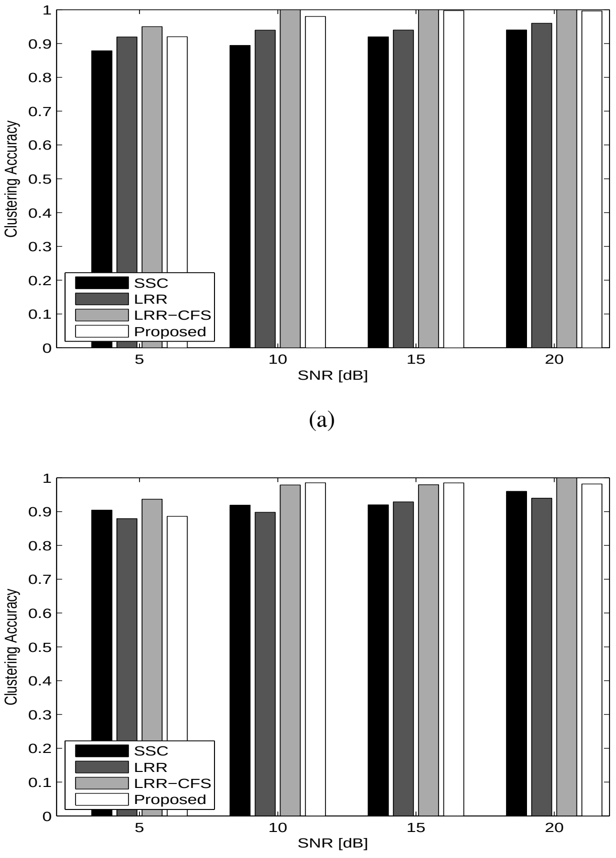


Figure 1. Computation time vs. the number of signals L, for K=32 subspaces, signals' dimension N=128 and M=128 atoms.

5. Conclusions

This paper presented a probabilistic subspace clustering approach that utilizes a mixture model in conjunction with sparse representation. Performance evaluation demonstrate comparable performance to state-of-the-art at a fraction of the computational load. We further plan to explore the relation between the number of atoms to clustering accuracy, estimation methods for the number of clusters and applications to data corrupted by missing entries and outliers.



(a)

(b)

Figure 2. Clustering accuracy for L=1000 signals in $\mathbb{R}^{128}$ drawn from 10 subspaces with dimension 10: (a) disjoint subspaces. (b) intersecting subspaces with 2 overlapping basis vectors.

References

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

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

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

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

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

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

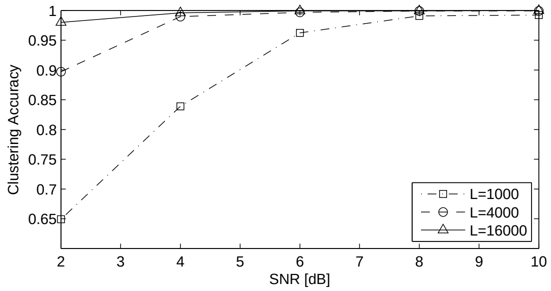


Figure 3. Clustering accuracy vs. dictionary training set size L: performance improves as L increases (M=128 atoms).

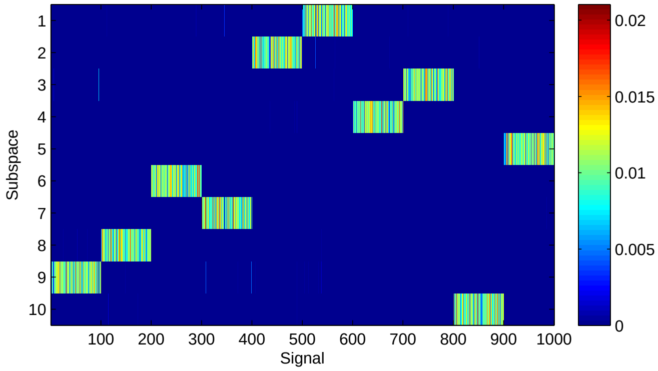


Figure 4. An example of $P(y_j|s_k)$ for L=1000 and K=10 disjoint subspaces (equal size clusters, SNR=10dB and M=128), as obtained from the NNMF of $V = \frac{|C|}{\sum_{ij}|C_{ij}|}$.

Liu, G., Lin, Z., and Yu, Y. Robust subspace segmentation by low-rank representation. *ICML*, 2010.

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

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

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

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

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

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