Provable Self-Representation Based Outlier Detection in a Union of Subspaces

Chong You Daniel P. Robinson René Vidal
Johns Hopkins University, Baltimore, MD, 21218, USA

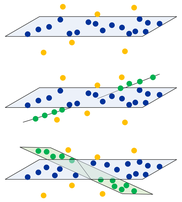
IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR 2017

Introduction

- Most collected visual data today are unlabeled/weakly labeled
- High-dimensional data often lie approx. in low-dimensional subspaces
- Subspace clustering is the problem of clustering data into subspaces
- This work addresses sensitivity of subspace clustering to **outliers**

Prior Work & Challenges

- RobustPCA methods (e.g. REAPER,OutlierPursuit) require inliers drawn from a single subspace
- Low-rank methods (e.g., LRR) require inlier subspaces to be independent
- Other methods (TSC, CoP, ℓ_1-thresholding) require dense inliers and/or incoherent outliers



Challenges: multiple subspaces, unknown number of subspaces and their dimensions, sparsity of inliers, close-by outliers, etc.

Contributions

- Outlier detection by using *self-representation* and *random walk*

✓ *Self-representation* allows our method to handle multiple subspaces, and the number of subspaces and their dimensions are not required

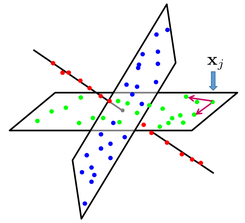
✓ *Random walk* allows our method to explore contextual information, hence our method can handle sparsity of inliers and close-by outliers

✓ Our method is provably correct in identifying outliers

This work was supported by the grant NSF-IIS 1447822.

Background: Sparse Subspace Clustering

- Task: cluster data $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ in a union of subspaces
- Observation: each data $\mathbf{x}_j$ can be expressed as a linear combination of a few others from its own subspace, i.e. $\mathbf{x}_j = X\mathbf{c}_j, c_{jj} = 0$
- Algorithm: find such representation by solving sparse optimization $\min_{\mathbf{c}_j} \|\mathbf{c}_j\|_1$ s.t. $\mathbf{x}_j = X\mathbf{c}_j, c_{jj} = 0$, define affinity between any two points $\mathbf{x}_i$, $\mathbf{x}_j$ to be $|[\mathbf{c}_j]_i|$, then apply spectral clustering
- This work: we extend the method to deal with outliers in data X



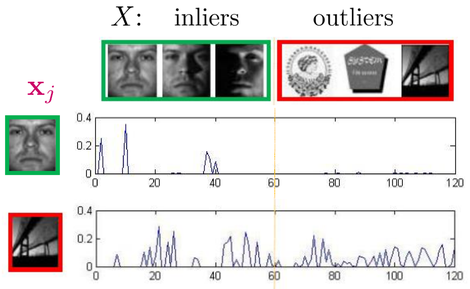
A Novel Outlier Detection Algorithm

Input: unlabeled dataset $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ that contains both inliers and outliers.

- Step 1. Compute data self-representation:

$$\min_{\mathbf{c}_j} \lambda\|\mathbf{c}_j\|_1 + \frac{1-\lambda}{2}\|\mathbf{c}_j\|_2^2 \text{ s.t. } \mathbf{x}_j = X\mathbf{c}_j, c_{jj} = 0$$

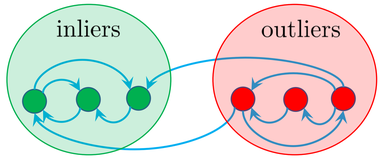
  - If $\mathbf{x}_j$ is inlier: $[\mathbf{c}_j]_i \neq 0 \rightarrow \mathbf{x}_i$ is inlier
  - If $\mathbf{x}_j$ is outlier: $[\mathbf{c}_j]_i \neq 0 \rightarrow \mathbf{x}_i$ can be either inlier or outlier



- Step 2. Define a random walk from $\{\mathbf{c}_j\}$:

$$[P]_{ij} = |[\mathbf{c}_i]_j| / \|\mathbf{c}_i\|_1$$

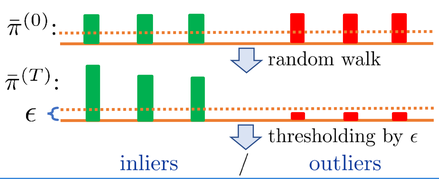
  - There is no transition from inlier to outlier
  - Any random walker will end up in inliers



- Step 3. Compute stationary distribution:

$$\bar{\pi}^{(T)} = \frac{1}{T}\sum_{t=1}^{T} \pi^{(0)} P^t$$

where $\pi^{(0)} = [1/N, \cdots, 1/N]$ is uniform



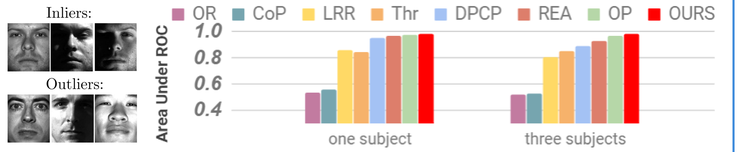
Analysis

Theorem: our outlier detection method correctly identifies all outliers as $T \rightarrow \infty$ and $\epsilon = 0$ if the following conditions are satisfied:

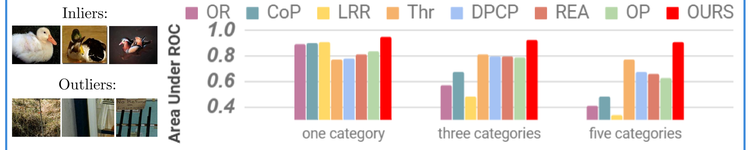
- Geometric condition: for every subspace $\mathcal{S}_\ell$, $\mu_\ell + \frac{1-\lambda}{\lambda} < r_\ell$
  μ_ℓ: coherence of points in subspace $\mathcal{S}_\ell$ and all other inliers & outliers
  r_ℓ: distribution of points in $\mathcal{S}_\ell$ measured by inradius
- Connectivity conditions: subgraph of each inlier subspace is strongly connected; no subgraph of outliers is closed

Experiments

- Experiments on Extended Yale B face database
  - Inliers: face of one/three subjects. Outliers: face of others



- Experiments on Caltech-256 object database
  - Inliers: images of one/multiple categories. Outliers: outlier category



- Our method achieves the best performance. Our method can effectively handle multiple inlier groups while many others cannot

[1] E. Elhamifar and R. Vidal., Sparse Subspace Clustering, In *CVPR*, 2009.
[2] H. X and C. Caramanis and S. Sanghavi, Robust PCA via Outlier Pursuit, In *NIPS*, 2010.
[3] M. Soltanolkotabi and E.J. Candes., A Geometric Analysis of Subspace Clustering with Outlier, In *Annals of Statistics*, 2013.
[4] C. You et al., Oracle Based Active Set Algorithm for Scalable Elastic Net Subspace Clustering, In *CVPR*, 2016.