Oracle Based Active Set Algorithm for Scalable Elastic Net Subspace Clustering

Chong You[†] Chun-Guang Li[*] Daniel P. Robinson[‡] René Vidal[†]

[†]Center for Imaging Science, Johns Hopkins University, Baltimore, USA
[*]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China
[‡]Department of Applied Mathematics and Statistics Johns Hopkins University, Baltimore, USA

Center for Imaging Science, Vision Lab @ JHU, http://www.vision.jhu.edu

IEEE 2016 Conference on Computer Vision and Pattern Recognition

CVPR 2016

Introduction

- Vision datasets often contain multiple classes, each lying in a low-dimensional subspace
- **Subspace clustering**: cluster data that lie in a union of subspaces

Prior Work

- Data from a union of subspaces is self-expressive, i.e., $\mathbf{x}_j = X\mathbf{c}_j$, $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$
  - find self-expression with regularization $f(\cdot)$: $\min_{\mathbf{c}_j} f(\mathbf{c}_j)$ s.t. $\mathbf{x}_j = X\mathbf{c}_j, c_{jj} = 0$
  - apply spectral clustering to data affinity $|c_{ij}| + |c_{ji}|$ to get the clusters

Sparse subspace clustering: $f(\cdot) = \|\cdot\|_1$

- — Sparse coefficient & few connections
- ✓ Guaranteed to give only correct connections under broad conditions
- ✗ Each cluster is not well connected
- ✗ Not scalable: difficult to solve

Least squares regression: $f(\cdot) = \|\cdot\|_2^2$

- — Dense coefficient & many connections
- ✗ There may exist many wrong connections in general cases
- ✓ Each cluster is well connected
- ✗ Not scalable: requires large memory

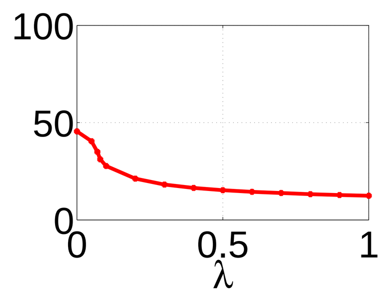
Contributions

- Elastic net Subspace Clustering (EnSC)

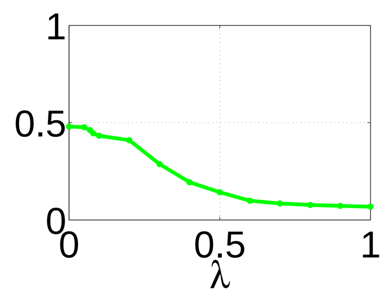
$$\min_{\mathbf{c}_j} \lambda\|\mathbf{c}_j\|_1 + \frac{1-\lambda}{2}\|\mathbf{c}_j\|_2^2 \quad \text{s.t.} \quad \mathbf{x}_j = X\mathbf{c}_j, \quad c_{jj} = 0 \tag{1}$$

- $\lambda \in [0, 1]$ controls a trade-off between correct connection and connectivity

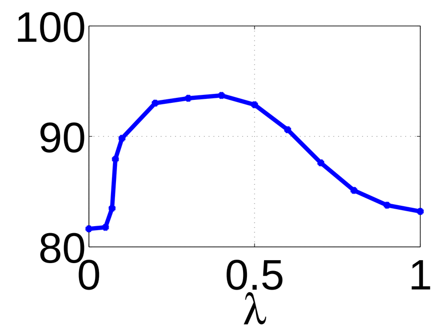
- ✓ Guaranteed correct connections.
- ✓ Improved connectivity by choosing λ
- ✓ We propose a new scalable algorithm



Connection error



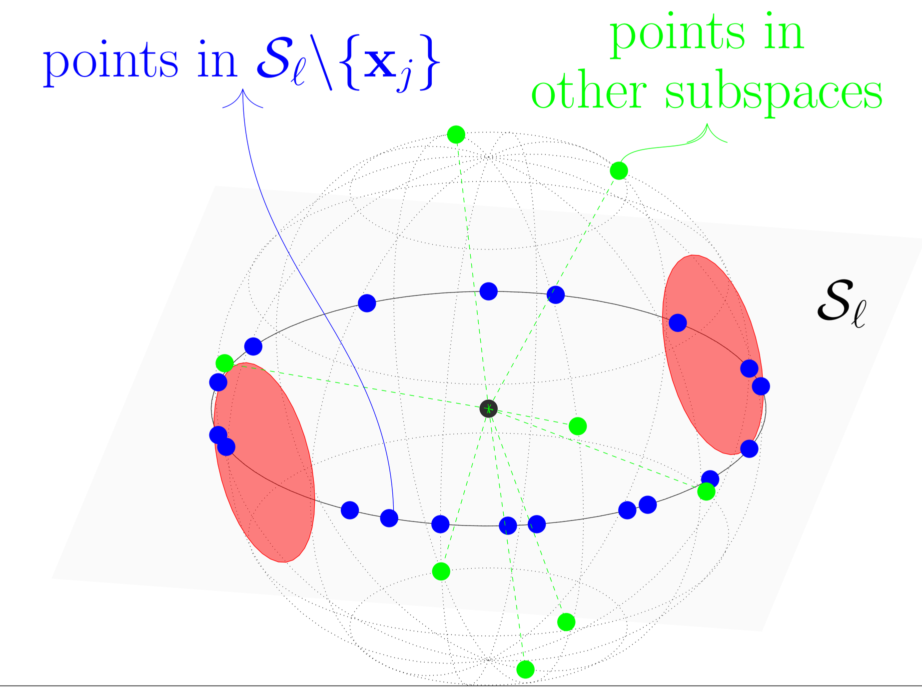
Connectivity



Clustering accuracy

Geometry of EnSC: Correct Connections vs. Connectivity

- Consider problem (1), assume $\mathbf{x}_j \in \mathcal{S}_\ell$ and that all data have unit ℓ_2 norm
- Lemma [Geometry of solution]: If no points in other subspaces lie in oracle region, then $c_{ij} \neq 0$ if and only if $\mathbf{x}_i$ lies in the oracle region
  - Oracle region is calculated by solving oracle problem using points in $\mathcal{S}_\ell$
- Lemma [Size of oracle region, informal]: An upper bound on the size of the oracle region *decreases* as the trade-off parameter λ *increases*
- Conclusion: λ provides a correct connection/connectivity tradeoff
  - Larger $\lambda \implies$ smaller oracle region $\implies$ easier to get only connect connections
  - Smaller $\lambda \implies$ larger oracle region $\implies$ more points in $\mathcal{S}_\ell$ are connected



- Conditions for guaranteed correct connection
  - δ_j: oracle point, lies in $\mathcal{S}_\ell$ and is the center of the oracle region
  - $\mu(\delta_j, X^{-\ell})$ / $\mu(\delta_j, X^\ell_{-j})$: coherence (max absolute inner product) between δ_j and points in other subspaces / in $\mathcal{S}_\ell \backslash \{\mathbf{x}_j\}$
  - Role of λ: condition is easier to be satisfied for larger λ

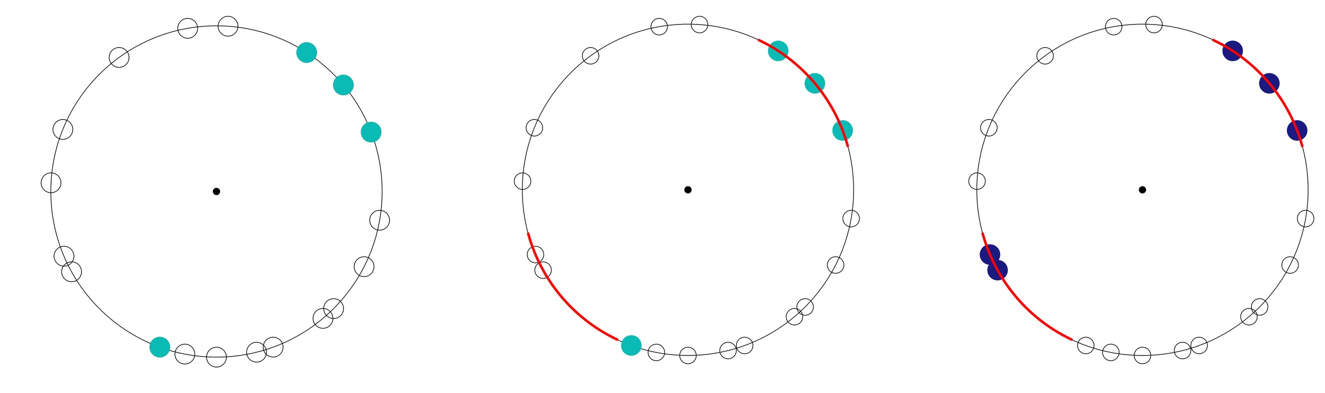
Theorem

Suppose that $\mathbf{x}_j \in \mathcal{S}_\ell$. Then, the solution $\mathbf{c}_j$ to (1) gives correct connections if

$$\mu(\delta_j, X^{-\ell}) < \frac{\mu(\delta_j, X^\ell_{-j})^2}{\mu(\delta_j, X^\ell_{-j}) + \frac{1-\lambda}{\lambda}}.$$

ORacle Guided Elastic Net (ORGEN) Algorithm

- Observation: if the support set T of the solution $\mathbf{c}_j$ to (1) is known, then problem (1) can be reduced to a small scale problem
- Algorithm: solve a sequence of small scale problems on small support sets T_k; the support sets are chosen such that T_{k+1} include points in the oracle region computed from T_k
- Convergence: T_k converges to T in finite number of iterations



Support set T_k Oracle region on T_k Support set T_{k+1}

Experiments

- Our EnSC achieves the best clustering accuracy
- EnSC with our ORGEN algorithm is efficient
  - Traditional SSC with ADMM algorithm cannot handle MNIST and CovType databases
  - Our method is mostly as efficient as the kNN method (TSC) and the greedy method (OMP)

		Clustering accuracy (%)					Running time (min.)				
	N	TSC	OMP	SSC	LRSC	EnSC	TSC	OMP	SSC	LRSC	EnSC
Coil-100	7,200	61.32	42.93	57.10	55.76	69.24	2	3	127	3	3
PIE	11,554	22.15	24.06	41.94	46.65	52.98	3	5	412	12	13
MNIST	70,000	85.00	93.07	-	-	93.79	30	6	-	-	28
CovType	581,012	35.45	48.76	-	-	53.52	999	783	-	-	1452

[1] E. Elhamifar and R. Vidal., Sparse Subspace Clustering, In *IEEE Conf. in Computer Vision and Pattern Recognition*, 2009.
[2] M. Soltanolkotabi and E.J. Candes., A Geometric Analysis of Subspace Clustering with Outlier, In *Annals of Statistics*, 2013.
[3] C. Lu et al., Robust and Efficient Subspace Segmentation via Least Squares Regression, In *European Conference on Computer Vision*, 2012.

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