Scalable Subspace Clustering

Chong You

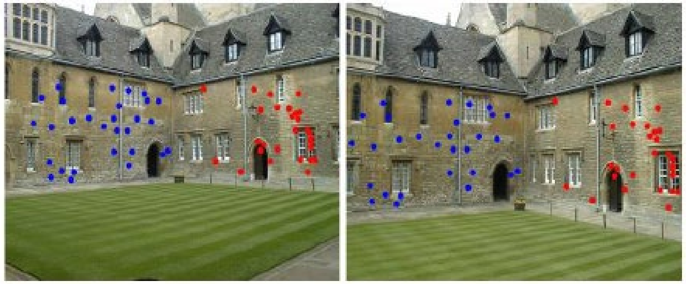
Johns Hopkins University

Joint work with Chun-guang Li, Daniel P. Robinson, and René Vidal

Motivation

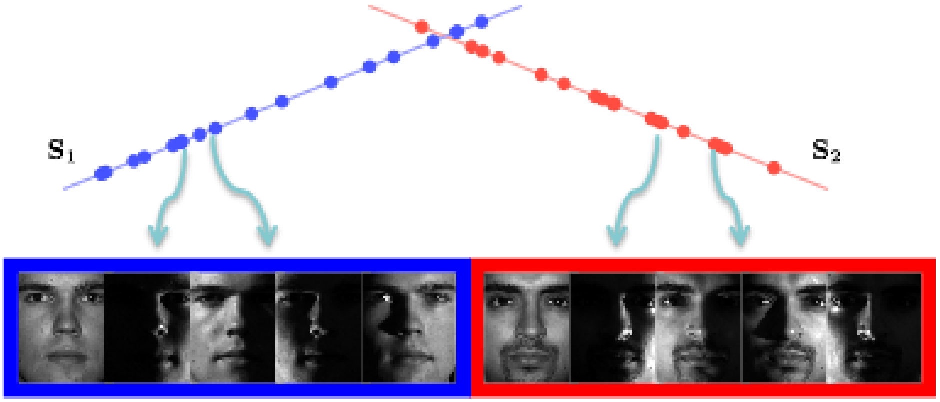
In many areas, we deal with large amount of data

- Data contains multiple classes
- Each class lies in a low-dimensional subspaces



Planar Segmentation

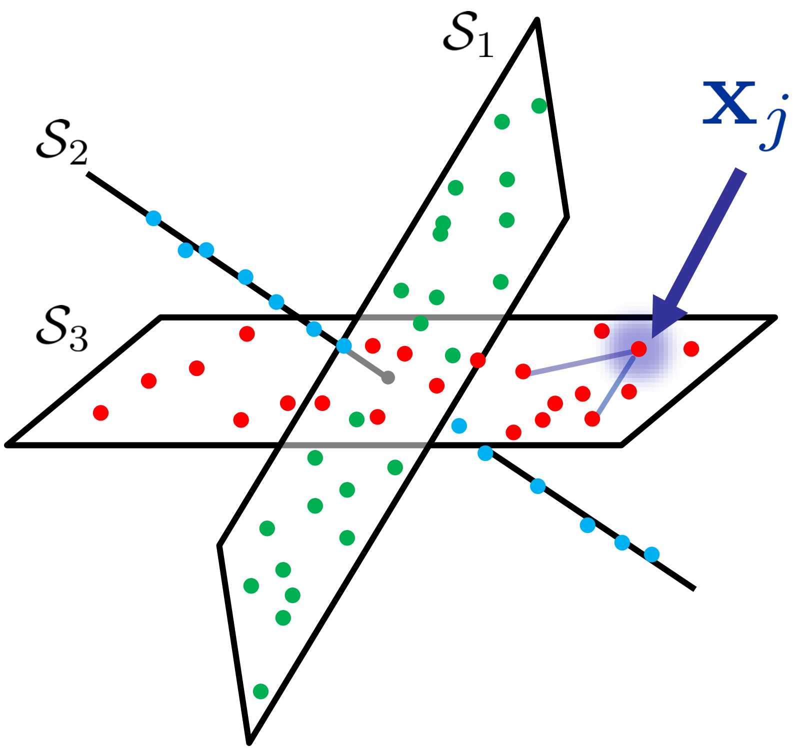
Motion Segmentation



Face Recognition/Clustering

Subspace clustering

Given data $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, find a union of subspaces that fits the data:



Two-step Approach

- Build data affinity
- Apply spectral clustering

Challenges

- Distance based affinity fails at the intersection of subspaces

Self-Expressive Model

- Compute affinity by data self-representation

[1] Elhamifar-Vidal, Sparse Subspace Clustering, CVPR 2009

Prior work: sparse subspace clustering (SSC)

Sparsity — Self-representation

$$\min_{\mathbf{c}_j} \|\mathbf{c}_j\|_0 \quad \text{s.t.} \quad \mathbf{x}_j = X\mathbf{c}_j, c_{jj} = 0$$

Convex relaxation

Method:

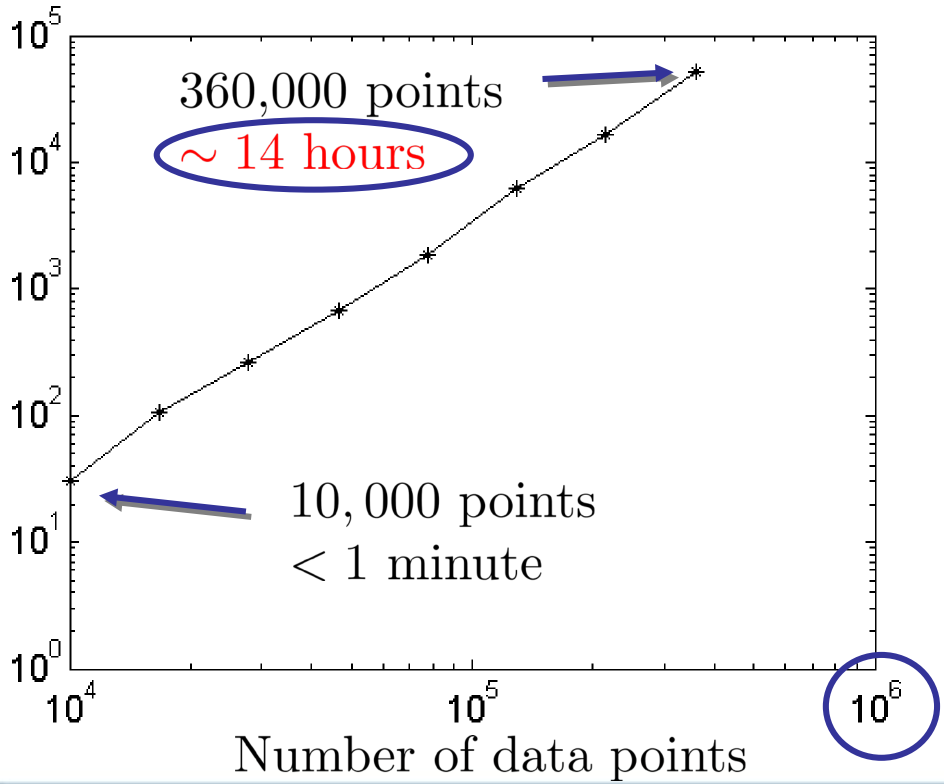
SSC by basis pursuit (SSC-BP)

Properties:

✓ Guaranteed correct connections

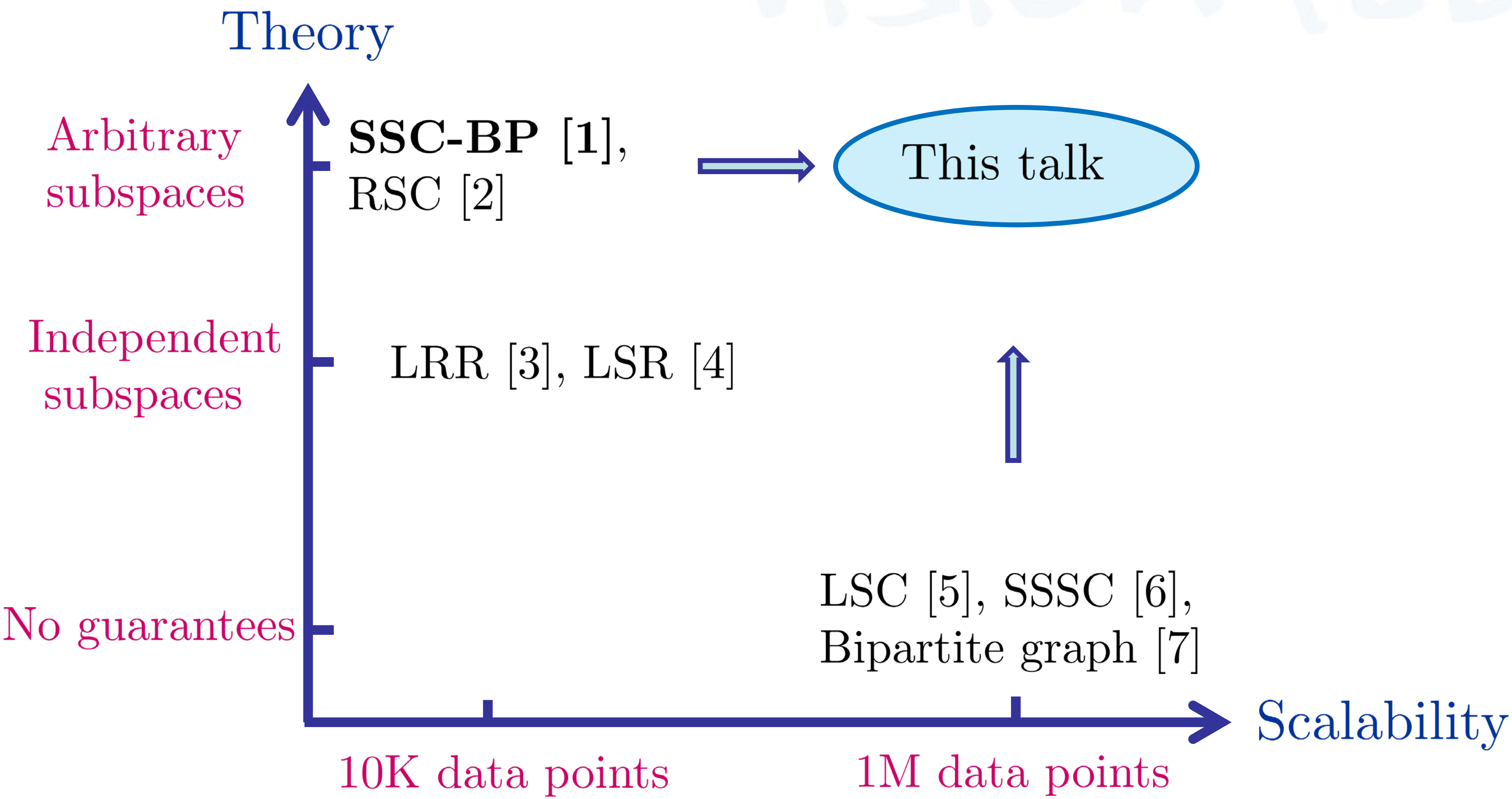
✗ Not scalable:
solved by CVX/ADMM
tested on ≤ 640 points

Running time (sec.)



[1] Elhamifar-Vidal, Sparse Subspace Clustering, CVPR 2009

Prior work: overview



[1] E. Elhamifar and R. Vidal, Sparse Subspace Clustering, CVPR'09
[2] M. Soltanolkotabi and E. Candes, Robust Subspace Clustering, Annual of Statistics'13
[3] G. Liu, Z. Lin, Y. Yu, Robust Subspace Segmentation by Low-Rank Representation, ICML'10
[4] Lu et al,, Robust and efficient subspace segmentation via least squares regression, ECCV 2012.
[5] X. Chen and D. Cai, Large Scale Spectral Clustering with Landmark-based Representation, AAAI'11
[6] X. Peng, L. Zhang, Z. Yi, Scalable Sparse Subspace Clustering, CVPR'13
[7] A. Adler, M. Elad, Y. Hel-Or, Linear-Time Subspace Clustering via Bipartite Graph Modeling

Scalable Sparse Subspace Clustering by Orthogonal Matching Pursuit

Chong You†, Daniel P. Robinson‡, René Vidal†

†Center for Imaging Science, Johns Hopkins University
‡Applied Mathematics and Statistics, Johns Hopkins University

Sparse subspace clustering

Sparsity — Self-representation

$$\min_{\mathbf{c}_j} \|\mathbf{c}_j\|_0 \quad \text{s.t.} \quad \mathbf{x}_j = X\mathbf{c}_j, c_{jj} = 0$$

Convex relaxation

Method:

SSC by *basis pursuit*
(SSC-BP)

Properties:

- ✓ Guaranteed correct connections
- ✗ Not scalable:
  solved by CVX/ADMM
  tested on ≤ 640 points

[1] Elhamifar-Vidal, Sparse Subspace Clustering, CVPR 2009
[2] Dyer et al, Greedy Feature Selection for Subspace Clustering, JMLR 2014

SSC by Orthogonal Matching Pursuit

Sparsity — Self-representation

$$\min_{\mathbf{c}_j} \|\mathbf{c}_j\|_0 \quad \text{s.t.} \quad \mathbf{x}_j = X\mathbf{c}_j, c_{jj} = 0$$

Convex relaxation — Greedy pursuit

Method:

SSC by *basis pursuit* (SSC-BP)

Properties:

✔ Guaranteed correct connections

✘ Not scalable:
solved by CVX/ADMM
tested on ≤ 640 points

Method:

SSC by *orthogonal matching pursuit* (SSC-OMP)

Properties:

? Guaranteed correct connections

? Scalable

[1] Elhamifar-Vidal, Sparse Subspace Clustering, CVPR 2009
[2] Dyer et al, Greedy Feature Selection for Subspace Clustering, JMLR 2014

SSC by Orthogonal Matching Pursuit

Sparsity | Self-representation

$$\min_{\mathbf{c}_j} \|\mathbf{c}_j\|_0 \quad \text{s.t.} \quad \mathbf{x}_j = X\mathbf{c}_j, c_{jj} = 0$$

Convex relaxation | Greedy pursuit

Method:
SSC by *basis pursuit* (SSC-BP)

Properties:

- ✔ Guaranteed correct connections
- ✘ Not scalable:
  solved by CVX/ADMM
  tested on ≤ 640 points

Method:
SSC by *orthogonal matching pursuit* (SSC-OMP)

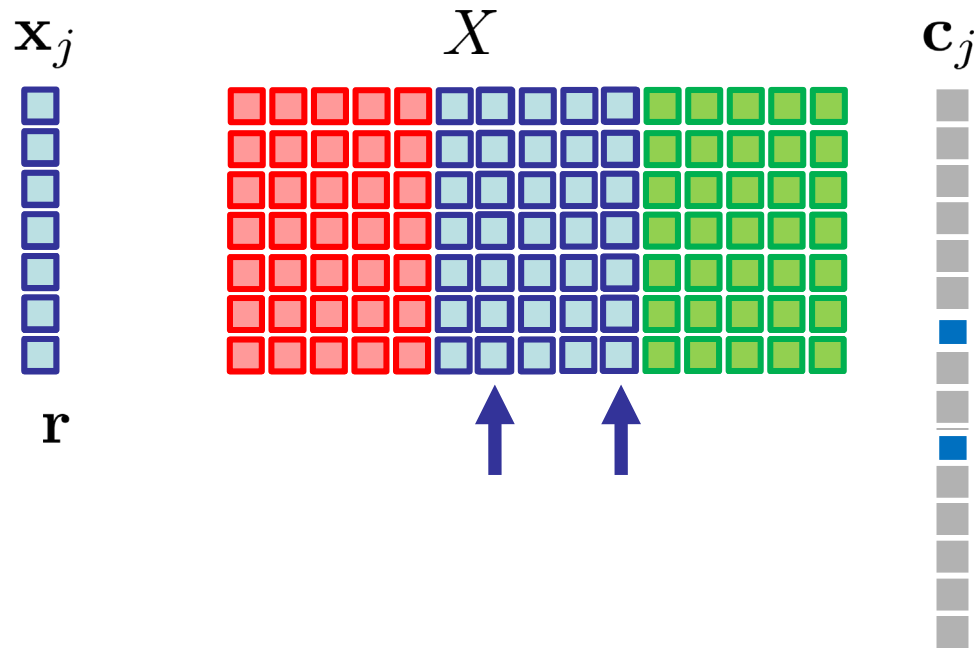
Contributions:

- ✔ Guaranteed correct connections
- ✔ Scalable:
  tested on 1,000,000 points

[1] Elhamifar-Vidal, Sparse Subspace Clustering, CVPR 2009
[2] Dyer et al, Greedy Feature Selection for Subspace Clustering, JMLR 2014

SSC by Orthogonal Matching Pursuit

Find representation $\mathbf{x}_j = X\mathbf{c}_j$ by greedy selection



What are the conditions for giving correct connections?

Each iteration picks a point from the same subspace

Guaranteed correct connections: deterministic model

Theorem

Suppose that $\mathbf{x}_j \in \mathcal{S}_\ell$. Then, $\mathbf{c}_j$ gives correct connections if

$$\mu(W_j^\ell, X^{-\ell}) < r^\ell,$$

where μ captures the similarity between $\mathcal{S}_\ell$ and all other subspaces, and r captures distribution of points in $\mathcal{S}_\ell$.

For SSC-BP[3]:

W_j^ℓ = dual points

For SSC-OMP:

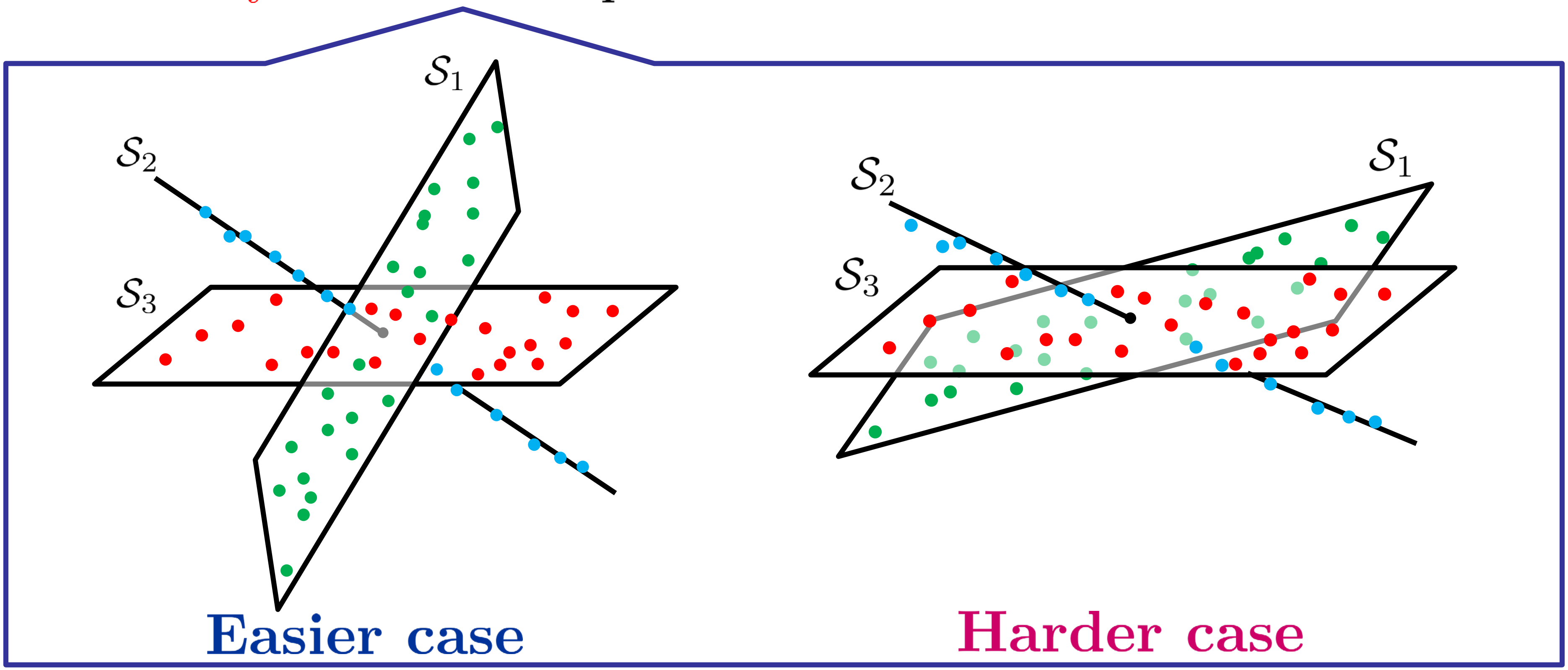
W_j^ℓ = residual points

[3] Soltanolkotabi-Candes, A geometric analysis of subspace clustering with outliers

Guaranteed correct connections: deterministic model

$$\mu(W_j^\ell, X^{-\ell}) < r^\ell$$

Similarity between subspaces

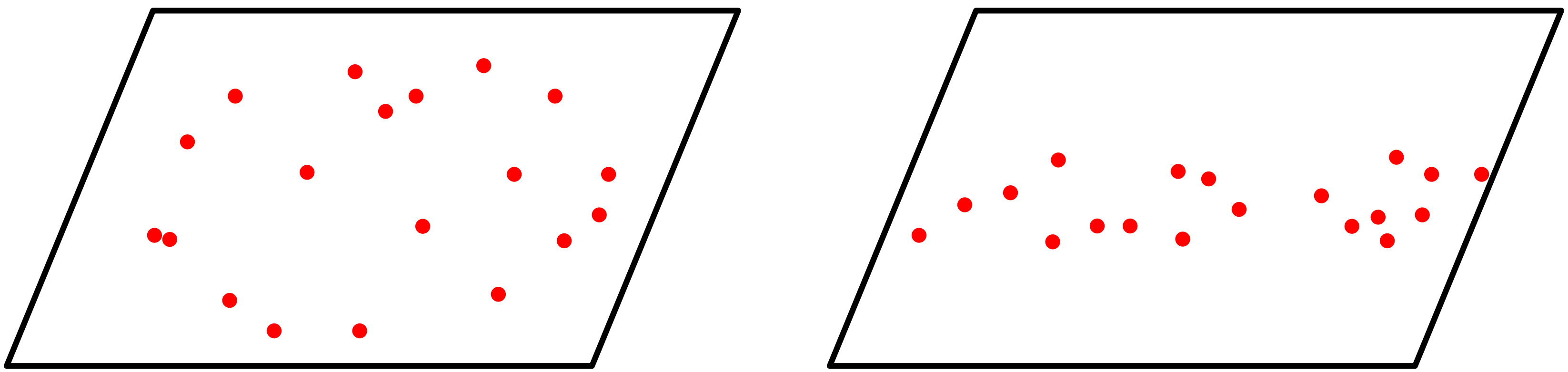


Guaranteed correct connections: deterministic model

$$\mu(W_j^\ell, X^{-\ell}) < r^\ell$$

Similarity between subspaces

Distribution of points



Easier case **Harder case**

Is this condition likely to be satisfied?

Guaranteed correct connections: random model

Random model:

- Draw n subspaces of dimension d in ambient dimension D
- Draw $\rho d + 1$ points from each subspace

Theorem

Under the random model, the solution $\{\mathbf{c}_j\}_{j=1}^N$ gives correct connections with overwhelming probability if

$$\frac{d}{D} < \frac{c^2(\rho)\log\rho}{12\log N}$$

For SSC-BP[3]:

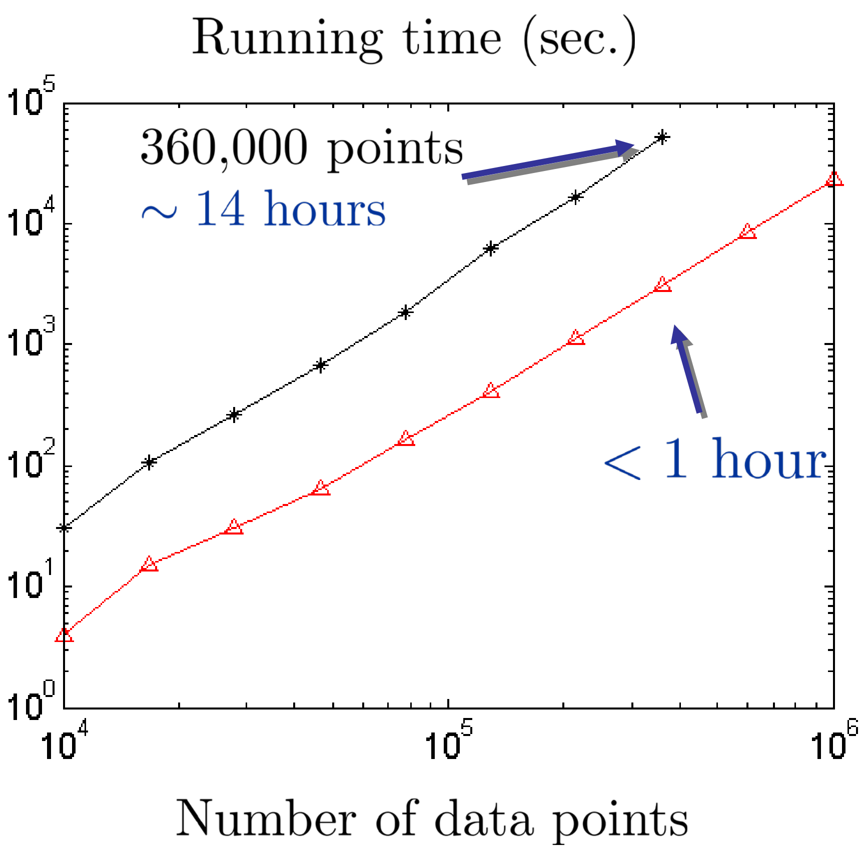
$$p > 1 - \frac{2}{N} - Ne^{-\sqrt{\rho}d}$$

For SSC-OMP:

$$p > 1 - \frac{2d}{N} - Ne^{-\sqrt{\rho}d}$$

[3] Soltanolkotabi-Candes, A geometric analysis of subspace clustering with outliers

Scalability: SSC-BP versus SSC-OMP

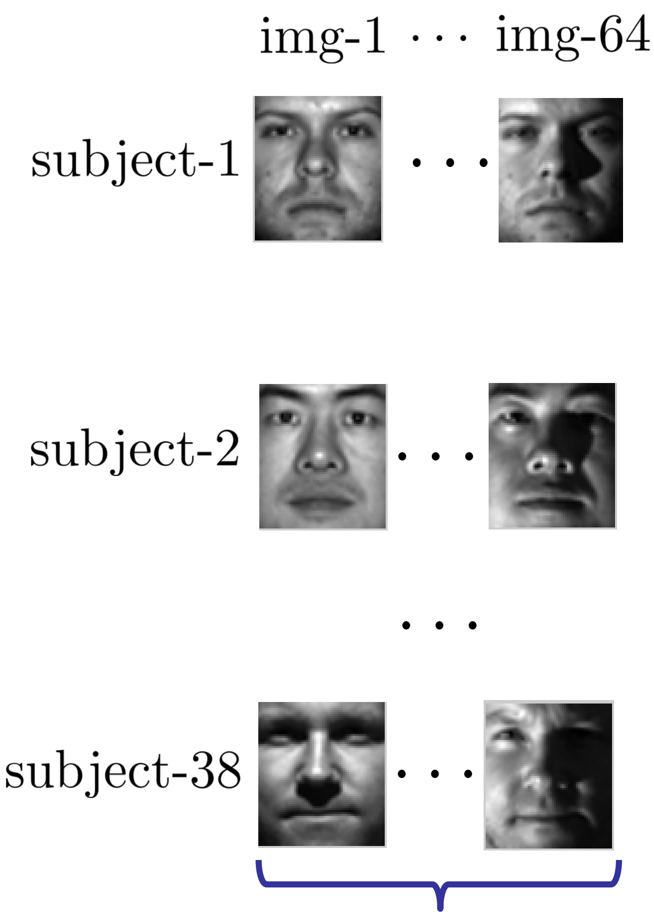


- SSC-BP (Baseline)

- SSC-OMP

SSC-OMP significantly reduces the time, and deals with 1 million data

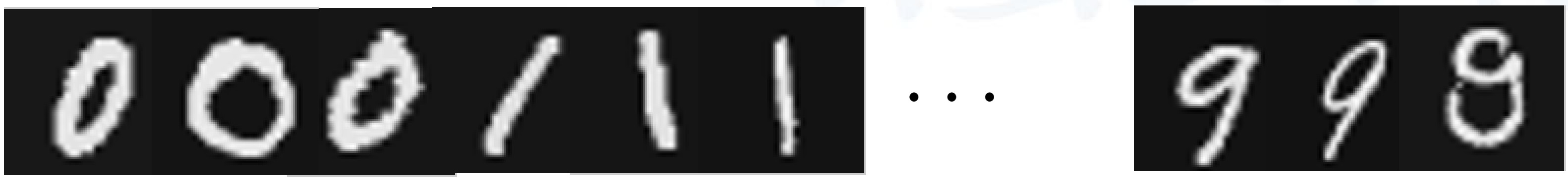
Experiment on extended Yale B



No. subjects	2	10	20	30	38
a%: average clustering accuracy					
SSC-OMP	99.21	88.43	**81.71**	**79.27**	**80.45**
SSC-BP	**99.45**	**91.85**	79.80	76.10	68.97
LSR	96.77	62.89	67.17	67.79	63.96
LRSC	94.32	66.98	66.34	67.49	66.78
SCC	78.91	NA	NA	14.15	12.80
t(sec.): running time					
SSC-OMP	0.3	1.7	4.7	9.4	**14.5**
SSC-BP	49.1	228.2	554.6	1240	1851
LSR	**0.1**	**0.8**	**3.1**	**8.3**	15.9
LRSC	1.1	1.9	6.3	14.8	26.5
SCC	50.0	NA	NA	520.3	750.7

$>$ 100 times faster

Experiment on MNIST



No. points	500	2,000	6,000	20,000	60,000
a%: average clustering accuracy					
SSC-OMP	**85.17**	**88.99**	**90.56**	**94.21**	**94.68**
SSC-BP	83.01	85.58	85.60	-	-
LSR	75.84	78.09	79.91	-	-
LRSC	75.02	79.44	79.88	-	-
SCC	53.45	66.43	70.60	-	-
t(sec.): running time					
SSC-OMP	**1.3**	**11.7**	**71.7**	**427**	**3219**
SSC-BP	20.1	635.2	13605	-	-
LSR	1.7	42.4	327.6	-	-
LRSC	1.9	43.0	312.9	-	-
SCC	31.2	101.3	366.8	-	-

Center for IMAGING SCIENCE

Conclusion

SSC by Orthogonal Matching Pursuit (OMP):

- ✓ theoretical guarantees for correct connections
- ✓ performance validation on large databases

JHU vision lab

Scalable Elastic Net Subspace Clustering

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

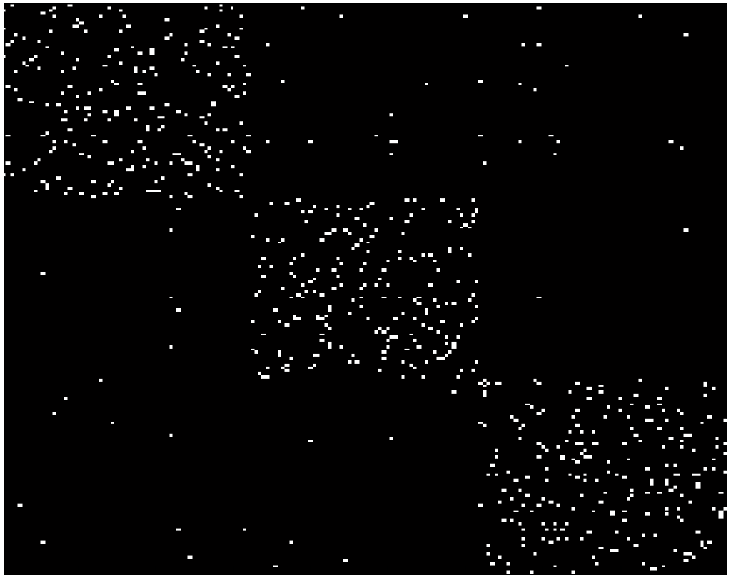
[†]Center for Imaging Science, Johns Hopkins University
[*]SICE, Beijing University of Posts and Telecommunications
[‡]Applied Mathematics and Statistics, Johns Hopkins University

Center for IMAGING SCIENCE

Motivation

SSC

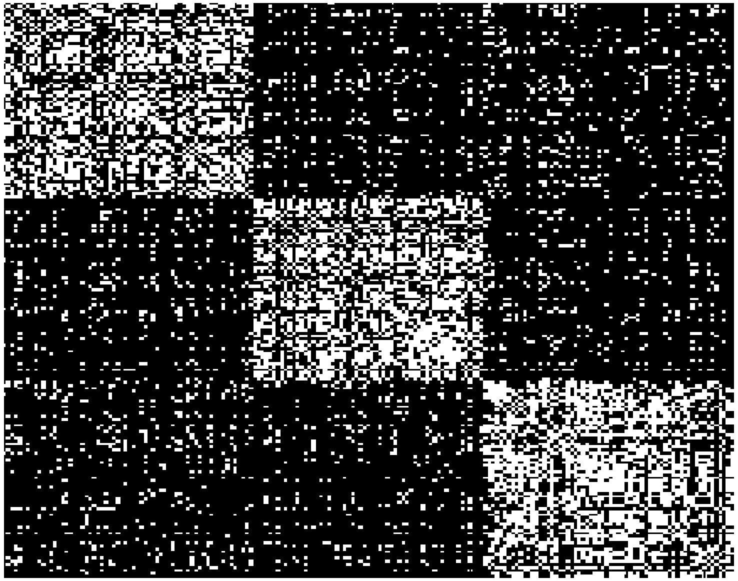
$$\min_{\mathbf{c}} \|\mathbf{c}\|_1 + \frac{\gamma}{2}\|\mathbf{x} - X\mathbf{c}\|_2^2$$



- ✓ Few wrong connections
- ✗ Not well connected

LSR

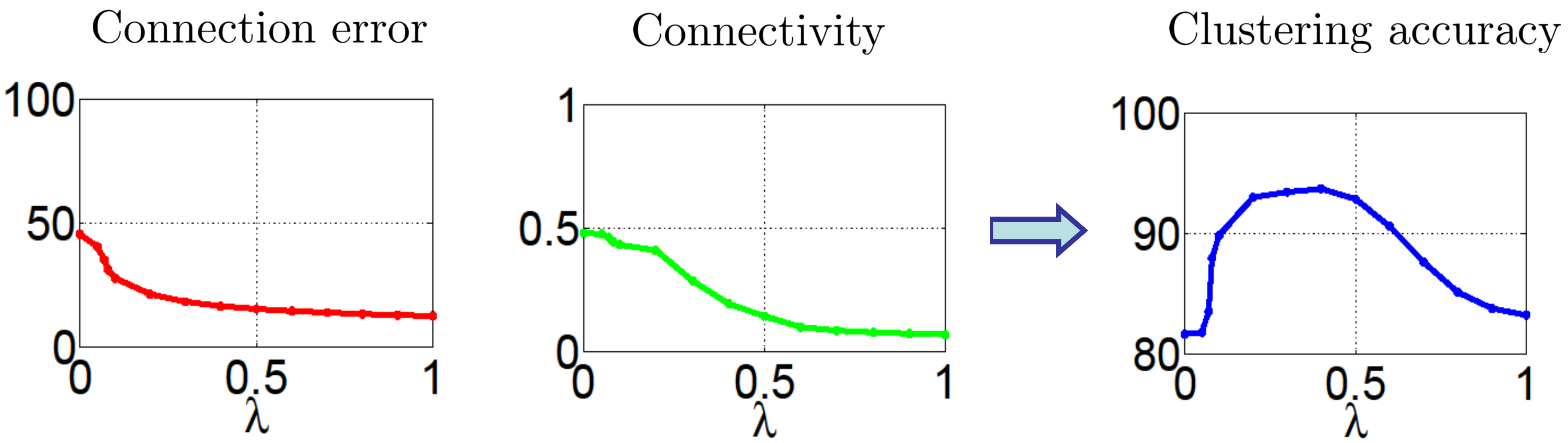
$$\min_{\mathbf{c}} \|\mathbf{c}\|_2^2 + \frac{\gamma}{2}\|\mathbf{x} - X\mathbf{c}\|_2^2$$



- ✗ Many wrong connections
- ✓ Well-connected

Elastic net Subspace Clustering (EnSC)

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



Key theoretical challenges:

- Is EnSC guaranteed to give correct connections
- How to explain the tradeoff with connectivity

Scalable Elastic net Subspace Clustering

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

- Prior methods
  - ADMM
  - Interior point
  - Solution path
  - Proximal gradient method
  - etc.

Scalability issue:
- Too many iterations to converge
- Access to full data matrix

Key challenge:

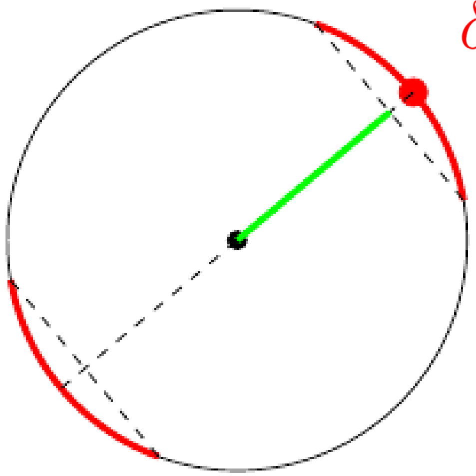
? Can we derive scalable algorithms that can handle 1 million data

Geometry of solution

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

Oracle point $\delta_j = \gamma(\mathbf{x}_j - X\mathbf{c}_j^*)$

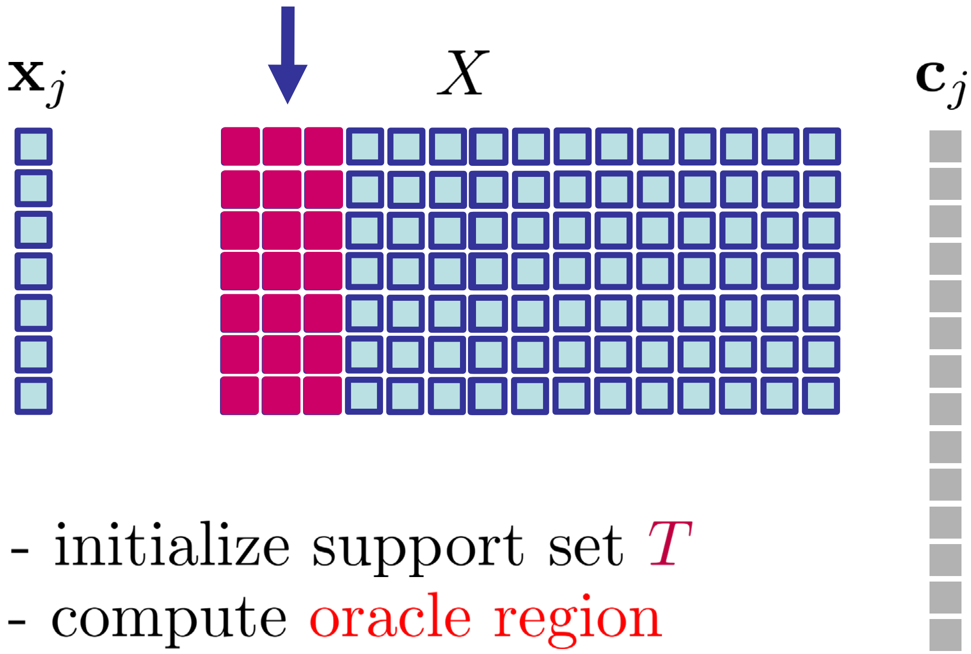
- If we know the solution $\mathbf{c}_j^*$, we can compute δ_j
- If we know δ_j, we can find the support of the solution $\mathbf{c}_j^*$



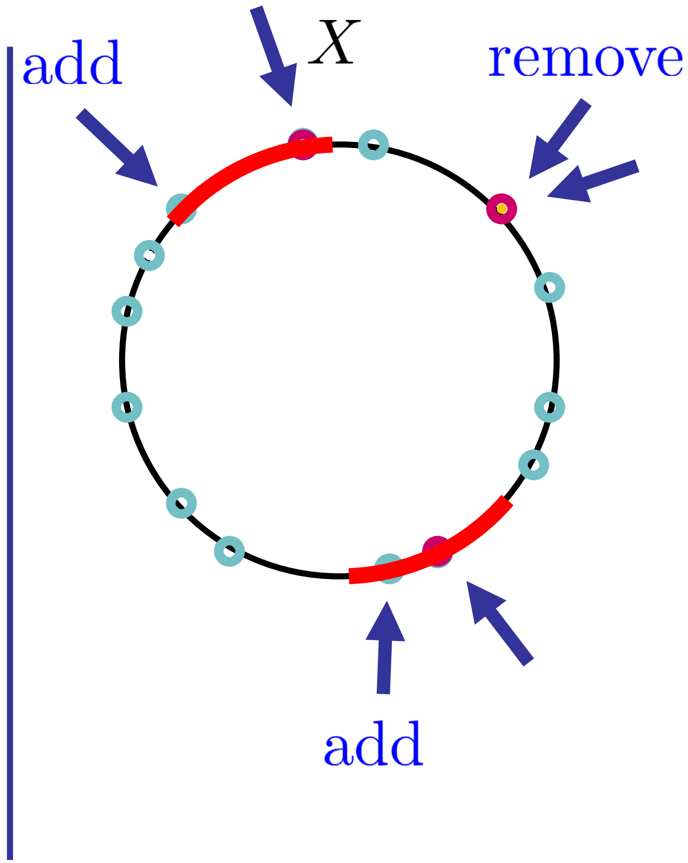
$|\langle \mathbf{x}_i, \delta_j \rangle| > \lambda$
$\Leftrightarrow [\mathbf{c}_j^*]_i \neq 0$

Oracle guided active set (ORGEN) algorithm

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



- initialize support set T
- compute oracle region



Oracle guided active set (ORGEN) algorithm

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

X

$\mathbf{x}_j$

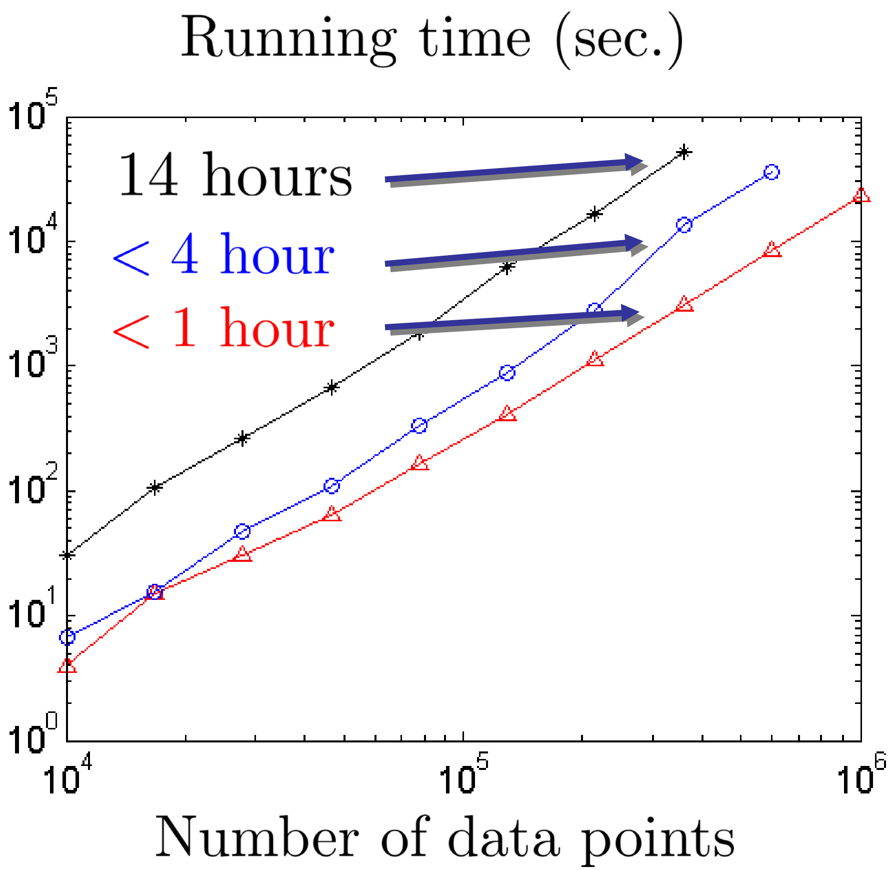
Theorem:

The support set T converges to the true support set in a finite number of iterations

- init
- con
- upc
- rep

Efficiency is gained by solving multiple small problems instead of one big problem

SSC-BP vs. SSC-OMP vs. EnSC-Oracle



- SSC-BP (Baseline)
- SSC-OMP
- EnSC-Oracle

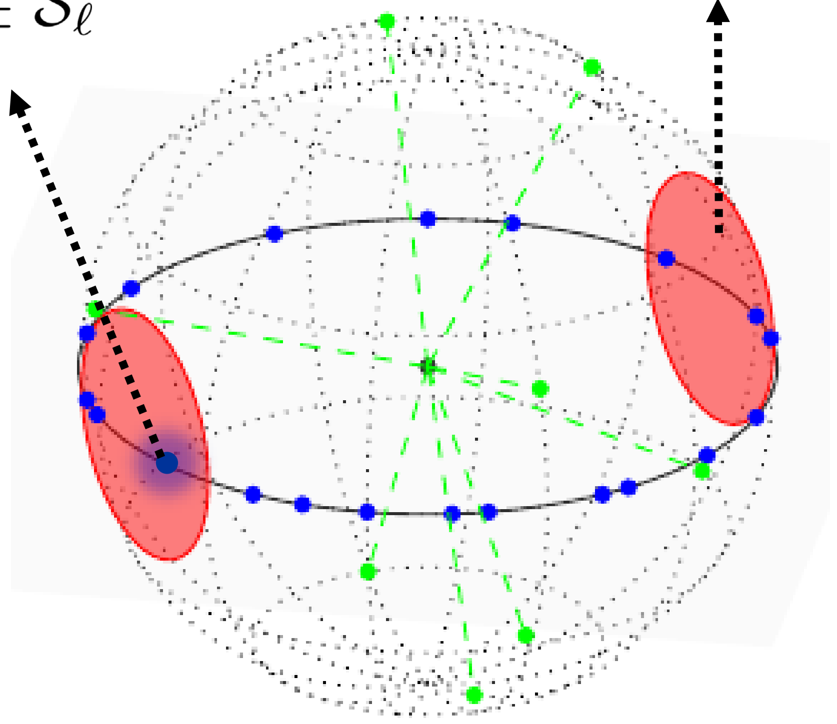
reduces the time for SSC-BP

Correct connections vs. connectivity

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

- λ is large
  - $\Longrightarrow$ oracle region is small
  - $\Longrightarrow$ correct connection

- λ is small
  - $\Longrightarrow$ oracle region is large
  - $\Longrightarrow$ well-connected



Guaranteed no wrong connections

$$\min_{\mathbf{c}_j} \|\mathbf{c}_j\|_1 + \frac{\gamma}{2}\|\mathbf{x}_j - X\mathbf{c}_j\|_2^2 \quad \text{s.t.} \quad \mathbf{c}_{jj} = 0$$

Theorem: (for SSC)

Condition for guaranteed no wrong connections:

$$\mu(W^{(\ell)}, X^{(-\ell)}) < r^{(\ell)}$$

Similarity between subspaces

Distribution of points

Guaranteed no wrong connections

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

Theorem: (for EnSC)

Condition for guaranteed no wrong connections:

$$\mu(W^{(\ell)}, X^{(-\ell)}) < r^{(\ell)} - \frac{1-\lambda}{\lambda}$$

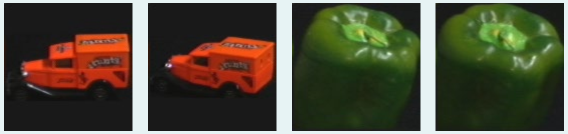
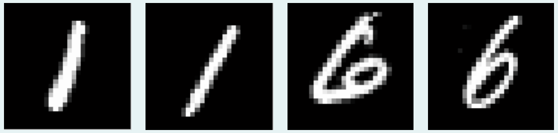
Similarity between subspaces

Distribution of points

Condition is harder to be satisfied
Graph has better connectivity
} Higher clustering accuracy

Experiments

Test of EnSC with ORGEN on real data

database	# data	ambient dim.	# clusters	Examples
Coil-100	7,200	1024	100	
PIE	11,554	1024	68	
MNIST	70,000	500	10	
CovType	581,012	54	7	

Experiments

Our method (EnSC) achieves the best clustering accuracy

database	# data	SSC-BP	SSC-OMP	EnSC
Coil-100	7,200	57.10%	42.93%	**69.24%**
PIE	11,554	41.94%	24.06%	**52.98%**
MNIST	70,000	-	93.07%	**93.79%**
CovType	581,012	-	48.76%	**53.52%**

Experiments

Our method (EnSC) is scalable

database	# data	SSC-BP	SSC-OMP	EnSC
Coil-100	7,200	127 mins	**3 mins**	**3 mins**
PIE	11,554	412 mins	**5 mins**	13 mins
MNIST	70,000	-	**6 mins**	28 mins
CovType	581,012	-	**783 mins**	1452 mins

Conclusion

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

- ✓ guaranteed correct connections
- ✓ improved connectivity

} better clustering

- ✓ efficient algorithm for large scale problems

Acknowledgement

Funding: NSF-IIS 1447822

Vision Lab @ Johns Hopkins University

http://www.vision.jhu.edu

Thank you!

Center for IMAGING SCIENCE