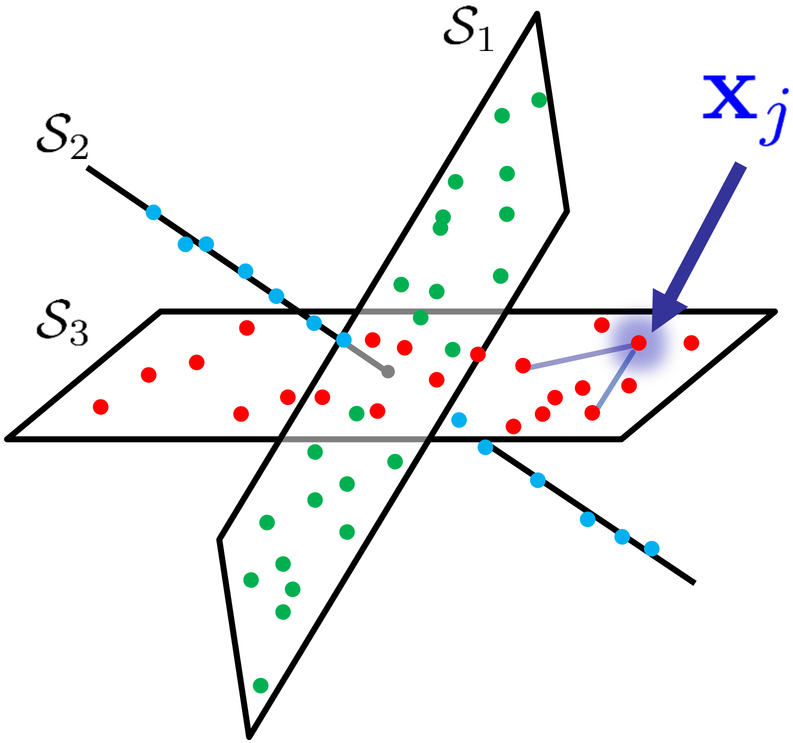
Scalable Sparse Subspace Clustering

René Vidal

Center for Imaging Science, Johns Hopkins University

Sparse Subspace Clustering



Sparse Subspace Clustering (SSC) [1]

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

- Nice theory
- Robust to noise, outliers and missing entries
- Can handle mid-size datasets ~ 1K-10K data points

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

Center for IMAGING SCIENCE

Scalable Sparse Subspace Clustering

- Subspace Clustering by Orthogonal Matching Pursuit (SSC-OMP)
  - Similar theoretical guarantees to SSC regarding correct connections
  - Three orders of magnitude faster than SSC
  - Scalable to 100K data points
- Subspace Clustering by Elastic Net Regularization (EnSC)
  - Similar theoretical guarantees to SSC regarding correct connections
  - Better theoretical guarantees regarding connectivity of each cluster
  - Three orders of magnitude faster than SSC
  - Novel algorithm that is scalable to 100K data points, and is also applicable to SSC problem

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 (SSC)

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

Basis pursuit[1] (BP)

Orthogonal matching pursuit[2] (OMP)

Method:
- Convex relaxation
- Replace $\|\mathbf{c}_j\|_0$ with $\|\mathbf{c}_j\|_1$

Properties:

✓ Guaranteed correct connections

✗ Not scalable: solved by CVX/ADMM

Method:
- Greedy pursuit
- Choose one at a time

Properties:

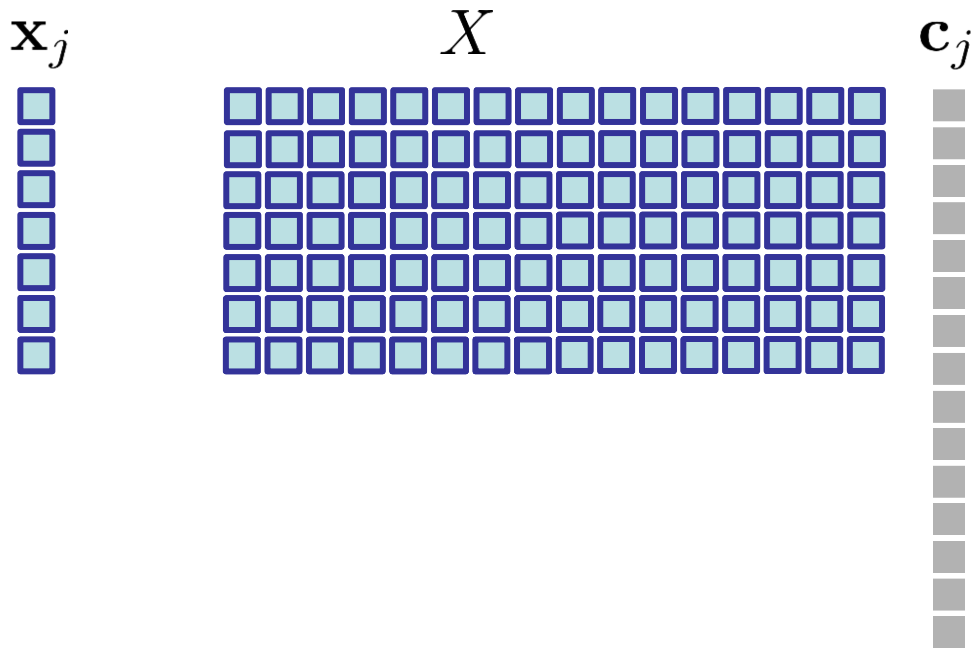
✓ Guaranteed correct connections

✓ Scalability: from 100 to 100,000 points

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

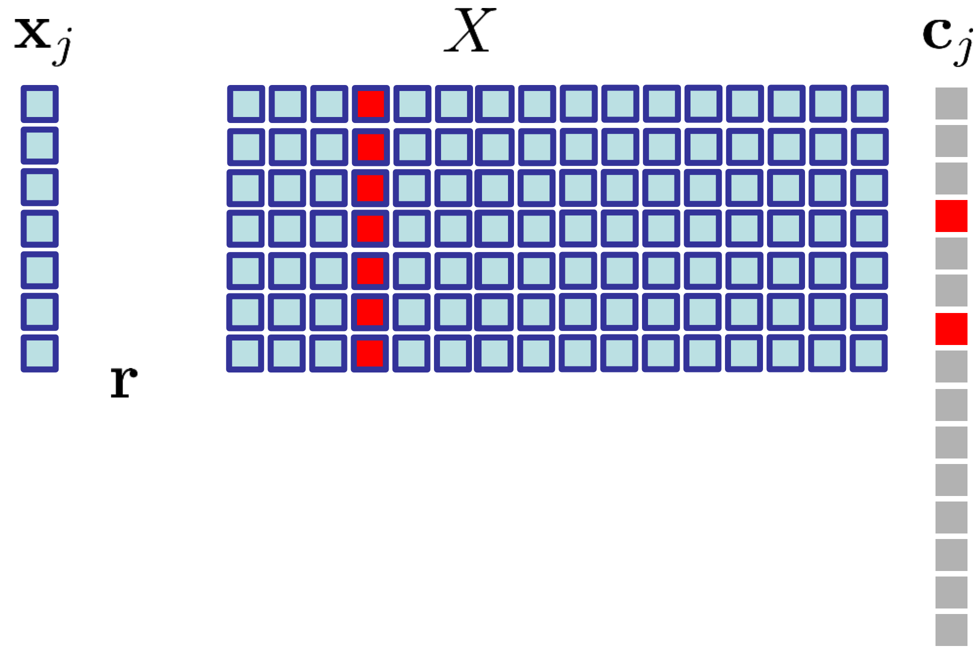
SSC by orthogonal matching pursuit

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



SSC by orthogonal matching pursuit

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



Give 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]:

$W_j^\ell =$ dual points

For SSC-OMP:

$W_j^\ell =$ residual points

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

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:

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

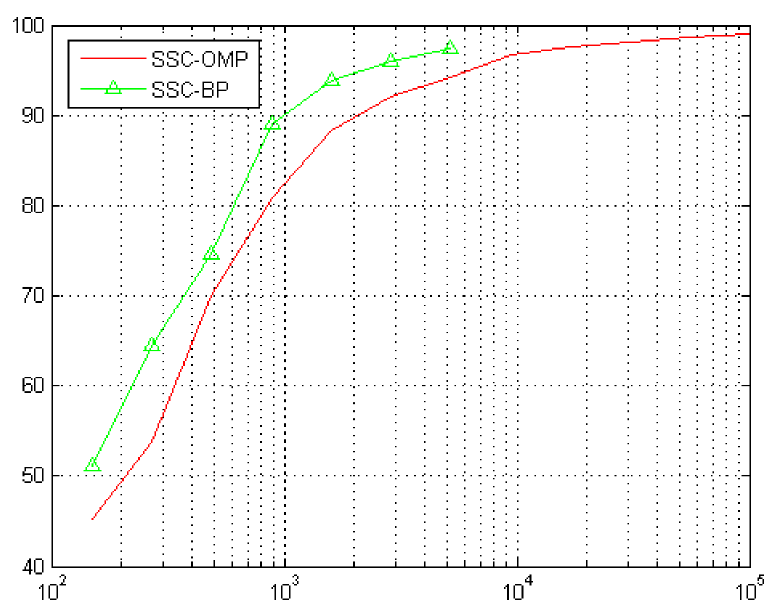
[3] Soltanolkotabi-Candes, A geometric analysis of subspace clustering with outliers

Synthetic experiments

Random model:

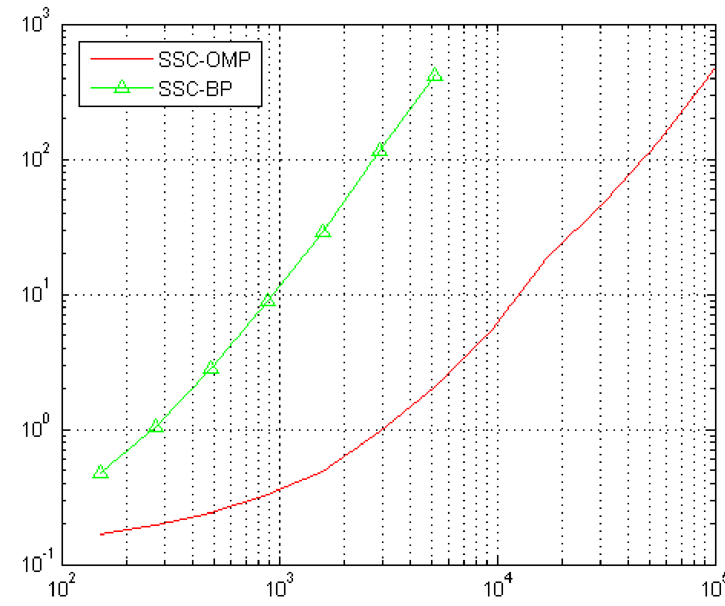
- 5 subspaces of dimension 6 in ambient dimension 9
- Vary the sample density

Clustering accuracy



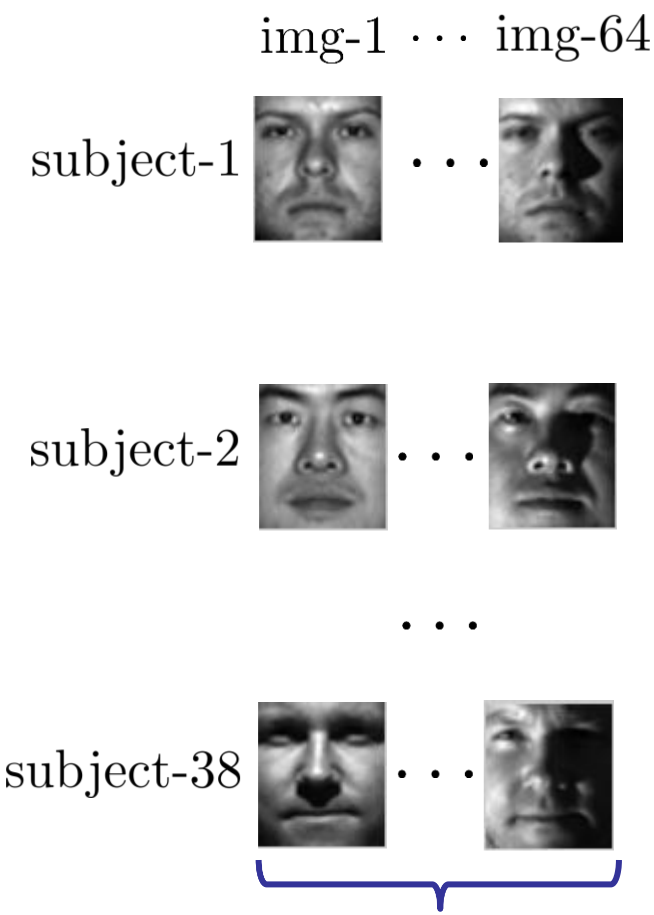
Number of data point

Running time



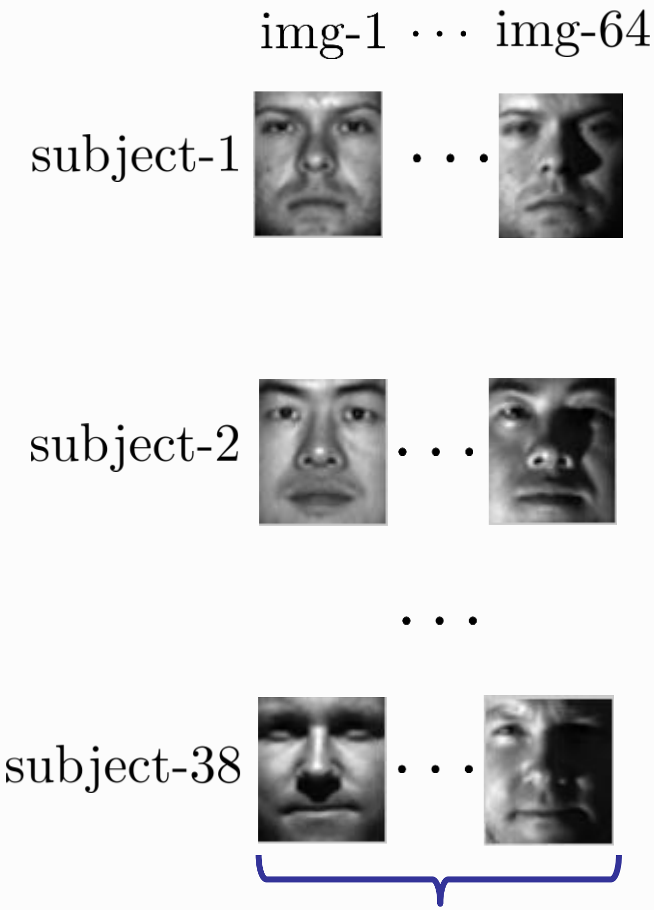
Number of data point

Experiment on extended Yale B



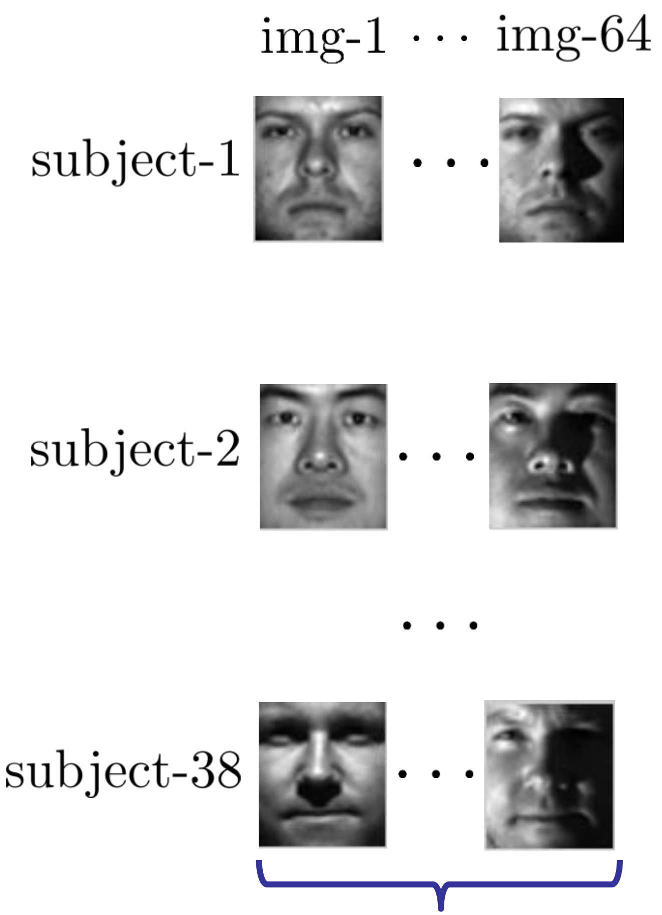
No. subjects	2	10	20	30	38
a%: average clustering accuracy					
SSC-OMP	99.21	88.43	**81.71**	**79.27**	**80.45**
SSC-BP	**99.45**	**91.85**	79.80	76.10	68.97
LSR	96.77	62.89	67.17	67.79	63.96
LRSC	94.32	66.98	66.34	67.49	66.78
SCC	78.91	NA	NA	14.15	12.80
t(sec.): running time					
SSC-OMP	0.3	1.7	4.7	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

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

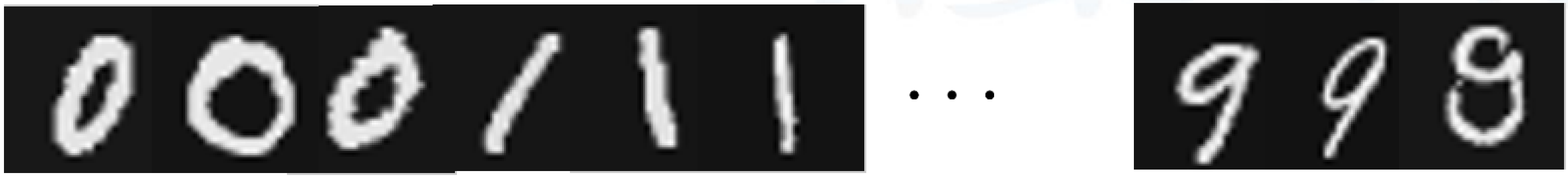
Experiment on extended Yale B



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	-	-

Conclusion

SSC by Orthogonal Matching Pursuit (OMP):

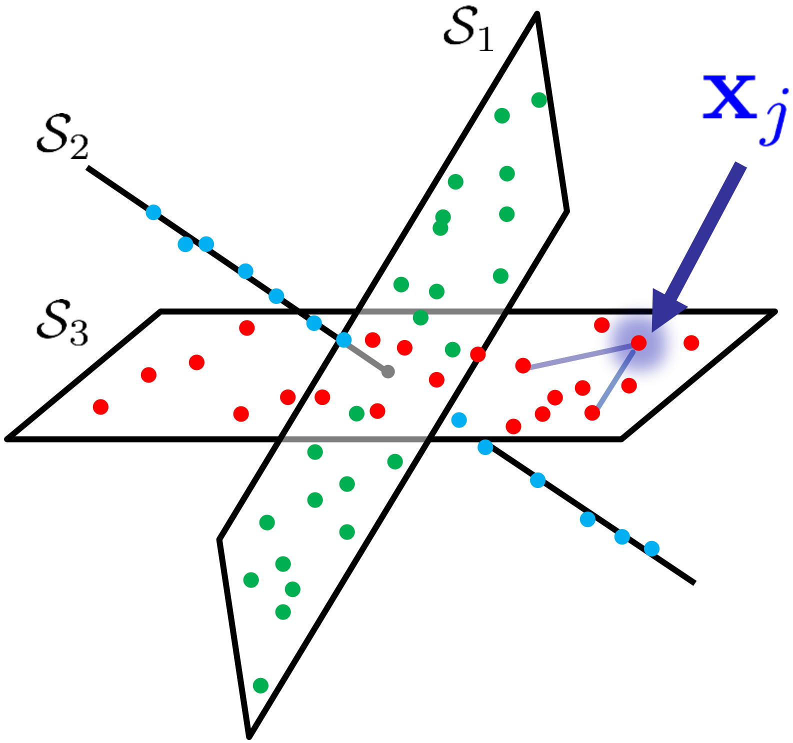
- ✓ stronger theoretical guarantees for correct connections
- ✓ performance validation on large databases

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

Prior work: Spectral Subspace Clustering



Sparse Subspace Clustering (SSC) [1]

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

$$\text{s.t.} \quad c_{jj} = 0$$

Least Squares Regression (LSR) [2]

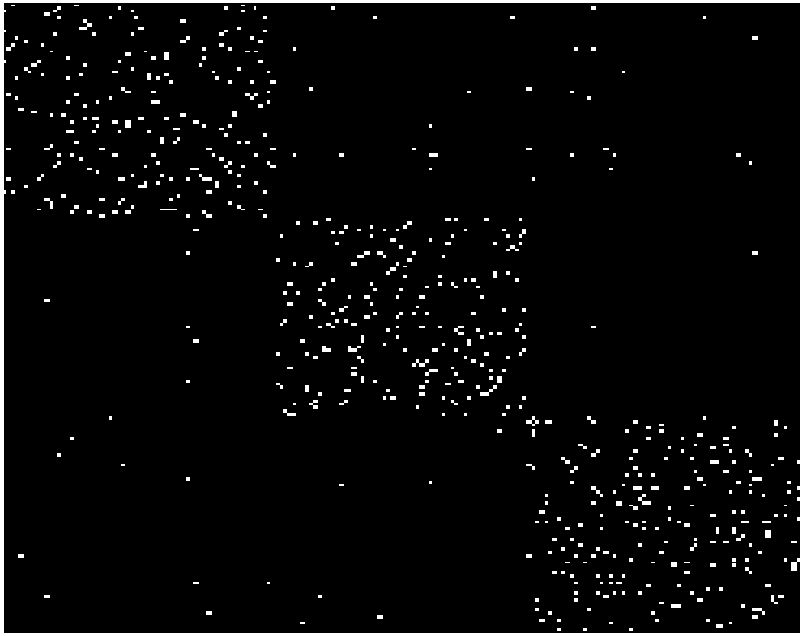
$$\min_{\mathbf{c}_j} {\color{red}\|\mathbf{c}_j\|_2^2} + \frac{\mu}{2}\|\mathbf{x}_j - X\mathbf{c}_j\|_2^2$$

$$\text{s.t.} \quad c_{jj} = 0$$

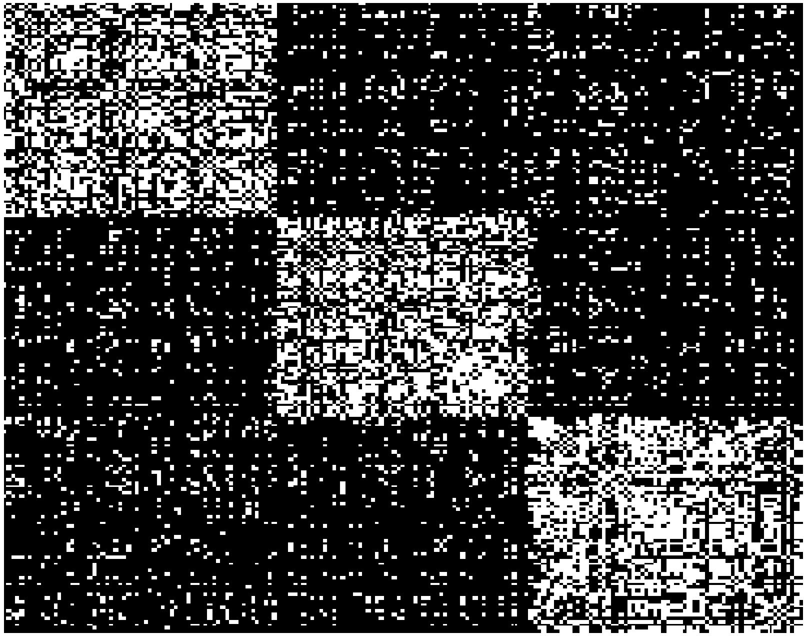
[1] Elhamifar-Vidal, Sparse Subspace Clustering, CVPR 2009
[2] Lu et al,, Robust and efficient subspace segmentation via least squares regression, ECCV 2012.

Prior work: Spectral Subspace Clustering

SSC ($\|\mathbf{c}_j\|_1$)



LSR ($\|\mathbf{c}_j\|_2^2$)



✓ Few wrong connections

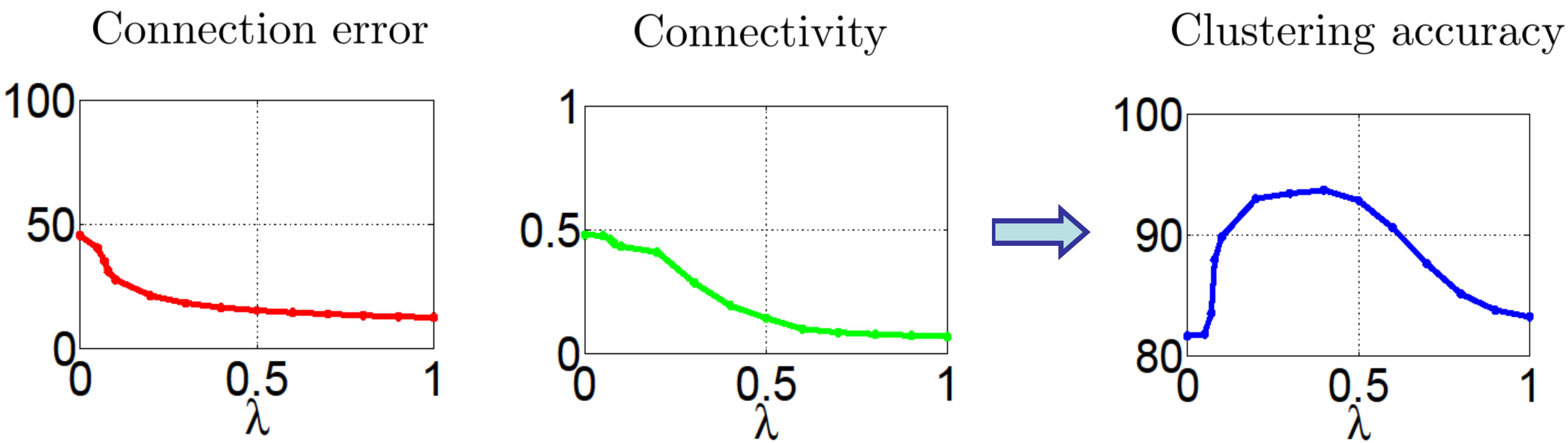
✗ Not well connected

✗ Many wrong connections

✓ Well-connected

Correct connection vs. connectivity

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



Key challenge: scalability

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

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

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

- We propose a new scalable algorithm
  - Exploits geometry of the EnSC problem to efficiently update the solution
  - Each iteration works with a small subset of data

Contributions

- We provide a geometric analysis of the solution, which explains the trade off between correct connection and connectivity effects.

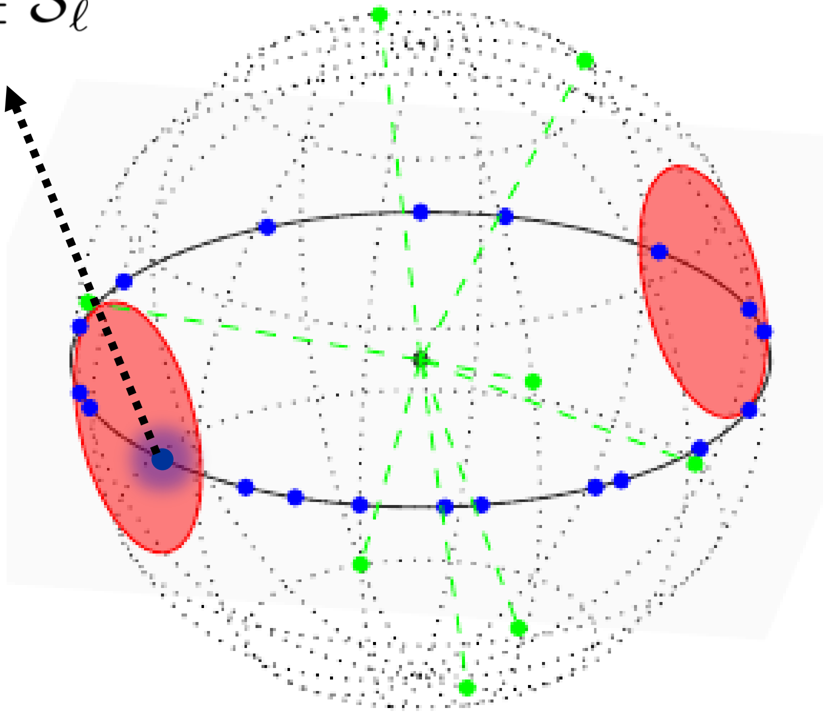
- We provide conditions under which the solution gives correct connection. Our condition improves upon prior result for SSC.

- We derive a scalable algorithm that can handle 1 million data points.

Correct connections vs. connectivity

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

$\mathbf{x}_j \in \mathcal{S}_\ell$



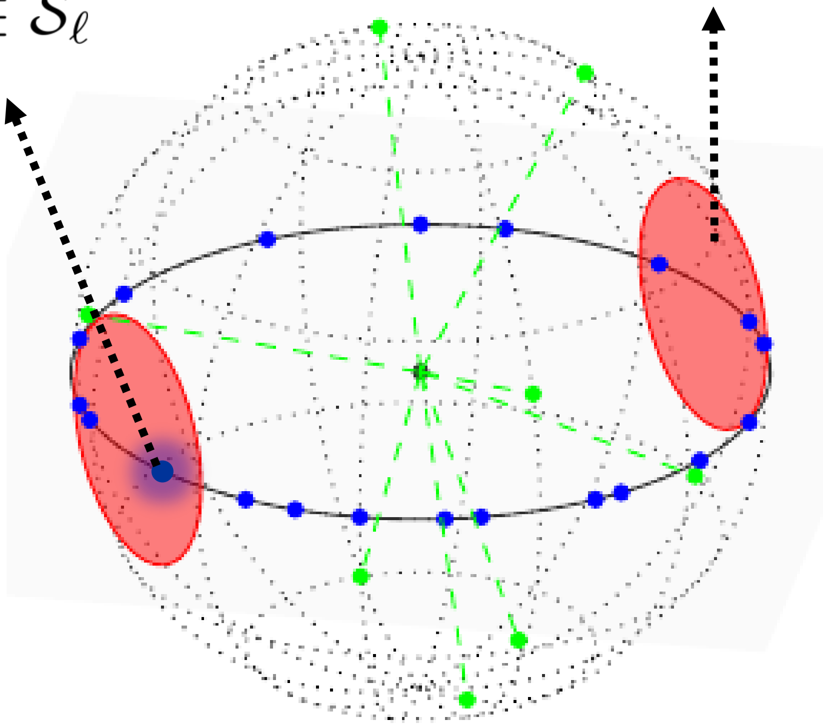
Correct connections vs. connectivity

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

- $c_{ij} \neq 0$ if and only if $\mathbf{x}_i$ belongs to the oracle region for $\mathbf{x}_j$

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

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



Guaranteed correct connection

Theorem

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

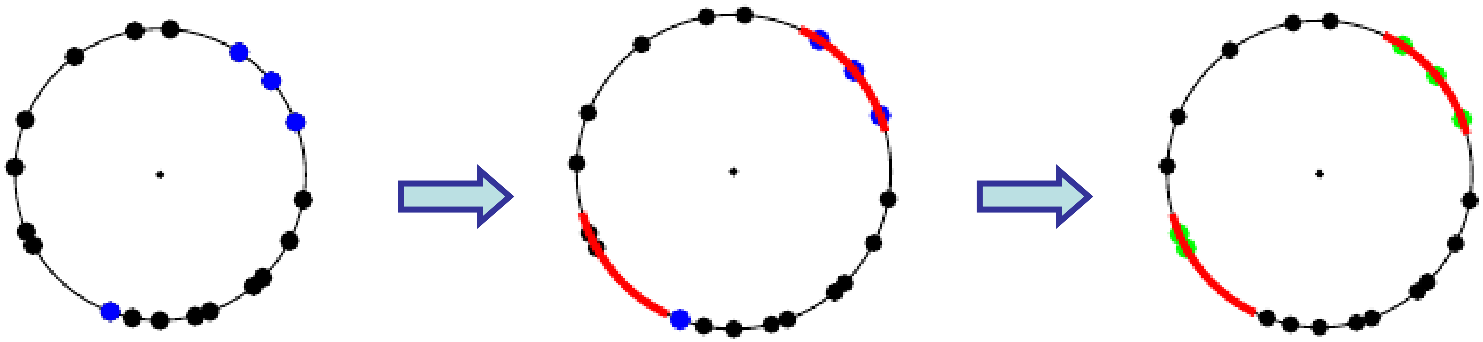
$$\mu(W_j^\ell, X^{-\ell}) \leq \frac{(r^\ell)^2}{r^\ell + \frac{1-\lambda}{\lambda}},$$

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

- RHS is an increasing function of λ
- Reduces to condition for SSC-BP if $\lambda = 1$

Oracle guided active set (ORGEN) algorithm

- Observation: if we know the *support set* T of the solution $\mathbf{c}_j$, it is reduced to a small scale problem
- Idea: solve a sequence of small scale problems on support T_k
- Algorithm: update T_k by using oracle region



Support set T_k

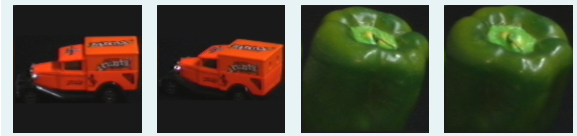
Oracle region on support T_k

Support set T_{k+1}

- Theorem: T_k converges to true support set T in a finite number of iterations

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 achieves the best clustering accuracy

	TSC	OMP	NSN	SSC[1]	SSC[2]	LRSC	ENSC	KMP	EnSC
Clustering accuracy (%)									
Coil-100	61.32	42.93	50.32	53.75	57.10	55.76	51.11	61.97	**69.24**
PIE	22.15	24.06	35.02	39.05	41.94	46.65	21.40	16.55	**52.98**
MNIST	85.00	93.07	85.82	92.46	-	-	-	-	**93.79**
CovType	35.45	48.76	38.04	-	-	-	-	-	**53.52**

1- SSC by the SPAMS package which implements the solution path method.
2- SSC by the ADMM algorithm

Experiments

- Our method achieves the best clustering accuracy

	TSC	OMP	NSN	SSC[1]	SSC[2]	LRSC	ENSC	KMP	EnSC
Clustering accuracy (%)									
Coil-100	61.32	42.93	50.32	53.75	57.10	55.76	51.11	61.97	**69.24**
PIE	22.15	24.06	35.02	39.05	41.94	46.65	21.40	16.55	**52.98**
MNIST	85.00	93.07	85.82	92.46	-	-	-	-	**93.79**
CovType	35.45	48.76	38.04	-	-	-	-	-	**53.52**

1- SSC by the SPAMS package which implements the solution path method.
2- SSC by the ADMM algorithm

Experiments

- Our method achieves the best clustering accuracy

	TSC	OMP	NSN	SSC[1]	SSC[2]	LRSC	ENSC	KMP	EnSC
Clustering accuracy (%)									
Coil-100	61.32	42.93	50.32	53.75	57.10	55.76	51.11	61.97	**69.24**
PIE	22.15	24.06	35.02	39.05	41.94	46.65	21.40	16.55	**52.98**
MNIST	85.00	93.07	85.82	92.46	-	-	-	-	**93.79**
CovType	35.45	48.76	38.04	-	-	-	-	-	**53.52**

- Our method is able to handle large databases efficiently

	TSC	OMP	NSN	SSC[1]	SSC[2]	LRSC	ENSC	KMP	EnSC
Running time (min.)									
Coil-100	**2**	3	11	16	127	3	8	63	3
PIE	**3**	5	25	67	412	12	25	361	13
MNIST	30	**6**	298	1350	-	-	-	-	28
CovType	999	**783**	3572	-	-	-	-	-	1452

1- SSC by the SPAMS package which implements the solution path method.
2- SSC by the ADMM algorithm

Conclusion

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

- ✓ guaranteed correct connections
- ✓ improved connectivity
- ✓ efficient algorithm for large scale problems

Acknowledgement

Funding: NSF-IIS 1447822

Vision Lab @ Johns Hopkins University
http://www.vision.jhu.edu

Thank You for your attention

Choice of Regularization

- Prior work

Method	$f(\mathbf{c}_j)$ or $f(C)$	Subspace preserving[1]	Connected[2]	Efficient
SSC [1]	$\|\mathbf{c}_j\|_1$	✓		
LRR/LRSC [2]	$\|C\|_*$		✓	
LSR [3]	$\|\mathbf{c}_j\|_2^2$		✓	
OMP/NSN [4]	$\|\mathbf{c}_j\|_0$ (Greedy)	✓		✓
LRSSC [5]	$\|C\|_1 + \|C\|_*$	✓	✓	
CASS [6]	$\|X\mathrm{Diag}(\mathbf{c}_j)\|_*$		✓	
KMP [7]	$\|\mathbf{c}_j\|_k^{sp}$		✓	
EnSC(Ours)	$\|\mathbf{c}_j\|_1 + \|\mathbf{c}_j\|_2^2$	✓	✓	✓

[1]there exists theoretical guaranteed subspace preserving property under general conditions.
[2]the solution is dense or have the grouping effect.

[1] Elhamifar-Vidal '09 '10 '13, Soltanolkotabi-Candes '12 '13, Wang-Xu '13 [2] Liu et al. '10 '13, Favaro '11, Vidal '13
[3] Lu et al. '12 [4] Dyer et al. '13, Park et al. '14, Adler et al. '15, You et al. '16 [5] Wang '13 [6] Lu et al. '13 [7] Lai et al. '14

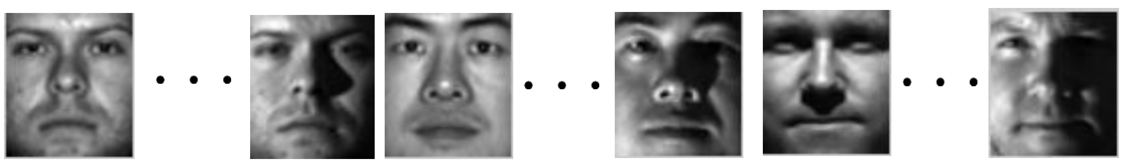
Center for IMAGING SCIENCE

Prior work: Sparse Subspace Clustering (SSC)

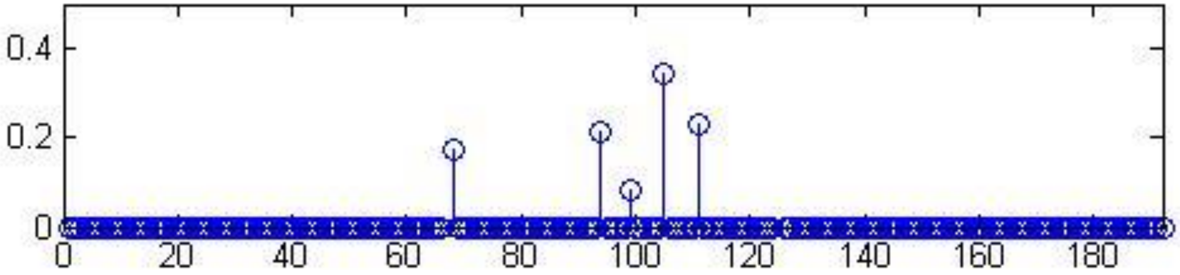
SSC: $\min_{\mathbf{c}_j} \|\mathbf{c}_j\|_1 \quad \text{s.t.} \quad \mathbf{x}_j = X\mathbf{c}_j, c_{jj} = 0$

solution is subspace preserving

X:



$\mathbf{x}_j$:



$\mathbf{c}_j$

Prior work: Sparse Subspace Clustering (SSC)

SSC: $\min_{\mathbf{c}_j} \|\mathbf{c}_j\|_1 \quad \text{s.t.} \quad \mathbf{x}_j = X\mathbf{c}_j, c_{jj} = 0$

solution is subspace preserving