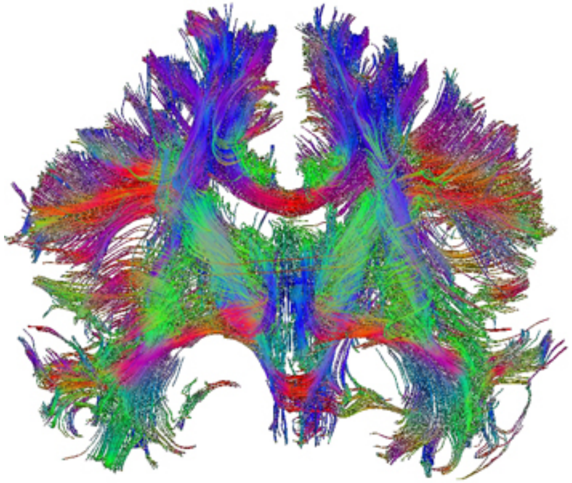
Algebraic, Sparse and Low Rank Subspace Clustering

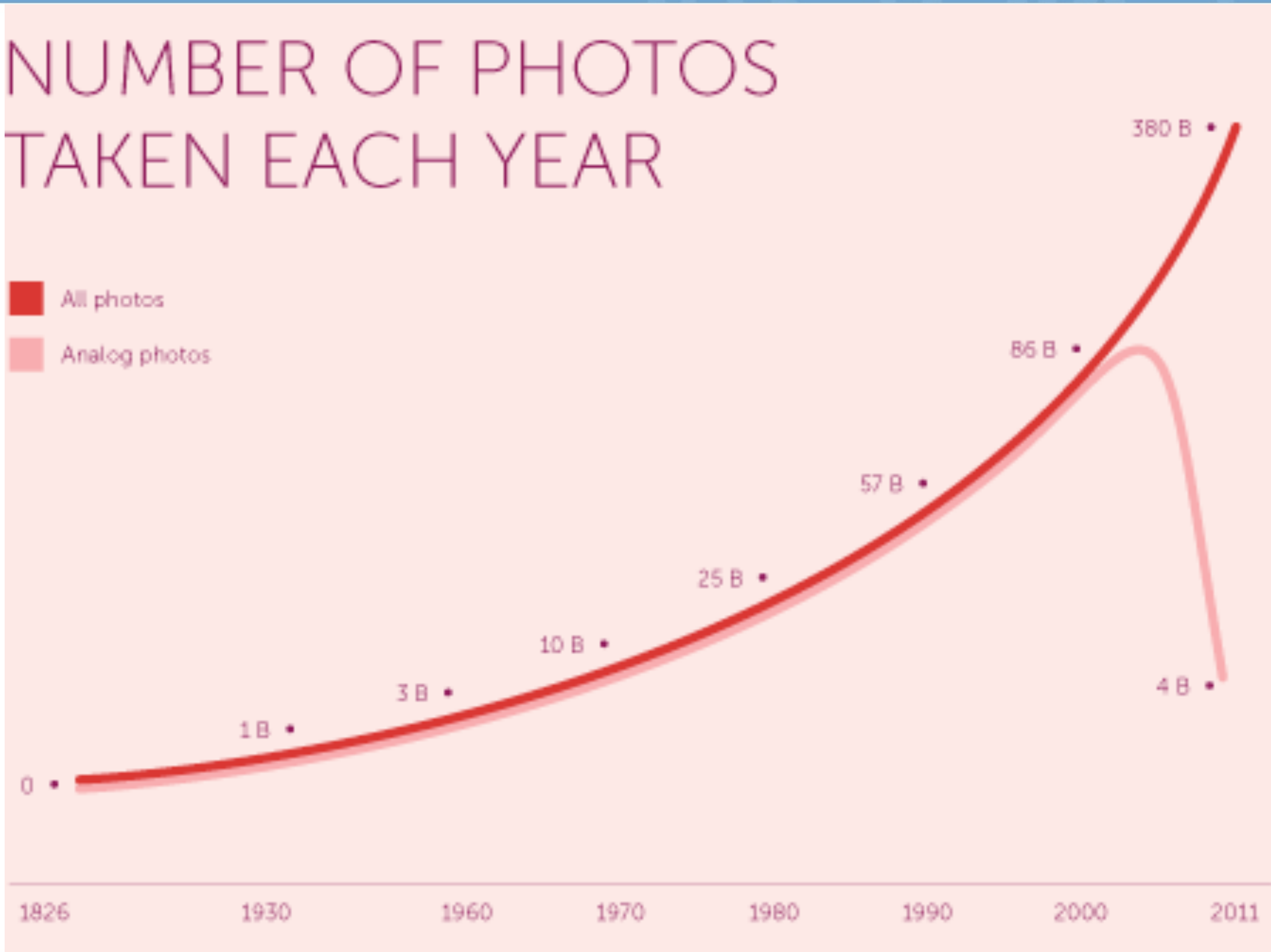
René Vidal
Center for Imaging Science
Johns Hopkins University

High-Dimensional Data

- In many areas, we deal with high-dimensional data
  - Computer vision
  - Medical imaging
  - Medical robotics
  - Signal processing
  - Bioinformatics



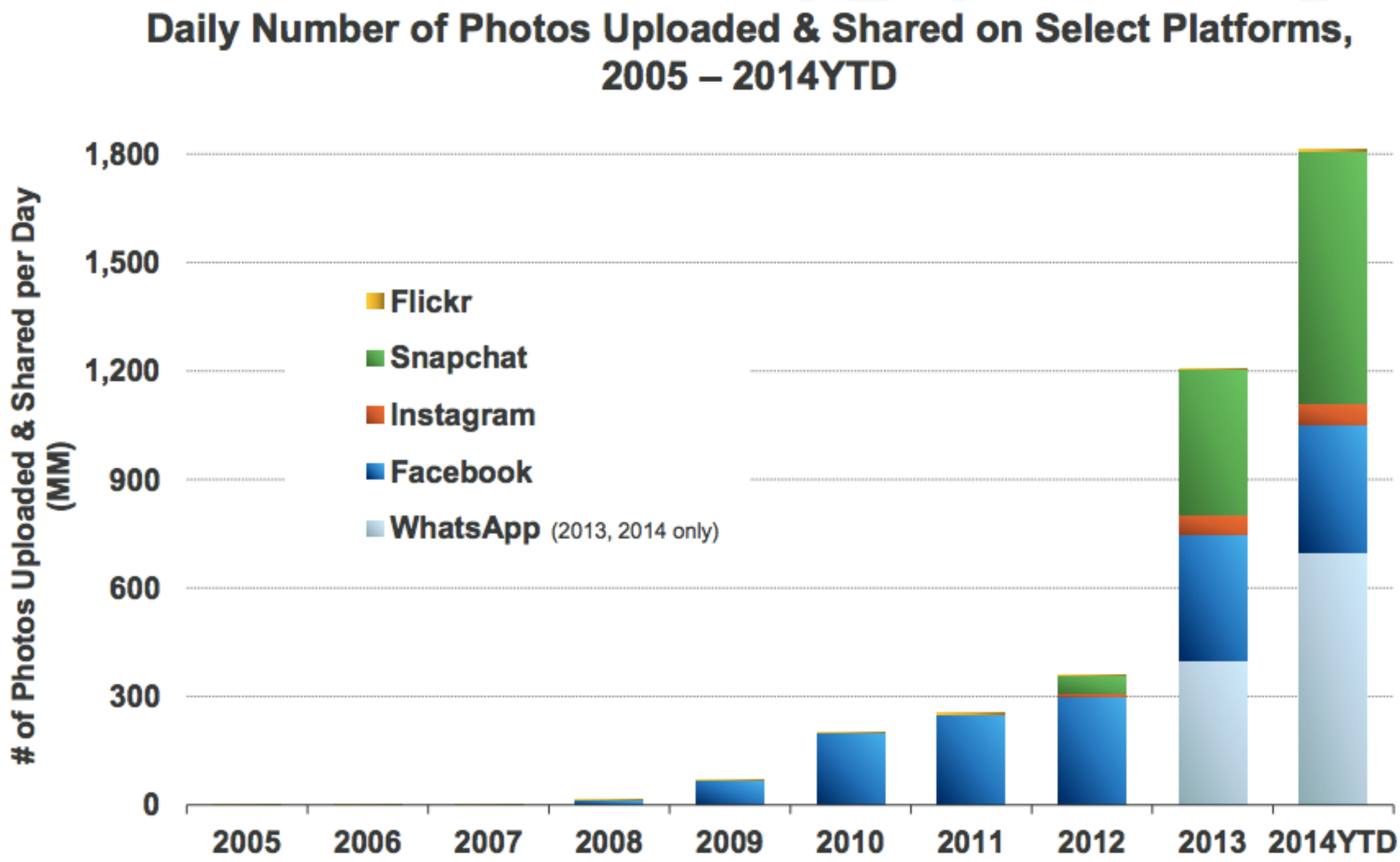
High-Dimensional Data in Computer Vision



http://blog.1000memories.com/94-number-of-photos-ever-taken-digital-and-analog-in-shoebox

Center for Imaging Science

High-Dimensional Data in Computer Vision



http://tech.firstpost.com/news-analysis/now-upload-share-1-8-billion-photos-everyday-meeker-report-224688.html

High-Dimensional Data in Computer Vision

facebook

- 140 billion images
- 350 million new photos/day

YouTube

- 120 million videos
- 300 hours of video/minute

- 3.8 trillion of photographs
- 10% in the past 12 months

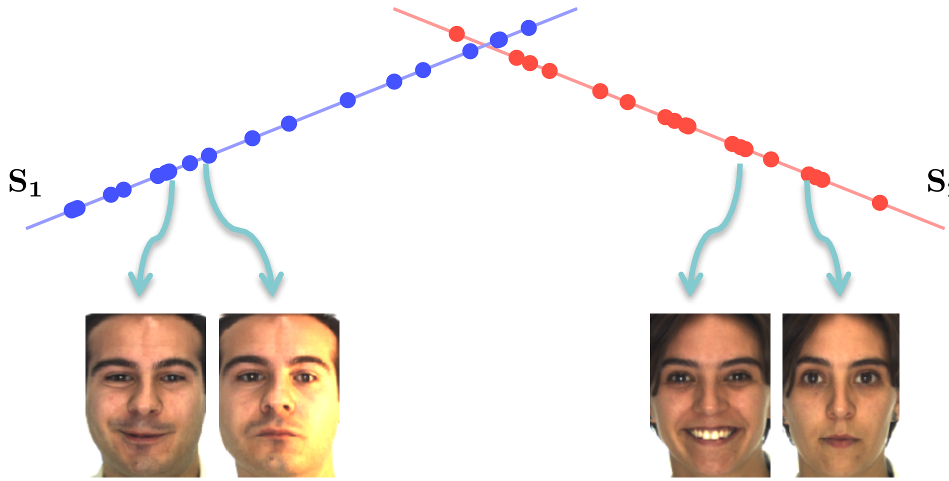
CISCO

- 90% of the internet traffic will be video by the end of 2017

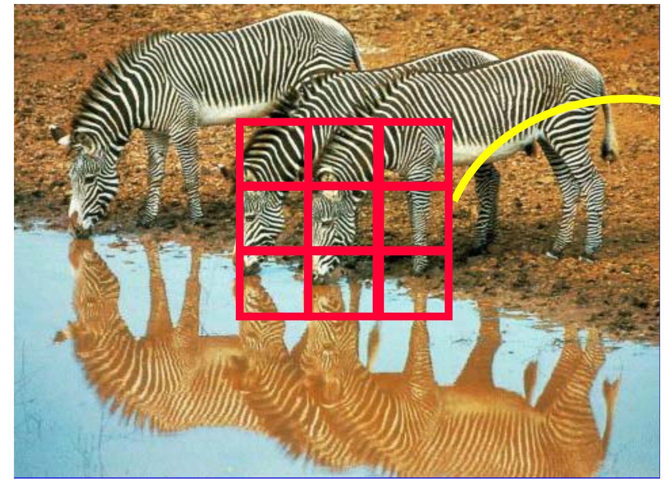
http://www.buzzfeed.com/hunterschwarz/how-many-photos-have-been-taken-ever-6zgv

Low-Dimensional Manifolds

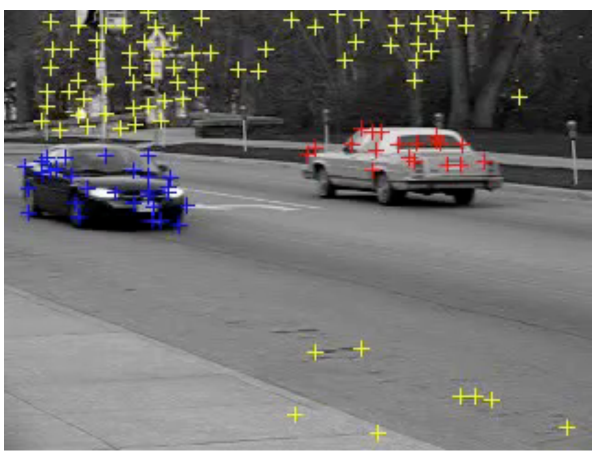
- Face clustering and classification



- Lossy image representation



- Motion segmentation

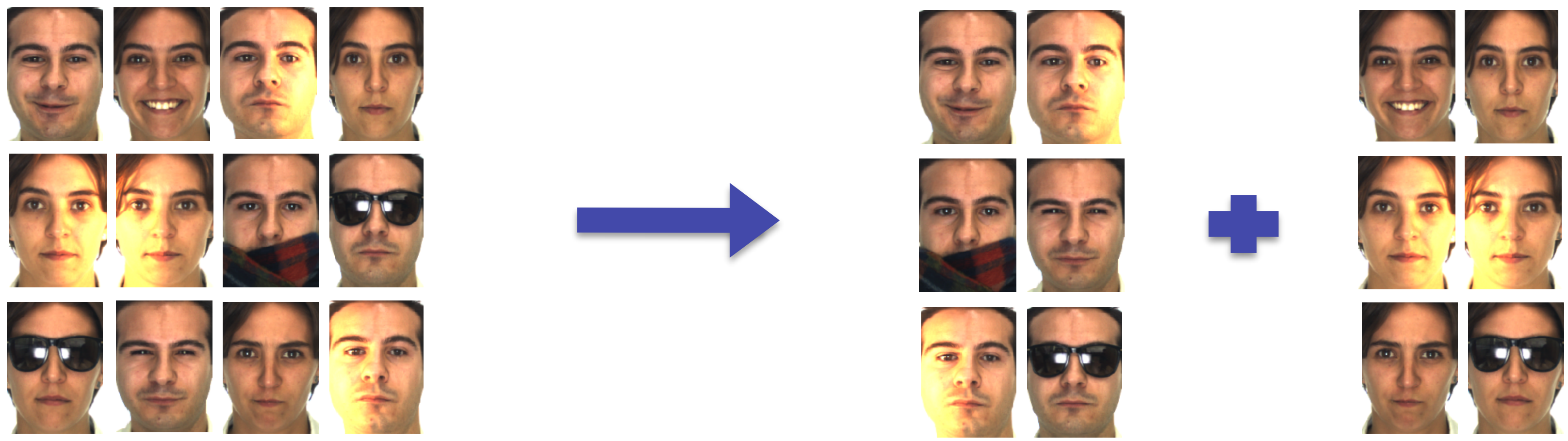


- DT segmentation

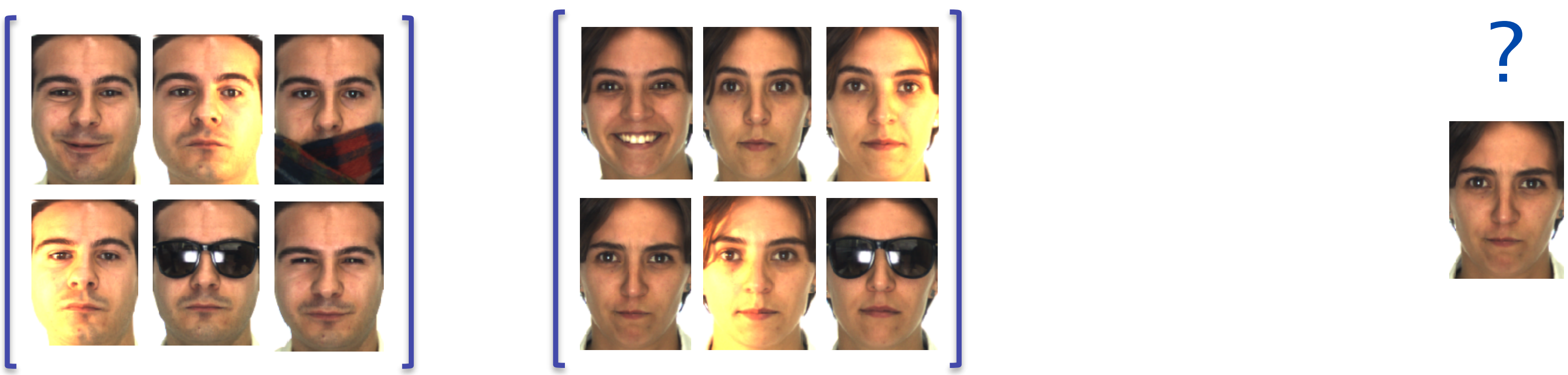
- Video segmentation

Two Fundamental Tasks

- Clustering of data in low-dimensional manifolds

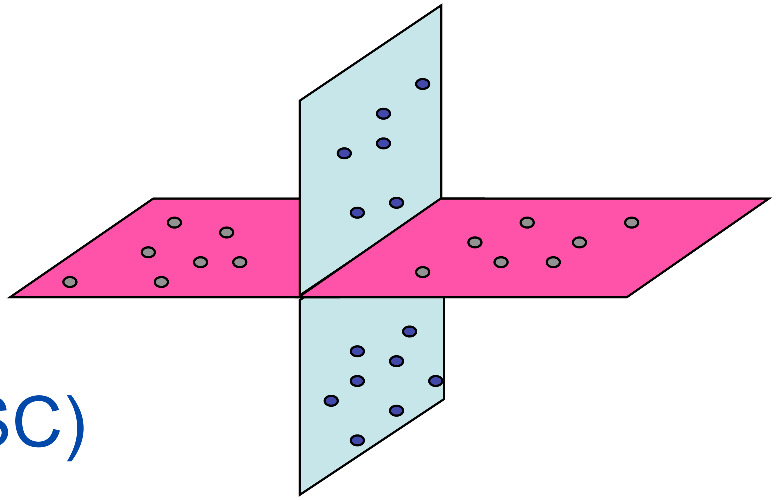


- Classification of data in low-dimensional manifolds



Talk Outline

- Introduction to Subspace Clustering
- Generalized Principal Component Analysis (GPCA)
  - Polynomial fitting and factorization
- Sparse Subspace Clustering (SSC)
  - Matrix of coefficients is sparse
- Low Rank Subspace Clustering (LRSC)
  - Matrix of coefficients is low-rank
- Applications:
  - Face clustering
  - Motion/video segmentation



R. Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA). CVPR 2003, CVPR 2004, PAMI 2005.
E. Elhamifar and R. Vidal. Sparse Subspace Clustering (SSC). CVPR 2009, ICASSP 2010, PAMI 2013.
P. Favaro, A. Ravichandran and R. Vidal. Low Rank Subspace Clustering (LRSC). CVPR 2011.

JHU vision lab

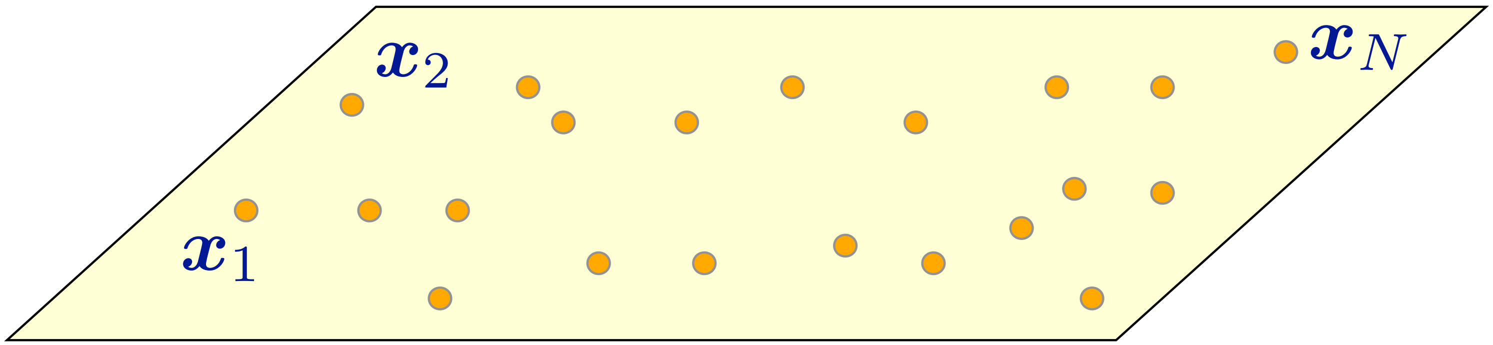
Introduction to Subspace Clustering

René Vidal

THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins
Center for Imaging Science

Principal Component Analysis (PCA)

- Given a set of points lying in one subspace, identify
  - Geometric PCA: find a subspace S passing through them
  - Statistical PCA: find projection directions that maximize the variance



- Solution (Beltrami'1873, Jordan'1874, Hotelling'33, Eckart-Householder-Young'36)

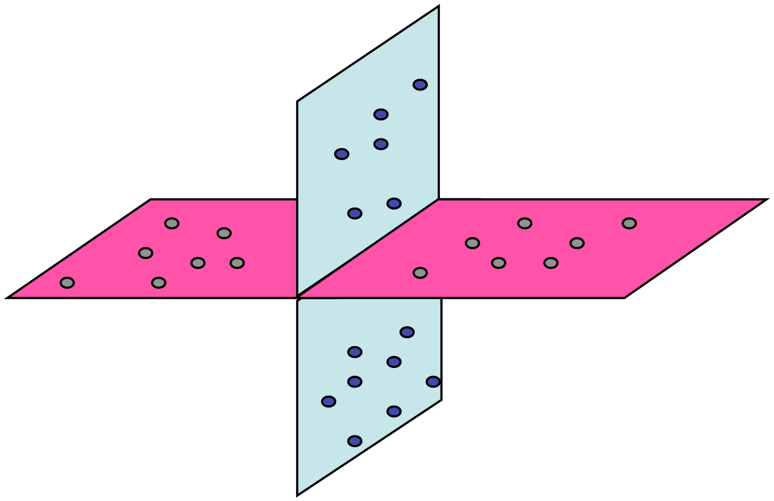
$$U\Sigma V^\top = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_N \end{bmatrix} \in \mathbb{R}^{D\times N}$$

- Applications:
  - Signal/image processing, computer vision (eigenfaces), machine learning, genomics, neuroscience (multi-channel neural recordings)

Subspace Clustering Problem

- Given a set of points lying in multiple subspaces, identify
  - The number of subspaces and their dimensions
  - A basis for each subspace
  - The segmentation of the data points

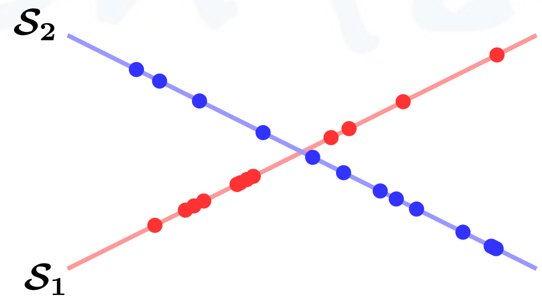
- Challenges
  - Model selection
  - Nonconvex
  - Combinatorial

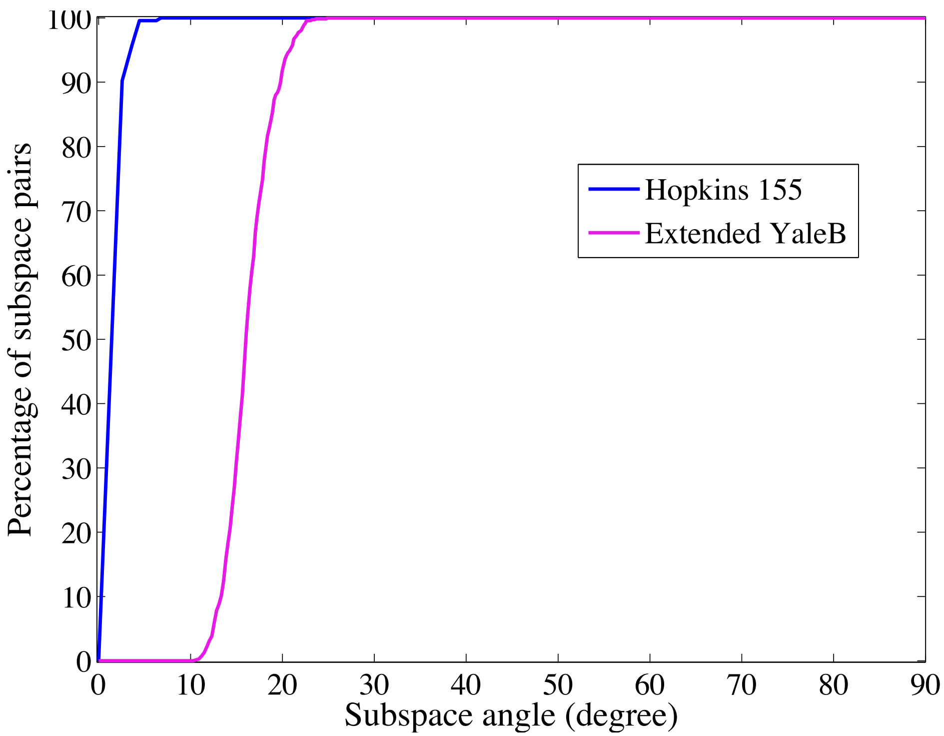


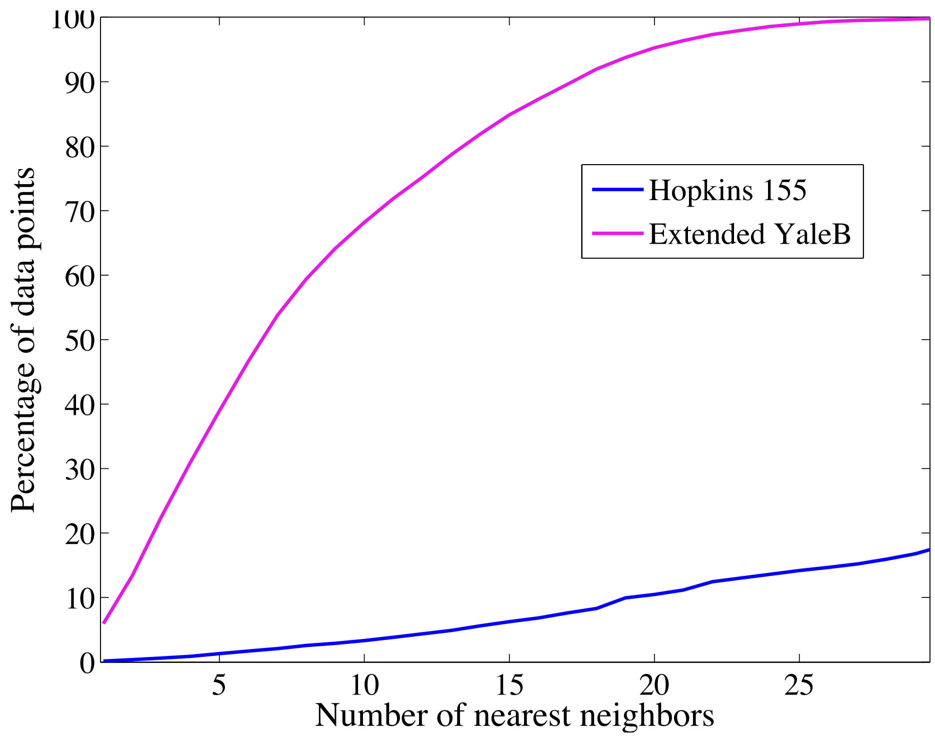
- More challenges
  - Noise
  - Outliers
  - Missing entries

Subspace Clustering Problem: Challenges

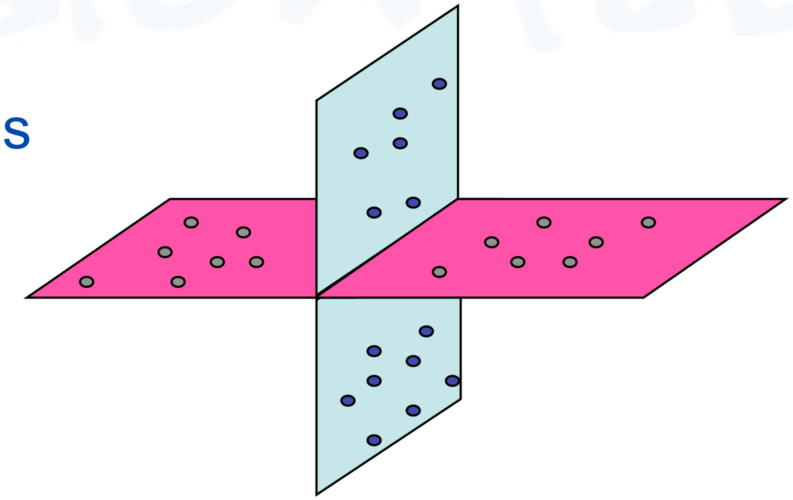
- Even more challenges
  - Angles between subspaces are small
  - Nearby points are in different subspaces







Prior Work: Iterative-Probabilistic Methods



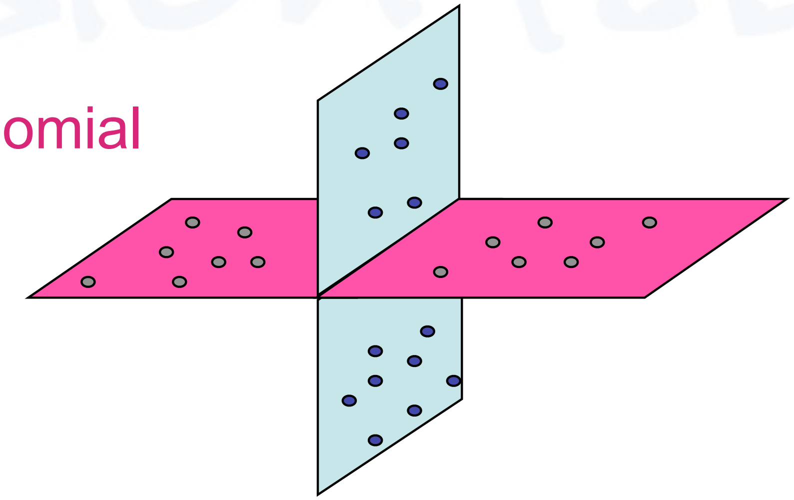
- Approach
  - Given segmentation, estimate subspaces
  - Given subspaces, segment the data
  - Iterate till convergence

- Representative methods
  - K-subspaces (Bradley-Mangasarian '00, Kambhatla-Leen '94, Tseng'00, Agarwal-Mustafa '04, Zhang et al. '09, Aldroubi et al. '09)
  - Mixtures of PPCA (Tipping-Bishop '99, Grubber-Weiss '04, Kanatani '04, Archambeau et al. '08, Chen '11)

Advantages	**Disadvantages / Open Problems**
Simple, intuitive	Known number of subspaces and dimensions
Missing data	Sensitive to initialization and outliers

Prior Work: Algebraic-Geometric Methods

- Approach
  - Number of subspaces = degree of polynomial
  - Subspaces = factors of polynomial

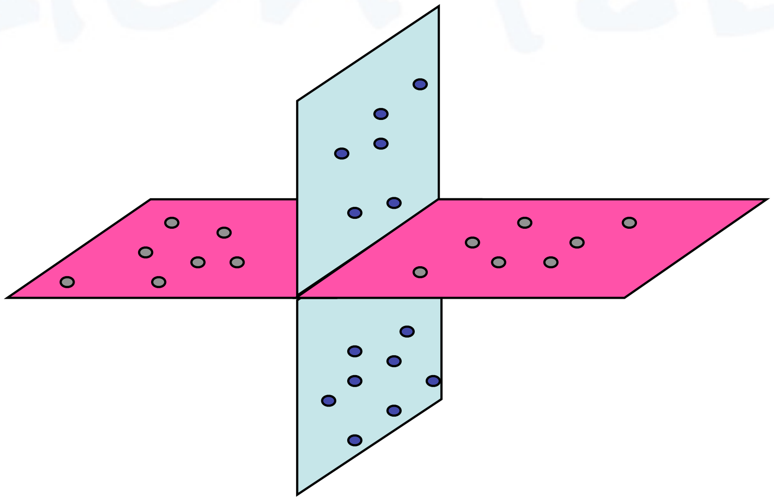


- Representative methods
  - Factorization (Boult-Brown'91, Costeira-Kanade'98, Gear'98, Kanatani et al.'01, Wu et al.'01, Sekmen'13)
  - GPCA (Shizawa-Maze '91, Vidal et al. '03 '04 '05, Huang et al. '05, Yang et al. '05, Derksen '07, Ma et al. '08, Ozay et al. '10)

Advantages	**Disadvantages / Open Problems**
Closed form	Complexity
Arbitrary dimensions	Sensitive to noise, outliers, missing entries

Prior Work: Spectral-Clustering Methods

- Approach
  - Data points = graph nodes
  - Pairwise similarity = edge weights
  - Segmentation = graph cut

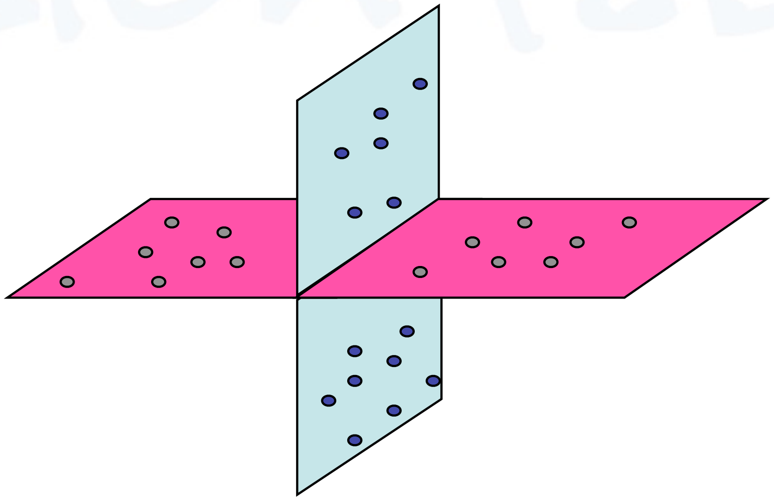


- Representative methods
  - Local (Zelnik-Manor '03, Yan-Pollefeys '06, Fan-Wu '06, Goh-Vidal '07, Sekmen'12)
  - Global (Govindu '05, Agarwal et al. '05, Chen-Lerman '08, Lauer-Schnorr '09, Zhang et al. '10)

Advantages	**Disadvantages / Open Problems**
Efficient	Known number of subspaces and dimensions
Robust	Global methods are complex

Prior Work: Sparse and Low-Rank Methods

- Approach
  - Data are self-expressive
  - Global affinity by convex optimization



- Representative methods
  - Sparse Subspace Clustering (SSC)
    (Elhamifar-Vidal ’09 ’10 ‘13, Candes ’12 ‘13)
  - Low-Rank Subspace Clustering (LRSC)
    (Liu et al. ’10 ‘13, Favaro-Vidal ’11 ’13)
  - Sparse + Low-Rank (Wang ‘13)

Advantages	**Disadvantages / Open Problems**
Efficient, Convex	Low-dimensional subspaces
Robust	Missing entries

Prior Work on Subspace Clustering

JHU vision lab

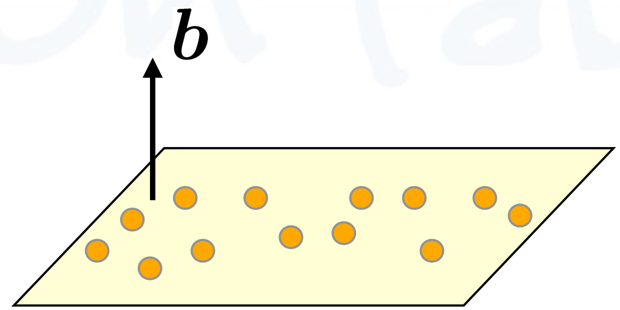
Generalized Principal Component Analysis (GPCA)

René Vidal, Yi Ma and Shankar Sastry

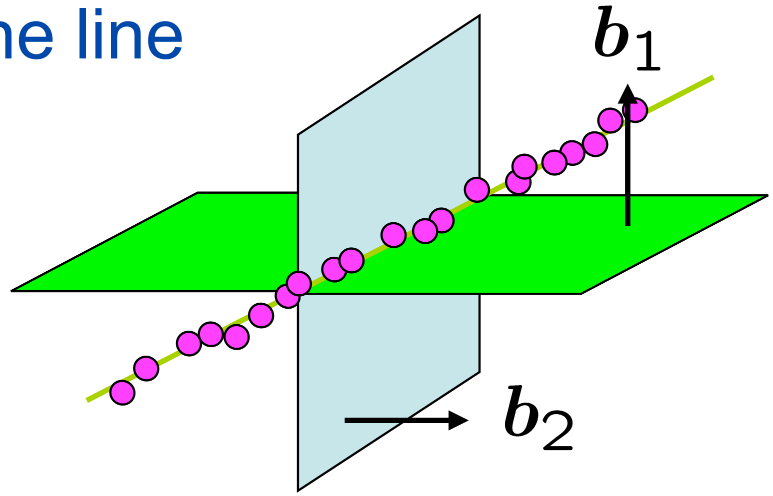
GPCA: Representing One Subspace

- One plane

$$\boldsymbol{b}^T \boldsymbol{x} = b_1 x_1 + b_2 x_2 + b_3 x_3 = 0$$



- One line



$$\boldsymbol{b}_1^T \boldsymbol{x} = b_1 x_1 + b_2 x_2 + b_3 x_3 = 0$$

$$\boldsymbol{b}_2^T \boldsymbol{x} = b_4 x_1 + b_5 x_2 + b_6 x_3 = 0$$

- One subspace can be represented with
  - Set of linear equations
  - Set of polynomials of degree 1

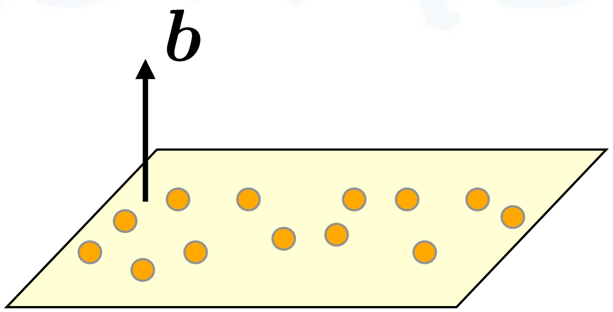
$$S = \{\boldsymbol{x} : B^T \boldsymbol{x} = 0\}$$

Vidal, Ma, Piazzi, Sastry. A new GPCA Algorithm for Clustering Subspaces by Fitting, Differentiating and Dividing Polynomials, CVPR 04.
R. Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA), PAMI 2005.

GPCA: Representing a Union of Subspaces

- One subspace

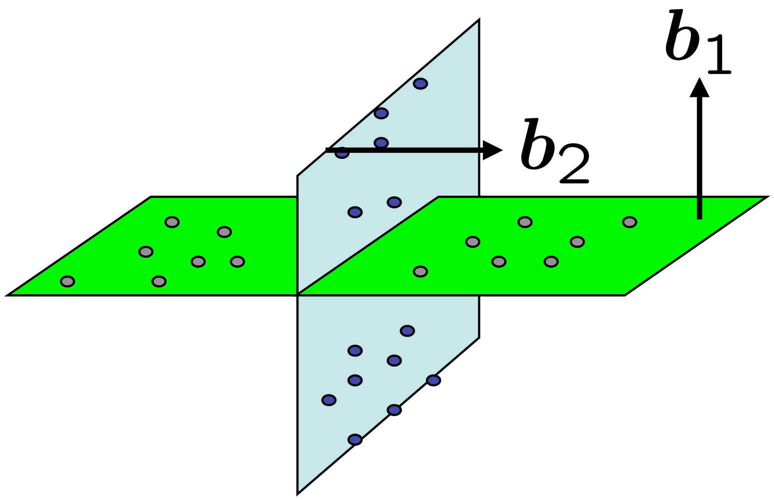
$$\boldsymbol{b}^T\boldsymbol{x} = b_1x_1 + b_2x_2 + b_3x_3 = 0$$



- Two subspaces

$$(\boldsymbol{b}_1^T\boldsymbol{x} = 0) \text{ or } (\boldsymbol{b}_2^T\boldsymbol{x} = 0)$$

$$p_2(\boldsymbol{x}) = (\boldsymbol{b}_1^T\boldsymbol{x})(\boldsymbol{b}_2^T\boldsymbol{x}) = 0$$



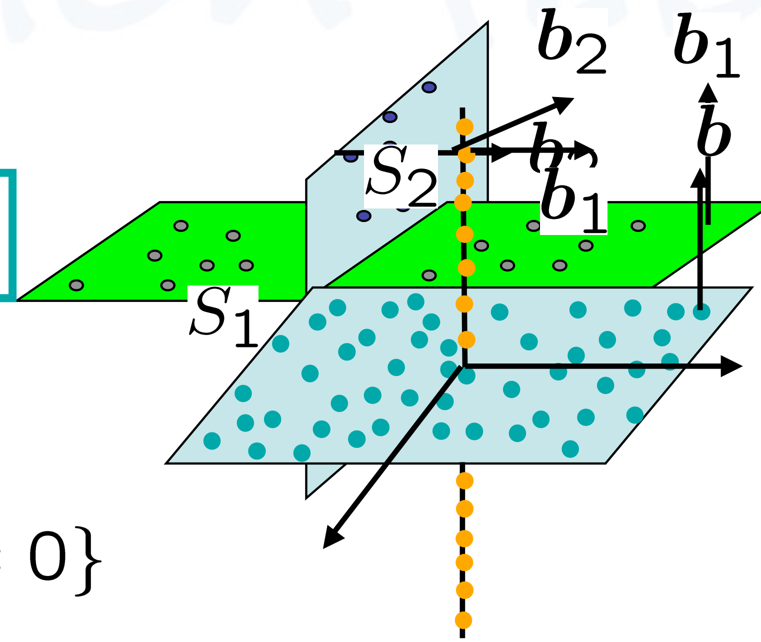
- A union of n subspaces can be represented with a set of homogeneous polynomials of degree n

R. Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA), CVPR 2003.
Vidal, Ma, Piazzi, Sastry. A new GPCA Algorithm for Clustering Subspaces by Fitting, Differentiating and Dividing Polynomials, CVPR 04.
R. Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA), PAMI 2005.

GPCA: Representing *n* Subspaces

- Two planes $(\boldsymbol{b}_1^T \boldsymbol{x} = 0)$ or $(\boldsymbol{b}_2^T \boldsymbol{x} = 0)$

$$p_2(\boldsymbol{x}) = (\boldsymbol{b}_1^T \boldsymbol{x})(\boldsymbol{b}_2^T \boldsymbol{x}) = 0$$



- One plane and one line
  - Plane: $S_1 = \{\boldsymbol{x} : \boldsymbol{b}^T \boldsymbol{x} = 0\}$
  - Line: $S_2 = \{\boldsymbol{x} : \boldsymbol{b}_1^T \boldsymbol{x} = \boldsymbol{b}_2^T \boldsymbol{x} = 0\}$

$$S_1 \cup S_2 = \{\boldsymbol{x} : (\boldsymbol{b}^T \boldsymbol{x} = 0) \text{ or } (\boldsymbol{b}_1^T \boldsymbol{x} = \boldsymbol{b}_2^T \boldsymbol{x} = 0\}$$

De Morgan's rule

$$S_1 \cup S_2 = \{\boldsymbol{x} : (\boldsymbol{b}^T \boldsymbol{x})(\boldsymbol{b}_1^T \boldsymbol{x}) = 0 \text{ and } (\boldsymbol{b}^T \boldsymbol{x})(\boldsymbol{b}_2^T \boldsymbol{x}) = 0\}$$

- A union of n subspaces can be represented with a set of homogeneous polynomials of degree n

Vidal, Ma, Piazzi, Sastry. A new GPCA Algorithm for Clustering Subspaces by Fitting, Differentiating and Dividing Polynomials, CVPR 04.
Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA), PAMI 2005.

GPCA: Fitting Polynomials to Data Points

- Polynomials are linear in their coefficients

$$(\boldsymbol{b}_1^\top \boldsymbol{x})(\boldsymbol{b}_2^\top \boldsymbol{x}) = c_1 x_1^2 + c_2 x_1 x_2 + c_3 x_2^2 = \boldsymbol{c}^\top \nu_n(\boldsymbol{x}) = 0$$

- Coefficients can be computed linearly from the nullspace of the embedded data matrix
  - Solve using least squares
  - N = #data points

$$L_n \boldsymbol{c} = \begin{bmatrix} \nu_n(\boldsymbol{x}_1)^\top \\ \vdots \\ \nu_n(\boldsymbol{x}_N)^\top \end{bmatrix} \boldsymbol{c} = \boldsymbol{0}$$

- Number of subspaces can be found from rank of embedded data matrix

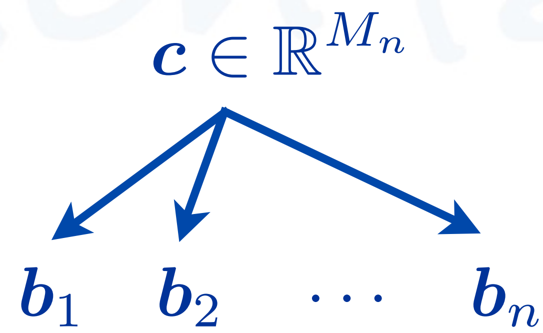
$$n = \min\{i : L_i \text{ drops rank}\}$$

R. Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA), CVPR 2003.
Vidal, Ma, Piazzi, Sastry. A new GPCA Algorithm for Clustering Subspaces by Fitting, Differentiating and Dividing Polynomials, CVPR 04.
R. Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA), PAMI 2005.

GPCA Algorithm by Polynomial Factorization

- Basis for each subspace

$$\boldsymbol{c}^T \nu_n(\boldsymbol{x}) = (\boldsymbol{b}_1^T \boldsymbol{x}) \cdots (\boldsymbol{b}_n^T \boldsymbol{x})$$



- Polynomial Factorization Algorithm
  - Find roots of polynomial of degree n in one variable
  - Solve D-2 linear systems in n variables

- Problems
  - Computing roots may be sensitive to noise
  - The estimated polynomial may not perfectly factor with noisy data

R. Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA), CVPR 2003.

GPCA Algorithm Polynomial Differentiation

$$\boldsymbol{c} \in \mathbb{R}^{M_n}$$

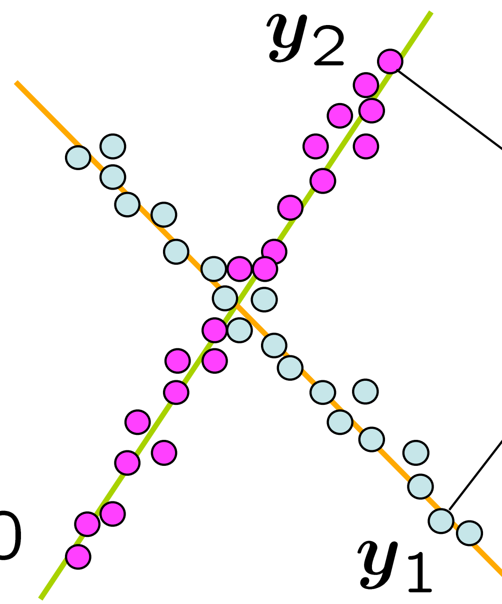
$$\boldsymbol{b}_1 \quad \boldsymbol{b}_2 \quad \dots \quad \boldsymbol{b}_n$$

$$\boldsymbol{b}_i = Dp_n(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{y}_i} \qquad \boldsymbol{y}_i \in S_i$$

$$p_n(\boldsymbol{x}) = 0$$

$$\boldsymbol{b}_1^T \boldsymbol{x} = 0$$

$$\boldsymbol{b}_2^T \boldsymbol{x} = 0$$



$$\boldsymbol{b}_2 \sim Dp_n(\boldsymbol{y}_2)$$

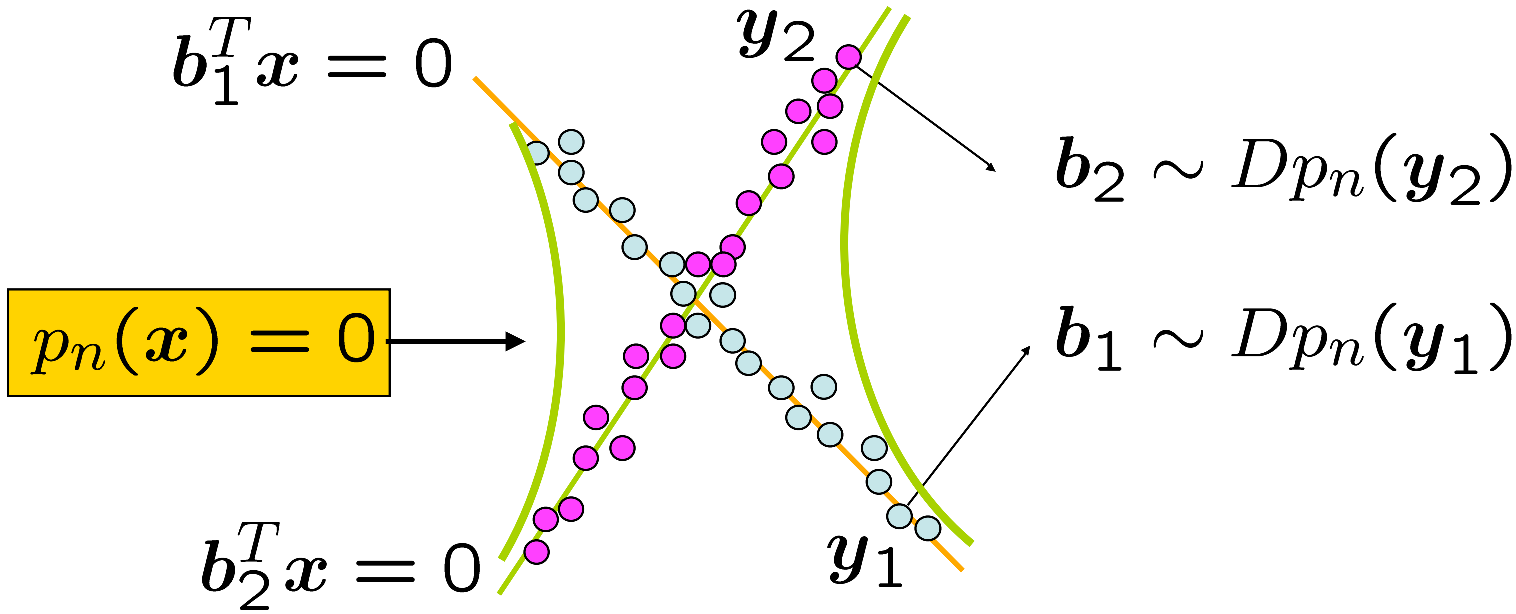
$$\boldsymbol{b}_1 \sim Dp_n(\boldsymbol{y}_1)$$

- To learn a mixture of subspaces we just need one positive example per class

Vidal, Ma, Piazzi, Sastry. A new GPCA Algorithm for Clustering Subspaces by Fitting, Differentiating and Dividing Polynomials, CVPR 04.
R. Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA), PAMI 2005.

GPCA Algorithm Polynomial Differentiation

- With noise and outliers
  – Polynomials may not be a perfect union of subspaces



  – Normals can estimated correctly by choosing points optimally
- Distance to closest subspace without knowing segmentation?

$$\|x - \tilde{x}\| = \sqrt{\frac{|p_n(x)|}{\|Dp_n(x)\|}} + O\left(\|x - \tilde{x}\|^2\right)$$

Vidal, Ma, Piazzi, Sastry. A new GPCA Algorithm for Clustering Subspaces by Fitting, Differentiating and Dividing Polynomials, CVPR 04.
R. Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA), PAMI 2005..

GPCA: Algorithm for Hyperplane Clustering

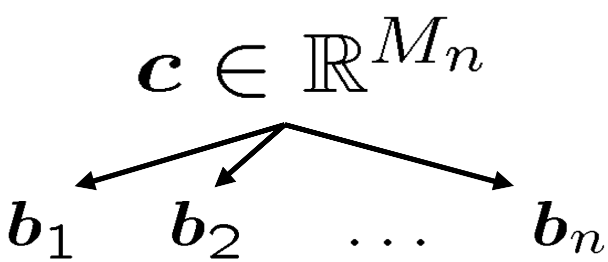
- Coefficients of the polynomial can be computed from null space of embedded data matrix
  - Solve using least squares
  - N = #data points

$$L_n \boldsymbol{c} = \begin{bmatrix} \nu_n(\boldsymbol{x}_1)^T \\ \vdots \\ \nu_n(\boldsymbol{x}_N)^T \end{bmatrix} \boldsymbol{c} = 0$$

- Number of subspaces can be computed from the rank of embedded data matrix

$$n = \min\{i : \text{rank}(L_i) = M_i - 1\}$$

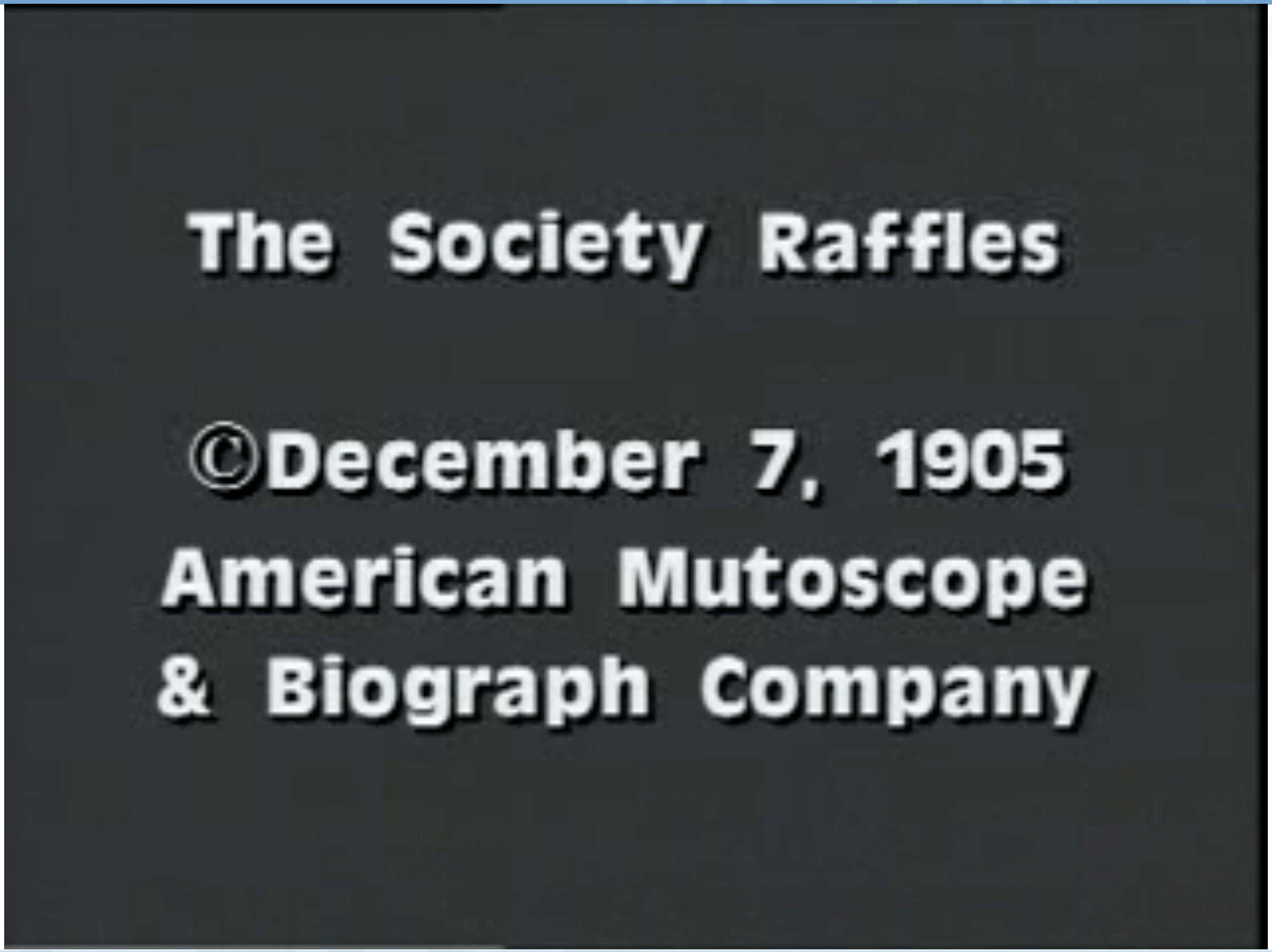
- Normal to the subspaces $\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots \boldsymbol{b}_n$ can be computed from the derivatives of the polynomial



$$\boldsymbol{b}_i = Dp_n(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{y}_i} \qquad \boldsymbol{y}_i \in S_i$$

Vidal, Ma, Piazzi, Sastry. A new GPCA Algorithm for Clustering Subspaces by Fitting, Differentiating and Dividing Polynomials, CVPR 04.
R. Vidal, Y. Ma, S. Sastry. Generalized Principal Component Analysis (GPCA), PAMI 2005.

Temporal Video Segmentation by GPCA



Temporal Video Segmentation by GPCA

- Empty living room
- Middle-aged man enters
- Woman enters
- Young man enters, introduces the woman and leaves
- Middle-aged man flirts with woman and steals her tiara
- Middle-aged man checks the time, rises and leaves
- Woman walks him to the door
- Woman returns to her seat
- Woman misses her tiara
- Woman searches her tiara
- Woman sits and dismays

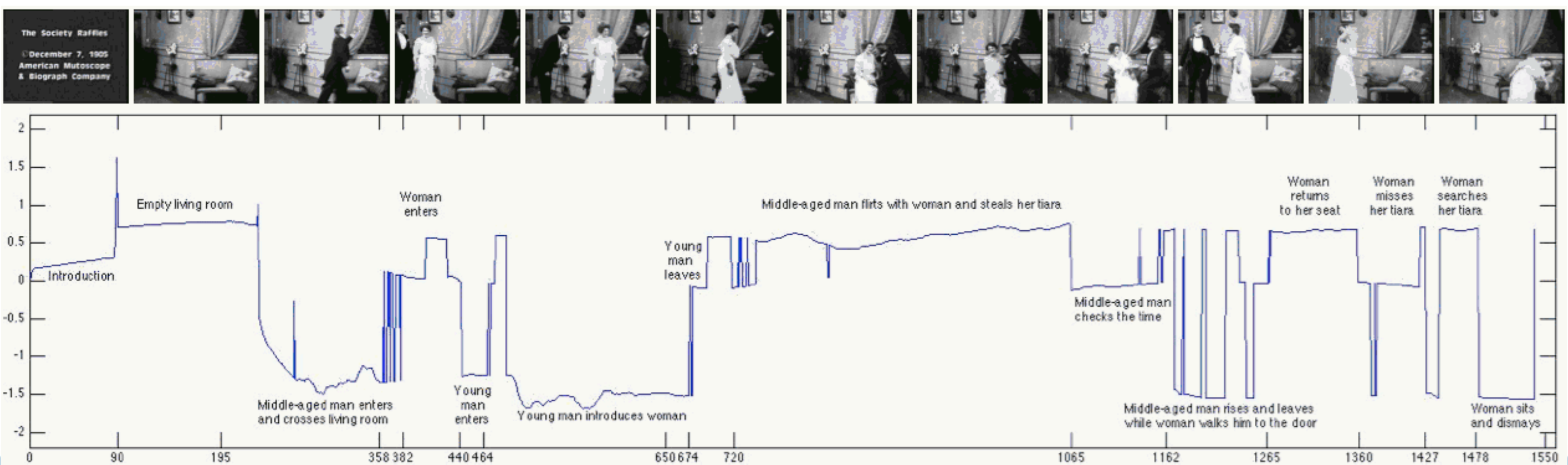


Fig. 5. Temporal segmentation of a scene from the movie *The society raffles*. The top row shows several key frames from the scene displaying different events. The bottom row shows the temporal evolution of the parameter $\hat{c}_t$ as a function of time.

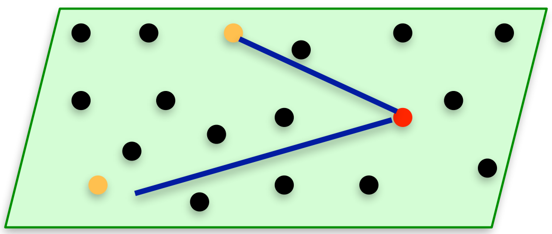
JHU vision lab

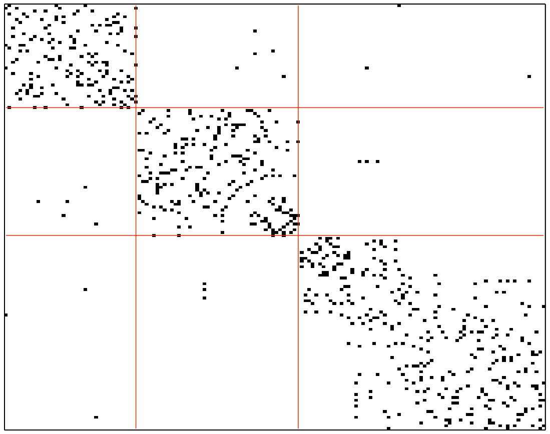
Sparse Subspace Clustering (SSC)

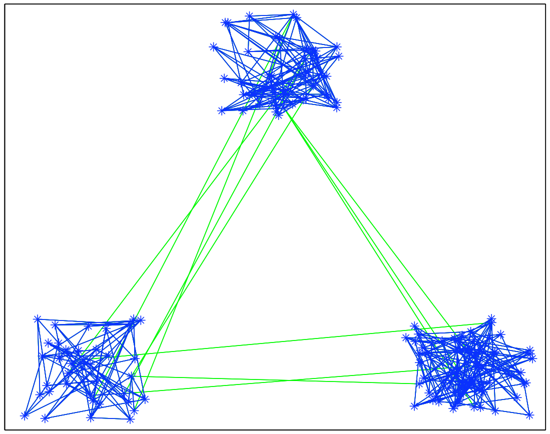
Ehsan Elhamifar and René Vidal

Sparse Subspace Clustering: Spectral Clustering

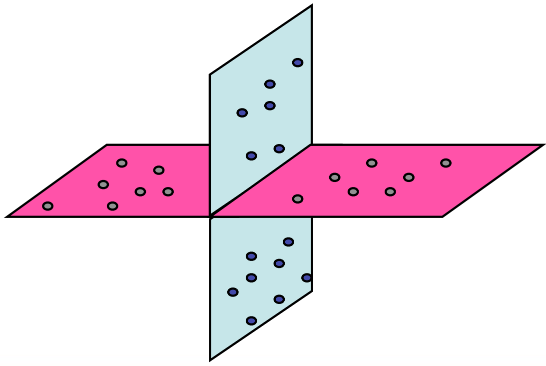
- Spectral clustering
  - Represent data points as nodes in graph G
  - Connect nodes i and j with weight c_{ij}
  - Infer clusters from Laplacian of G





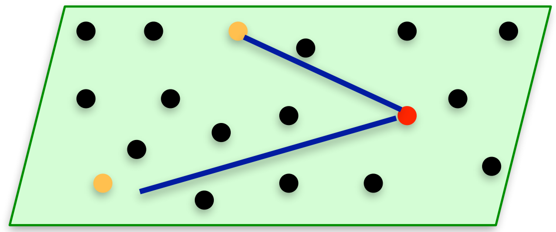


- How to define a good affinity matrix C for subspaces?
  - points in the same subspace: $c_{ij} \neq 0$
  - points in different subspaces: $c_{ij} = 0$

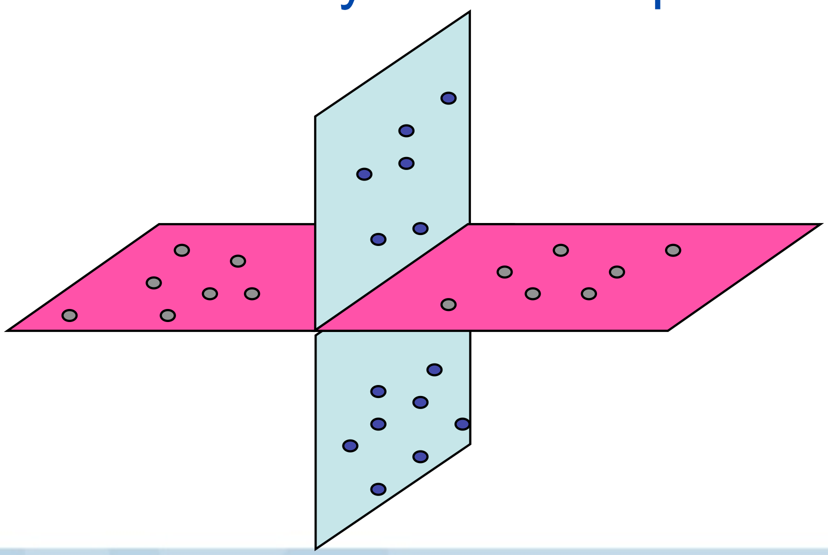


Sparse Subspace Clustering: Spectral Clustering

- Spectral curvature clustering (SCC) (Chen-Lerman '08)
  - Define multiway similarity as normalized volume of d+1 points



- Local subspace affinity (LSA) (Yan-Pollefeys '06)
  - Use the angles between locally fitted subspaces as similarity

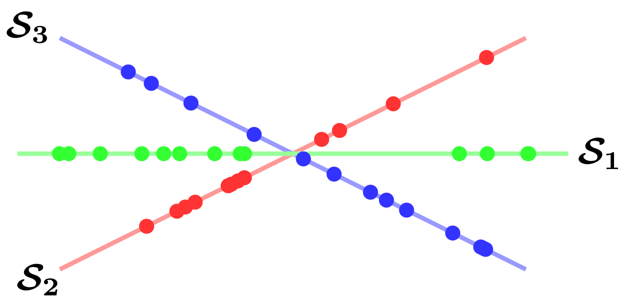


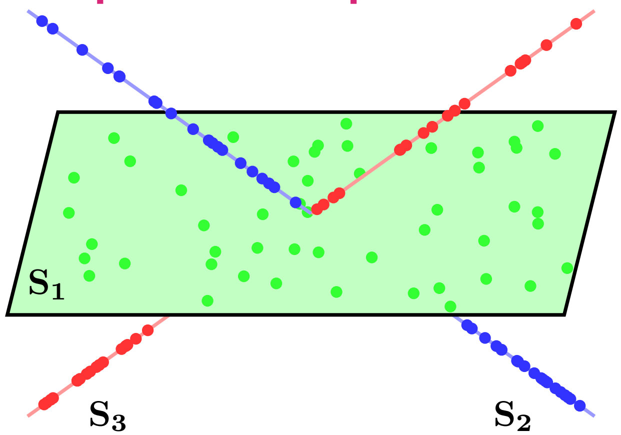
Sparse Subspace Clustering: Intuition

- Data in a union of subspaces are self-expressive

$$\boldsymbol{y}_i = \sum_{j=1}^{N} c_{ji} \boldsymbol{y}_j \implies \boldsymbol{y}_i = Y\boldsymbol{c}_i \implies Y = YC$$

- Union of subspaces admits subspace-sparse representation





- Under what conditions on the subspaces and the data
  - L0 = subspace sparse?
  - L1 = subspace sparse? $P_1 : \min \|\boldsymbol{c}_i\|_1 \ \text{ s.t. } \ \boldsymbol{y}_i = Y\boldsymbol{c}_i, \quad c_{ii} = 0$

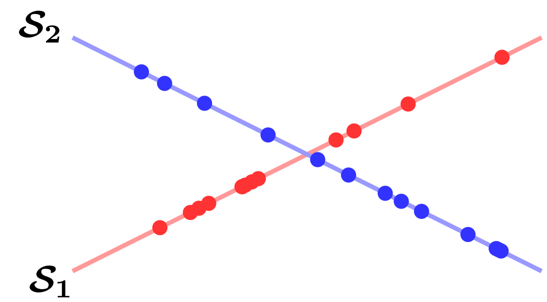
E. Elhamifar and R. Vidal. Sparse Subspace Clustering. CVPR 2009.
E. Elhamifar and R. Vidal. Clustering Disjoint Subspaces via Sparse Representation. ICASSP 2010.
E. Elhamifar and R. Vidal. Sparse Subspace Clustering: Algorithm, Theory and Applications. TPAMI 2013.

Center for Imaging Science

Sparse Subspace Clustering: Noiseless Data

- **Theorem 1**: P_1 recovers a subspace-sparse representation if
  - Subspaces are independent:

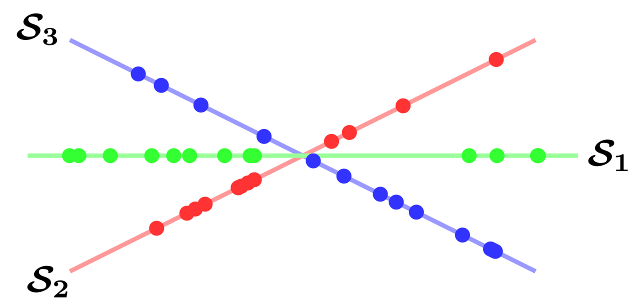
$$\dim\left(\bigoplus_{i=1}^{n} S_i\right) = \sum_{i=1}^{n} \dim(S_i)$$



$$P_1 : \min \|\boldsymbol{c}_i\|_1 \ \text{ s.t. } \ \boldsymbol{y}_i = Y\boldsymbol{c}_i, \ \ c_{ii} = 0$$

Sparse Subspace Clustering: Noiseless Data

- **Theorem 2**: P_1 recovers a subspace-sparse representation if
  - Subspaces are disjoint: $S_i \cap S_j = \{0\}$
  - Subspaces are sufficiently well separated and data are sufficiently well distributed



$$\max_{\text{rank}(\bar{\boldsymbol{Y}}_i)=d_i} \sigma_{d_i}(\bar{\boldsymbol{Y}}_i) > \sqrt{d_i} \max_{j \neq i} \cos(\theta_{ij})$$

- θ_{ij} is the smallest subspace angle between subspaces i and j
  - subspace angles decrease ➡ harder recovery
- $\sigma_{d_i}(\bar{Y}_i)$ is the smallest singular value in each subspace
  - data closer to a degenerate subspace ➡ harder recovery

$$P_1 : \min \|\boldsymbol{c}_i\|_1 \quad \text{s.t.} \quad \boldsymbol{y}_i = Y\boldsymbol{c}_i, \quad c_{ii} = 0$$

Sparse Subspace Clustering: Noiseless Data

- **Theorem 3**:
  - *n* d-dimensional subspaces chosen independently, uniformly at random
  - *r d + 1* points per subspace chosen independently, uniformly at random
  - P_1 recovers a subspace-sparse representation with high probability if

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

$$P_1 : \min \|\boldsymbol{c}_i\|_1 \quad \text{s.t.} \quad \boldsymbol{y}_i = Y\boldsymbol{c}_i, \quad c_{ii} = 0$$

A geometric analysis of subspace clustering with outliers. M. Soltanolkotabi and E. J. Candes. Annals of Statistics 40(4), 2195–2238.

Sparse Subspace Clustering: Data with Outliers

- **Assumptions**
  - n d-dimensional subspaces chosen independently, uniformly at random
  - $r\,d + 1$ inliers per subspace chosen independently, uniformly at random
  - $N_{outliers}$ outliers chosen independently and uniformly at random
  - Declare point i as an outlier if the solution to P1 satisfies

$$\|\boldsymbol{c}_i\|_1 > \lambda(\gamma)\sqrt{D}$$

- **Theorem 4**:
  - P_1 correctly detects all outliers with high probability if

$$N_{outliers} < \frac{1}{D} e^{c\sqrt{D}} - N_{inliers}$$

  - P_1 does not detect any inlier as an outlier if

$$P_1 : \min \|\boldsymbol{c}_i\|_1 \quad \text{s.t.} \quad \boldsymbol{y}_i = Y\boldsymbol{c}_i, \quad c_{ii} = 0$$

A geometric analysis of subspace clustering with outliers. M. Soltanolkotabi and E. J. Candes. Annals of Statistics 40(4), 2195–2238.

Sparse Subspace Clustering: Corrupted Data

- When the data are corrupted with noise $\tilde{\mathbf{y}} = \mathbf{y} + \mathbf{e}$

$$\min \|\mathbf{c}_i\|_1 + \mu \, \|\mathbf{y}_i - Y\mathbf{c}_i\|_2$$

- When the data have missing entries
  - Let $I \subset \{1, \dots, D\}$ be the indices of the missing entries in $\mathbf{y} \in \mathbb{R}^D$
  - Form $\tilde{\mathbf{y}} \in \mathbb{R}^{D-|I|}$ and $\tilde{Y} \in \mathbb{R}^{D-|I| \times N}$ by eliminating rows of $\mathbf{y}$ and Y indexed by I, and solve the same optimization problems

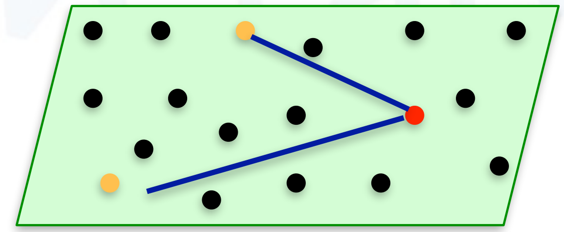
- When the data are corrupted with outlying entries
  - Let $\tilde{\mathbf{y}} = Y\mathbf{c} + \mathbf{e} = \begin{bmatrix} Y & I_D \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{e} \end{bmatrix}$ be corrupted by a vector $\mathbf{e} \in \mathbb{R}^D$
  - The vector $\begin{bmatrix} \mathbf{c}^\top & \mathbf{e}^\top \end{bmatrix}^\top$ is still sparse and can be recovered from

$$\min \, \left\|\begin{bmatrix} \mathbf{c} \\ \mathbf{e} \end{bmatrix}\right\|_1 + \mu \, \left\|\tilde{\mathbf{y}} - \begin{bmatrix} Y & I_D \end{bmatrix}\begin{bmatrix} \mathbf{c} \\ \mathbf{e} \end{bmatrix}\right\|_2$$

S. Rao, R. Tron, R. Vidal and Y. Ma. Robust Motion Segmentation, CVPR 2008, PAMI 2009.
E. Elhamifar and R. Vidal. Sparse Subspace Clustering. CVPR 2009.
E. Elhamifar and R. Vidal. Clustering Disjoint Subspaces via Sparse Representation. ICASSP 2010.

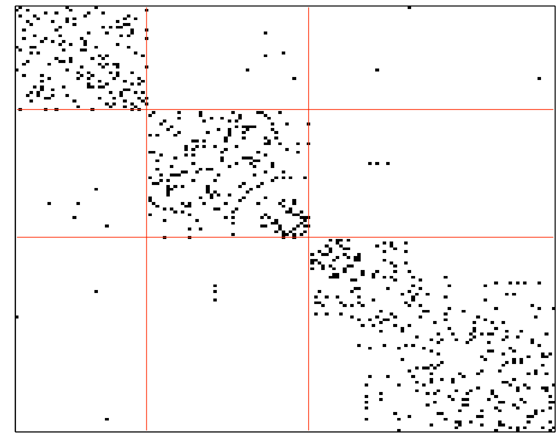
Sparse Subspace Clustering: Algorithm

- Represent data points as nodes in graph G



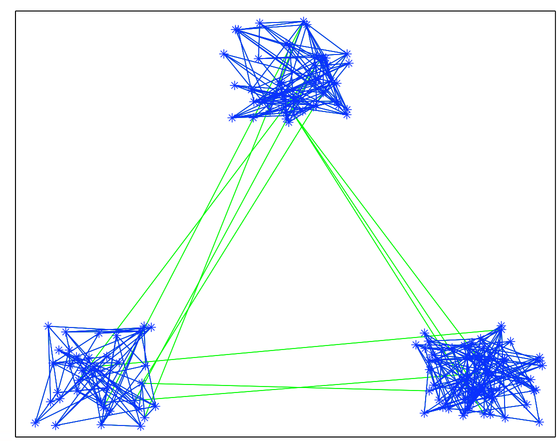
- Find the sparse coefficient vectors $\{\mathbf{c}_i\}_{i=1}^N$

$$\min \|\mathbf{c}_i\|_1 + \mu \|\mathbf{y}_i - Y\mathbf{c}_i\|_2$$



- Connect nodes i and j by an edge with weight

$$|c_{ij}| + |c_{ji}|$$



- Spectral clustering: apply K-means to the smallest eigenvectors of the Laplacian of G

Center for IMAGING SCIENCE

Low Rank Subspace Clustering (LRSC)

Paolo Favaro and René Vidal

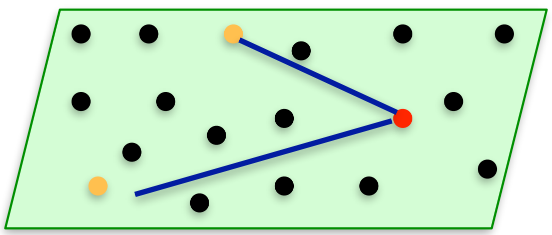
JHU vision lab

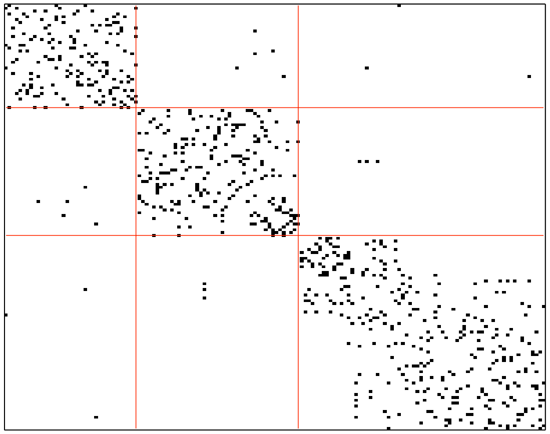
THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

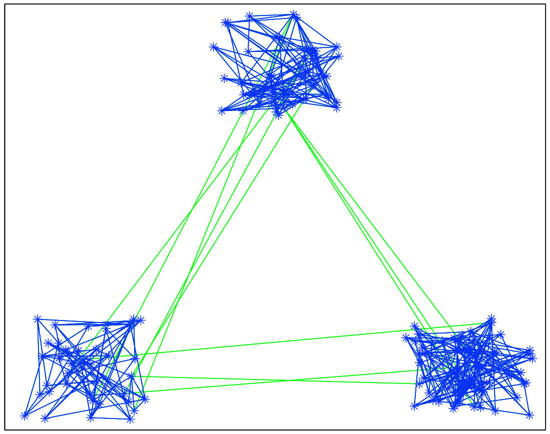
Center for Imaging Science

Sparse Subspace Clustering: Spectral Clustering

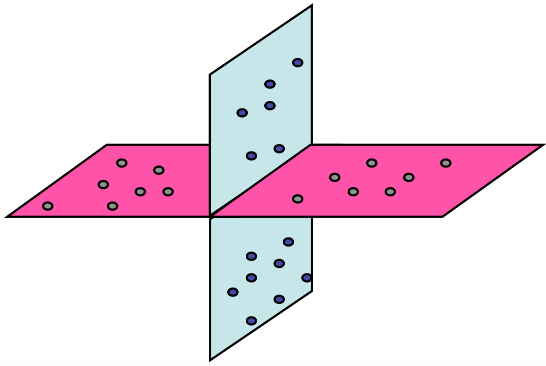
- Spectral clustering
  - Represent data points as nodes in graph G
  - Connect nodes i and j with weight c_{ij}
  - Infer clusters from Laplacian of G







- How to define a good affinity matrix C for subspaces?
  - points in the same subspace: $c_{ij} \neq 0$
  - points in different subspaces: $c_{ij} = 0$

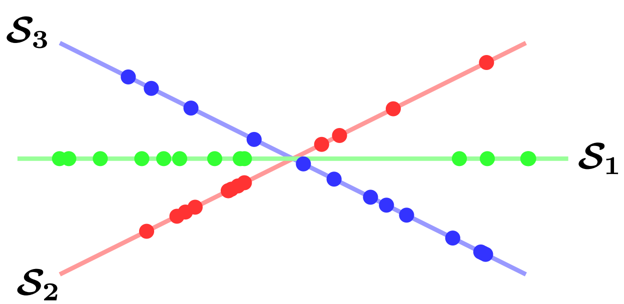


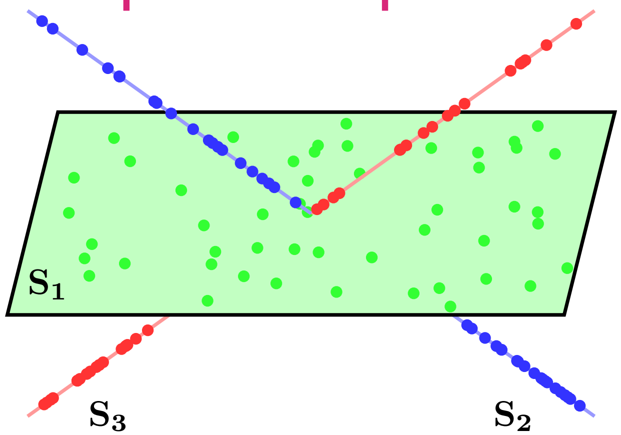
Sparse Subspace Clustering: Intuition

- Data in a union of subspaces are self-expressive

$$\boldsymbol{y}_i = \sum_{j=1}^{N} c_{ji} \boldsymbol{y}_j \implies \boldsymbol{y}_i = Y \boldsymbol{c}_i \implies Y = YC$$

- Union of subspaces admits subspace-sparse representation





- Sparse Subspace Clustering

$$P_1 : \min \|\boldsymbol{c}_i\|_1 \quad \text{s.t.} \quad \boldsymbol{y}_i = Y \boldsymbol{c}_i, \quad c_{ii} = 0$$

E. Elhamifar and R. Vidal. Sparse Subspace Clustering. CVPR 2009.
E. Elhamifar and R. Vidal. Clustering Disjoint Subspaces via Sparse Representation. ICASSP 2010.
E. Elhamifar and R. Vidal. Sparse Subspace Clustering: Algorithm, Theory and Applications. TPAMI 2013.

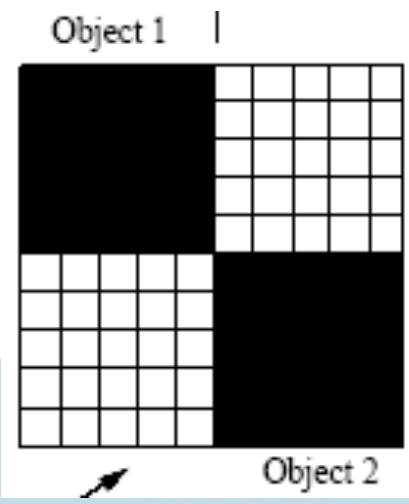
Subspace Clustering by Matrix Factorization

- Data from i-th subspace can be factorized as $Y_i = U_i V_i^\top$

$$Y\Gamma = [Y_1, Y_2, \dots, Y_n] = [U_1, U_2, \dots, U_n] \begin{bmatrix} V_1^\top & & & \\ & V_2^\top & & \\ & & \ddots & \\ & & & V_n^\top \end{bmatrix}$$

- Segmentation of the data can be obtained from
  - Leading singular vector of $Y = \mathcal{U}\Sigma\mathcal{V}^\top$ (Boult and Brown '91)
  - Shape interaction matrix $C = \mathcal{V}\mathcal{V}^\top$ (Costeira & Kanade '95, Gear '94)

- $C_{ij} = 0$ if points i and j lie in two independent subspaces (Kanatani et al. '01, Vidal et al. '08)



Low Rank Subspace Clustering

- Data in a union of subspaces are self-expressive

$$\boldsymbol{y}_i = \sum_{j=1}^{N} c_{ji} \boldsymbol{y}_j \implies \boldsymbol{y}_j = Y \boldsymbol{c}_i \implies Y = YC$$

  - C is sparse
  - C is low-rank

- Low Rank Subspace Clustering (noiseless case)

$$\min_C \|C\|_* \ \text{ s.t. } \ Y = YC \implies \begin{array}{l} Y = \mathcal{U}\Sigma\mathcal{V}^\top \\ C = \mathcal{V}\mathcal{V}^\top \end{array}$$

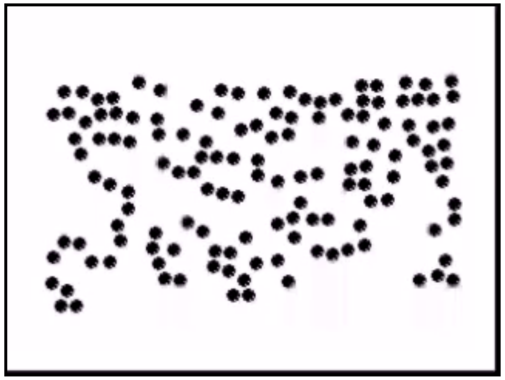
- Low Rank Subspace Clustering (noisy case)

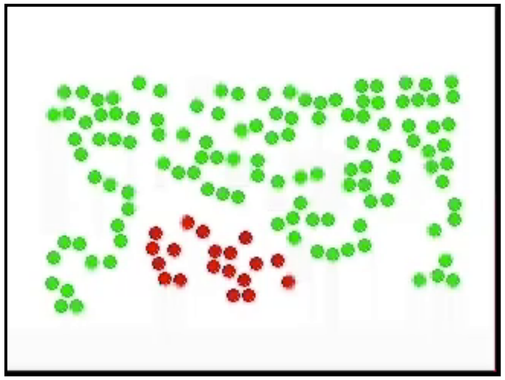
$$\min_C \|C\|_* + \frac{\tau}{2}\|Y - YC\|_F^2 \implies C = \mathcal{V}(I - \frac{1}{\tau}\Sigma^{-2})\mathcal{V}^\top$$

Applications in Computer Vision

Experiments on 3D Motion Segmentation

- Motion segmentation problem
  - Input: multiple images of a scene with multiple rigid-body motions
  - Output: number of motions, motion model parameters, segmentation





- Motion of a rigid-body: 4D subspace (Boult and Brown '91, Tomasi and Kanade '92)
  - P = #points
  - F = #frames

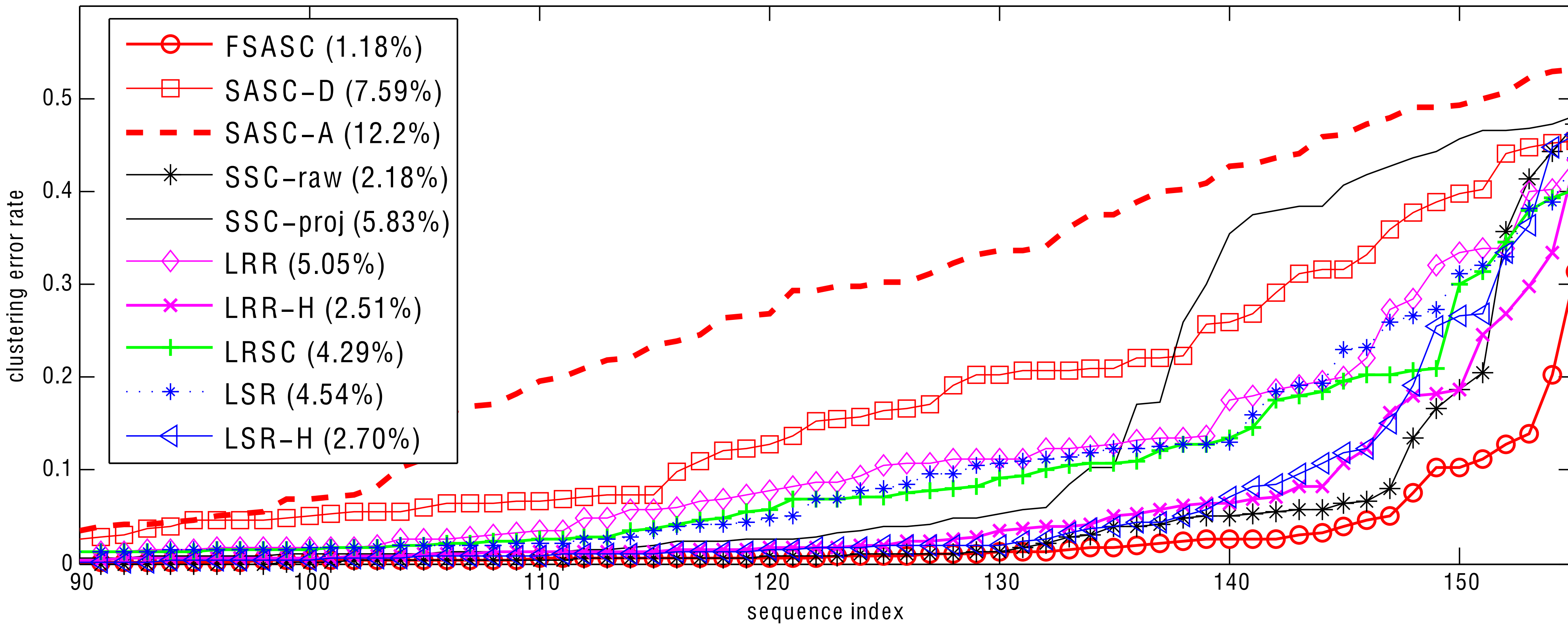
$$\underbrace{\begin{bmatrix} \boldsymbol{x}_{11} & \cdots & \boldsymbol{x}_{1P} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_{F1} & \cdots & \boldsymbol{x}_{FP} \end{bmatrix}}_{2F \times P} = \underbrace{\begin{bmatrix} \boldsymbol{A}_1 \\ \vdots \\ \boldsymbol{A}_F \end{bmatrix}}_{2F \times 4} \underbrace{\begin{bmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_P \end{bmatrix}}_{4 \times P}$$

Vidal et al., ECCV02, IJCV06; Vidal, Ma and Sastry CVPR03, PAMI05; Vidal and Sastry CVPR03; Vidal and Ma ECCV04, JMIV06; Vidal and Hartley, CVPR04; Tron and Vidal, CVPR07; Li et al. CVPR07; Goh and Vidal CVPR07; Vidal and Hartley, PAMI08; Vidal, Tron and Hartley IJCV08; Rao et al. CVPR 08, PAMI 09; Elhamifar and Vidal, CVPR 09, TPAMI 13; Vidal SPM11; Tsakiris '15

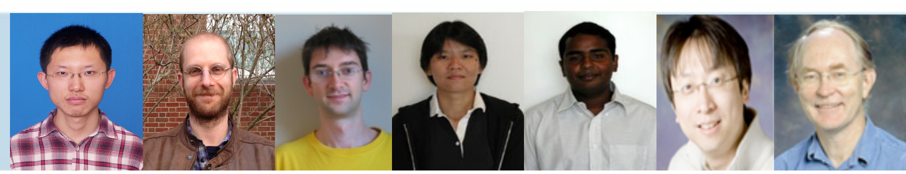
Experiments on 3D Motion Segmentation

- Misclassification rates on Hopkins 155 database

R. Tron and R. Vidal. A Benchmark for the Comparison of 3-D Motion Segmentation Algorithms. CVPR 2007.

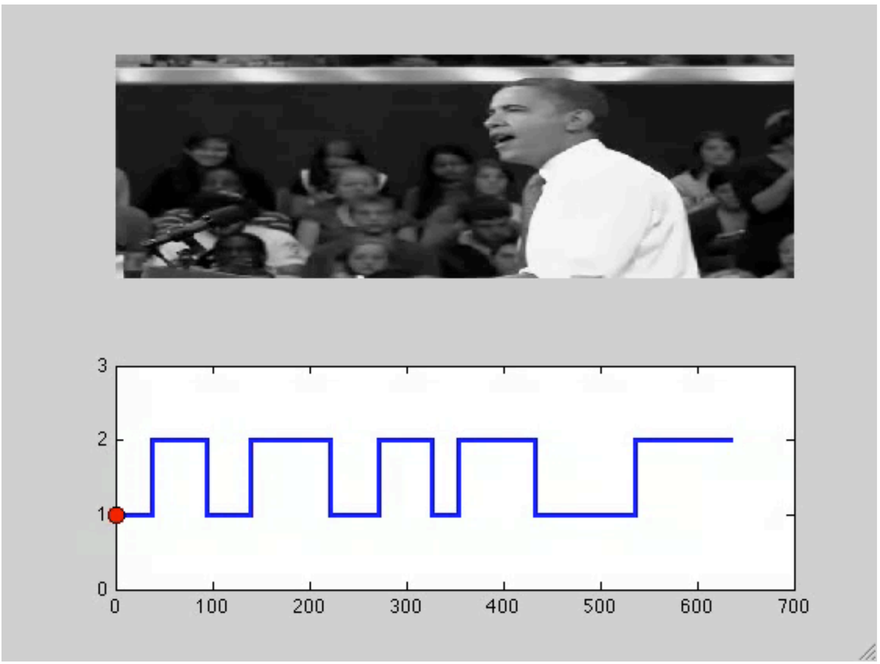


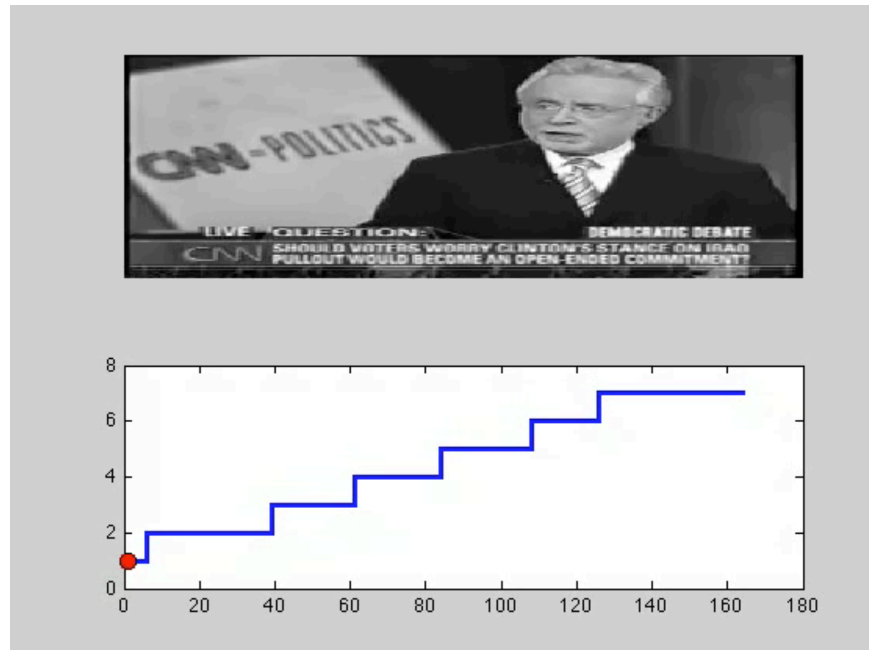
Vidal et al., ECCV02, IJCV06; Vidal, Ma and Sastry CVPR03, PAMI05; Vidal and Sastry CVPR03; Vidal and Ma ECCV04, JMIV06; Vidal and Hartley, CVPR04; Tron and Vidal, CVPR07; Li et al. CVPR07; Goh and Vidal CVPR07; Vidal and Hartley, PAMI08; Vidal, Tron and Hartley IJCV08; Rao et al. CVPR 08, PAMI 09; Elhamifar and Vidal, CVPR 09, TPAMI 13; Vidal SPM11; Tsakiris '15



Experiments on Video Segmentation

- Model each video segment as a low-dimensional subspace
- Cluster video frames into multiple segments

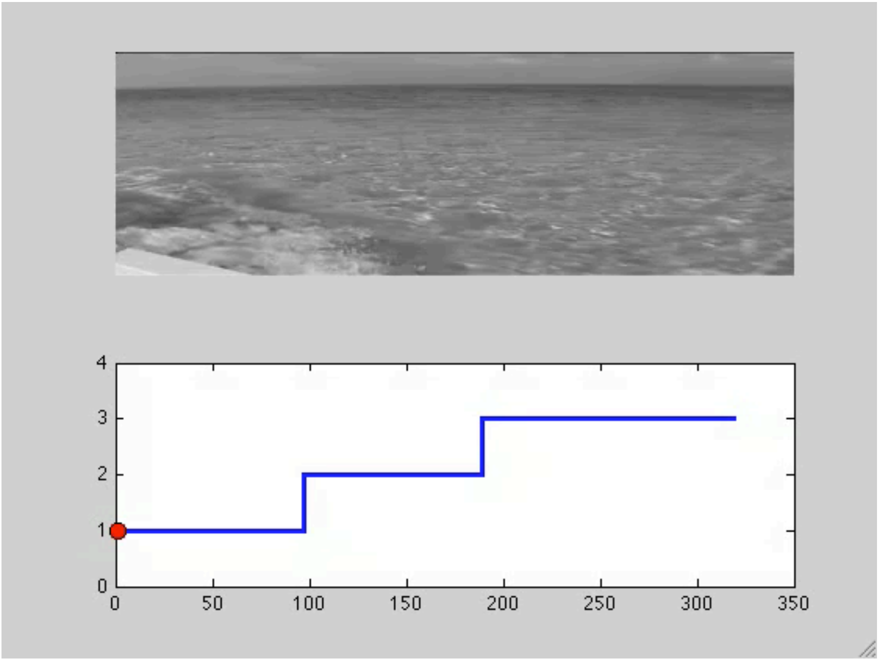


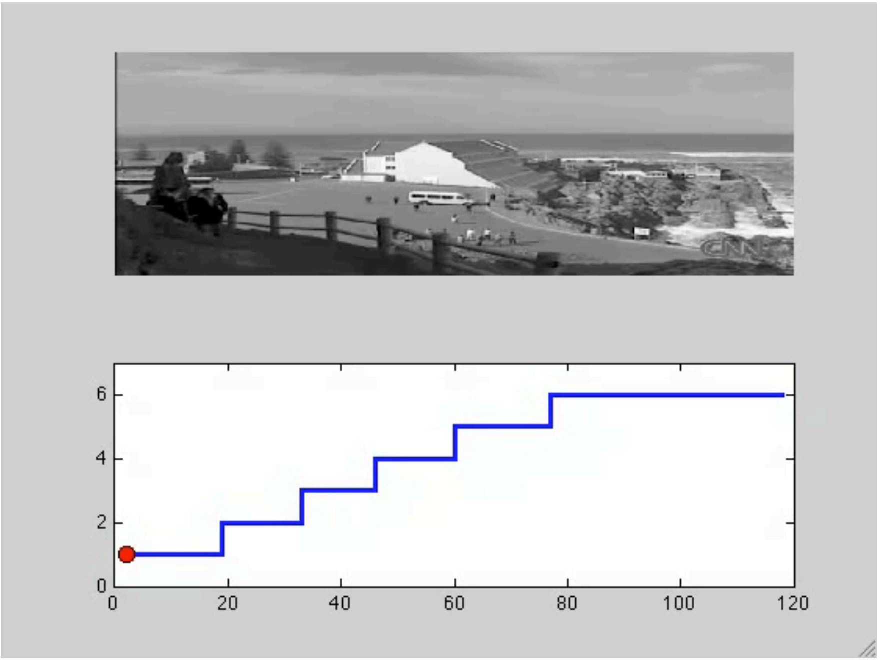


- Advantages
  – SSC easily detects sharp transitions in the video
  – SSC can handle camera motion and scene variations

Experiments on Video Segmentation

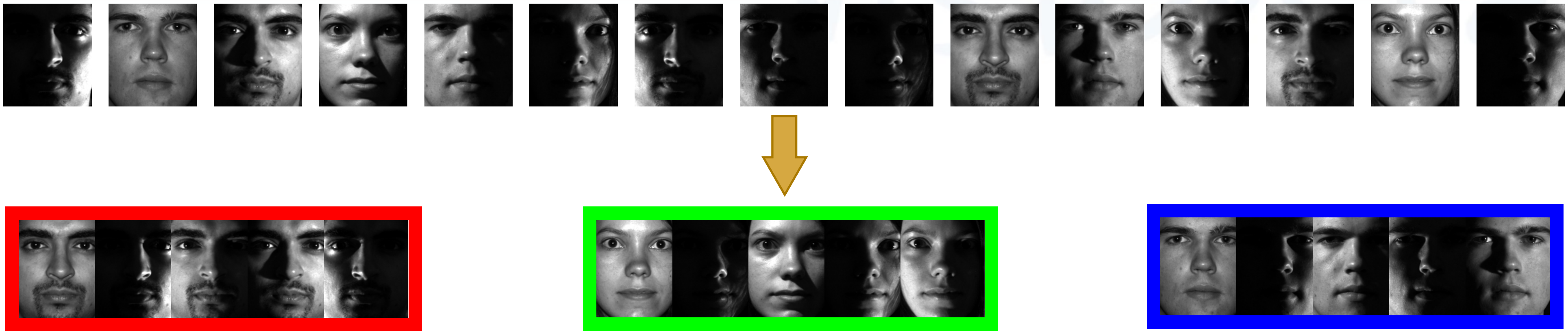
- Model each video segment as a low-dimensional subspace
- Cluster video frames into multiple segments





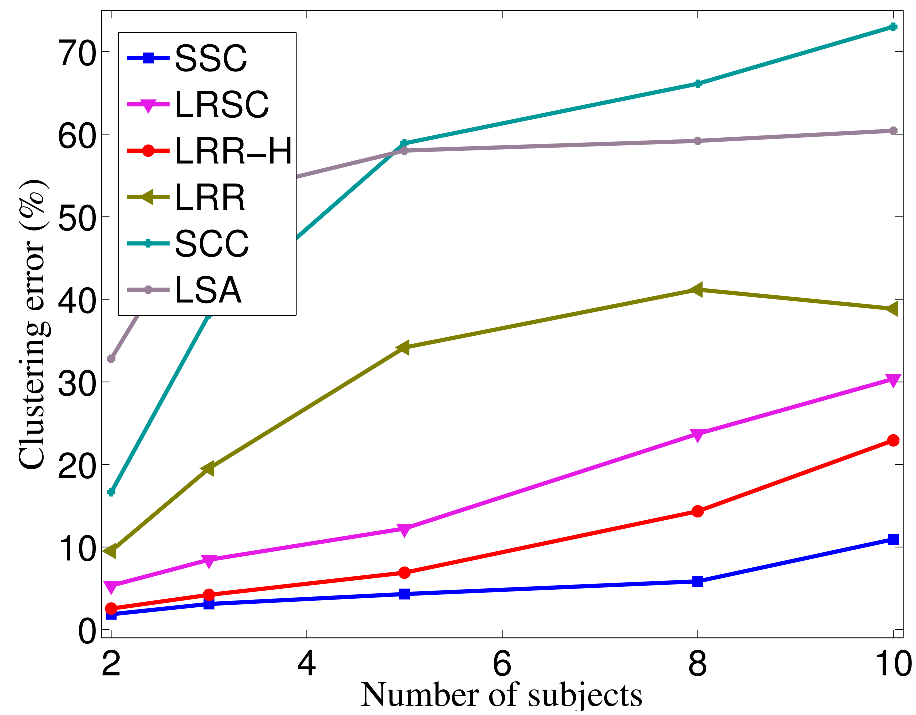
- Advantages
  - SSC easily detects sharp transitions in the video
  - SSC can handle camera motion and scene variations

Experiments on Face Clustering



- Faces under varying illumination
  - 9D subspace
- Extended Yale B dataset
  - 38 subjects
  - 64 images per subject
- Clustering error
  - SSC < 2.0% error for 2 subjects
  - SSC < 11.0% error for 10 subjects

D = 2,016 dimensional data



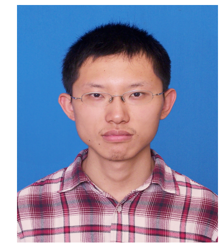
Conclusions

- Many problems in computer vision can be posed as subspace clustering and classification problems
  - Spatial and temporal video segmentation
  - Face clustering under varying illumination
  - Face classification
- These problems can be solved using
  - Generalized Principal Component Analysis (GPCA)
  - Sparse Subspace Clustering (SSC)
  - Low Rank Subspace Clustering (LRSC)
- This algorithms is provably correct when
  - Subspaces are sufficiently separated
  - Data are well distributed within each subspace

What’s Next

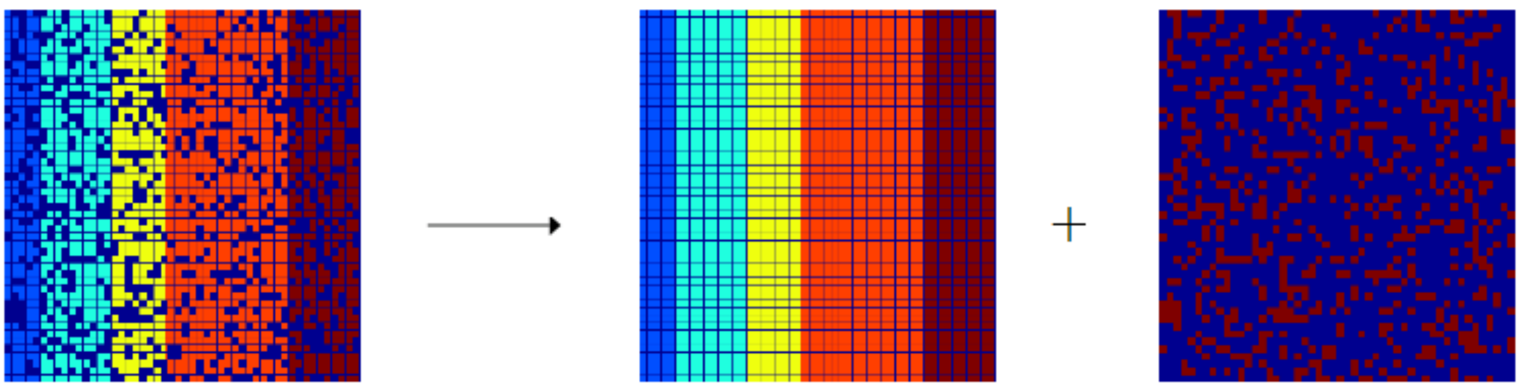
- **Big Data** (Peng ’13, Dyer ’13, You ’15)

	GPCA	SSC	OMP	?
Dimension of the data	10	10,000	10,000	1M
Number of data points	1000	10,000	100,000	1M

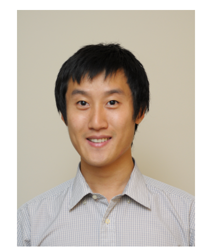


Chong You

- **Missing Data:** (Grubber ’04, Eriksson ’12, Balzano ’12, Pimentel ’14, Candes ’14, Yang’15)



Matrix of corrupted observations *Underlying low-rank matrix* *Sparse error matrix*



Congyuan Yang

Acknowledgements

- Algebraic Methods
  - Y. Ma, S. Sastry, M. Tsakiris



- Sparse and Low Rank
  - E. Elhamifar, P. Favaro, C. You



- Funding
  - Sloan Research Fellowship
  - ONR Young Investigator Award
  - NSF CAREER Award 0447739

- More information/code
  - Vision Lab @ Johns Hopkins University http://www.vision.jhu.edu

Thank You!

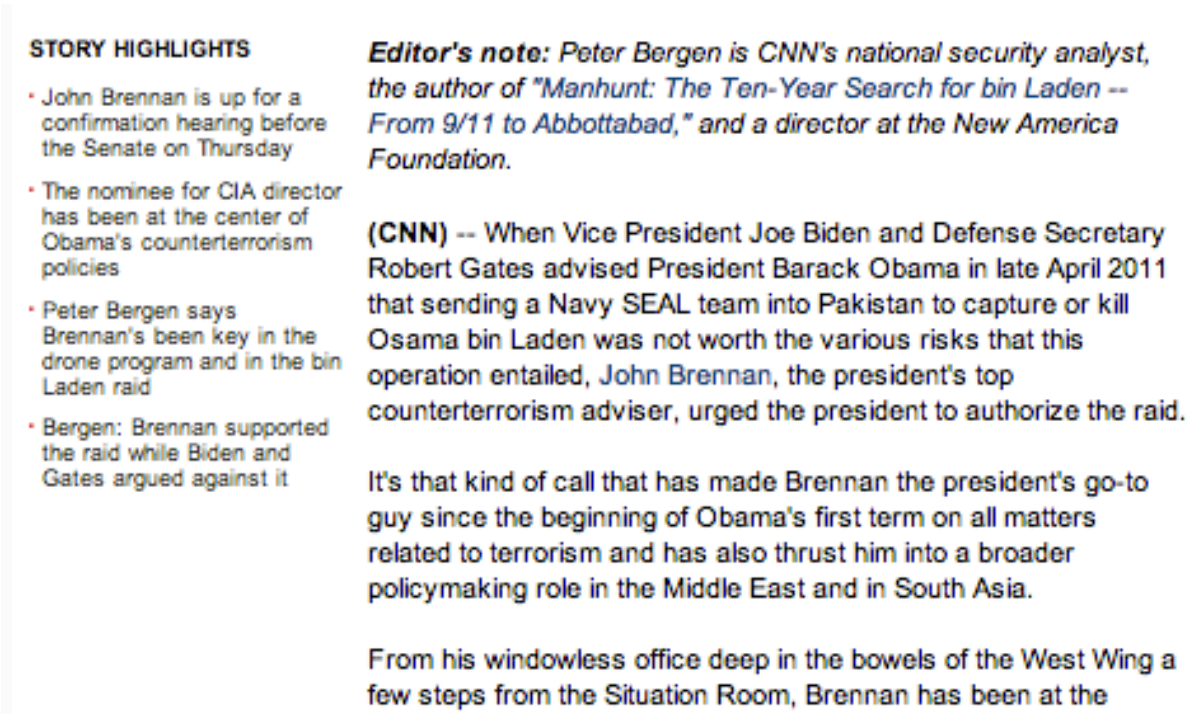
JHU vision lab

See All by Looking at A Few: Sparse Modeling for Finding Data Exemplars

Ehsan Elhamifar (Berkeley), Guillermo Sapiro (Duke) and René Vidal (Hopkins)

Problem Statement

- Given a set of points $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ select a subset of $k \ll N$ points that efficiently represent the whole data set
  - Summarize/visualize text/images/videos



STORY HIGHLIGHTS

- John Brennan is up for a confirmation hearing before the Senate on Thursday
- The nominee for CIA director has been at the center of Obama's counterterrorism policies
- Peter Bergen says Brennan's been key in the drone program and in the bin Laden raid
- Bergen: Brennan supported the raid while Biden and Gates argued against it

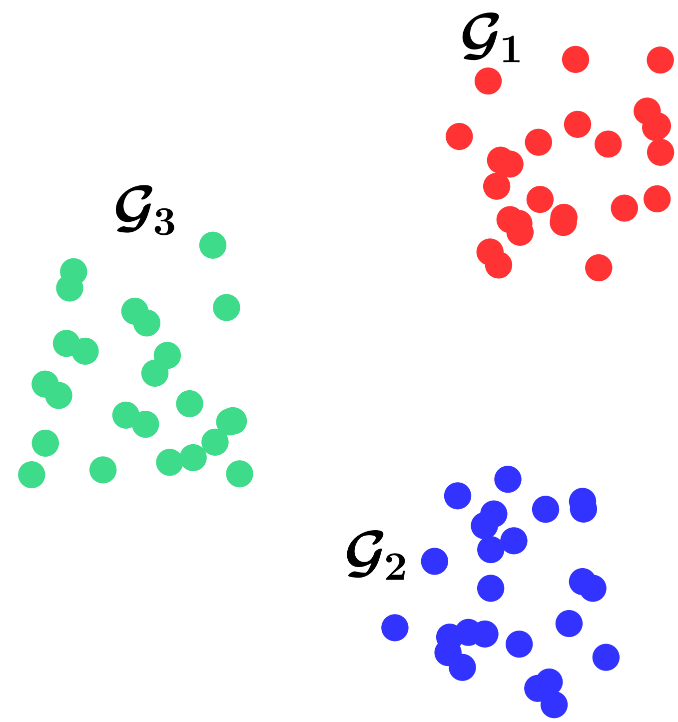
***Editor's note:** Peter Bergen is CNN's national security analyst, the author of "Manhunt: The Ten-Year Search for bin Laden -- From 9/11 to Abbottabad," and a director at the New America Foundation.*

(CNN) -- When Vice President Joe Biden and Defense Secretary Robert Gates advised President Barack Obama in late April 2011 that sending a Navy SEAL team into Pakistan to capture or kill Osama bin Laden was not worth the various risks that this operation entailed, John Brennan, the president's top counterterrorism adviser, urged the president to authorize the raid.

It's that kind of call that has made Brennan the president's go-to guy since the beginning of Obama's first term on all matters related to terrorism and has also thrust him into a broader policymaking role in the Middle East and in South Asia.

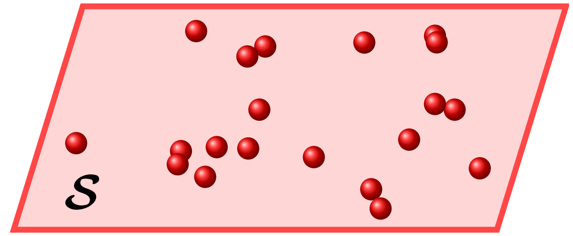
From his windowless office deep in the bowels of the West Wing a few steps from the Situation Room, Brennan has been at the

  - Produce a clustering of the data
  - Reduce computational time and memory requirements of classification algorithms

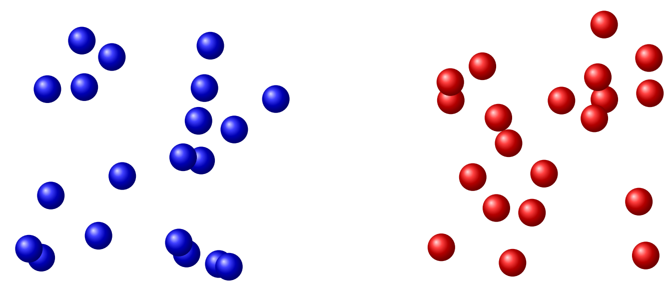


State of the Art

- Methods based on low-rank representations
  - Rank revealing QR [Chan '87, Gu-Eisenstat '96]
  - NNMF [Esser et al. '12, Bittorf et al. '12]
  - CUR [Mahoney-Drineas '09]
  - Randomized/greedy algorithms [Tropp '09, Boutsidis et al. '09, Balzano '10]



- Methods based on clustering
  - Central clustering: k-medoids [Kaufman '87]
  - Set cover optimization [Bien-Tibshirani '11]
  - Affinity propagation [Frey-Duek '06,'07; Givoni et al. '11]

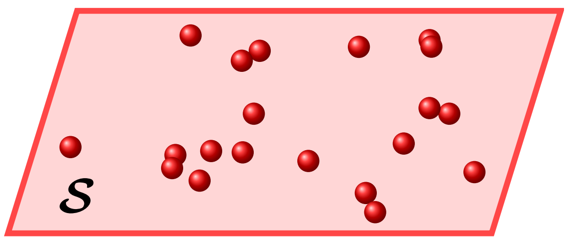


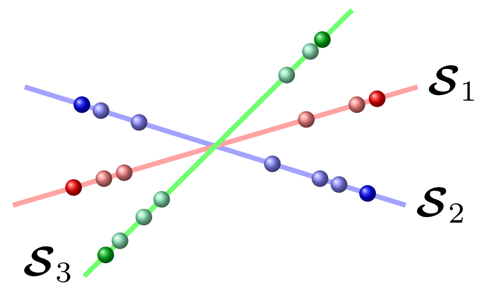
- Challenges
  - Depend on initialization (local minima), return approximate solutions
  - Require prior knowledge about the dimensions, number of groups, etc.

Contributions

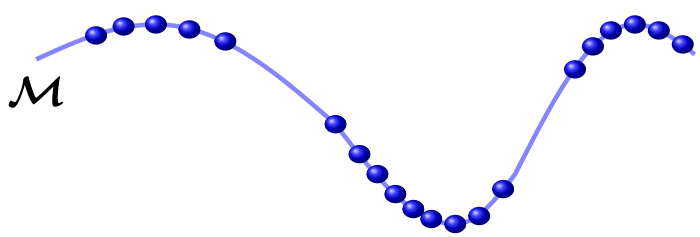
- Goals
  - Develop efficient (convex) algorithms
  - Analyze the geometry of solution
  - Have theoretical guarantees

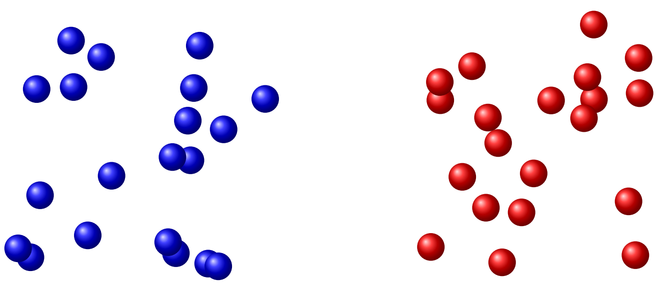
- Part I: Sparse Representation of the Data [1]





- Part II: Sparse Representation of Dissimilarities [2]





[1] E. Elhamifar, G. Sapiro, and R. Vidal. See All by Looking at A Few: Sparse Modeling for Finding Representative Objects. CVPR 2012.
[2] E. Elhamifar, G. Sapiro, and R. Vidal. Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery. NIPS 2012.

Center for IMAGING SCIENCE

Exemplars from Linear Data Relationships

- Input: set of data points $Y = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N] \in \mathbb{R}^{m \times N}$
- Output: set of exemplars $U = [\boldsymbol{y}_{i_1}, \cdots, \boldsymbol{y}_{i_k}] \in \mathbb{R}^{m \times k}$

- Classical PCA: find $U \in \mathbb{R}^{m \times k}$ and $C \in \mathbb{R}^{k \times N}$ such that
  - $\min_{U,C} \|Y - UC\|_F^2 \quad \text{such that} \quad U^\top U = I_k$
  - columns of *U* need not coincide with the data

- Our approach:
  - Choose the smallest number of columns *k* such that

$$[\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N] \approx [\boldsymbol{y}_{i_1}, \cdots, \boldsymbol{y}_{i_k}] \begin{bmatrix} \boldsymbol{c}^{i_1} \\ \vdots \\ \boldsymbol{c}^{i_k} \end{bmatrix}$$

E. Elhamifar, G. Sapiro, and R. Vidal. See All by Looking at A Few: Sparse Modeling for Finding Representative Objects. CVPR 2012.

Exemplars from Linear Data Relationships

- Use the entire data matrix as a dictionary and let the nonzero rows indicate the exemplars

$$[\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N] \approx [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N] \begin{bmatrix} \boldsymbol{c}^1 \\ \vdots \\ \boldsymbol{c}^N \end{bmatrix}$$

- Choose smallest k => minimize number of nonzero rows of C
  [Chen-Huo'05, Tropp'06, Jenatton-Audibert-Bach'11]

$$\min_C \|C\|_{0,q} = \sum_{i=1}^N I(\|\boldsymbol{c}^i\|_q \neq 0) \implies \min_C \|C\|_{1,q} = \sum_{i=1}^N \|\boldsymbol{c}^i\|_q$$

- Find exemplars by solving the convex problem

$$\min_C \quad \|C\|_{1,q} \quad \text{s.t.} \quad Y = YC, \quad 1^\top C = 1^\top \quad (q \geq 1)$$

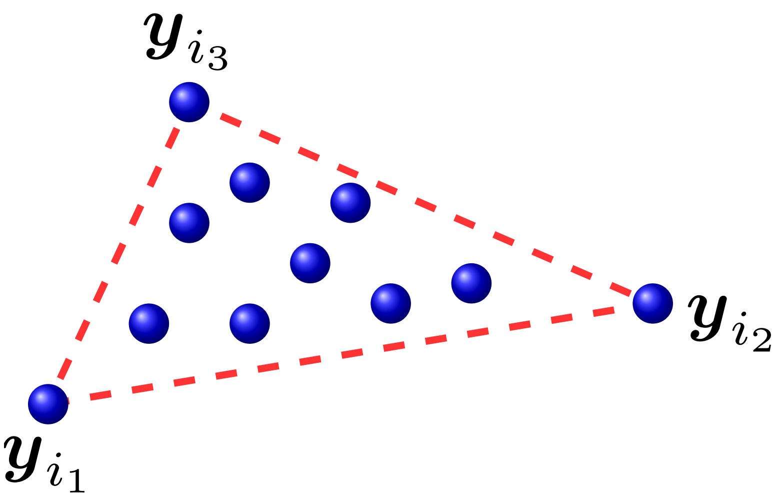
E. Elhamifar, G. Sapiro, and R. Vidal. See All by Looking at A Few: Sparse Modeling for Finding Representative Objects. CVPR 2012.

Theoretical Guarantees

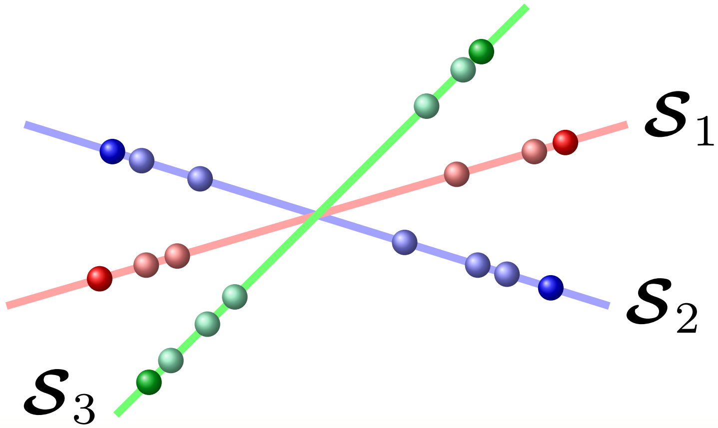
$$\min_{C} \quad \|C\|_{1,q} \quad \text{s.t.} \quad Y = YC, \quad 1^\top C = 1^\top$$

- **Theorem 1:**
  - H = convex hull of data Y
  - k = number of vertices of H
  - Data lie in an affine subspace of dim k-1
  - k nonzero rows of C* = k vertices of H

$$C^* = \Gamma \begin{bmatrix} I_k & \Delta \\ 0 & 0 \end{bmatrix} \qquad \Delta \in [0,1)^k$$



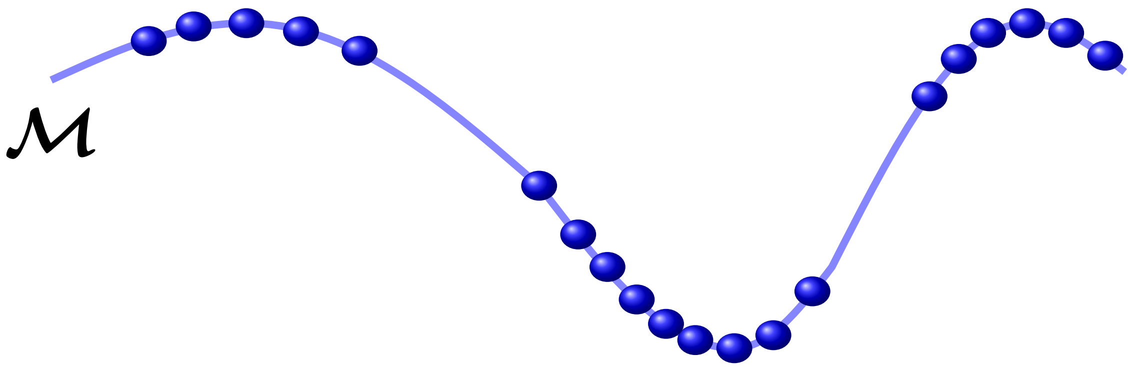
- **Theorem 2:**
  - Data lie in union of independent subspaces
  - Nonzero rows of C* include at least dim(Si) + 1 exemplars for subspace Si



E. Elhamifar, G. Sapiro, and R. Vidal. See All by Looking at A Few: Sparse Modeling for Finding Representative Objects. CVPR 2012.

Beyond Linear Relationships

- Linear relationship model can be restrictive



- Consider dissimilarities between pairs of data points

$$D \triangleq \begin{bmatrix} \boldsymbol{d}_1^\top \\ \vdots \\ \boldsymbol{d}_N^\top \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ \vdots & \vdots & & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{bmatrix} \in \mathbb{R}^{N \times N}$$

  - d_{ij} = cost of encoding point $\boldsymbol{y}_j$ with exemplar $\boldsymbol{y}_i$
  - Euclidean/geodesic distance, KL divergence, etc.
  - Dissimilarities need not come from a metric

E. Elhamifar, G. Sapiro, and R. Vidal. Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery. NIPS 2012.

Exemplars from Pairwise Dissimilarities

- Let $z_{ij} \in \{0,1\}$ denote whether $\boldsymbol{y}_i$ is chosen to encode $\boldsymbol{y}_j$

- The total encoding cost is given by $\operatorname{tr}(D^\top Z) = \sum_{ij} z_{ij} d_{ij}$

- Choose smallest *k* => minimize number of nonzero rows of Z
[Chen-Huo'05, Tropp'06, Jenatton-Audibert-Bach'11]

$$\min_Z \|Z\|_{0,q} = \sum_{i=1}^N I(\|\boldsymbol{z}^i\|_q \neq 0) \implies \min_Z \|Z\|_{1,q} = \sum_{i=1}^N \|\boldsymbol{z}^i\|_q$$

- Find exemplars by solving the convex problem

$$\min_Z \ \operatorname{tr}(D^\top Z) + \lambda \|Z\|_{1,q} \quad \text{s.t.} \quad Z \geq 0, \ \ 1^\top Z = 1^\top$$

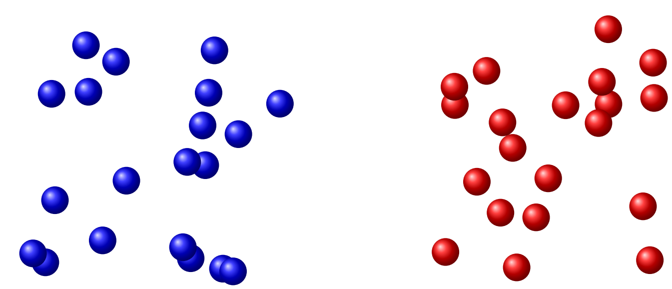
E. Elhamifar, G. Sapiro, and R. Vidal. Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery. NIPS 2012.

Theoretical Guarantees

$$\min_{Z} \ \operatorname{tr}(D^\top Z) + \lambda \|Z\|_{1,q} \quad \text{s.t.} \quad Z \geq 0, \ \ 1^\top Z = 1^\top$$

- **Theorem 1:** If λ is too big, only one exemplar is chosen; and if λ is too small, each point chooses itself as an exemplar
  - $\lambda \geq \lambda_{\max,q}(D) \implies Z = \boldsymbol{e}_\ell 1^\top \quad \text{where} \quad \ell = \arg\min_i 1^\top \boldsymbol{d}^i$
  - $\lambda \leq \lambda_{\min,q}(D) \implies Z = I$

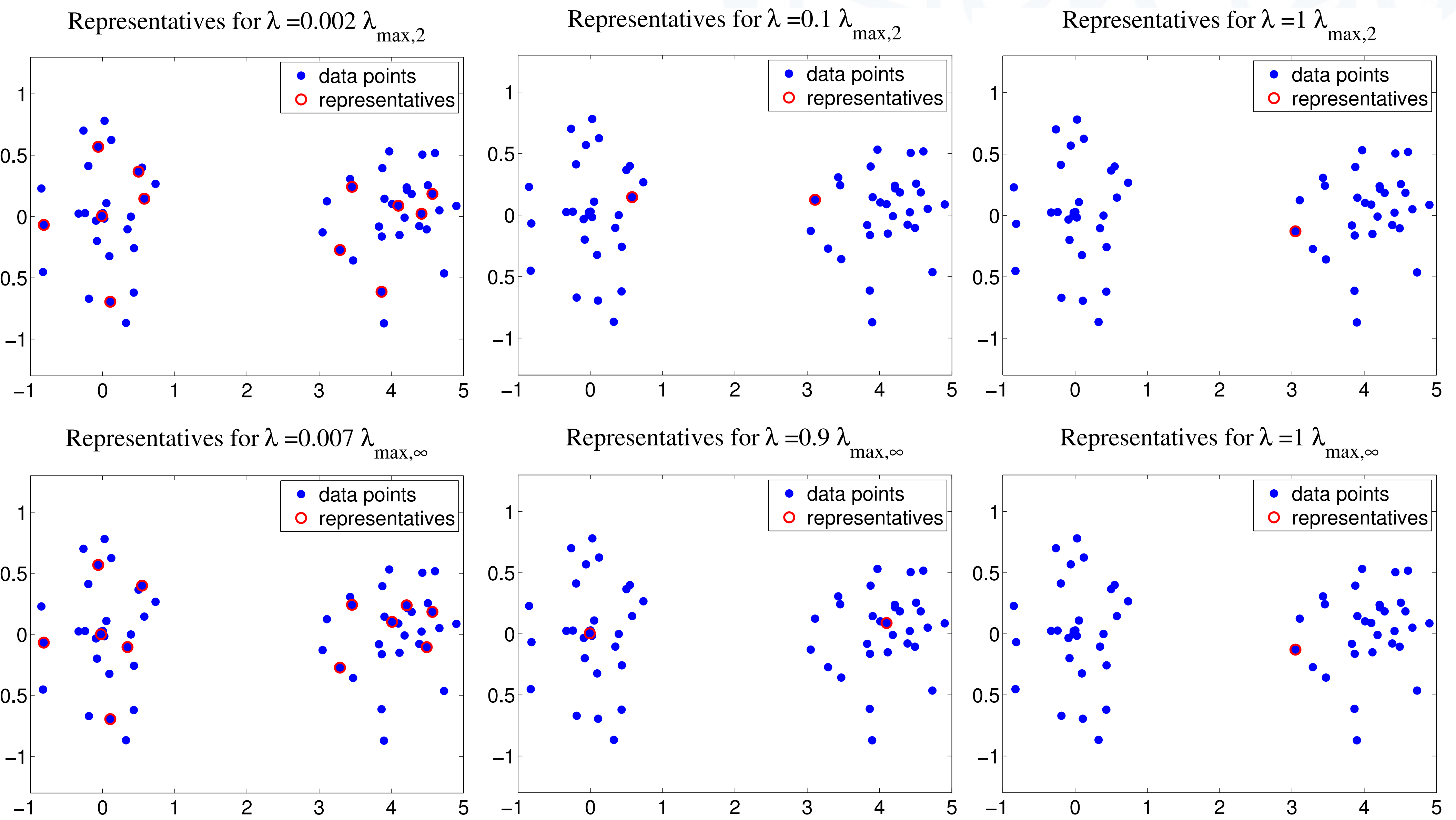
- **Theorem 2**: if $\lambda \leq \lambda_c(D)$ and the data partitions into *n* clusters, the optimal *Z* is such that data points within each cluster select exemplars from that cluster only



E. Elhamifar, G. Sapiro, and R. Vidal. Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery. NIPS 2012.

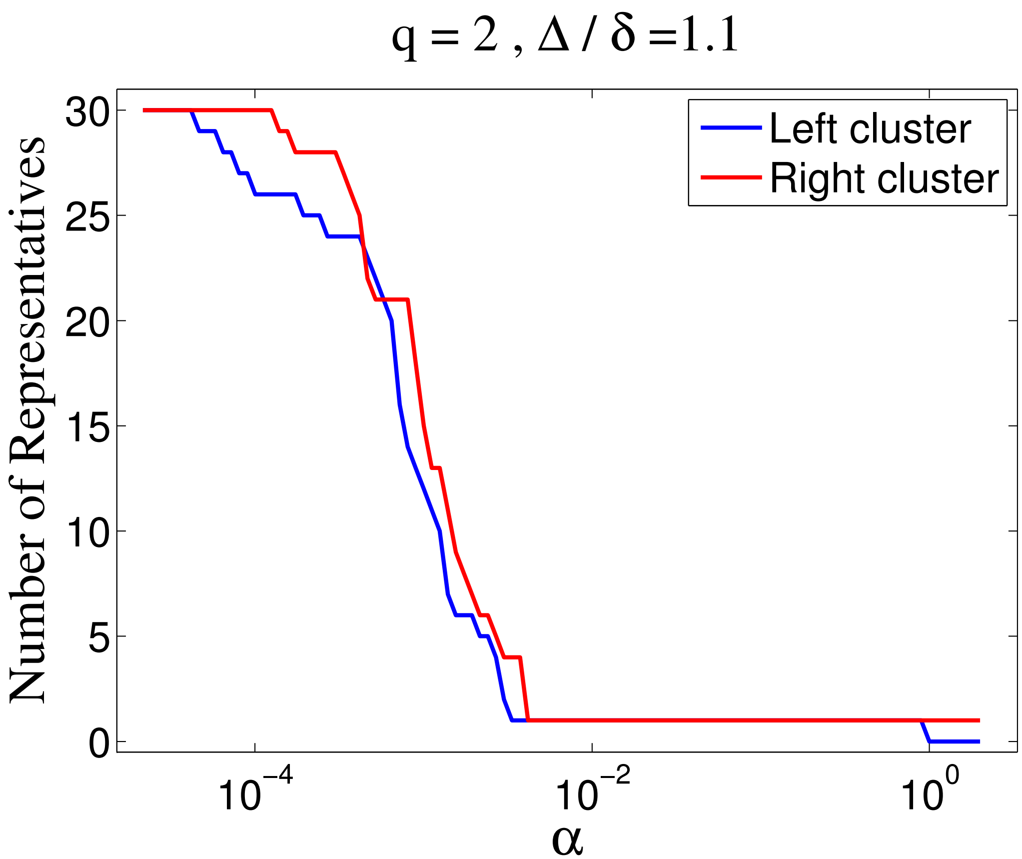
$\lambda_{\max,2} \triangleq \max_{i \neq \ell} \frac{\sqrt{N}}{2} \cdot \frac{\|\boldsymbol{d}_i - \boldsymbol{d}_\ell\|_2^2}{1^\top(\boldsymbol{d}_i - \boldsymbol{d}_\ell)}, \quad \lambda_{\max,\infty} \triangleq \max_{i \neq \ell} \frac{\|\boldsymbol{d}_i - \boldsymbol{d}_\ell\|_1}{2}, \quad \lambda_{\min,q} \triangleq \min_j (\min_{i \neq j} d_{ij} - d_{jj}), \quad \lambda_c \triangleq \min_j \min_{j' \in \mathcal{C}_j} (\min_{i \neq j} \min_{i' \in \mathcal{C}_i} d_{i'j'} - \max_{i' \in \mathcal{C}_j} d_{i'j'})$

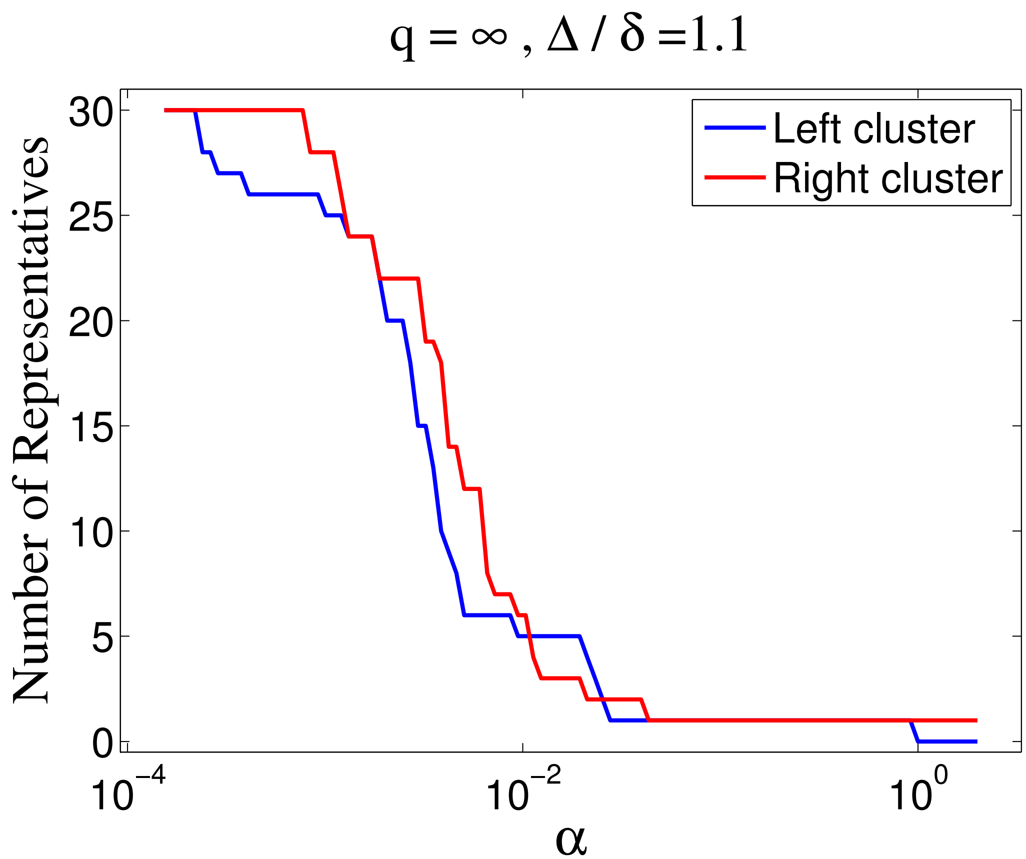
Experiments on Synthetic Data



E. Elhamifar, G. Sapiro, and R. Vidal. Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery. NIPS 2012.

Number of Exemplars/Cluster vs Lambda



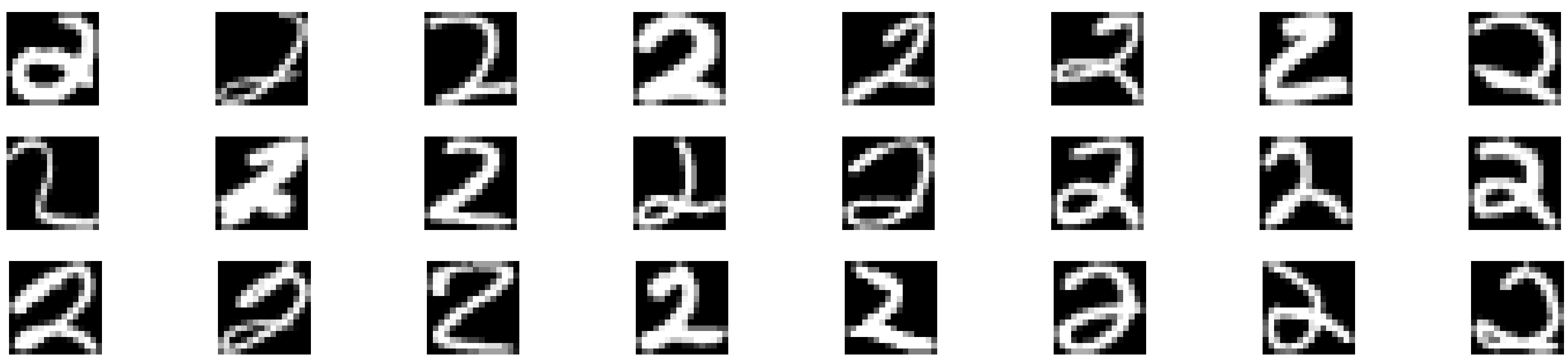


E. Elhamifar, G. Sapiro, and R. Vidal. Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery. NIPS 2012.

Applications: Classification with Exemplars

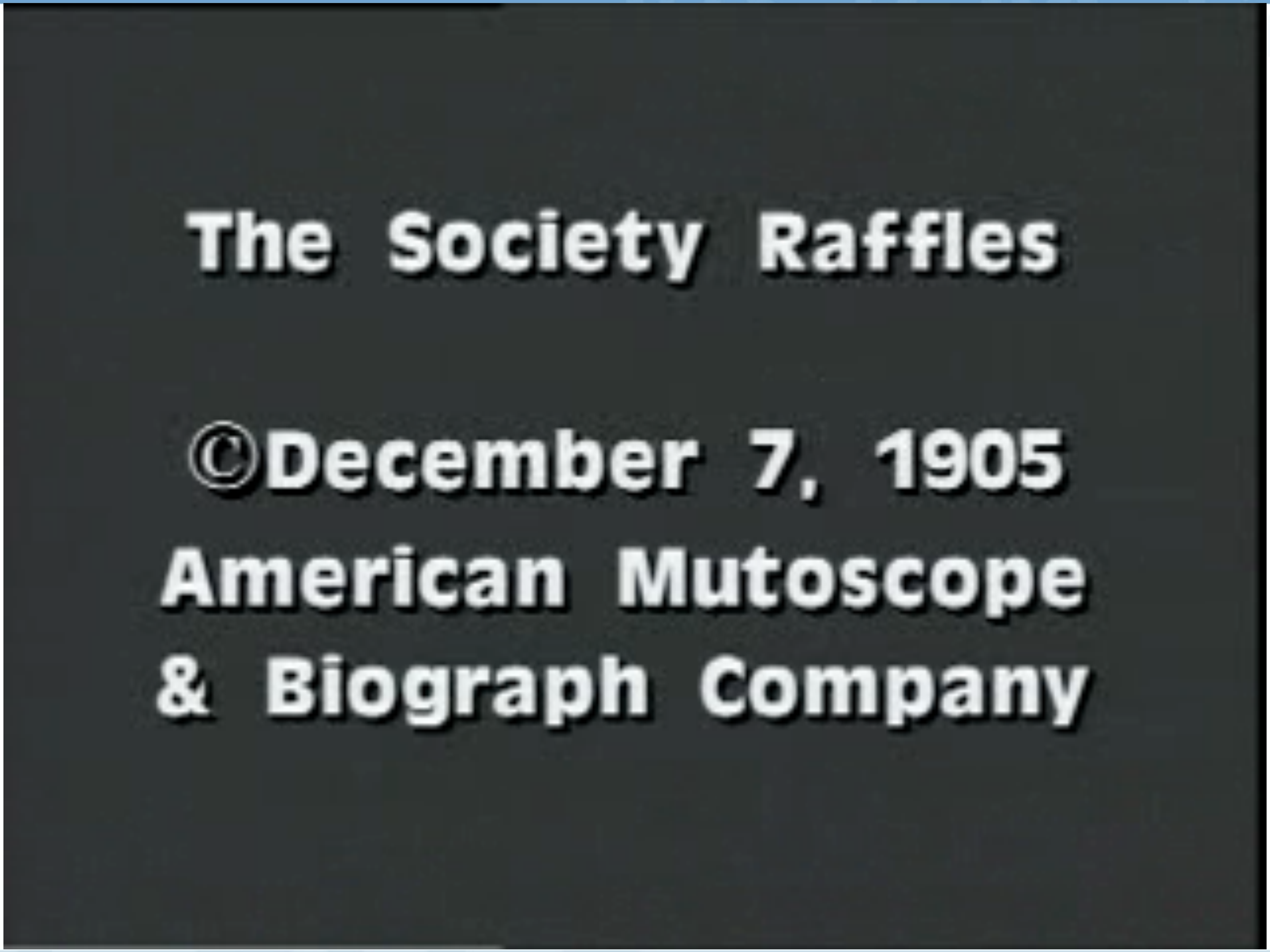
- Classification Results on the USPS digit database using 25 representatives of the 1,000 training samples in each class

	NN	NS	SRC	SVM
Rand	76.4%	84.9%	83.5%	98.6%
Kmedoids	**86.0%**	89.7%	89.6%	99.2%
RRQR	59.1%	81.3%	78.3%	94.3%
SMRS	83.4%	**93.8%**	**91.7%**	**99.7%**
All Data	96.2%	96.4%	98.9%	99.7%



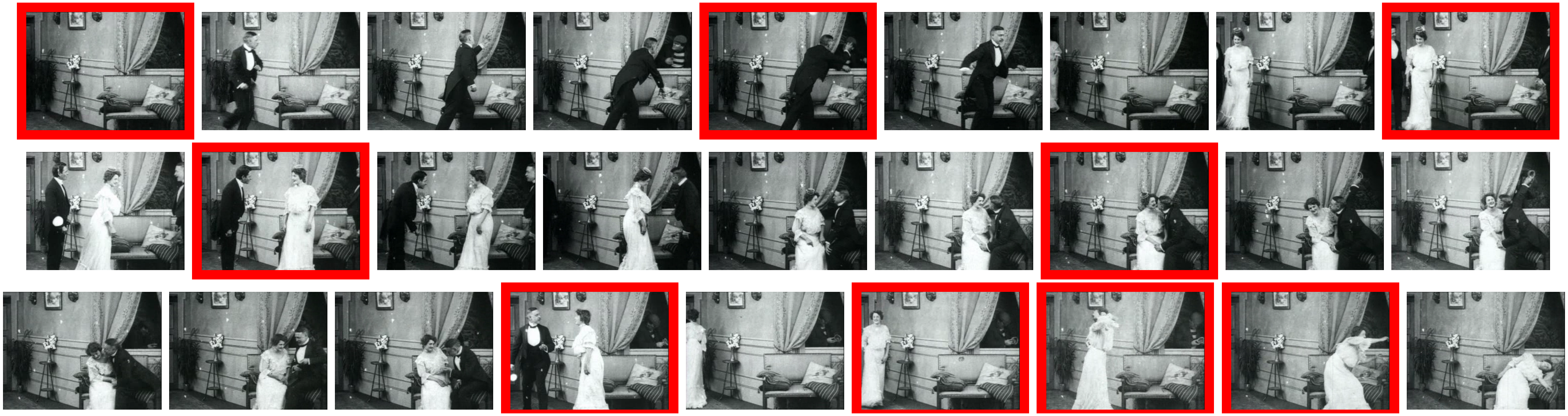
E. Elhamifar, G. Sapiro, and R. Vidal. See All by Looking at A Few: Sparse Modeling for Finding Representative Objects. CVPR 2012.

Applications: Exemplar Frames in a Video



Applications: Exemplar Frames in a Video

- Empty living room
- Middle-aged man enters
- Woman enters
- Young man enters, introduces the woman and leaves
- Middle-aged man flirts with woman and steals her tiara
- Middle-aged man checks the time, rises and leaves
- Woman walks him to the door
- Woman returns to her seat
- Woman misses her tiara
- Woman searches her tiara
- Woman sits and dismays



Applications: Exemplar Frames in a Video

E. Elhamifar, G. Sapiro, and R. Vidal. See All by Looking at A Few: Sparse Modeling for Finding Representative Objects. CVPR 2012.

Applications: Summarizing our NIPS paper!

– Given pairwise dissimilarities between data points, we consider the problem of finding a subset of data points, called representatives or exemplars, that can efficiently describe the data collection.

– We obtain the range of the regularization parameter for which the solution of the proposed optimization program changes from selecting one representative for all data points to selecting all data points as representatives.

– When there is a clustering of data points, defined based on their dissimilarities, we show that, for a suitable range of the regularization parameter, the algorithm finds representatives from each cluster.

– As the results show, the classification performance using the representatives found by our proposed algorithm is close to that of using all the training samples.

E. Elhamifar, G. Sapiro, and R. Vidal. Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery. NIPS 2012.

Acknowledgements

- Funding
  - ONR N00014-09-10839
  - NSF CNS-0931805,
  - NSF ECCS-0941463,
  - NSF OIA-0941362

Vision Lab @ Johns Hopkins University

http://www.vision.jhu.edu

Thank You!

Center for IMAGING SCIENCE