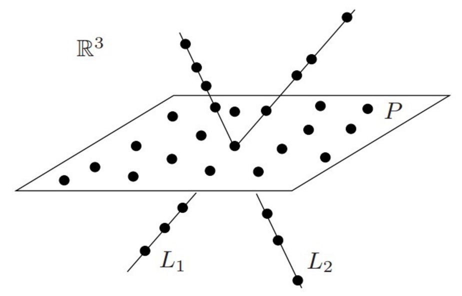
Correlation Adaptive Subspace Segmentation by Trace Lasso

Canyi Lu[1], *Jiashi Feng*[1], *Zhouchen Lin*[2], *Shuicheng Yan*[1]

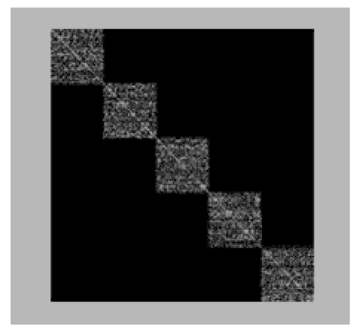
[1] *National University of Singapore*, [2] *Peking University*

Subspace Segmentation Problem

Given a set of data vectors $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_k] = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ drawn from a union of k subspaces $\{\mathcal{S}_i\}_{i=1}^k$. Let $\mathbf{X}_i$ be a collection of n_i data vectors drawn from the subspace $\mathcal{S}_i$, $n = \sum_{i=1}^k n_i$. The task is to segment the data according to the underlying subspaces they are drawn from.



An example of three subspaces.



Block diagonal affinity matrix.

Related Works

- Spectral Clustering (SC) is used as the framework for subspace segmentation.
- The main challenge by using SC is to define a "good" affinity matrix (or graph) $\mathbf{Z} \in \mathbb{R}^{n \times n}$. Each entry z_{ij} measures the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$.
- Ideally, the affinity matrix should be block diagonal **(sparsity between-cluster)**, and has **grouping effect** (sufficient connections within-cluster).

SSC (Sparse Subspace Clustering) (Elhamifar, CVPR 2009, TPAMI 2013) or **ℓ_1-Graph** (S. Yan, SDM 2009, TIP 2010) :

$$\min ||\mathbf{Z}||_1 \ \text{ s.t. } \mathbf{X} = \mathbf{X}\mathbf{Z}, \ \ \text{diag}(\mathbf{Z}) = 0.$$

$$\min \lambda||\mathbf{Z}||_1 + \frac{1}{2}||\mathbf{X} - \mathbf{X}\mathbf{Z}||_F^2, \ \text{ s.t. diag}(\mathbf{Z}) = 0.$$

LRR (Low-Rank Representation) (G. Liu et al., ICML 2010, TPAMI 2012):

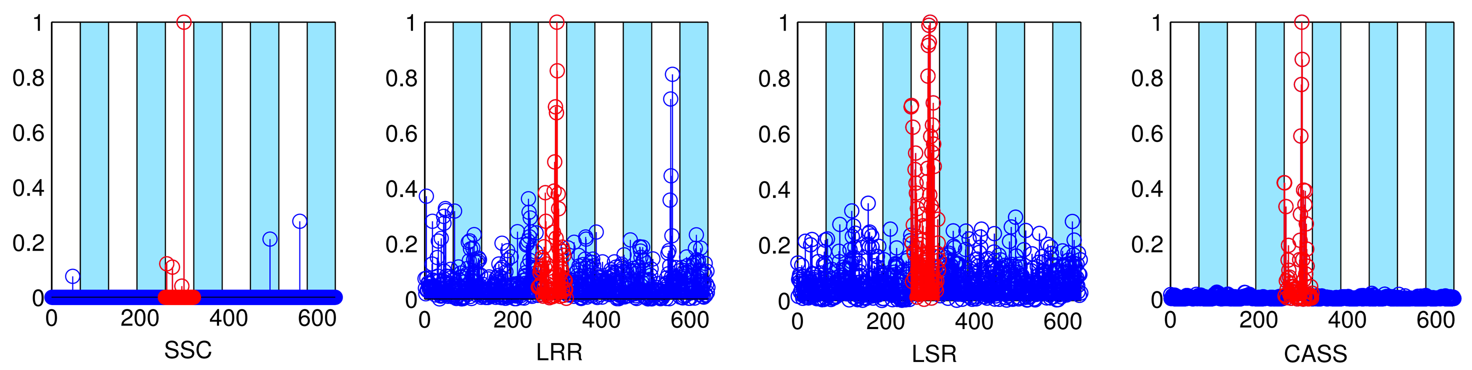
$$\min ||\mathbf{Z}||_* \ \text{ s.t. } \ \mathbf{X} = \mathbf{X}\mathbf{Z}.$$

$$\min \lambda||\mathbf{Z}||_* + ||\mathbf{X} - \mathbf{X}\mathbf{Z}||_{2,1}.$$

LSR (Least Squares Regression) (C. Lu and S. Yan, ECCV 2012):

$$\min ||\mathbf{Z}||_F \ \text{ s.t. } \ \mathbf{X} = \mathbf{X}\mathbf{Z}.$$

$$\min \lambda||\mathbf{Z}||_F^2 + \frac{1}{2}||\mathbf{X} - \mathbf{X}\mathbf{Z}||_F^2.$$



SSC pursues sparsity, but has no grouping effect.

LRR and LSR have strong grouping effect, but weak in sample selection.

A better choice, balances SSC (ℓ_1) and LSR (ℓ_2).

Correlation Adaptive Subspace Segmentation

Trace Lasso is an adaptive norm which balances the ℓ_1-norm and ℓ_2-norm. It is formally defined as

$$\Omega(\mathbf{w}) = ||\mathbf{X}\,\text{Diag}(\mathbf{w})||_*.$$

If the data are uncorrelated (the data points are orthogonal, $\mathbf{X}^T\mathbf{X} = \mathbf{I}$), trace Lasso is equal to the ℓ_1-norm:

$$||\mathbf{X}\,\text{Diag}(\mathbf{w})||_* = ||\text{Diag}(\mathbf{w})||_* = \sum_{i=1}^n |w_i| = ||\mathbf{w}||_1.$$

If the data are highly correlated (the data points are all the same, $\mathbf{X} = \mathbf{x}_1\mathbf{1}^T$, $\mathbf{X}^T\mathbf{X} = \mathbf{1}\mathbf{1}^T$), trace Lasso is equal to the ℓ_2-norm:

$$||\mathbf{X}\,\text{Diag}(\mathbf{w})||_* = ||\mathbf{x}_1\mathbf{w}^T||_* = ||\mathbf{x}_1||_2||\mathbf{w}||_2 = ||\mathbf{w}||_2.$$

For other cases, trace Lasso interpolates between the ℓ_2-norm and ℓ_1-norm:

$$||\mathbf{w}||_2 \leq ||\mathbf{X}\,\text{Diag}(\mathbf{w})||_* \leq ||\mathbf{w}||_1.$$

Generally speaking, trace Lasso is adaptive to the correlation of $\mathbf{X}$. If $\mathbf{X}$ is of low correlation, trace Lasso is close to the ℓ_1-norm. If $\mathbf{X}$ is of high correlation, trace Lasso is close to the ℓ_2-norm.

Correlation Adaptive Subspace Segmentation (CASS)

CASS with clean data:

$$\min_{\mathbf{w} \in \mathbb{R}^n} ||\mathbf{X}\,\text{Diag}(\mathbf{w})||_* \ \text{ s.t. } \ \mathbf{y} = \mathbf{X}\mathbf{w}. \qquad (1)$$

Enforced Block Sparse (EBS) Conditions. Given a matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ and a vector $\mathbf{w} = [\mathbf{w}_a; \mathbf{w}_b; \mathbf{w}_c] \in \mathbb{R}^n$, $\mathbf{w} \neq \mathbf{0}$. Let $\mathbf{w}^B = [\mathbf{0}; \mathbf{w}_b; \mathbf{0}] \in \mathbb{R}^n$. Saying $f(\mathbf{X}, \mathbf{w})$ satisfies the EBS conditions if

(1) $f(\mathbf{X}, \mathbf{w}) = f(\mathbf{X}\mathbf{P}, \mathbf{P}^{-1}\mathbf{w})$, for any permutation matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$;

(2) $f(\mathbf{X}, \mathbf{w}) \geq f(\mathbf{X}, \mathbf{w}^B)$, and the equality holds if and only if $\mathbf{w} = \mathbf{w}^B$.

Theorem 1 *If $f(\mathbf{X}, \mathbf{w})$ satisfies the EBS conditions, and the data in $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_k]$ are drawn from independent subspaces. $\mathbf{y}$ is a new point in $\mathcal{S}_i$. The solution to the following problem*

$$\min_{\mathbf{w}} f(\mathbf{X}, \mathbf{w}) \ \text{ s.t. } \ \mathbf{y} = \mathbf{X}\mathbf{w},$$

is block sparse, i.e., *$\mathbf{w}_i^* \neq 0$ and $\mathbf{w}_j^* = 0$, for all $j \neq i$.*

Theorem 2 *Trace Lasso $||\mathbf{X}\,Diag(\mathbf{w})||_*$ satisfies the EBS conditions. Thus the solution to problem (1) is block sparse when the subspaces are independent.*

The block sparse property guarantees that CASS achieves perfect segmentation under certain condition as other methods, e.g. LRR and LSR.

CASS with noisy data:

$$\min_{\mathbf{w}} \lambda||\mathbf{X}\,\text{Diag}(\mathbf{w})||_* + \frac{1}{2}||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2. \qquad (2)$$

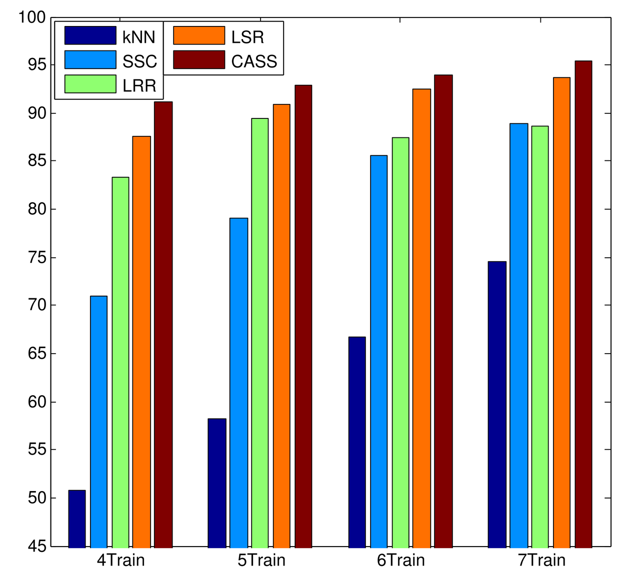
Theorem 3 *(The Grouping Effect). Given a data vector $\mathbf{y} \in \mathbb{R}^d$, data points $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ and parameter $\lambda > 0$. Let $\mathbf{w}^* = [w_1^*, \cdots, w_n^*]^T \in \mathbb{R}^n$ be the optimal solution to problem (2). If $\mathbf{x}_i \rightarrow \mathbf{x}_j$, then $w_i^* \rightarrow w_j^*$.*

The grouping effect of CASS guarantees that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are highly correlated, their corresponding coefficients are almost the same. Then they tends to be grouped together.

Segmentation Algorithm.

1. For each $\mathbf{x}_i$ in $\mathbf{X}$, solve problem (2) based on the dictionary $\mathbf{X}_{\hat{i}} = [\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_n]$. Denote the coefficient as $\mathbf{w}_i^*$. Let $\mathbf{W}^* = [\mathbf{w}_1^*, \cdots, \mathbf{w}_n^*]$.
2. Construct the affinity matrix by $(|\mathbf{W}^*| + |\mathbf{W}^{*T}|)/2$.
3. Segment the data into k groups by Normalized Cuts.

Experiments



semi-supervised learning

Segmentation errors (%) on the Hopkins 155 data.

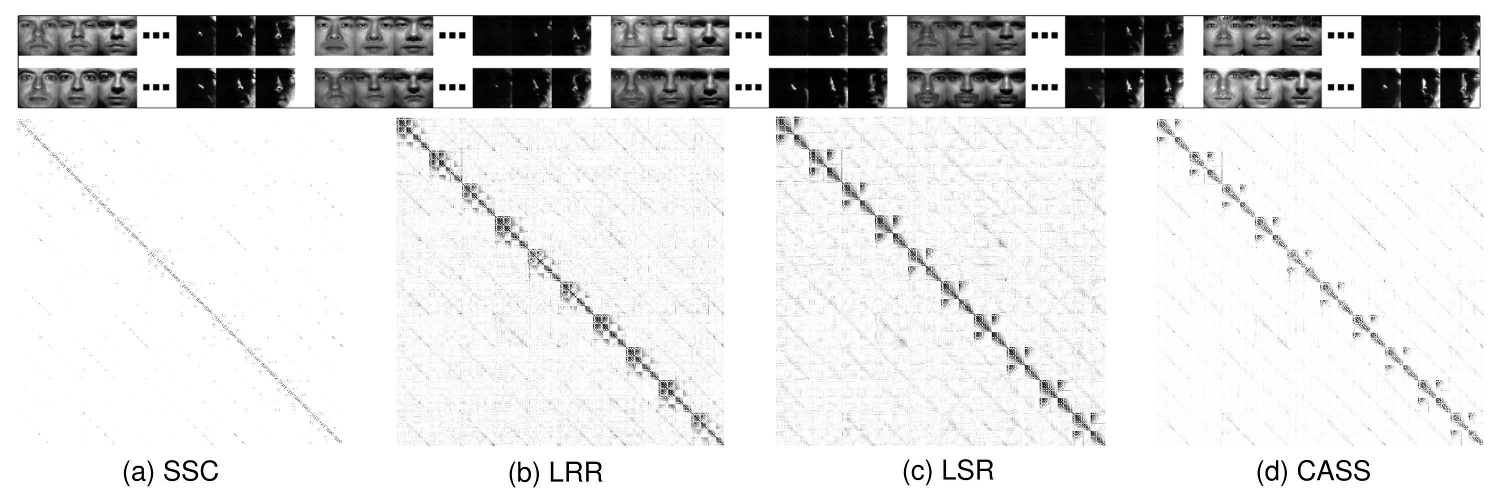
	kNN	SSC	LRR	LSR	CASS
MAX	45.59	39.53	36.36	36.36	**32.85**
MEAN	13.44	4.02	3.23	2.50	**2.42**
STD	12.90	10.04	6.60	**5.62**	5.84

Face clustering accuracy on the Yalbe B database.

	kNN	SSC	LRR	LSR	CASS
5 subjects	56.88	80.31	86.56	92.19	**94.03**
8 subjects	52.34	62.90	78.91	80.66	**91.41**
10 subjects	50.94	52.19	65.00	73.59	**81.88**

Digit clustering accuracy on the MNIST database.

	kNN	SSC	LRR	LSR	CASS
ACC.	61.00	62.60	66.80	68.00	**73.80**



Affinity matrices obtained by different methods on the Yale B database.

Take Home Message

- We propose the Correlation Adaptive Subspace Segmentation (CASS) method by using the trace Lasso.
- CASS is adaptive to the data correlation. When the data correlation is low, it is close to SSC. When the data correlation is high, it is close to LSR.
- CASS leads to a block diagonal representation and has the grouping effect.