ℓ^0-Sparse Subspace Clustering

Yingzhen Yang[1], Jiashi Feng[2], Nebojsa Jojic[3], Jianchao Yang[4], Thomas S. Huang[1]

[1] Beckman Institute, University of Illinois at Urbana-Champaign, USA
[2] Department of ECE, National University of Singapore, Singapore
[3] Microsoft Research, USA
[4] Snapchat, USA

Introduction

- Sparse Subspace Clustering (SSC) aims to partition the data according to their underlying subspaces.

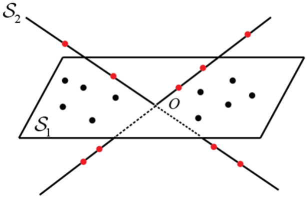


Figure 1: Black dots and red dots indicate the data that lie in subspace $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively.

Sparse Subspace Clustering

- Sparse Subspace Clustering (SSC) aims to partition the data according to their underlying subspaces.
- SSC and its robust version solve the following sparse representation problems:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \quad s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{\alpha},\ \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{0}$$

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{\alpha}\|_F^2 + \lambda_{\ell 1}\|\boldsymbol{\alpha}\|_1 \quad s.t.\ \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{0}$$

- Under certain assumptions on the underlying subspaces and the data, $\boldsymbol{\alpha}$ satisfies Subspace Detection Property (SDP): its nonzero elements correspond to the data that lie in the same subspace as point $\mathbf{x}_i$.

ℓ^0-induced Sparse Subspace Clustering

- Subspace Detection Property (SDP) is crucial for its success: data belonging to different subspaces are disconnected in the sparse graph.

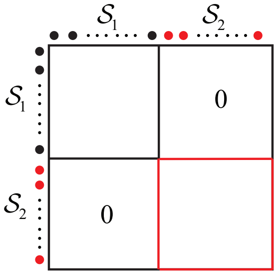


Figure 2: Block-diagonal similarity matrix due to SDP

- We propose ℓ^0-induced Sparse Subspace Clustering (ℓ^0-SSC), which solves the ℓ^0 problem:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0 \quad s.t. \ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{\alpha}, \ \operatorname{diag}(\boldsymbol{\alpha}) = \boldsymbol{0}$$

Models for Analyzing the Subspace Detection Property

- **Deterministic Model:** the subspaces and the data in each subspace are fixed.
- **Randomized Model:**
  - **Semi-Random Model:** the subspaces are fixed but the data are distributed at random in each of the subspaces.
  - **Full-Random Model:** the subspaces and the data of each subspace are random.

ℓ^0-induced Sparse Subspace Clustering

- The sparse subspace clustering literature does not have the answer to the fundamental problem: what is the relationship between sparse representation and SDP?
- Almost surely equivalence between ℓ^0-sparsity and SDP, under the mildest assumption to the best of our knowledge.

Theorem 1 (*ℓ^0-sparsity $\Rightarrow$ SDP*)

Under semi-random or full-random model, suppose data in each subspace are generated i.i.d. according to any continuous distribution. Then with probability 1 over the data for semi-random model, or over both the data and the subspaces for the full-random model, the optimal solution to the ℓ^0 sparse representation problem satisfies the subspace detection property.

ℓ^0-induced Sparse Subspace Clustering

- Inter-subspace hyperplane: the hyperplane spanned by data from different subspaces. The source where the confusion comes from.
- Key element in the proof: the probability of the intersection of the inter-subspace hyperplane and any associated subspace is 0.

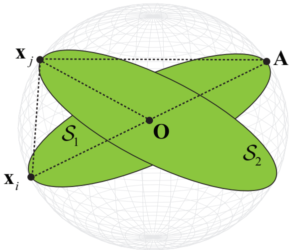


Figure 3: Illustration of a inter-subspace hyperplane spanned by $\mathbf{x}_i$ and $\mathbf{x}_j$.

ℓ^0-induced Sparse Subspace Clustering

- Compared to previous subspace clustering methods, ℓ^0-SSC achieves SDP under far less restrictive assumptions on both the underlying subspaces and the random data generation.

Assumption on Subspaces	Explanation
S_1:Independent Subspaces	$\mathrm{Dim}[\mathcal{S}_1 \oplus \mathcal{S}_2 \ldots \mathcal{S}_K] = {}_k \mathrm{Dim}[\mathcal{S}_k]$
S_2:Disjoint Subspaces	$\mathcal{S}_k \cap \mathcal{S}_{k'} = \mathbf{0}$ for $k \neq k'$
S_3:Overlapping Subspaces	$1 \leq \mathrm{Dim}[\mathcal{S}_k \cap \mathcal{S}_{k'}] < \min\{\mathrm{Dim}[\mathcal{S}_k], \mathrm{Dim}[\mathcal{S}_{k'}]\}$ for $k \neq k'$
S_4:Distinct Subspaces (ℓ^0-SSC)	$\mathcal{S}_k \neq \mathcal{S}_{k'}$ for $k \neq k'$

Assumption on Random Data Generation	Explanation
D_1:Semi-Random Model or Full-Random Model	i.i.d. uniformly on the unit sphere.
D_2:IID (ℓ^0-SSC)	i.i.d. from arbitrary continuous distribution.

- No requirement for other complex geometric conditions, such as ingradius and subspace incoherence.

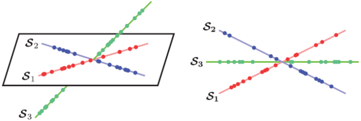


Figure 4: Independent (left) and disjoint (right) subspaces

ℓ^0-induced Sparse Subspace Clustering

- No free lunch! The price we pay for SDP under such much milder assumptions is solving the NP-hard ℓ^0 problem.
- No better deal! The converse of Theorem 1:

Theorem 2 (*No free lunch: SDP $\Rightarrow$ ℓ^0-sparsity*)

Under the semi-random or full-random model and the assumptions of Theorem 1, if there is an algorithm which, for any data point $\mathbf{x}_i \in \mathcal{S}_k$, $1 \le i \le n, 1 \le k \le K$, can find the data from the same subspace as $\mathbf{x}_i$ that linearly represent $\mathbf{x}_i$, i.e.

$$\mathbf{x}_i = \boldsymbol{X}\boldsymbol{\beta} \quad (\boldsymbol{\beta}_i = 0) \tag{1}$$

where nonzero elements of $\boldsymbol{\beta}$ correspond to the data that lie in the subspace $\mathcal{S}_k$. Then, with probability 1, solution to the ℓ^0 problem (for $\mathbf{x}_i$) can be obtained from $\boldsymbol{\beta}$ in $\mathcal{O}(\hat{n}^3)$ time, where $\hat{n}$ is the number of nonzero elements in $\boldsymbol{\beta}$.

Approximate ℓ^0-SSC (Aℓ^0-SSC)

- Allowing for some tolerance to noise, the optimization problem of ℓ^0-SSC is

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{n \times n}, \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{0}} L(\boldsymbol{\alpha}) = \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{\alpha}\|_F^2 + \lambda\|\boldsymbol{\alpha}\|_0$$

- Optimization by proximal gradient descent, using SSC as initialization

$$\boldsymbol{\alpha}^{i(t)} = h_{\sqrt{\frac{2\lambda}{\tau s}}}\left(\boldsymbol{\alpha}^{i(t-1)} - \frac{2}{\tau s}\left(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\alpha}^{i(t-1)} - \boldsymbol{X}^\top \mathbf{x}_i\right)\right)$$

where h is an element-wise hard thresholding operator.

Approximate ℓ^0-SSC

- The objective value $\{L(\boldsymbol{\alpha}^{i^{(t)}})\}_t$ is non-increasing and consequently it converges.
- But does $\{\boldsymbol{\alpha}^{i^{(t)}}\}_t$ converge?
- If $\{\boldsymbol{\alpha}^{i^{(t)}}\}_t$ converges, how far is the resultant sub-optimal solution from the globally optimal solution?

Approximate ℓ^0-SSC

- Definition of sparse eigenvalues

$$\kappa_-(m) := \min_{\|\mathbf{u}\|_0 \le m; \|\mathbf{u}\|_2 = 1} \|\boldsymbol{X}\mathbf{u}\|_2^2 \quad \kappa_+(m) := \max_{\|\mathbf{u}\|_0 \le m, \|\mathbf{u}\|_2 = 1} \|\boldsymbol{X}\mathbf{u}\|_2^2$$

Proposition 1

If $\kappa_-(|\mathrm{supp}(\boldsymbol{\alpha}^{i(0)})|) > 0$, $\{\boldsymbol{\alpha}^{i(t)}\}_t$ *is a bounded sequence that converges to a critical point of* L, *denoted by* $\hat{\boldsymbol{\alpha}}^i$.

Approximate ℓ^0-SSC

- Now how far is $\hat{\boldsymbol{\alpha}}^i$ from $\boldsymbol{\alpha}^{i*}$ (the globally optimal solution)?
- Roadmap: prove that both are local solutions to a capped-ℓ^1 problem, and then we can obtain the following bound:

Theorem 3

(Bounded distance between sub-optimal solution and the globally optimal solution) Under certain assumptions on the sparse eigenvalues of the data matrix, the sequence $\{\boldsymbol{\alpha}^{i(t)}\}_t$ *converges to a critical point of* $L(\boldsymbol{\alpha}^i)$, $\hat{\boldsymbol{\alpha}}^i$. *Then*

$$\|(\hat{\boldsymbol{\alpha}}^i - \boldsymbol{\alpha}^{i*})\|_2^2 \le \frac{2}{(\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|) - \kappa)^2}$$

$$\Big(\sum_{j \in \hat{\mathbf{S}}_i} (\max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j^i - b|\})^2 + |\mathbf{S}_i^* \setminus \hat{\mathbf{S}}_i|(\max\{0, \frac{\lambda}{b} - \kappa b\})^2\Big)$$

Approximate ℓ^0-SSC

- Remember that

$$\boldsymbol{\alpha}^{i\,(t)} = h_{\sqrt{\frac{2\lambda}{\tau s}}}\left(\boldsymbol{\alpha}^{i\,(t-1)} - \frac{2}{\tau \boxed{s}}\left(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\alpha}^{i\,(t-1)} - \boldsymbol{X}^\top \mathbf{x}_i\right)\right)$$

Proposition 2

If $s > \max\{2|\mathrm{supp}(\boldsymbol{\alpha}^{i\,(0)})|, \frac{2(1+\lambda|\mathrm{supp}(\boldsymbol{\alpha}^{i\,(0)})|)}{\lambda\tau}\}$, *then*

$$\mathrm{supp}(\boldsymbol{\alpha}^{i\,(t)}) \subseteq \mathrm{supp}(\boldsymbol{\alpha}^{i\,(t-1)}), t \geq 1$$

- Significantly reduces computational cost with efficient optimization:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n, \boldsymbol{\alpha}^i_i=0} \|\mathbf{x}_i - \boxed{\boldsymbol{X}}\boldsymbol{\alpha}^i\|_2^2 + \lambda\|\boldsymbol{\alpha}^i\|_0 \overset{PGD}{\Leftrightarrow} \min_{\boldsymbol{\alpha}\in\mathbb{R}^n, \boldsymbol{\alpha}^i_i=0} \|\mathbf{x}_i - \boxed{\boldsymbol{X}_{\mathbf{S}_i}}\boldsymbol{\alpha}^i\|_2^2 + \lambda\|\boldsymbol{\alpha}^i\|_0$$

Approximate ℓ^0-SSC

Algorithm 1 (Data Clustering by Aℓ^0-SSC)

Input:
The data set $\boldsymbol{X} = \{\mathbf{x}_i\}_{i=1}^n$, *the number of clusters* c, *the parameter* λ *for A*ℓ^0*-SSC, maximum iteration number* M, *stopping threshold* ε.

1: *Obtain the sub-optimal solution* $\tilde{\boldsymbol{\alpha}}$ *by proximal gradient descent.*
2: *Build the sparse similarity matrix by symmetrizing* $\tilde{\boldsymbol{\alpha}}$: $\tilde{\mathbf{W}} = \frac{|\tilde{\boldsymbol{\alpha}}|+|\tilde{\boldsymbol{\alpha}}^\top|}{2}$
3: *Apply spectral clustering method to* $\tilde{\mathbf{W}}$.

Output: *The cluster labels.*

Clustering Results

Table 1: Clustering Results on Various Image Data Sets

Data Set	Measure	KM	SC	SSC	SMCE	SSC-OMP	Aℓ^0-SSC
MNIST (random sampling)	AC	0.5621	0.4922	0.4948	0.5784	0.5754	**0.6590**
	NMI	0.5113	0.4755	0.5210	0.6332	0.5463	**0.6709**
COIL-20	AC	0.6554	0.4278	0.7854	0.7549	0.3389	**0.8472**
	NMI	0.7630	0.6217	0.9148	0.8754	0.4853	**0.9428**
COIL-100	AC	0.4996	0.2835	0.5275	0.5639	0.1667	**0.7683**
	NMI	0.7539	0.5923	0.8041	0.8064	0.3757	**0.9182**
Extended Yale-B	AC	0.0954	0.1077	0.7850	0.3293	0.6529	**0.8480**
	NMI	0.1258	0.1485	0.7760	0.3812	0.7024	**0.8612**
UMIST Face	AC	0.4275	0.4052	0.4904	0.4487	0.4835	**0.6730**
	NMI	0.6426	0.6159	0.6885	0.6696	0.6310	**0.7924**
CMU PIE	AC	0.0845	0.0729	0.2287	0.1733	0.0821	**0.2591**
	NMI	0.1884	0.1789	0.3659	0.3343	0.1494	**0.4435**
AR Face	AC	0.2752	0.2957	0.5914	0.3543	0.4229	**0.6086**
	NMI	0.5941	0.6248	0.8060	0.6573	0.6835	**0.8117**
MPIE S1	AC	0.1164	0.1285	0.5892	0.1721	0.1695	**0.6741**
	NMI	0.5049	0.5292	0.7653	0.5514	0.3395	**0.8622**
MPIE S2	AC	0.1315	0.1410	0.6994	0.1898	0.2093	**0.7527**
	NMI	0.4834	0.5128	0.8149	0.5293	0.4292	**0.8939**
MPIE S3	AC	0.1291	0.1459	0.6316	0.1856	0.1787	**0.7050**
	NMI	0.4811	0.5185	0.7858	0.5155	0.3415	**0.8750**
MPIE S4	AC	0.1308	0.1463	0.6803	0.1823	0.1680	**0.7246**
	NMI	0.4866	0.5280	0.8063	0.5294	0.3345	**0.8837**
Georgia Face	AC	0.4987	0.5187	0.5413	0.6053	0.4733	**0.6187**
	NMI	0.6856	0.7014	0.6968	0.7394	0.6622	**0.7400**

Parameter Sensitivity

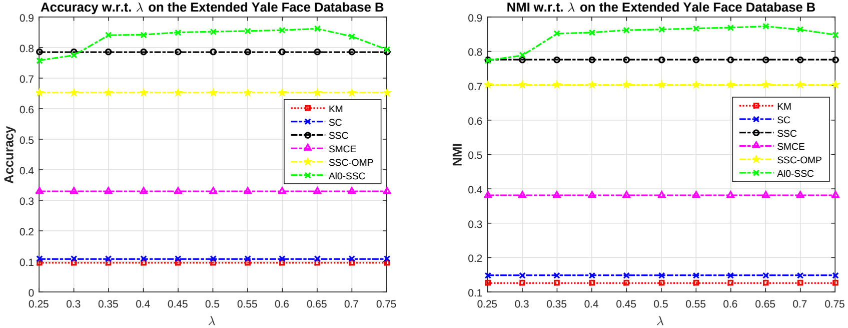
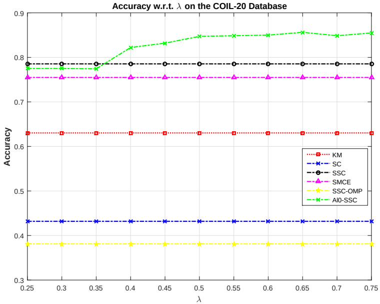


Figure 5: The performance change with varying λ on Extended Yale B

Parameter Sensitivity



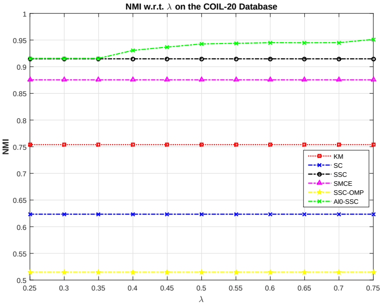


Figure 6: The performance change with varying λ on COIL-20

Summary

- Theory: Almost surely equivalence between ℓ^0-sparsity and the subspace detection property, under the mildest assumption to the best of our knowledge.

- Practice: Implemented by both MATLAB and CUDA C++ for extreme efficiency, with effectiveness evidenced by extensive experiments.

Thank you!