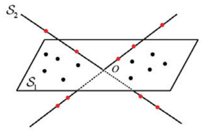
ℓ^0-Sparse Subspace Clustering

Yingzhen Yang, Jiashi Feng, Nebojsa Jojic, Jianchao Yang, Thomas S. Huang. UIUC, NUS, MSR, Snapchat.

Introduction

- High dimensional data often lie in a set of low-dimensional subspaces in many practical scenarios.
- Subspace clustering: partition the data such that data belonging to the same subspace are identified as one cluster.
- Among various subspace clustering algorithms, the ones that employ sparsity prior, such as Sparse Subspace Clustering (SSC), are effective in subspace clustering under certain assumptions.
- Typical algorithms build similarity matrix in accordance with the subspace-sparse representation, then apply spectral clustering on this similarity matrix to obtain clustering result.
- We present ℓ^0-Sparse Subspace Clustering (ℓ^0-SSC) with theoretical advantages and compelling empirical results compared to SSC and other competing subspace clustering methods.



Contributions

Key element: subspace-sparse representation or the representation that satisfies subspace detection property, which specifies data in the same subspace for each data point.

Main contributions of ℓ^0-SSC:

- Almost surely equivalence between ℓ^0-sparsity and the subspace detection property, under the mildest assumption to the best of our knowledge.
- Although the optimization of ℓ^0-SSC is NP-hard, we present an efficient proximal algorithm with theoretical guarantee. More concretely, under certain assumptions on the sparse eigenvalues of the data, the proximal algorithm guarantees convergence to the critical point of the objective, and the obtained sub-optimal solution is close to the globally optimal solution.

Formulation

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0 \quad s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{\alpha},\ \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{0} \tag{1}$$

Theorem 1 *(ℓ^0-sparsity $\Rightarrow$ subspace detection property) Suppose the data $\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ lie in a union of K distinct subspaces $\{\mathcal{S}_k\}_{k=1}^K$ of dimensions $\{d_k\}_{k=1}^K$. Let $\boldsymbol{X}^{(k)} \in \mathbb{R}^{d \times n_k}$ denote the data that belong to subspace $\mathcal{S}_k$, and $\sum_{k=1}^{K} n_k = n$. When $n_k \ge d_k + 1$, if the data belonging to each subspace are generated i.i.d. from arbitrary unknown continuous distribution supported on that subspace, then with probability 1, the optimal solution to (1) satisfies the subspace detection property, i.e. nonzero elements of $\boldsymbol{\alpha}^{i*}$ corresponds to the data that lie in the same subspace as $\mathbf{x}_i$.*

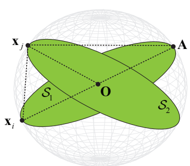


Illustration of a inter-subspace hyperplane. The hyperplane spanned by $\mathbf{x}_i \in \mathcal{S}_1$ and $\mathbf{x}_j \in \mathcal{S}_2$ is a inter-subspace hyperplane, and the intersection of this inter-subspace hyperplane and $\mathcal{S}_1$ is the dashed line $\mathbf{x}_i\mathbf{O}A$.

Formulation

- Key observation: the probability measure of inter-subspace hyperplane is 0.
- The only assumptions are i.i.d. random data generation and non-degenerate data distribution, which are the mildest assumptions required for subspace detection property so far to the best of our knowledge. Most other methods, such as SSC, noisy SSC and their geometric analysis, need more and stronger assumptions on the subspaces and the geometric properties of the data, such as conditions on the inradius and subspace incoherence.

Theorem 2 *(No free lunch: subspace detection property $\Rightarrow$ ℓ^0-sparsity) Under 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.*

$$\mathbf{x}_i = \boldsymbol{X}\boldsymbol{\beta} \quad (\boldsymbol{\beta}_i = 0) \tag{2}$$

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 (1) (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 $\boldsymbol{\beta}$.

- Proximal method for optimizing the following ℓ^0-SSC problem (Approximate ℓ^0-SSC)

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{n \times n}, \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{0}} L(\boldsymbol{\alpha}) = \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{\alpha}\|_F^2 + \lambda\|\boldsymbol{\alpha}\|_0 \tag{3}$$

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

where h is an element-wise hard thresholding operator.

Theorem 3 *(Bounded distance between sub-optimal solution and the globally optimal solution. Informal Statement, refer to the paper for more details) Under certain assumptions on the sparse eigenvalues of the data matrix, the sequence $\{\boldsymbol{\alpha}^{i^{(t)}}\}_t$ converges to a critical point of $L(\boldsymbol{\alpha}^i)$, which is denoted by $\hat{\boldsymbol{\alpha}}^i$. Let $\boldsymbol{\alpha}^{i^*}$ be the globally optimal solution to (3), then*

$$\|(\hat{\boldsymbol{\alpha}}^i - \boldsymbol{\alpha}^{i^*})\|_2^2 \le \frac{2}{(\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|) - \kappa)^2}$$
$$\Big(\sum_{j \in \hat{\mathbf{S}}_i} (\max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j^i - b|\})^2 + |\mathbf{S}_i^* \setminus \hat{\mathbf{S}}_i|(\max\{0, \frac{\lambda}{b} - \kappa b\})^2\Big)$$

Experimental Results

Clustering result on several face data sets, with comparison to several competing methods. KM: K-means, SC: Spectral Clustering, SMCE: Sparse Manifold Clustering and Embedding, SSC-OMP: using Orthogonal Matching Pursuit (OMP) to solve (1).

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**