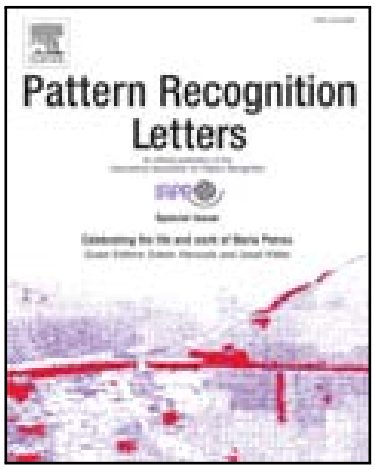

Sparse subspace clustering via smoothed ℓ_p minimization

Wenhua Dong, Xiao-jun Wu, Josef Kittler

PII: S0167-8655(19)30125-4
DOI: https://doi.org/10.1016/j.patrec.2019.04.018
Reference: PATREC 7500

To appear in: *Pattern Recognition Letters*

Received date: 28 November 2018
Revised date: 19 April 2019
Accepted date: 21 April 2019

Please cite this article as: Wenhua Dong, Xiao-jun Wu, Josef Kittler, Sparse subspace clustering via smoothed ℓ_p minimization, *Pattern Recognition Letters* (2019), doi: https://doi.org/10.1016/j.patrec.2019.04.018

Highlights

- A unified formulation for different practical clustering problems.
- An effective algorithm for multi-constrained ‘p minimization problem.
- The convergence of our algorithm was verified theoretically and experimentally.
- We applied our method to synthetic and real datasets and obtained good results.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Sparse subspace clustering via smoothed ℓ_p minimization

Wenhua Dong[a], Xiao-jun Wu[a,**], Josef Kittler[b]

[a]*Jiangnan University, No. 1800, LiHu, Avenue, Wuxi and 214122, China*
[b]*Centre for Vision, Speech and Signal Processing,University of Surrey, Guildford, GU2 7XH,U.K.*

ABSTRACT

In this letter, we formulate sparse subspace clustering as a smoothed ℓ_p $(0 < p < 1)$ minimization problem (SSC-SLp) and present a unified formulation for different practical clustering problems by introducing a new pseudo norm. Generally, the use of ℓ_p $(0 < p < 1)$ norm approximating the ℓ_0 one can lead to a more effective approximation than the ℓ_1 norm, while the ℓ_p-regularization also causes the objective function to be non-convex and non-smooth. Besides, better adapting to the property of data representing real problems, the objective function is usually constrained by multiple factors (such as spatial distribution of data and errors). In view of this, we propose a computationally efficient method for solving the multi-constrained non-smooth ℓ_p minimization problem, which smooths the ℓ_p norm and minimizes the objective function by alternately updating a block (or a variable) and its weight. In addition, the convergence of the proposed algorithm is theoretically proven. Extensive experimental results on real datasets demonstrate the effectiveness of the proposed method.

Keywords: Sparse subspace clustering; ℓ_p minimization; Unified formulation.

1. Introduction

The processing of high-dimensional data has become a part of the necessary daily work in numerical applications of pattern recognition [18] and data mining [5]. Subspace clustering, as a key technology in high dimensional data processing, has naturally become an important research topic. The so-called subspace clustering is the problem that groups high-dimensional data into different clusters (corresponding to different subspaces).

Recently, various subspace clustering methods have been developed, which can be roughly divided into four categories: statistical [22], algebraic [3], iterative [23], and spectral-type [2] methods. For more progress on subspace clustering, please refer to [20, 15, 12, 24, 4, 10, 13, 28, 27, 25, 26, 1, 7, 21, 30]. In this letter, we focus on the spectral-type method.

In the spectral-type paradigm, the representation based clustering method has attracted a wide attention due to its computational effectiveness and superior clustering performance. Let $X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{m\times n}$ denote a given data matrix drawn from a union of t linear subspaces $\{\mathcal{S}_i\}_{i=1}^t$ with dimensions $\{d_i\}_{i=1}^t$. The premise is that each data point x_i can be reconstructed by a linear combination of all data points, i.e., $x_i = Xc_i$, where $c_i \in \mathbb{R}^n$ is the representation coefficient of x_i. For example, nuclear norm based regularized LRR [12] and LRSC [24] attempt to find a low rank representation of data. In contrast to LRR and LRSC, sparse representation based SSC [4] encourages model coefficients to be sparse by ℓ_1-regularization. After that, various variants have been proposed, such as S^3C [10], LapLRR [13], LRSSC [26] and MLRSSC [1]. To better model practical problems, these low rank and sparse representation based clustering problems naturally involve multiple variables and constraints. The Alternating Direction Method (ADM) [29] with theoretical guarantees is widely used to solve such problems. Although rigorous mathematical analysis is used to construct the optimization process, the ADM needs to introduce several auxiliary variables which may sacrifice convergence rate, or even affect the convergence of the algorithm when there exist too many variables. To alleviate the limitation, Linearized ADM (LADM) [11] is proposed to improve the convergence rate and reduce the number of auxiliary variables, but it also

**Corresponding author:
e-mail: wu_xiaojun@jiangnan.edu.cn (Xiao-jun Wu)

encounters convergence problems. Recent studies [8, 14] show that the Iteratively Reweighted Least Squares (IRLS) method for the unconstrained ℓ_p $(0 < p \leqslant 1)$ minimization problem can provide a fast solver and good sparse recovery performance, which motivates us to formulate sparse subspace clustering as a smoothed ℓ_p $(0 < p < 1)$ minimization problem with multiple constraints. The main contributions of this letter include:

1. A unified formulation for different practical clustering problems via multi-constrained ℓ_p minimization.
2. An effective method for solving the proposed ℓ_p minimization problem.
3. A theoretically and experimentally verified convergence of the proposed algorithm.
4. Superior performance on real datasets against the state-of-the-art algorithms.

2. Related work

In this section, we present a brief review of sparse subspace clustering (SSC) methods. In recent years, SSC has attracted a lot of interest due to its superior clustering performance (e.g., [10, 28, 27]). The commonality of these methods is to encourage the model coefficients to be sparse such that the inter-cluster affinity is zero and the intra-cluster one is sparse. According to [4], the following practical clustering problems are considered:

($\boldsymbol{\mathcal{P}_1}$) Clustering of clean data points drawn from a union of t linear subspaces. For each x_i, SSC is formulated as:

$$\min_{c_i} \|c_i\|_0 \quad \text{s.t. } x_i = Xc_i,\ c_{ii} = 0, \tag{1}$$

where $\|c_i\|_0 = \#\{c_{ji}\}$ denotes the ℓ_0-norm of $c_i \in \mathbb{R}^n$ and the constraint $c_{ii} = 0$ is used to avoid the trivial solution. In general, the ℓ_0 minimization is an NP-hard problem. A common approach is to employ the convex surrogate ℓ_1-norm instead, which leads to the following minimization problem:

$$\min_{c_i} \|c_i\|_1 \quad \text{s.t. } x_i = Xc_i,\ c_{ii} = 0, \tag{2}$$

where $\|c_i\|_1 = \sum_{j=1}^{n} |c_{ji}|$ denotes the ℓ_1-norm of c_i.

($\boldsymbol{\mathcal{P}_2}$) Clustering of data points drawn from a union of t affine subspaces and corrupted by noise. Here, SSC is formulated as:

$$\min_{c_i, z_i} \|c_i\|_1 \quad \text{s.t. } x_i = Xc_i + z_i,\ \mathbf{1}^T c_i = 1,\ c_{ii} = 0. \tag{3}$$

where $z_i \in \mathbb{R}^m$ is the additive noise, $\mathbf{1}$ denotes the all ones vector, and $\mathbf{1}^T c_i = 1$ is the linear equality constraint of affine subspaces [4].

($\boldsymbol{\mathcal{P}_3}$) Clustering of data points drawn from a union of t affine subspaces and contaminated by noise and sparse outlying entries [4]. For this scenario, SSC can be formulated as follows

$$\min_{c_i, z_i, o_i} \|c_i\|_1 \quad \text{s.t. } x_i = Xc_i + z_i + o_i,\ \mathbf{1}^T c_i = 1,\ c_{ii} = 0. \tag{4}$$

where $o_i \in \mathbb{R}^m$ is the sparse outlying entry with a few large nonzero components.

The problems in (2)-(4) are all convex and can be solved by the ADM, respectively. After that the optimized solution $C = [c_1, c_2, \cdots, c_n]$ is employed to construct an affinity matrix $\mathcal{A} = |C| + |C|^T$ from which we can obtain clustering assignments using the spectral clustering [20] approach.

3. Sparse subspace clustering via smoothed ℓ_p minimization

In this section, we present a unified formulation for different practical clustering problems ($\boldsymbol{\mathcal{P}_1}$)-($\boldsymbol{\mathcal{P}_3}$) via smoothed ℓ_p minimization. For simplification, we start with the clustering problem ($\boldsymbol{\mathcal{P}_2}$) and regard ($\boldsymbol{\mathcal{P}_1}$) as its special case. We then incorporate the formulation of ($\boldsymbol{\mathcal{P}_3}$) into the obtained framework.

3.1. Methods

Usually, the ℓ_1-norm is employed for sparse recovery as a convex relaxation of the ℓ_0-norm in compressed sensing. The problem is that this relaxation may cause large error on large coefficients [31]. Besides, it is well known that the ℓ_p $(0 < p < 1)$ is a pseudo norm between ℓ_0 and ℓ_1 norms. When decreasing the value of p from 1 toward 0, the ℓ_p norm approaches ℓ_0 norm. Accordingly, we formulate sparse subspace clustering as a multi-constrained ℓ_p minimization problem.

To start, we consider the clustering problem ($\boldsymbol{\mathcal{P}_2}$) and formulate it as follows

$$\min_{c_i, z_i} \|c_i\|_p^p + \frac{1}{2\lambda}\|z_i\|_2^2 \quad \text{s.t. } x_i = Xc_i + z_i, \mathbf{1}^T c_i = 1,\ c_{ii} = 0, \tag{5}$$

where $\|c_i\|_p^p = \sum_{j=1}^{n} |c_{ji}|^p$ denotes the ℓ_p $(0 < p < 1)$ norm of c_i and $\lambda > 0$ is a balance parameter. Obviously, the variation without the term z_i and affine constraint is the formulation of the problem ($\boldsymbol{\mathcal{P}_1}$). It is equivalent to

$$\min_{c_i} \|c_i\|_p^p + \frac{1}{2\lambda}\|x_i - Xc_i\|_2^2 \quad \text{s.t. } \mathbf{1}^T c_i = 1,\ c_{ii} = 0. \tag{6}$$

According to the strategy used in [16], the problem in (6) is equivalent to the following minimization problem:

$$\min_{c_i} \|c_i\|_p^p + \frac{1}{2\lambda}\|x_i - Xc_i\|_2^2 + \frac{\beta}{2}(\mathbf{1}^T c_i - 1)^2 \quad \text{s.t. } c_{ii} = 0. \tag{7}$$

When β is large enough, the optimal solution c_i of (7) will make the third term $(\mathbf{1}^T c_i - 1)^2$ to be zero due to $(\mathbf{1}^T c_i - 1)^2 \geqslant 0$. Therefore, the problems in (6) and (7) are equivalent w.r.t. a large enough value of β [16].

As the quadratic terms in (7) can be combined together, we can rewrite (7) as

$$\min_{c_i} \mathcal{H}(c_i) \triangleq \|c_i\|_p^p + \frac{1}{2\lambda}\left\|\hat{x}_i - \hat{X}c_i\right\|_2^2 \quad \text{s.t. } c_{ii} = 0, \tag{8}$$

where $\hat{X} \triangleq [X; (\lambda\beta)^{1/2}\mathbf{1}^T]$ with the i-th column $\hat{x}_i$. The main challenge for solving (8) originates in the non-convex and non-smooth nature of the objective. An efficient way is to smooth the ℓ_p-norm by introducing a regularization parameter $\delta > 0$:

$$\min_{c_i} \mathcal{H}(c_i, \delta) \triangleq \left\|(c_i)^2 + \delta\mathbf{1}\right\|_{\frac{p}{2}}^p + \frac{1}{2\lambda}\left\|\hat{x}_i - \hat{X}c_i\right\|_2^2 \quad \text{s.t. } c_{ii} = 0, \tag{9}$$

where $\left\|(c_i)^2 + \delta\mathbf{1}\right\|_{\frac{p}{2}}^p = \sum_{j=1}^{n}[(c_{ji})^2 + \delta]^{p/2}$. The problems in (8) and (9) are ϵ-equivalent w.r.t. δ, which is based on the fact: when $\delta \downarrow 0$, $\mathcal{H}(c_i, \delta) \downarrow \mathcal{H}(c_i)$, i.e., for any given $\epsilon > 0$, there always exists $\delta > 0$ such that

$$\mathcal{H}(c_i) < \mathcal{H}(c_i, \delta) \leqslant \mathcal{H}(c_i) + \epsilon. \tag{10}$$

The Lagrangian function of the problem in (9) is

$$\mathcal{L}(c_i,\delta)=\left\|(c_i)^2+\delta\mathbf{1}\right\|_{\frac{p}{2}}^{p}+\frac{1}{2\lambda}\left\|\hat{x}_i-\hat{X}c_i\right\|_2^2+\gamma e_i^T c_i, \tag{11}$$

where γ is the Lagrange multiplier and $e_i \in \mathbb{R}^n$ denotes the i-th standard basis vector whose i-th component is one and the others are zero. The derivative of Eq. (11) w.r.t. c_i is

$$\frac{\partial\mathcal{L}(c_i,\delta)}{\partial c_i}=Wc_i-\frac{1}{\lambda}\hat{X}^T(\hat{x}_i-\hat{X}c_i)+\gamma e_i, \tag{12}$$

where W is the weight matrix (a diagonal matrix) of c_i with the j-th diagonal entry $p[(c_{ji})^2+\delta]^{\frac{p-2}{2}}$. Setting the derivative to zero, we have

$$Wc_i-\frac{1}{\lambda}\hat{X}^T(\hat{x}_i-\hat{X}c_i)+\gamma e_i=0. \tag{13}$$

Solving the above system of linear equations, we have

$$c_i=(\hat{X}^T\hat{X}+\hat{W})^{-1}(\hat{X}^T\hat{x}_i-\lambda\gamma e_i), \tag{14}$$

where $\hat{W}$ is a diagonal matrix with the j-th diagonal entry:

$$[\hat{W}]_{jj}=\lambda p[(c_{ji})^2+\delta]^{\frac{p-2}{2}}. \tag{15}$$

Multiplying both sides of Eq. (14) by e_i^T, and noting that $e_i^T c_i = 0$, we can obtain

$$\gamma=\frac{e_i^T(\hat{X}^T\hat{X}+\hat{W})^{-1}\hat{X}^T\hat{x}_i}{\lambda e_i^T(\hat{X}^T\hat{X}+\hat{W})^{-1}e_i}. \tag{16}$$

Substituting (16) into (14), we find

$$c_i=\mathcal{M}(\hat{X}^T\hat{x}_i-\frac{e_i^T\mathcal{N}\hat{x}_i}{e_i^T\mathcal{M}e_i}e_i), \tag{17}$$

where $\mathcal{M}=(\hat{X}^T\hat{X}+\hat{W})^{-1}$ and $\mathcal{N}=\mathcal{M}\hat{X}^T$. It is worth noting that our solution can deal with the case of $0<p<2$ ($p-2$ is negative in (15)).

Finally, (15) and (17) are alternately updated until the algorithm converges or the number of iterations exceeds a predetermined number. The algorithm process is summarized in Algorithm 1, where v^k denotes the k-th iteration of v.

3.2. A clustering problem with sparse outlying entries

Typically, when the objective involves too many variables the algorithm analysis will become difficult. So, we expect to incorporate the formulation of ($\boldsymbol{\mathcal{P}_3}$) into the framework of the model (5). For each x_i, the problem ($\boldsymbol{\mathcal{P}_3}$) is formulated as:

$$\begin{aligned}&\min_{c_i,z_i,o_i}\ (\|c_i\|_p^p+\|o_i\|_q^q)+\frac{1}{2\lambda}\|z_i\|_2^2\\&\text{s.t. }\ x_i=Xc_i+z_i+o_i,\ \mathbf{1}^Tc_i=1,\ \ c_{ii}=0,\end{aligned} \tag{18}$$

where $\|o_i\|_q^q=\sum_{j=1}^m|o_{ji}|^q$ denotes the ℓ_q ($0<q<1$) norm of $o_i\in\mathbb{R}^m$. Note that the problem in (18) involves multiple variables and constraints, and its first two terms are non-smooth. The work in [7] proposed an ADM [29] framework based algorithm to solve such problems, where the IRLS is embedded into the ADM for solving the ℓ_p and ℓ_q subproblems, respectively. As the previous analysis shows, the traditional ADM framework based algorithm may encounter the problem of convergence rate and convergence. To effectively optimize the problem in (18), we first introduce some symbols used in subsequent sections.

Algorithm 1 : Sparse subspace clustering by (9)

Input: Data set $X\in\mathbb{R}^{m\times n}$, number t of subspaces, $\lambda>0$, $p\in\{0.3,0.5,0.7\}$, $\rho>1$ and $\zeta=10^{-6}$.

Initialization: Representation matrix $C^0=0$ and $\delta^0=1$.

for $k=0,1,2,\cdots$

1. Update the weight matrix $\hat{W}^k$ of each c_i by $[\hat{W}^k]_{jj}=\lambda p[(c_{ji}^k)^2+\delta]^{\frac{p-2}{2}}$.
2. Calculate matrices $\mathcal{M}$ and $\mathcal{N}$, and update $\hat{C}^{k+1}$ by $c_i^{k+1}=\mathcal{M}(\hat{X}^T\hat{x}_i-\frac{e_i^T\mathcal{N}\hat{x}_i}{e_i^T\mathcal{M}e_i}e_i)$.
3. Update δ^{k+1} by $\delta^{k+1}=\delta^k/\rho$.
4. If $\left\|C^k-C^{k+1}\right\|_\infty\le\zeta$, break.

end for

5. Build an affinity matrix $\mathcal{A}=|C|+|C|^T$.
6. Apply spectral clustering to the affinity matrix $\mathcal{A}$.

Output: Cluster assignments of X.

Let $\tilde{X}\triangleq[X,I]$ and $\tilde{c}_i\triangleq[c_i;o_i]$. Thus, the constraint $x_i=Xc_i+z_i+o_i$ in (18) is reduced to $x_i=\tilde{X}\tilde{c}_i+z_i$. Note that the block $\tilde{c}_i$ is still sparse when c_i and o_i are sparse. Furthermore, we define a function satisfying the following properties:

$$\begin{aligned}&\|[a;b]\|_{p|q}\triangleq(\sum_{i=1}^n|a_i|^p)^{\frac{1}{p}}+(\sum_{j=1}^m\left|b_j\right|^q)^{\frac{1}{q}},\ \text{and}\\&\|[a;b]\|_{p|q}^{p|q}\triangleq\sum_{i=1}^n|a_i|^p+\sum_{j=1}^m\left|b_j\right|^q,\end{aligned} \tag{19}$$

where $a\in\mathbb{R}^n$ and $b\in\mathbb{R}^m$ with the i-th elements a_i and b_i, respectively, and $0<p,q<1$. It is easy to prove that the above function is a pseudo norm since it violates the triangle inequality, and has similar properties to the ℓ_p norm. We call it $p|q$ norm in this letter. Obviously, when $p=q$, the $p|q$ norm is reduced to ℓ_p norm. That is to say, the $p|q$ norm is an extension of ℓ_p norm, where p and q can take different values. Based on the $p|q$ norm and the block $\tilde{c}_i$, $(\|c_i\|_p^p+\|o_i\|_q^q)$ in (18) is reduced to $\|[c_i;o_i]\|_{p|q}^{p|q}=\|\tilde{c}_i\|_{p|q}^{p|q}$. Thus, we can rewrite (18) as

$$\min_{\tilde{c}_i,z_i}\|\tilde{c}_i\|_{p|q}^{p|q}+\frac{1}{2\lambda}\|z_i\|_2^2\ \ \text{s.t. }x_i=\tilde{X}\tilde{c}_i+z_i,\mathbf{1}^Tc_i=1,c_{ii}=0. \tag{20}$$

It is noteworthy that the model (20) is reduced to (5). Owning to the similarity of the $p|q$ and ℓ_p norms, Algorithm 1 can be also applied to the model in (20) by alternately updating a block and its weight. Moreover, Algorithm 1 is applicable to the clustering of data points contaminated by noise or sparse outlying entries for linear subspaces as well. Therefore, it is a unified formulation for different practical clustering problems. Importantly, the unified formulation makes the algorithm analysis of the problem in (18) easier, although it contains multiple variables. In subsequent sections, we mainly focus on it.

4. Algorithm analysis

In this section, we present the convergence and complexity analysis of Algorithm 1.

4.1. Convergence Analysis

We conduct the convergence analysis of Algorithm 1 for non-convex and non-smooth optimization.

Theorem 3.1. For given λ and $0 < p < 1$,
1) $\mathcal{L}(c_i^k, \delta^k)$ is non-increasing;
2) The sequence $\{c_i^k\}$ is bounded;
3) $\lim_{k\to\infty} \left\|c_i^k - c_i^{k+1}\right\|_2 = 0$.

Proof. According to (17), it is easy to verify that $e_i^T c_i^k = 0$, and thus we have $\mathcal{L}(c_i^k, \delta^k) \geqslant 0$ for all $k \geqslant 0$. We calculate

$$\begin{aligned}
&\mathcal{L}(c_i^k, \delta^k) - \mathcal{L}(c_i^{k+1}, \delta^{k+1}) \\
&= \left\|(c_i^k)^2 + \delta^k \mathbf{1}\right\|_{\frac{p}{2}}^p - \left\|(c_i^{k+1})^2 + \delta^{k+1}\mathbf{1}\right\|_{\frac{p}{2}}^p \\
&+ \frac{1}{2\lambda}\left\|\hat{X}c_i^k - \hat{X}c_i^{k+1}\right\|_2^2 - \frac{1}{\lambda}(\hat{X}c_i^k - \hat{X}c_i^{k+1})^T(\hat{x}_i - \hat{X}c_i^{k+1}) \\
&= \sum_{j=1}^{n} \frac{(c_{ji}^k)^2 + \delta^k}{[(c_{ji}^k)^2 + \delta^k]^{1-\frac{p}{2}}} - \frac{[(c_{ji}^{k+1})^2 + \delta^{k+1}]^{\frac{p}{2}}[(c_{ji}^k)^2 + \delta^k]^{1-\frac{p}{2}}}{[(c_{ji}^k)^2 + \delta^k]^{1-\frac{p}{2}}} \qquad (21) \\
&+ \frac{1}{2\lambda}\left\|\hat{X}c_i^k - \hat{X}c_i^{k+1}\right\|_2^2 - (c_i^k - c_i^{k+1})^T(W^k c_i^{k+1} + \gamma^{k+1} e_i) \\
&\geqslant \frac{1}{2}(c_i^k - c_i^{k+1})^T W^k (c_i^k + c_i^{k+1}) + \frac{1}{2\lambda}\left\|\hat{X}c_i^k - \hat{X}c_i^{k+1}\right\|_2^2 \\
&- (c_i^k - c_i^{k+1})^T W^k c_i^{k+1} \\
&= \frac{1}{2}(c_i^k - c_i^{k+1})^T (W^k + \frac{1}{\lambda}\hat{X}^T\hat{X})(c_i^k - c_i^{k+1}) \geqslant 0, \qquad (22)
\end{aligned}$$

where the first equality follows (13) and $e_i^T c_i^k = 0$. The first inequality holds by using the Young's inequality for the second term in (21) and the fact that $\delta^{k+1} < \delta^k$ in Algorithm 1. The last inequality follows the positivity of the matrix $(W^k + \frac{1}{\lambda}\hat{X}^T X)$. The above equation indicates that $\mathcal{L}(c_i, \delta)$ is non-increasing, which leads to $\lim_{k\to\infty} \left\|c_i^k - c_i^{k+1}\right\|_2 = 0$. Thus, we have

$$\left\|(c_i^k)^2\right\|_{\frac{p}{2}}^p < \left\|(c_i^k)^2 + \delta^k\right\|_{\frac{p}{2}}^p \leqslant \mathcal{L}(c_i^k, \delta^k) \leqslant \mathcal{L}(c_i^0, \delta^0), \qquad (23)$$

which implies that the sequence $\{c_i\}$ is bounded.

Theorem 3.2. For given λ and $0 < p < 1$, the limit c_i^* of any convergent subsequence is a stationary point of the problem in (9).

Proof. Since the sequence $\{c_i^k\}$ is bounded there exists a subsequence $\{c_i^{k_j}\}$, such that $\lim_{j\to\infty} c_i^{k_j} = c_i^*$. Note the fact that $\lim_{k\to\infty} \left\|c_i^k - c_i^{k+1}\right\|_2 = 0$ in Theorem 3.1, which implies that the subsequence $\{c_i^{k_j+1}\}$ also converging to c_i^*. At the k_j-th iteration, $c_i^{k_j+1}$ solves (13), i.e.,

$$W^{k_j} c_i^{k_j+1} - \frac{1}{\lambda}\hat{X}^T(\hat{x}_i - \hat{X}c_i^{k_j+1}) + \gamma^{k_j+1} e_i = 0. \qquad (24)$$

Let $j \to \infty$. We have

$$W^* c_i^* - \frac{1}{\lambda}\hat{X}^T(\hat{x}_i - \hat{X}c_i^*) + \gamma^* e_i = 0, \qquad (25)$$

where $[W^*]_{jj} = p[(c_{ji}^*)^2 + \delta^*]^{\frac{p-2}{2}}$, and γ^* is defined in (16). Namely, c_i^* is a stationary point of (9).

In Algorithm 1, we decrease δ from a larger value 1 by $\delta^{k+1} = \delta^k/\rho$ with $\rho > 1$, which intuitively verifies that the smoothed ℓ_p minimization problem (9) is close to the non-smoothed one (8).

4.2. Complexity analysis

We evaluate the main time complexity of SSC-SLp and other related methods SSC [4] and ℓ_qSSC [7]. SSC adopts the ADM method with the main time complexity of $O(3mn^2 + n^3)$, and ℓ_qSSC uses the integration of the ADM and IRLS with the main time complexity of $O(4mn^2 + 2n^3)$. Our SSC-SLp adopts a simple and effective iterative algorithm with the main time complexity of $O(2(m+1)n^2 + n^3)$, and requires fewer iterations. Numerical results in the next section also validate that our SSC-SLp is faster than SSC and ℓ_qSSC.

5. Experimental verification and analysis

In this section, we evaluate the effectiveness of our SSC-SLp on synthetic data and real datasets including Extended Yale B (Yale B) [9], AR [17], COIL-20 (COIL) [19] and USPS [6]. For comparison, several stat-of-the-art algorithms, LSR [15] (Least Squares Regression), LRR [12] (ADM), SSC [4] (ADM), OMP-SSC [28] (Orthogonal Match Pursuit), ℓ_0-SSC [27] (Proximal Gradient Descent) and ℓ_qSSC [7] (ADM and IRLS), are introduced. Clustering Error (CE) (or Average Clustering Error (ACE) and Median (Med.)) [4] are used as performance measures. For a fair comparison, all experiments are conducted on a PC with Intel (R) Core (TM) i5-4460 CPU @ 3.20 GHz and 12.0 GB RAM, and all reported experimental results are the average of 30 experiments.

5.1. Experimental details

In all experiments, the parameters of the competing algorithms correspond to the smallest clustering error in most cases, where $q = 0.3$ for ℓ_qSSC. For our SSC-SLp, we set $\lambda = 7$, $\rho = 1.1$ for Yale B; $\lambda = 0.05, \rho = 1.1$ for AR; $\lambda = 3, \rho = 1.5$ for COIL; $\lambda = 8, \rho = 1.5$ for synthetic data and USPS; $\beta = 10^6$ and $p \in \{0.3, 0.5, 0.7\}$ for all the experiments. Although a smaller p can make the ℓ_p norm closer to the ℓ_0 one, it also leads to (8) more nonconvex, and thus more difficult to solve. Moreover, for LSR, ℓ_qSSC and SSC-SLp, we reduce the dimensionality of real datasets by PCA.

5.2. Experiment on synthetic data

In this experiment, we evaluate the effect of noise levels on the convergence and clustering performance of the proposed algorithm. According to [11] and [14], we generate 5 independent subspaces $\{S_i\}_{i=1}^5 \in \mathbb{R}^{200}$ and sample 20 data points from each of them to form synthetic data, where each subspace has a rank of 5. We randomly select 50% samples to corrupt using Gaussian noise via $z = x + \beta\mathbf{n}$, where x is the selected sample, β denote the corruption ratio (%) increasing from 10% to 90% with a 20% interval, and $\mathbf{n}$ is the noise which follows standard normal distribution. Thus we can obtain 6 synthetic datasets (including clean dataset). We apply SSC-SLp to each dataset.

Fig.1 (A)-(C) illustrate the convergence curves of SSC-SLp on synthetic data with corruption ratio $\beta = 0, 50\%$ and 90%, respectively, and Fig.1 (D) shows the clustering error of SSC-SLp under different corruption ratios. From Fig.1, we have the following observations: First, SSC-SLp can converge well under different noise levels and values of p, which demonstrates the convergence of SSC-SLp, as shown in Theorems 3.1 and 3.2. Second, the number of iterations increases slightly with the decrease of p because a smaller p makes the objective function more difficult to solve. Third, for the case of $p = 0.3$, the clustering error of SSC-SLp increases from 0 to 33.0% with the increase of β from 0 to 90%. Compared with $p = 0.3$, the clustering error is slightly larger for $p = 0.5$ and 0.7. This indicates that the proposed algorithm is considerably robust to noise.

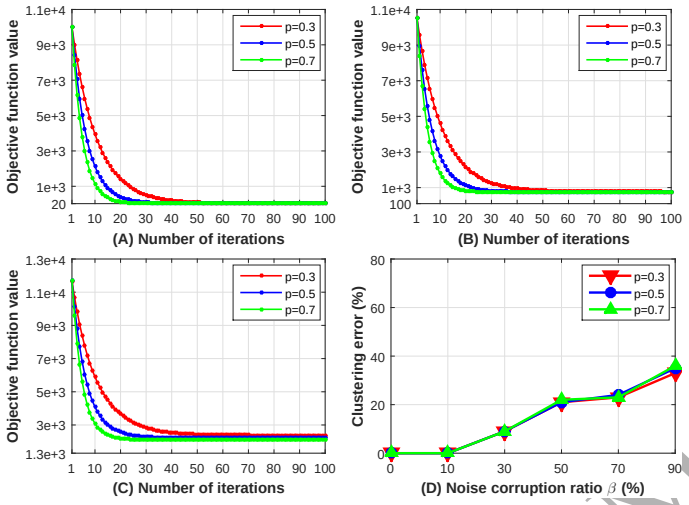

Fig. 1. Objective function values of SSC-SLp versus the number of iterations obtained on synthetic data with corruption ratio $\beta = 0, 50\%$ and 90% from (A) to (C), and clustering error of SSC-SLp versus corruption ratio $\beta \in \{0, 10\%, 30\%, 50\%, 70\%, 90\%\}$ in (D).

5.3. Experiments on real datasets

The Yale B is a face dataset which contains 2432 cropped frontal face images taken from 38 subjects under 64 illumination conditions. Each image is downsized from 192×168 to 48×42. According to subjects, we divide them into 4 groups: 1-10, 11-20, 21-30 and 31-38, and consider all the possible choices for $N \in \{2, 3, 5, 8, 10\}$ in each group, as done in [4]. Thus we can obtain a set of data for each N. For AR dataset, we use all face images (1300 images) of the first 100 subjects taken in the first stage. The COIL dataset consists of 1440 images with 32×32 pixels of 20 different objects taken 5 degrees apart, where the image background is discarded. For the USPS dataset, we adopt the first 100 images with 16×16 pixels for each digit from 0 to 9. Fig. 2 shows some sample images drawn from the above datasets.

5.4. Analysis of experimental results

In this section, extensive experiments are conducted to demonstrate the superior performance of SSC-SLp.

Fig. 2. Some sample images drawn from the following datasets, from top to bottom: Yale B, AR, COIL and USPS.

1) Performance Comparison. From Table 1 for the Yale B dataset, we can observe that: (A) Our SSC-SLp achieves favourable clustering performance; (B) Compared with SSC, the clustering error of SSC-SLp shows a significant decrease except when clustering 2 subjects with $p = 0.7$. This indicates that the smoothed ℓ_p $(0 < p < 1)$ norm is more helpful for enhancing the clustering performance than the ℓ_1 one; (C) When p decreases from 0.7 to 0.3, the clustering error of SSC-SLp also decreases, which resonates with the intuition that a smaller p makes the ℓ_p norm closer to the ℓ_0 one. (D) In contrast to other methods based on sparse representation (SSC-OMP, ℓ_0-SSC and ℓ_qSSC), our method obtains better clustering performance; (E) The above observation can also be verified by the extensive experimental results in Table 2 on the AR, COIL and USPS datasets. In addition, $p = 0.5$ gives a slightly lower accuracy than $p = 0.7$ on both COIL and USPS datasets. This do not counter the above intuition because a smaller ρ can further improve the clustering performance (see the next section).

Table 1. Average clustering error (%) on the Yale B dataset

Method		LSR	LRR	SSC	SSC-OMP	ℓ_0-SSC	ℓ_q SSC	SSC-SLp (p) 0.3	0.5	0.7
2	ACE	5.9	2.1	1.9	5.2	7.9	1.6	**1.4**	1.7	2.4
	Med.	6.3	0.8	**0.0**	0.8	0.8	**0.0**	0.8	0.8	0.8
3	ACE	9.3	3.5	3.3	5.4	11.0	2.5	**2.2**	2.5	3.2
	Med.	9.4	2.1	**0.5**	2.1	4.2	1.0	1.0	1.6	1.6
5	ACE	17.9	5.9	4.3	7.4	13.8	4.6	**2.9**	3.4	3.9
	Med.	18.4	5.0	2.7	3.4	7.2	2.5	**2.2**	2.5	2.8
8	ACE	29.1	11.1	5.9	9.8	14.5	7.5	**3.2**	3.8	4.2
	Med.	29.5	7.4	4.5	5.9	8.6	4.0	**2.6**	3.0	3.9
10	ACE	32.6	16.9	10.9	11.3	14.7	7.6	**3.7**	3.8	4.3
	Med.	35.6	18.9	5.6	13.9	10.0	4.1	**2.7**	2.8	3.4

Table 2. Clustering error (%) on AR, COIL and USPS datasets

Method		LSR	LRR	SSC	SSC-OMP	ℓ_0-SSC	ℓ_q SSC	SSC-SLp (p) 0.3	0.5	0.7
AR	CE	30.5	46.0	30.9	47.1	51.4	34.2	**18.5**	20.0	21.8
COIL	CE	38.0	41.0	12.2	49.5	14.7	14.1	**8.1**	8.3	8.1
USPS	CE	26.9	25.5	27.7	18.5	25.6	24.0	**7.6**	10.6	9.5

2) Sensitivity Analysis. In this experiment, we examine the sensitivity of the parameters λ and ρ in Algorithm 1. The parameter λ is used to balance the two parts of the objective function (9), and ρ is used to control the descent rate of the regularization parameter δ. We fix one parameter and vary the other. The evaluation results over the first 10 subjects in the Yale B dataset are shown in Fig. 3 from which we can see that, for all $p \in \{0.3, 0.5, 0.7\}$, the obtained solution is considerably stable when $\lambda \in [0.01, 14]$; but ρ is sensitive because too large ρ may cause the algorithm to fall into local minima, and thus degenerate clustering performance. Moreover, in our experimental details, we do not select too small ρ, since it may slow down the convergence rate.

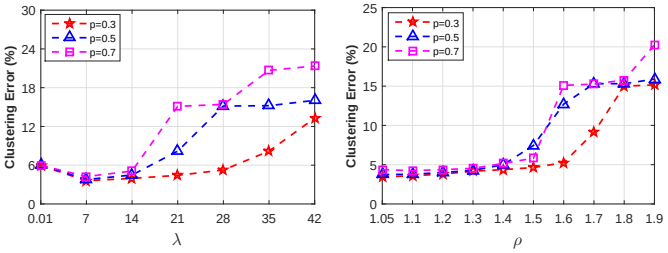

Fig. 3. Clustering error of SSC-SLp versus parameters involving 10 subjects in the Yale B dataset. From left to right: λ varying and ρ varying.

3) Time Comparison. As solving the representation matrix is a dominant time-consuming factor in the evaluation of the objective function (9), we compare the computational time for solving the representation matrix incurred by all algorithms on the Yale B dataset. Fig. 4 shows that our method finds the solution relatively fast compared to the ADM based LRR, SSC and ℓ_qSSC, which conforms the observation in Section 5.2.

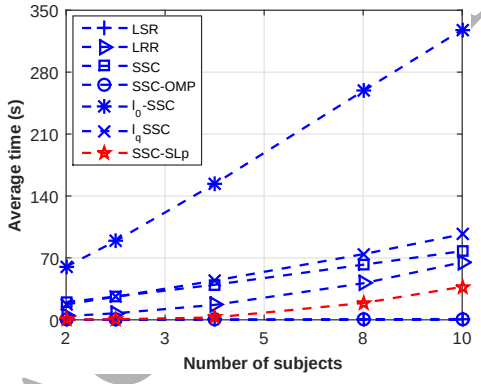

Fig. 4. Average computational time (s) required to find the representation matrix on Yale B, where $q = 0.3$ for ℓ_qSSC and $p = 0.3$ for SSC-SLp.

6. Conclusion

We evaluated sparse subspace clustering via smoothed ℓ_p minimization and presented a unified formulation for different practical clustering problems. Importantly, the proposed unified formulation simplifies the analysis of the algorithm involving multiple variables and non-smooth terms. A simple and effective iterative algorithm underpinned by a theoretical analysis is presented for the proposed unified formulation. The numerical experimental results demonstrate the advantages of our method over most state-of-the-art algorithms. In addition, the proposed method can be extended to many applications in the fields of compressed sensing and signal processing.

Acknowledgments

This paper is jointly supported by the 111 Project of Chinese Ministry of Education under Grant B12018 and the National Natural Science Foundation of China under Grant 61373055; 61672265.

References

[1] Brbic, M., Kopriva, I., 2018. Multi-view low-rank sparse subspace clustering. Pattern Recognition 73, 247–258.
[2] Chen, G., Lerman, G., 2009. Spectral curvature clustering (SCC). International Journal of Computer Vision 81, 317–330.
[3] Costeira, J.P., Kanade, T., 1998. A multibody factorization method for independently moving objects. International Journal of Computer Vision 29, 159–179.
[4] Elhamifar, E., Vidal, R., 2013. Sparse subspace clustering: Algorithm, theory, and applications. IEEE transactions on pattern analysis and machine intelligence 35, 2765–2781.
[5] Han, J., Pei, J., Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.
[6] Hull, J.J., 1994. A database for handwritten text recognition research. IEEE Transactions on pattern analysis and machine intelligence 16, 550–554.
[7] Kuang, S., Chao, H., Yang, J., 2017. Efficient lq norm based sparse subspace clustering via smooth IRLS and ADMM. Multimedia Tools and Applications 76, 23163–23185.
[8] Lai, M.J., Xu, Y., Yin, W., 2013. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. SIAM Journal on Numerical Analysis 51, 927–957.
[9] Lee, K.C., Ho, J., Kriegman, D.J., 2005. Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on Pattern Analysis & Machine Intelligence , 684–698.
[10] Li, C.G., Vidal, R., 2015. Structured sparse subspace clustering: A unified optimization framework, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 277–286.
[11] Lin, Z., Liu, R., Su, Z., 2011. Linearized alternating direction method with adaptive penalty for low-rank representation, in: Advances in neural information processing systems, pp. 612–620.
[12] Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y., 2013. Robust recovery of subspace structures by low-rank representation. IEEE transactions on pattern analysis and machine intelligence 35, 171–184.
[13] Liu, J., Chen, Y., Zhang, J., Xu, Z., 2014. Enhancing low-rank subspace clustering by manifold regularization. IEEE Transactions on Image Processing 23, 4022–4030.
[14] Lu, C., Lin, Z., Yan, S., 2015. Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. IEEE Transactions on Image Processing 24, 646–654.
[15] Lu, C.Y., Min, H., Zhao, Z.Q., Zhu, L., Huang, D.S., Yan, S., 2012. Robust and efficient subspace segmentation via least squares regression, in: European conference on computer vision, Springer. pp. 347–360.
[16] Luo, M., Nie, F., Chang, X., Yang, Y., Hauptmann, A.G., Zheng, Q., 2018. Adaptive unsupervised feature selection with structure regularization. IEEE transactions on neural networks and learning systems 29, 944–956.
[17] Martinez, A.M., 1998. The ar face database. CVC Technical Report24 .
[18] Nasrabadi, N.M., 2007. Pattern recognition and machine learning. Journal of electronic imaging 16, 049901.
[19] Nene, S.A., Nayar, S.K., Murase, H., et al., 1996. Columbia object image library (coil-20) .

[20] Ng, A.Y., Jordan, M.I., Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm, in: Advances in neural information processing systems, pp. 849–856.
[21] Shi, D., Wang, J., Cheng, D., Gao, J., 2017. A global-local affinity matrix model via eigengap for graph-based subspace clustering. Pattern Recognition Letters 89, 67–72.
[22] Tipping, M.E., Bishop, C.M., 1999. Mixtures of probabilistic principal component analyzers. Neural computation 11, 443–482.
[23] Tseng, P., 2000. Nearest q-flat to m points. Journal of Optimization Theory and Applications 105, 249–252.
[24] Vidal, R., Favaro, P., 2014. Low rank subspace clustering (LRSC). Pattern Recognition Letters 43, 47–61.
[25] Wang, Y.X., Xu, H., 2016. Noisy sparse subspace clustering. The Journal of Machine Learning Research 17, 320–360.
[26] Wang, Y.X., Xu, H., Leng, C., 2013. Provable subspace clustering: When LRR meets SSC, in: Advances in Neural Information Processing Systems, pp. 64–72.
[27] Yang, Y., Feng, J., Jojic, N., Yang, J., Huang, T.S., 2016. ℓ_0-sparse subspace clustering, in: European Conference on Computer Vision, Springer. pp. 731–747.
[28] You, C., Robinson, D., Vidal, R., 2016. Scalable sparse subspace clustering by orthogonal matching pursuit, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3918–3927.
[29] Zhang, Y., 2010. Recent advances in alternating direction methods: Practice and theory, in: IPAM Workshop on Continuous Optimization.
[30] Zhu, W., Lu, J., Zhou, J., 2018. Nonlinear subspace clustering for image clustering. Pattern Recognition Letters 107, 131–136.
[31] Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. Annals of statistics 36, 1509.