Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9216351, 12 pages
https://doi.org/10.1155/2020/9216351

Research Article

A Novel Convex Clustering Method for High-Dimensional Data Using Semiproximal ADMM

Huangyue Chen,[1] **Lingchen Kong**,[1] **and Yan Li**[2]

[1]*Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, China*
[2]*School of Insurance and Economics, University of International Business and Economics, Beijing 100029, China*

Correspondence should be addressed to Huangyue Chen; hychen@bjtu.edu.cn

Received 5 April 2020; Revised 2 July 2020; Accepted 7 July 2020; Published 21 September 2020

Academic Editor: Bogdan Smolka

Clustering is an important ingredient of unsupervised learning; classical clustering methods include K-means clustering and hierarchical clustering. These methods may suffer from instability because of their tendency prone to sink into the local optimal solutions of the nonconvex optimization model. In this paper, we propose a new convex clustering method for high-dimensional data based on the sparse group lasso penalty, which can simultaneously group observations and eliminate noninformative features. In this method, the number of clusters can be learned from the data instead of being given in advance as a parameter. We theoretically prove that the proposed method has desirable statistical properties, including a finite sample error bound and feature screening consistency. Furthermore, the semiproximal alternating direction method of multipliers is designed to solve the sparse group lasso convex clustering model, and its convergence analysis is established without any conditions. Finally, the effectiveness of the proposed method is thoroughly demonstrated through simulated experiments and real applications.

1. Introduction

Clustering is an important ingredient of unsupervised learning. It assigns samples into different clusters by minimizing the differences in the same cluster and maximizing the differences between different clusters. As an exploratory data analysis technique, it has been applied in many fields, such as image processing, energy engineering, and social networks [1–3]. To date, a wide variety of clustering methods have been introduced, including classical clustering methods such as K-means clustering [4, 5] and hierarchical clustering [6, 7]. However, the convexity of the corresponding optimization models cannot be guaranteed in general, so their global optimal solutions are hard to find. In addition, the performances of these methods are strongly dependent on their initial settings. To overcome the aforementioned disadvantages, convex clustering (CC), which provides a convex optimization perspective toward the task of clustering via shrinkage or regularization techniques, has been considered by many researchers, e.g., [8–11]. Its corresponding optimization model is given by

$$\min_{X\in\mathbb{R}^{n\times p}} f(X) := \frac{1}{2}\|X-A\|_F^2 + \gamma_1 \sum_{\ell\in\Theta} \omega_\ell \left\| X_{\ell_1\cdot} - X_{\ell_2\cdot} \right\|_q, \quad (1)$$

where A is a given data matrix with n observations and p features; $\|\cdot\|_F$ denotes the Frobenius norm of the matrix; $\gamma_1 \geq 0$ is a tuning parameter that controls the balance between the model fit and the number of clusters; $\Theta = \{\ell = (\ell_1, \ell_2) \mid 1 \leq \ell_1 < \ell_2 \leq n\}$ is the index set; $\omega_\ell \geq 0, \ell \in \Theta$ are given weights that are generally chosen based on the given data matrix A; $X_{\ell_1\cdot}(X_{\cdot\ell_1})$ is the ℓ_1th row (column) of matrix X; and the most common choices of q are 1, 2, and ∞, which ensure the convexity of model (1). Let $\widehat{X}$ be an optimal solution of (1); then, $\widehat{X}_{i\cdot} = \widehat{X}_{j\cdot}$ indicates that the observations $A_{i\cdot}$ and $A_{j\cdot}$ are assigned to the same cluster. Note from (1) that if we set $\gamma_1 = 0$, this will lead to the trivial solution $\widehat{X} = A$, i.e., the partition of the observations into

singletons. As γ_1 increases, some rows of $\widehat{X}$ become identical. If we set γ_1 to be sufficiently large, all rows of $\widehat{X}$ become identical, which indicates that all observations are lumped into a single cluster. Hence, the clustering path can be obtained by solving (1) with a range of tuning parameters, and then the number of clusters is determined by learning from the data rather than being given in advance. This fascinating characteristic of convex clustering has attracted increasing attention, see, e.g., [12–21].

The theoretical results on cluster recovery of (1) with $\omega_\ell = 1$ had been established in [12, 13]. Later, Sun et al. [14] extended these results to the general weighted convex clustering model. Furthermore, in [15], the authors analyzed the statistical properties of (1) with $\omega_\ell = 1$, such as an unbiased estimator of the degrees of freedom and a finite sample bound for the prediction error. The large sample behavior of (1) with l_1 fusion penalization is presented in [16]. In [17], the authors present conditions that guarantee that the convex clustering solution path recovers a tree and explicitly describes how the affinity parameters modulate the structure of the recovered tree. Meanwhile, some researchers concentrated on the solution algorithms of (1). For example, Chi and Lange [10] adopted the Alternating Direction Method of Multipliers (ADMM) and Alternating Minimization Algorithm (AMA) to solve (1). More recently, in [14, 18], the semismooth Newton-based augmented Lagrangian method was applied to (1) and outperformed ADMM and AMA in efficiency, scalability, and robustness. Moreover, the idea of convex clustering was applied to process missing data [19], binary data [20], and outlier detection [21].

Although convex clustering (1) has desirable theoretical results and performs well computationally, Wang et al. [22] pointed out that the clustering performance may be destroyed in high-dimensional scenarios where the number of features is much more than observations. The noninformative features included in the clustering are the key reason behind why convex clustering (1) is disheartening. Hence, eliminating some noninformative features is indispensable for high-dimensional data clustering. This motivates us to propose a more reasonable convex clustering method that can perform cluster analysis and feature screening simultaneously.

In this paper, we propose the Sparse Group Lasso Convex Clustering (SGLCC) method by adopting the sparse group lasso penalty [23]. The optimization problem is summarized as

$$\min_{X\in\mathbb{R}^{n\times p}} f(X) + \gamma_2 \sum_{j=1}^{p}\left[(1-\alpha)u_j\left\|X_{\cdot j}\right\|_2 + \alpha\left\|X_{\cdot j}\right\|_1\right], \qquad (2)$$

where the tuning parameter γ_2 controls the number of informative features and $u = (u_1, u_2, \ldots, u_p)^T \in \mathbb{R}^p$ is the weight vector. The form of the penalty term is much like the elastic-net penalty [24], and $\|X_{\cdot j}\|_2$ and $\|X_{\cdot j}\|_1$ denote the Euclidean norm and ℓ_1 norm of vector $X_{\cdot j}$, respectively. The parameter $0 \le \alpha \le 1$ controls the balance between the group lasso [25, 26] and the lasso [27]. Obviously, convex clustering (1) is a special case of (2) if we set $\gamma_2 = 0$. Unlike (1), model (2) can perform variable selection. When $\alpha = 0$, model (2) reduces to sparse convex clustering (SCC) [22]. SCC uses the group lasso penalty to promote group sparsity, that is, all components of the solution $\widehat{X}_{\cdot j}$ are either zero or nonzero. However, it cannot induce ubiquitous within-group sparsity in genetic data. For example, a biological pathway may be implicated in the progression of a particular type of cancer, but not all genes in the pathway need to be active [28]. The sparse group lasso penalty is designed to achieve such within-group sparsity. Therefore, we believe that the proposed clustering model (2) can significantly improve the clustering performance of model (1). It is worth noting that solving the optimization problem (2) is more challenging than convex clustering (1) and SCC. If the optimization algorithms utilized in [10, 22] are adopted to solve (2) directly, the computation efficiency may be poor because the subproblem might have no closed-form solution and the step length is too small. Hence, we developed the steps in this paper along with the theory, algorithm, and applications of SGLCC.

The main contributions of the paper are summarized as follows:

(1) For the proposed SGLCC, we obtain a finite sample error bound and feature screening consistency under mild conditions. Our results are not only applicable to the more practical SGLCC model but also subsume the known results of (1) and SCC as special cases.

(2) We adopt the semiproximal alternating direction method of multipliers to solve (2), and its convergence analysis is established without any additional conditions. The subproblem per iteration of this algorithm has a closed-form solution and can be implemented efficiently.

(3) The experimental results on both synthetic and real datasets illustrate that SGLCC provides superior clustering performance and feature selection abilities to other clustering methods.

The remainder of this paper is organized as follows. In Section 2, we summarize some preliminaries and notations. We study the finite sample error bound and feature screening consistency of our proposed method in Section 3. In Section 4, we introduce an efficient optimization algorithm for solving (2). After that, in Section 5, we conduct numerical experiments to verify the effectiveness of the proposed method in synthetic datasets and four microarray datasets. Conclusions are given in Section 6. The proof of the main results can be found in the Supplementary Materials (available here).

2. Preliminaries

In this section, we introduce some preliminaries that are used later in the paper.

For convenience, we start with some necessary notations. For a vector $x \in \mathbb{R}^n$ and $q \ge 1$, $\|x\|_q$ denotes the ℓ_q norm of x and $1_n \in \mathbb{R}^n$ denotes a vector with all components equal to one. Let $|x| = (|x_1|, |x_2|, \ldots, |x_n|)^T$. For a matrix $X \in \mathbb{R}^{n\times m}$, $X_{i\cdot}$ ($X_{\cdot j}$) is the ith row (the jth column) of X, $\|X\|_F$ denotes

the Frobenius norm of X, $\|X\|_2$ denotes the spectral norm of X, and the vectorization of X is defined by $\mathrm{vec}(X) = (X_{11}, X_{21}, \ldots, X_{n1}, X_{12}, \ldots, X_{nm})^T$. The symbols o and $\otimes$ denote the Hadamard product and Kronecker product, respectively. The linear operator $\mathscr{H}: \mathbb{R}^{n\times p} \longrightarrow \mathbb{R}^{|\Theta|\times p}$ is given by

$$\mathscr{H}(X) = \left[X_{\ell_1\cdot} - X_{\ell_2\cdot}\right]_{(\ell_1,\ell_2)\in\Theta} = HX. \tag{3}$$

The ℓth row of H is denoted as $H_{\ell\cdot} = [e_{\ell_1} - e_{\ell_2}]^T$, $\ell = (\ell_1, \ell_2) \in \Theta$, and e_{ℓ_1} is a vector with the ℓ_1th component equal to one and the rest equal to zero. The adjoint operator of $\mathscr{H}$ is given by $\mathscr{H}^*: \mathbb{R}^{|\Theta|\times p} \longrightarrow \mathbb{R}^{n\times p}$, that is, $\mathscr{H}^*(Y) = H^T Y$.

Next, some definitions and results from convex analysis [29] are provided. Proximal mapping is important for designing optimization algorithms and has been well studied. For a proper closed convex function $g: \mathbb{R}^n \longrightarrow (-\infty; +\infty]$, the proximal mapping $\mathrm{Prox}_{tg}(x)$ for g at any $x \in \mathbb{R}^n$ with $t > 0$ is defined by

$$\mathrm{Prox}_{tg}(x) = \arg\min_y \left\{ tg(y) + \frac{1}{2}\|x - y\|_2^2 \right\}. \tag{4}$$

In particular, if g is an indicator function of the set $\mathscr{C}$ ($g(x) = 0$ if $x \in \mathscr{C}$; otherwise $g(x) = \infty$); then, the proximal mapping reduces to the projection operator onto $\mathscr{C}$. The Fenchel conjugate of g is defined by

$$g^*(y) = \sup_x \{\langle x, y\rangle - g(x)\}. \tag{5}$$

A key property that connects the proximal mapping and Fenchel conjugate is the so-called Moreau decomposition:

$$x = \mathrm{Prox}_{tg}(x) + t\mathrm{Prox}_{t^{-1}g^*}\left(t^{-1}x\right). \tag{6}$$

For example, if $g(x) = \|x\|_q$ with $q \geq 1$, from Moreau decomposition, we obtain

$$\mathrm{Prox}_{tg}(x) = x - t\mathrm{Prox}_{t^{-1}g^*}\left(t^{-1}x\right) = x - \Pi_{t\mathscr{B}}(x), \tag{7}$$

where $\mathscr{B} = \left\{x \mid \|x\|_{q^*} \leq 1\right\}$ and $(1/q) + (1/q^*) = 1$.

3. Statistical Properties

In this section, we study the finite sample error bound and feature screening consistency of sparse group lasso convex clustering (2). We start with the following conditions that facilitate the technical proofs.

A1: assume that $\mathrm{vec}(A) = x^* + \varepsilon$, where $x^* = \mathrm{vec}(X^*) \in \mathbb{R}^{np}$ is the mean vector and $\varepsilon \in \mathbb{R}^{np}$ is an error vector of independent sub-Gaussian random variables with mean zero and variance δ^2

A2: only s ($s \ll p$) features are informative; the informative feature set and its complementary set are then denoted as $\mathscr{I}$ and $\mathscr{I}^c$, respectively

The condition **A1** implies that the data matrix A is composed of sub-Gaussian random variables and is commonly used in the literature, see, e.g., [15, 22]. In high-dimensional scenarios, only a few features are active in general. Thus, condition **A2** has been widely used in high-dimensional data analysis, see, e.g., [22, 30]. For simplicity, as described in the literature [12, 13, 15, 22], we consider the case with $\omega_\ell = 1$.

3.1. Bounds for Prediction Error. The following theorems provide the finite sample bounds for prediction error of model (2), and our theoretical results subsume the existing results for CC and SCC as special cases.

Theorem 1. *Suppose that condition A1 is satisfied. Let $\widehat{X}$ be the solution of model (2) with $q = 1$. If $\gamma_1 \geq 2\delta\sqrt{(\log(p|\Theta|))/n}$, then there exists positive constants b_1 and b_2 such that*

$$P\left(\frac{c_1}{np}\left\|\widehat{X} - X^*\right\|_F^2 \leq \frac{4\gamma_1}{np}\left\|\mathrm{vec}\left(HX^*\right)\right\|_1 + \frac{2\gamma_2\alpha}{np}\left\|\mathrm{vec}\left(X^*\right)\right\|_1 + c_2 + c_3\right) \geq 1 - c_4. \tag{8}$$

Here,

$$\begin{aligned} c_1 &= 1 + \gamma_2(\alpha - 1), \\ c_2 &= 2\delta^2\left[\frac{1}{n} + \sqrt{\frac{\log(np)}{n^2 p}}\right] + \frac{2}{\sqrt{np}}, \\ c_3 &= \frac{\gamma_2(1-\alpha)}{np}\|u\|_2^2, \\ c_4 &= \exp\left\{-\min\left(b_1\log(np), b_2\sqrt{p\log(np)}\right)\right\} \\ &\quad + \frac{\left(\gamma_2(1-\alpha)\|u\|_1 + \gamma_2\alpha\right)^2\delta^2}{np} + \frac{2}{p|\Theta|}. \end{aligned} \tag{9}$$

From Theorem 1, we can observe that the average prediction error is bounded, and c_2 tends to zero as $n, p \longrightarrow \infty$. We know that $\|\mathrm{vec}(X^*)\|_1 = O(n)$ and $\|\mathrm{vec}(HX^*)\|_1 = O(n^2)$ by combining condition **A2** with $|\Theta| = n(n-1)/2$. Model (2) with $q = 1$ has prediction consistency only if $0 \leq \gamma_2 < (1/1-\alpha)$, $\gamma_2 = O(1)$, $(\gamma_2(1-\alpha)/np)\|u\|_1^2 = o(1)$, and $\sqrt{(n\log(p|\Theta|)/p^2)} = o(1)$. To be specific, we find that $c_1 > 0, c_1 = O(1)$ and $(2\gamma_2\alpha/np)\|\mathrm{vec}(X^*)\|_1$ tend to zero as $n, p \longrightarrow \infty$ by $0 \leq \gamma_2 < (1/1-\alpha)$ and $\gamma_2 = O(1)$. If $(\gamma_2(1-\alpha)/np)\|u\|_1^2 = o(1)$, then

$$0 \leq c_3 \leq \frac{\gamma_2(1-\alpha)}{np}\|u\|_1^2 = o(1), \tag{10}$$

and c_4 decays to zero as $n, p \longrightarrow \infty$. Additionally, by including $\gamma_1 \geq 2\delta\sqrt{(\log(p|\Theta|)/n)}$, the condition $\sqrt{(n\log(p|\Theta|)/p^2)} = o(1)$ implies that $(4\gamma_1/np)\|\mathrm{vec}\ (HX^*)\|_1 = o(1)$.

The next theorem presents the finite sample bound for the prediction error of model (2) with $q = 2$.

Theorem 2. *Suppose that condition A1 is satisfied. Let $\widehat{X}$ be the solution of model (2) with $q = 2$. If $\gamma_1 \geq 2\delta\sqrt{(p\log(p|\Theta|)/n)}$, then there exists positive constants b_1 and b_2 such that*

$$P\left(\frac{c_1}{np}\left\|\widehat{X} - X^*\right\|_F^2 \leq \frac{4\gamma_1}{np}\sum_{\ell\in\Theta}\left\|H_{\ell\cdot}X^*\right\|_2 + \frac{2\gamma_2\alpha}{np}\left\|\mathrm{vec}\left(X^*\right)\right\|_1 + c_2 + c_3\right) \geq 1 - c_4. \tag{11}$$

We only discuss the first term on the right-hand side of (11) in Theorem 2 because all other terms are the same as (8) in Theorem 1. We find that $\|H_{\ell\cdot}X^*\|_2 = O(1)$ by condition **A2**. Note that $|\Theta| = (n(n-1)/2)$, and thus $\sum_{\ell\in\Theta}\|H_{\ell\cdot}X^*\|_2 = O(n^2)$. Suppose that $\sqrt{(n\log(p|\Theta|)/p)} = o(1)$, we obtain that $(4\gamma_1/np)\sum_{\ell\in\Theta}\|H_{\ell\cdot}X^*\|_2 = o(1)$ by combining $\gamma_1 \geq 2\delta\sqrt{(p\log(p|\Theta|)/n)}$.

Remark 1

(1) In high-dimensional scenarios, the number of features p is considerably larger than the number of observations n. If $p = O(n^r)$, $r > 1$, then conditions $\sqrt{(n\log(p|\Theta|)/p^2)} = o(1)$ and $\sqrt{(n\log(p|\Theta|)/p)} = o(1)$ hold automatically, and this condition can be found in [15].

(2) This intuitive illustration implies that the condition $(\gamma_2(1-\alpha)/np)\|u\|_1^2 = o(1)$ requires that the weight u not be too large.

(3) Theorems 1 and 2 subsume the known results for CC and SCC by taking $\gamma_2 = 0$ and $\alpha = 0$, respectively. In addition, we obtain encouraging results provided in Corollary 1 when $\alpha = 1$.

The following corollary follows directly from Theorems 1 and 2 by letting $\alpha = 1$.

Corollary 1. *Suppose that conditions A1-A2 are satisfied. Let $\widehat{X}$ be the solution of model (2) with $\alpha = 1$. We obtain the following statements.*

(1) When $q = 1$, if $\gamma_1 \geq 2\delta\sqrt{(\log(p|\Theta|)/n)}$, $\gamma_2 = O(1)$, and $\sqrt{(n\log(p|\Theta|)/p^2)} = o(1)$, then, as $n, p \longrightarrow \infty$,

$$P\left(\frac{1}{np}\left\|\widehat{X} - X^*\right\|_F^2 \leq \frac{4\gamma_1}{np}\left\|\mathrm{vec}\left(HX^*\right)\right\|_1 + \frac{2\gamma_2}{np}\left\|\mathrm{vec}\left(X^*\right)\right\|_1 + c_2\right) \longrightarrow 1. \tag{12}$$

(2) When $q = 2$, if $\gamma_1 \geq 2\delta\sqrt{(p\log(p|\Theta|)/n)}$, $\gamma_2 = O(1)$, and $\sqrt{\log(p|\Theta|)/np} = o(1)$, then, as $n, p \longrightarrow \infty$,

$$P\left(\frac{1}{np}\left\|\widehat{X} - X^*\right\|_F^2 \leq \frac{4\gamma_1}{np}\sum_{\ell\in\Theta}\left\|H_{\ell\cdot}X^*\right\|_2 + \frac{2\gamma_2}{np}\left\|\mathrm{vec}\left(X^*\right)\right\|_1 + c_2\right) \longrightarrow 1. \tag{13}$$

As we can see from Corollary 1, SGLCC with $\alpha = 1$ has prediction consistency and only need some conditions that are easy to satisfy.

3.2. Feature Screening Consistency. Feature screening consistency is a desirable property in high-dimensional scenarios where many noninformative features need to be eliminated. This section demonstrates the asymptotic feature screening consistency of SGLCC.

Theorem 3. *Let $\widehat{X}$ be the solution of model (2) with $q = 1$. Suppose that conditions A1-A2 are hold. If $\gamma_1 \geq 2\delta\sqrt{(\log(p|\Theta|)/n)}$, $(4\gamma_1/np)\|vec(HX^*)\|_1 = o(1)$, $0 \leq \gamma_2 < (1/(1-\alpha))$, $\gamma_2 = O(1)$, and $(\gamma_2((1-\alpha)/n)p)\|u\|_1^2 = o(1)$, then $P(\|\widehat{X}_{\cdot j}\|_2 = 0) \longrightarrow 1$, for any $j \in \mathcal{I}^c$.*

Theorem 4. *Let $\widehat{X}$ be the solution of model (2) with $q = 2$. Suppose that conditions A1-A2 are hold. If $\gamma_1 \geq 2\delta\sqrt{(p\log(p|\Theta|)/n)}$, $(4\gamma_1/np)\sum_{\ell\in\Theta}\|H_{\ell\cdot}X^*\|_2 = o(1)$, $0 \leq \gamma_2 < (1/1-\alpha)$, $\gamma_2 = O(1)$, and $(\gamma_2((1-\alpha)/n)p)\|u\|_1^2 = o(1)$, then $P(\|\widehat{X}_{\cdot j}\|_2 = 0) \longrightarrow 1$, for any $j \in \mathcal{I}^c$.*

Theorems 3 and 4 give the asymptotic consistency of feature selection of SGLCC with $q = 1$ and $q = 2$, respectively. In this sense, SGLCC possesses the ability to correctly eliminate the noninformative features that distort clustering performance. Hence, we believe that the clustering performance of SGLCC is promising.

4. Algorithmic Design

In this section, we adopt the semiproximal alternating direction method of multipliers to solve (2), and its global convergence can be established. For simplicity, we focus on designing an algorithm to solve (2) with $q = 2$. The same algorithmic design and implementation can be applied to cases $q = 1$ and $q = \infty$ without difficulty.

4.1. Optimality Conditions. By ignoring the terms with $\omega_\ell = 0$, model (2) can be rewritten as

$$\min_{X\in\mathbb{R}^{n\times p}}\frac{1}{2}\|X - A\|_F^2 + \gamma_1\sum_{\ell\in\mathrm{E}}\omega_\ell\left\|X_{\ell_1\cdot} - X_{\ell_2\cdot}\right\|_2 + \gamma_2\sum_{j=1}^{p}\left[(1-\alpha)u_j\left\|X_{\cdot j}\right\|_2 + \alpha\left\|X_{\cdot j}\right\|_1\right], \tag{14}$$

where $\mathcal{E} = \{\ell = (\ell_1, \ell_2) \mid \omega_\ell > 0\}$. For the convenience of mathematical expression, we define the linear operator

$\mathcal{J}: \mathbb{R}^{n\times p} \longrightarrow \mathbb{R}^{|\mathcal{E}|\times p}$ and its adjoint operator $\mathcal{J}^*: \mathbb{R}^{|\mathcal{E}|\times p} \longrightarrow \mathbb{R}^{n\times p}$, respectively, by

$$\mathcal{J}(X) = \left[X_{\ell_1\cdot} - X_{\ell_2\cdot}\right]_{(\ell_1,\ell_2)\in E} = JX \text{ and } \mathcal{J}^*(Y) = J^T Y, \tag{15}$$

where $J_{\ell\cdot} = [e_{\ell_1} - e_{\ell_2}]^T$. Furthermore, model (13) can be reformulated as the following linearly constrained convex optimization problem with a separable structure:

$$\begin{aligned} \min_{X,Y,Z,W} \; & F(X,Y,Z,W) := \frac{1}{2}\|X - A\|_F^2 \\ & \qquad + Q_1(Y) + Q_2(Z) + Q_3(W), \\ & JX - Y = 0, \\ \text{s.t.} \quad & X - Z = 0, \\ & X - W = 0, \end{aligned} \tag{16}$$

where $Q_1(Y) = \gamma_1 \sum_{\ell\in\mathcal{E}} \omega_\ell \|Y_{\ell\cdot}\|_2$, $Q_2(Z) = \gamma_2(1-\alpha)\sum_{j=1}^p u_j \|Z_{\cdot j}\|_2$, and $Q_3(W) = \gamma_2\alpha\sum_{j=1}^p \|X_{\cdot j}\|_1$. The Lagrangian function associated with (16) is defined by

$$\begin{aligned} L(X,Y,Z,W;\lambda,\mu,\xi) = {} & \frac{1}{2}\|X - A\|_F^2 + Q_1(Y) + Q_2(Z) \\ & + Q_3(W) + \langle \lambda, JX - Y\rangle \\ & + \langle \mu, X - Z\rangle + \langle \xi, X - W\rangle, \end{aligned} \tag{17}$$

where $\lambda \in \mathbb{R}^{|\mathcal{E}|\times p}$, $\mu \in \mathbb{R}^{n\times p}$, and $\xi \in \mathbb{R}^{n\times p}$ are Lagrangian multipliers. For a given parameter $\sigma > 0$, the augmented Lagrangian function for (16) is defined by

$$\begin{aligned} L_\sigma(X,Y,Z,W;\lambda,\mu,\xi) = {} & L(X,Y,Z,W;\lambda,\mu,\xi) \\ & + \frac{\sigma}{2}\|JX - Y\|_F^2 + \frac{\sigma}{2}\|X - Z\|_F^2 \\ & + \frac{\sigma}{2}\|X - W\|_F^2. \end{aligned} \tag{18}$$

The Karush–Kuhn–Tucker (KKT) conditions for (16) are given by

$$\begin{cases} X - A + J^T\lambda + \mu + \xi = 0, \\ Y - \text{Prox}_{Q_1}(Y + \lambda) = 0, \\ Z - \text{Prox}_{Q_2}(Z + \mu) = 0, \\ W - \text{Prox}_{Q_3}(W + \xi) = 0, \\ JX - Y = 0,\ X - Z = 0,\ X - W = 0. \end{cases} \tag{19}$$

Let $\mathcal{F} := \{(X,Y,Z,W) \mid JX - Y = 0,\ X - Z = 0,\ X - W = 0\}$ be a feasible set of (16). Obviously, $\mathcal{F} \neq \emptyset$ and $F(X,Y,Z,W) \geq 0$. Then, we know from Corollaries 28.2.2 and 28.3.1 in [29] that $(\widehat{X},\widehat{Y},\widehat{Z},\widehat{W})$ is an optimal solution for (16) if and only if there exists a Lagrangian multiplier $(\widehat{\lambda},\widehat{\mu},\widehat{\xi})$ such that $(\widehat{X},\widehat{Y},\widehat{Z},\widehat{W},\widehat{\lambda},\widehat{\mu},\widehat{\xi})$ satisfies the KKT conditions (19). Hence, we could design a stopping criterion based on the KKT conditions in terms of guaranteeing the optimality of a point generated by the algorithm proposed in Section 4.2.

4.2. Algorithm and Convergence Analysis. Although the optimization problem (16) is a convex optimization problem with linear constraints, it is difficult to minimize all four variables at the same time. Hence, we divide all four variables into two groups and update the two groups of variables alternately. According to the separable structure, we employ the semiproximal alternating direction method of multipliers (sPADMM) to solve (16) and view X as one block and $V = (Y,Z,W)$ as another.

Next, we discuss the iteration schemes of sPADMM for solving the optimization problem (16). Given the kth iteration $(X^k, V^k, \lambda^k, \mu^k, \xi^k)$ and two self-adjoint positive semidefinite operators $\mathcal{P}_1$ and $\mathcal{P}_2$, first update variables X and V as follows:

$$\begin{cases} X^{k+1} = \arg\min_X \left\{ L_\sigma\left(X, V^k; \lambda^k, \mu^k, \xi^k\right) + \frac{\sigma}{2}\left\|X - X^k\right\|_{\mathcal{P}_1}^2 \right\}, \\ V^{k+1} = \arg\min_V \left\{ L_\sigma\left(X^{k+1}, V; \lambda^k, \mu^k, \xi^k\right) + \frac{\sigma}{2}\left\|V - V^k\right\|_{\mathcal{P}_2}^2 \right\}. \end{cases} \tag{20}$$

Then, update Lagrangian multipliers:

$$\begin{cases} \lambda^{k+1} = \lambda^k + \tau\sigma\left(JX^{k+1} - Y^{k+1}\right), \\ \mu^{k+1} = \mu^k + \tau\sigma\left(X^{k+1} - Z^{k+1}\right), \\ \xi^{k+1} = \xi^k + \tau\sigma\left(X^{k+1} - W^{k+1}\right). \end{cases} \tag{21}$$

Here, τ is step length and $\|X - X^k\|_{\mathcal{P}_1}^2 := \sum_{j=1}^p (X_{\cdot j} - X_{\cdot j}^k)^T \mathcal{P}_1 (X_{\cdot j} - X_{\cdot j}^k)$. It is well known that the sPADMM reduces to the classical ADMM introduced by [31, 32] when $\mathcal{P}_1 = 0$ and $\mathcal{P}_2 = 0$. A comprehensive review of sPADMM can be found in Appendix B of [33].

Remark 2. The choice of the two self-adjoint positive semidefinite linear operators $\mathcal{P}_1$ and $\mathcal{P}_2$ depends greatly on the problem. The general principle is that both $\mathcal{P}_1$ and $\mathcal{P}_2$ should be as small as possible, while the corresponding subproblems are still relatively easy to solve.

Each block of (20) involves solving a convex optimization problem. Next, we show the closed-form solutions of each block. First, we deal with the first block in (20):

$$\begin{aligned} X^{k+1} &= \arg\min_X \left\{ L_\sigma\left(X, V^k; \lambda^k, \mu^k, \xi^k\right) + \frac{\sigma}{2}\left\|X - X^k\right\|_{\mathscr{P}_1}^2 \right\} \\ &= \arg\min_X \ \frac{1}{2}\|X - A\|_F^2 + \frac{\sigma}{2}\left\|JX - Y^k + \frac{\lambda^k}{\sigma}\right\|_F^2 \\ &\quad + \frac{\sigma}{2}\left\|X - Z^k + \frac{\mu^k}{\sigma}\right\|_F^2 \\ &\quad + \frac{\sigma}{2}\left\|X - W^k + \frac{\xi^k}{\sigma}\right\|_F^2 + \frac{\sigma}{2}\left\|X - X^k\right\|_{\mathscr{P}_1}^2. \end{aligned} \tag{22}$$

The X-subproblem is a smooth, strongly convex optimization that has a unique global minimizer X^{k+1}. Finding the minimizer X^{k+1} is equivalent to finding the solution of the following linear system:

$$MX = A + \sigma\left(J^T Y^k + Z^k + W^k + \mathscr{P}_1 X^k\right) - J^T\lambda^k - \mu^k - \xi^k, \tag{23}$$

where $M = \sigma(J^T J + \mathscr{P}_1) + (1 + 2\sigma)I_n$. To solve this linear system efficiently, we design an appropriate $\mathscr{P}_1$. In our experiment, let $\mathscr{P}_1 = nI_n - 1_n 1_n^T - J^T J$, which is a positive semidefinite matrix. Then, $M = (1 + 2\sigma + n\sigma)I_n - \sigma 1_n 1_n^T$, and

$$M^{-1} = \frac{1}{1 + 2\sigma + n\sigma}\left(I_n + \frac{\sigma}{1 + 2\sigma} 1_n 1_n^T\right), \tag{24}$$

by the Sherman–Morrison–Woodbury formula [34]. Hence, the closed-form solution for this linear system can be obtained as

$$\begin{aligned} X^{k+1} = M^{-1}\{ & A + \sigma\left(J^T Y^k + Z^k + W^k + \mathscr{P}_1 X^k\right) \\ & - J^T\lambda^k - \mu^k - \xi^k\}. \end{aligned} \tag{25}$$

For the second block in (20), we let the linear operator $\mathscr{P}_2 = 0$ since the linear operator of V in (16) is an identity operator that ensures the well-defined V-subproblem:

$$\begin{aligned} V^{k+1} &= \arg\min_V \left\{ L_\sigma\left(X^{k+1}, V; \lambda^k, \mu^k, \xi^k\right)\right\} \\ &= \arg\min_{Y,Z,W} \ Q_1(Y) + Q_2(Z) + Q_3(W) \\ &\quad + \frac{\sigma}{2}\left\|JX^{k+1} + \frac{\lambda^k}{\sigma} - Y\right\|_F^2 + \frac{\sigma}{2}\left\|X^{k+1} + \frac{\mu^k}{\sigma} - Z\right\|_F^2 \\ &\quad + \frac{\sigma}{2}\left\|X^{k+1} + \frac{\xi^k}{\sigma} - W\right\|_F^2, \end{aligned} \tag{26}$$

which can be divided into three parts:

$$\begin{cases} Y^{k+1} = \arg\min_Y \ Q_1(Y) + \dfrac{\sigma}{2}\left\|JX^{k+1} + \dfrac{\lambda^k}{\sigma} - Y\right\|_F^2, \\ Z^{k+1} = \arg\min_Z \ Q_2(Z) + \dfrac{\sigma}{2}\left\|X^{k+1} + \dfrac{\mu^k}{\sigma} - Z\right\|_F^2, \\ W^{k+1} = \arg\min_W \ Q_3(W) + \dfrac{\sigma}{2}\left\|X^{k+1} + \dfrac{\xi^k}{\sigma} - W\right\|_F^2. \end{cases} \tag{27}$$

Together with the definition of proximal mapping and Moreau decomposition, we show the closed-form solutions of the above subproblems. For any $\ell \in \mathscr{E}$, the Y-subproblem has a closed-form solution that is given by

$$Y_{\ell\cdot}^{k+1} = \max\left\{1 - \frac{\gamma_1\omega_\ell}{\sigma\left\|J_{\ell\cdot}X^{k+1} + \left(\lambda_{\ell\cdot}^k/\sigma\right)\right\|_2}, 0\right\}\left(J_{\ell\cdot}X^{k+1} + \frac{\lambda_{\ell\cdot}^k}{\sigma}\right). \tag{28}$$

Similarly, for any $j = 1, 2, \ldots, p$, the solution of the Z-subproblem is given by

$$Z_{\cdot j}^{k+1} = \max\left\{1 - \frac{\gamma_2 u_j(1-\alpha)}{\sigma\left\|X_{\cdot j}^{k+1} + \left(\mu_{\cdot j}^k/\sigma\right)\right\|_2}, 0\right\}\left(X_{\cdot j}^{k+1} + \frac{\mu_{\cdot j}^k}{\sigma}\right). \tag{29}$$

The last subproblem is a soft threshold operator for each column, that is, for any $j = 1, 2, \ldots, p$,

$$W_{\cdot j}^{k+1} = \operatorname{sign}\left\{X_{\cdot j}^{k+1} + \frac{\xi_{\cdot j}^k}{\sigma}\right\} \circ \max\left(\left|X_{\cdot j}^{k+1} + \frac{\xi_{\cdot j}^k}{\sigma}\right| - \frac{\gamma_2\alpha}{\sigma}, 0\right). \tag{30}$$

Based on the above analysis, the algorithm framework of sPADMM for solving (16) is summarized in Algorithm 1.

The following theorem provides the convergence analysis for Algorithm 1. The proof is inspired by Theorem B.1 in [33], but no additional conditions are required in our proof.

Theorem 5. *If sequence $\left\{(X^k, V^k, \lambda^k, \mu^k, \xi^k)\right\}$ is generated from Algorithm 1, then sequence $\{(X^k, V^k)\}$ converges to an optimal solution for (16) and any accumulation point of $\left\{(X^k, V^k, \lambda^k, \mu^k, \xi^k)\right\}$ such that KKT conditions (19) hold.*

Proof. We first justify that the conditions required in Theorem B.1 in [33] automatically hold for (16).

(a) The existence of a solution for (16): note that the objective function $F(X, Y, Z, W)$ tends to infinity as $\|(X, Y, Z, W)\|_F \longrightarrow \infty$. Therefore, the level set

$$\mathscr{L}_0 := \{(X, Y, Z, W) \mid F(X, Y, Z, W) < F(0)\}. \tag{31}$$

is bounded by Theorem 4.11 in [35]. Hence, we further know that the minimizer of (16) can be attained by Theorem 4.10 in [35].

(b) The constraints of (16) can be rewritten as $\mathscr{A}^*(X) - \mathscr{B}^*(V) = 0$, where

$$\mathscr{A}^*(X) = \begin{pmatrix} J \\ I \\ I \end{pmatrix} X \text{ and } \mathscr{B}^*(V) = V. \tag{32}$$

Hence, we find that $\mathscr{A}\mathscr{A}^* = J^T J + 2I > 0$, so $\sigma(\mathscr{P}_1 + \mathscr{A}\mathscr{A}^*)$ is positive definite. Moreover, $\sigma(\mathscr{P}_2 + \mathscr{B}\mathscr{B}^*)$ is positive definite because $\mathscr{B}^*$ is the identity operator.

Based on these ingredients, the remainder of the proofs can be obtained easily (see Appendix B of [33] for details).

5. Numerical Results

The aim of this section is to illustrate the practical performance of sparse group lasso convex clustering (SGLCC), which performs well in theory. We conduct experiments both on synthetic and real datasets and compare the proposed method to K-means clustering, CC [10] and SCC [22]. The clustering results of K-means clustering can be obtained by built-in functions (kmeans) in MATLAB. All numerical experiments were performed in MATLAB R2017b on a personal computer with a 3.30 GHz processor and 4 GB of RAM.

In our experiments, we stop sPADMM based on the relative KKT residual of (16), which is given by

$$\varepsilon = \max\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5\} \le \text{Tol}, \tag{33}$$

where

$$\begin{aligned}
\varepsilon_1 &= \frac{\|JX - Y\|_F + \|X - Z\|_F + \|X - W\|_F}{1 + \|Y\|_F + \|Z\|_F + \|W\|_F}, \\
\varepsilon_2 &= \frac{\left\|X - A + J^T\lambda + \mu + \xi\right\|_F}{1 + \|A\|_F}, \\
\varepsilon_3 &= \frac{\left\|Y - \text{Prox}_{Q_1}(Y + \lambda)\right\|_F}{1 + \|Y\|_F}, \\
\varepsilon_4 &= \frac{\left\|Z - \text{Prox}_{Q_2}(Z + \mu)\right\|_F}{1 + \|Z\|_F}, \\
\varepsilon_5 &= \frac{\left\|W - \text{Prox}_{Q_3}(W + \xi)\right\|_F}{1 + \|W\|_F},
\end{aligned} \tag{34}$$

and Tol is a given tolerance. If the requirement $\varepsilon \le \text{Tol}$ is not satisfied after a maximum number of iterations (maxiter), we also terminate the algorithm. In our experiments, we set $\text{Tol} = 10^{-3}$ and maxiter = 10000. Moreover, we use the same method to select the weights ω_ℓ and the adaptive weights u_j in all experiments. For the weights ω_ℓ, we select the most popular and practical method proposed by [10]:

$$\omega_\ell = \begin{cases} \exp\left(-\phi\left\|A_{\ell_1\cdot} - A_{\ell_2\cdot}\right\|_2^2\right), & \ell_2 \text{ is among } \ell_1' sm - \text{nearest neighbors}, \\ 0, & \text{otherwise}, \end{cases} \tag{35}$$

where $\phi > 0$ and $m < n$. We set $\phi = 0.5$ and $m = 5$ in all experiments. Inspired by [36], we let the adaptive weights $u_j = (1/\|\widehat{X}_{\cdot j}^0\|_2)$, where $\widehat{X}_{\cdot j}^0$ is the estimator of $X_{\cdot j}$ in (1). As can be seen from models (1) and (2), the tuning parameters γ_1 and γ_2, α control the number of clusters and the number of informative features, respectively. In our experiments, the principle for selecting γ_1 and γ_2, α is to make the results match the true number of clusters and the true number of informative features as much as possible, and the Rand index as large as possible. This principle is also used in the literature [21] to determine the tuning parameters.

5.1. Synthetic Data. In this section, we evaluate the performance of SGLCC on synthetic datasets by comparing the results with those obtained with K-means clustering, CC and SCC. We consider three synthetic datasets. Each synthetic dataset consists of $n = 200$ observations with either $K = 2$ or 4 clusters. The number of features is p and ranges from 2000 to 5000, where the percentage of informative features is 2%. The cluster label $L_i (i = 1, 2, \ldots, n)$ is uniformly sampled from $\{1, 2, \ldots, K\}$. Without loss of generality, we assume that the first $s(2\%p)$ features are informative and the remaining features are noninformative. The first s informative features are generated from an s-dimensional normal distribution with mean $\overline{x}_K(L_i)$ and covariance Σ_s, whose (i, j) entry is $\Sigma_s(i, j) = \rho^{|i-j|}$, $1 \le i, j \le s$, and the noninformative features are generated from a $(p - s)$-dimensional standard normal distribution. Here,

$$\overline{x}_K(L_i) = \begin{cases} 1_s \mathbf{I}(L_i = 1) - 1_s \mathbf{I}(L_i = 2), & K = 2, \\ \left(1_{(s/2)}, 1_{(s/2)}\right)\mathbf{I}(L_i = 1) + \left(1_{(s/2)}, -1_{(s/2)}\right)\mathbf{I}(L_i = 2) & \\ + \left(-1_{(s/2)}, 1_{(s/2)}\right)\mathbf{I}(L_i = 3) + \left(-1_{(s/2)}, -1_{(s/2)}\right)\mathbf{I}(L_i = 4), & K = 4, \end{cases} \tag{36}$$

(1) **Initialization** Let $\tau = 1.618$ and $\sigma > 0$. Given a start point $(X^0, V^0, \lambda^0, \mu^0, \xi^0)$, for $k = 0, 1, 2, \cdots$,
(2) **repeat**
(3) Step 1. Update X^{k+1} according to (25)
(4) Step 2. Update V^{k+1} according to (28), (29) and (30)
(5) Step 3. Update the Lagrangian multipliers by (21)
(6) **until** Stopping criterion is satisfied.

Algorithm 1: sPADMM for solving (16).

Table 1: Results for Case I. Empirical mean and standard deviation (SD) of the RI, FMI, FNR, and FPR based on 50 repetitions.

Case I: $K = 2, \Sigma_s(i, j) = 0.5^{\lvert i-j \rvert}$									
p	Methods	RI		FMI		FNR		FPR	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
2000	K-means	0.970	0.120	0.980	0.079	0.000	0.000	1.000	0.000
	CC	0.816	0.077	0.789	0.096	0.000	0.000	1.000	0.000
	SCC	1.000	0.000	1.000	0.000	0.000	0.000	0.223	0.010
	SGLCC	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
3000	K-means	0.980	0.099	0.986	0.069	0.000	0.000	1.000	0.000
	CC	0.804	0.076	0.773	0.096	0.000	0.000	1.000	0.000
	SCC	1.000	0.000	1.000	0.000	0.000	0.000	0.221	0.008
	SGLCC	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
4000	K-means	1.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000
	CC	0.784	0.083	0.747	0.104	0.000	0.000	1.000	0.000
	SCC	0.999	0.004	0.999	0.005	0.000	0.000	0.223	0.007
	SGLCC	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
5000	K-means	0.990	0.071	0.994	0.042	0.000	0.000	1.000	0.000
	CC	0.762	0.071	0.717	0.093	0.000	0.000	1.000	0.000
	SCC	0.998	0.006	0.998	0.006	0.000	0.000	0.224	0.008
	SGLCC	1.000	0.000	1.000	0.000	0.000	0.001	0.000	0.000

Table 2: Results for Case II. Empirical mean and standard deviation (SD) of the RI, FMI, FNR, and FPR based on 50 repetitions.

Case II: $K = 4, \Sigma_s(i, j) = 0.5^{\lvert i-j \rvert}$									
p	Methods	RI		FMI		FNR		FPR	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
2000	K-means	0.907	0.073	0.848	0.117	0.000	0.000	1.000	0.000
	CC	0.970	0.045	0.942	0.085	0.000	0.000	1.000	0.000
	SCC	0.998	0.002	0.995	0.004	0.000	0.000	0.550	0.127
	SGLCC	0.995	0.021	0.991	0.034	0.015	0.045	0.000	0.000
3000	K-means	0.887	0.087	0.815	0.138	0.000	0.000	1.000	0.000
	CC	0.928	0.029	0.841	0.068	0.000	0.000	1.000	0.000
	SCC	0.997	0.015	0.993	0.032	0.000	0.000	0.681	0.071
	SGLCC	1.000	0.000	1.000	0.000	0.000	0.003	0.000	0.000
4000	K-means	0.892	0.076	0.823	0.122	0.000	0.000	1.000	0.000
	CC	0.925	0.030	0.834	0.069	0.000	0.000	1.000	0.000
	SCC	0.998	0.006	0.996	0.012	0.000	0.000	0.685	0.007
	SGLCC	1.000	0.000	1.000	0.000	0.000	0.001	0.000	0.000
5000	K-means	0.885	0.077	0.810	0.123	0.000	0.000	1.000	0.000
	CC	0.911	0.028	0.800	0.068	0.000	0.000	1.000	0.000
	SCC	0.986	0.033	0.968	0.077	0.000	0.000	0.711	0.085
	SGLCC	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000

where $\mathbf{I}(\cdot)$ is a function whose value is one if the event happens and zero otherwise. Let us illustrate this definition with an example. Specifically, if $K = 2$, the first s informative features of the observations with cluster labels 1 and 2 are generated from s-dimensional normal distributions with mean $\overline{x}_2(1) = 1_s$ and $\overline{x}_2(2) = -1_s$, respectively. In this part,

Table 3: Results for Case III. Empirical mean and standard deviation (SD) of the RI, FMI, FNR, and FPR based on 50 repetitions.

Case III: $K = 4, \Sigma_s(i,j) = 0.8^{\lvert i-j \rvert}$									
p	Methods	RI		FMI		FNR		FPR	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
2000	K-means	0.903	0.081	0.837	0.134	0.000	0.000	1.000	0.000
	CC	0.966	0.063	0.950	0.085	0.000	0.000	1.000	0.000
	SCC	0.988	0.027	0.977	0.045	0.015	0.075	0.506	0.115
	SGLCC	0.993	0.006	0.987	0.012	0.029	0.115	0.128	0.070
3000	K-means	0.914	0.085	0.860	0.136	0.000	0.000	1.000	0.000
	CC	0.947	0.071	0.924	0.096	0.000	0.000	1.000	0.000
	SCC	0.989	0.015	0.978	0.031	0.007	0.035	0.775	0.125
	SGLCC	0.993	0.019	0.986	0.030	0.029	0.114	0.089	0.041
4000	K-means	0.897	0.080	0.834	0.122	0.000	0.000	1.000	0.000
	CC	0.915	0.097	0.793	0.253	0.000	0.000	1.000	0.000
	SCC	0.933	0.080	0.825	0.228	0.025	0.101	0.823	0.231
	SGLCC	0.983	0.060	0.979	0.070	0.030	0.119	0.077	0.028
5000	K-means	0.892	0.085	0.825	0.136	0.000	0.000	1.000	0.000
	CC	0.900	0.048	0.764	0.115	0.000	0.000	1.000	0.000
	SCC	0.938	0.058	0.856	0.137	0.000	0.000	0.780	0.214
	SGLCC	0.999	0.003	0.999	0.006	0.010	0.068	0.003	0.001

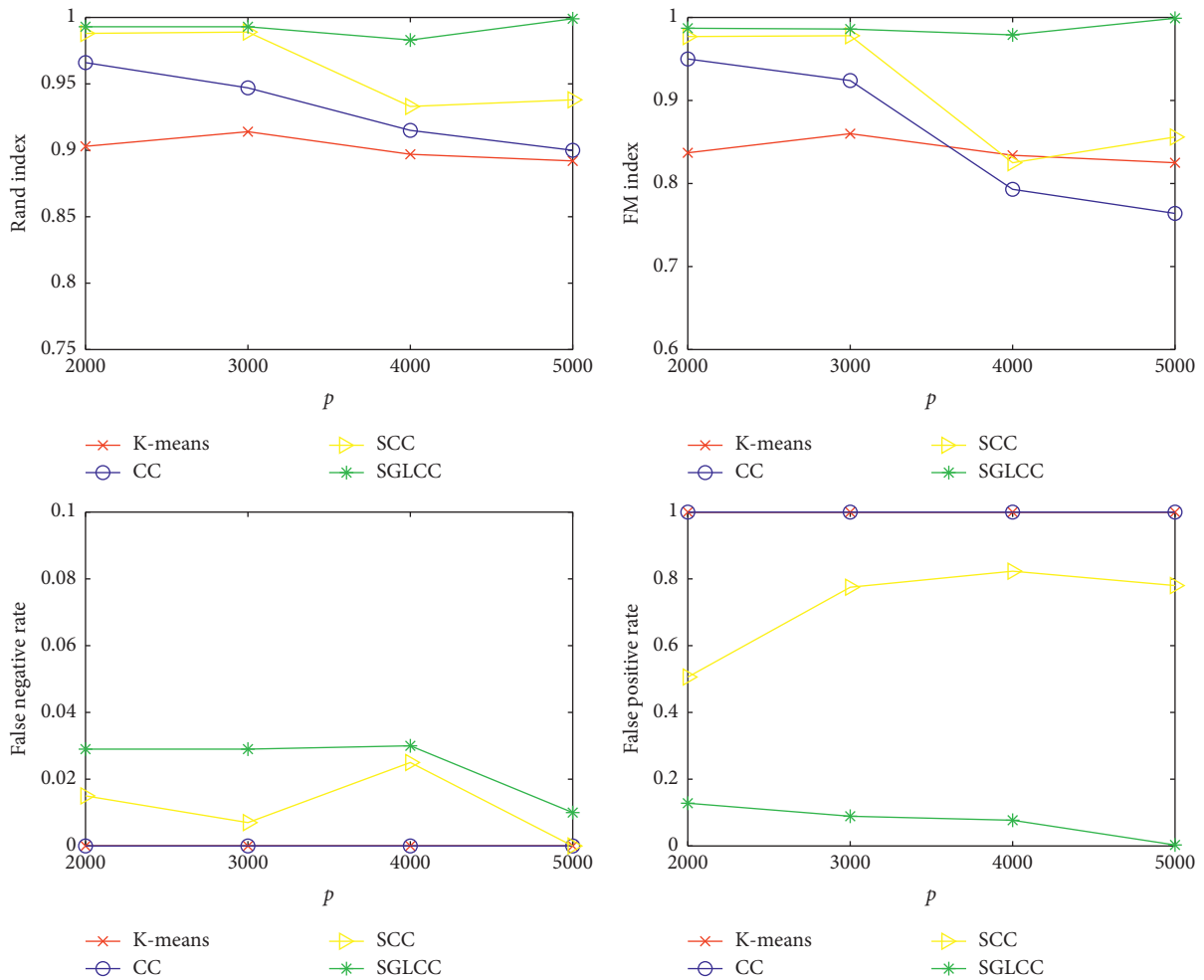

Figure 1: Results for Case III: $K = 4, \Sigma_s(i,j) = 0.8^{\lvert i-j \rvert}$.

Table 4: Details of the real datasets.

Datasets	Features	Observations	Clusters	Points of per cluster
Brain a [40]	5597	42	5	10, 10, 10, 4, 8
SRBCT [41]	2308	63	4	23, 8, 12, 20
Prostate [42]	6033	102	2	50, 52
Colon [43]	2000	62	2	22, 40
ARB	8266	590	8	46, 81, 162, 189, 31, 12, 50, 19

we consider the following three synthetic datasets. Case I: $K = 2, \Sigma_s(i, j) = 0.5^{|i-j|}$; Case II: $K = 4, \Sigma_s(i, j) = 0.5^{|i-j|}$; Case III: $K = 4, \Sigma_s(i, j) = 0.8^{|i-j|}$.

We repeat each experiment 50 times and evaluate the performance through four common criteria for cluster analysis. The Rand index [37] and Fowlkes–Mallows index [38] are two measures of the similarity between the estimated clustering result and the underlying true clustering assignment. Their values range from 0 to 1, and a larger value indicates better performance. For simplicity, we shall abbreviate the Rand index and Fowlkes–Mallows index as RI and FMI, respectively. Moreover, the false negative rate (FNR) and false positive rate (FPR) are reported to evaluate the performance of feature screening. All experimental results for the synthetic datasets are summarized in Tables 1–3.

From Table 1, several interesting observations can be made. (a) When the feature dimensions are high ($p = 2000, 3000, 4000, 5000$), CC does not perform well, and its clustering results worsen as p increases. (b) Because CC and K-means do not have the ability to eliminate noninformative features, their FNR and FPR are zero and one, respectively. (c) The clustering results of SGLCC and SCC are significantly better than those of CC when the feature dimensions are high because the penalty term is incorporated in convex clustering (1). (d) SGLCC has more stable performance than K-means and CC because it has the smallest standard deviation (SD). (e) SGLCC selects informative features with greater accuracy than the SCC, that is, with a lower FPR and almost the same FNR. The phenomenon mentioned above can also be observed from the results for Case II, see Table 2 for details. It is worth noting that SCC and SGLCC have similar performance for Case II. This reflects the ability of SCC to conduct feature selection, especially for data with group sparsity and dimensions that are not very large. However, compared with that of SGLCC, the performance of SCC becomes unstable as the number of dimensions increases. The reason behind this difference is that SGLCC allows a more nuanced selection of informative features.

The experimental results for Case III are reported in Figure 1 and Table 3. As we can see from Figure 1, SGLCC is significantly better than SCC when there are highly correlated variables in the dataset, and the improvement becomes increasingly evident as p increases. The main reason behind this may be the form of the sparse group lasso penalty, which, much like the elastic net, is good at handling the collinearity problem.

5.2. Real Data. In this section, five public datasets are used to validate the performance of our proposed SGLCC method. These datasets have been broadly applied to numerous studies in the literature, which is the primary reason why we selected them. All real datasets are summarized in Table 4. **Brain A** contains 42 observations with 5597 genes of five clusters that contain 10, 10, 10, 4, and 8 observations. **SRBCT** contains 63 observations of four clusters, and each observation has 2308 genes. The four clusters contain 23, 8, 12, and 20 observations. **Prostate** contains 102 observations with 5597 genes of two clusters that contain 50 and 52. **Colon** contains 62 observations of two clusters that contain 22 and 40 observations, each of which has 2000 genes. **ARB** contains 590 observations of eight clusters, and each observation has 8266 variables. The eight clusters contain 46, 81, 162, 189, 31, 12, 50, and 19 observations. We used the preprocessing versions of the first four datasets based on [39]. In our experiments, each feature from all of the datasets is centered. The Rand index of K-means, hierarchical clustering (Hclust),

Table 5: Results for the real datasets.

Datasets	Methods	Rand index
Brain A	K-means	0.798
	Hclust	0.317
	Sclust	0.506
	CC	0.577
	SCC	0.793
	SGLCC	**0.890**
Prostate	K-means	0.505
	Hclust	0.496
	Sclust	0.497
	CC	0.511
	SCC	0.518
	SGLCC	**0.530**
ARB	K-means	0.539
	Hclust	0.233
	Sclust	0.635
	CC	0.485
	SCC	0.562
	SGLCC	**0.675**
SRBCT	K-means	0.593
	Hclust	0.308
	Sclust	0.525
	CC	0.595
	SCC	0.634
	SGLCC	**0.770**
Colon	K-means	0.556
	Hclust	0.527
	Sclust	0.511
	CC	0.492
	SCC	0.596
	SGLCC	**0.640**

spectral clustering (Sclust), CC, SCC, and SGLCC for all five real datasets are reported in Table 5, and the best results are highlighted in bold. Here, the clustering results of Hclust and Sclust are obtained through two built-in functions in MATLAB, namely, clusterdata and spectralcluster.

As we can see from Table 5, the performance of CC is unsatisfactory and worse than that of K-means for **Brain A**, **Colon,** and **ARB**. A similar phenomenon can also be observed in [15, 22]. SGLCC is significantly superior to K-means, Hclust, Sclust, and CC in term of clustering performance, which implies that feature screening is an indispensable part of high-dimensional data analysis. Moreover, SGLCC is better than SCC because a more reasonable penalty term is used. In summary, our proposed SGLCC method has the largest Rand index among the six methods.

6. Conclusion

In this paper, we propose the Sparse Group Lasso Convex Clustering method for high-dimensional datasets, which can determine the number of clusters by learning from the data. SGLCC not only divides the observation set but also eliminates noninformative features. For the proposed clustering method, we provide the theoretical finite sample bounds of the prediction error and feature screening consistency. Moreover, an efficient sPADMM is designed to implement our proposed estimator, and each subproblem yields closed-form solutions. Convergence analysis of the sPADMM is established without any additional conditions. The experimental results illustrate that SGLCC provides superior clustering performance and feature selection abilities than other compared clustering methods. In particular, our method is also very effective for real data applications.

Data Availability

The real datasets used in Section 5.2 are available from http://www.stat.cmu.edu/~jiashun/Research/software/GenomicsData/ and http://archive.ics.uci.edu/ml/datasets.php.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

All authors read and approved the final manuscript. H. Chen mainly contributed to the statistical properties, algorithm design, and numerical results; L. Kong mainly contributed to the idea of the model and algorithm design; and Y. Li mainly contributed to the numerical results.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (11431002 and 11671029).

Supplementary Materials

All technical details for proofs of Theorems 3.1–3.4 are included in in this section, which can be found in https://figshare.com/articles/online_resource/Supplementary_Materials_pdf/12579164. (*Supplementary Materials*)

References

[1] J. Miao, X. Zhou, and T.-Z. Huang, "Local segmentation of images using an improved fuzzy c-means clustering algorithm based on self-adaptive dictionary learning," *Applied Soft Computing*, vol. 91, Article ID 106200, 2020.

[2] P.-H. Li, S. Pye, and I. Keppo, "Using clustering algorithms to characterise uncertain long-term decarbonisation pathways," *Applied Energy*, vol. 268, Article ID 114947, 2020.

[3] R. Xu, Y. Che, X. Wang, J. Hu, and Y. Xie, "Stacked autoencoder-based community detection method via an ensemble clustering framework," *Information Sciences*, vol. 526, pp. 151–165, 2020.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, Berkeley, CA, USA, June 1967.

[5] J. J. Wu, *Advances in K-Means Clustering: A Data Mining Thinking*, Springer, Berlin, Germany, 2012.

[6] L. Hubert, "Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures," *Journal of the American Statistical Association*, vol. 69, no. 347, pp. 698–704, 1974.

[7] Y. H. Liu, D. Liu, F. Yu, and Z. M. Ma, "A novel local density hierarchical clustering algorithm based on reverse nearest neighbors," *Mathematical Problems in Engineering*, vol. 2019, Article ID 2959017, 10 pages, 2019.

[8] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor, "Convex clustering shrinkage," in *Proceedings of the PASCAL Workshop on Statistics and Optimization of Clustering Workshop*, vol. 1–6, London, UK, July 2005.

[9] T. D. Hocking, A. Joulin, F. Bach, and J. P. Vert, "Clusterpath: an algorithm for clustering using convex fusion penalties," in *Proceedings of the 28th International Conference on Machine Learning*, pp. 745–752, Bellevue, WA, USA, June 2011.

[10] E. C. Chi and K. Lange, "Splitting methods for convex clustering," *Journal of Computational and Graphical Statistics*, vol. 24, no. 4, pp. 994–1013, 2015.

[11] F. Lindsten, H. Ohlsson, and L. Ljung, "Clustering using sum-of-norms regularization with application to particle filter output computation," in *Proceedings of the IEEE Statistical Signal Processing Workshop*, pp. 201–204, Nice, France, June 2011.

[12] C. B. Zhu, H. Xu, C. L. Leng, and S. C. Yan, "Convex optimization procedure for clustering: theoretical revisit," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 27, pp. 1619–1627, Montreal, Canada, December 2014.

[13] A. Panahi, D. Dubhashi, F. D. Johansson, and C. Bhattacharyya, "Clustering by sum of norms: stochastic incremental algorithm, convergence and cluster recovery," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 2769–2777, Sydney, Australia, August 2017.

[14] D. F. Sun, K. C. Toh, and Y. C. Yuan, "Convex clustering: model, theoretical guarantee and efficient algorithm," 2018, https://arxiv.org/abs/1810.02677.

[15] K. M. Tan and D. Witten, "Statistical properties of convex clustering," *Electronic Journal of Statistics*, vol. 9, no. 2, pp. 2324–2347, 2015.

[16] P. Radchenko and G. Mukherjee, "Convex clustering via l1-fusion penalization," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 5, pp. 1527–1546, 2017.

[17] E. C. Chi and S. Steinerberger, "Recovering trees with convex clustering," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 3, pp. 383–407, 2019.

[18] Y. C. Yuan, D. F. Sun, and K. C. Toh, "An efficient semi-smooth Newton based algorithm for convex clustering," in *Proceedings of the 35th International Conference on Machine Learning*, pp. 5718–5726, Stockholm, Sweden, July 2018.

[19] S. Poddar and M. Jacob, "Convex clustering and recovery of partially observed data," in *Proceedings of IEEE International Conference on Image Processing*, pp. 3498–3502, Phoenix, AZ, USA, December 2016.

[20] H. Choi and S. Lee, "Convex clustering for binary data," *Advances in Data Analysis and Classification*, vol. 13, no. 4, pp. 991–1018, 2019.

[21] X. L. Sui, L. Xu, X. Qian, and T. Liu, "Convex clustering with metric learning," *Pattern Recognition*, vol. 81, pp. 575–584, 2018.

[22] B. Wang, Y. Zhang, W. W. Sun, and Y. Fang, "Sparse convex clustering," *Journal of Computational and Graphical Statistics*, vol. 27, no. 2, pp. 393–403, 2018.

[23] S. Noah, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse group lasso," *Journal of Computational & Graphical Statistics*, vol. 22, no. 2, pp. 213–245, 2013.

[24] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[25] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[26] H. Wang and C. Leng, "A note on adaptive group lasso," *Computational Statistics & Data Analysis*, vol. 52, no. 12, pp. 5277–5286, 2008.

[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[28] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, Taylor & Francis Group, Boca Raton, FL, USA, 2015.

[29] T. R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, USA, 1970.

[30] L. Wang, Y. Wu, and R. Li, "Quantile regression for analyzing heterogeneity in ultra-high dimension," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 214–222, 2012.

[31] R. Glowinski and A. Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires," *Revue Française D'automatique, Informatique, Recherche Opérationnelle. Analyse Numérique*, vol. 9, no. R2, pp. 41–76, 1975.

[32] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.

[33] M. Fazel, T. K. Pong, D. F. Sun, and P. Tseng, "Hankel matrix rank minimization with applications to system identification and realization," *SIAM Journal on Matrix Analysis & Applications*, vol. 34, no. 3, pp. 946–977, 2012.

[34] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, USA, 3rd edition, 1996.

[35] B. S. Mordukhovich and N. M. Nam, *An Easy Path to Convex Analysis and Applications*, Morgan & Claypool Publishers, San Rafael, CA, USA, 2014.

[36] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[37] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[38] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 553–569, 1983.

[39] M. Dettling, "Bagboosting for tumor classification with gene expression data," *Bioinformatics*, vol. 20, no. 18, pp. 3583–3593, 2004.

[40] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[41] J. Khan, J. S. Wei, M. Ringnér et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[42] D. Singh, P. G. Febbo, K. Ross et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[43] U. Alon, N. Barkai, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.