Unsupervised Feature Selection via Nonnegative Orthogonal Constrained Regularized Minimization

Yan Li   LI-YAN20@MAILS.TSINGHUA.EDU.CN
Department of Mathematical Sciences
Tsinghua University
Beijing 100084, China

Defeng Sun   DEFENG.SUN@POLYU.EDU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong

Liping Zhang∗   LIPINGZHANG@TSINGHUA.EDU.CN
Department of Mathematical Sciences
Tsinghua University
Beijing 100084, China

Editor:

Abstract

Unsupervised feature selection has drawn wide attention in the era of big data since it is a primary technique for dimensionality reduction. However, many existing unsupervised feature selection models and solution methods were presented for the purpose of application, and lack of theoretical support, e.g., without convergence analysis. In this paper, we first establish a novel unsupervised feature selection model based on regularized minimization with nonnegative orthogonal constraints, which has advantages of embedding feature selection into the nonnegative spectral clustering and preventing overfitting. An effective inexact augmented Lagrangian multiplier method is proposed to solve our model, which adopts the proximal alternating minimization method to solve subproblem at each iteration. We show that the sequence generated by our method globally converges to a Karush-Kuhn-Tucker point of our model. Extensive numerical experiments on popular datasets demonstrate the stability and robustness of our method. Moreover, comparison results of algorithm performance show that our method outperforms some existing state-of-the-art methods.

Keywords: unsupervised feature selection, orthogonal constraint, augmented Lagrangian multiplier method, alternating minimization method, Karush-Kuhn-Tucker point

1. Introduction

Due to large amounts of data produced by rapid development of technology, processing high-dimensional data is one of the most challenging problems in many fields, such as action recognition (Klaser et al., 2011), image classification (Gui et al., 2014), computational biology (Chen et al., 2020). Generally, not all the features are equally important for the data with high-dimensional features. There are some redundant, irrelevant and noisy

---

∗. Liping Zhang is the corresponding author (lipingzhang@tsinghua.edu.cn).

features, which not only increase computational cost and storage burden, but also reduce the performance of learning tasks. Dimensionality reduction methods can be roughly divided into two types: feature extraction (Lee and Seung, 1999; Charte et al., 2021; Lian et al., 2018) and feature selection (Kittler, 1986; Li et al., 2021; Roffo et al., 2020; Yu et al., 2019). They project the high-dimensional feature space to a low-dimensional space to squeeze features. The low-dimensional space generated by the former is usually composed of linear or nonlinear combinations of original features, but irrelevant, redundant and even noisy features are involved in the process of reducing dimension, which may affect the subsequent learning tasks to some extent. However, the latter evaluates each dimension feature of high dimensional data and directly select the optimal feature subset from the original high-dimensional feature set by using certain criteria to achieve compact and accurate data representation (Liu et al., 2004; Molina et al., 2002). Compared with the former, the latter has better interpretability. Feature selection maintains the semantic information of the original features and it aims to select valuable and discriminative feature subsets from the original high-dimensional feature set, while feature extraction changes the original meanings of the feature and the new features usually lose the physical meanings of the original features. Therefore, feature selection enjoys tremendous popularity in a wide range of applications from data mining to machine learning. Many feature selection methods (Nie et al., 2016; Hou et al., 2013; Nie et al., 2019) are proposed to better explore the properties of high-dimensional data.

According to whether the class label information is available or not, feature selection methods can be roughly grouped into two categories, i.e., supervised feature selection, and unsupervised feature selection (Dash et al., 1997; He et al., 2005). Benefiting from the sample-wise annotations, supervised feature selection algorithms, e.g., Fisher score (Duda and Hart, 2001), robust regression (Nie et al., 2010), minimum redundancy maximum relevance (Peng et al., 2005) and trace radio (Nie et al., 2008), are able to select discriminative features and achieve superior classification accuracy and reliability. With the fact that the labeled data is often inadequate or completely unobtainable in many practical applications, traditional supervised feature selection methods cannot deal with such problems. In addition, annotating the unlabeled data requires an excessive cost in human resources and is time-consuming. Therefore, for the high-dimensional data with missing labels, it is an effective means to solve above mentioned problems by using unsupervised approaches to reduce the feature dimension. Compared to supervised feature selection, unsupervised feature selection is a more challenging task since the label information of the training data is unavailable (He et al., 2005). Many studies have been conducted on unsupervised feature selection methods, such as spectral analysis (Zhao and Liu, 2007; Cai et al., 2010; Li et al., 2012), matrix factorization (Wang et al., 2015; Qian and Zhai, 2013), dictionary learning (Zhu et al., 2016) and so on.

Unsupervised feature selection methods (Nie et al., 2019; Chen et al., 2022) generally select features according to the intrinsic structural characteristics of data and have achieved pretty good performance, which can alleviate the undesirable influence of noise and redundant features in the original data. For example, MaxVar (Krzanowski, 1987) is a statistical method, which selects features corresponding to the maximum variance. Laplacian Score (He et al., 2005) is a similarity preserving method, which evaluates the importance of a feature by its power of locality preservation; SPEC (Zhao and Liu, 2007) selects features using spectral regression. RSR (Zhu et al., 2015) is a data reconstruction method, which uses

the $l_{2,1}$-norm to measure the fitting error and to promote sparsity; CPFS (Masaeli et al., 2010) relaxes the feature selection problem into a continuous convex optimization problem; REFS (Li et al., 2017) embeds the reconstruction function learning process to feature selection. MCFS (Cai et al., 2010) selects features based on spectral analysis and sparse regression problem, UDFS (Yang et al., 2011a) which selects features by preserving the structure based on discriminative information; UDPFS (Wang et al., 2020) introduces fuzziness into sub-space learning to learn a discriminative projection for feature selection; NDFS (Li et al., 2012) selects features by leveraging a joint framework of nonnegative spectral analysis and $l_{2,1}$-norm regularization. However, numerical algorithms proposed in these unsupervised feature selection methods are often without global convergence analysis and then lack of theoretical support (Shi et al., 2016). Furthermore, these methods may be greatly affected by disturbance and do not have good performance, and then they may not have good stability and strong robustness.

Motivated by this, we establish a novel unsupervised feature selection model based on regularized minimization with nonnegative orthogonal constraints, which has two advantages of embedding feature selection into the nonnegative spectral clustering and preventing overfitting. In our model, the $l_{2,1}$-regularized term will enable the subproblem from our proposed algorithm has closed-form optimal solution, and the Frobenius-norm regularization will explicitly control the overfitting, which is the main difference from NDFS (Li et al., 2012). And the nonnegative orthogonal constraints can embed feature selection into the nonnegative spectral clustering. However, it is hard to handle the orthogonal constraints in general (Absil et al., 2009). Some existing popular solution methods, such as the multiplicative update method (Ding et al., 2006; Yoo and Choi, 2008) and the greedy orthogonal pivoting algorithm (Zhang et al., 2019), require the objective function to be differentiable and have the special formulation. So, they are not applicable to our model. This prompts us to design an effective solution method for our model. Based on the algorithmic frameworks of the augmented Lagrangian method (Andreani et al., 2008) and the proximal alternating minimization (Attouch et al., 2013), we propose an effective inexact augmented Lagrangian multiplier (ALM) method to solve our model, which uses the proximal alternating minimization (PAM) method to solve subproblems at each iteration. We show that the sequence generated by our ALM method globally converges to a Karush-Kuhn-Tucker point of our model. Numerical experiments on popular datasets demonstrate the stability and robustness of our method. Moreover, comparison results of algorithm performance show that our method outperforms some existing state-of-the-art methods.

The main contribution of this paper is summarized as follows:

- We establish a novel $l_{2,1}$-regularized optimization model with nonnegative orthogonal constraints for unsupervised feature selection, which has two advantages of embedding feature selection into the nonnegative spectral clustering and preventing overfitting. Specifically, we use the spectral clustering technique to learn pseudo class labels, and then select features which are most discriminative to pseudo class labels.

- We propose an effective inexact ALM method to solve our model. At each iteration, we use the PAM method to solve subproblems, which has the advantage of making each subproblem have a closed form solution. This helps us to show that the sequence generated by our ALM method globally converges to a Karush-Kuhn-Tucker point of

our model without any further assumption. Numerical results on popular datasets are reported to show the efficiency, stability and robustness of our method.

The rest of this paper is organized as follows. In Section 2, some preliminaries for nonsmooth optimization are collected. In Section 3, we establish a novel model for unsupervised feature selection. In Section 4, an inexact ALM method is proposed to solve our model, and its convergence analysis is also given. Numerical experiments and concluding remarks are given in the last two sections.

2. Preliminaries

In this section, we recall some preliminaries on nonsmooth optimization and give some notations. Throughout this paper, matrices are written as capital letters (e.g., $A, B, \cdots$) and vectors are denoted as boldface lowercase letters (e.g., $\mathbf{x}, \mathbf{y}, \cdots$). For any positive integer n, denote $[n] = \{1, 2, \ldots, n\}$. Given a matrix $Y = (Y_{i,j}) \in \mathbb{R}^{n \times m}$, its maximum (elementwise) norm is denoted by

$$\|Y\|_\infty := \max\{|Y_{i,j}| : \ i \in [n], \ j \in [m]\}.$$

The Frobenius norm of Y is denoted by

$$\|Y\|_F := \sqrt{\sum_{i=1}^n \sum_{j=1}^m Y_{i,j}^2},$$

and its $l_{2,1}$-norm is defined as

$$\|Y\|_{2,1} := \sum_{i=1}^n \sqrt{\sum_{j=1}^m Y_{i,j}^2} = \sum_{i=1}^n \|Y_i\|_2,$$

where Y_i is the i-th row of Y and $\|\cdot\|_2$ is Euclidean norm. Let $\mathrm{Vec}(Y)$ be an $mn \times 1$ vector with $\mathrm{Vec}(Y) := [Y_1^\intercal, Y_2^\intercal, \cdots, Y_m^\intercal]^\intercal$. For any $\mathbf{v} \in \mathbb{R}^n$, let $[\mathbf{v}]_i$ denote its ith component, and let $\mathrm{diag}(\mathbf{v}) \in \mathbb{R}^{n \times n}$ denote the diagonal matrix with diagonal entries $\{[\mathbf{v}]_i\}_{i=1}^n$. Given a square matrix Y, $Y \succ 0$ denotes that Y is a positive definite matrix and the trace of Y, i.e., the sum of the diagonal elements of Y, is denoted by $\mathrm{Tr}(Y)$. $\mathbf{E}$ is a matrix whose elements are all 1. $\mathbf{O}$ is a matrix whose elements are all 0. $\mathbf{O} \leq X \leq \mathbf{E}$ denotes that each element of X satisfies $0 \leq X_{i,j} \leq 1$. $\mathbf{0} \leq \mathbf{v} \leq \mathbf{1}$ denotes that each element of $\mathbf{v}$ satisfies $0 \leq [\mathbf{v}]_i \leq 1$. Given a set Ω, $\Pi_\Omega Y$ denotes the projection of Y on Ω. For an index sequence $\mathcal{K} = \{k_0, k_1, k_2, \ldots\}$ that satisfies $k_{j+1} > k_j$ for any $j \geq 0$, we denote $\lim_{k \in \mathcal{K}} x_k := \lim_{j \to \infty} x_{k_j}$. For any set S, its indicator function is defined by

$$\delta_S(X) = \left\{ \begin{array}{ll} 0, & \text{if } X \in S, \\ +\infty, & \text{otherwise.} \end{array} \right. \tag{1}$$

Let us recall some definitions of sub-differential calculus (see, e.g., Rockafellar and Wets, 2009).

Definition 1 *Let $C \subseteq \mathbb{R}^n$ and $\overline{x} \in C$. A vector v is normal to C at $\overline{x}$ in the regular sense, or a regular normal, written $v \in \hat{N}_C(\overline{x})$, if*

$$\langle v, x - \overline{x} \rangle \leq o(\|x - \overline{x}\|) \ \text{for} \ x \in C.$$

A vector is normal to C at $\overline{x}$ in the general sense, written $v \in N_C(\overline{x})$, if there exists sequence $\{x_k\}_k \subset C, \{v_k\}_k$ such that $x_k \to \overline{x}$ and $v_k \to v$ with $v_k \in \hat{N}_C(x_k)$. The cone $N_C(\overline{x})$ is called the normal cone to C at $\overline{x}$.

Definition 2 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function.*

1) *The domain of f is defined and denoted by $\text{dom} \, f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$.*

2) *For each $x \in \text{dom} \, f$, the vector $x^* \in \mathbb{R}^n$ is said to be a regular subgradient of f at x, written $x^* \in \hat{\partial} f(x)$, if $f(y) \geq f(x) + \langle x^*, y - x \rangle + o(\|x - y\|)$.*

3) *The vector $x^* \in \mathbb{R}^n$ is said to be a (limiting) subgradient of f at $x \in \text{dom} \, f$, written $x^* \in \partial f(x)$, if there exists $\{x_n\}_n, \{x_n^*\}_n$ such that $x_n \to x, f(x_n) \to f(x)$ and $x_n^* \in \hat{\partial} f(x_n)$ with $x_n^* \to x^*$.*

4) *For each $x \in \text{dom} \, f$, x is called (limiting)-critical if $0 \in \partial f(x)$.*

Remark 3 (Closedness of ∂f) *Let $(x_k, x_k^*) \in \text{Graph} \, \partial f$ be a sequence that converges to (x, x^*). By the definition of $\partial f(x)$, if $f(x_k)$ converges to $f(x)$ then $(x, x^*) \in \text{Graph} \, \partial f$.*

Remark 4 *(Rockafellar and Wets, 2009, Example 6.7) Let S be a closed nonempty subset of $\mathbb{R}^n$, then*

$$\partial \delta_S(\overline{x}) = N_S(\overline{x}), \ \overline{x} \in S.$$

Furthermore, for a smooth mapping $G : \mathbb{R}^n \to \mathbb{R}^m$, i.e., $G(x) := (g_1(x), \cdots, g_m(x))^\mathsf{T}$, define $S = G^{-1}(0) \subset \mathbb{R}^n$. Set $\nabla G(x) := [\frac{\partial g_j}{\partial x_i}(x)]_{i,j=1}^{n,m} \in \mathbb{R}^{n \times m}$. If $\nabla G(\overline{x})$ has full rank m at a point $\overline{x} \in S$, with $G(\overline{x}) = 0$, then its normal cone to S can be explicitly written as

$$N_S(\overline{x}) = \{\nabla G(\overline{x}) y \mid y \in \mathbb{R}^m\}.$$

3. A New Unsupervised Feature Selection Model

Let $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ be the data matrix with each column $\mathbf{x}_i \in \mathbb{R}^{d \times 1}$ being the i-th data point. Let d and n be the number of features and the number of sample, respectively. Suppose these n samples are sampled from c classes. Denote $F = [\mathbf{f}_1, \cdots, \mathbf{f}_n]^\mathsf{T} \in \{0, 1\}^{n \times c}$, where $\mathbf{f}_i \in \{0, 1\}^{c \times 1}$ is the cluster indicator vector for $\mathbf{x}_i$. That is, the j-th element of $\mathbf{f}_i$ is 1, if $\mathbf{x}_i$ is assigned to the j-th cluster, otherwise 0. Following the notation in Yang et al. (2011b), the scaled cluster indicator matrix Y is defined as

$$Y = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]^\mathsf{T} = F(F^\mathsf{T} F)^{-\frac{1}{2}},$$

where y_i is the scaled cluster indicator of $\mathbf{x}_i$. It turns out that

$$Y^\mathsf{T} Y = (F^\mathsf{T} F)^{-\frac{1}{2}} F^\mathsf{T} F (F^\mathsf{T} F)^{-\frac{1}{2}} = I_c,$$

where $I_c \in \mathbb{R}^{c \times c}$ is an identity matrix.

At first, we use the clustering techniques to learn the scaled cluster indicators of data points, which can be regarded as pseudo class labels. Given a set of data points $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ and some notion of similarity $s_{i,j} \geq 0$ between all pairs of data points $\mathbf{x}_i$ and $\mathbf{x}_j$, the intuitive goal of clustering is to divide the data points into several groups such that points in the same group are similar and points in different groups are dissimilar to each other. Spectral clustering is widely used in that it can effectively generate the pseudo labels from the graphs. In our method, we construct a k-nearest neighbors graph and choose the Gaussian kernel as the weight (see Cai et al., 2005). Specially, we define the affinity graph S as follows:

$$
S_{i,j} = \left\{ \begin{array}{ll} \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}), & \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ 0, & \text{otherwise}, \end{array} \right.
$$

where $\mathcal{N}_k(\mathbf{x})$ is the set of k-nearest neighbors of $\mathbf{x}$. The corresponding degree matrix can be constructed to D with $D_{ii} = \sum_j S_{i,j}$, and Laplacian matrix L of the normalized graph (see Von Luxburg, 2007) is calculated with $L = D^{-\frac{1}{2}}(D - S)D^{-\frac{1}{2}}$. Therefore, the local geometrical structure of data points can be obtained by:

$$
\min_Y \text{Tr}(Y^{\intercal} L Y) \quad \text{s.t.} \quad Y = F(F^{\intercal}F)^{-\frac{1}{2}}. \tag{2}
$$

This is a discrete optimization problem as the entries of the feasible solution are only allowed to take two particular values, and of course it is a NP-hard problem. A well-known method is to discard the discreteness condition and relax the problem by allowing the entries of the matrix Y to take arbitrary real values. Then, the relaxed problem becomes:

$$
\min_{Y \in \mathbb{R}^{n \times c}} \text{Tr}(Y^{\intercal} L Y) \quad \text{s.t.} \quad Y^{\intercal}Y = I_c. \tag{3}
$$

The next stage is to construct a sparse transformation W on the data matrix X by employing the scaled cluster indicator matrix Y, joined with two regularization terms. We can formulate it as:

$$
\min_{W \in \mathbb{R}^{d \times c}} \|Y - X^{\intercal}W\|_{2,1} + \beta\|W\|_{2,1} + \gamma\|W\|_F^2, \tag{4}
$$

where W is a linear and low dimensional transformation matrix, and β and γ are the regularization parameters. In the objection function of the problem (4), the first term represents the linear transformation model to measure the association between the features and the pseudo class labels. The second term constructs the sparsity on the rows of the transformation matrix W, which is beneficial for selecting discriminative features. The third term is to avoid overfitting.

By integrating the spectral clustering (3) and sparse regression (4) in a joint objective function, the model we proposed can be obtained as follows:

$$
\begin{aligned}
&\min_{W,Y} \text{Tr}(Y^{\intercal} L Y) + \alpha\|Y - X^{\intercal}W\|_{2,1} + \beta\|W\|_{2,1} + \gamma\|W\|_F^2 \\
&\text{s.t.} \quad Y^{\intercal}Y = I_c, \ Y \geq \mathbf{O},
\end{aligned} \tag{5}
$$

where α is a tuning parameter.

4. Algorithm Description of Our Inexact ALM Method

In this section, we develop our inexact augmented Lagrangian method for solving problem (5), which is a nonconvex optimization with a nonsmooth objective function. By introducing auxiliary variables $U, V, \widehat{Y}, F$, the problem (5) can be transformed into the following equivalent:

$$\min_{W,U,V,Y,F,\widehat{Y}} \operatorname{Tr}(Y^\intercal LY) + \alpha\|U\|_{2,1} + \beta\|V\|_{2,1} + \gamma\|W\|_F^2 + \delta_{\mathbb{S}_1}(\widehat{Y}) + \delta_{\mathbb{S}_2}(F)$$

$$\text{s.t.} \quad \begin{cases} Y = F \\ U = Y - X^\intercal W \\ V = W \\ Y = \widehat{Y} \end{cases} \tag{6}$$

where $\mathbb{S}_1 = \{\,\widehat{Y} \mid \widehat{Y}^\intercal\widehat{Y} = I_c\,\}$, $\mathbb{S}_2 = \{\,F \mid \mathbf{O} \le F \le \mathbf{E}\,\}$.

Set $\lambda := (\lambda_1, \lambda_2, \lambda_3, \lambda_4) \in \mathbb{R}^{n\times c} \times \mathbb{R}^{d\times c} \times \mathbb{R}^{n\times c} \times \mathbb{R}^{n\times c}$. The augmented Lagrangian function for (6) is defined by

$$\begin{aligned} L(W,U,V,Y,F,\widehat{Y},\lambda;\rho) :=& \operatorname{Tr}(Y^\intercal LY) + \alpha\|U\|_{2,1} + \beta\|V\|_{2,1} + \gamma\|W\|_F^2 + \delta_{\mathbb{S}_1}(\widehat{Y}) \\ &+ \delta_{\mathbb{S}_2}(F) + \langle\lambda_1, Y - X^\intercal W - U\rangle + \langle\lambda_2, V - W\rangle \\ &+ \langle\lambda_3, Y - F\rangle + \langle\lambda_4, \widehat{Y} - Y\rangle + \frac{\rho}{2}(\|\widehat{Y} - Y\|_F^2 + \|V - W\|_F^2 \\ &+ \|Y - F\|_F^2 + \|Y - X^\intercal W - U\|_F^2), \end{aligned} \tag{7}$$

where ρ is a positive penalty parameter.

The ALM method can be used to alternately update the $(W, U, V, Y, F, \widehat{Y})$, the multiplier λ, and the penalty parameter ρ to satisfy the accuracy condition (10). We describe our inexact ALM method for solving (5) in details as follows.

Remark 5 *Set the parameters in Algorithm 1 as follows: $\tau \in [0, 1)$; $\rho^1 > 0$; $r > 1$; the sequence of positive tolerance parameters $\{\epsilon_k\}_{k\in\mathbb{N}}$ is chosen such that $\lim_{k\to+\infty} \epsilon_k = 0$. The parameters $\overline{\lambda}_1^1, \overline{\lambda}_2^1, \overline{\lambda}_3^1, \overline{\lambda}_4^1, \overline{\lambda}_{N,min}, \overline{\lambda}_{N,max}$ are finite-valued matrices satisfying*

$$-\infty < [\overline{\lambda}_{N,min}]_{i,j} < [\overline{\lambda}_{N,max}]_{i,j} < +\infty \ \forall i,j, \quad N = 1, 2, 3, 4.$$

In Algorithm 1, the most important is how to solve (8-10). That is, given the current iterate $(W^k, U^k, V^k, Y^k, F^k, \widehat{Y}^k)$, how to generate the next iterate $(W^{k+1}, U^{k+1}, V^{k+1}, Y^{k+1}, F^{k+1}, \widehat{Y}^{k+1})$ is very critical. We propose a PAM method to solve (8-10) and show that there exists a solution for (8-10) and such a solution can be efficiently computed as $\epsilon_k \downarrow 0$, i.e., Step 1 in Algorithm 1 is well defined. We will establish the PAM method and its convergence in the following two subsections.

4.1 PAM for Augmented Lagrangian Subproblem

In this subsection, we present more details on implementing Algorithm 1 and construct a PAM method to solve the augmented Lagrangian subproblem with arbitrarily given accuracy.

Algorithm 1 Inexact ALM Method for (5)

Input. Data matrix $X \in \mathbb{R}^{d \times n}$. Given predefined parameters $\{\epsilon_k\}_{k \in \mathbb{N}}$, ρ^1, τ, r, $\overline{\lambda}_{N,min}$, $\overline{\lambda}_{N,max}$ ($N = 1, 2, 3, 4$), and $\overline{\lambda}^1 := (\overline{\lambda}_1^1, \overline{\lambda}_2^1, \overline{\lambda}_3^1, \overline{\lambda}_4^1)$ that satisfy the condition in Remark 5, for $k = 1, 2, \dots$,

Output. Sort all the d features according to $\|W_i\|_2$ ($i \in [d]$) and select the top q ranked features.

Step 1: Compute the subproblem

$$(W^k, U^k, V^k, Y^k, F^k, \widehat{Y}^k) \approx \underset{W,U,V,Y,F,\widehat{Y}}{\arg\min} \; L(W, U, V, Y, F, \widehat{Y}, \overline{\lambda}^k; \rho^k) \tag{8}$$

such that

$$\mathbf{O} \leq F^k \leq E, (\widehat{Y}^k)^\intercal \widehat{Y}^k = I_c, \tag{9}$$

and there exists $\Theta^k \in \partial L(W^k, U^k, V^k, Y^k, F^k, \widehat{Y}^k, \overline{\lambda}^k; \rho^k)$ satisfying

$$\|\Theta^k\|_\infty \leq \epsilon_k. \tag{10}$$

Step 2: Update the multiplier as:

$$\lambda_1^{k+1} = \overline{\lambda}_1^k + \rho^k (Y^k - X^\intercal W^k - U^k)$$
$$\lambda_2^{k+1} = \overline{\lambda}_2^k + \rho^k (V^k - W^k)$$
$$\lambda_3^{k+1} = \overline{\lambda}_3^k + \rho^k (Y^k - F^k)$$
$$\lambda_4^{k+1} = \overline{\lambda}_4^k + \rho^k (\widehat{Y}^k - Y^k)$$

where $\overline{\lambda}_N^{k+1} = \Pi_\Omega \lambda_N^{k+1}$ and $\Omega = \{\lambda_N : \overline{\lambda}_{N,min} \leq \lambda_N \leq \overline{\lambda}_{N,max}\}$, $N = 1, 2, 3, 4$.

Step 3: Update the penalty parameter:

$$\rho^{k+1} = \begin{cases} \rho^k & \text{if } \|R_i^k\|_\infty \leq \tau \|R_i^{k-1}\|_\infty \; (i = 1, 2, 3, 4) \\ r\rho^k & \text{otherwise,} \end{cases} \tag{11}$$

where $R_1^k = Y^k - X^\intercal W^k - U^k$, $R_2^k = V^k - W^k$, $R_3^k = Y^k - F^k$, $R_4^k = \widehat{Y}^k - Y^k$.

It can be seen that the constraint (10) is an ϵ^k-perturbation of the critical point property

$$0 \in \partial L(W^k, U^k, V^k, Y^k, F^k, \widehat{Y}^k, \overline{\lambda}^k; \rho^k). \tag{12}$$

In fact, the algorithm we proposed to deal with (12) is a regularized proximal six-block Gauss-Seidel method. At the kth outer iteration, the problem (12) can be solved with arbitrary accuracy using the following alternating minimizing procedure:

(a) Update $W^{k,j}$:

$$W^{k,j} \in \arg\min \{L(W, U^{k,j-1}, V^{k,j-1}, Y^{k,j-1}, F^{k,j-1}, \widehat{Y}^{k,j-1}, \overline{\lambda}^k; \rho^k) + \frac{C_1^{k,j-1}}{2} \|W - W^{k,j-1}\|_F^2\}, \tag{13}$$

(b) Update $U^{k,j}$:

$$U^{k,j} \in \arg\min \{L(W^{k,j}, U, V^{k,j-1}, Y^{k,j-1}, F^{k,j-1}, \widehat{Y}^{k,j-1}, \overline{\lambda}^k; \rho^k) + \frac{C_2^{k,j-1}}{2}\|U - U^{k,j-1}\|_F^2\},$$
(14)

(c) Update $V^{k,j}$:

$$V^{k,j} \in \arg\min \{L(W^{k,j}, U^{k,j}, V, Y^{k,j-1}, F^{k,j-1}, \widehat{Y}^{k,j-1}, \overline{\lambda}^k; \rho^k) + \frac{C_3^{k,j-1}}{2}\|V - V^{k,j-1}\|_F^2\},$$
(15)

(d) Update $Y^{k,j}$:

$$Y^{k,j} \in \arg\min \{L(W^{k,j}, U^{k,j}, V^{k,j}, Y, F^{k,j-1}, \widehat{Y}^{k,j-1}, \overline{\lambda}^k; \rho^k) + \frac{C_4^{k,j-1}}{2}\|Y - Y^{k,j-1}\|_F^2\}, \quad (16)$$

(e) Update $F^{k,j}$:

$$F^{k,j} \in \arg\min \{L(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F, \widehat{Y}^{k,j-1}, \overline{\lambda}^k; \rho^k) + \frac{C_5^{k,j-1}}{2}\|F - F^{k,j-1}\|_F^2\}, \quad (17)$$

(f) Update $\widehat{Y}^{k,j}$:

$$\widehat{Y}^{k,j} \in \arg\min \{L(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F^{k,j}, \widehat{Y}, \overline{\lambda}^k; \rho^k) + \frac{C_6^{k,j-1}}{2}\|\widehat{Y} - \widehat{Y}^{k,j-1}\|_F^2\}, \quad (18)$$

where the proximal parameters $\{C_i^{k,j}\}_{k,j}$ need to satisfy

$$0 < \underline{C} \le C_i^{k,j} < \overline{C} < \infty, \ k, j \in \mathbb{N}, \ i = 1, 2, 3, 4, 5, 6,$$

for some predetermined positive constants $\underline{C}$ and $\overline{C}$.

By direct calculation, the subproblems in (13-18) have closed-form solutions as follows:

(a) For (13):

$$W^{k,j} = \left(\frac{1}{a}I_d - \frac{\rho^k}{a^2}X(I_n + \frac{\rho^k}{a}X^\mathsf{T}X)^{-1}X^\mathsf{T}\right) Z,$$

where $a = 2\gamma + \rho^k + C_1^{k,j-1}$ and

$$Z = X\overline{\lambda}_1^k + \overline{\lambda}_2^k + \rho^k XY^{k,j-1} - \rho^k XU^{k,j-1} + \rho^k V^{k,j-1} + C_1^{k,j-1}W^{k,j-1}.$$

(b) For (14): $U^{k,j} = (U_i^{k,j})_{i \in [n]}$, where $U_i^{k,j}$ is the row vector of $U^{k,j}$.

Set

$$N = Y^{k,j-1} - X^\mathsf{T}W^{k,j} + \frac{\overline{\lambda}_1^k}{\rho^k}$$

and denote n_i as the i-th row vector of N. Then,

$$U_i^{k,j} = \max\left\{0, 1 - \frac{\alpha}{\|\rho^k n_i + C_2^{k,j-1}U_i^{k,j-1}\|_2}\right\}\left[\frac{\rho^k}{\rho^k + C_2^{k,j-1}}n_i + \frac{C_2^{k,j-1}}{\rho^k + C_2^{k,j-1}}U_i^{k,j-1}\right].$$

(c) For (15): $V^{k,j} = (V_i^{k,j})_{i \in [d]}$, where $V_i^{k,j}$ is the row vector of $V^{k,j}$.

Set
$$M = W^{k,j} - \frac{\overline{\lambda}_2^k}{\rho^k}$$

and its row vector is denoted by m_i. Then,

$$V_i^{k,j} = \max \left\{ 0, 1 - \frac{\beta}{\|\rho^k m_i + C_3^{k,j-1} V_i^{k,j-1}\|_2} \right\} \left[\frac{\rho^k}{\rho^k + C_3^{k,j-1}} m_i + \frac{C_3^{k,j-1}}{\rho^k + C_3^{k,j-1}} V_i^{k,j-1} \right].$$

(d) For (16):
$$Y^{k,j} = [2L + (3\rho^k + C_4^{k,j-1})I]^{-1} P,$$

where
$$P = \overline{\lambda}_4^k - \overline{\lambda}_3^k - \overline{\lambda}_1^k + \rho^k X^\mathsf{T} W^{k,j} + \rho^k U^{k,j} + \rho^k F^{k,j-1} + \rho^k \widehat{Y}^{k,j-1} + C_4^{k,j-1} Y^{k,j-1}.$$

(e) For (17):
$F^{k,j} = (F_{s,t}^{k,j})_{s \in [n], t \in [c]}$ and $F_{s,t}^{k,j} = \Pi_{[0,1]} A_{st}$, where

$$A = (A_{s,t})_{s \in [n], t \in [c]} = \frac{\rho^k (Y^{k,j} + \frac{\overline{\lambda}_3^k}{\rho^k}) + C_5^{k,j-1} F^{k,j-1}}{\rho^k + C_5^{k,j-1}}.$$

(f) For (18):
$\widehat{Y}^{k,j} = U I_{n \times c} V^\mathsf{T}$, where $U \in \mathbb{R}^{n \times n}, V \in \mathbb{R}^{c \times c}$ are two orthogonal matrices and $\sum \in \mathbb{R}^{n \times c}$ is a diagonal matrix satisfying the SVD factorization

$$\frac{\rho^k (Y^{k,j} - \frac{\overline{\lambda}_4^k}{\rho^k}) + C_6^{k,j-1} \widehat{Y}^{k,j-1}}{\rho^k + C_6^{k,j-1}} = U \sum V^\mathsf{T}.$$

The iteration is terminated if there exists $\Theta^{k,j} \in \partial L(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F^{k,j}, \widehat{Y}^{k,j}, \overline{\lambda}^k; \rho^k)$ satisfying
$$\|\Theta^{k,j}\|_\infty \le \epsilon_k, \quad \mathbf{O} \le F^{k,j} \le \mathbf{E}, \quad (\widehat{Y}^{k,j})^\mathsf{T} \widehat{Y}^{k,j} = I_c,$$
where $\Theta^{k,j} := (\Theta_1^{k,j}, \Theta_2^{k,j}, \Theta_3^{k,j}, \Theta_4^{k,j}, \Theta_5^{k,j}, \Theta_6^{k,j}) \in \mathbb{R}^{d \times c} \times \mathbb{R}^{n \times c} \times \mathbb{R}^{d \times c} \times \mathbb{R}^{n \times c} \times \mathbb{R}^{n \times c} \times \mathbb{R}^{n \times c}$ is concretely expressed in the form

$$\begin{cases}
\Theta_1^{k,j} := \rho^k X(Y^{k,j-1} - Y^{k,j}) + \rho^k X(U^{k,j} - U^{k,j-1}) + \rho^k(V^{k,j-1} - V^{k,j}) \\
\qquad + C_1^{k,j-1}(W^{k,j-1} - W^{k,j}) \\
\Theta_2^{k,j} := \rho^k(Y^{k,j-1} - Y^{k,j}) + C_2^{k,j-1}(U^{k,j-1} - U^{k,j}) \\
\Theta_3^{k,j} := C_3^{k,j-1}(V^{k,j-1} - V^{k,j}) \\
\Theta_4^{k,j} := \rho^k(F^{k,j-1} - F^{k,j}) + \rho^k(\widehat{Y}^{k,j-1} - \widehat{Y}^{k,j}) + C_4^{k,j-1}(Y^{k,j-1} - Y^{k,j}) \\
\Theta_5^{k,j} := C_5^{k,j-1}(F^{k,j-1} - F^{k,j}) \\
\Theta_6^{k,j} := C_6^{k,j-1}(\widehat{Y}^{k,j-1} - \widehat{Y}^{k,j}).
\end{cases} \tag{19}$$

We summarize the algorithmic framework of PAM in Algorithm 2, whose convergence analysis is established in the next subsection.

Algorithm 2 PAM Method for (8-10)

Input:
Let $(W^{1,0}, U^{1,0}, V^{1,0}, Y^{1,0}, F^{1,0}, \widehat{Y}^{1,0})$ be any initialization;
For $k \geq 2$, set $(W^{k,0}, U^{k,0}, V^{k,0}, Y^{k,0}, F^{k,0}, \widehat{Y}^{k,0}) = (W^{k-1}, U^{k-1}, V^{k-1}, Y^{k-1}, F^{k-1}, \widehat{Y}^{k-1})$;

Output: $(W^k, U^k, V^k, Y^k, F^k, \widehat{Y}^k)$;
Step 1: Reiterate on j until $\|\Theta^{k,j}\|_\infty \leq \epsilon_k$, where $\Theta^{k,j}$ is defined by (19);

1. Compute $W^{k,j}$ by (13);

2. Compute $U^{k,j}$ by (14);

3. Compute $V^{k,j}$ by (15);

4. Compute $Y^{k,j}$ by (16);

5. Compute $F^{k,j}$ by (17);

6. Compute $\widehat{Y}^{k,j}$ by (18);

Step 2: Set

$$(W^k, U^k, V^k, Y^k, F^k, \widehat{Y}^k) := (W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F^{k,j}, \widehat{Y}^{k,j})$$

and $\Theta^k := \Theta^{k.j}$.

4.2 Convergence Analysis for Algorithm 2

For the sake of notation simplicity, we fix some notations. We define $T := (W, U, V, Y, F, \widehat{Y})$ and $L_k(T) := L(W, U, V, Y, F, \widehat{Y}, \overline{\lambda}^k; \rho^k)$ for the k-th outer iteration. In this part, we will establish the global convergence for Algorithm 2, in other words, we can derive that the solution set of (8-10) is nonempty and hence Algorithm 1 is well defined with using Algorithm 2 to solve the subproblem in Step 1.

We first claim that $\Theta^{k,j} := (\Theta_1^{k,j}, \Theta_2^{k,j}, \Theta_3^{k,j}, \Theta_4^{k,j}, \Theta_5^{k,j}, \Theta_6^{k,j})$ defined by (19) must satisfy

$$\Theta^{k,j} \in \partial L_k(T^{k,j}) \quad \forall j \in \mathbb{N}.$$

for each $k \in \mathbb{N}$.

Considering the structure of $L_k(T)$, it can be split as

$$L_k(T) = f_1(W) + f_2(Y) + f_3(U) + f_4(V) + f_5(F) + f_6(\widehat{Y}) + g_k(T), \tag{20}$$

where

$$\begin{cases} f_1(W) := \gamma \|W\|_F^2; \quad f_2(Y) := \text{Tr}(Y^\mathsf{T} L Y); \quad f_3(U) := \alpha \|U\|_{2,1}; \\ f_4(V) := \beta \|V\|_{2,1}; \quad f_5(F) := \delta_{\mathbb{S}_2}(F); \quad f_6(\widehat{Y}) := \delta_{\mathbb{S}_1}(\widehat{Y}); \\ g_k(T) := \langle \overline{\lambda}_1^k, Y - X^\mathsf{T} W - U \rangle + \langle \overline{\lambda}_2^k, V - W \rangle + \langle \overline{\lambda}_3^k, Y - F \rangle + \langle \overline{\lambda}_4^k, \widehat{Y} - Y \rangle \\ \qquad + \dfrac{\rho^k}{2} \Big(\|Y - X^\mathsf{T} W - U\|_F^2 + \|\widehat{Y} - Y\|_F^2 + \|V - W\|_F^2 + \|Y - F\|_F^2 \Big). \end{cases}$$

Then, a direct calculation shows that $\Theta^{k,j} := (\Theta_1^{k,j}, \Theta_2^{k,j}, \Theta_3^{k,j}, \Theta_4^{k,j}, \Theta_5^{k,j}, \Theta_6^{k,j})$ defined by (19) can be expressed in terms of partial derivatives of $g := g_k$ as

$$
\begin{cases}
\Theta_1^{k,j} = -\nabla_W g(W^{k,j}, U^{k,j-1}, V^{k,j-1}, Y^{k,j-1}, F^{k,j-1}, \widehat{Y}^{k,j-1}) - C_1^{k,j-1}(W^{k,j} - W^{k,j-1}) + \nabla_W g(T^{k,j}), \\
\Theta_2^{k,j} = -\nabla_U g(W^{k,j}, U^{k,j}, V^{k,j-1}, Y^{k,j-1}, F^{k,j-1}, \widehat{Y}^{k,j-1}) - C_2^{k,j-1}(U^{k,j} - U^{k,j-1}) + \nabla_U g(T^{k,j}), \\
\Theta_3^{k,j} = -\nabla_V g(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j-1}, F^{k,j-1}, \widehat{Y}^{k,j-1}) - C_3^{k,j-1}(V^{k,j} - V^{k,j-1}) + \nabla_V g(T^{k,j}), \\
\Theta_4^{k,j} = -\nabla_Y g(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F^{k,j-1}, \widehat{Y}^{k,j-1}) - C_4^{k,j-1}(Y^{k,j} - Y^{k,j-1}) + \nabla_Y g(T^{k,j}), \\
\Theta_5^{k,j} = -\nabla_F g(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F^{k,j}, \widehat{Y}^{k,j-1}) - C_5^{k,j-1}(F^{k,j} - F^{k,j-1}) + \nabla_F g(T^{k,j}), \\
\Theta_6^{k,j} = -\nabla_{\widehat{Y}} g(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F^{k,j}, \widehat{Y}^{k,j}) - C_6^{k,j-1}(\widehat{Y}^{k,j} - \widehat{Y}^{k,j-1}) + \nabla_{\widehat{Y}} g(T^{k,j}).
\end{cases}
$$
(21)

Moreover, given $(W^{k,j-1}, U^{k,j-1}, V^{k,j-1}, Y^{k,j-1}, F^{k,j-1}, \widehat{Y}^{k,j-1})$, using 8.8(c) in Rockafellar and Wets (2009), the necessary first-order optimality conditions for the subproblem (13-18) are the following system:

$$
\begin{cases}
\nabla_W g(W^{k,j}, U^{k,j-1}, V^{k,j-1}, Y^{k,j-1}, F^{k,j-1}, \widehat{Y}^{k,j-1}) + \nabla f_1(W^{k,j}) + C_1^{k,j-1}(W^{k,j} - W^{k,j-1}) = \mathbf{O}, \\
\xi^{k,j} + \nabla_U g(W^{k,j}, U^{k,j}, V^{k,j-1}, Y^{k,j-1}, F^{k,j-1}, \widehat{Y}^{k,j-1}) + C_2^{k,j-1}(U^{k,j} - U^{k,j-1}) = \mathbf{O}, \\
\zeta^{k,j} + \nabla_V g(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j-1}, F^{k,j-1}, \widehat{Y}^{k,j-1}) + C_3^{k,j-1}(V^{k,j} - V^{k,j-1}) = \mathbf{O}, \\
\nabla f_2(Y^{k,j}) + \nabla_Y g(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F^{k,j-1}, \widehat{Y}^{k,j-1}) + C_4^{k,j-1}(Y^{k,j} - Y^{k,j-1}) = \mathbf{O}, \\
\vartheta^{k,j} + \nabla_F g(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F^{k,j}, \widehat{Y}^{k,j-1}) + C_5^{k,j-1}(F^{k,j} - F^{k,j-1}) = \mathbf{O}, \\
\varsigma^{k,j} + \nabla_{\widehat{Y}} g(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F^{k,j}, \widehat{Y}^{k,j}) + C_6^{k,j-1}(\widehat{Y}^{k,j} - \widehat{Y}^{k,j-1}) = \mathbf{O},
\end{cases}
$$
(22)

where $\xi^{k,j} \in \partial f_3(U^{k,j})$, $\zeta^{k,j} \in \partial f_4(V^{k,j})$, $\vartheta^{k,j} \in \partial f_5(F^{k,j})$, and $\varsigma^{k,j} \in \partial f_6(\widehat{Y}^{k,j})$. Combining (21) with (22), we have

$$
\begin{cases}
\Theta_1^{k,j} = \nabla f_1(W^{k,j}) + \nabla_W g(T^{k,j}), \\
\Theta_2^{k,j} = \xi^{k,j} + \nabla_U g(T^{k,j}), \\
\Theta_3^{k,j} = \zeta^{k,j} + \nabla_V g(T^{k,j}), \\
\Theta_4^{k,j} = \nabla f_2(Y^{k,j}) + \nabla_Y g(T^{k,j}), \\
\Theta_5^{k,j} = \vartheta^{k,j} + \nabla_F g(T^{k,j}), \\
\Theta_6^{k,j} = \varsigma^{k,j} + \nabla_{\widehat{Y}} g(T^{k,j}).
\end{cases}
$$

By Proposition 2.1 in Attouch et al. (2010), for each $k \in \mathbb{N}$, we get

$$
\Theta^{k,j} \in \partial L_k(W^{k,j}, U^{k,j}, V^{k,j}, Y^{k,j}, F^{k,j}, \widehat{Y}^{k,j}), \quad \forall j \in \mathbb{N}.
$$

Thus, we can obtain the following theorem which shows that Algorithm 2 converges, which means the Step 1 of Algorithm 1 is well defined. The proof is based on a general result established in Attouch et al. (2013, Theorem 6.2).

Theorem 6 *Set parameters $r > 1, \rho^1 > 0$ in Algorithm 1. For each $k \in \mathbb{N}$, we have the sequence $\{T^{k,j}\}_{j\in\mathbb{N}}$ produced by Algorithm 2 converges and*

$$
\|\Theta^{k,j}\|_\infty \to 0 \quad as \ j \to \infty.
$$

Proof We know that $\mathbb{S}_1$ and $\mathbb{S}_2$ are semi-algebraic sets and their indicator functions are semi-algebraic (Bolte et al., 2014). The quadratic functions $\mathbf{x}^{\intercal} L \mathbf{x}$ and $\|\mathbf{x}\|_p$ (p is rational) are also semi-algebraic. Using the fact that composition of semi-algebraic functions is semi-algebraic, we derive that L_k is a semi-algebraic function. Also known is that the semi-algebraic function is a Kurdyka-Łojasiewicz (KL) function (Bolte et al., 2014, Appendic). Thus, L_k is a KL function. From the expression (20) of L_k, it can be seen that the function L_k satisfies: (i)f_i is a proper lower semicontinuous function, $i = 1, 2, 3, 4, 5, 6$; (ii) g_k is a C^1-function with locally Lipschitz continuous gradient.

Next, we will verify that for each $k \in \mathbb{N}$, L_k is bounded below and the sequence $\{T^{k,j}\}_{j \in \mathbb{N}}$ is bounded. For each $k \in \mathbb{N}$, the lower boundness of L_k is proved by showing that L_k is a coercive function (i.e., $L_k(T) \to +\infty$ when $\|T\|_\infty \to \infty$), provided that the parameters $r > 1, \rho^1 > 0$. Clearly, the five terms f_1, f_3, f_4, f_5, f_6 of L_k in (20) are coercive. Then the residual terms are

$$f_2(Y) + g_k(W, U, V, Y, F, \widehat{Y}) = \mathrm{Tr}(Y^{\intercal} L Y) + \langle \overline{\lambda}_1^k, Y - X^{\intercal} W - U \rangle + \langle \overline{\lambda}_2^k, V - W \rangle + \langle \overline{\lambda}_3^k, Y - F \rangle$$
$$+ \langle \overline{\lambda}_4^k, \widehat{Y} - Y \rangle + \frac{\rho^k}{2} \Big(\|\widehat{Y} - Y\|_F^2 + \|V - W\|_F^2 + \|Y - F\|_F^2$$
$$+ \|Y - X^{\intercal} W - U\|_F^2 \Big).$$

We can rewrite it as

$$f_2(Y) + g_k(W, U, V, Y, F, \widehat{Y}) = g_{1,k}(W, U, Y) + g_{2,k}(W, V, Y, F, \widehat{Y}),$$

where

$$g_{1,k}(W, U, Y) = \mathrm{Tr}(Y^{\intercal} L Y) + \langle \overline{\lambda}_1^k, Y - X^{\intercal} W - U \rangle + \frac{\rho^k}{2} \|Y - X^{\intercal} W - U\|_F^2$$

and

$$g_{2,k}(W, V, Y, F, \widehat{Y}) = \langle \overline{\lambda}_2^k, V - W \rangle + \langle \overline{\lambda}_3^k, Y - F \rangle + \langle \overline{\lambda}_4^k, \widehat{Y} - Y \rangle + \frac{\rho^k}{2} (\|\widehat{Y} - Y\|_F^2 + \|V - W\|_F^2$$
$$+ \|Y - F\|_F^2).$$

Let us observe that

$$g_{1,k}(W, U, Y) = \mathrm{Tr}(Y^{\intercal} L Y) + \frac{\rho^k}{2} \|Y - X^{\intercal} W - U + \frac{\overline{\lambda}_1^k}{\rho^k}\|_F^2 - \frac{\rho^k}{2} \|\frac{\overline{\lambda}_1^k}{\rho^k}\|_F^2$$

and

$$g_{2,k}(W, V, Y, F, \widehat{Y}) = \frac{\rho^k}{2} \Big[\|V - W + \frac{\overline{\lambda}_2^k}{\rho^k}\|_F^2 + \|Y - F + \frac{\overline{\lambda}_3^k}{\rho^k}\|_F^2 + \|\widehat{Y} - Y + \frac{\overline{\lambda}_4^k}{\rho^k}\|_F^2 - (\|\frac{\overline{\lambda}_2^k}{\rho^k}\|_F^2$$
$$+ \|\frac{\overline{\lambda}_3^k}{\rho^k}\|_F^2 + \|\frac{\overline{\lambda}_4^k}{\rho^k}\|_F^2) \Big].$$

Thus, $g_{1,k}(W, U, Y)$ and $g_{2,k}(W, V, Y, F, \widehat{Y})$ are all bounded below. Furthermore, the functions $\{L_k\}_k \in \mathbb{N}$ defined by (20) are all coercive.

13

The boundedness of the sequence $\{T^{k,j}\}_{j\in\mathbb{N}}$ is proved by contradiction. On the one hand, suppose that the sequence $\{T^{k_0,j}\}_{j\in\mathbb{N}}$ is unbounded, and so $\lim_{j\to\infty}\|T^{k_0,j}\|=\infty$. Then, it follows from the coercive of $L_{k_0}(T)$ that the sequence $\{L_{k_0}(T^{k_0,j})\}_{j\in\mathbb{N}}$ should diverge to infinity. On the other hand, let

$$\widetilde{L}_{k_0,j}^1 = L_{k_0}(W^{k_0,j+1}, U^{k_0,j}, V^{k_0,j}, Y^{k_0,j}, F^{k_0,j}, \widehat{Y}^{k_0,j}),$$
$$\widetilde{L}_{k_0,j}^2 = L_{k_0}(W^{k_0,j+1}, U^{k_0,j+1}, V^{k_0,j}, Y^{k_0,j}, F^{k_0,j}, \widehat{Y}^{k_0,j}),$$
$$\widetilde{L}_{k_0,j}^3 = L_{k_0}(W^{k_0,j+1}, U^{k_0,j+1}, V^{k_0,j+1}, Y^{k_0,j}, F^{k_0,j}, \widehat{Y}^{k_0,j}),$$
$$\widetilde{L}_{k_0,j}^4 = L_{k_0}(W^{k_0,j+1}, U^{k_0,j+1}, V^{k_0,j+1}, Y^{k_0,j+1}, F^{k_0,j}, \widehat{Y}^{k_0,j}),$$
$$\widetilde{L}_{k_0,j}^5 = L_{k_0}(W^{k_0,j+1}, U^{k_0,j+1}, V^{k_0,j+1}, Y^{k_0,j+1}, F^{k_0,j+1}, \widehat{Y}^{k_0,j}).$$

By (13-18), we deduce that

$$\widetilde{L}_{k_0,j}^1 + \frac{C_1^{k_0,j}}{2}\|W^{k_0,j+1} - W^{k_0,j}\|_F^2 \leq L_{k_0}(T^{k_0,j});$$

$$\widetilde{L}_{k_0,j}^2 + \frac{C_2^{k_0,j}}{2}\|U^{k_0,j+1} - U^{k_0,j}\|_F^2 \leq \widetilde{L}_{k_0,j}^1;$$

$$\widetilde{L}_{k_0,j}^3 + \frac{C_3^{k_0,j}}{2}\|V^{k_0,j+1} - V^{k_0,j}\|_F^2 \leq \widetilde{L}_{k_0,j}^2;$$

$$\widetilde{L}_{k_0,j}^4 + \frac{C_4^{k_0,j}}{2}\|Y^{k_0,j+1} - Y^{k_0,j}\|_F^2 \leq \widetilde{L}_{k_0,j}^3;$$

$$\widetilde{L}_{k_0,j}^5 + \frac{C_5^{k_0,j}}{2}\|F^{k_0,j+1} - F^{k_0,j}\|_F^2 \leq \widetilde{L}_{k_0,j}^4;$$

$$L_{k_0}(T^{k_0,j+1}) + \frac{C_6^{k_0,j}}{2}\|\widehat{Y}^{k_0,j+1} - \widehat{Y}^{k_0,j}\|_F^2 \leq \widetilde{L}_{k_0,j}^5.$$

Summing up these inequalities, we have

$$L_{k_0}(T^{k_0,j+1}) + \frac{C}{2}\|T^{k_0,j+1} - T^{k_0,j}\|_F^2 \leq L_{k_0}(T^{k_0,j}), \ j\in\mathbb{N},$$

which implies that $\{L_{k_0}(T^{k_0,j})\}_{j\in\mathbb{N}}$ is a nonincreasing sequence, leading to a contradiction.

Based on a general result established in Attouch et al. (2013, Theorem 6.2), we know that for each $k\in\mathbb{N}$, the sequence $\{T^{k,j}\}_{j\in\mathbb{N}}$ has finite length, i.e., $\sum_{j=1}^{\infty}\|T^{k,j+1} - T^{k,j}\|_F < \infty$, and the sequence $\{T^{k,j}\}_{j\in\mathbb{N}}$ converges to a critical point of L_k. Since $\Theta^{k,j}$ is given by (19), we conclude that for each $k\in\mathbb{N}$, $\|\Theta^{k,j}\|_\infty \to 0$ as $j\to\infty$. The proof is complete. ∎

5. Convergence Analysis of Our Inexact ALM Method

In this section, we discuss the convergence for our inexact ALM method given in Algorithm 1.

In the following, we rewrite (6) using the notation of vectors. Let $\mathbf{x}\in\mathbb{R}^{2dc+4nc}$ denote the column vector formed by concatenating the columns of $W, U, V, Y, F, \widehat{Y}$, i.e.,

$$\mathbf{x} := \text{Vec}([W|U|V|Y|F|\widehat{Y}]). \tag{23}$$

14

Then, problem (6) can be rewritten as follows:

$$\min_{\mathbf{x}\in\mathbb{R}^{2dc+4nc}} f(x) \quad \text{s.t. } h_1(\mathbf{x}) = 0 \text{ and } h_2(\mathbf{x}) = 0, \tag{24}$$

where $h_1(\mathbf{x}) \in \mathbb{R}^{3nc+dc}$ denotes $\text{Vec}([Y - X^\mathsf{T}W - U|V - W|Y - F|\widehat{Y} - Y])$, $h_2(\mathbf{x})$ denotes the $\frac{c(c+1)}{2} \times 1$ vector obtained by vectorizing only the lower triangular part of the symmetric matrix $\widehat{Y}^\mathsf{T}\widehat{Y} - I_c$, and

$$f(\mathbf{x}) := \sum_{j=1}^{c} Y_j^\mathsf{T} L Y_j + \gamma\|W_j\|_2^2 + \delta_{S'}(F_j) + \sum_{i=1}^{n} \alpha\|U_i\|_2 + \sum_{i=1}^{d} \beta\|V_i\|_2.$$

In this case, Y_j, W_j, F_j are the column vectors of Y, W and F, respectively; U_i and V_i are the row vectors of U and V, respectively; $S' = \{F_j \mid \mathbf{0} \leq F_j \leq \mathbf{1}\}$. Let $\Lambda :=$ $\text{Vec}([\lambda_1|\lambda_2|\lambda_3|\lambda_4])$. Then, the corresponding augmented Lagrangian function of (24) is

$$L(\mathbf{x}, \Lambda; \rho) := f(\mathbf{x}) + \sum_{i=1}^{m_1}[\Lambda]_i[h_1(\mathbf{x})]_i + \frac{\rho}{2}\sum_{i=1}^{m_1}[h_1(\mathbf{x})]_i^2,$$

where $\mathbf{x} \in \Gamma$, $m_1 := 3nc + dc$, $m_2 := \frac{c(c+1)}{2}$ and

$$\Gamma := \{\mathbf{x} \mid h_2(\mathbf{x}) = 0\}. \tag{25}$$

Therefore, $(W^*, U^*, V^*, Y^*, F^*, \widehat{Y}^*)$ is a KKT point for optimization problem (6) if and only if the vector $\mathbf{x}$ defined by (23) is a KKT point for optimization problem (24), i.e., there exist $\theta^* \in \partial f(\mathbf{x}^*)$, $\Lambda^* \in \mathbb{R}^{m_1}$, $\eta^* \in \mathbb{R}^{m_2}$ such that the following system is fulfilled

$$\begin{cases} \theta^* + \sum_{i=1}^{m_1}[\Lambda^*]_i\nabla[h_1(\mathbf{x}^*)]_i + \sum_{i=1}^{m_2}[\eta^*]_i\nabla[h_2(\mathbf{x}^*)]_i = 0, \\ h_1(\mathbf{x}^*) = 0, \\ h_2(\mathbf{x}^*) = 0. \end{cases} \tag{26}$$

Suppose that $\{T^k\}_{k\in\mathbb{N}}$ is a sequence generated by Algorithm 1. We will show first that the sequence $\{T^k\}_{k\in\mathbb{N}}$ is bounded. Then, there exists at least one convergent subsequence of $\{T^k\}_{k\in\mathbb{N}}$. We will next show that it converges to a KKT point of the optimization problem (24). Thus, we have the following main convergence result for Algorithm 1.

Theorem 7 *Suppose that the parameters $r > 1$ and $\rho^1 > 0$ in Algorithm 1. Let $\{T^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Then, the limit point set of $\{T^k\}_{k\in\mathbb{N}}$ is nonempty, and every limit point is a KKT point of the original problem (6).*

To show Theorem 7, we need the following two lemmas.

Lemma 8 *Let $\{T^k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Suppose that the parameters r, ρ^1 in Algorithm 1 are chosen so that $r > 1$ and $\rho^1 > 0$. Then, $\{T^k\}_{k\in\mathbb{N}}$ is bounded and thus contains at least one convergent sequence.*

Proof It follows from (9) that the sequence $\{F^k\}_{k\in\mathbb{N}}$ and $\{\widehat{Y}^k\}_{k\in\mathbb{N}}$ are bounded. The first four partial subdifferentials of L in (10) guarantee the following: there exist $\xi^k \in \partial\alpha\|U\|_{2,1}$, $\zeta^k \in \partial\beta\|V\|_{2,1}$ and $\aleph^k = (\aleph_1^k, \aleph_2^k, \aleph_3^k, \aleph_4^k) \in \mathbb{R}^{d\times c} \times \mathbb{R}^{n\times c} \times \mathbb{R}^{d\times c} \times \mathbb{R}^{n\times c}$ such that

$$
\begin{cases}
\aleph_1^k = 2\gamma W^k - X\overline{\lambda}_1^k - \overline{\lambda}_2^k - \rho^k X(Y^k - X^\intercal W^k - U^k) - \rho^k(V^k - W^k), \\
\aleph_2^k = \xi^k - \overline{\lambda}_1^k - \rho^k(Y^k - X^\intercal W^k - U^k), \\
\aleph_3^k = \zeta^k + \overline{\lambda}_2^k + \rho^k(V^k - W^k), \\
\aleph_4^k = 2LY^k + \overline{\lambda}_1^k + \overline{\lambda}_3^k - \overline{\lambda}_4^k + \rho^k(Y^k - X^\intercal W^k - U^k) + \rho^k(Y^k - F^k) - \rho^k(\widehat{Y}^k - Y^k),
\end{cases}
\tag{27}
$$

where $\|\aleph^k\|_\infty \le \epsilon^k$. By adding $\aleph_2^k$ and $\aleph_4^k$, we obtain that

$$
\aleph_2^k + \aleph_4^k = \xi^k + (2L + 2\rho^k I)Y^k + \overline{\lambda}_3^k - \overline{\lambda}_4^k - \rho^k F^k - \rho^k \widehat{Y}^k.
$$

This implies

$$
Y^k = [2(L + \rho^k I)]^{-1}(\aleph_2^k + \aleph_4^k - \xi^k - \overline{\lambda}_3^k + \overline{\lambda}_4^k + \rho^k F^k + \rho^k \widehat{Y}^k).
\tag{28}
$$

Let $L = D\mathrm{diag}(\sigma_1, \cdots, \sigma_n)D^\intercal$ denotes the SVD decomposition of the symmetric and positive semi-definite matrix L. Hence (28) yields

$$
\begin{aligned}
Y^k =& D\mathrm{diag}\left(\frac{1}{2(\sigma_1 + \rho^k)}, \frac{1}{2(\sigma_2 + \rho^k)}, \cdots, \frac{1}{2(\sigma_n + \rho^k)}\right) D^\intercal(\aleph_2^k + \aleph_4^k - \xi^k - \overline{\lambda}_3^k + \overline{\lambda}_4^k) \\
&+ D\mathrm{diag}\left(\frac{\rho^k}{2(\sigma_1 + \rho^k)}, \frac{\rho^k}{2(\sigma_2 + \rho^k)}, \cdots, \frac{\rho^k}{2(\sigma_n + \rho^k)}\right) D^\intercal(F^k + \widehat{Y}^k).
\end{aligned}
\tag{29}
$$

Using the fact $\{\rho^k\}_{k\in\mathbb{N}}$ is a nondecreasing sequence and $2(L + \rho^1 I) \succ 0$, for $k \in \mathbb{N}$, we have $2(L + \rho^k I) \succ 0$, which derives $2(\sigma_i + \rho^k) > 0$, $i = 1, 2, \cdots, n$. Then, we can show that for each $k \in \mathbb{N}$

$$
\begin{cases}
0 < \frac{1}{2(\sigma_i + \rho^k)} \le \frac{1}{2(\sigma_i + \rho^1)} < +\infty, & i = 1, 2, \cdots, n; \\
0 < \frac{\rho^k}{2(\sigma_i + \rho^k)} \le \frac{1}{2}, & i = 1, 2, \cdots, n.
\end{cases}
\tag{30}
$$

Note that $\{\xi^k\}_{k\in\mathbb{N}}$, $\{\aleph_2^k\}_{k\in\mathbb{N}}$, $\{\aleph_4^k\}_{k\in\mathbb{N}}$, $\{\overline{\lambda}_3^k\}_{k\in\mathbb{N}}$ and $\{\overline{\lambda}_4^k\}_{k\in\mathbb{N}}$ are bounded. It follows from (29) and (30) that the sequence $\{Y^k\}_{k\in\mathbb{N}}$ is bounded.

Likewise, according to the expression of $\aleph_3^k$ and $\aleph_2^k$ in (27), we conclude that $\{\rho^k(V^k - W^k)\}_{k\in\mathbb{N}}$ and $\{\rho^k(Y^k - X^\intercal W^k - U^k)\}_{k\in\mathbb{N}}$ are bounded. Then, from the expression of $\aleph_1^k$ in (27), we must have that the sequence $\{W^k\}_{k\in\mathbb{N}}$ is bounded. Using the fact that $\rho^k \ge \rho^1$ again, we obtain that $\{V^k - W^k\}_{k\in\mathbb{N}}$ and $\{Y^k - X^\intercal W^k - U^k\}_{k\in\mathbb{N}}$ are bounded. Therefore, the sequence $\{V^k\}_{k\in\mathbb{N}}$ and $\{U^k\}_{k\in\mathbb{N}}$ are bounded. In a conclusion, the sequence $\{(W^k, U^k, V^k, Y^k, F^k, \widehat{Y}^k)\}_{k\in\mathbb{N}}$ is bounded. The proof is complete. ∎

Lemma 9 *Suppose that $\bar{\boldsymbol{x}} \in \Gamma$. Then $\{\nabla[h_1(\bar{\boldsymbol{x}})]_i\}_{i=1}^{m_1} \cup \{\nabla[h_2(\bar{\boldsymbol{x}})]_i\}_{i=1}^{m_2}$ are linearly independent, where h_1 and h_2 are defined as in (24).*

$$G(\mathbf{x}) = \begin{bmatrix} 2\widehat{Y}_1 & \widehat{Y}_2 & \widehat{Y}_3 & \cdots & \widehat{Y}_c & \mathbf{O}_{n\times 1} & \mathbf{O}_{n\times 1} & \cdots & \mathbf{O}_{n\times 1} & \mathbf{O}_{n\times 1} & \mathbf{O}_{n\times 1} & \mathbf{O}_{n\times 1} \\ \mathbf{O}_{n\times 1} & \widehat{Y}_1 & \mathbf{O}_{n\times 1} & \cdots & \mathbf{O}_{n\times 1} & 2\widehat{Y}_2 & \widehat{Y}_3 & \cdots & \widehat{Y}_c & \vdots & \vdots & \vdots \\ \mathbf{O}_{n\times 1} & \mathbf{O}_{n\times 1} & \widehat{Y}_1 & \cdots & \mathbf{O}_{n\times 1} & \mathbf{O}_{n\times 1} & \widehat{Y}_2 & \cdots & \mathbf{O}_{n\times 1} & \cdots & \mathbf{O}_{n\times 1} & \mathbf{O}_{n\times 1} & \vdots \\ \vdots & \vdots & \ddots & \ddots & \mathbf{O}_{n\times 1} & \vdots & \ddots & \ddots & \vdots & 2\widehat{Y}_{c-1} & \widehat{Y}_c & \mathbf{O}_{n\times 1} \\ \mathbf{O}_{n\times 1} & \mathbf{O}_{n\times 1} & \cdots & \mathbf{O}_{n\times 1} & \widehat{Y}_1 & \mathbf{O}_{n\times 1} & \cdots & \mathbf{O}_{n\times 1} & \widehat{Y}_2 & \mathbf{O}_{n\times 1} & \widehat{Y}_{c-1} & 2\widehat{Y}_c \end{bmatrix} \tag{31}$$

Proof For convenience, we define the block diagonal matrix $A \in \mathbb{R}^{dc\times nc}$, $B \in \mathbb{R}^{dc\times dc}$ and $C \in \mathbb{R}^{nc\times nc}$ as follows:

$$A = \begin{bmatrix} -X & & & \\ & -X & & \\ & & \ddots & \\ & & & -X \end{bmatrix}, \quad B = \begin{bmatrix} I_d & & & \\ & I_d & & \\ & & \ddots & \\ & & & I_d \end{bmatrix}, \quad C = \begin{bmatrix} I_n & & & \\ & I_n & & \\ & & \ddots & \\ & & & I_n \end{bmatrix}.$$

By the structure of $\mathbf{x}$ defined in (23), we have

$$\nabla h_1(\mathbf{x}) = \left[\begin{array}{c|c|c|c} A & -B & \mathbf{O}_{dc\times nc} & \mathbf{O}_{dc\times nc} \\ \hline -C & \mathbf{O}_{nc\times dc} & \mathbf{O}_{nc\times nc} & \mathbf{O}_{nc\times nc} \\ \hline \mathbf{O}_{dc\times nc} & B & \mathbf{O}_{dc\times nc} & \mathbf{O}_{dc\times nc} \\ \hline C & \mathbf{O}_{nc\times dc} & C & -C \\ \hline \mathbf{O}_{nc\times nc} & \mathbf{O}_{nc\times dc} & -C & \mathbf{O}_{nc\times nc} \\ \hline \mathbf{O}_{nc\times nc} & \mathbf{O}_{nc\times dc} & \mathbf{O}_{nc\times nc} & C \end{array} \right] \quad \text{and} \quad \nabla h_2(\mathbf{x}) = \begin{bmatrix} \mathbf{O}_{dc\times m_2} \\ \hline \mathbf{O}_{nc\times m_2} \\ \hline \mathbf{O}_{dc\times m_2} \\ \hline \mathbf{O}_{nc\times m_2} \\ \hline \mathbf{O}_{nc\times m_2} \\ \hline G(\mathbf{x}) \end{bmatrix},$$

where $G(\mathbf{x})$ is given in (31) and $\{\widehat{Y}_i\}_{i=1}^c$ are the cloumn vectors of $\widehat{Y}$.

As $\mathbf{x} \in \Gamma$, we must have that the column vectors $\{\widehat{Y}_i\}_{i=1}^c$ are orthogonal to each other, and then the columns of $G(\mathbf{x})$ are orthogonal to each other. Note that the first $3nc + 2dc$ rows of $\nabla h_2(\mathbf{x})$ constitute a zero matrix. Therefore, it follows from the structure of $\nabla h_1(\mathbf{x})$ and $\nabla h_2(\mathbf{x})$ that $\{\nabla[h_1(\bar{\mathbf{x}})]_i\}_{i=1}^{m_1} \cup \{\nabla[h_2(\bar{\mathbf{x}})]_i\}_{i=1}^{m_2}$ are linearly independent for any $\mathbf{x} \in \Gamma$. The proof is complete. ∎

By Lemmas 8 and 9, we can show that any accumulation point $\mathbf{x}^*$ of the corresponding sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ with respect to $\{T^k\}_{k\in\mathbb{N}}$ is a KKT point of problem (24). As shown in Remark 4, the normal cone $\partial\delta_{\mathbb{S}_1}(T) = N_{\mathbb{S}_1}(T)$ in vector notation is

$$N_\Gamma(\bar{\mathbf{x}}) = \{\nabla h_2(\bar{\mathbf{x}})\upsilon | \upsilon \in \mathbb{R}^{m_2}\} = \{\sum_{i=1}^{m_2}[\upsilon]_i \nabla[h_2(\bar{\mathbf{x}})]_i | \upsilon \in \mathbb{R}^{m_2}\}.$$

According to the well-definedness of (10), in view of the vector notation, we can obtain a solution $\mathbf{x}^k$ such that there exist two vectors $\theta^k \in \partial f(\mathbf{x}^k)$ and υ^k to satisfy

$$\|\theta^k + \sum_{i=1}^{m_1}([\bar{\Lambda}^k]_i + \rho^k[h_1(\mathbf{x}^k)]_i)\nabla[h_1(\mathbf{x}^k)]_i + \sum_{i=1}^{m_2}[\upsilon^k]_i\nabla[h_2(\mathbf{x}^k)]_i\|_\infty \le \epsilon^k$$

for each $k \in \mathbb{N}$. The following result is central to this paper.

Theorem 10 *Let $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ be the iteration sequence generated by Algorithm 1 and $\boldsymbol{x}^*$ be its accumulation point, i.e., there exists a subsequence $\mathcal{K} \subseteq \mathbb{N}$ such that $\lim_{k\in\mathcal{K}} \boldsymbol{x}^k = \boldsymbol{x}^*$. Then $\boldsymbol{x}^*$ is also a KKT point of problem (24).*

Proof We first show that $\mathbf{x}^*$ satisfies the feasibility of problem (24), i.e., $h_1(\mathbf{x}^*) = 0$ and $h_2(\mathbf{x}^*) = 0$. By (9), we conclude that $h_2(\mathbf{x}^k) = 0$ for each $k \in \mathbb{N}$. The continuity of h_2 yields $h_2(\mathbf{x}^*) = 0$, i.e., $\mathbf{x}^* \in \Gamma$. The proof of feasibility $h_1(\mathbf{x}^*) = 0$ is divided into two parts, according to the boundedness of the sequence $\{\rho^k\}_{k\in\mathbb{N}}$.

Part I. Suppose first that the penalty sequence $\{\rho^k\}_{k\in\mathbb{N}}$ is bounded. By the penalty parameter update rule (11), it follows that ρ^k stabilizes after some k_0, which implies that $\|h_1(\mathbf{x}^{k+1})\|_\infty \leq \tau\|h_1(\mathbf{x}^k)\|_\infty$ for all $k \geq k_0$ and the constant $\tau \in [0,1)$. By a standard continuity argument, we obtain that $h_1(\mathbf{x}^*) = 0$.

Part II. In the following, we assume that $\{\rho^k\}_{k\in\mathbb{N}}$ is unbounded. For each $k \in \mathcal{K}$, there exist vectors $\{\delta^k\}_{k\in\mathbb{N}}$ with $\|\delta^k\|_\infty \leq \epsilon^k$ and $\epsilon^k \downarrow 0$ such that

$$\theta^k + \sum_{i=1}^{m_1}([\overline{\Lambda}^k]_i + \rho^k[h_1(\mathbf{x}^k)]_i)\nabla[h_1(\mathbf{x}^k)]_i + \sum_{i=1}^{m_2}[\upsilon^k]_i\nabla[h_2(\mathbf{x}^k)]_i = \delta^k \tag{32}$$

for some $\theta^k \in \partial f(\mathbf{x}^k)$. Dividing both sides of (32) by ρ^k, we obtain that

$$\sum_{i=1}^{m_1}([\frac{\overline{\Lambda}^k}{\rho^k}]_i + [h_1(\mathbf{x}^k)]_i)\nabla[h_1(\mathbf{x}^k)]_i + \sum_{i=1}^{m_2}[\hat{\upsilon}^k]_i\nabla[h_2(\mathbf{x}^k)]_i = \frac{\delta^k - \theta^k}{\rho^k}, \tag{33}$$

where $\hat{\upsilon}^k = \frac{\upsilon^k}{\rho^k}$. Define

$$H(\mathbf{x})^\intercal := [\nabla h_1(\mathbf{x}) \; \nabla h_2(\mathbf{x})]$$

and

$$\eta^k := ([\frac{\overline{\Lambda}^k}{\rho^k}]_1 + [h_1(\mathbf{x}^k)]_1, \cdots, [\frac{\overline{\Lambda}^k}{\rho^k}]_{m_1} + [h_1(\mathbf{x}^k)]_{m_1}, [\hat{\upsilon}^k]_1, \cdots, [\hat{\upsilon}^k]_{m_2})^\intercal.$$

Hence we can rewrite (33) in the following way:

$$H(\mathbf{x})^\intercal\eta^k = \frac{\delta^k - \theta^k}{\rho^k}.$$

A straightforward application of Lemma 9 yields that $\{\nabla[h_1(\mathbf{x}^*)]_i\}_{i=1}^{m_1} \cup \{\nabla[h_2(\mathbf{x}^*)]_i\}_{i=1}^{m_2}$ are independent as $\mathbf{x}^* \in \Gamma$. In addition, we notice that the gradient vectors $\nabla h_1, \nabla h_2$ are continuous and $h_2(\mathbf{x}^k) = 0$ for all $k \in \mathcal{K}$. This means that $H(\mathbf{x}^k) \to H(\mathbf{x}^*)$ and $H(\mathbf{x}^*)$ has full rank as $\mathbf{x}^* \in \Gamma$. Therefore, we have that $H(\mathbf{x}^k)H(\mathbf{x}^k)^\intercal \to H(\mathbf{x}^*)H(\mathbf{x}^*)^\intercal \succ 0$. By the fact that eigenvalues of a symmetric matrix vary continuously with its matrix values. We then conclude that $H(\mathbf{x}^k)H(\mathbf{x}^k)^\intercal$ is nonsingular for sufficiently large $k \in \mathcal{K}$, which yields

$$\eta^k = [H(\mathbf{x}^k)H(\mathbf{x}^k)^\intercal]^{-1}H(\mathbf{x}^k)\frac{\delta^k - \theta^k}{\rho^k}.$$

Since f is a convex function, the set $\cup_{x\in\mathcal{X}}\partial f(\mathbf{x})$ is bounded whenever $\mathcal{X}$ is bounded. A nice proof of this result can be found in Bertsekas (1999, Proposition B.24(b)). It is then straightforward to see that $\{\theta^k\}_{k\in\mathcal{K}}$ is bounded by setting $\mathcal{X} = \{\mathbf{x}^k\}_{k\in\mathcal{K}}$, where the

boundedness of $\{\mathbf{x}^k\}_{k \in \mathcal{K}}$ is motivated by Proposition 8. Combining the previous result $\|\delta^k\|_\infty \leq \epsilon^k \downarrow 0$, we obtain for $k \in \mathcal{K}$

$$\eta^k \to 0 \quad as \; k \to \infty.$$

Finally, with the boundedness of Lagrange multipliers $\{\overline{\lambda}^k\}_k$, $[h_1(\mathbf{x}^*)]_i = 0 = [\hat{v}]_j$ is guaranteed for all i, j. Hence we conclude that $h_1(\mathbf{x}^*) = 0$.

Next, we show that $\mathbf{x}^*$ is a KKT point. The boundedness of $\{\theta^k\}_{k \in \mathcal{K}}$ implies that there exists a subsequence $\mathcal{K}_1 \subseteq \mathcal{K}$ such that $\lim_{k \in \mathcal{K}_1} \theta^k = \theta^*$. Together with $\lim_{k \in \mathcal{K}_1} \mathbf{x}^k = \mathbf{x}^*$ and $\theta^k \in \partial f(\mathbf{x}^k)$, it can be follows from the closedness property of subdifferential that

$$\theta^* \in \partial f(\mathbf{x}^*).$$

By the fact that $[\lambda^{k+1}]_i = [\overline{\Lambda}^k]_i + \rho^k[h_1(\mathbf{x}^k)]_i$ for all i, we have that for $k \in \mathcal{K}_1$

$$\theta^k + \sum_{i=1}^{m_1}[\lambda^{k+1}]_i \nabla[h_1(\mathbf{x}^k)]_i + \sum_{i=1}^{m_2}[v^k]_i \nabla[h_2(\mathbf{x}^k)]_i = \delta^k \tag{34}$$

for some vector δ^k with $\|\delta^k\|_\infty \leq \epsilon^k \downarrow 0$ and $\theta^k \in \partial f(\mathbf{x}^k)$. Define

$$\pi^k := ([\lambda^{k+1}]_1, \cdots, [\lambda^{k+1}]_{m_1}, [v^k]_1, \cdots, [v^k]_{m_2})^\mathsf{T}. \tag{35}$$

We then deduce from (34)

$$H(\mathbf{x}^k)^\mathsf{T}\pi^k = \delta^k - \theta^k.$$

Likewise, the matrix $H(\mathbf{x}^k)H(\mathbf{x}^k)^\mathsf{T}$ is nonsingular for sufficiently large $k \in \mathcal{K}_1$, and

$$\pi^k = [H(\mathbf{x}^k)H(\mathbf{x}^k)^\mathsf{T}]^{-1}H(\mathbf{x}^k)(\delta^k - \theta^k).$$

Taking limitations within $\mathcal{K}_1$ on both sides of the expression above for π^k, we have then

$$\pi^k \to \pi^* = -[H(\mathbf{x}^*)H(\mathbf{x}^*)^\mathsf{T}]^{-1}H(\mathbf{x}^*)\theta^*.$$

Taking limitations for $k \in \mathcal{K}_1$ again on both sides of (34), it follows from (35) that

$$\theta^* + \sum_{i=1}^{m_1}[\Lambda^*]_i \nabla[h_1(\mathbf{x}^*)]_i + \sum_{i=1}^{m_2}[v^*]_i \nabla[h_2(\mathbf{x}^*)]_i = 0,$$

where Λ^* and v^* are guaranteed by π^*. Therefore, $\mathbf{x}^*$ is a KKT point of problem (24). ∎

By Theorem 10 and (23), we can immediately obtain Theorem 7. Our numerical experiments in the next section testify that Algorithm 1 works well and can output KKT point of the original problem (6).

6. Experiment Study

In this section, we conduct numerical experiments to show effectiveness of Algorithm 1 by using MATLAB (2020a) on a laptop of 16G of memory and Inter Core i7 2.3Ghz CPU against several state-of-the-art unsupervised feature selection methods on six real-world datasets, including one speech signal dataset (Isolet*), two face image datasets (ORL*,COIL20*), three microarray datasets (lung*, TOX-171*, 9_Tumors†). Table 1 summarizes the details of these 6 benchmark datasets used in the experiments. In addition to verifying the effectiveness of our method on the above datasets, we also show the stability analysis, robustness analysis and parameter sensitivity analysis on some datasets.

Table 1: Dataset Description

Dataset	Size	# of Features	# of Classes
lung	203	3312	5
TOX-171	171	5748	4
9_Tumors	60	5726	9
Isolet	1560	617	26
ORL	400	1024	40
COIL20	1440	1024	20

Methods to Compare. We compare the performance of Algorithm 1 with the following state-of-the-art unsupervised feature selection methods:

- **Baseline**: All of the original features are adopted.

- **MaxVar** (Krzanowski, 1987): Features corresponding to the maximum variance are selected to obtain the expressive features.

- **LS** (He et al., 2005): Laplacian Score, in which features are selected with the most consistency with Gaussian Laplacian matrix.

- **SPEC** (Zhao and Liu, 2007): According to spectrum of the graph to select features.

- **MCFS** (Cai et al., 2010): Multi-cluster feature selection, it uses the l_1-norm to regularize the feature selection process as a spectral information regression problem.

- **NDFS** (Li et al., 2012):Non-negative discriminative feature selection, which addressed feature discriminability and correlation simultaneously.

- **UDFS** (Yang et al., 2011a):Unsupervised discriminative feature selection incorporated discriminative analysis as well as $l_{2,1}$-norm minimization, which is formalized as a unified framework.

∗. https://jundongl.github.io/scikit-feature/datasets.html

†. https://github.com/primekangkang/Genedata

- **UDPFS** (Wang et al., 2020): Unsupervised discriminative projection for feature selection to select discriminative features by conducting fuzziness learning and sparse learning simultaneously.

Evaluation Measures. Similar to previous work, and basing on the attained clustering results and the ground truth information, we evaluate the performance of the unsupervised feature selection methods by two widely utilized evaluation metrics, i.e., clustering ACCuracy (ACC) and Normalized Mutual Information (NMI) (Yang et al., 2011a). The higher the ACC and NMI are, the better the clustering performance is.

Given one sample $\mathbf{x}_i \in \{\mathbf{x}_i\}_{i=1}^n$, denote y_i be the ground truth label and l_i be the predicted clustering label. The ACC is defined as

$$\text{ACC} = \frac{1}{n} \sum_{i=1}^n \delta(y_i, map(l_i)),$$

where $\delta(a,b) = 1$ if $a = b$; otherwise $\delta(a,b) = 0$, and $map(l_i)$ is the permutation mapping function that maps each cluster label l_i to the equivalent label from the data set.

Given two random variables P and Q, P denotes the true labels and Q represents clustering results. The NMI of P and Q is defined as:

$$\text{NMI}(P, Q) = \frac{I(P; Q)}{\sqrt{H(P)H(Q)}},$$

where $I(P; Q)$ is the mutual information between P and Q, $H(P)$ and $H(Q)$ are the entropies of P and Q, respectively.

Experiment Setting. In our experiments, the parameters of Algorithm 1 are set as follows:

$$\tau = 0.99, \quad r = 1.01, \quad \rho^1 = c/2, \quad \overline{\lambda}_1^1 = \overline{\lambda}_3^1 = \overline{\lambda}_4^1 = \mathbf{O}_{n \times c}, \quad \overline{\lambda}_2^1 = \mathbf{O}_{d \times c},$$

and

$$\overline{\lambda}_{N,min} = -100\mathbf{E}, \quad \overline{\lambda}_{N,max} = 100\mathbf{E} \quad (N = 1,2,3,4), \quad \epsilon^k = 0.995^k \quad (k \in \mathbb{N}).$$

The parameters in Algorithm 2 are set as $\underline{C} = C_i^{k,j} = \overline{C} = 0.5$. The iteration is terminated if the iteration number exceeds 20.

In the compared methods, there are some hyper-parameters to be set in advance. We fix number of neighboring parameter $k = 5$ for LS, SPEC, MCFS, UDFS, NDFS, and our proposed method. In order to make fair comparison of different unsupervised feature selection methods, we tuned the parameters for all methods by a grid-search strategy from $\{10^{-6}, 10^{-5}, 10^{-4}, \cdots, 10^4, 10^5, 10^6\}$, and the best clustering results from the optimal parameters are reported for all the algorithms. Because the optimal number of selected features is unknown, we set different number of selected features for all datasets, the selected feature number was tuned from $\{50, 100, 150, 200, 250, 300\}$. After completing the feature selection process, we use K-means algorithm to cluster the data into c groups. Since the initial center points have great impact on the performance of K-means algorithm, we conduct

K-means algorithm 20 times repeatedly with random initialization to report the mean and standard deviation values of ACC and NMI.

In the next subsections, we will illustrate the algorithmic performance, stability, robustness and parameter sensitivity, respectively.

6.1 Algorithmic Performance

The experiments results of different methods on the datasets are summarized in Tables 2 and 3. The best results are highlighted in bold fonts.

In view of the averaging of all numerical results, it can be seen that the performance of our method is superior to other state-of-the-art methods. Its good performance is mainly attributed to the following aspects: Firstly, we adopt the technology similar to NDFS to establish the model, i.e., learning the pseudo class label indicators and the feature selection matrix simultaneously. However, the difference is that we use $l_{2,1}$-norm to characterize the linear loss function between features and pseudo labels and also take into account the prevention of overfitting. Secondly, different from the commonly used processing methods, we apply a convergent algorithm that can simultaneously optimize all variables in the feature selection model. In the previous section, we have proven the convergence property of our algorithm. Since the iterative sequence of our algorithm converges to KKT points, it achieves better results than other methods.

Table 2: Clustering results (ACC±STD%) of different feature selection algorithms on six real-world datasets. The best results are highlighted in bold.

Dataset	All features	LS	Maxvar	MCFS	NDFS	SPEC	UDFS	UDPFS	Ours
lung	65.0±3.6	74.9±0.2	68.0±9.4	77.6±11.0	63.3±6.9	64.1±7.9	72.3±10.9	69.6±7.7	**82.4±7.9**
ORL	49.7±3.2	49.9±2.4	50.8±1.4	**55.7±3.7**	50.5±3.0	51.4± 2.2	53.3±4.1	53.1±3.8	52.9±3.4
Isolet	60.9±2.1	58.7±1.5	56.9±2.3	64.5±4.3	61.6±4.4	56.5±3.0	57.8± 3.1	58.3±2.9	**65.8±3.9**
COIL20	62.7±3.1	62.2±1.9	61.4±1.6	63.0±3.7	58.7± 4.1	**65.5±3.8**	60.2±4.2	58.2 ±4.6	61.6±3.8
TOX-171	42.8±2.1	43.1±1.4	42.9±1.6	42.9±1.6	43.4±3.3	40.4±0.0	48.2±2.1	**54.0± 3.2**	49.2±4.1
9_Tumors	40.8±3.7	42.3±2.6	41.2±2.6	42.4±3.6	44.0±3.7	35.8±2.4	43.0± 4.3	**44.2±4.3**	44.1±4.1
Mean	53.7±3.0	55.2±1.7	53.5±3.2	57.7±4.7	53.6±4.2	52.3±3.2	55.8±4.8	56.2±4.4	**59.3±4.5**

Table 3: Clustering results (NMI±STD%) of different feature selection algorithms on six real-world datasets. The best results are highlighted in bold.

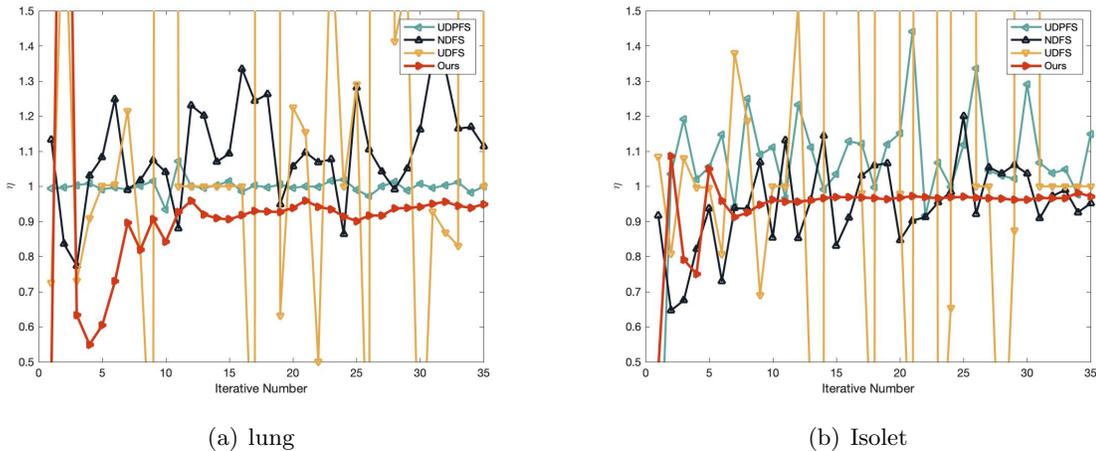| Dataset	All features	LS	Maxvar	MCFS	NDFS	SPEC	UDFS	UDPFS	Ours
lung	51.6±1.9	53.1± 0.5	57.8± 3.9	67.5±7.0	53.0±3.5	52.5± 5.6	61.3±5.8	59.0±4.0	**69.0±4.4**
ORL	70.0±1.7	71.1± 1.3	70.7± 2.1	**76.8±1.8**	73.2±1.9	71.4± 1.3	74.7±1.6	74.8±1.6	74.9±1.7
Isolet	75.7±0.8	73.2±0.9	74.8±1.3	77.7±1.7	77.1± 2.2	72.4±1.1	74.7±1.8	74.4±1.3	**80.5±1.3**
COIL20	**77.1±1.3**	72.5±1.1	71.9±0.7	76.5±1.7	74.0± 1.6	75.3±1.6	75.4±1.3	73.9±2.0	76.3±2.3
TOX-171	13.6±2.3	12.5±1.7	11.4±3.2	12.7±0.4	16.4 ±5.9	9.7 ± 0.0	22.8±3.5	**29.9±1.2**	25.3±4.4
9_Tumors	39.5±3.1	41.0±2.3	40.2±2.5	41.1±2.7	44.7±4.5	34.5±2.4	44.1±4.3	**46.7±3.6**	44.8±3.2
Mean	54.6±1.9	53.9±1.3	54.5±2.3	58.7±2.6	56.4±3.3	52.6±2	58.8±3.1	59.8±2.3	**61.8±2.9**

Figure 1: Stability curves over lung and Isolet.

6.2 Stability Analysis

Now we will illustrate that our algorithm is more stable than other iterative algorithms including: UDPFS (Wang et al., 2020), NDFS (Li et al., 2012) and UDFS (Yang et al., 2011a). Following the symbol in Li et al. (2012); Yang et al. (2011a); Wang et al. (2020), we denote the feature selection matrix as W in these methods and define

$$\eta = \frac{\|W_{k+1} - W_k\|_F}{\|W_k - W_{k-1}\|_F},$$

where W_k is the k-th iterative point. To demonstrate fully that our algorithm is more stable, we randomly initialize cluster indicator matrix Y and W 20 times. Under the parameter setting of the optimal results obtained by corresponding method, we record the average results of η. The experimental results are shown in Fig. 1.

It can be seen that the value of η of the other three methods are always changing irregularly , while ours starts to stabilize after fewer iterations and then always less than 1. Furthermore, we know that , with the increase of iterative number k, $\|W_{k+1} - W_k\|_F$ decreases gradually in our method, which shows that our iterative sequence $\{W_k\}_{k\in\mathbb{N}}$ keeps the "distance" of the adjacent two points gradually reduced and it is changed regularly according to the iterative rules. Following the previous theoretical proof, iterative sequence $\{W_k\}_{k\in\mathbb{N}}$ will eventually converge to the KKT points. Compared with our method, since the values of η of UDPFS, NDFS and UDFS are ruleless, iterative sequence $\{W_k\}_{k\in\mathbb{N}}$ is "jumping" irregularly and does not have a convergence trend. Therefore, our method is more stable.
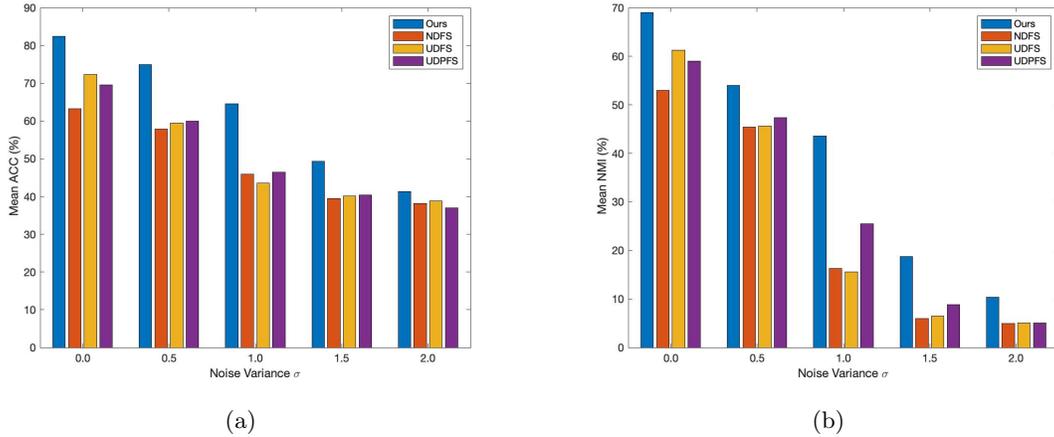
6.3 Robustness Analysis

Figure 2: Robustness comparison to data perturbation between our method and other iterative methods on lung.
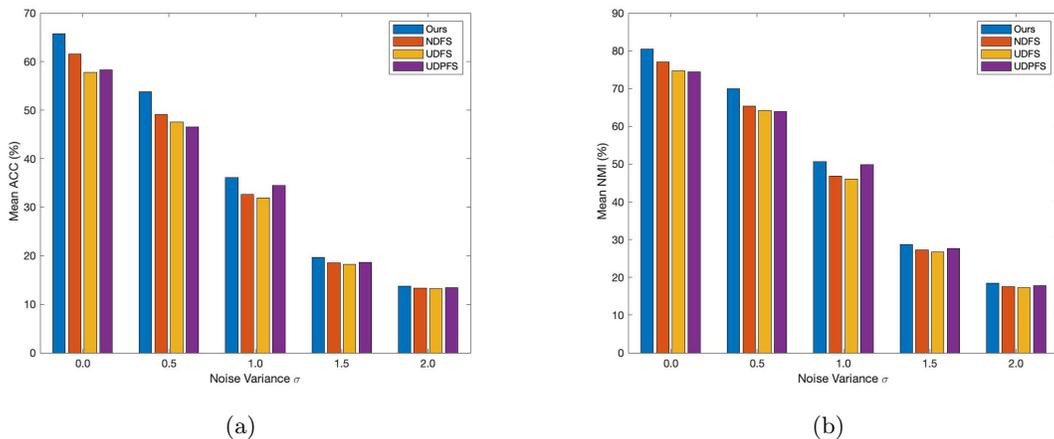

Figure 3: Robustness comparison to data perturbation between our method and other iterative methods on Isolet.

In this subsection, we summarize the main results for our robustness analysis. We consider the effect of varying the amount of perturbation introduced in the datasets, i.e., the effect of performance if we fine-tune the Gaussian noise from the distribution $\mathcal{N}(\mathbf{0}, \sigma^2)$ where σ is sampled from the set $\{0.0, 0.5, 1.0, 1.5, 2.0\}$ and add the Gaussian noise to the input data. In order to make a fair comparison, we conduct the experiments under the parameter setting of the optimal results obtained by each method for the chosen dataset. Meanwhile, in order to avoid the influence of the randomness of noise in a single experiment, we uniformly do ten experiments for each noise variance, and then average the results as the final result.
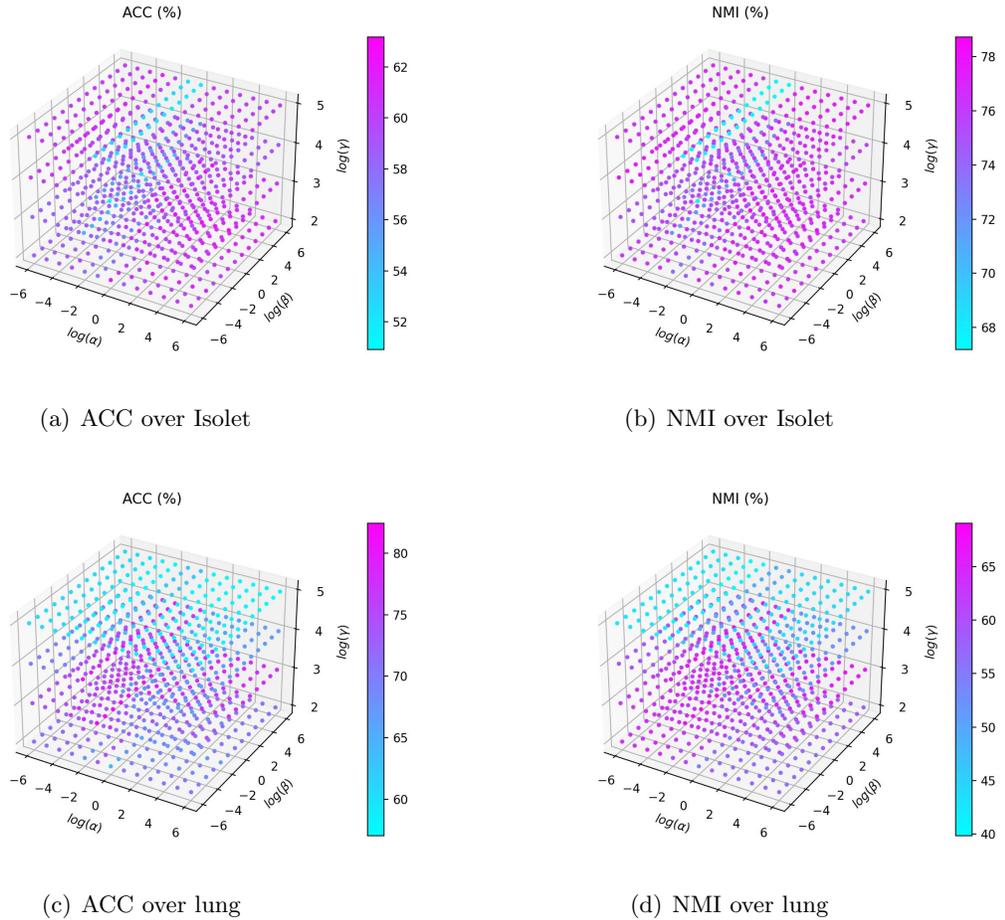
(a) ACC over Isolet

(b) NMI over Isolet

(c) ACC over lung

(d) NMI over lung

Figure 4: Performance with different α, β, γ values on Isolet and lung with a grid search strategy.

Fig. 2 and Fig. 3 show the robustness of the iterative methods here considered on the lung dataset and Isolet dataset with different levels of noise. Note that with the increase of disturbance, the robustness of all iterative methods falls off, while the performance of our method is always the best. Therefore, compared with other methods, our method has a strong robustness.

6.4 Parameter Sensitivity Analysis

Like many other feature selection algorithms, our proposed method also requires several parameters α, β, γ to be set in advance. Next, we will discuss their sensitivity. In our experiments, we observe that the parameters α and β have more effect on the performance than the parameter γ on the given datasets. Therefore, we focus on discussing the parameters α and β. We will conduct the parameter sensitivity study in terms of α, β when γ is fixed

to some values. α and β are tuned from $\{10^{-6}, 10^{-5}, \cdots, 10^5, 10^6\}$. The results on lung and Isolet are presented in Fig. 4. It can be seen that our method is not sensitive to α, β and γ with relatively wide ranges.

7. conclusion

In this paper, we firstly have explored an ideal feature selection model: $l_{2,1}$-norm regularized regression optimization problem with non-negative orthogonal constraint, which well captures the most representative features from the original high-dimensional data. Then, we propose an inexact augmented Lagrangian multiplier method to solve our feature selection model. Moreover, a proximal alternating minimization method is utilized to solve the augmented Lagrangian subproblem with the benefit being that each subproblem has a closed form solution. It is shown that our algorithm has the subsequence convergence property, which is not provided in the state-of-the-art unsupervised feature selection methods. Quantitative and qualitative experimental results have shown the effectiveness of our proposed method.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Roberto Andreani, Ernesto G Birgin, José Mario Martínez, and María Laura Schuverdt. On augmented lagrangian methods with general lower-level constraints. *SIAM Journal on Optimization*, 18(4):1286–1309, 2008.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

DP Bertsekas. Nonlinear programming, 2nd edn (belmont, ma: Athena scientific). 1999.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1): 459–494, 2014.

Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12):1624–1637, 2005.

Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 333–342, 2010.

David Charte, Francisco Charte, and Francisco Herrera. Reducing data complexity using autoencoders with class-informed loss functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Hong Chen, Feiping Nie, Rong Wang, and Xuelong Li. Fast unsupervised feature selection with bipartite graph and $l_{2,0}$-norm constraint. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Zheng Chen, Meng Pang, Zixin Zhao, Shuainan Li, Rui Miao, Yifan Zhang, Xiaoyue Feng, Xin Feng, Yexian Zhang, Meiyu Duan, et al. Feature selection may improve deep neural networks for the bioinformatics problems. *Bioinformatics*, 36(5):1542–1552, 2020.

Manoranjan Dash, Hua Liu, and Jun Yao. Dimensionality reduction of unsupervised data. In *Proceedings ninth ieee international conference on tools with artificial intelligence*, pages 532–539. IEEE, 1997.

Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135, 2006.

Richard O Duda and Peter E Hart. Stork. dg, pattern classification. *John Willey and Sons, New York*, 2001.

Jie Gui, Dacheng Tao, Zhenan Sun, Yong Luo, Xinge You, and Yuan Yan Tang. Group sparse multiview patch alignment framework with view consistency for image classification. *IEEE transactions on image processing*, 23(7):3126–3137, 2014.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. *Advances in Neural Information Processing Systems*, 18, 2005.

Chenping Hou, Feiping Nie, Xuelong Li, Dongyun Yi, and Yi Wu. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE Transactions on Cybernetics*, 44(6):793–804, 2013.

Josef Kittler. Feature selection and extraction. *Handbook of Pattern Recognition and Image Processing*, 1986.

A Klaser, C Schmid, and CL Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition*, volume 42, pages 3169–3176, 2011.

Wojtek J Krzanowski. Selection of variables to preserve multivariate data structure, using principal components. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(1):22–33, 1987.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Jundong Li, Jiliang Tang, and Huan Liu. Reconstruction-based unsupervised feature selection: An embedded approach. In *IJCAI*, pages 2159–2165, 2017.

Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. Unsupervised feature selection using nonnegative spectral analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 2012.

Zhengxin Li, Feiping Nie, Jintang Bian, Danyang Wu, and Xuelong Li. Sparse pca via $l_{2,p}$-norm regularization for unsupervised feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. doi: 10.1109/TPAMI.2021.3121329.

Chunfeng Lian, Mingxia Liu, Jun Zhang, and Dinggang Shen. Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):880–893, 2018.

Huan Liu, Hiroshi Motoda, and Lei Yu. A selective sampling approach to active feature selection. *Artificial Intelligence*, 159(1-2):49–74, 2004.

Mahdokht Masaeli, Yan Yan, Ying Cui, Glenn Fung, and Jennifer G Dy. Convex principal feature selection. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 619–628. SIAM, 2010.

Luis Carlos Molina, Lluís Belanche, and Àngela Nebot. Feature selection algorithms: A survey and experimental evaluation. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 306–313. IEEE, 2002.

Feiping Nie, Shiming Xiang, Yangqing Jia, Changshui Zhang, and Shuicheng Yan. Trace ratio criterion for feature selection. In *AAAI*, volume 2, pages 671–676, 2008.

Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. *Advances in neural information processing systems*, 23, 2010.

Feiping Nie, Wei Zhu, and Xuelong Li. Unsupervised feature selection with structured graph optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Feiping Nie, Wei Zhu, and Xuelong Li. Structured graph optimization for unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 33(3):1210–1222, 2019.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

Mingjie Qian and Chengxiang Zhai. Robust unsupervised feature selection. In *Twenty-third international joint conference on artificial intelligence*, 2013.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Giorgio Roffo, Simone Melzi, Umberto Castellani, Alessandro Vinciarelli, and Marco Cristani. Infinite feature selection: a graph-based feature filtering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4396–4410, 2020.

Hao-Jun Michael Shi, Shenyinying Tu, Yangyang Xu, and Wotao Yin. A primer on coordinate descent algorithms. *arXiv preprint arXiv:1610.00040*, 2016.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007.

Rong Wang, Jintang Bian, Feiping Nie, and Xuelong Li. Unsupervised discriminative projection for feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

Suhang Wang, Jiliang Tang, and Huan Liu. Embedded unsupervised feature selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011a.

Yi Yang, Heng Tao Shen, Feiping Nie, Rongrong Ji, and Xiaofang Zhou. Nonnegative spectral clustering with discriminative regularization. In *Twenty-fifth AAAI Conference on Artificial Intelligence*, 2011b.

Jiho Yoo and Seungjin Choi. Orthogonal nonnegative matrix factorization: Multiplicative updates on stiefel manifolds. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 140–147. Springer, 2008.

Kui Yu, Lin Liu, Jiuyong Li, Wei Ding, and Thuc Duy Le. Multi-source causal feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2240– 2256, 2019.

Kai Zhang, Sheng Zhang, Jun Liu, Jun Wang, and Jie Zhang. Greedy orthogonal pivoting algorithm for non-negative matrix factorization. In *International Conference on Machine Learning*, pages 7493–7501. PMLR, 2019.

Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1151–1157, 2007.

Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Qinghua Hu, and Simon CK Shiu. Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446, 2015.

Pengfei Zhu, Qinghua Hu, Changqing Zhang, and Wangmeng Zuo. Coupled dictionary learning for unsupervised feature selection. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.