

Data Clustering via Uncorrelated Ridge Regression

Rui Zhang^{ID}, Member, IEEE, Xuelong Li^{ID}, Fellow, IEEE, Tong Wu^{ID}, and Yi Zhao^{ID}, Student Member, IEEE

Abstract—Ridge regression is frequently utilized by both supervised and semisupervised learnings. However, the trivial solution might occur, when ridge regression is directly applied for clustering. To address this issue, an uncorrelated constraint is introduced to the ridge regression with embedding the manifold structure. In particular, we choose uncorrelated constraint over orthogonal constraint, since the closed-form solution can be obtained correspondingly. In addition to the proposed uncorrelated ridge regression, a soft pseudo label is utilized with ℓ_1 ball constraint for clustering. Moreover, a brand new strategy, i.e., a rescaled technique, is proposed such that optimal scaling within the uncorrelated constraint can be achieved automatically to avoid the inconvenience of tuning it manually. Equipped with the rescaled uncorrelated ridge regression with the soft label, a novel clustering method can be developed based on solving the related clustering model. Consequently, extensive experiments are provided to illustrate the effectiveness of the proposed method.

Index Terms—Clustering, rescaling, ridge regression, uncorrelated constraint.

I. INTRODUCTION

Speedy development of industry leads to an increasing amount of data to deal with. Accordingly, how to efficiently examine these data serves as the mainstream for a spectrum of real-world applications. Particularly, a large number of real-world data usually are generated without the label information. In other words, it is either laborious or expensive to obtain the label information in reality. Therefore, the issue concerning clustering the given data points into certain clusters performs significantly for the data investigation in multiple fields.

Recently, a lot of works regarding clustering approaches [1]–[6] have been proposed. Admittedly, these works mentioned above do contribute to some extent. Some of them may lead to satisfactory clustering performance under specially designated stipulation. However, most of the current unsupervised learning techniques are associated with k -means [7] to some extent by either utilizing it as a metric or incorporating it into the learning process. Due to the efficiency and simplicity, the k -means type clustering methods have been widely investigated. To visualize the cluster structures of data, self-organizing map approaches [8], [9] have been researched. In addition, clustering algorithms such as spectral clustering [10]–[13], support vector clustering [14], and kernel-based clustering [15], [16] have been proposed to examine nonlinear cluster structures.

Manuscript received March 3, 2019; revised July 20, 2019 and November 14, 2019; accepted February 23, 2020. Date of publication April 7, 2020; date of current version January 5, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61871470 and Grant U1801262, in part by the China Postdoctoral Science Foundation under Grant 2018M643765 and Grant 2019T120960, and in part by the Xi'an Postdoctoral Innovation Base Funding. (*Corresponding author: Xuelong Li*)

Rui Zhang and Xuelong Li are with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: ruizhang8633@gmail.com; xuelong_li@nwpu.edu.cn).

Tong Wu is with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: wutong034@stu.xjtu.edu.cn).

Yi Zhao is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: zhaoyi16@mails.tsinghua.edu.cn).

Color versions of one or more of the figures in this article are available online at <https://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNNLS.2020.2978755

More specifically, k -means is frequently exploited to serve as a metric for evaluating the performance of unsupervised feature selection [17]–[22] methods. Another example to illustrate here is fuzzy k -means [23], [24]. The distinction between k -means and fuzzy k -means reflects on the membership. In other words, k -means is a hard clustering method, while fuzzy k -means [25], [26] is a soft one, via which each data point is assigned with degrees of membership in all clusters. To sum up, k -means is fundamentally significant for being a basic clustering method.

By sharing the similar least square form with k -means, ridge regression [27]–[29] on the other hand is hardly applied to direct clustering due to the trivial solution it will trigger. In other words, all the data points are grouped to the same cluster, while the subspace is a null space. According to its efficiency under both semisupervised and supervised learnings, it is unfortunate that ridge regression cannot be directly investigated from the perspective of unsupervised learning. To address this issue, two common criteria can be adopted to prevent the potential trivial solution: 1) constraining the hard binary label matrix to be full rank and 2) constraining the subspace to avoid the potential zero space.

Motivation and Contribution: To apply ridge regression for clustering successfully and effectively, the previously mentioned options are taken into account. There are two common constraints with manifold structure, which are known as orthogonal and uncorrelated constraints, respectively. The uncorrelated constraint [30] is often utilized to ensure the data on the subspace to be statistically uncorrelated. In this brief, we choose an uncorrelated constraint over an orthogonal constraint since ridge regression under uncorrelated constraint has the closed-form solution, while ridge regression under orthogonal constraint does not. In sum, the contributions of this brief are listed as the following items.

- 1) The manifold structure of uncorrelated constraint is incorporated into the subspace to avoid the potential trivial solution to the ridge regression clustering model.
- 2) Soft label is utilized with ℓ_1 ball constraint in the uncorrelated ridge regression for the full rank label matrix.
- 3) A novel rescaled strategy is proposed for the uncorrelated ridge regression with soft label (URR-SL), such that optimal scaling in the uncorrelated constraint can be obtained automatically instead of conventional tuning technique.

Consequently, the modified ridge regression can be successfully applied to the clustering via obtaining the closed form of the associated subspace with manifold structure.

Notation: Suppose $\mathbf{W} = [\mathbf{w}_{ij}]_{i \times j} \in \mathbb{R}^{d \times c}$, Frobenius-norm of \mathbf{W} is defined as $\|\mathbf{W}\|_F = (\text{Tr}(\mathbf{W}^T \mathbf{W}))^{1/2} = (\sum_{i=1}^d \sum_{j=1}^c \mathbf{w}_{ij}^2)^{1/2}$. $\mathbf{1}_i$ and $\mathbf{0}_{ij}$ denote $\mathbf{1}_i = [1, \dots, 1]^T \in \mathbb{R}^{i \times 1}$ and $\mathbf{0}_{ij} = [0]_{i \times j} = [\mathbf{0}_{i1}, \dots, \mathbf{0}_{i1}] \in \mathbb{R}^{i \times j}$, respectively. \mathbf{I} denotes the identity matrix. The Karush–Kuhn–Tucker (KKT) conditions (also known as the Kuhn–Tucker conditions) are first-order necessary conditions for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied. For any given real function $\mathbf{h}(x)$, nonlinear operator $(\cdot)_+$ is defined as $(\mathbf{h}(x))_+ = \max(\mathbf{h}(x), 0)$. For an arbitrary matrix $\mathbf{K} \in \mathbb{R}^{p \times q}$, $\mathbf{K} \geq \mathbf{0}_{ij}$ denotes that \mathbf{K} element-wisely satisfies the condition, i.e., any element of \mathbf{K} satisfies $\mathbf{K}_{ij} \geq 0 \forall i, j$.

II. POTENTIAL PROBLEM WHEN RIDGE REGRESSION IS DIRECTLY APPLIED FOR CLUSTERING

Given the data set $\mathcal{X} = \{\mathbf{x}_i | \mathbf{x}_i \in \mathbb{R}^{d \times 1}, i = 1, 2, \dots, n\}$ and related data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, data points are distributed into c different clusters with dimension d and data number n ($d \geq c$).

As for the supervised learning, i.e., ground truth label \mathbf{L} is pre-given for the associated input data \mathbf{X} , ridge regression can be represented as

$$\min_{\mathbf{W}, \mathbf{e}} \|\mathbf{X}^T \mathbf{W} + \mathbf{1}_n \mathbf{e}^T - \mathbf{L}\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{d \times c}$ is the subspace, $\mathbf{e} \in \mathbb{R}^{c \times 1}$ is the bias, and $\mathbf{L} \in \{0, 1\}^{n \times c}$ is the binary label matrix with the regularization parameter $\lambda \in \mathbb{R}$. The purpose of the supervised problem in (1) is to find the optimal subspace \mathbf{W} , which projects the data onto the same space of the label matrix \mathbf{L} , such that the least square fitting error is obtained.

According to the supervised ridge regression in (1), when direct clustering is applied, unsupervised ridge regression could be represented as

$$\begin{aligned} & \min_{\mathbf{W}, \mathbf{e}, \mathbf{Y}} \|\mathbf{X}^T \mathbf{W} + \mathbf{1}_n \mathbf{e}^T - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \geq 0 \\ & \text{s.t. } \mathbf{Y} \in \{0, 1\}^{n \times c}, \mathbf{Y} \mathbf{1}_c = \mathbf{1}_n \end{aligned} \quad (2)$$

where the binary pseudo label matrix $\mathbf{Y} \in \{0, 1\}^{n \times c}$ serves as an optimization variable with satisfying $\mathbf{Y} \mathbf{1}_c = \mathbf{1}_n$, namely, each row of \mathbf{Y} has only one 1, where the rest of the elements in each row are 0. In particular, cluster number c has to be known in advance for initializing both \mathbf{W} and \mathbf{Y} . As for the clustering model in (2), the special case regarding $\mathbf{W} = \mathbf{0}_{dc}$, $\mathbf{e} = [1, 0, \dots, 0]^T$, and $\mathbf{Y} = [\mathbf{1}_n, \mathbf{0}_{n1}, \dots, \mathbf{0}_{n1}]$ serves as the solution to problem (2) by reaching its lower bound 0 as

$$\begin{aligned} & \|\mathbf{X}^T \mathbf{W} + \mathbf{1}_n \mathbf{e}^T - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \\ &= \|\mathbf{X}^T \mathbf{0}_{dc} + \mathbf{1}_n [1, 0, \dots, 0] - [\mathbf{1}_n, \mathbf{0}_{n1}, \dots, \mathbf{0}_{n1}]\|_F^2 + \lambda 0 \\ &= 0. \end{aligned} \quad (3)$$

Obviously, it is irrational that all the data points are grouped to the same cluster based on this pseudo label $\mathbf{Y} = [\mathbf{1}_n, \mathbf{0}_{n1}, \dots, \mathbf{0}_{n1}]$. Therefore, unsupervised ridge regression (2) leads to the trivial solution, i.e., projection matrix \mathbf{W} is null/zero matrix and the rank of label matrix \mathbf{Y} is 1 due to the unconstrained variables. In other words, ridge regression cannot be directly applied for clustering.

III. UNCORRELATED RIDGE REGRESSION WITH SOFT LABEL

To avoid the potential trivial solution previously mentioned, i.e., $\mathbf{W} = \mathbf{0}_{dc}$, it is natural to constrain subspace \mathbf{W} to be full rank. Accordingly, the discussion leads to two general options: orthogonal constraint [31] $\mathbf{W}^T \mathbf{W} = (1/\alpha^2)\mathbf{I}$ and uncorrelated constraint [30] $\mathbf{W}^T \mathbf{S}_t \mathbf{W} = (1/\alpha^2)\mathbf{I}$, where $\mathbf{S}_t = \mathbf{X} \mathbf{H} \mathbf{X}^T$ is the total scatter matrix and α is the scaling term with the centering matrix $\mathbf{H} = \mathbf{I} - (1/n)\mathbf{1}_n \mathbf{1}_n^T \in \mathbb{R}^{n \times n}$. It is worth noting that both of the constraints perform with manifold structure, such that the structure of data remains unchanged during the projection. In this brief, we choose uncorrelated constraint over orthogonal constraint due to the simple fact that orthogonal constrained ridge regression or known as orthogonal least square regression (since regularization $\|\mathbf{W}\|_F^2$ vanishes) does not have closed-form solution with respect to \mathbf{W} . In addition, The conventional uncorrelated constraint is usually utilized to find the most uncorrelated data in the subspace, such that the projected dimensions are orthonormal to ensure that the data on the subspace are uncorrelated to each other. More specifically, the uncorrelated data structure can be further

explored on the subspace. Nevertheless, when the number of samples is less than features [32], the total scatter matrix \mathbf{S}_t is positive semidefinite. The required inverse operation of the scatter matrix is unavailable, such that model under uncorrelated constraint might lead to potential trivial solutions. Therefore, in this brief, the total scatter matrix is modified as $\mathbf{S}_t = \mathbf{X} \mathbf{H} \mathbf{X}^T + \lambda \mathbf{I}$ to ensure a positive definite \mathbf{S}_t . Moreover, we utilize soft label with ℓ_1 ball constraint instead of hard binary label, since soft label is more representative and practical for real-world applications.

Accordingly, the URR-SL problem can be proposed for clustering as

$$\begin{aligned} & \min_{\mathbf{W}, \mathbf{e}, \mathbf{Y}} \|\mathbf{X}^T \mathbf{W} + \mathbf{1}_n \mathbf{e}^T - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \\ & \text{s.t. } \mathbf{W}^T \mathbf{S}_t \mathbf{W} = \frac{1}{\alpha^2} \mathbf{I}, \quad \mathbf{Y} \mathbf{1}_c = \mathbf{1}_n, \quad \mathbf{Y} \geq \mathbf{0}_{nc} \end{aligned} \quad (4)$$

where the total scatter matrix $\mathbf{S}_t = \mathbf{X} \mathbf{H} \mathbf{X}^T + \lambda \mathbf{I} \in \mathbb{R}^{d \times d}$.

IV. RESCALED URR-SL FOR THE OPTIMAL SCALING

Although URR-SL in (4) can prevent the trivial case of clustering as represented in (2), an existing deficiency of the proposed URR-SL in (4) is that scaling α in the uncorrelated constraint can only be selected via conventional tuning technique. Note that the uncorrelated constraint $\mathbf{W}^T \mathbf{S}_t \mathbf{W} = (1/\alpha^2)\mathbf{I}$ in (4) can be rewritten as $((1/\alpha)\mathbf{Z}^T)\mathbf{S}_t((1/\alpha)\mathbf{Z}) = (1/\alpha^2)\mathbf{I} \Rightarrow \mathbf{Z}^T \mathbf{S}_t \mathbf{Z} = \mathbf{I}$ under $\mathbf{Z} = \alpha \mathbf{W}$, then URR-SL in (4) can be rescaled into the following equivalent counterpart:

$$\begin{aligned} & \min_{\mathbf{Z}, \mathbf{e}, \mathbf{Y}} \left\| \mathbf{X}^T \left(\frac{1}{\alpha} \mathbf{Z} \right) + \mathbf{1}_n \mathbf{e}^T - \mathbf{Y} \right\|_F^2 + \lambda \left\| \frac{1}{\alpha} \mathbf{Z} \right\|_F^2 \\ &= \min_{\mathbf{Z}, \mathbf{b}, \mathbf{Y}} \frac{1}{\alpha^2} \|\mathbf{X}^T \mathbf{Z} + \mathbf{1}_n \mathbf{b}^T - \alpha \mathbf{Y}\|_F^2 + \frac{\lambda}{\alpha^2} \|\mathbf{Z}\|_F^2 \\ &\Rightarrow \min_{\mathbf{Z}, \mathbf{b}, \mathbf{Y}} \|\mathbf{X}^T \mathbf{Z} + \mathbf{1}_n \mathbf{b}^T - \alpha \mathbf{Y}\|_F^2 + \lambda \|\mathbf{Z}\|_F^2 \\ & \text{s.t. } \mathbf{Z}^T \mathbf{S}_t \mathbf{Z} = \mathbf{I}, \quad \mathbf{Y} \mathbf{1}_c = \mathbf{1}_n, \quad \mathbf{Y} \geq \mathbf{0}_{nc} \end{aligned} \quad (5)$$

where the bias $\mathbf{b} = \alpha \mathbf{e}$ is also a free variable as the bias \mathbf{e} in (4) with the subspace $\mathbf{Z} \in \mathbb{R}^{d \times c}$. Based on the abovementioned rescaled dual problem, we further attempt to achieve the optimal scaling automatically by optimizing α as a variable in (5). Accordingly, the rescaled URR-SL (RURR-SL) problem can be eventually formulated as

$$\begin{aligned} & \min_{\mathbf{Z}, \alpha, \mathbf{b}, \mathbf{Y}} \|\mathbf{X}^T \mathbf{Z} + \mathbf{1}_n \mathbf{b}^T - \alpha \mathbf{Y}\|_F^2 + \lambda \|\mathbf{Z}\|_F^2 \\ & \text{s.t. } \mathbf{Z}^T \mathbf{S}_t \mathbf{Z} = \mathbf{I}, \quad \mathbf{Y} \mathbf{1}_c = \mathbf{1}_n, \quad \mathbf{Y} \geq \mathbf{0}_{nc} \end{aligned} \quad (6)$$

which can be utilized for the direct clustering.

V. OPTIMIZATION METHODOLOGY WITH THEORETICAL ANALYSIS

How to solve the proposed clustering model, i.e., RURR-SL in (6), then takes the first priority. Hereinafter, coordinate blocking method, i.e., alternating method, is employed.

A. Optimize \mathbf{Y} With Fixing \mathbf{Z} , \mathbf{b} , and α

When \mathbf{Z} , \mathbf{b} , and α are fixed, problem (6) can be rewritten as

$$\min_{\mathbf{Y} \mathbf{1}_c = \mathbf{1}_n, \mathbf{Y} \geq \mathbf{0}_{nc}} \|\mathbf{V} - \alpha \mathbf{Y}\|_F^2 \quad (7)$$

which is equivalent to

$$\min_{\mathbf{Y}^{(\alpha)} \mathbf{1}_c = \alpha \mathbf{1}_n, \mathbf{Y}^{(\alpha)} \geq \mathbf{0}_{nc}} \|\mathbf{V} - \mathbf{Y}^{(\alpha)}\|_F^2 \quad (8)$$

where $\mathbf{V} = \mathbf{X}^T \mathbf{Z} + \mathbf{1}_n \mathbf{b}^T$ and $\mathbf{Y}^{(a)} = \alpha \mathbf{Y}$. Due to the independence of each soft label, i.e., each row vector of \mathbf{Y} , problem (8) can be individually solved by

$$\min_{(\mathbf{y}_i^{(a)})^T \mathbf{1}_c = \alpha, \mathbf{y}_i^{(a)} \geq \mathbf{0}_{c1}} \frac{1}{2} \|\mathbf{y}_i^{(a)} - \mathbf{v}_i\|_2^2, \quad (1 \leq i \leq n) \quad (9)$$

where $\mathbf{y}_i^{(a)} \in \mathbb{R}^{c \times 1}$ and $\mathbf{v}_i \in \mathbb{R}^{c \times 1}$ are the i th columns of $(\mathbf{Y}^{(a)})^T$ and \mathbf{V}^T , respectively. Hence, the Lagrangian function of (9) can be represented as

$$\mathcal{L} = \frac{1}{2} \|\mathbf{y}_i^{(a)} - \mathbf{v}_i\|_2^2 - \beta \left((\mathbf{y}_i^{(a)})^T \mathbf{1}_c - \alpha \right) - \sigma^T \mathbf{y}_i^{(a)} \quad (10)$$

where $\beta \in \mathbb{R}$ and $\sigma \in \mathbb{R}^{c \times 1} \geq \mathbf{0}_{c1}$ are Lagrangian multipliers. According to \mathcal{L} in (10), associated KKT conditions can be illustrated as

$$\begin{cases} \forall j, \mathbf{y}_{ij}^{(a)} - \mathbf{v}_{ij} - \beta - \sigma_j = 0 \\ \forall j, \mathbf{y}_{ij}^{(a)} \geq 0 \\ \forall j, \sigma_j \geq 0 \\ \forall j, \mathbf{y}_{ij}^{(a)} \sigma_j = 0. \end{cases} \quad (11)$$

According to the constraint $(\mathbf{y}_i^{(a)})^T \mathbf{1}_c = \alpha$ in (9) and (11), we have

$$\beta = \frac{\alpha - \mathbf{1}_c^T \mathbf{v}_i - \mathbf{1}_c^T \sigma}{c}. \quad (12)$$

Based on (11) and (12), we could further achieve that

$$\mathbf{y}_i^{(a)} = \mathbf{v}_i + \frac{\alpha}{c} \mathbf{1}_c - \frac{\mathbf{1}_c^T \mathbf{v}_i}{c} \mathbf{1}_c - \frac{\mathbf{1}_c^T \sigma}{c} \mathbf{1}_c + \sigma. \quad (13)$$

Denote $\bar{\sigma} = (\mathbf{1}_c^T \sigma / c)$ and $\mathbf{p}_i = \mathbf{v}_i + (\alpha / c) \mathbf{1}_c - (\mathbf{1}_c^T \mathbf{v}_i / c) \mathbf{1}_c$, then (13) could be reformulated into

$$\mathbf{y}_i^{(a)} = \mathbf{p}_i - \bar{\sigma} \mathbf{1}_c + \sigma. \quad (14)$$

In other words, (14) could be element-wisely illustrated as

$$\mathbf{y}_{ij}^{(a)} = \mathbf{p}_{ij} - \bar{\sigma} + \sigma_j \quad \forall j. \quad (15)$$

According to (11) and (15), the discussion will be further decomposed into the following cases:

$$\begin{cases} \text{Case 1: } \mathbf{y}_{ij}^{(a)} > 0, \sigma_j = 0 \Leftrightarrow \mathbf{y}_{ij}^{(a)} = \mathbf{p}_{ij} - \bar{\sigma} > 0 \\ \text{Case 2: } \mathbf{y}_{ij}^{(a)} = 0, \sigma_j > 0 \Leftrightarrow -\sigma_j = \mathbf{p}_{ij} - \bar{\sigma} < 0 \\ \text{Case 3: } \mathbf{y}_{ij}^{(a)} = \sigma_j = 0 \Leftrightarrow \mathbf{p}_{ij} - \bar{\sigma} = 0. \end{cases} \quad (16)$$

From (16), the closed-form solutions can be simply summarized as

$$\begin{cases} \mathbf{y}_{ij}^{(a)} = (\mathbf{p}_{ij} - \bar{\sigma})_+ \\ \sigma_j = (\bar{\sigma} - \mathbf{p}_{ij})_+. \end{cases} \quad (17)$$

From (17), it is easy to notice that $\mathbf{y}_{ij}^{(a)}$ can be determined if and only if $\bar{\sigma}$ can be obtained. Actually, based on the formulas of both $\bar{\sigma}$ and σ , we further have

$$\begin{aligned} \bar{\sigma} &= \frac{\mathbf{1}_c^T \sigma}{c} = \frac{\sum_{j=1}^c \sigma_j}{c} \\ &= \frac{1}{c} \sum_{j=1}^c (\bar{\sigma} - \mathbf{p}_{ij})_+. \end{aligned} \quad (18)$$

In order to solve $\bar{\sigma}$ in (18), we introduce the following function:

$$\mathbf{f}(\bar{\sigma}) = \frac{1}{c} \sum_{j=1}^c (\bar{\sigma} - \mathbf{p}_{ij})_+ - \bar{\sigma}. \quad (19)$$

Apparently, solving (18) is equivalent to finding the root of $\mathbf{f}(\bar{\sigma}) = 0$ in (19), which can be easily obtained by Newton method as

$$\bar{\sigma}_{t+1} = \bar{\sigma}_t - \frac{\mathbf{f}(\bar{\sigma}_t)}{\mathbf{f}'(\bar{\sigma}_t)} \quad (20)$$

where t represents the t th iteration.

B. Optimize \mathbf{Z} , \mathbf{b} , and α With Fixing \mathbf{Y}

When soft label \mathbf{Y} is fixed, problem (6) degenerates to the following supervised case:

$$\min_{\mathbf{Z}, \mathbf{b}, \alpha} \|\mathbf{X}^T \mathbf{Z} + \mathbf{1}_n \mathbf{b}^T - \alpha \mathbf{Y}\|_F^2 + \lambda \|\mathbf{Z}\|_F^2. \quad (21)$$

Note that bias \mathbf{b} is a free variable in (21), thus closed-form solution with respect to \mathbf{b} can be derived as

$$\begin{aligned} \frac{\partial (\|\mathbf{X}^T \mathbf{Z} + \mathbf{1}_n \mathbf{b}^T - \alpha \mathbf{Y}\|_F^2 + \lambda \|\mathbf{Z}\|_F^2)}{\partial \mathbf{b}} &= 0 \\ \Rightarrow (\mathbf{Z}^T \mathbf{X} - \alpha \mathbf{Y}^T) \mathbf{1}_n + \mathbf{b} \mathbf{1}_n^T \mathbf{1}_n &= 0 \\ \Rightarrow \mathbf{b} &= \frac{1}{n} (\alpha \mathbf{Y}^T - \mathbf{Z}^T \mathbf{X}) \mathbf{1}_n. \end{aligned} \quad (22)$$

By further substituting (22) into (21), supervised case of rescaled problem (6) can be reformulated into

$$\min_{\mathbf{Z}, \mathbf{b}, \alpha} \|\mathbf{H}(\mathbf{X}^T \mathbf{Z} - \alpha \mathbf{Y})\|_F^2 + \lambda \|\mathbf{Z}\|_F^2 \quad (23)$$

where the centering matrix $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \in \mathbb{R}^{n \times n}$ is idempotent, i.e., $\mathbf{H}^2 = \mathbf{H}$.

Based on (23), the discussion is decomposed into two subcases as follows.

1) *Optimize α With Fixing \mathbf{Z} and \mathbf{Y}* : When \mathbf{Z} and \mathbf{Y} are fixed, problem (23) can be rewritten as

$$\min_{\alpha} \alpha^2 \text{Tr}(\mathbf{Y}^T \mathbf{H} \mathbf{Y}) - 2\alpha \text{Tr}(\mathbf{Z}^T \mathbf{X} \mathbf{H} \mathbf{Y}) \quad (24)$$

which leads to the solution

$$\frac{\partial (\alpha^2 \text{Tr}(\mathbf{Y}^T \mathbf{H} \mathbf{Y}) - 2\alpha \text{Tr}(\mathbf{Z}^T \mathbf{X} \mathbf{H} \mathbf{Y}))}{\partial \alpha} = 0 \Rightarrow \alpha = \frac{\text{Tr}(\mathbf{Z}^T \mathbf{X} \mathbf{H} \mathbf{Y})}{\text{Tr}(\mathbf{Y}^T \mathbf{H} \mathbf{Y})}. \quad (25)$$

2) *Optimize \mathbf{Z} With Fixing α and \mathbf{Y}* : When α and \mathbf{Y} are fixed, problem (6) can be rewritten as

$$\max_{\mathbf{Z}, \mathbf{b}} 2\alpha \text{Tr}(\mathbf{Z}^T \mathbf{X} \mathbf{H} \mathbf{Y}) \quad (26)$$

which further leads to

$$\max_{\mathbf{Q}' \mathbf{Q} = \mathbf{I}} \text{Tr}(\mathbf{Q}^T \mathbf{M}) \quad (27)$$

where $\mathbf{Q} = \mathbf{S}_t^{\frac{1}{2}} \mathbf{Z}$ and $\mathbf{M} = \mathbf{S}_t^{-\frac{1}{2}} \mathbf{X} \mathbf{H} \mathbf{Y}$.

In particular, a closed-form solution can be achieved for problem (27) according to the following theorem.

Theorem 1: As for problem (27), closed-form solution with respect to \mathbf{Q} can be obtained as

$$\mathbf{U} \mathbf{V}^T = \arg \max_{\mathbf{Q}^T \mathbf{Q} = \mathbf{I}} \text{Tr}(\mathbf{Q}^T \mathbf{M}) \quad (28)$$

where $\mathbf{M} = \mathbf{U} \mathbf{S} \mathbf{V}^T$ via the compact SVD method with $\mathbf{U} \in \mathbb{R}^{d \times c}$, $\mathbf{S} \in \mathbb{R}^{c \times c}$, and $\mathbf{V} \in \mathbb{R}^{c \times c}$.

Due to Theorem 1, problem (26) has the closed-form solution as

$$\mathbf{Z} = \mathbf{S}_t^{-\frac{1}{2}} \mathbf{U} \mathbf{V}^T. \quad (29)$$

According to the closed-form solutions as previously derived in (17), (22), (25), and (29), RURR-SL method can be summarized in Algorithm 1 to serve as a clustering method.

Theorem 2: Algorithm 1 decreases the objective value of problem (6) monotonically until convergence.

Theorem 3: The label matrix \mathbf{Y} in (6) is row-wisely penalized by a latent ℓ_1 regularization.

Theorem 4: The solution \mathbf{Z} to problem (6) is a nonzero matrix with the rank c .

Algorithm 1 RURR-SL Method Under Problem (6)

Input: input data \mathbf{X} , total scatter \mathbf{S}_t , the cluster number c , and trade-off parameter λ .

Output: projection matrix $\mathbf{Z} \in \mathbb{R}^{d \times c}$ and soft label $\mathbf{Y} \in \mathbb{R}^{n \times c}$.

- 1 Initialize random soft matrix \mathbf{Y} satisfying $\mathbf{Y}\mathbf{1}_c = \mathbf{1}_n$;
- 2 **while** not converge **do**
- 3 Update $\mathbf{M} \leftarrow \mathbf{S}_t^{-\frac{1}{2}}\mathbf{X}\mathbf{H}\mathbf{Y}$;
- 4 Calculate $\mathbf{USV}^T = \mathbf{M}$ via compact SVD of \mathbf{M} according to Theorem 1;
- 5 Update $\mathbf{Z} \leftarrow \mathbf{S}_t^{-\frac{1}{2}}\mathbf{UV}^T$;
- 6 Update $\alpha \leftarrow \frac{\text{Tr}(\mathbf{Z}^T\mathbf{X}\mathbf{H}\mathbf{Y})}{\text{Tr}(\mathbf{Y}^T\mathbf{H}\mathbf{Y})}$;
- 7 Update $\mathbf{b} \leftarrow \frac{1}{n}(\alpha\mathbf{Y}^T - \mathbf{Z}^T\mathbf{X})\mathbf{1}_n$;
- 8 Update $\mathbf{V} \leftarrow \mathbf{X}^T\mathbf{Z} + \mathbf{1}_n\mathbf{b}^T$;
- 9 **for** $i = 1 : n$ **do**
- 10 Update $\mathbf{p}_i \leftarrow \mathbf{v}_i + \frac{\alpha}{c}\mathbf{1}_c - \frac{\mathbf{1}_c^T\mathbf{v}_i}{c}\mathbf{1}_c$;
- 11 Update $\bar{\sigma}$ via Newton method in (20);
- 12 **for** $j = 1 : c$ **do**
- 13 Update $\mathbf{y}_{ij}^{(a)} \leftarrow (\mathbf{p}_{ij} - \bar{\sigma})_+$;
- 14 **end**
- 15 **end**
- 16 Calculate $\mathbf{Y} = \frac{1}{\alpha}\mathbf{Y}^{(a)}$;
- 17 **end**
- 18 return \mathbf{Z} and \mathbf{Y} ;

The proofs of Theorems 1–4 are provided in the Appendix. According to Theorem 4, the trivial solution previously mentioned in (3) can be prevented.

Complexity Analysis: Suppose Algorithm 1 converges after l iterations. In each iteration, inverse computation is performed to solve problem (28), such that the cost is $O(d^3)$. The calculation of solution \mathbf{Z} costs $O(d^2n + d^2c + dn^2)$ and the cost of updating α is $O(dn^2 + dnc)$. Due to the fact that $n > c$, the total cost of Algorithm 1 is $O(l(d^3 + d^2n + dn^2))$. Since URR-SL and RURR-SL has the only difference regarding the adaptive scaling α , the cost of URR-SL is similar as RURR-SL in Algorithm 1.

VI. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed RURR-SL method in terms of two typical clustering evaluation metrics, namely, clustering accuracy (Accuracy) and normalized mutual information (NMI).

A. Clustering Accuracy

The clustering accuracy in the experiment can be computed by $\text{Accuracy} = (\sum_{i=1}^n \delta(\text{map}(\mathbf{r}_i), \mathbf{l}_i)/n)$, where \mathbf{r}_i represents the pseudo-cluster label of \mathbf{x}_i , \mathbf{l}_i represents the true class label, n is the number of data samples, $\delta(x, y)$ is the delta function, and $\text{map}(\cdot)$ is the optimal map function, which maps each cluster index to the best class label. Note that $\delta(x, y) = 1$, if $x = y$; $\delta(x, y) = 0$, otherwise. A larger **Accuracy** implies a better clustering performance.

B. Normalized Mutual Information

The NMI serves as an index to determine the quality of the clusters. Let $\{\mathbf{C}_i\}$, $(1 \leq i \leq c)$ denotes the set of clusters obtained from the ground truth and $\{\mathbf{C}'_j\}$, $(1 \leq j \leq c)$ denotes the clusters obtained from the proposed method. Their NMI NMI is defined as $\text{NMI} = (\sum_{i=1}^c \sum_{j=1}^c n_{ij} \log(n_{ij}/n_i \hat{n}_j))/((\sum_{i=1}^c n_i \log(n_i/n))(\sum_{j=1}^c \hat{n}_j \log(\hat{n}_j/n)))^{1/2}$, where n_i denotes the number of data in the cluster \mathbf{C}_i , $(1 \leq i \leq c)$ and \hat{n}_j denotes the number of data belonging to the

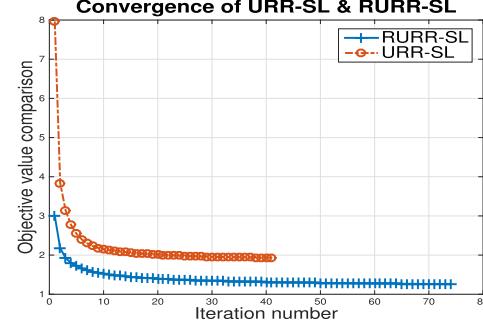


Fig. 1. Convergent curves of URR-SL ($\alpha = 1$) and RURR-SL are performed under random generated data.

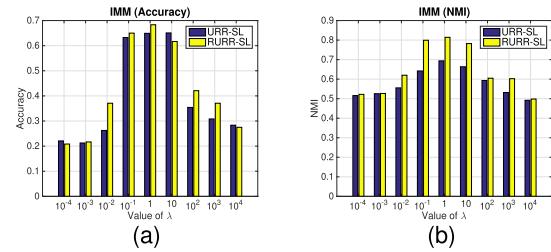


Fig. 2. Parameter tuning for URR-SL and RURR-SL on the data set IMM under the two metrics, namely, Accuracy (IMM). (a) Accuracy (IMM). (b) NMI (IMM).

class \mathbf{C}'_j , $(1 \leq j \leq c)$. In particular, n_{ij} is the number of data, which are in the intersection between cluster \mathbf{C}_i and class \mathbf{C}'_j . Similarly, a larger **NMI** represents a more consistent clustering performance. Accordingly, **NMI** ranges from 0 to 1. More specifically, **NMI** = 1 if two sets of clusters (set from ground truth and set from cluster algorithm) are identical, whereas **NMI** = 0 if they are independent.

C. Comparative Methods and Parameter Tuning

Four clustering approaches are utilized for the comparison, which includes the following.

- 1) k -means method [7] assigns the given data samples into clusters such that the data point is assigned to only one cluster with 100% probability.
- 2) Robust k -means method (RMKMC) [33] utilizes $\ell_{2,1}$ -norm loss such that the cluster centroids are in the form of weighted mean.
- 3) Fuzzy k -means clustering (FKM) [24] allows each object to possess a certain degree of membership to each cluster rather than having a membership to only one cluster as k -means.
- 4) Robust and sparse fuzzy k -means clustering (RSFKM) [26] applies robust loss function to tackle data outliers with adding the regularization for the proper sparseness. Since the RSFKM with capped ℓ_2 -norm has a better performance than $\ell_{2,1}$ -norm, we only employ RSFKM (capped ℓ_2) as a competitor in this brief.

Particularly, URR-SL in (4) is also utilized with $\alpha = 1$. As for all the methods, real cluster number c is pre-given as *a priori*.

On one hand, from Fig. 1, we notice that the RURR-SL converges to a less objective value than URR-SL does. In other words, RURR-SL is numerically better than URR-SL. On the other hand, both URR-SL and RURR-SL methods have the regularization parameter λ to tune. More specifically, regularization parameter λ is searched in the grid of interval $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3, 10^4]$. As for the comparative methods, k -means and RMKMC are parameter-free with the input cluster prior c . The parameter of FKM, i.e., the fuzzy level is chosen as 2.5 for the optimal performance, while RSFKM has only an integer parameter, which is tuned in the grid

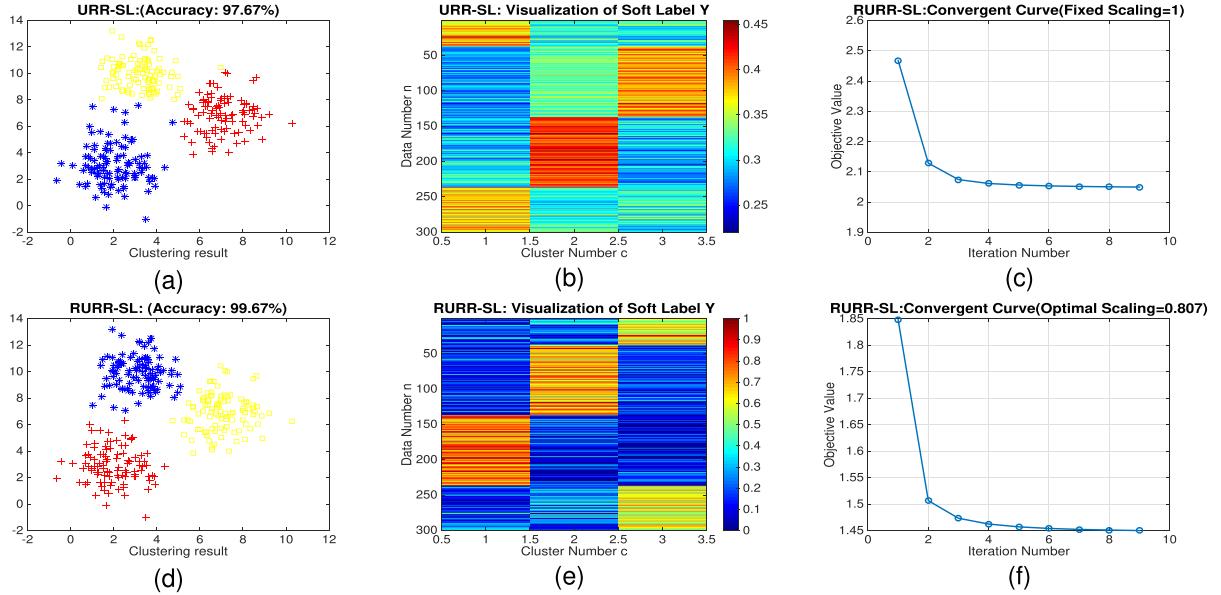


Fig. 3. Comparison of URR-SL ($\alpha = 1$) and RURR-SL regarding clustering result, soft label, and convergence under the three-cluster Gaussian distributed data. (a) URR-SL: clustering. (b) URR-SL: label. (c) URR-SL: convergence. (d) RURR-SL: clustering. (e) RURR-SL: label. (f) RURR-SL: convergence.

TABLE I
COMPARISON OF CLUSTERING ACCURACY AND NMI UNDER EIGHT BENCHMARK DATA SETS

Dataset		IMM		AT&T		FLOWER ₁₇		COLON	
Method	Accuracy(%)	NMI(%)	Accuracy(%)	NMI(%)	Accuracy(%)	NMI(%)	Accuracy(%)	NMI(%)	
<i>k</i> -means	63.77	68.90	61.25	68.90	23.24	36.79	57.06	68.87	
RMKMC	66.28	71.61	63.51	71.72	27.76	40.91	58.79	68.93	
FKM	60.83	66.77	61.13	68.83	25.88	39.52	57.19	66.76	
RSFKM ($\ell_{2,1}$)	64.11	68.04	62.08	70.04	27.30	38.52	57.23	62.93	
URR-SL ($\alpha = 1$)	65.38	70.58	62.50	70.56	25.93	37.68	58.68	65.41	
RURR-SL	68.95	81.24	64.33	74.99	30.44	41.86	59.80	69.99	
Dataset		GT		ETHZ ₅₃		FEI		GLIOMA	
Method	Accuracy(%)	NMI(%)	Accuracy(%)	NMI(%)	Accuracy(%)	NMI(%)	Accuracy(%)	NMI(%)	
<i>k</i> -means	55.73	63.06	52.74	61.77	44.32	55.87	68.52	81.38	
RMKMC	58.29	65.13	54.81	65.22	46.02	59.72	70.56	80.09	
FKM	56.61	62.91	53.09	64.00	45.99	54.10	68.83	78.21	
RSFKM ($\ell_{2,1}$)	57.28	63.03	53.26	60.14	46.76	56.05	68.96	77.35	
URR-SL ($\alpha = 1$)	57.87	65.25	54.32	62.78	46.18	57.29	69.77	80.16	
RURR-SL	59.91	65.77	55.43	66.89	48.21	59.35	72.11	80.00	

of [10, 15, 20, 25, 30, 35, 40, 45, 50] for selecting the optimal neighbors. Besides that, an illustration of parameter tuning for the proposed URR-SL and RURR-SL is demonstrated in Fig. 2 under the data set IMM. From Fig. 2, we notice that both URR-SL and RURR-SL achieve the optimal Accuracy and NMI around $\lambda \in [1, 10]$ for the data set IMM.

D. Clustering Results

1) *Synthetic/Toy Data Sets:* In Fig. 3, the proposed RURR-SL method is compared with URR-SL method for clustering 3-cluster Gaussian distributed data. In addition, the associated soft label and convergent curve are also provided. In Fig. 4, the proposed RURR-SL method is compared with URR-SL method and *k*-means method for clustering multicluster data. Accordingly, the following conclusions can be drawn:

- From Fig. 3, RURR-SL is better than URR-SL with 2% Accuracy improvement under 3-cluster toy data. From Fig. 4, both RURR-SL and URR-SL methods have better clustering

results than *k*-means under the multicluster toy data, i.e., 7% and 4.5% Accuracy improvements, respectively.

- From Figs. 3 and 4, we could observe that the RURR-SL method performs consistently better than the URR-SL method in the aspects of both objective value and clustering results due to the adaptive scaling in the rescaled counterpart of URR-SL.

2) *Benchmark Data Sets:* In Fig. 5, soft labels of RURR-SL and URR-SL methods are illustrated. In Table I, average clustering accuracy and NMI of the clustering approaches previously mentioned are reported under eight benchmark data sets, where each experiment is run for 10 times. Therefore, we could conclude that:

- From Figs. 3 and 5, we could observe a row sparse soft label, which corresponds to the analysis of Theorem 3.
- From Table I, the RURR-SL method performs consistently better than other clustering approaches regarding clustering accuracy. Especially for the data set FLOWER₁₇, RURR-SL has 2.68% improvement than the runner-up method RMKMC.

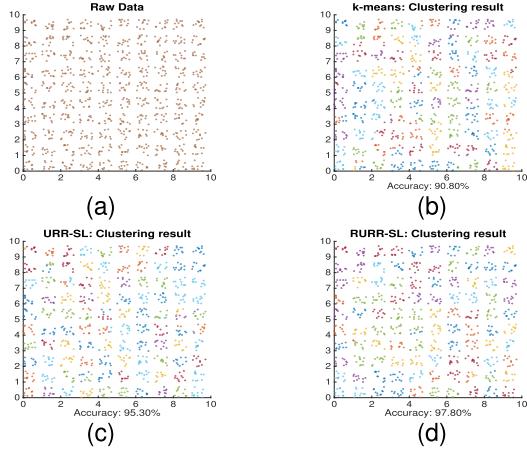


Fig. 4. Comparison of the clustering results for k -means, URR-SL, and RURR-SL under multicluster data. (a) Raw (Multiclusters). (b) k -means. (c) URR-SL ($\alpha = 1$). (d) RURR-SL ($\alpha = 0.375$).

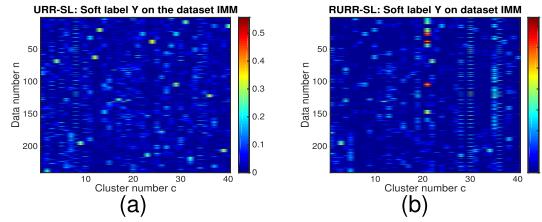


Fig. 5. Illustration of the soft label for URR-SL and RURR-SL under the data set IMM. (a) URR-SL (IMM). (b) RURR-SL (IMM).

TABLE II

DETAILS OF THE SELECTED BENCHMARK DATA SETS

Dataset	Data No.	Class No.	Feature No.
AT&T	400	40	1024
COLON	62	2	2000
EHTZ ₅₃	265	53	1024
FEI	2800	200	1024
FLOWER ₁₇	1360	17	1024
GLIOMA	50	4	4434
GT	750	50	1024
IMM	240	40	1024

- 3) From Table I, the RURR-SL method performs better than other clustering approaches in most cases regarding NMI. Especially for the data set IMM, RURR-SL has 9.63% improvement than the runner-up method RMKMC.
- 4) From Table I, the RURR-SL method performs better than the URR-SL in terms of both clustering metrics due to the rescaled strategy, i.e., obtaining the optimal scaling.

E. Data Sets

First, two synthetic or toy data sets are utilized including 3-cluster Gaussian distributed data and multicluster data. Second, eight benchmark data sets are used including AT&T,¹ COLON,² EHTZ₅₃,³ FEI,⁴ GLIOMA,⁵ GT,⁶ FLOWER₁₇,⁷ and IMM.⁸ More specifics of each benchmark data set are listed in Table II.

¹http://www.cl.cam.ac.uk/Research/DTG/attarchive/pub/data/att_faces.zip

²<http://penglab.janelia.org/proj/mRMR/index.htm#data>

³http://www.vision.ee.ethz.ch/datasets_extra/Obj_DB.tar.gz

⁴<http://fei.edu.br/cet/facedatabase.html>

⁵<http://featureselection.asu.edu/datasets.php>

⁶http://www.anefian.com/research/gt_db.zip

⁷<http://www.robots.ox.ac.uk/vgg/data0.html>

⁸<http://www2.imm.dtu.dk/aam/datasets/datasets.html>

VII. CONCLUSION

In this brief, URR-SL is proposed for data clustering, such that the potential trivial solution triggered by unsupervised ridge regression model can be avoided. In addition, manifold structure is incorporated into the subspace via the uncorrelated constraint. Moreover, a row sparse soft label is obtained correspondingly due to the latent ℓ_1 regularization. To further strengthen the URR-SL model, a novel rescaled strategy is provided, such that an RURR-SL is achieved. Consequently, an effective clustering method is proposed with obtaining the optimal scaling automatically. Extensive experiments are performed to validate the effectiveness of the proposed method.

APPENDIX

A. Proof of Theorem 1

Proof: Suppose full SVD of \mathbf{M} is $\mathbf{M} = \mathbb{U}\Sigma\mathbb{V}^T$ with $\mathbb{U} \in \mathbb{R}^{d \times d}$, $\Sigma \in \mathbb{R}^{d \times c}$ and $\mathbb{V} \in \mathbb{R}^{c \times c}$, then we have

$$\text{Tr}(\mathbf{Q}^T \mathbf{M}) = \text{Tr}(\mathbf{Q}^T \mathbb{U} \Sigma \mathbb{V}^T) = \text{Tr}(\Sigma \mathbf{G}) = \sum_{i=1}^c \sigma_{ii} g_{ii}$$

where $\mathbf{G} = \mathbb{V}^T \mathbf{Q}^T \mathbb{U} \in \mathbb{R}^{c \times d}$ with g_{ii} and σ_{ii} being the (i, i) th elements of the matrix \mathbf{G} and Σ , respectively.

Note that $\mathbf{G}\mathbf{G}^T = \mathbf{I}$, thus $|g_{ii}| \leq 1$. On the other hand, $\sigma_{ii} \geq 0$ since σ_{ii} is a singular value of the matrix \mathbf{M} . Therefore, we have $\text{Tr}(\mathbf{Q}^T \mathbf{M}) = \sum_{i=1}^c g_{ii} \sigma_{ii} \leq \sum_{i=1}^c \sigma_{ii}$.

Apparently, the equality holds when $g_{ii} = 1, (1 \leq i \leq c)$. In other words, $\text{Tr}(\mathbf{Q}^T \mathbf{M})$ reaches the maximum when the matrix $\mathbf{G} = [\mathbf{I}, \mathbf{0}_{c(d-c)}] \in \mathbb{R}^{c \times d}$. Recall that $\mathbf{G} = \mathbb{V}^T \mathbf{Q}^T \mathbb{U}$, thus the optimal solution to problem (28) can be represented as

$$\mathbf{Q} = \mathbb{U} \mathbf{G}^T \mathbb{V}^T = \mathbb{U} [\mathbf{I}; \mathbf{0}_{(d-c)c}] \mathbb{V}^T. \quad (30)$$

Since (30) stems from full SVD of matrix \mathbf{M} , (30) can be rewritten as $\mathbf{Q} = \mathbf{U} \mathbf{V}^T$ via compact SVD of matrix \mathbf{M} . \square

B. Proof of Theorem 2

Proof: Suppose \mathbf{Z}_{t+1} , α_{t+1} , and \mathbf{Y}_{t+1} are updated by \mathbf{Z}_t , α_t , and \mathbf{Y}_t where t stands for the t th iteration; then, according to steps 5 and 6 in Algorithm 1 and Theorem 1, we have

$$\begin{aligned} & \text{Tr}(\mathbf{Z}_{t+1}^T \mathbf{X} \mathbf{H} \mathbf{Y}_t) \\ & \geq \text{Tr}(\mathbf{Z}_t^T \mathbf{X} \mathbf{H} \mathbf{Y}_t) \\ & \Rightarrow \alpha_t^2 \text{Tr}(\mathbf{Y}_t^T \mathbf{H} \mathbf{Y}_t) - 2\alpha_t \text{Tr}(\mathbf{Z}_{t+1}^T \mathbf{X} \mathbf{H} \mathbf{Y}_t) + \text{Tr}(\mathbf{I}) \\ & \leq \alpha_t^2 \text{Tr}(\mathbf{Y}_t^T \mathbf{H} \mathbf{Y}_t) - 2\alpha_t \text{Tr}(\mathbf{Z}_t^T \mathbf{X} \mathbf{H} \mathbf{Y}_t) + \text{Tr}(\mathbf{I}) \\ & \Rightarrow \alpha_{t+1}^2 \text{Tr}(\mathbf{Y}_t^T \mathbf{H} \mathbf{Y}_t) - 2\alpha_{t+1} \text{Tr}(\mathbf{Z}_{t+1}^T \mathbf{X} \mathbf{H} \mathbf{Y}_t) + \text{Tr}(\mathbf{Z}_{t+1}^T \mathbf{S}_t \mathbf{Z}_{t+1}) \\ & \leq \alpha_t^2 \text{Tr}(\mathbf{Y}_t^T \mathbf{H} \mathbf{Y}_t) - 2\alpha_t \text{Tr}(\mathbf{Z}_t^T \mathbf{X} \mathbf{H} \mathbf{Y}_t) + \text{Tr}(\mathbf{Z}_t^T \mathbf{S}_t \mathbf{Z}_t) \\ & \Rightarrow \|\mathbf{H}(\mathbf{X}^T \mathbf{Z}_{t+1} - \alpha_{t+1} \mathbf{Y}_t)\|_F^2 + \lambda \|\mathbf{Z}_{t+1}\|_F^2 \\ & \leq \|\mathbf{H}(\mathbf{X}^T \mathbf{Z}_t - \alpha_t \mathbf{Y}_t)\|_F^2 + \lambda \|\mathbf{Z}_t\|_F^2. \end{aligned} \quad (31)$$

In addition, from step 7 to step 17 in Algorithm 1, we could infer that

$$\begin{aligned} & \|\mathbf{V}_{t+1} - \alpha_{t+1} \mathbf{Y}_{t+1}\|_F^2 \\ & \leq \|\mathbf{V}_{t+1} - \alpha_{t+1} \mathbf{Y}_t\|_F^2 \\ & \Rightarrow \|\mathbf{H}(\mathbf{X}^T \mathbf{Z}_{t+1} - \alpha_{t+1} \mathbf{Y}_{t+1})\|_F^2 + \lambda \|\mathbf{Z}_{t+1}\|_F^2 \\ & \leq \|\mathbf{H}(\mathbf{X}^T \mathbf{Z}_{t+1} - \alpha_{t+1} \mathbf{Y}_t)\|_F^2 + \lambda \|\mathbf{Z}_{t+1}\|_F^2. \end{aligned} \quad (32)$$

Therefore, (31) and (32) eventually lead to

$$\begin{aligned} & \|\mathbf{H}(\mathbf{X}^T \mathbf{Z}_{t+1} - \alpha_{t+1} \mathbf{Y}_{t+1})\|_F^2 + \lambda \|\mathbf{Z}_{t+1}\|_F^2 \\ & \leq \|\mathbf{H}(\mathbf{X}^T \mathbf{Z}_t - \alpha_t \mathbf{Y}_t)\|_F^2 + \lambda \|\mathbf{Z}_t\|_F^2 \end{aligned}$$

which indicates that Algorithm 1 decreases the objective value of problem (6) monotonically. Moreover, 0 serves as a lower bound of problem (6). In sum, Theorem 2 can be proved. \square

C. Proof of Theorem 3

Proof: The soft label \mathbf{Y} is achieved by individually solving the transpose of its each row vector in (9) as

$$\min_{\mathbf{y}_i^T \mathbf{1}_{c=1}, \mathbf{y}_i \geq \mathbf{0}_{cl}} \frac{1}{2} \|\alpha \mathbf{y}_i - \mathbf{v}_i\|_2^2$$

which implies a latent ℓ_1 regularization. Particularly, we could infer that

$$\min_{\mathbf{y}_i^T \mathbf{1}_{c=1}, \mathbf{y}_i \geq \mathbf{0}_{cl}} \frac{1}{2} \|\alpha \mathbf{y}_i - \mathbf{v}_i\|_2^2 = \min_{\mathbf{y}_i^T \mathbf{1}_{c=1}, \mathbf{y}_i \geq \mathbf{0}_{cl}} \frac{1}{2} \|\alpha \mathbf{y}_i - \mathbf{v}_i\|_2^2 + \xi \|\mathbf{y}_i\|_1$$

which indicates that each row of \mathbf{Y} is penalized by latent ℓ_1 regularization. \square

D. Proof of Theorem 4

Proof: As for the solution to problem (6), \mathbf{Z} satisfies the uncorrelated constraint $\mathbf{Z}^T \mathbf{S}_t \mathbf{Z} = \mathbf{I}$. On one hand, we have

$$\text{rank}(\mathbf{Z}^T \mathbf{S}_t \mathbf{Z}) = \text{rank}(\mathbf{I}) = c$$

where **rank** denotes the rank of the matrix.

On the other hand, since total scatter matrix \mathbf{S}_t is positive definite, we could infer that

$$\text{rank}(\mathbf{Z}^T \mathbf{S}_t \mathbf{Z}) = \text{rank}\left(\left(\mathbf{S}_t^{\frac{1}{2}} \mathbf{Z}\right)^T \mathbf{S}_t^{\frac{1}{2}} \mathbf{Z}\right) = \text{rank}\left(\mathbf{S}_t^{\frac{1}{2}} \mathbf{Z}\right) = \text{rank}(\mathbf{Z}).$$

In sum, Theorem 4 can be proved. \square

REFERENCES

- [1] R. Zhang, F. Nie, and X. Li, "Self-weighted spectral clustering with parameter-free constraint," *Neurocomputing*, vol. 241, pp. 164–170, Jun. 2017.
- [2] X. Chen, F. Nie, J. Z. Huang, and M. Yang, "Scalable normalized cut with improved spectral rotation," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1518–1524.
- [3] F. Nie, W. Zhu, and X. Li, "Unsupervised large graph embedding," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2422–2428.
- [4] F. Nie, X. Wang, M. I. Jordan, and H. Huang, "The constrained Laplacian rank algorithm for graph-based clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1969–1976.
- [5] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 977–986.
- [6] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.
- [7] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.
- [8] K. Tasdemir and E. Merenyi, "Exploiting data topology in visualization and clustering of self-organizing maps," *IEEE Trans. Neural Netw.*, vol. 20, no. 4, pp. 549–562, Apr. 2009.
- [9] K. Tasdemir, "Graph based representations of density distribution and distances for self-organizing maps," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 520–526, Mar. 2010.
- [10] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
- [11] H. Hu, Z. Lin, J. Feng, and J. Zhou, "Smooth representation clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3834–3841.
- [12] C.-G. Li and R. Vidal, "Structured sparse subspace clustering: A unified optimization framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 277–286.
- [13] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.
- [14] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *J. Mach. Learn. Res.*, vol. 2, pp. 125–137, Mar. 2002.
- [15] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, pp. 780–784, May 2002.
- [16] A. Szymkowiak-Have, M. A. Girolami, and J. Larsen, "Clustering via kernel decomposition," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 256–264, Jan. 2006.
- [17] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1490–1507, Jul. 2017.
- [18] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1302–1308.
- [19] D. Wang, F. Nie, and H. Huang, "Feature selection via global redundancy minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2743–2755, Oct. 2015.
- [20] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2010, pp. 333–342.
- [21] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou, " $\ell_{2,1}$ -norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1589–1594.
- [22] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proc. AAAI*, 2012, pp. 1026–1032.
- [23] E. H. Ruspini, "A new approach to clustering," *Inf. Control*, vol. 15, no. 1, pp. 22–32, 1969.
- [24] J. C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 1, pp. 1–8, Jan. 1980.
- [25] M. J. Li, M. K. Ng, Y.-M. Cheung, and J. Z. Huang, "Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 11, pp. 1519–1534, Nov. 2008.
- [26] J. Xu, J. Han, K. Xiong, and F. Nie, "Robust and sparse fuzzy k-means clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2224–2230.
- [27] Z. Deng, K.-S. Choi, Y. Jiang, and S. Wang, "Generalized hidden-mapping ridge regression, knowledge-leveraged inductive transfer learning for neural networks, fuzzy systems and kernel methods," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2585–2599, Dec. 2014.
- [28] N. Kim, Y.-S. Jeong, M.-K. Jeong, and T. M. Young, "Kernel ridge regression with lagged-dependent variable: Applications to prediction of internal bond strength in a medium density fiberboard process," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1011–1020, Nov. 2012.
- [29] F. Nie, S. Xiang, Y. Liu, C. Hou, and C. Zhang, "Orthogonal vs. uncorrelated least squares discriminant analysis for feature extraction," *Pattern Recognit. Lett.*, vol. 33, no. 5, pp. 485–491, Apr. 2012.
- [30] J. Ye, R. Janardan, Q. Li, and H. Park, "Feature reduction via generalized uncorrelated linear discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1312–1322, Oct. 2006.
- [31] F. Nie, R. Zhang, and X. Li, "A generalized power iteration method for solving quadratic problem on the stiefel manifold," *Sci. China Inf. Sci.*, vol. 60, no. 11, pp. 1–10, Nov. 2017.
- [32] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, Mar. 1991.
- [33] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2598–2604.