

Chapter 2

Understanding-Oriented Unsupervised Feature Selection

Abstract In many image processing and pattern recognition problems, visual contents of images are currently described by high-dimensional features, which are often redundant and noisy. Toward this end, we propose two novel understanding-oriented unsupervised feature selection schemes. For exploring discriminative information, nonnegative spectral analysis is proposed to learn more accurate cluster labels of the input images. For feature selection, the hidden structure shared by different features and the redundancy among different features are explored, respectively. Row-wise sparse models with the $\ell_{2,p}$ -norm ($0 < p \leq 1$) are leveraged to make the proposed models suitable for feature selection and robust to noise.

2.1 Introduction

In many image processing and multimedia problems, images are usually represented by high-dimensional visual features, such as local features (such as SIFT [28]). In practice, it is well known that all features that characterize images are not usually equal important for a given task and most of them are often correlated or redundant to each other, and sometimes noisy [13]. Besides, it is hard to discriminate images of different classes from each other in the high-dimensional space of visual features. That is, these high-dimensional features may bring some disadvantages, such as over-fitting, low efficiency, and poor performance, to the traditional learning models [42]. As a consequence, it is necessary and challenging to select an optimal feature subset from high-dimensional image to remove irrelevant and redundant features, increase learning accuracy, and improve the performance comprehensibility.

The task of selecting the “best” feature subset is known as *feature selection*, which is an important and widely used method. The importance of feature selection in improving both the efficiency and accuracy of image processing is three-fold. First, it can result in computationally efficient algorithms since the dimensionality of selected feature subset is much lower. Second, it enables to provide a better understanding of the underlying structure of the data. Finally, it can improve the

©2017 IEEE. Reprinted, with permission, from Li et al. [25], and Li and Tang [26].

performance by removing noisy and redundant features. Therefore, many feature selection methods have been proposed and studied [5, 17, 19, 21, 27, 33, 34, 39, 47]. These algorithms can be categorized as supervised algorithms, semi-supervised algorithms and unsupervised algorithms, according to the way of utilizing label information. Since the discriminative information is encoded in the labels, supervised and semi-supervised approaches can generally achieve good performance. However, the labels of data annotated by human experts are typically expensive and time-consuming and there is usually no shortage of unlabeled data in many real-world applications. Consequently, it is quite promising and demanding to develop unsupervised feature selection techniques, which may be more practical.

In unsupervised feature selection, features are selected based on a frequently used criterion which evaluates features by their capability of keeping certain properties of the data, such as the data distribution, the redundancy of features, or local structure. The whole features contain necessary features (which are essential for the task), redundant features (which are useful but dependent on each other. Thus, not all of the redundant features are not necessary.), noisy features (which degrade the performance), and indifferent features (which do not matter for the task). The goal of feature selection is to select necessary features, discard noisy or indifferent features, and control the use of redundant features. The previous methods do not jointly consider these four features. Besides, they fail to exploit discriminative information from data. On the other hand, due to the absence of labels that would guide the search for discriminative features, unsupervised feature selection is considered as a much harder problem [14], which evaluates feature relevance by their capability of keeping certain properties of the data.

In light of all these factors, we propose a novel understanding-oriented unsupervised feature selection framework to explore discriminative information from data and the latent structural analysis, select necessary features, discard noisy or indifferent features and control the use of redundant features simultaneously. Due to the importance of discriminative information, it is necessary and beneficial to exploit discriminative information in unsupervised feature selection. As a consequence, we propose a novel nonnegative spectral analysis scheme to uncover discriminative information by learning more accurate cluster indicators. With nonnegative and orthogonality constraints, the learned cluster indicators are much closer to the ideal ones and can be readily utilized to obtain more accurate cluster labels, which can be utilized to guide feature selection. The joint learning of the cluster labels and feature selection matrix enables to select the most discriminative features. For the sake of feature selection, the predictive matrix is constrained to be sparse in rows, which is formulated as a general $\ell_{2,p}$ -norm ($0 < p \leq 1$) minimization term. Furthermore, on one hand, the features are correlated as they jointly reflect the semantic components. It is reasonable to assume that the features share a common structure in a low-dimensional space. The cluster indicators are predicted by the original features together with the features in the low-dimensional subspace. The latent structural analysis can uncover the feature correlations to make the results more reliable. On the other hand, the redundancy between features is explicitly exploited to control the redundancy of the selected features. The proposed problems are formulated

as optimization problems with well-defined objective functions. To solve the proposed problems, simple yet efficient iterative algorithms are proposed. Extensive experiments are conducted on face data, handwritten digit data, document data, and biomedical data. The experimental results show that compared with several representative algorithms, the proposed approaches achieve encouraging performance. Most of the work in this chapter has been published in [25, 26].

2.2 Related Work

According to the availability of label information, feature selection algorithms can be classified into three broad categories: supervised, semi-supervised, and unsupervised approaches. More details can be obtained in [19, 48]. In this section, we will elaborate unsupervised feature selection methods.

From the perspective of selection strategy, the unsupervised feature selection approaches can be broadly categorized as the *filter*, *wrapper*, and *embedded* ones. For filter methods [9, 17, 30, 47], a proxy measure is utilized to score a feature subset instead of the error rate. The simplest measure may be the variance score with the assumption that larger variance means better representation ability. However, there is no reason to assume that these features are useful for discriminating data in different classes. Laplacian Score [17] selects features which can best reflect the underlying manifold structure. However, the redundancy among features is not exploited, which may result in redundant features and compromise the performance. Filters are usually less computationally intensive than wrappers, but produce a feature set which is not tuned to a specific type of predictive model. Wrapper methods [14, 42, 46] score feature subsets using a predictive model. They wrap feature search around the learning algorithms and utilize the learned results to select features. Clustering is a commonly utilized learning algorithm [5, 14, 46]. The clusterability of the input data points is measured by analyzing the spectral properties of the affinity matrix. MCFS [5] uses a two-step spectral regression approach to unsupervised feature selection. Embedded methods [10, 23] perform feature selection as a part of the model construction process, which fall in between filters and wrappers in terms of computational complexity.

State-of-the-art algorithms exploit discriminative information and feature correlation to select features [12, 27, 35, 38, 43]. Nonnegative Discriminative Feature Selection (NDFS) [27] proposes nonnegative spectral clustering to guide feature selection and selects features over the whole feature space. In [12], a global and a set of locally linear regression model are integrated into a unified learning framework. Qian et al. [35] extended NDFS to handle outliers or noise data. The graph embedding and sparse spectral regression are improved in [38]. However, the above methods do not explicitly control the redundancy between features, which may lead to redundancy existing in the selected features.

Some methods have been designed to consider the dependency between features. In [45], the redundancy between selected features is removed using a

correlation-based filter. Peng et al. [34] proposed a mutual information-based two-stage feature selection approach to choose features with least redundancy by minimizing the mutual information among the selected features. A multilayer perceptron neural network is designed for feature selection with consideration a measure of linear dependency to control the redundancy in [6]. However, they only focus on considering the dependency between features, and fail to select discriminative features. In this work, we select features by considering the dependency between features and the discriminant information simultaneously. The most discriminant features with controlled redundancy are selected.

Different from previous work, the proposed framework exploits nonnegative spectral analysis, the underlying structure analysis and explicitly controls the redundancy between features in a joint framework for unsupervised feature learning. One general sparse model with $\ell_{2,p}$ -norm ($0 < p \leq 1$) is adopted to learn a better sparsity matrix.

2.3 Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection

Assume that we have n samples $\mathbb{X} = \{\mathbf{x}_i\}_{i=1}^n$. Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ denote the data matrix, in which $\mathbf{x}_i \in \mathbb{R}^d$ is the feature descriptor of the i th sample. Suppose these n samples are sampled from c classes. Denote $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^T \in \{0, 1\}^{n \times c}$, where $\mathbf{y}_i \in \{0, 1\}^{c \times 1}$ is the cluster indicator vector for \mathbf{x}_i . That is, $Y_{ij} = 1$ if the sample \mathbf{x}_i is assigned to the j th cluster, and $Y_{ij} = 0$ otherwise. The scaled cluster indicator matrix \mathbf{F} is defined:

$$\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_c] = \mathbf{Y}(\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}}, \quad (2.1)$$

It turns out that

$$\mathbf{F}^T \mathbf{F} = (\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}} \mathbf{Y}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}} = \mathbf{I}_c, \quad (2.2)$$

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

To select the discriminative features for unsupervised learning, we propose a Clustering-Guided Sparse Structural Learning (CGSSL) method to jointly exploit the cluster analysis and sparse structural analysis simultaneously. Clustering techniques are adopted to learn the cluster indicators (which can be regarded as pseudo class labels), which are used to guide the process of structural learning. Meanwhile, the pseudo class labels are also predicted by the structural learning with predictive functions, which correlate the samples and the pseudo class labels. To conduct effective feature selection, we impose the sparse feature selection models on the regularization term. Therefore, CGSSL is formulated as

$$\begin{aligned} \min_{\mathbf{F}, h} J(\mathbf{F}) + \sum_{i=1}^c \left(\alpha \sum_{j=1}^n l(h_i(\mathbf{x}_j), \mathbf{f}_i) + \Omega(h_i) \right) \\ \text{s.t. } \mathbf{F} = \mathbf{Y}(\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}}, \end{aligned} \quad (2.3)$$

where $J(\mathbf{F})$ is a clustering criterion, $l(\cdot, \cdot)$ is the loss function, $h_i(\cdot)$ is a predictive function for the i th cluster, and $\Omega(\cdot)$ is a regularization function with sparsity. α is a trade-off parameter.

2.3.1 Nonnegative Spectral Clustering

In cluster analysis, graph-theoretic methods have been well studied and utilized in many applications. As one of graph-theoretic methods, spectral clustering has been verified to be effective to detect the cluster structure of data and has received significant research attention. Therefore, we adopt spectral clustering as the cluster analysis technique.

Clearly, an effective cluster indicator matrix is more capable to reflect the discriminative information of the input data. The local geometric structure of data plays an important role in data clustering, which has been exploited by many spectral clustering algorithms [32, 37, 40]. Note that there are many different algorithms to uncover local data structure. In this work, we use the strategy proposed in [37] to be the criterion for its simplicity. The local geometric structure can be effectively modeled by a nearest neighbor graph on a scatter of data points. To construct the affinity graph \mathbf{S} , we define

$$S_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right) & \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ 0 & \text{otherwise,} \end{cases} \quad \text{None}$$

where $\mathcal{N}_k(\mathbf{x})$ is the set of k -nearest neighbors of \mathbf{x} . The local geometrical structure can be exploited by

$$\min_{\mathbf{F}} \frac{1}{2} \sum_{i,j=1}^n S_{ij} \left\| \frac{\mathbf{f}_i}{\sqrt{E_{ii}}} - \frac{\mathbf{f}_j}{\sqrt{E_{jj}}} \right\|_2^2 = \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}], \quad (2.5)$$

where \mathbf{E} is a diagonal matrix with $E_{ii} = \sum_{j=1}^n S_{ij}$ and $\mathbf{L} = \mathbf{E}^{-1/2}(\mathbf{E} - \mathbf{S})\mathbf{E}^{-1/2}$ is the normalized graph Laplacian matrix. Therefore $J(\mathbf{F})$ is defined as

$$J(\mathbf{F}) = \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}]. \quad (2.6)$$

According to the definition of \mathbf{F} , its elements are constrained to be discrete values, making the problem (2.3) an NP-hard problem [37]. A well-known solution is to relax it from discrete values to continuous ones while keeping the property of Eq. (2.2) [37], i.e., the objective function (2.3) is relaxed to

$$\begin{aligned} \min_{\mathbf{F}, h} \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \sum_{i=1}^c \left(\alpha \sum_{j=1}^n l(h_i(\mathbf{x}_j), \mathbf{f}_i) + \Omega(h_i) \right) \\ \text{s.t. } \mathbf{F}^T \mathbf{F} = \mathbf{I}_c. \end{aligned} \quad (2.7)$$

Note that according to the definition of the cluster indicator matrix \mathbf{F} , each element F_{ij} indicates the relationship between the i th sample and the j th cluster, which is nonnegative by nature. Unfortunately, the optimal \mathbf{F} of the problem (2.7) has mixed signs, which violates its definition. Moreover, the mixed signs make it difficult to get the cluster labels. Discrete process, such as spectral rotation or Kmeans, is performed in previous works to obtain the cluster labels. However, our work is a one-step model and contains no discrete process, which makes the learned \mathbf{F} severely deviate from the ideal cluster indicators. To address this problem, it is natural and reasonable to impose nonnegative constraints on \mathbf{F} . When both nonnegative and orthogonal constraints are satisfied, only one element in each row of \mathbf{F} is greater than zero and all of the others are zeros, which makes the results more appropriate for clustering. Note that if there exists one row with at least two positive elements, \mathbf{F} cannot satisfy the orthogonality constraint because it results in positive nondiagonal elements in $\mathbf{F}^T \mathbf{F}$. Let us assume that there are m ($m \geq 2$) positive elements in the i th row of \mathbf{F} : $\{F_{ik_1}, \dots, F_{ik_m}\}$. When j and l are within $\{k_1, \dots, k_m\}$, and $j \neq l$, we obtain:

$$(\mathbf{F}^T \mathbf{F})_{jl} = \sum_{q=1}^n F_{qj} F_{ql} \geq F_{ij} F_{il} > 0, \quad (2.8)$$

which conflicts the orthogonality condition. Because of this characteristic, the learned \mathbf{F} is more accurate, and more capable to provide discriminative information. Therefore, we rewrite (2.7) as

$$\begin{aligned} \min_{\mathbf{F}, h} \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \sum_{i=1}^c \left(\alpha \sum_{j=1}^n l(h_i(\mathbf{x}_j), \mathbf{f}_i) + \Omega(h_i) \right) \\ \text{s.t. } \mathbf{F}^T \mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0. \end{aligned} \quad (2.9)$$

It is worth noting that we adopt \mathbf{L} defined in (2.5) for simplicity while other sophisticated Laplacian matrices can be used as well.

2.3.2 Sparse Structural Analysis

In CGSSL, the features which are most discriminative to the pseudo class labels are selected. To this end, we adopt a linear model to predict the pseudo labels. Since the features are correlated to jointly reflect the semantic components that can represent some semantic meaning, we propose to exploit feature combinations as well as the original features for the pseudo label prediction. Motivated by [4, 20], the semantic components are uncovered by a shared structure learning model, which enables to learn a more discriminative predictors to make the learned results more reliable. For simplicity, we assume that the shared structure is a hidden low-dimensional subspace in this work. Therefore, the original data features together with the features in the low-dimensional subspace are both used to predict the pseudo labels of samples:

$$h_i(\mathbf{x}_j) = \mathbf{v}_i^T \mathbf{x}_j + \mathbf{p}_i^T \mathbf{Q}^T \mathbf{x}_j, \quad (2.10)$$

where $\mathbf{v}_i \in \mathbb{R}^d$ and $\mathbf{p}_i \in \mathbb{R}^r$ are the weight vectors, and $\mathbf{Q} \in \mathbb{R}^{d \times r}$ is the linear transformation to parameterize the shared r -dimensional subspace. To make the problem tractable, the orthogonal constraint $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_r$ is imposed. Denote $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_c] \in \mathbb{R}^{d \times c}$ and $\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_c] \in \mathbb{R}^{r \times c}$. Thus, we have

$$\begin{aligned} & \sum_{i=1}^c \left(\sum_{j=1}^n l(h_i(\mathbf{x}_j), \mathbf{f}_j) + \Omega(h_i) \right) \\ &= l((\mathbf{V} + \mathbf{QP})^T \mathbf{X}, \mathbf{F}) + \Omega(\mathbf{V}, \mathbf{P}). \end{aligned} \quad (2.11)$$

By defining $\mathbf{W} = \mathbf{V} + \mathbf{QP}$ and combining (2.9) and (2.11), our formulation becomes

$$\begin{aligned} & \min_{\mathbf{V}, \mathbf{W}, \mathbf{Q}, \mathbf{F}} \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha l(\mathbf{W}^T \mathbf{X}, \mathbf{F}) + \Omega(\mathbf{V}, \mathbf{W}) \\ & \text{s.t. } \mathbf{F}^T \mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0; \mathbf{Q}^T \mathbf{Q} = \mathbf{I}_r. \end{aligned} \quad (2.12)$$

To solve the optimization problem in (2.12), we first decide which loss function is chosen for $l(\cdot, \cdot)$ and which regularization functions used for Ω . In this work, we utilize the least square loss $l(x, y) = (x - y)^2$ for simplicity. For \mathbf{V} , the quadratic regularization is used, that is, $\|\mathbf{V}\|_F^2 = \|\mathbf{W} - \mathbf{QP}\|_F^2$. To achieve feature selection across all samples, $\ell_{2,1}$ -norm regularization is adopted for \mathbf{W} to guarantee that \mathbf{W} is sparse in rows. So we have

$$\begin{aligned} & \min_{\mathbf{P}, \mathbf{W}, \mathbf{Q}, \mathbf{F}} \mathcal{O} = \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha \|\mathbf{F} - \mathbf{X}^T \mathbf{W}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} + \gamma \|\mathbf{W} - \mathbf{QP}\|_F^2 \\ & \text{s.t. } \mathbf{F}^T \mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0; \mathbf{Q}^T \mathbf{Q} = \mathbf{I}_r. \end{aligned} \quad (2.13)$$

β and γ are two regularization parameters. The joint minimization of the regression model and $\ell_{2,1}$ -norm regularization term enables \mathbf{W} to evaluate the correlation between pseudo labels and features, making it particularly suitable for feature selection. More specifically, \mathbf{w}_i , the i th row of \mathbf{W} , shrinks to zero if the i th feature is less discriminative to the pseudo labels \mathbf{F} . Once \mathbf{W} is learned, we can select the top p ranked features by sorting all d features according to $\|\mathbf{w}_i\|_2$ ($i = 1, \dots, d$) in descending order. Therefore, the features corresponding to zero rows of \mathbf{W} will be discarded when performing feature selection.

2.3.3 Optimization

The optimization problem (2.13) involves the $\ell_{2,1}$ -norm which is nonsmooth and cannot have a closed form solution. Consequently, we propose an iterative optimization algorithm.

We can see that the optimal \mathbf{P} in the optimization problem (2.13) can be expressed in terms of \mathbf{W} and \mathbf{Q} . By setting the derivative $\partial \mathcal{O} / \partial \mathbf{P} = 0$, we obtain

$$2\gamma(\mathbf{Q}^T \mathbf{Q} \mathbf{P} - \mathbf{Q}^T \mathbf{W}) = 0 \Rightarrow \mathbf{P} = \mathbf{Q}^T \mathbf{W}. \quad (2.14)$$

Because we have the property that $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_r$.

Now, by substituting \mathbf{P} in \mathcal{O} with Eq. (2.14), the objective function \mathcal{O} is written as follows:

$$\begin{aligned} \mathcal{O} &= \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha \|\mathbf{F} - \mathbf{X}^T \mathbf{W}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} \\ &\quad + \gamma \text{Tr}[(\mathbf{W} - \mathbf{Q} \mathbf{Q}^T \mathbf{W})^T (\mathbf{W} - \mathbf{Q} \mathbf{Q}^T \mathbf{W})] \\ &= \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha \|\mathbf{F} - \mathbf{X}^T \mathbf{W}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} \\ &\quad + \gamma \text{Tr}[\mathbf{W}^T (\mathbf{I}_d - \mathbf{Q} \mathbf{Q}^T) (\mathbf{I}_d - \mathbf{Q} \mathbf{Q}^T) \mathbf{W}] \end{aligned} \quad (2.15)$$

Since $(\mathbf{I}_d - \mathbf{Q} \mathbf{Q}^T)(\mathbf{I}_d - \mathbf{Q} \mathbf{Q}^T) = \mathbf{I}_d - \mathbf{Q} \mathbf{Q}^T$, by setting the derivative $\partial \mathcal{O} / \partial \mathbf{W} = 0$, we get

$$\begin{aligned} &\alpha \mathbf{X}(\mathbf{X}^T \mathbf{W} - \mathbf{F}) + \beta \mathbf{D} \mathbf{W} + \gamma (\mathbf{I}_d - \mathbf{Q} \mathbf{Q}^T) \mathbf{W} = 0 \\ \Leftrightarrow &(\alpha \mathbf{X} \mathbf{X}^T + \beta \mathbf{D} + \gamma (\mathbf{I}_d - \mathbf{Q} \mathbf{Q}^T)) \mathbf{W} = \alpha \mathbf{X} \mathbf{F} \\ \Leftrightarrow &\mathbf{W} = \alpha (\mathbf{G} - \gamma \mathbf{Q} \mathbf{Q}^T)^{-1} \mathbf{X} \mathbf{F} \\ \Leftrightarrow &\mathbf{W} = \alpha \mathbf{H}^{-1} \mathbf{X} \mathbf{F} \end{aligned} \quad (2.16)$$

Here \mathbf{D} is a diagonal matrix with $D_{ii} = \frac{1}{2\|\mathbf{w}_i\|_2}$.¹ $\mathbf{G} = \alpha \mathbf{X} \mathbf{X}^T + \beta \mathbf{D} + \gamma \mathbf{I}_d$ and $\mathbf{H} = \mathbf{G} - \gamma \mathbf{Q} \mathbf{Q}^T$.

Owing to $\|\mathbf{A}\|_F^2 = \text{Tr}(\mathbf{A}^T \mathbf{A})$ for any arbitrary matrix \mathbf{A} , we can rewrite Eq. (2.15) as follows:

$$\begin{aligned} \mathcal{O} &= \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha \text{Tr}[(\mathbf{X}^T \mathbf{W} - \mathbf{F})^T (\mathbf{X}^T \mathbf{W} - \mathbf{F})] \\ &\quad + \beta \|\mathbf{W}\|_{2,1} + \gamma \text{Tr}[\mathbf{W}^T (\mathbf{I}_d - \mathbf{Q} \mathbf{Q}^T) \mathbf{W}] \\ &= \text{Tr}[\mathbf{W}^T (\alpha \mathbf{X} \mathbf{X}^T + \beta \mathbf{D} + \gamma \mathbf{I}_d - \gamma \mathbf{Q} \mathbf{Q}^T) \mathbf{W}] \\ &\quad - 2\alpha \text{Tr}[\mathbf{W}^T \mathbf{X} \mathbf{F}] + \alpha \text{Tr}[\mathbf{F}^T \mathbf{F}] + \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] \\ &= \text{Tr}[\mathbf{W}^T (\mathbf{H} \mathbf{W} - 2\alpha \mathbf{X} \mathbf{F})] + \text{Tr}[\mathbf{F}^T (\alpha \mathbf{I}_n + \mathbf{L}) \mathbf{F}] \end{aligned} \quad (2.17)$$

By substituting the expression for \mathbf{W} in Eq. (2.16) into Eq. (2.17), since $\mathbf{H} = \mathbf{H}^T$, we obtain the following equation:

¹In practice, $\|\mathbf{w}_i\|_2$ could be close to zero but not zero. Theoretically, it could be zeros. For this case, we can regularize $D_{ii} = \frac{1}{2\sqrt{(\mathbf{w}_i^T \mathbf{w}_i + \varepsilon)}}$, where ε is very small constant. When $\varepsilon \rightarrow 0$, we can

see that $\frac{1}{2\sqrt{(\mathbf{w}_i^T \mathbf{w}_i + \varepsilon)}}$ approximates $\frac{1}{2\sqrt{\mathbf{w}_i^T \mathbf{w}_i}}$.

$$\begin{aligned}
\mathcal{O} &= \text{Tr}[\alpha^2 \mathbf{F}^T \mathbf{X}^T (\mathbf{H}^{-1} \mathbf{H} \mathbf{H}^{-1} - 2\mathbf{H}^{-1}) \mathbf{X} \mathbf{F}] + \text{Tr}[\mathbf{F}^T (\alpha \mathbf{I}_n + \mathbf{L}) \mathbf{F}] \\
&= \text{Tr}[\mathbf{F}^T (\alpha \mathbf{I}_n + \mathbf{L}) \mathbf{F}] - \alpha^2 \text{Tr}[\mathbf{F}^T \mathbf{X}^T \mathbf{H}^{-1} \mathbf{X} \mathbf{F}]
\end{aligned} \tag{2.18}$$

By substituting Eq. (2.18) into the problem (2.13), we have the following optimization problem *w.r.t.* \mathbf{Q} :

$$\max_{\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_r} \text{Tr}[\mathbf{F}^T \mathbf{X}^T \mathbf{H}^{-1} \mathbf{X} \mathbf{F}] \tag{2.19}$$

To compute the matrix inverse, using the Sherman–Morrison–Woodbury formula [18]: $(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{U} (\mathbf{C}^{-1} + \mathbf{VA}^{-1} \mathbf{U})^{-1} \mathbf{VA}^{-1}$, we have

$$\begin{aligned}
\mathbf{H}^{-1} &= (\mathbf{G} - \gamma \mathbf{Q} \mathbf{Q}^T)^{-1} \\
&= \mathbf{G}^{-1} + \gamma \mathbf{G}^{-1} \mathbf{Q} (\mathbf{I}_r - \gamma \mathbf{Q}^T \mathbf{G}^{-1} \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{G}^{-1}.
\end{aligned} \tag{2.20}$$

Thus, by using the property that $\text{Tr}[\mathbf{AB}] = \text{Tr}[\mathbf{BA}]$ for any arbitrary matrices \mathbf{A} and \mathbf{B} , the optimization problem (2.19) is equivalent to

$$\begin{aligned}
&\max \text{Tr}[\mathbf{F}^T \mathbf{X}^T \mathbf{G}^{-1} \mathbf{Q} (\mathbf{I}_r - \gamma \mathbf{Q}^T \mathbf{G}^{-1} \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{G}^{-1} \mathbf{X} \mathbf{F}] \\
&\Leftrightarrow \max \text{Tr}[(\mathbf{I}_r - \gamma \mathbf{Q}^T \mathbf{G}^{-1} \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{G}^{-1} \mathbf{X} \mathbf{F} \mathbf{F}^T \mathbf{X}^T \mathbf{G}^{-1} \mathbf{Q}] \\
&\Leftrightarrow \max \text{Tr}[(\mathbf{Q}^T (\mathbf{I}_d - \gamma \mathbf{G}^{-1}) \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{T} \mathbf{Q}] \\
&\Leftrightarrow \max \text{Tr}[\mathbf{Q}^T \mathbf{N}^{-1} \mathbf{T} \mathbf{Q}] \\
&\text{s.t. } \mathbf{Q}^T \mathbf{Q} = \mathbf{I}_r,
\end{aligned} \tag{2.21}$$

where $\mathbf{T} = \mathbf{G}^{-1} \mathbf{X} \mathbf{F} \mathbf{F}^T \mathbf{X}^T \mathbf{G}^{-1}$ and $\mathbf{N} = \mathbf{I}_d - \gamma \mathbf{G}^{-1}$. Note that \mathbf{N} is positive definite [20], thus \mathbf{Q} can be easily obtained by the eigen-decomposition of $\mathbf{N}^{-1} \mathbf{T}$.

Substituting the expression for \mathcal{O} in Eq.(2.18) into Eq.(2.13), we obtain the following optimization problem *w.s.t.* \mathbf{F} .

$$\begin{aligned}
&\min_{\mathbf{F}} \text{Tr}[\mathbf{F}^T (\mathbf{L} + \alpha \mathbf{I}_n - \alpha^2 \mathbf{X}^T \mathbf{H}^{-1} \mathbf{X}) \mathbf{F}] \\
&\text{s.t. } \mathbf{F}^T \mathbf{F} = \mathbf{I}_c; \mathbf{F} \geq 0
\end{aligned} \tag{2.22}$$

Then we relax the orthogonal constraint and rewrite the above optimization problem as follows:

$$\min_{\mathbf{F} \geq 0} \text{Tr}[\mathbf{F}^T \mathbf{M} \mathbf{F}] + \frac{\lambda}{2} \|\mathbf{F}^T \mathbf{F} - \mathbf{I}_c\|_F^2. \tag{2.23}$$

Here $\mathbf{M} = \mathbf{L} + \alpha \mathbf{I}_n - \alpha^2 \mathbf{X}^T \mathbf{H}^{-1} \mathbf{X}$ and $\lambda > 0$ is a parameter to control the orthogonality condition. In practice, λ should be large enough to insure the orthogonality satisfied. Let ϕ_{ij} be the Lagrange multiplier for constraint $F_{ij} \geq 0$ and $\Phi = [\phi_{ij}]$. Since $\|\mathbf{A}\|_F^2 = \text{Tr}(\mathbf{A}^T \mathbf{A})$, the Lagrange function is

Algorithm 1 CGSSL for Feature Selection**Input:**

Data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$; Parameters $\alpha, \beta, \gamma, \lambda, k, c, r$ and p

- 1: Construct the k -nearest neighbor graph and calculate \mathbf{L} ;
- 2: The iteration step $t = 0$; Initialize $\mathbf{F}_0 \in \mathbb{R}^{n \times c}$ and set $\mathbf{D}_0 \in \mathbb{R}^{d \times d}$ as an identity matrix;
- 3: **repeat**
- 4: $\mathbf{G}_t = \alpha \mathbf{X} \mathbf{X}^T + \beta \mathbf{D}_t + \gamma \mathbf{I}_d$;
- 5: $\mathbf{N}_t = \mathbf{I}_d - \gamma \mathbf{G}_t^{-1}$;
- 6: $\mathbf{T}_t = \mathbf{G}_t^{-1} \mathbf{X} \mathbf{F}_t \mathbf{F}_t^T \mathbf{X}^T \mathbf{G}_t^{-1}$;
- 7: Obtain \mathbf{Q}_{t+1} by the eigen-decomposition of $\mathbf{N}_t^{-1} \mathbf{T}_t$;
- 8: $\mathbf{H}_t = \mathbf{G}_t - \gamma \mathbf{Q}_{t+1} \mathbf{Q}_{t+1}^T$;
- 9: $\mathbf{M}_t = \mathbf{L} + \alpha \mathbf{I}_n - \alpha^2 \mathbf{X}^T \mathbf{H}_t^{-1} \mathbf{X}$;
- 10: $(F_{t+1})_{ij} = (F_t)_{ij} \frac{(\lambda \mathbf{F}_t)_{ij}}{(\mathbf{M}_t \mathbf{F}_t + \lambda \mathbf{F}_t \mathbf{F}_t^T \mathbf{F}_t)_{ij}}$;
- 11: $\mathbf{W}_{t+1} = \mathbf{H}_t^{-1} \mathbf{X} \mathbf{F}_{t+1}$;
- 12: Update the diagonal matrix \mathbf{D} as

$$\mathbf{D}_{t+1} = \begin{bmatrix} \frac{1}{2\|(\mathbf{w}_{t+1})_1\|_2} & & \\ & \dots & \\ & & \frac{1}{2\|(\mathbf{w}_{t+1})_d\|_2} \end{bmatrix};$$
- 13: $t=t+1$;
- 14: **until** Convergence criterion satisfied

Output:

Sort all d features according to $\|(\mathbf{w}_t)_i\|_2$ in descending order and select the top p ranked features.

$$\text{Tr}[\mathbf{F}^T \mathbf{M} \mathbf{F}] + \frac{\lambda}{2} \text{Tr}[(\mathbf{F}^T \mathbf{F} - \mathbf{I}_c)^T (\mathbf{F}^T \mathbf{F} - \mathbf{I}_c)] + \text{Tr}[\Phi \mathbf{F}^T]. \quad (2.24)$$

Setting its derivative with respect to \mathbf{F} to 0, we have

$$2\mathbf{M} \mathbf{F} + 2\lambda \mathbf{F}(\mathbf{F}^T \mathbf{F} - \mathbf{I}_c) + \Phi = 0. \quad (2.25)$$

Using the Karush–Kuhn–Tuckre (KKT) condition [22] $\phi_{ij} F_{ij} = 0$, we obtain the updating rules:

$$\begin{aligned} 2[\mathbf{M} \mathbf{F} + \lambda \mathbf{F}(\mathbf{F}^T \mathbf{F} - \mathbf{I}_c)]_{ij} F_{ij} + \Phi_{ij} F_{ij} &= 0 \\ \Rightarrow F_{ij} &\leftarrow F_{ij} \frac{(\lambda \mathbf{F})_{ij}}{(\mathbf{M} \mathbf{F} + \lambda \mathbf{F} \mathbf{F}^T \mathbf{F})_{ij}}. \end{aligned} \quad (2.26)$$

Then we normalize \mathbf{F} with $(\mathbf{F}^T \mathbf{F})_{ii} = 1, i = 1, \dots, c$.

From the above analysis, we can see that \mathbf{D} related to \mathbf{W} is required to solve \mathbf{Q} and \mathbf{F} and it is still not straightforward to obtain \mathbf{W} , \mathbf{Q} , and \mathbf{F} . To this end, we design an iterative algorithm to solve the proposed formulation, which is summarized in Algorithm 1.

Now, we briefly analyze the computational complexity. In our case, $c \ll n, c \ll d$ and $r < d$. The complexity of calculating the inverse of a few matrices is $O(d^3)$ and the eigen-decomposition of $\mathbf{N}^{-1} \mathbf{T}$ also needs $O(d^3)$ in complexity. In each

iteration step, the cost for updating \mathbf{Q} is $O(d^3 + nd^2)$. It takes $O(d^3 + nd^2 + dn^2)$ to update \mathbf{F} and $O(d^3)$ to update \mathbf{W} , respectively. Thus the overall cost for CGSSL is $O(t(d^3 + nd^2 + dn^2))$, where t is the number of iterations.

2.3.4 Convergence Analysis

The proposed iterative procedure in Algorithm 1 can be verified to converge to the optimal solutions by the following theorem.

Theorem 2.1 *The alternative updating rules in Algorithm 1 monotonically decrease the objective function value of (2.13) in each iteration.*

Proof For convenience, let us denote

$$\begin{aligned} \mathcal{L}(\mathbf{Q}, \mathbf{F}, \mathbf{W}) = & \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha \|\mathbf{F} - \mathbf{X}^T \mathbf{W}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} \\ & + \gamma \|\mathbf{W} - \mathbf{Q} \mathbf{Q}^T \mathbf{W}\|_F^2 + \frac{\lambda}{2} \|\mathbf{F}^T \mathbf{F} - \mathbf{I}_c\|_F^2 \end{aligned} \quad (2.27)$$

From the above analysis, the problem (2.13) can be relaxed into the following problem:

$$\min_{\mathbf{Q}^T \mathbf{Q} = \mathbf{I}, \mathbf{F} \geq 0, \mathbf{W}} \mathcal{L}(\mathbf{Q}, \mathbf{F}, \mathbf{W}) \quad (2.28)$$

With \mathbf{F}_t and \mathbf{W}_t fixed, we can see that

$$\begin{aligned} \mathbf{Q}_{t+1} = & \arg \max_{\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_t} \text{Tr}[\mathbf{Q}^T \mathbf{N}_t^{-1} \mathbf{T}_t \mathbf{Q}] \\ \Rightarrow & \text{Tr}[\mathbf{Q}_{t+1}^T \mathbf{N}_t^{-1} \mathbf{T}_t \mathbf{Q}_{t+1}] \geq \text{Tr}[\mathbf{Q}_t^T \mathbf{N}_t^{-1} \mathbf{T}_t \mathbf{Q}_t]. \end{aligned} \quad (2.29)$$

Thus we obtain

$$\mathcal{L}(\mathbf{Q}_{t+1}, \mathbf{F}_t, \mathbf{W}_t) \leq \mathcal{L}(\mathbf{Q}_t, \mathbf{F}_t, \mathbf{W}_t). \quad (2.30)$$

With \mathbf{W}_t and \mathbf{Q}_{t+1} fixed, by introducing an auxiliary function of $\mathcal{L}(\mathbf{Q}_{t+1}, \mathbf{F}_t, \mathbf{W})$ as in [24], it is easy to prove

$$\mathcal{L}(\mathbf{Q}_{t+1}, \mathbf{F}_{t+1}, \mathbf{W}_t) \leq \mathcal{L}(\mathbf{Q}_{t+1}, \mathbf{F}_t, \mathbf{W}_t). \quad (2.31)$$

It can easily verified that Eq. (2.16) is the solution to the following problem with \mathbf{Q}_{t+1} and \mathbf{F}_{t+1} fixed.

$$\begin{aligned} \min_{\mathbf{W}} & \alpha \|\mathbf{X}^T \mathbf{W} - \mathbf{F}_{t+1}\|_F^2 + \beta \text{Tr}[\mathbf{W}^T \mathbf{D}_t \mathbf{W}] \\ & + \gamma \|\mathbf{W} - \mathbf{Q}_{t+1} \mathbf{Q}_{t+1}^T \mathbf{W}\|_F^2 \end{aligned} \quad (2.32)$$

For the ease of representation, let us define $g(\mathbf{W}) = \alpha \|\mathbf{X}^T \mathbf{W} - \mathbf{F}_{t+1}\|_F^2 + \gamma \|\mathbf{W} - \mathbf{Q}_{t+1} \mathbf{P}_{t+1}\|_F^2$. Accordingly, in the t th iteration, we have

$$\begin{aligned}
\mathbf{W}^{t+1} &= \arg \min_{\mathbf{W}} g(\mathbf{W}) + \beta \text{Tr}[\mathbf{W}^T \mathbf{D}_t \mathbf{W}] \\
&\Rightarrow g(\mathbf{W}_{t+1}) + \beta \text{Tr}[\mathbf{W}_{t+1}^T \mathbf{D}_t \mathbf{W}_{t+1}] \leq g(\mathbf{W}_t) + \beta \text{Tr}[\mathbf{W}_t^T \mathbf{D}_t \mathbf{W}_t] \\
&\Rightarrow g(\mathbf{W}_{t+1}) + \beta \sum_i \frac{\|(\mathbf{w}_{t+1})_i\|_2^2}{2\|(\mathbf{w}_t)_i\|_2} \leq g(\mathbf{W}_t) + \beta \sum_i \frac{\|(\mathbf{w}_t)_i\|_2^2}{2\|(\mathbf{w}_t)_i\|_2} \\
&\Rightarrow g(\mathbf{W}_{t+1}) + \beta \|\mathbf{W}_{t+1}\|_{2,1} - \beta (\|\mathbf{W}_{t+1}\|_{2,1} - \sum_i \frac{\|(\mathbf{w}_{t+1})_i\|_2^2}{2\|(\mathbf{w}_t)_i\|_2}) \\
&\leq g(\mathbf{W}_t) + \beta \|\mathbf{W}_t\|_{2,1} - \beta (\|\mathbf{W}_t\|_{2,1} - \sum_i \frac{\|(\mathbf{w}_t)_i\|_2^2}{2\|(\mathbf{w}_t)_i\|_2}). \tag{2.33}
\end{aligned}$$

According to the Lemmas in [33], $\|\mathbf{W}_{t+1}\|_{2,1} - \sum_i \frac{\|(\mathbf{w}_{t+1})_i\|_2^2}{2\|(\mathbf{w}_t)_i\|_2} \leq \|\mathbf{W}^t\|_{2,1} - \sum_i \frac{\|(\mathbf{w}_t)_i\|_2^2}{2\|(\mathbf{w}_t)_i\|_2}$. Thus,

$$g(\mathbf{W}_{t+1}) + \beta \|\mathbf{W}^{t+1}\|_{2,1} \leq g(\mathbf{W}_t) + \beta \|\mathbf{W}^t\|_{2,1}. \tag{2.34}$$

Therefore, we arrive at

$$\mathcal{L}(\mathbf{Q}_{t+1}, \mathbf{F}_{t+1}, \mathbf{W}_{t+1}) \leq \mathcal{L}(\mathbf{Q}_{t+1}, \mathbf{F}_{t+1}, \mathbf{W}_t). \tag{2.35}$$

Based on Eqs. (2.30), (2.31) and (2.35), we obtain

$$\begin{aligned}
\mathcal{L}(\mathbf{Q}_{t+1}, \mathbf{F}_{t+1}, \mathbf{W}_{t+1}) &\leq \mathcal{L}(\mathbf{Q}_{t+1}, \mathbf{F}_{t+1}, \mathbf{W}_t) \\
&\leq \mathcal{L}(\mathbf{Q}_{t+1}, \mathbf{F}_t, \mathbf{W}_t) \leq \mathcal{L}(\mathbf{Q}_t, \mathbf{F}_t, \mathbf{W}_t). \tag{2.36}
\end{aligned}$$

Thus, the objective function monotonically decreases using the updating rules in Algorithm 1 and Theorem 2.1 is proved.

According to Theorem 2.1, we can see that the iterative approach in Algorithm 1 converges to local optimal solutions. The proposed optimization algorithm is efficient. In the experiment, we observe that our algorithm usually converges around only 20 iterations.

2.3.5 Discussions

In this section, we discuss the relationships between the proposed method and several algorithms, including SPFS [48], MCFS [5], UDFS [43], and NDFS [27].

Connection with SPFS: SPFS [48] performs feature selection by preserving sample similarity, which can handle feature redundancy. It is formulated as:

$$\min_{\|\mathbf{W}\|_{2,1} \leq \tau} \sum_{i,j=1}^n (\mathbf{x}_i^T \mathbf{W} \mathbf{W}^T \mathbf{x}_j - S_{ij})^2. \quad (2.37)$$

Here τ ($\tau > 0$) is a hyper-parameter. The connection between SPFS and CGSSL is discovered as follows.

Proposition 2.1 *SPFS has the similar fashion with the proposed CGSSL when $\alpha \rightarrow +\infty$, $\gamma = 0$ and the orthogonal and nonnegative constraints are removed.*

Proof When $\alpha \rightarrow +\infty$, and the orthogonal and nonnegative constraints are removed, we have $\mathbf{F} = \mathbf{X}^T \mathbf{W}$. Then, with $\gamma = 0$ CGSSL becomes

$$\begin{aligned} & \min_{\mathbf{W}} \text{Tr}[\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}] + \beta \|\mathbf{W}\|_{2,1} \\ &= \frac{1}{2} \sum_{i,j=1}^n S_{ij} \|\mathbf{x}_i^T \mathbf{W} - \mathbf{x}_j^T \mathbf{W}\|^2 + \beta \|\mathbf{W}\|_{2,1}. \end{aligned} \quad (2.38)$$

Compared the problems (2.37) with (2.38), they both try to keep data similarity with different criteria. That is, Proposition 2.1 is proved.

Connection with MCFS: MCFS [5] uses a two-step strategy to select features according to spectral analysis and is formulated as the following form.

$$\min_{\mathbf{F}^T \mathbf{F} = \mathbf{I}_c} \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] \quad (2.39)$$

$$\min_{\mathbf{w}_i} \|\mathbf{f}_i - \mathbf{X}^T \mathbf{w}_i\| + \beta \|\mathbf{w}_i\|_1 \quad (2.40)$$

$\|\mathbf{w}_i\|_1$ is the ℓ_1 -norm of \mathbf{w}_i .

Proposition 2.2 *MCFS and CGSSL have similar fashions with different regularization forms on \mathbf{W} , when $\gamma = 0$ and the nonnegative constraint is removed.*

Proof When $\gamma = 0$ and the nonnegative constraint removed, the proposed formulation becomes

$$\min_{\mathbf{W}, \mathbf{F}^T \mathbf{F} = \mathbf{I}_c} \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha \|\mathbf{F} - \mathbf{X}^T \mathbf{W}\| + \beta \|\mathbf{W}\|_{2,1}. \quad (2.41)$$

If we set $\alpha \rightarrow 0$ and $\beta \rightarrow 0$, the above problem leads to a two-step algorithm. That is,

$$\min_{\mathbf{F}^T \mathbf{F} = \mathbf{I}_c} \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] \quad (2.42)$$

$$\min_{\mathbf{W}} \sum_{i=1}^c \|\mathbf{f}_i - \mathbf{X}^T \mathbf{w}_i\| + \frac{\beta}{\alpha} \|\mathbf{w}_i\|_2. \quad (2.43)$$

We can see that the regularization function for \mathbf{w}_i is different. Thus, Proposition 2.2 is proved.

Different from MCFS, CGSSL is an one-step algorithm. Thus, CGSSL is more general. Second, \mathbf{F} is constrained to be nonnegative. When both nonnegative and orthogonal constraints are satisfied, the learned \mathbf{F} is much closer to the ideal result, and the solution can be directly obtained without discretization. Finally, we perform clustering and feature selection simultaneously, which explicitly enforces that \mathbf{F} can be linearly approximated by the selected features, making the results more accurate.

Connection with UDFS: UDFS [43] was proposed to select discriminative features by optimizing the following objective function

$$\min_{\mathbf{W}^T \mathbf{W} = \mathbf{I}_c} \text{Tr}[\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}] + \beta \|\mathbf{W}\|_{2,1}. \quad (2.44)$$

Proposition 2.3 *UDFS and CGSSL have similar fashions when $\alpha \rightarrow +\infty$, $\gamma = 0$ and the nonnegative constraint is removed.*

Proof With $\alpha \rightarrow +\infty$ and the nonnegative constraint removed, we have $\mathbf{F} = \mathbf{X}^T \mathbf{W}$. Then when $\gamma = 0$, the proposed CGSSL formulation becomes

$$\min_{\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W} = \mathbf{I}_c} \text{Tr}[\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}] + \beta \|\mathbf{W}\|_{2,1}. \quad (2.45)$$

Therefore, UDFS and CGSSL have similar fashions with different orthogonal constraints.

In this extreme case, \mathbf{F} is enforced to be linear, i.e., $\mathbf{F} = \mathbf{X}^T \mathbf{W}$. However, as indicated in [37], it is likely that \mathbf{F} is nonlinear in many applications. Hence, CGSSL is superior to UDFS due to its flexibility of linearity. Additionally, \mathbf{F} is constrained to be nonnegative, making it more accurate than the one with mixed signs. Therefore, CGSSL is more capable to select discriminative features, verified by our experiments.

Connection with NDFS: NDFS [27] is our preliminary version, which does not exploit the underlying structure. Its formulation is presented as follows

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{F}} \quad & \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha \|\mathbf{X}^T \mathbf{W} - \mathbf{F}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{F}^T \mathbf{F} = \mathbf{I}, \mathbf{F} \geq 0. \end{aligned} \quad (2.46)$$

By setting $\gamma = 0$, the problem (2.13) leads to the above problem. Thus we have the following proposition.

Proposition 2.4 *NDFS is a special case of the proposed CGSSL algorithm, when $\lambda = 0$.*

2.3.6 Experiments

In this section, we evaluate the performance of the proposed formulation, which can be applied to many applications, such as clustering and classification. Following previous unsupervised feature selection work [5, 43], we only evaluate the performance of CGSSL for feature selection and compared with representative algorithms in terms of clustering. In our experiments, we first select the top p features and then utilize Kmeans algorithm to cluster samples based on the selected features.

2.3.6.1 Data Sets

The experiments are conducted on 12 publicly available datasets, including three face image data sets (i.e., UMIST [1], JAFFE [29], and Pointing4 [16]), three handwritten digit image data sets [i.e., a subset of MNIST used in [44], Binary Alphabet (BA) [1] and a subset of USPS with 40 samples randomly selected for each class [1]], three text data sets (i.e., WebKB collected by the University of Texas [11], tr11 [2], and oh15 [2]), and three biomedical data sets (i.e., Tox-171 [3], Tumors9 [3], and Leukemia1 [3]). Data sets from different areas serve as a good test bed for a comprehensive evaluation. Table 2.1 summarizes the details of these 12 data sets used in experiments.

Table 2.1 Dataset description

Domain	Dataset	n	d	c
Face images	UMIST	575	644	20
	JAFFE	213	676	10
	Pointing4	2790	1120	15
Handwritten digits images	MNIST	5000	784	10
	BA	1404	320	36
	USPS	400	256	10
Text data	WebKB	814	4029	7
	tr11	414	6429	9
	oh15	913	3100	10
Biomedical data	TOX-171	171	5748	4
	Tumors9	60	5726	9
	Leukemia1	72	5327	3

2.3.6.2 Compared Scheme

To validate the effectiveness of CGSSL for feature selection, we compare it with one baseline and several unsupervised feature selection methods. The compared algorithms are enumerated as follows:

1. **Baseline**: All original features are adopted;
2. **MaxVar**: Features corresponding to the maximum variance are selected to obtain the best expressive features;
3. **LS** [17]: Features consistent with Gaussian Laplacian matrix are selected to best preserve the local manifold structure [17];
4. **SPEC** [47]: Features are selected using spectral regression;
5. **SPFS-SFS** [48]: The traditional forward search strategy is utilized for similarity preserving feature selection in the SPFS framework.
6. **MCFS** [5]: Features are selected based on spectral analysis and sparse regression problem;
7. **UDFS** [43]: Features are selected by a joint framework of discriminative analysis and $\ell_{2,1}$ -norm minimization.
8. **CGSSL**: The proposed Cluster-Guided Sparse Structural Learning method.

2.3.6.3 Parameter Setting

There are some parameters to be set in advance. For LS, SPEC, MCFS, UDFS and CGSSL, we set $k = 5$ for all the datasets to specify the size of neighborhoods. For CGSSL, to guarantee the orthogonality satisfied, we fix $\lambda = 10^8$ in our experiments. To fairly compare different unsupervised feature selection algorithms, we tune the parameters for all methods by a “grid-search” strategy from $\{10^{-8}, 10^{-6}, \dots, 10^8\}$. The dimensionality of the low-dimensional space is set $r = \min(5 \times \max(\lfloor \frac{c-1}{5} \rfloor, 1), c - 1)$ in the experiments since the performance is not very sensitive to it. The numbers of selected features are set as $\{50, 100, 150, 200, 250, 300\}$ for all the datasets except USPS. Because the total feature number of USPS is 256, we set the number of selected features as $\{50, 80, 110, 140, 170, 200\}$. For all the algorithms, we report the the best clustering results from the optimal parameters. Different parameters may be used for different databases. In our experiments, we adopt Kmeans algorithm to cluster samples based on the selected features. The performance of Kmeans clustering depends on initialization. Following [5, 43], we repeat the clustering 20 times with random initialization for each setup. The average results with standard deviation (std) are reported. In real applications, it is impossible to tune parameters using the “grid-search” strategy. But it is an acceptable method to tune parameters for experimental comparisons since all the compared methods are with the well-chosen parameter values. The parameter sensitivity study and convergence study for CGSSL will be shown in the following subsection.

2.3.6.4 Evaluation Metrics

With the selected features, we evaluate the performance in terms of clustering by two widely used evaluation metrics, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI). The larger ACC and NMI are, the better performance is. ACC is defined by

$$\text{ACC} = \frac{1}{n} \sum_{i=1}^n \delta(c_i, \text{map}(g_i)), \quad (2.47)$$

where c_i is the clustering label and g_i is the ground truth label of \mathbf{x}_i . $\text{map}(g_i)$ is the optimal mapping function that permutes clustering labels and the ground truth labels. The optimal mapping can be obtained by using the Kuhn–Munkres algorithm. $\delta(c_i, g_i)$ is an indicator function that equals to 1 if $c_i = g_i$ and equals to 0 otherwise. NMI is defined as

$$\text{NMI} = \frac{\sum_{l,h=1}^c t_{l,h} \log(\frac{n \times t_{l,h}}{t_l \hat{t}_h})}{\sqrt{(\sum_{l=1}^c t_l \log \frac{t_l}{n}) (\sum_{h=1}^c \hat{t}_h \log \frac{\hat{t}_h}{n})}}, \quad (2.48)$$

where t_l is the number of samples in the l th cluster \mathcal{C}_l according to clustering results and \hat{t}_h is the number of samples in the h th ground truth class \mathcal{G}_h . $t_{l,h}$ is the number of overlap between \mathcal{C}_l and \mathcal{G}_h .

2.3.6.5 Performance Comparison

We now empirically evaluate the performance of these nine feature selection algorithms in terms of ACC and NMI. The detailed results on the face, handwritten digit, text, and biomedical data sets are summarized in Tables 2.2, 2.3, 2.4, and 2.5, respectively. The results demonstrate that compared to the compared algorithms, CGSSL achieves the best performance on all the 12 data sets, which validates its effectiveness.

From the above four tables, we have the following observations. (1) Compared with the baseline, it can be observed that feature selection is necessary and effective by removing the noise and redundancy. It can not only reduce the number of features and make the algorithms more efficient, but also enhance the performance. (2) It is better to perform feature selection jointly. The joint feature selection algorithms, such as MCFS, UDFS, and CGSSL are always superior to the methods selecting features one after another, such as MaxVar and SPEC. (3) By utilizing the local geometric structure of data distribution, LS, SPEC, MCFS, UDFS, and CGSSL usually yield superior performance. (4) MCFS, UDFS, and CGSSL achieve more accurate clustering performance by exploiting discriminative information, which demonstrates that it is crucial to uncover the discriminative information in the unsupervised case. (5) CGSSL outperforms SPEC and MCFS. SPEC and MCFS adopt a two-step approach to introduce spectral analysis into feature selection while CGSSL is an one-step

Table 2.2 Clustering results comparison on the face data sets. The best results are highlighted in bold

Dataset	Baseline	Max Var	LS	SPFS-SFS	SPEC	MCFS	UDFS	CGSSL
ACC \pm std (%)								
UMIST	41.8 \pm 2.7	45.8 \pm 2.8	45.9 \pm 2.9	44.8 \pm 3.5	47.9 \pm 3.0	46.3 \pm 3.6	48.6 \pm 3.7	53.4 \pm 3.1
JAFPE	72.5 \pm 9.2	67.3 \pm 5.8	74.0 \pm 7.6	74.6 \pm 9.4	76.9 \pm 7.2	78.8 \pm 9.1	76.7 \pm 7.1	82.3 \pm 7.5
Pointing4	35.9 \pm 2.2	44.0 \pm 2.8	37.1 \pm 1.6	37.4 \pm 1.3	38.6 \pm 2.2	46.2 \pm 2.9	45.1 \pm 2.4	51.1 \pm 2.6
NMI \pm std (%)								
UMIST	62.3 \pm 2.3	63.5 \pm 1.5	63.9 \pm 1.8	63.2 \pm 3.3	65.2 \pm 2.0	66.7 \pm 1.9	67.3 \pm 3.0	70.9 \pm 2.2
JAFPE	80.0 \pm 5.7	70.3 \pm 4.2	79.4 \pm 7.0	82.1 \pm 4.9	82.8 \pm 3.8	83.4 \pm 5.0	82.3 \pm 6.5	87.5 \pm 5.1
Pointing4	41.7 \pm 1.4	50.8 \pm 1.8	42.7 \pm 1.2	43.4 \pm 1.4	40.5 \pm 1.0	53.1 \pm 1.1	52.4 \pm 1.7	57.7 \pm 1.3

Table 2.3 Clustering results comparison on the handwritten digit data sets. The best results are highlighted in bold

Dataset	Baseline	Max Var	LS	SPFS-SFS	SPEC	MCFS	UDFS	CGSSL
ACC \pm std (%)								
MNIST	52.2 \pm 5.0	53.3 \pm 2.7	54.3 \pm 4.8	54.6 \pm 3.2	55.6 \pm 5.2	56.5 \pm 4.1	56.6 \pm 4.2	59.3 \pm 3.6
BA	40.3 \pm 2.0	40.7 \pm 1.7	42.1 \pm 1.7	41.6 \pm 1.7	42.2 \pm 2.2	41.5 \pm 1.8	42.7 \pm 1.8	45.1 \pm 1.8
USPS	62.6 \pm 5.3	63.8 \pm 4.3	64.9 \pm 5.1	65.1 \pm 2.9	65.5 \pm 3.8	64.4 \pm 3.1	66.2 \pm 4.7	68.3 \pm 4.5
NMI \pm std (%)								
MNIST	47.8 \pm 2.3	48.6 \pm 1.1	48.6 \pm 2.0	49.1 \pm 1.8	49.7 \pm 2.0	50.0 \pm 1.8	50.8 \pm 1.6	52.8 \pm 1.8
BA	56.5 \pm 1.3	56.9 \pm 1.3	57.3 \pm 0.8	57.2 \pm 1.2	57.9 \pm 1.1	57.5 \pm 0.8	58.1 \pm 1.0	60.0 \pm 1.1
USPS	56.9 \pm 3.1	58.1 \pm 2.7	58.7 \pm 3.0	58.1 \pm 1.9	59.5 \pm 2.1	59.3 \pm 2.9	60.1 \pm 4.3	62.0 \pm 2.8

Table 2.4 Clustering results comparison on the text data sets. The best results are highlighted in bold

Dataset	Baseline	Max Var	LS	SPFS-SFS	SPEC	MCFS	UDFS	CGSSL
ACC \pm std (%)								
WebKB	56.7 \pm 2.7	54.6 \pm 2.8	56.8 \pm 2.9	60.7 \pm 0.1	61.1 \pm 2.8	61.3 \pm 2.3	61.7 \pm 3.2	63.1 \pm 3.0
tr11	30.9 \pm 2.0	30.5 \pm 1.0	31.6 \pm 1.6	30.6 \pm 1.1	35.1 \pm 1.8	36.3 \pm 4.1	35.9 \pm 2.5	41.1 \pm 4.1
oh15	30.4 \pm 3.0	32.9 \pm 3.2	33.8 \pm 2.7	34.0 \pm 3.4	33.8 \pm 2.1	33.8 \pm 2.8	32.9 \pm 2.3	37.2 \pm 2.5
NMI \pm std (%)								
WebKB	11.4 \pm 5.0	17.1 \pm 1.4	10.6 \pm 4.0	11.4 \pm 3.7	17.2 \pm 3.1	17.6 \pm 0.8	18.1 \pm 3.3	19.0 \pm 1.7
tr11	7.0 \pm 1.4	7.6 \pm 1.1	8.0 \pm 2.0	6.2 \pm 0.6	11.5 \pm 2.9	13.5 \pm 3.3	13.7 \pm 1.9	19.5 \pm 3.9
oh15	17.7 \pm 3.0	22.1 \pm 2.7	23.2 \pm 2.8	23.2 \pm 3.0	23.6 \pm 2.2	23.1 \pm 3.0	21.8 \pm 2.0	24.9 \pm 1.8

Table 2.5 Clustering results comparison on the biomedical data sets. The best results are highlighted in bold

Dataset	Baseline	Max Var	LS	SPFS-SFS	SPEC	MCFS	UDFS	CGSSL
ACC \pm std (%)								
TOX-171	40.9 \pm 3.6	41.0 \pm 2.7	41.8 \pm 1.5	41.7 \pm 5.0	44.9 \pm 2.7	42.7 \pm 2.1	45.7 \pm 2.0	48.6 \pm 1.4
Tumors9	40.2 \pm 5.3	41.0 \pm 5.2	42.3 \pm 3.6	42.8 \pm 4.6	41.0 \pm 3.4	41.8 \pm 4.9	42.9 \pm 4.5	46.8 \pm 4.6
Leukemia1	56.7 \pm 9.2	58.5 \pm 11.6	70.2 \pm 7.2	70.3 \pm 6.4	60.2 \pm 4.7	70.8 \pm 10.1	71.2 \pm 10.6	74.7 \pm 8.7
NMI \pm std (%)								
TOX-171	12.7 \pm 4.0	12.4 \pm 2.6	13.5 \pm 0.8	12.2 \pm 5.4	18.1 \pm 1.0	13.4 \pm 2.6	19.2 \pm 2.2	25.7 \pm 2.8
Tumors9	38.2 \pm 4.9	40.2 \pm 2.8	41.9 \pm 3.3	42.0 \pm 3.4	38.7 \pm 2.5	40.3 \pm 6.0	42.2 \pm 4.2	44.4 \pm 3.4
Leukemia1	23.4 \pm 1.4	27.2 \pm 19.3	34.0 \pm 11.1	34.0 \pm 11.7	31.1 \pm 3.3	39.0 \pm 11.8	40.8 \pm 12.0	48.5 \pm 8.9

framework and performs spectral analysis and feature selection simultaneously. (6) CGSSL achieves higher ACC and NMI than MCFS by imposing the nonnegative constraint, which makes the scaled cluster indicators more accurate. In summary, CGSSL achieves best performance on all data sets by exploiting nonnegative spectral analysis and structural learning with $\ell_{2,1}$ -norm regularization simultaneously for feature selection.

2.4 Nonnegative Spectral Analysis and Redundancy Control for Unsupervised Feature Selection

The proposed CGSSL method ignores the redundancy among the features. The select subset may be not compact. To select the discriminative features with controlled redundancy, we propose a new Nonnegative Spectral analysis with Constrained Redundancy (NSCR) method to exploit clustering analysis and explicitly consider the redundancy between features simultaneously. NSCR is formulated as

$$\begin{aligned} \min_{\mathbf{F}, \mathbf{W}} J(\mathbf{F}) + \alpha l(h(\mathbf{W}; \mathbf{X}), \mathbf{F}) + \beta \Omega(\mathbf{W}) + \lambda g(\mathbf{W}) \\ \text{s.t. } \mathbf{F} = \mathbf{Y}(\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}}, \end{aligned} \quad (2.49)$$

where $J(\mathbf{F})$ is a clustering criterion, $l(\cdot, \cdot)$ is the loss function, $h(\cdot)$ is a predictive function, $\Omega(\cdot)$ is a regularization function with row sparsity, and $g(\cdot)$ is a function to control the redundancy. α , β , and λ are three nonnegative trade-off parameter.

2.4.1 The Objective Function

For the clustering criterion $J(\mathbf{F})$, following the above section, we utilize the proposed nonnegative spectral analysis.

$$J(\mathbf{F}) = \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}], \quad \text{s.t. } \mathbf{F}^T \mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0. \quad (2.50)$$

In NSCR, the features which are most discriminative to the cluster indicators are selected. To this end, we assume that there is a linear transformation between features and the cluster indicators and adopt a linear model to predict the cluster indicators. Therefore, we have the following function:

$$h(\mathbf{W}; \mathbf{X}) = \mathbf{X}^T \mathbf{W} \quad (2.51)$$

where $\mathbf{W} = [\mathbf{w}^1, \dots, \mathbf{w}^c] \in \mathbb{R}^{d \times c}$ is the linear transformation matrix to predict the cluster indicators. To learn a more discriminative predictors for more reliable results

and make our method robust to noisy features, we impose a more general and better sparse model on \mathbf{W} . It has been verified by extensive computational studies that ℓ_p -norm ($0 < p < 1$) can lead to sparser solution than using ℓ_1 -norm [7, 8], and $\ell_{2,p}$ -norm based minimization can also achieve a better sparsity solution than $\ell_{2,1}$ -norm [41]. Thus, we introduce a $\ell_{2,p}$ -norm based regularization for Ω to guarantee that \mathbf{W} is sparse in rows. It can discard noisy or indifferent features

$$\Omega(\mathbf{W}) = \sum_{i=1}^d \|\mathbf{w}_i\|_2^p = \|\mathbf{W}\|_{2,p}^p. \quad (2.52)$$

The proposed problem in (2.49) can be rewritten as

$$\min_{\mathbf{F}, \mathbf{W}} J(\mathbf{F}) + \alpha l(\mathbf{X}^T \mathbf{W}, \mathbf{F}) + \beta \|\mathbf{W}\|_{2,p}^p + \lambda g(\mathbf{W}) \text{ s.t. } \mathbf{F}^T \mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0. \quad (2.53)$$

Under the guide of nonnegative spectral clustering, the feature selection matrix with $\ell_{2,p}$ -norm regularization can select necessary features and discard noisy or indifferent features. However, correlated features may be selected simultaneously since currently we do not penalize the proposed method for redundant features. For example, if the i th feature is highly correlated to the j th feature, we do not need to select both of them simultaneously. Toward this end, we introduce a penalty factor $g(\mathbf{W})$ into our feature selection scheme to control the redundancy while selecting features. Many strategies can be used to define the penalty for using redundant features. In this work, we adopt the correlation between features to define $g(\mathbf{W})$.

$$g(\mathbf{W}) = \frac{1}{d(d-1)} \sum_{i=1}^d \|\mathbf{w}_i\|_2 \sum_{j=1, j \neq i}^d \|\mathbf{w}_j\|_2 C_{ij} \quad (2.54)$$

$C_{ij} \geq 0$ is a measure of correlation between the i th feature and the j th feature. $\|\mathbf{w}_i\|_2$ is a weight to measure the importance of the i th feature. The correlation can be measured linearly or nonlinearly. For a linear measure, the Pearsons correlation coefficient between the i th feature and the j th feature can be used. The mutual information between the i th feature and the j th feature can be used to measure the nonlinear correlation. In this work, the mutual information is adopted. If we set $C_{ii} = 0$, we have

$$g(\mathbf{W}) = \frac{1}{d(d-1)} \sum_{i,j=1}^d \|\mathbf{w}_i\|_2 \|\mathbf{w}_j\|_2 C_{ij}. \quad (2.55)$$

The normalized factor $\frac{1}{d(d-1)}$ is used just to make the regularization term independent of the number of features. By taking the redundancy into account, our method can avoid selected many members of a redundant set of features.

By incorporating the nonnegative spectral clustering, sparse prediction model and the redundancy control into a unified framework, we obtain the following optimization problem:

$$\begin{aligned} \min_{\mathbf{F}, \mathbf{W}} \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha l(\mathbf{X}^T \mathbf{W}, \mathbf{F}) + \beta \|\mathbf{W}\|_{2,p}^p + \frac{\lambda}{d(d-1)} \sum_{i,j=1}^d \|\mathbf{w}_i\|_2 \|\mathbf{w}_j\|_2 C_{ij} \\ \text{s.t. } \mathbf{F}^T \mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0. \end{aligned} \quad (2.56)$$

To solve the optimization problem in (2.56), we first decide which loss function is chosen for $l(\cdot, \cdot)$. In this work, we utilize the least square loss $l(x, y) = \frac{1}{2}(x - y)^2$ for simplicity and set $\gamma = \frac{\lambda}{d(d-1)}$. Hence we have

$$\begin{aligned} \min_{\mathbf{F}, \mathbf{W}} \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \frac{\alpha}{2} \|\mathbf{F} - \mathbf{X}^T \mathbf{W}\|_F^2 + \beta \|\mathbf{W}\|_{2,p}^p + \gamma \sum_{i,j=1}^d \|\mathbf{w}_i\|_2 \|\mathbf{w}_j\|_2 C_{ij} \\ \text{s.t. } \mathbf{F}^T \mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0. \end{aligned} \quad (2.57)$$

The joint minimization of the regression model and $\ell_{2,p}$ -norm regularization term enables \mathbf{W} to evaluate the correlation between pseudo labels and features, making it particularly suitable for feature selection. More specifically, \mathbf{w}_i , the i th row of \mathbf{W} , shrinks to zero if the i th feature is less discriminative to the pseudo labels \mathbf{F} . It can guarantee that the necessary features are selected and the noisy or indifferent features are discarded. The consideration of the redundancy can explicitly control the redundancy between the selected features. Once \mathbf{W} is learned, we can select the top r ranked features by sorting all d features according to $\|\mathbf{w}_i\|_2$ ($i = 1, \dots, d$) in descending order. Therefore, the features corresponding to zero rows of \mathbf{W} will be discarded when performing feature selection.

2.4.2 Optimization

Since ℓ_p ($0 < p < 1$) vector norm is neither convex nor Lipschitz continuous, $\ell_{2,p}$ matrix pseudo norm is not convex or Lipschitz continuous yet. The optimization problem (2.57) involves the $\ell_{2,p}$ -norm which is not convex and nonsmooth. Consequently, we propose an iterative optimization algorithm to solve the optimization problem (2.57). For the ease of representation, let us define

$$\mathcal{L}(\mathbf{F}, \mathbf{W}) = \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \frac{\alpha}{2} \|\mathbf{F} - \mathbf{X}^T \mathbf{W}\|_F^2 + \beta \|\mathbf{W}\|_{2,p}^p + \gamma \sum_{i,j=1}^d \|\mathbf{w}_i\|_2 \|\mathbf{w}_j\|_2 C_{ij}. \quad (2.58)$$

By computing the derivative of \mathcal{L} with respect to \mathbf{w}_i ,² we obtain:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} = \alpha \mathbf{X}(\mathbf{X}^T \mathbf{w}_i - \mathbf{f}_i) + \beta \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|_2} + \gamma \frac{\sum_j \|\mathbf{w}_j\|_2 C_{ij}}{\|\mathbf{w}_i\|_2} \mathbf{w}_i. \quad (2.59)$$

The following equation can be easily induced

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \alpha \mathbf{X}(\mathbf{X}^T \mathbf{W} - \mathbf{F}) + \beta \mathbf{D} \mathbf{W} + \gamma \mathbf{H} \mathbf{W}, \quad (2.60)$$

where \mathbf{D} is a diagonal matrix with $D_{ii} = \frac{p}{2\|\mathbf{w}_i\|_2^{2-p}}$ and \mathbf{H} is another diagonal matrix with $H_{ii} = \frac{\sum_j \|\mathbf{w}_j\|_2 C_{ij}}{2\|\mathbf{w}_i\|_2}$. Setting $\frac{\partial \mathcal{L}(\mathbf{F}, \mathbf{W})}{\partial \mathbf{W}} = 0$, we have

$$\begin{aligned} \alpha \mathbf{X}(\mathbf{X}^T \mathbf{W} - \mathbf{F}) + \beta \mathbf{D} \mathbf{W} + \gamma \mathbf{H} \mathbf{W} &= 0 \\ \Rightarrow \mathbf{W} &= \alpha (\alpha \mathbf{X} \mathbf{X}^T + \beta \mathbf{D} + \gamma \mathbf{H})^{-1} \mathbf{X} \mathbf{F} \\ &= \mathbf{G}^{-1} \mathbf{X} \mathbf{F} \end{aligned} \quad (2.61)$$

Here $\mathbf{G} = \mathbf{X} \mathbf{X}^T + \frac{\beta}{\alpha} \mathbf{D} + \frac{\gamma}{\alpha} \mathbf{H}$.

Owing to $\|\mathbf{A}\|_F^2 = \text{Tr}(\mathbf{A}^T \mathbf{A})$ for any arbitrary matrix \mathbf{A} , we can rewrite Eq. (2.58) as follows:

$$\begin{aligned} \mathcal{L} &= \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha \text{Tr}[(\mathbf{X}^T \mathbf{W} - \mathbf{F})^T (\mathbf{X}^T \mathbf{W} - \mathbf{F})] + \beta \text{Tr}[\mathbf{W}^T \mathbf{D} \mathbf{W}] + \gamma \text{Tr}[\mathbf{W}^T \mathbf{H} \mathbf{W}] \\ &= \text{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha \text{Tr}[\mathbf{F}^T \mathbf{F}] - 2\alpha \text{Tr}[\mathbf{W}^T \mathbf{X} \mathbf{F}] + \text{Tr}[\mathbf{W}^T (\alpha \mathbf{X} \mathbf{X}^T + \beta \mathbf{D} + \gamma \mathbf{H}) \mathbf{W}]. \end{aligned} \quad (2.62)$$

By substituting the expression for \mathbf{W} in Eq. (2.61) into the above equation, we have

$$\mathcal{L}(\mathbf{F}) = \text{Tr}[\mathbf{F}^T (\mathbf{L} + \alpha \mathbf{I}_n - \alpha \mathbf{X}^T \mathbf{G}^{-1} \mathbf{X}) \mathbf{F}]. \quad (2.63)$$

Thus, we obtain the following optimization problem *w.s.r.* \mathbf{F}

$$\min_{\mathbf{F}} \text{Tr}[\mathbf{F}^T \mathbf{M} \mathbf{F}] \quad \text{s.t.} \quad \mathbf{F}^T \mathbf{F} = \mathbf{I}_c; \quad \mathbf{F} \geq 0. \quad (2.64)$$

Here $\mathbf{M} = \mathbf{L} + \alpha \mathbf{I}_n - \alpha \mathbf{X}^T \mathbf{G}^{-1} \mathbf{X}$. Then we relax the orthogonal constraint by incorporating the orthogonal constraint of \mathbf{F} into the objective function via Lagrange multiplier and obtain the optimization problem as follows:

$$\min_{\mathbf{F} \geq 0} \text{Tr}[\mathbf{F}^T \mathbf{M} \mathbf{F}] + \frac{\mu}{2} \|\mathbf{F}^T \mathbf{F} - \mathbf{I}_c\|_F^2 \quad (2.65)$$

²In practice, $\|\mathbf{w}_i\|_2$ could be close to zero but not zero. Theoretically, it could be zeros. For this case, we can regularize $\|\mathbf{w}_i\|_2 \leftarrow \|\mathbf{w}_i\|_2 + \varepsilon$, where ε is a very small constant. When $\varepsilon \rightarrow 0$, we can see that $\|\mathbf{w}_i\|_2 + \varepsilon$ approximates $\|\mathbf{w}_i\|_2$.

$\mu > 0$ is a parameter to control the orthogonality condition. In practice, λ should be large enough to insure the orthogonality satisfied. Let ϕ_{ij} be the Lagrange multiplier for constraint $F_{ij} \geq 0$ and $\Phi = [\phi_{ij}]$. The Lagrange function is

$$\text{Tr}[\mathbf{F}^T \mathbf{M} \mathbf{F}] + \frac{\mu}{2} \text{Tr}[(\mathbf{F}^T \mathbf{F} - \mathbf{I}_c)^T (\mathbf{F}^T \mathbf{F} - \mathbf{I}_c)] + \text{Tr}[\Phi \mathbf{F}^T]. \quad (2.66)$$

Setting its derivative with respect to \mathbf{F} to 0, we have

$$2\mathbf{M}\mathbf{F} + 2\mu\mathbf{F}(\mathbf{F}^T \mathbf{F} - \mathbf{I}_c) + \Phi = 0. \quad (2.67)$$

Using the Karush–Kuhn–Tuckre (KKT) condition [22] $\phi_{ij} F_{ij} = 0$, we obtain the updating rules:

$$\begin{aligned} 2[\mathbf{M}\mathbf{F} + \lambda\mathbf{F}(\mathbf{F}^T \mathbf{F} - \mathbf{I}_c)]_{ij} F_{ij} + \Phi_{ij} F_{ij} &= 0 \\ \Rightarrow [\mathbf{M}\mathbf{F} + \lambda\mathbf{F}(\mathbf{F}^T \mathbf{F} - \mathbf{I}_c)]_{ij} F_{ij} &= 0. \end{aligned} \quad (2.68)$$

There may exist mix-signed elements in \mathbf{M} . To guarantee the nonnegative property of \mathbf{F} , by introducing $\mathbf{M} = \mathbf{M}^+ - \mathbf{M}^-$, where $M_{ij}^+ = (|M_{ij}| + M_{ij})/2$ and $M_{ij}^- = (|M_{ij}| - M_{ij})/2$, the above equation is equivalent to

$$[(\mathbf{M}^+ - \mathbf{M}^-)\mathbf{F} + \mu\mathbf{F}(\mathbf{F}^T \mathbf{F} - \mathbf{I}_c)]_{ij} F_{ij} = 0. \quad (2.69)$$

Here $|\cdot|$ denotes the absolute value function. Thus, we have

$$F_{ij} \leftarrow F_{ij} \frac{(\mathbf{M}^+ \mathbf{F} + \mu\mathbf{F})_{ij}}{(\mathbf{M}^+ \mathbf{F} + \mu\mathbf{F}\mathbf{F}^T \mathbf{F})_{ij}}. \quad (2.70)$$

Then we normalize \mathbf{F} with $(\mathbf{F}^T \mathbf{F})_{ii} = 1, i = 1, \dots, c$.

From the above analysis, we can see that \mathbf{D} and \mathbf{H} related to \mathbf{W} is required to solve \mathbf{F} and it is still not straightforward to obtain \mathbf{W} and \mathbf{F} . To this end, we design an iterative algorithm to solve the proposed formulation, which is summarized in Algorithm 2.

The alternative updating rules in Algorithm 2 monotonically decrease the objective function value of (2.57) in each iteration. That is, the proposed iterative procedure in Algorithm 2 can be verified to be convergent. The convergence is also experimentally verified in our experiments. Besides, the proposed optimization algorithm is efficient. In the experiments, we observe that our algorithm usually converges around only 20 iterations.

Now, we briefly analyze the computational complexity. In our case, $c \ll n$ and $c \ll d$. It takes $O(nd^2)$ to obtain \mathbf{C} . The complexity of calculating the inverse of a matrix is $O(d^3)$. In each iteration step, the cost for updating \mathbf{G} based on \mathbf{W} and \mathbf{C} is

Algorithm 2 The Proposed NSCR Method**Input:**Data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$;Parameters $\alpha, \beta, \gamma, \mu, k, c$ and p 1: Construct the k -nearest neighbor graph and calculate \mathbf{L} ;2: Construct the correlation matrix between features \mathbf{C} ;3: The iteration step $t = 1$; Initialize $\mathbf{F}_t \in \mathbb{R}^{n \times c}$, set $\mathbf{D}_t \in \mathbb{R}^{d \times d}$ as an identity matrix and $\mathbf{H}_t \in \mathbb{R}^{d \times d}$ as a zero matrix;4: **repeat**5: $\mathbf{G}_t = \mathbf{X}\mathbf{X}^T + \frac{\beta}{\alpha}\mathbf{D}_t + \frac{\gamma}{\alpha}\mathbf{H}_t$;6: $\mathbf{M}_t = \mathbf{L} + \alpha\mathbf{I}_n - \alpha\mathbf{X}^T\mathbf{G}_t^{-1}\mathbf{X}$;7: $\mathbf{M}_t^+ = (|\mathbf{M}_t| + \mathbf{M}_t)/2$ and $\mathbf{M}_t^- = (|\mathbf{M}_t| - \mathbf{M}_t)/2$;8: $(F_{t+1})_{ij} = (F_t)_{ij} \frac{(\mathbf{M}_t^-\mathbf{F}_t + \mu\mathbf{F}_t)_{ij}}{(\mathbf{M}_t^+\mathbf{F}_t + \mu\mathbf{F}_t\mathbf{F}_t^T)_{ij}}$;9: $\mathbf{W}_{t+1} = \mathbf{G}_t^{-1}\mathbf{X}\mathbf{F}_{t+1}$;10: Update the diagonal matrix \mathbf{D} as

$$\mathbf{D}_{t+1} = \begin{bmatrix} \frac{p}{2\|(\mathbf{w}_{t+1})_1\|_2^{2-p}} & & \\ & \dots & \\ & & \frac{p}{2\|(\mathbf{w}_{t+1})_d\|_2^{2-p}} \end{bmatrix};$$

11: Update the diagonal matrix \mathbf{H} as

$$\mathbf{H}_{t+1} = \begin{bmatrix} \frac{\sum_j \|(\mathbf{w}_{t+1})_j\|_2 C_{1j}}{2\|(\mathbf{w}_{t+1})_1\|_2} & & \\ & \dots & \\ & & \frac{\sum_j \|(\mathbf{w}_{t+1})_j\|_2 C_{dj}}{2\|(\mathbf{w}_{t+1})_d\|_2} \end{bmatrix};$$

12: $t=t+1$;13: **until** Convergence criterion satisfied**Output:**Sort all d features according to $\|(\mathbf{w}_t)_i\|_2$ in descending order and select the top r ranked features.

$O(cd + d^2)$. It needs $O(d^3)$ to obtain \mathbf{G}^{-1} . The cost for updating \mathbf{M} is $O(nd^2 + n^2d)$. It takes $O(cn^2)$ to update \mathbf{F} and $O(nd^2)$ to update \mathbf{W} , respectively. Thus the overall cost for the proposed NSCR is $O(T(d^3 + nd^2 + dn^2))$, where T is the number of iterations.

2.4.3 Experiments

In this section, we experimentally evaluate the performance of the proposed NSCR method for unsupervised feature selection, which can be applied to many applications, such as clustering and classification. We only evaluate the performance of NSCR and compared with representative algorithms in terms of clustering. In our experiments, we first select the top r features and then utilize Kmeans algorithm to cluster images based on the selected features.

2.4.3.1 Data Sets

The experiments are conducted on nine publicly available image datasets, including four face image data sets, i.e., UMIST [1], AT&T [36], JAFFE [29], and Pointing4 [44], three handwritten digit data sets, i.e., MNIST used in [44], Binary Alphabet (BA) [1] and a subset of USPS with 40 samples randomly selected for each class [1], and two object image databases, i.e., COIL20 [31] and Caltech101 [15]. Data sets from different areas serve as a good test bed for a comprehensive evaluation.

Some datasets have been introduced in Sect. 2.3. In the AT&T face image dataset [36], there are 10 gray scale images for each of the 40 human objects. There were taken at different times, varying the lighting, facial expressions and facial details. The image size is 32×32 . The COIL20 [31] database contains 32×32 gray scale images of 20 objects viewed from varying angles and each object has 72 images. The Caltech101 dataset [15] contains 9144 images of 101 classes and an additional class of background images. In our experiments, we select the 10 largest categories, except the BACKGROUND_GOOGLE category. The SIFT descriptor is extracted and then 1000-dimensional bag of visual word is generated to represent each image. Table 2.6 summarizes the details of these nine benchmark data sets used in the experiments in terms of the total number n of images, the total number c of clusters and the feature dimension d .

2.4.3.2 Compared Scheme

To validate the effectiveness of the proposed NSCR for feature selection, we compare it with one baseline and several unsupervised feature selection methods. The compared algorithms are enumerated as follows:

- **Baseline:** All original visual features are adopted.

Table 2.6 Image dataset description. n is the number of images; d denotes the dimension of features; c is the number of clusters.

Domain	Dataset	n	d	c
Face image	UMIST	575	644	20
	AT&T	400	1024	40
	JAFFE	213	676	10
	Pointing4	2790	1120	15
Handwritten digits image	MNIST	5000	784	10
	BA	1404	320	36
	USPS	400	256	10
Object image	Coil20	1440	1024	20
	Caltech101	3379	1000	10

- **MaxVar**: Features corresponding to the maximum variance are selected to obtain the expressive features.
- **LS** [17]: Features consistent with Gaussian Laplacian matrix are selected to preserve the local manifold structure.
- **SPEC** [47]: Features are selected using spectral regression based on pairwise image similarity.
- **SPFS-SFS** [48]: The traditional forward search strategy is utilized for similarity preserving feature selection in the SPFS framework.
- **MCFS** [5]: Features are selected based on spectral analysis and sparse regression in a two-step scheme;
- **UDFS** [43]: Features are selected by exploiting the local structure for local discriminative information and row-sparse models for feature correlations simultaneously.
- **SCR**: A special case of the proposed method without considering the redundancy constraint for unsupervised feature selection, i.e., $\gamma = 0$.
- **NSCR**: The proposed method with Nonnegative Spectral analysis and Controlled Redundancy for unsupervised feature selection.

In the compared methods, there are some hyper-parameters to be set in advance. The same strategy used in Sect. 2.3 is adopted.

2.4.3.3 Results on Synthetic Data

To well evaluate the effectiveness of the proposed NSCR method on the feature selection task, we conduct experiments on one widely used synthetic dataset, i.e., Corral [45]. It contains six Boolean features ($A0$, $A1$, $B0$, $B1$, I , R), in which the relevant features, irrelevant features, and redundant features are provided. Specifically, the class labels of data points are defined by $(A0 \wedge A1) \vee (B0 \wedge B1)$ while $A0$, $A1$, $B0$, and $B1$ are independent to each other. Feature R is redundant by matching the class label 75% of the time, while feature I is uniformly random. That is, features $A0$, $A1$, $B0$, and $B1$ are necessary features, feature R is the redundant feature while feature I is the noisy feature. The results of features ranked by different methods are presented in Table 2.7. For each method, features are selected from left to right, top to bottom. It can be seen that if the top four features are selected, only the proposed method can remove the noisy feature and the redundant feature simultaneously while other methods fail to filter out the redundant feature, which demonstrates the effectiveness of the proposed method for feature selection.

Table 2.7 The rank of features by different methods on the Corral data. Features are selected from left to right, top to bottom.

	MaxVar	LS	SPEC	SPFS-SFS	MCFS	UDFS	SCR	NSCR
Rank	$R, A0, A1,$	$R, A0, B0,$	$R, A0, A1,$	$R, B1, A0,$	$R, A0, B0,$	$A1, R, B0,$	$B0, B1, A1,$	$B1, B0, A1,$
	$B0, B1, I$	$A1, B1, I$	$B0, B1, I$	$B0, A1, I$	$A1, B1, I$	$A0, B1, I$	$R, A0, I$	$A0, R, I$

2.4.3.4 Performance Comparison on Real-world Data

We now empirically evaluate the performance of these nine feature selection algorithms for clustering in terms of ACC and NMI. The detailed results on the face, handwritten digit and object data sets are summarized in Tables 2.8, 2.9, and 2.10, respectively. The results demonstrate that NSCR achieves the best performance on all the nine image sets compared to other eight feature selection algorithms, which validates its effectiveness.

From the above experimental results, we have the following observations. First, it is observed that NSCR achieves better performance than SCR by considering the nonnegative constraint. It demonstrates that it is necessary and effective to introduce the nonnegative constraint. The improved NSCR enables to remove the redundant features while preserving the necessary features. Second, NSCR is both superior to MCFS by introducing the nonnegative constraint, which makes the scaled cluster indicators more accurate. They can select more necessary features. Third, NSCR, UDFS, and MCFS achieve larger values of ACC and NMI by exploiting discriminative information from data. It demonstrates that it is crucial to uncover the discriminative information in the unsupervised case, which can remove noisy features and indifferent features. Fourth, NSCR achieves more accurate clustering performance than SPEC and MCFS. SPEC and MCFS adopt a two-step approach to introduce spectral analysis into feature selection while NSCR is an one-step framework and perform spectral analysis and feature selection simultaneously. Fifth, by exploiting the local geometric structure of data distribution, LS, SPEC, MCFS, UDFS, and NSCR usually yield superior performance. Besides, it can be seen that it is necessary to select features jointly rather than one by one. The joint feature selection algorithms, such as MCFS, UDFS, and NSCR are always superior to the methods selecting features one after another, such as MaxVar and SPEC. Finally, compared with the baseline, it can be observed that feature selection is necessary and effective by removing the noise. It can not only reduce the number of features and make the algorithms more efficient, but also improve the performance. In conclusion, NSCR achieves the best performance on all data sets by exploiting nonnegative spectral analysis and redundancy between features simultaneously for feature selection, which can select necessary features, control the use of redundant features and discard noisy or indifferent features.

2.5 Discussions

In this chapter, we propose a novel understanding-oriented unsupervised feature selection framework, which exploits nonnegative spectral analysis, the latent structure analysis, and explicitly control the redundancy between features while a sparse model with the $\ell_{2,p}$ -norm is introduced. The proposed framework can select necessary features, remove noisy, or indifferent features while control the redundancy between the selected features. The cluster indicators learned by nonnegative spectral

Table 2.8 Clustering results comparison on the face image data sets. The best results are highlighted in bold

Dataset	Baseline	MaxVar	LS	SPEC	SPFS-SFS	MCFS	UDFS	SCR	NSCR
ACC \pm std (%)									
UMIST	41.3 \pm 2.9	45.6 \pm 4.9	46.1 \pm 1.9	48.5 \pm 4.4	47.7 \pm 2.9	47.5 \pm 3.6	49.5 \pm 3.2	54.6 \pm 1.1	56.7 \pm 2.6
AT&T	50.8 \pm 3.1	46.5 \pm 3.2	47.6 \pm 2.2	50.7 \pm 3.2	51.9 \pm 2.5	52.9 \pm 2.9	52.3 \pm 2.1	53.7 \pm 1.9	56.2 \pm 1.3
JAFPE	73.6 \pm 9.2	72.2 \pm 5.1	74.9 \pm 6.4	75.6 \pm 5.8	76.6 \pm 6.3	77.4 \pm 6.1	77.9 \pm 4.9	82.2 \pm 4.2	82.9 \pm 4.0
Pointing4	35.6 \pm 1.7	44.8 \pm 2.0	38.0 \pm 1.7	38.9 \pm 2.5	39.3 \pm 1.1	47.2 \pm 1.9	45.4 \pm 3.0	50.1 \pm 1.8	52.3 \pm 1.1
NMI \pm std (%)									
UMIST	62.7 \pm 1.8	63.7 \pm 3.7	64.1 \pm 1.2	67.6 \pm 1.9	62.7 \pm 2.2	67.1 \pm 2.3	69.3 \pm 3.5	71.6 \pm 1.5	73.4 \pm 2.5
AT&T	71.2 \pm 1.8	68.9 \pm 2.4	69.5 \pm 1.4	71.4 \pm 1.6	72.6 \pm 1.5	73.5 \pm 1.5	72.0 \pm 1.9	71.3 \pm 0.6	74.5 \pm 1.7
JAFPE	80.8 \pm 5.9	74.7 \pm 4.1	81.2 \pm 4.4	82.3 \pm 4.4	83.1 \pm 3.8	81.7 \pm 4.5	82.1 \pm 3.8	86.2 \pm 2.2	87.7 \pm 3.1
Pointing4	41.4 \pm 1.0	51.6 \pm 1.6	55.3 \pm 1.5	56.3 \pm 1.4	42.7 \pm 1.2	55.8 \pm 1.6	52.4 \pm 1.6	57.0 \pm 1.5	57.9 \pm 1.0

Table 2.9 Clustering results comparison on the handwritten digit data sets. The best results are highlighted in bold

Dataset	Baseline	MaxVar	LS	SPEC	SPFS-SFS	MCFS	UDFS	SCR	NSCR
ACC \pm std (%)									
MNIST	50.1 \pm 5.3	54.2 \pm 3.1	52.9 \pm 3.4	53.3 \pm 4.7	54.9 \pm 4.7	55.8 \pm 3.7	57.3 \pm 2.5	58.7 \pm 1.9	60.8 \pm 2.7
BA	40.7 \pm 1.7	41.7 \pm 1.1	42.6 \pm 2.3	41.8 \pm 1.4	42.5 \pm 2.1	42.3 \pm 1.7	42.8 \pm 1.9	42.1 \pm 1.0	45.7 \pm 1.6
USPS	62.8 \pm 5.1	64.1 \pm 1.3	64.5 \pm 5.0	65.9 \pm 4.0	63.4 \pm 3.3	65.9 \pm 2.8	65.2 \pm 3.6	68.1 \pm 1.5	72.0 \pm 2.6
NMI \pm std (%)									
MNIST	47.2 \pm 2.4	48.1 \pm 1.8	48.2 \pm 1.5	49.1 \pm 2.2	49.6 \pm 2.1	50.4 \pm 1.6	51.1 \pm 1.4	50.9 \pm 1.3	53.9 \pm 2.4
BA	56.4 \pm 0.9	57.3 \pm 0.6	58.1 \pm 0.6	57.4 \pm 0.8	57.6 \pm 1.2	57.9 \pm 1.0	58.1 \pm 1.0	57.3 \pm 0.5	60.0 \pm 1.3
USPS	58.5 \pm 3.1	60.2 \pm 0.8	58.9 \pm 2.8	59.2 \pm 1.8	56.5 \pm 1.7	59.4 \pm 2.4	60.1 \pm 3.4	61.8 \pm 1.1	63.9 \pm 1.1

Table 2.10 Clustering results comparison on the object image data sets. The best results are highlighted in bold

Dataset	Baseline	Max Var	LS	SPEC	SPFS-SFS	MCFS	UDFS	SCR	NSCR
ACC \pm std (%)									
COIL20	59.0 \pm 5.7	58.4 \pm 4.0	57.3 \pm 3.0	61.0 \pm 2.1	62.5 \pm 2.6	61.7 \pm 4.3	62.9 \pm 2.6	65.3 \pm 3.6	67.2 \pm 1.5
Caltech101	47.7 \pm 3.8	40.1 \pm 1.9	47.6 \pm 2.3	48.6 \pm 3.2	49.2 \pm 2.7	50.2 \pm 5.0	50.8 \pm 2.9	52.3 \pm 1.8	54.3 \pm 2.4
NMI \pm std (%)									
COIL20	72.9 \pm 2.8	70.5 \pm 0.9	70.4 \pm 1.1	74.1 \pm 2.4	76.2 \pm 1.6	74.7 \pm 2.3	75.9 \pm 1.1	77.3 \pm 1.7	78.9 \pm 1.2
Caltech101	34.5 \pm 4.5	35.0 \pm 2.0	37.5 \pm 2.7	36.3 \pm 2.2	33.4 \pm 2.3	38.0 \pm 5.3	38.4 \pm 2.5	50.9 \pm 1.3	42.2 \pm 1.4

clustering are used to provide label information for unsupervised feature selection. To facilitate feature selection, the predictive matrix is constrained to be sparse in rows. By imposing the $\ell_{2,p}$ -norm regularization, the proposed methods jointly selects the most discriminative features across the entire feature space. For future work, we will focus on extending our methods in the kernel learning framework and the local learning framework. Besides, how to select the adaptive hyper-parameters and the number of selected features are also our directions for future research.

By the proposed understanding-oriented feature selection framework, better feature subsets can be chosen to represent the contents of images. However, the contents of images are still described by the low-level visual features. The semantic gap between low-level features and high-level semantic still exist, although it may be reduced to a certain degree by the understanding-oriented feature selection methods. The better way may be to learn a good representation for data, i.e., feature extraction. The understanding-oriented feature representation is desired to be studied.

On the other hand, we known that social images are associated with user-provided tags. Although these tags are imperfect, they can reflect the semantic information of images to a certain degree. It is believed that users express their individual understanding through tags. Thus, the user-provided tags enable to help us to select better feature subsets or learn better data representation. Meanwhile, the imperfect problem of the user-provided tags should be handled during the learning procedure. Better data representations can be found by exploring the user-provided tags.

References

1. <http://cs.nyu.edu/~roweis/data.html>
2. <http://tunedit.org/repo/Data/Text-wc>
3. <http://featureselection.asu.edu/datasets.php>
4. Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.* **6**, 1817–1853 (2005)
5. Cai, D., Zhang, C., He, X.: Unsupervised feature selection for multi-cluster data. In: *Proceedings of ACM SIGKDD International Conference Knowledge Discovery and Data Mining*, pp. 333–342 (2010)
6. Chakraborty, R., Pal, N.: Feature selection using a neural framework with controlled redundancy. *IEEE Trans. Neural Netw. Learn. Syst.* **26**(1), 35–50 (2015)
7. Chartrand, R.: Exact reconstructions of sparse signals via nonconvex minimization. *IEEE Signal Process Lett.* **14**(10), 2832–2852 (2007)
8. Chen, X., Xu, F., Ye, Y.: Lower bound theory of nonzero entries in solutions of ℓ_2 - ℓ_p minimization. *SIAM J. Sci. Comput.* **32**(5), 2832–2852 (2010)
9. Cheung, Y.M., Zeng, H.: Local kernel regression score for selecting features of high-dimensional data. *IEEE Trans. Knowl. Data Eng.* **21**(12), 1798–1802 (2009)
10. Constantinopoulos, C., Titsias, M.K., Likas, A.: Bayesian feature and model selection for gaussian mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(6), 1013–1018 (2006)
11. Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T.M., Nigam, K., Slattey, S.: Learning to extract symbolic knowledge from the world wide web. In: *Proceedings of AAAI Conference on Artificial Intelligence* (1998)

12. Du, L., Shen, Z., Li, X., Zhou, P., Shen, Y.D.: Local and global discriminative learning for unsupervised feature selection. In: *Proceedings of IEEE International Conference on Data Mining*, pp. 131–140 (2013)
13. Duda, R., Hart, P., Stork, D.: *Pattern Recognition*, 2nd edn. Wiley, New York (2001)
14. Dy, J.G., Brodley, C.E.: Feature selection for unsupervised learning. *J. Mach. Learn. Res.* **5**, 845–889 (2004)
15. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: *Proceedings of IEEE Computer Vision and Pattern Recognition Workshop on Generative-Model Based Vision*, pp. 178–186 (2004)
16. Gourier, N., Hall, D., Crowley, J.: Estimating face orientation from robust detection of salient facial features. In: *Proceedings of ICPR Workshop on Visual Observation of Deictic Gestures*, pp. 1–9 (2004)
17. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: *Advances in Neural Information Processing Systems*, pp. 507–514 (2005)
18. Higham, N.J.: *Accuracy and Stability of Numerical Algorithms*, 2nd edn. Society for Industrial and Applied Mathematics, Manchester (2002)
19. Jain, A., Zongker, D.: Feature selection: evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(2), 153–158 (1997)
20. Ji, S., Zhang, L., Yu, S., Ye, J.: A shared-subspace learning framework for multi-label classification. *ACM Trans. Knowl. Discov. Data* **4**(2), 1817–1853 (2010)
21. Jiang, W., Er, G., Dai, Q., Gu, J.: Similarity-based online feature selection in content-based image retrieval. *IEEE Trans. Image Process.* **15**(3), 702–712 (2006)
22. Kuhn, H., Tucker, A.: Nonlinear programming. In: *Berkeley Symposium on Mathematical Statistics and Probabilistics* (1951)
23. Law, M.H., Figueirdo, M.A., Jain, A.K.: Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(9), 1154–1166 (2004)
24. Lee, D., Seung, H.: Learning the parts of objects by nonnegative matrix factorization. *Nature* **401**, 788–791 (1999)
25. Li, Z., Liu, J., Yang, Y., Zhou, X., Lu, H.: Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE Trans. Knowl. Data Eng.* **9**(26), 2138–2150 (2014)
26. Li, Z., Tang, J.: Unsupervised feature selection via nonnegative spectral analysis and redundancy control. *IEEE Trans. Image Process.* **24**(12), 5343–5355 (2015)
27. Li, Z., Yang, Y., Liu, J., Zhou, X., Lu, H.: Unsupervised feature selection using nonnegative spectral analysis. In: *Proceedings of National Conference on Artificial Intelligence (AAAI)*, pp. 1026–1032 (2012)
28. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
29. Lyons, M.J., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. *IEEE Trans. Pattern Anal. Mach. Intell.* **21**(12), 1357–1362 (1999)
30. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(3), 301–312 (2002)
31. Nene, S., Nayar, S., Murase, H.: *Columbia object image library*. Tech. Rep. Tech. Rep. CUCS-005-96, Columbia University (1996)
32. Ng, A.Y., Jordan, M., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: *Advances in Neural Information Processing Systems* (2001)
33. Nie, F., Huang, H., Cai, X., Ding, C.: Efficient and robust feature selection via joint $\ell_{2,1}$ -norms minimization. In: *Advances in Neural Information Processing Systems*, pp. 1813–1821 (2010)
34. Peng, H.C., Long, F.H., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundant. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1226–1238 (2005)
35. Qian, M., Zhai, C.: Robust unsupervised feature selection. In: *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 1621–1627 (2014)

36. Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification. In: Proceedings of IEEE Workshop on Applications of Computer Vision, pp. 1–9 (1994)
37. Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 888–905 (2000)
38. Shi, L., Du, L., Shen, Y.D.: Robust spectral learning for unsupervised feature selection. In: Proceedings of IEEE International Conference on Data Mining, pp. 131–140 (2014)
39. Song, L., Smola, A., Gretton, A., Bedo, J., Borgwardt, K.: Feature selection via dependence maximization. *J. Mach. Learn. Res.* **13**, 1393–1434 (2012)
40. Tang, Z., Zhang, Y., Li, Z., Lu, H.: Face clustering in videos with proportion prior. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 2191–2197 (2015)
41. Wang, L., Chen, S., Wang, Y.: A unified algorithm for mixed $\ell_{2,p}$ -minimizations and its application in feature selection. *Comput. Optim. Appl.* **58**(2), 409–421 (2014)
42. Wolf, L., Shashua, A.: Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weight-based approach. *J. Mach. Learn. Res.* **6**, 1855–1887 (2005)
43. Yang, Y., Shen, H.T., Ma, Z., Huang, Z., Zhou, X.: $\ell_{2,1}$ -norm regularized discriminative feature selection for unsupervised learning. In: Proceedings of International Joint Conference on Artificial Intelligence (2011)
44. Yang, Y., Xu, D., Nie, F., Yan, S., Zhuang, Y.: Image clustering using local discriminant models and global integration. *IEEE Trans. Image Process.* **19**(10), 2761–2773 (2010)
45. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.* **5**(10), 1205–1224 (2004)
46. Zeng, H., Cheung, Y.M.: Feature selection and kernel learning for local learning-based clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(8), 1532–1547 (2011)
47. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: Proceedings of International Conference on Machine Learning, pp. 1151–1157 (2007)
48. Zhao, Z., Wang, L., Liu, H., Ye, J.: On similarity preserving feature selection. *IEEE Trans. Knowl. Data Eng.* **25**(3), 619–632 (2013)

Understanding-Oriented Multimedia Content Analysis

Zhou, H.

2017, XIX, 156 p. 36 illus., 33 illus. in color., Hardcover

ISBN: 978-981-10-3688-0