UNSUPERVISED FEATURE SELECTION USING A BASIS OF FEATURE SPACE AND SELF-REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, there has been extensive research into unsupervised feature selection methods based on self-representation. However, there exists a major gap in the mathematical principles that underlie these approaches and their capacity to represent the feature space. In this paper, a novel representation learning method, Graph Regularized Self-Representation and Sparse Subspace Learning (GRSSLFS), is proposed for the unsupervised feature selection. Firstly, GRSSLFS expresses the self-representation problem based on the concept of "a basis of feature space" to represent the original feature space as a low-dimensional space made of linearly independent features. Furthermore, the manifold structure corresponding to the newly constructed subspace is learned in order to preserve the geometric structure of the feature vectors. Secondly, the objective function of GRSSLFS is developed based on a self-representation framework that combines subspace learning and matrix factorization of the basis matrix. Finally, the effectiveness of GRSSLFS is explored through experiments on widely-used datasets. Results show that GRSSLFS achieves a high level of performance in comparison with several classic and state-of-the-art feature selection methods.

1 INTRODUCTION

Feature selection is a pivotal tool for handling high-dimensional data, which facilitates the identification of essential features within the original feature set Dinh & Ho (2020); Majumdar & Chatterjee (2022). This process provides valuable interpretations across a spectrum of applications, such as image processing Shi et al. (2023), bioinformatics Liang et al. (2018), genomics Sun et al. (2022), and so on. In recent years, feature selection methods have been designed with different perspectives such as subspace learning Li et al. (2021), graph learning Roffo et al. (2021) and self-representation learning Chen et al. (2022).

Subspace Learning is a comprehensive concept that encompasses a variety of techniques with the goal of revealing lower-dimensional subspaces within datasets characterized by high dimensions Ren et al. (2021). The main objective of subspace learning is to depict data points in a dimensionally reduced space while preserving as much relevant information as feasible. These methods can be categorized into two primary types: (1) Linear Subspace Learning, and (2) Nonlinear Subspace Learning. The first category, which comprises well-established techniques such as Local Preserving Projection (LPP) He & Niyogi (2003), Principal Component Analysis (PCA) Marukatat (2023), is focused on recognizing linear combinations of the initial features that define a lower-dimensional subspace. Nonetheless, in situations where the inherent data structure departs from linearity, nonlinear subspace learning methods become relevant. Methods like Kernel PCA Ghojogh et al. (2023) and Locally Linear Embedding (LLE) Roweis & Saul (2000) serve as examples of approaches for nonlinear subspace learning. Although subspace learning and feature selection can be employed separately, their combination can result in even more potent approaches to dimensionality reduction in data and modeling strategies. For example, MFFS Wang et al. (2015a) serves as a prominent instance of unsupervised feature selection methods developed within the context of subspace learning and matrix factorization techniques. Following that, numerous approaches that integrate subspace learning and feature selection have been introduced, such as Shang et al. (2020); Sheng et al. (2021); Yuan et al. (2022); Li et al. (2022).

In addition to subspace learning, representation learning has been widely used in various parts of dimensionality reduction methods Shang et al. (2021); Wang et al. (2023). When it comes to deal with high-dimensional data with a large amount of redundant features, it will be a complex task to analyze and mine such data, which will generally lead to imprecise results. One preliminary aim of representation learning is to discover the basic information such as interior and characteristic structure from the data Wu et al. (2023); Shao et al. (2023). As a popular category of representation learning, self-representation learning achieves high performance in combination with other methods such as subspace learning for feature selection Chen et al. (2022). Self-representation learning indeed originates from one of the earliest mathematical concepts called a basis for vector spaces and is rested on the assumption that every sample or feature vector of the original data can be described in terms of other sample or feature vectors. According to this merit of self-representation, some effective feature selection methods have been recently developed to focus on utilizing the correlations between features Tang et al. (2019); Zhang et al. (2022).

However, the main drawback of feature selection methods based on self-representation is that they use all the features of the original data, especially the redundant ones, in their representation. In order to get the root of this issue, the philosophy behind the basis for vector spaces can be beneficial Lin et al. (2022). In linear algebra, it is known that a basis for a vector space can be defined as a set of linearly independent vectors that can uniquely produce the whole space. In fact, because of the linear independency among the members of a basis, it will be very likely that the redundant vectors will be left out of the basis. Therefore, by having a basis for the space generated by features of the original data, redundant features will play a much lesser role in the self-representation process.

These two comprehensive properties of the basis for the feature space, i.e., the linear independence of its members and the unique representation of the feature space, are the motivation for us to introduce a more efficient form of self-representation and subspace learning problems according to the basis of features. The aim of this paper is to establish a new type of feature selection method which is called Graph Regularized Self-Representation and Sparse Subspace Learning (GRSSLFS). This method simultaneously considers both subspace learning and self-representation problems with respect to the use of a basis for the feature space. Extensive experimental results indicate the efficiency of the GRSSLFS method. The main contributions of the paper are summarized below:

- To the best of our knowledge, GRSSLFS is the first feature selection method that applies a basis of linearly independent features into a unified framework of subspace learning and self-representation.
- GRSSLFS characterizes the challenges of subspace learning and self-representation by utilizing a basis matrix derived from the feature space. The objective is to eliminate unnecessary features and identify informative ones within high-dimensional data.
- GRSSLFS constructs a basis for the feature space by incorporating the variance information of features into the Basis Extension method. The goal is to create a feature basis with the highest variance score. This ensures that the basis for the feature space comprises elements with the most dispersion in space, thereby minimizing the accumulation of features.

2 Related Work

Feature selection techniques that fall under either the subspace learning or self-representation learning framework are a prevalent category within the realm of dimensionality reduction. Improving this issue can be achieved by introducing different assumptions, such as orthogonality constraints, graph, or sparse regularization techniques. In this context, there is a relevant area of research that investigates the combination of subspace learning and matrix factorization methods, as evidenced by the studies such as Qi et al. (2018); Zheng et al. (2019); Lin et al. (2022); Luo et al. (2022). An additional area of research investigates the simultaneous use of self-representation and sparse matrix approximation techniques, for example Chen et al. (2022); Tang et al. (2023). However, these works encompass all the features from the initial dataset, including redundant ones, in their representations, which could potentially harm the effectiveness of feature selection. The key characteristics of some unsupervised feature selection methods based on subspace learning or self-representation are outlined in Table 2 shown in Appendix B. Drawing inspiration from the idea of establishing a basis for the feature space, our proposed approach redefines both the subspace learning and self-representation learning frameworks. It does so by building upon a feature basis consisting of a set of linearly independent features, resulting in a reduced level of redundancy.

3 METHODOLOGY

In this section, the details of the proposed GRSSLFS method are explained. In what follows, $\mathbf{X} \in \mathbb{R}_+^{m\times n}$ indicates the data matrix with m samples and n features. Here, $\mathbb{R}_+^{m\times n}$ denotes the set of non-negative matrices of size $m \times n$. For an m dimension vector space W, the set $B \subseteq W$ is called a basis if the span of elements of B can produce W and B is linearly independent. Here, the span of some vectors refers to all the possible linear combinations of them. The Frobenius norm, the $L_{2,1}$-norm, the trace, and the transpose of a matrix $\mathbf{A}$ are denoted by $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_{2,1}$, $\mathrm{Tr}(\mathbf{A})$, and $\mathbf{A}^T$, respectively. Finally, $\langle \mathbf{a}, \mathbf{b}\rangle = \mathbf{a}^T\mathbf{b}$ is the Euclidean inner product of the vectors $\mathbf{a}$ and $\mathbf{b}$.

3.1 SELF-REPRESENTATION PROBLEM AND THE BASIS

The self-representation problem of features is considered as a linear representation of features in a dataset. In light of the linear combination concept, the self-representation problem of features $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$ is defined as:

$$\min_{\mathbf{C}\in\mathbb{R}_+^{n\times n}} \|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2, \tag{1}$$

which can also be expressed as:

$$\begin{cases} \mathbf{f}_1 \simeq c_{11}\mathbf{f}_1 + c_{21}\mathbf{f}_2 + \cdots + c_{n1}\mathbf{f}_n, \\ \mathbf{f}_2 \simeq c_{12}\mathbf{f}_1 + c_{22}\mathbf{f}_2 + \cdots + c_{n2}\mathbf{f}_n, \\ \vdots \\ \mathbf{f}_n \simeq c_{1n}\mathbf{f}_1 + c_{2n}\mathbf{f}_2 + \cdots + c_{nn}\mathbf{f}_n, \end{cases} \tag{2}$$

where $\mathbf{C} = [c_{ij}]$ is the coefficient matrix. In the following, we assume, without loss of generality, that $\mathrm{rank}(\mathbf{X}) = m$ where $m \leq n$. In cases where $m > n$, a similar discussion can be also applied. Under this assumption, it can be easily shown that there exists a basis B with linearly independent features $\{\mathbf{f}_{i_1}, \ldots, \mathbf{f}_{i_m}\}$ such that the set $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$ can be represented by the basis B. More precisely, compared to the linear system of equations (2), the following more compact equations are obtained:

$$\begin{cases} \mathbf{f}_1 \simeq g_{11}\mathbf{f}_{i_1} + g_{12}\mathbf{f}_{i_2} + \cdots + g_{m1}\mathbf{f}_{i_m}, \\ \mathbf{f}_2 \simeq g_{12}\mathbf{f}_{i_1} + g_{22}\mathbf{f}_{i_2} + \cdots + g_{m2}\mathbf{f}_{i_m}, \\ \vdots \\ \mathbf{f}_n \simeq g_{1n}\mathbf{f}_{i_1} + g_{m2}\mathbf{f}_{i_2} + \cdots + g_{mn}\mathbf{f}_{i_m}, \end{cases} \tag{3}$$

where $\mathbf{G} = [g_{ij}] \in \mathbb{R}^{m\times n}$ is the basis coefficient matrix. Now, considering the representation of the features for the original dataset, the self-representation problem can be defined based on the basis as:

$$\min_{\mathbf{G}\in\mathbb{R}_+^{m\times n}} \|\mathbf{X} - \mathbf{B}\mathbf{G}\|_F^2, \tag{4}$$

where $\mathbf{B} = [\mathbf{f}_{i_1}, \ldots, \mathbf{f}_{i_m}] \in \mathbb{R}^{m\times m}$ is a basis matrix. Problem (4) clarifies the learning of the matrix $\mathbf{G}$. According to the relation $\mathbf{X} \simeq \mathbf{B}\mathbf{G}$ and assuming that $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_n]$, it is concluded that

$$\mathbf{X} \simeq \mathbf{B}\mathbf{G} \Rightarrow [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n] \simeq [\mathbf{B}\mathbf{g}_1, \ldots, \mathbf{B}\mathbf{g}_n]. \tag{5}$$

As a result, according to the basic principle of preserving the geometric structure in the feature manifold, if the two features $\mathbf{f}_l$ and $\mathbf{f}_r$ have a similar structure in the feature space, then it can be expected that their corresponding representations $\mathbf{B}\mathbf{g}_l$ and $\mathbf{B}\mathbf{g}_r$ will also have the similar structure. This observation can be expressed as follows:

$$\min_{\mathbf{G}\in\mathbb{R}_+^{m\times n}} \mathrm{Tr}\left(\mathbf{B}\mathbf{G}\mathbf{L}\mathbf{G}^T\mathbf{B}^T\right) = \frac{1}{2}\min_{\mathbf{G}\in\mathbb{R}_+^{m\times n}} \sum_{q,r=1}^{n} a_{qr}\|\mathbf{B}\mathbf{g}_q - \mathbf{B}\mathbf{g}_r\|_2^2,$$

where $\mathbf{A} = [a_{qr}] \in \mathbb{R}^{n\times n}$ is the similarity matrix for the features, and the Laplacian matrix $\mathbf{L}$ is defined as $\mathbf{L} = \mathbf{P} - \mathbf{A}$ such that $\mathbf{P} = \mathrm{diag}(p_{qq})$ is a diagonal matrix with the diagonal entries $p_{qq} = \sum_{r=1}^{n} a_{qr}$, for $q = 1, \ldots, n$. Here, we consider $a_{qr} = e^{\frac{-\|\mathbf{f}_q - \mathbf{f}_r\|_2^2}{t^2}}$, if $\mathbf{f}_q \in N_k(\mathbf{f}_r)$ or $\mathbf{f}_r \in N_k(\mathbf{f}_q)$, where $N_k(\mathbf{f}_r)$ refers to the set of k feature vectors that are closest to $\mathbf{f}_r$, and t denotes the scale parameter of the Gaussian distribution, and $a_{qr} = 0$ otherwise.

Now, the graph regularized self-representation problem can be defined based on the basis as:

$$\min_{\mathbf{G}\in\mathbb{R}_+^{m\times n}} f_{\text{GRS}}(\mathbf{G}) = \|\mathbf{X} - \mathbf{BG}\|_F^2 + \alpha \text{Tr}\left(\mathbf{BGLG}^T\mathbf{B}^T\right),$$

in which α is called a feature graph regularization parameter.

3.2 Subspace Learning Problem and the Basis

One of the most popular feature selection techniques in the machine learning community is subspace learning so that a low-dimensional representation of high-dimensional data can be learned through linear or non-linear mappings. Several subspace learning models using Euclidean distance have been introduced in the past few years Wang et al. (2015a); Qi et al. (2018); Shang et al. (2020). Among the efficient subspace learning models is the following one:

$$\text{dist}_{\text{SL}}(\mathbf{X}, \mathbf{XU}) = \min_{\mathbf{V}} \|\mathbf{X} - \mathbf{XUV}\|_F^2, \tag{6}$$

where $\mathbf{U} \in \mathbb{R}_+^{n\times k}$ and $\mathbf{V} \in \mathbb{R}_+^{k\times n}$ are called the feature weight and the representation matrices, respectively. We shall note that the subspace learning model (6) indeed calculates the distance between the space of features generated by $\mathbf{X}$ and the space of the selected subset of features generated by $\mathbf{XU}$. A direct result of this model is that if $\mathbf{U}$ is chosen such that $\mathbf{XU} = \mathbf{B}$, then it can be seen that the distance between $\mathbf{X}$ and $\mathbf{B}$ is zero, i.e., $\text{dist}_{\text{SL}}(\mathbf{X}, \mathbf{B}) = 0$, which means that the space of $\mathbf{X}$ and the space of $\mathbf{B}$ can align or be identical; more information regarding this issue is provided in the following theorem.

Theorem 3.1. *Let $\mathbf{X} = [\mathbf{f}_1, \dots, \mathbf{f}_n] \in \mathbb{R}^{m\times n}$ be a dataset with n features such that $\text{rank}(\mathbf{X}) = m$. Let us also assume that $\mathbf{B} = [\mathbf{f}_{i_1}, \dots, \mathbf{f}_{i_m}]$ is a basis for the space generated by $\mathbf{X}$. Then, according to the distance-based subspace learning problem introduced in (6), the distance between $\mathbf{X}$ and $\mathbf{B}$ is zero, that is $\text{dist}_{SL}(\mathbf{X}, \mathbf{B}) = 0$.*

Proof. The detailed proof is provided in Appendix C due to space constraints. □

Remark. It should be noted that in the case where $\text{rank}(\mathbf{X}) < m$ or $m > n$, it is still possible to create a basic matrix for the feature space, and this will not cause any issues in designing our feature selection framework.

Similar to expressing the self-representation problem based on the basis matrix $\mathbf{B} \in \mathbb{R}^{m\times m}$, we can also state the subspace learning model (6) in terms of the basis matrix. In order to accomplish this goal, let us consider the representation relation (5). Therefore, the expression $\mathbf{X} - \mathbf{XUV}$ can be shown as $\mathbf{X} - \mathbf{BGUV}$. On the other hand, according to the vector space generation property of the basis, since the distance between $\mathbf{X}$ and $\mathbf{B}$ is zero, it is straightforward to see that $\mathbf{X}$ can be substituted using the basis matrix $\mathbf{B}$. As a result of what has been discussed so far, the subspace learning problem (6) based on the basis can be established as follows:

$$\min_{\mathbf{G},\mathbf{U},\mathbf{V}} \|\mathbf{B} - \mathbf{BGUV}\|_F^2, \tag{7}$$

where $\mathbf{G} \in \mathbb{R}_+^{m\times n}, \mathbf{U} \in \mathbb{R}_+^{n\times k}$, and $\mathbf{V} \in \mathbb{R}_+^{k\times m}$, such that $k \le n$ is the dimension of the selected feature space.

3.3 Sparsity Regularization term on U and V

The framework defined in (7) can be employed as a bedrock for the feature selection process, in which the feature weight matrix $\mathbf{U}$ and the representation matrix $\mathbf{V}$ play a major role in selecting underlying features. First, $\mathbf{U}$ is used as the weight matrix in the feature selection process, that is $\mathbf{X}_{\text{selected}} = \mathbf{BGU} = \sum_{i=1}^n (\mathbf{BG})_i \mathbf{U}_{i:}$, where $(\mathbf{BG})_i$ is the i-th column of $\mathbf{BG}$, and $\mathbf{U}_{i:}$ is the i-th row of $\mathbf{U}$. In fact, the sparser the construction of rows in $\mathbf{U}$, the higher the likelihood of selecting effective features. This description allows the utilization of the $L_{2,1}$-norm to promote row sparsity within the matrix $\mathbf{U}$, which can be demonstrated as $\|\mathbf{U}\|_{2,1}$. The importance of row sparsity in the $L_{2,1}$-norm lies in its effectiveness in managing structured sparsity, where the objective is to choose discriminative features. Second, let us consider that $\mathbf{B} = [\mathbf{f}_{i_1}, \dots, \mathbf{f}_{i_m}]$ and $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_m]$. It can be concluded from Problem (7) that $\mathbf{B} \simeq \mathbf{BGUV}$, which means that $[\mathbf{f}_{i_1}, \dots, \mathbf{f}_{i_m}] \simeq$

$[\mathbf{BGU}\mathbf{v}_1, \ldots, \mathbf{BGU}\mathbf{v}_m]$. This emplies that each feature $\mathbf{f}_{i_l}$ of the basis can be represented by $\mathbf{BGU}\mathbf{v}_l$, for $l = 1, \ldots, m$. To be more precise, we can see that $\mathbf{f}_{i_l} \simeq \sum_{r=1}^{k} (\mathbf{BGU})_r \mathbf{v}_{rl}$. Hence, the more sparse the columns of $\mathbf{V}$ are, then fewer members of $\mathbf{BGU}$ are used in the representation of $\mathbf{f}_{i_l}$ of the basis matrix $\mathbf{B}$, and as a result, a reduction in redundancy is more likely to occur. In order to reflect the importance of the representation matrix $\mathbf{V}$ in our proposed feature selection process, the idea of inner product regularization can be used. Let us consider the representation matrix $\mathbf{V}$. The Gram matrix of $\mathbf{V}$ can be written as $\mathbf{V}^T\mathbf{V} = [\langle \mathbf{v}_l, \mathbf{v}_p \rangle]$, where $\mathbf{v}_l$ is the l-th column of $\mathbf{V}$, for $l = 1, \ldots, m$. Now, it can be easily seen that the expression $\mathrm{Tr}(\mathbf{1}_{m\times m}\mathbf{V}^T\mathbf{V}) - \mathrm{Tr}(\mathbf{V}^T\mathbf{V})$ readily corresponds to the summation of the off-diagonal elements within the Gram matrix $\mathbf{V}^T\mathbf{V}$, that is to say $\mathrm{Tr}(\mathbf{1}_{m\times m}\mathbf{V}^T\mathbf{V}) - \mathrm{Tr}(\mathbf{V}^T\mathbf{V}) = \sum_{l,p=1,l\neq p}^{m} \langle \mathbf{v}_l, \mathbf{v}_p \rangle$. Therefore, by the subsequent problem formulated as follows:

$$\min_{\mathbf{V}\in\mathbb{R}^{k\times m}} \mathrm{Tr}(\mathbf{1}_{m\times m}\mathbf{V}^T\mathbf{V}) - \mathrm{Tr}(\mathbf{V}^T\mathbf{V}) = \min_{\mathbf{V}\in\mathbb{R}^{k\times m}} \sum_{l,p=1,l\neq p}^{m} \langle \mathbf{v}_l, \mathbf{v}_p \rangle, \tag{8}$$

and under the assumption that $\mathbf{V} \in \mathbb{R}_+^{k\times m}$, it can be deduced that the optimal solution to Problem (8) is attained when the inner product terms $\langle \mathbf{v}_l, \mathbf{v}_p \rangle$ tend toward zero for all pairs of l and p except when $l = p$. On the other hand, since $\langle \mathbf{v}_l, \mathbf{v}_p \rangle = (\mathbf{v}_l)^T \mathbf{v}_p$ and considering the non-negativity assumption for $\mathbf{v}_l$ and $\mathbf{v}_p$, the value of $\langle \mathbf{v}_l, \mathbf{v}_p \rangle$ tends towards zero when either $\mathbf{v}_l$ or $\mathbf{v}_p$ become zero or extremely sparse vectors. Consequently, the optimization Problem (8) can result in a significant degree of sparsity within the coefficient matrix $\mathbf{V}$.

Considering the importance of the role of $\mathbf{U}$ and $\mathbf{V}$, we integrate two efficient sparsity regularizations into the objective function (7) in order to make both $\mathbf{U}$ and $\mathbf{V}$ sparse. Two of the most useful and simple regularizations are the $L_{2,1}$-norm and the inner product regularization, which benefit from the robustness in a row and column sparsity, respectively. We can now state the sparse form of the subspace learning problem based on the basis as follows:

$$\min_{\mathbf{G},\mathbf{U},\mathbf{V}} f_{\mathrm{SL}}(\mathbf{G}, \mathbf{U}, \mathbf{V}) = \|\mathbf{B} - \mathbf{BGUV}\|_F^2 + \beta\|\mathbf{U}\|_{2,1} + \frac{\gamma}{2}\left(\mathrm{Tr}(\mathbf{1}_{m\times m}\mathbf{V}^T\mathbf{V}) - \mathrm{Tr}(\mathbf{V}^T\mathbf{V})\right),$$

where β and γ are the sparsity regularization parameters.

3.4 Constructing Basis for Feature Space

This section introduces a successful technique, called the Variance Basis Extension (VBE) method, to build a basis for the space generated by the features of a dataset. Let us consider the data matrix $\mathbf{X} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n]$ so that the rank of $\mathbf{X}$ is r. The Basis Extension (BE) method Meyer (2000) starts by choosing an arbitrary feature like $\mathbf{f}_{i_1}$ from among all the features and sets $B = \{\mathbf{f}_{i_1}\}$. It is clear that B is linearly independent. If $r = 1$, the basis building process by the BE method terminates at this point, and B will be a basis. Otherwise, we have $r \leq n$, and we can choose a new feature like $\mathbf{f}_{i_2}$ from $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\} - \{\mathbf{f}_{i_1}\}$ so that $B \cup \{\mathbf{f}_{i_2}\} = \{\mathbf{f}_{i_1}, \mathbf{f}_{i_2}\}$ is a linearly independent set. The process described in these two steps continues in the same way, and it can be easily proved that during the r steps, the BE method is able to construct the set $\{\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \ldots, \mathbf{f}_{i_r}\}$, which is a basis for the space generated by the whole features.

We are now in a position to make a modification to the BE method, that can lead to the selection of a suitable basis for the feature space. As seen in the first step of the BE method, we are allowed to choose any arbitrary vector from $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$. Therefore, there is an ideal situation to choose features that contain useful information from the data and can be effective in the feature selection process. As one of the most popular tools in data mining, the variance information of the data has gained widespread acceptance in some well-known dimensionality reduction methods such as the PCA and variance score methods. The primary motivation for the use of the variance information stems from its simple implementation and its power to display the amount of data dispersion. According to this advantage, the variance score method simply but effectively maintains a number of features with the highest variance score and removes the remaining features.

In the following, we introduce our framework, called the Variance Basis Extension (VBE) method, for integrating the variance information of features into the BE method with the aim of constructing a basis of features with the highest variance score. The VBE method involves the following three steps:

- For each feature $\mathbf{f}_i$, $i = 1, 2, \ldots, n$, its variance score is calculated. Then, the set of features $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$ is sorted in descending order based on their variance score. Next, we will have $\{\mathbf{f}'_1, \mathbf{f}'_2, \ldots, \mathbf{f}'_n\}$ where $\mathbf{f}'_i$ refers to the case that $\mathbf{f}'_i$ has the ith-ranked among all features based on the variance scores. Now, we set $\mathbf{X}' = [\mathbf{f}'_1, \mathbf{f}'_2, \ldots, \mathbf{f}'_n]$.
- Set $B = \{\mathbf{f}'_1\}$ so that $\mathbf{f}'_1$ has the highest variance score.
- Let $\mathrm{rank}(\mathbf{X}) = r$. For $k = 2, \ldots, n$, if $\mathbf{f}'_k \notin \mathrm{span}(B)$, set $B = B \cup \{\mathbf{f}'_k\}$. In this case, if the number of members of B is equal to r, then the VBE method terminates at this point, and B will is a basis for the space generated by the features $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$.

The main difference between the proposed VBE method and its original BE method is that the features selected by VBE have the highest variance score. This method gives a basis with the members that have the most dispersion in space and avoide the accumulation of the features as much as possible.

3.5 The Proposed GRSSLFS Method

Besed on the the previous sections, we establish the novel GRSSLFS method according to the combination of "self-representation based on the basis" and "sparse subspace learning based on the basis" as follows:

$$\min_{\mathbf{G},\mathbf{U},\mathbf{V}\geq 0} \|\mathbf{X} - \mathbf{B}\mathbf{G}\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{B}\mathbf{G}\mathbf{L}\mathbf{G}^T\mathbf{B}^T\right) + \|\mathbf{B} - \mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V}\|_F^2 + \beta\|\mathbf{U}\|_{2,1} + \frac{\gamma}{2}\left(\mathrm{Tr}(\mathbf{1}_{m\times m}\mathbf{V}^T\mathbf{V}) - \mathrm{Tr}(\mathbf{V}^T\mathbf{V})\right). \tag{9}$$

Since $\mathbf{X} \simeq \mathbf{B}\mathbf{G}$, based on the term $\|\mathbf{B} - \mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V}\|_F^2$ and in the context of subspace learning in section 3.2, we can interpret the matrix $\mathbf{X}_I \simeq \mathbf{X}\mathbf{U}$ as the desired feature selected sub-matrix. Moreover, since $\mathbf{X}_I = \mathbf{X}\mathbf{U} = [\mathbf{f}_1, \ldots, \mathbf{f}_n]\mathbf{U} = \mathbf{f}_1\mathbf{u}^{(1)} + \cdots + \mathbf{f}_n\mathbf{u}^{(n)}$, the rows $\mathbf{u}^{(i)}, i = 1, \ldots, n$ play as the feature weights. To solve Problem (9), let us consider the function $J(\mathbf{G}, \mathbf{U}, \mathbf{V})$ as follows:

$$J(\mathbf{G}, \mathbf{U}, \mathbf{V}) = \|\mathbf{X} - \mathbf{B}\mathbf{G}\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{B}\mathbf{G}\mathbf{L}\mathbf{G}^T\mathbf{B}^T\right) + \|\mathbf{B} - \mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V}\|_F^2 + \beta \mathrm{Tr}(\mathbf{U}^T\mathbf{E}\mathbf{U}) + \gamma\left(\mathrm{Tr}(\mathbf{1}_{m\times m}\mathbf{V}^T\mathbf{V}) - \mathrm{Tr}(\mathbf{V}^T\mathbf{V})\right) + \mathrm{Tr}(\mathbf{M}\mathbf{G}^T) + \mathrm{Tr}(\mathbf{N}\mathbf{U}^T) + \mathrm{Tr}(\mathbf{W}\mathbf{V}^T),$$

where $\mathbf{E} \in \mathbb{R}_+^{n\times n}$ is a diagonal matrix with the diagonal elements $\mathbf{E}_{ii} = 1/2\|\mathbf{U}_i\|_2$, for $i = 1, \ldots, n$. Moreover, $\mathbf{M} \in \mathbb{R}^{m\times n}$, $\mathbf{N} \in \mathbb{R}^{n\times n}$ and $\mathbf{W} \in \mathbb{R}^{k\times n}$ are the Lagrange multipliers. The problem mentioned above can be effectively solved by applying an alternative iterative procedure involving $\mathbf{G}$, $\mathbf{U}$, and $\mathbf{V}$. To achieve this, one variable needs to be fixed, whereas the other variables need to be determined. That is to say

- **Update G with fixed U and V.** Taking the partial derivative of J in terms of $\mathbf{G}$ shows

$$\frac{\partial J}{\partial \mathbf{G}} = -2\mathbf{B}^T\mathbf{X} + 2\mathbf{B}^T\mathbf{B}\mathbf{G} + 2\alpha\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{L} - 2\mathbf{B}^T\mathbf{B}\mathbf{V}^T\mathbf{U}^T + 2\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T + \mathbf{M}.$$

- **Update U with fixed G and V.** Taking the partial derivative of J in terms of $\mathbf{U}$ shows

$$\frac{\partial J}{\partial \mathbf{U}} = -2\mathbf{G}^T\mathbf{B}^T\mathbf{B}\mathbf{V}^T + 2\mathbf{G}^T\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V}\mathbf{V}^T + 2\beta\mathbf{E}\mathbf{U} + \mathbf{N}.$$

- **Update V with fixed G and U.** Finally, taking the partial derivative of J in terms of $\mathbf{V}$ shows

$$\frac{\partial J}{\partial \mathbf{V}} = -2\mathbf{U}^T\mathbf{G}^T\mathbf{B}^T\mathbf{B} + 2\mathbf{U}^T\mathbf{G}^T\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V} + 2\gamma(\mathbf{V}\mathbf{1}_{m\times m} - \mathbf{V}) + \mathbf{W}.$$

Now, putting the Karush-Kuhn-Tucker conditions together with $\partial J/\partial \mathbf{G} = \partial J/\partial \mathbf{U} = \partial J/\partial \mathbf{V} = 0$ results in the following updating rule:

$$\mathbf{G}_{ij} \leftarrow \mathbf{G}_{ij}\sqrt{\frac{(\mathbf{B}^T\mathbf{X} + \alpha\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{A} + \mathbf{B}^T\mathbf{B}\mathbf{V}^T\mathbf{U}^T)_{ij}}{(\mathbf{B}^T\mathbf{B}\mathbf{G} + \alpha\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{P} + \mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T)_{ij}}}, \tag{10}$$

$$\mathbf{U}_{ij} \leftarrow \mathbf{U}_{ij}\sqrt{\frac{(\mathbf{G}^T\mathbf{B}^T\mathbf{B}\mathbf{V}^T)_{ij}}{(\mathbf{G}^T\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V}\mathbf{V}^T + \beta\mathbf{E}\mathbf{U})_{ij}}}, \tag{11}$$

$$\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij}\sqrt{\frac{(\mathbf{U}^T\mathbf{G}^T\mathbf{B}^T\mathbf{B} + \gamma\mathbf{V})_{ij}}{(\mathbf{U}^T\mathbf{G}^T\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V} + \gamma\mathbf{V}\mathbf{1}_{m\times m})_{ij}}}. \tag{12}$$

Remark. Regarding the square root operation applied in the updating rules (10), (11), and (12), it is essential to highlight that the objective function J incorporates the second-order matrix polynomials concerning the variables $\mathbf{G}$, $\mathbf{U}$, and $\mathbf{V}$. To derive efficient update procedures for solving these variables, prior studies Ding et al. (2005); Ding & et al. (2006) suggest the utilization of update rules (like the ones given in 10), (11), and (12) that are founded on the concept of the square root. Algorithm 1 is a summary of the procedure developed above to solve the objective function of GRSSLFS. Moreover, the framework of the proposed GRSSLFS method is displayed in Figure 1.

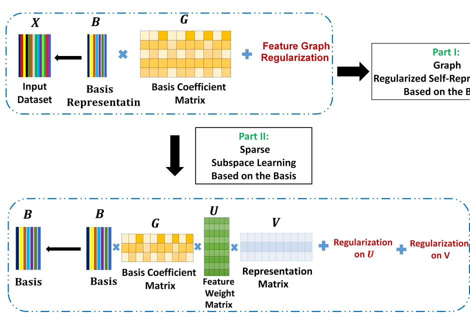


Figure 1: Framework of the proposed GRSSLFS method which integrates the self-representation based on the basis with the sparse subspace learning based on the basis.

Algorithm 1: The proposed iterative algorithm for the GRSSLFS method.

1: **Input:** Data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$; the number of selected features k.
2: Construct the basis matrix $\mathbf{B}$ using the VBE method discussed in Section 3.4.
3: Initialize the matrices $\mathbf{G}$, $\mathbf{U}$, and $\mathbf{V}$.
4: **while** not converged **do steps 4 and 6**:
5: Update $\mathbf{G}$ according to the rule (10).
6: Update $\mathbf{U}$ according to the rule (11).
7: Update $\mathbf{V}$ according to the rule (12).
8: **Output:** Put the rows of $\mathbf{U}$ in descending order of value according to the 2-norm. Next, sort the features of $\mathbf{X}$ so that the k features have the highest 2-norm score in $\mathbf{U}$.

3.6 Convergence Analysis

In the following, a detailed analysis of the monotone property of the objective function for the GRSSLFS method is conducted.

Theorem 3.2. *The objective function of the GRSSLFS method monotonously decreases according to the update rules (10), (11), and (12).*

Proof. The demonstration of the descending behavior of the objective function of the GRSSLFS method can be found in Appendix D. □

3.7 Complexity Analysis

Let m, n and k be the number of samples, the number of features and the number of selected features, respectively. In order to update the matrices $\mathbf{G} \in \mathbb{R}^{m\times n}$, $\mathbf{U} \in \mathbb{R}^{n\times k}$, and $\mathbf{V} \in \mathbb{R}^{k\times m}$, several matrix multiplications must be done in GRSSLFS. From the updating rules 10), (11), and (12, it can be found that among all the operations, the most time-consuming parts to update of $\mathbf{G}$, $\mathbf{U}$ and $\mathbf{V}$ are $\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T$, $\mathbf{G}^T\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V}\mathbf{V}^T$, and $\mathbf{U}^T\mathbf{G}^T\mathbf{B}^T\mathbf{B}\mathbf{G}\mathbf{U}\mathbf{V}$, respectively, so that the time complexity of each of them is almost equal to $O(k^2m^2n^2)$ per iteration. Besides, the required time complexity is $O(n^2)$ for constructing the feature graph. In summary, assuming that $k \leq \{m, n\}$, the computational complexity of the proposed GRSSLFS method is almost equal to $O(m^2n^2)$.

4 EXPERIMENTAL STUDIES

Results and Analysis: We evaluate the performance of the unsupervised feature selection algorithms in terms of how the selected features impact the downstream task of clustering in various datasets (refer to Appendix E for detailed information on benchmark datasets, Appendix F for the compared feature selection algorithms, Appendix G for the evaluation metrics, and Appendix H for the experimental settings details). The best value of ACC and NMI metrics at a specific number of selected features is listed in Table 1.

It is evident from these results (see Table 1, and Appendix I for more illustration) that our proposed method, GRSSLFS, outperforms all other eight methods in terms of selecting features that result in better clustering performance. There are only two exceptions (on the GLIOMA and ORL datasets) where the CAE and CD-LSR methods led to a marginally higher ACC compared to our method. For example, the ACC corresponding to CAE is 54.52 versus 54.10 in our method (only 0.7% difference in the average values). Despite the superiority of our method over other methods, its behavior across all datasets is not the same. For example, we can refer to the CNS dataset, where the NMI values are considerably lower compared to the ACC values. This trend is also observed for the other eight feature selection methods. It is mainly due to the nature of datasets, some of which pose difficulties in selecting distinguishable features and hence have a higher clustering accuracy.

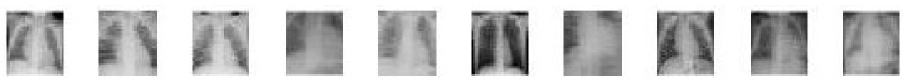
On average, the results of the VCSDFS, CAE and CD-LSR methods are closer to those of the proposed method, compared to other methods. Nevertheless, the proposed method is still superior to such methods for two main reasons. First, the basis vectors in the GRSSLFS method are constructed as an independent linear combination of the selected features. Second, the GRSSLFS method uses self-representation learning along with subspace learning in which the basis of features is used to build both the self-representation and subspace feature spaces. This combination of self-representation and subspace learning with the underlying basis of features makes the proposed method efficient compared to other methods. The effectiveness is observable in Table 1 where the consistency of our method compared to the other methods across all the different datasets is evident. For instance, the ACC values in Table 1 show that while CAE has a comparable performance with respect to GRSSLFS on the GLIOMA dataset, its performance degrades on the other datasets (refer to the Appendix section for Clustering Performance Results in Figure 3, Convergence Test in I.1, Running Time Comparison in I.2, Non-parametric Statistical Test in I.3, and Ablation Study in I.4).

Table 1: Comparing the best values of the clustering ACC and NMI of nine algorithms on all datasets. The values in parentheses denote the selected number of features at which the ACC and NMI are reported.

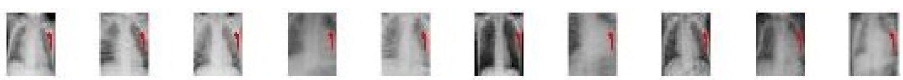
ACC										
Datasets	**Baseline**	**LS**	**CNAFS**	**OCLSP**	**RNE**	**RMFFS**	**VCSDFS**	**CAE**	**CD-LSR**	**GRSSLFS**
CNS	53.33 ± 1.45	58.66 ± 1.02 (80)	63.84 ± 1.62 (50)	65.31 ± 1.94 (70)	58.33 ± 0.00 (40)	63.33 ± 0.00 (60)	66.49 ± 2.11 (20)	66.61±0.05(60)	61.67±1.83(60)	**73.33 ± 0.01 (20)**
GLIOMA	44.05 ± 0.00	44.00 ± 0.00 (10)	50.11 ± 2.83 (40)	50.77 ± 1.03 (40)	47.11 ± 2.19 (20)	49.80 ± 5.30 (30)	51.23 ± 0.77 (30)	**54.52±3.13(40)**	48.70±3.92(10)	54.10 ± 2.71(30)
TOX171	41.25 ± 0.72	40.35 ± 0.72 (10)	48.06 ± 1.32 (90)	48.67 ± 0.89 (40)	51.69 ± 0.48 (40)	49.64 ± 1.59 (20)	52.69 ± 1.26(40)	53.36±1.24(40)	48.77±4.18(10)	**57.39 ± 1.27 (40)**
SRBCT	25.66 ± 2.82	46.02 ± 2.96 (100)	44.17 ± 1.55 (60)	44.27 ± 2.07 (60)	42.12 ± 3.37 (80)	48.97 ± 4.67 (90)	44.39 ± 0.91 (30)	46.08±4.18(90)	52.63±4.39(10)	**53.49 ± 3.11 (60)**
SMK	51.34 ± 0.56	55.11± 2.35 (80)	60.12 ± 1.72 (40)	62.31 ± 3.00 (100)	61.51 ± 0.34 (10)	52.94± 2.31 (50)	62.91 ± 2.62 (40)	61.80±2.06(40)	64.71±2.72(30)	**65.72 ± 2.59 (10)**
ATT	63.84 ± 4.03	46.75 ±3.87(60)	58.68 ±2.78(100)	59.71 ±1.94(100)	59.08 ± 3.12(50)	46.25±3.48(30)	61.97 ± 1.83 (100)	56.86±2.79(90)	58.03±3.33(100)	**64.11 6± 2.65(100)**
ORL	50.25 ± 2.91	38.51 ± 1.66(90)	49.17±2.52(90)	50.45 ±3.41(100)	51.31 ±3.01(100)	47.71±2.73(70)	47.51±1.38(100)	50.09±2.42(100)	**54.05±3.51(100)**	53.45 ± 3.13(80)
warpAR10P	26.88 ± 2.18	21.50 ±0.94(10)	38.83 ±2.28(10)	41.68±3.31(10)	34.62 ±2.45(10)	41.55±3.21(90)	36.85 ±0.73(10)	39.26±2.83(10)	31.94±2.78(30)	**46.66 ± 3.51(10)**

NMI										
Datasets	**Baseline**	**LS**	**CNAFS**	**OCLSP**	**RNE**	**RMFFS**	**VCSDFS**	**CAE**	**CD-LSR**	**GRSSLFS**
CNS	1.17 ± 0.00	2.44 ± 0.45 (90)	8.15 ± 3.09 (70)	10.76 ± 3.68 (70)	2.29 ± 0.00 (50)	9.35 ± 0.08 (50)	10.35± 2.62 (50)	11.05±0.41(30)	13.76±0.06(20)	**18.56 ± 0.00 (70)**
GLIOMA	17.94± 0.69	18.67± 0.29 (50)	27.21 ± 2.43 (20)	28.93 ± 1.86 (30)	25.84 ± 4.20 (20)	29.81± 1.46 (30)	28.92 ± 1.29 (40)	**32.83±1.83(40)**	25.35±1.68(70)	32.09 ± 2.14 (40)
TOX171	13.54± 0.23	11.68± 1.97 (90)	23.47 ± 1.98(80)	25.27 ± 2.31 (80)	26.78 ± 0.08 (40)	24.60± 0.41 (20)	30.98 ± 0.77 (100)	31.12±0.31(40)	31.28±0.93(10)	**36.96 ± 0.60 (100)**
SRBCT	11.55 ± 3.47	56.23± 4.06 (60)	41.21 ± 4.36 (60)	42.44 ± 2.68 (90)	29.48 ± 3.52 (80)	47.82± 4.13 (90)	48.00 ± 4.01 (90)	45.28±3.59(80)	49.53±3.86(30)	**59.67 ± 4.18 (90)**
SMK	0.01 ± 0.00	2.17± 0.05 (80)	3.75 ± 0.09 (40)	4.72 ± 0.22 (30)	3.77 ± 0.08 (10)	3.21± 0.19 (70)	5.63 ± 1.55 (40)	3.51±0.09(80)	7.12±1.05(30)	**8.20 ± 0.53 (10)**
ATT	81.19±1.77	68.46 ± 1.37(70)	77.05± 1.44(90)	78.61 ±2.53(100)	77.17 ± 1.04(80)	65.27±2.03(50)	78.62 ±0.81(100)	75.27±2.24(90)	77.02±2.19(100)	**81.71± 1.31(100)**
ORL	72.68 ±1.27	63.85 ±1.23(100)	71.44 ±1.16(100)	72.19 ±2.34(100)	72.59±1.33(100)	69.72 ±1.69(70)	73.02 ±1.49(100)	73.64±1.41(100)	72.57±2.02(100)	**74.56 ± 1.72(80)**
warpAR10P	28.48 ± 3.31	19.39 ±1.84(40)	43.00 ±1.64(80)	45.46 ±3.08(20)	35.85±4.38(20)	46.60±2.94(90)	41.62 ±1.22(80)	42.35±2.23(80)	34.22±3.37(20)	**48.63 ± 2.83(10)**

Application to the PneumoniaMNIST dataset: In patients diagnosed with pneumonia, chest X-rays reveal various diagnostic features, particularly in the cardiac section. Notable among these characteristics are an augmented cardiac shadow, alterations in cardiac positioning, the presence of effusion, the pericardial effusion, and the instances of cardiomegaly Reed (2017). In this section, the application of our proposed feature selection method to the PneumoniaMNIST dataset Yang et al. (2021) is assessed with the goal of precisely identifying and analyzing the cardiac silhouette in chest X-ray images. The PneumoniaMNIST dataset, designed for Chest X-Ray radiographs and forming part of the MedMNIST collection Yang et al. (2023), comprises 5,856 Chest X-Ray radiograph images, each measuring 28x28 pixels in grayscale. To this end, we test the representativeness of the selected features using GRSSLFS by running it on the PneumoniaMNIST image database. Here, the first ten Chest X-Ray radiograph images are chosen from this dataset. Moreover, we select $\{10, 50, 100\}$ pixels by running GRSSLFS. Figure 2 shows the outputs, where the original and the feature-selected images for each Chest X-Ray radiograph image are plotted in the corresponding row from left to right. In Figure 2, the selected features are depicted by red dots and the un-selected features remain at their original pixel colour. As can be seen, when the number of selected features increases from 10 to 100, subsequently GRSSLFS is able to capture the most *representative* parts of the images, like the central part of the chest, which contains the most discriminative features of the Chest X-Ray radiograph images and can detect the cardiac silhouette. The performance of the obtained feature-selected images has been validated through *assessment by radiologists*, which demonstrates the proficiency of our proposed method in capturing the fundamental features of the PneumoniaMNIST dataset. This capability shows the *interpretability and explainability* power of our model. The reason for this selection may lie in the cooperation of the basis matrix in self-representation and subspace learning terms of our objective function. Finally, Section J reports the results pertaining to the selection of 100 features by other comparative methods. Examining this figure, it is evident that some methods, including our proposed method, RMFFS, and VCSDFS, exhibit a tendency to identifythe central area of chest X-ray images. Conversely, some other methods, while successfully identifying portions of the central areas, may overlook additional regions that could provide valuable information. These methods fail to capture such information in radiological analysis.

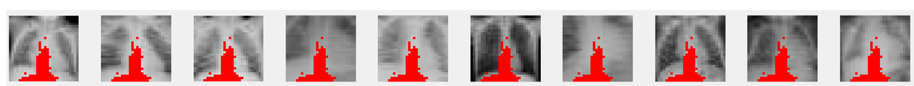


(a) PneumoniaMNIST - Original Images



(b) PneumoniaMNIST - Selected Features = 10

(c) PneumoniaMNIST - Selected Features = 50



(d) PneumoniaMNIST - Selected Features = 100

Figure 2: The visualization of selected features on PneumoniaMNIST images.

5 CONCLUSION

In this paper, we addressed the self-representation learning problem by establishing a basis within the feature space, with the primary aim of eliminating unnecessary and redundant features for the representation of the initial data. The motivation behind this work stems from a fundamental concept in linear algebra known as the basis for a given vector space. Essentially, this concept implies that every feature vector can be expressed as a linear combination of a basis set of features, which are linearly independent. Leveraging this notion, we also designed a subspace learning framework to select features with minimal redundancy. Through the introduction of these novel variations in self-representation and subspace learning, we proposed an unsupervised feature selection method called Graph Regularized Self-Representation and Sparse Subspace Learning (GRSSLFS). Comparative experiments conducted in this study demonstrated that the proposed GRSSLFS method exhibits a high level of effectiveness in feature selection.

The performance of our proposed GRSSLFS method is closely tied to the choice of the feature space basis. As discussed, GRSSLFS employs the variance information of features to establish a basis for the feature space. However, it raises the question: what alternative mathematical or statistical tools could be used instead of variance information for constructing such a basis? This prompts us to recognize the challenge of devising a meaningful basis for the feature space through a potentially more optimized process. As such, the pursuit of constructing an optimal basis for the feature space remains an open problem within the domain of unsupervised feature selection, warranting further research in future studies.

REFERENCES

Muhammed Fatih Balın, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 444–453. PMLR, 2019.

Verónica Bolón-Canedo, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282:111–135, 2014.

Hao Chen, Hongmei Chen, Weiyi Li, Tianrui Li, Chuan Luo, and Jihong Wan. Robust dual-graph regularized and minimum redundancy based on self-representation for semi-supervised feature selection. *Neurocomputing*, 490:104–123, 2022.

Chris Ding and et al. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

Chris Ding, Xiaofeng He, and Horst D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2005.

Vu C Dinh and Lam S Ho. Consistent feature selection for analytic deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2420–2431. Curran Associates, Inc., 2020.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Benyamin Ghojogh, Mark Crowley, Fakhri Karray, and Ali Ghodsi. Principal component analysis. In *Elements of Dimensionality Reduction and Manifold Learning*, pp. 123–154. Springer International Publishing, Cham, 2023.

Xiaofei He and Partha Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems*, volume 16, pp. 153–160, 2003.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems*, volume 18, pp. 507–514, 2005.

Saeed Karami, Farid Saberi-Movahed, Prayag Tiwari, Pekka Marttinen, and Sahar Vahdat. Unsupervised feature selection based on variance–covariance subspace distance. *Neural Networks*, 166(1): 188–203, 2023.

Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, volume 13, pp. 556–562, 2001.

Changsheng Li, Kaihang Mao, Lingyan Liang, Dongchun Ren, Wei Zhang, Ye Yuan, and Guoren Wang. Unsupervised active learning via subspace learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8332–8339, 2021.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):94, 2018.

Weiyi Li, Hongmei Chen, Tianrui Li, Jihong Wan, and Binbin Sang. Unsupervised feature selection via self-paced learning and low-redundant regularization. *Knowledge-Based Systems*, 240:108150, 2022.

Sen Liang, Anjun Ma, Sen Yang, Yan Wang, and Qin Ma. A review of matched-pairs feature selection methods for gene expression data analysis. *Computational and structural biotechnology journal*, 16:88–97, 2018.

Xiaochang Lin, Jiewen Guan, Bilian Chen, and Yifeng Zeng. Unsupervised feature selection via orthogonal basis clustering and local structure preserving. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, 33(11):6881–6892, 2022.

Yanfang Liu, Dongyi Ye, Wenbin Li, Huihui Wang, and Yang Gao. Robust neighborhood embedding for unsupervised feature selection. *Knowledge-Based Systems*, 193:105462, 2020.

Chuan Luo, Jian Zheng, Tianrui Li, Hongmei Chen, Yanyong Huang, and Xi Peng. Orthogonally constrained matrix factorization for robust unsupervised feature selection with local preserving. *Information Sciences*, 586:662–675, 2022.

Subhabrata Majumdar and Snigdhansu Chatterjee. Feature selection using e-values. In *International Conference on Machine Learning*, pp. 14753–14773. PMLR, 2022.

Sanparith Marukatat. Tutorial on PCA and approximate kernel PCA. *Artificial Intelligence Review*, 56(6):5445–5477, 2023.

Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. SIAM, 2000.

Mohsen Ghassemi Parsa, Hadi Zare, and Mehdi Ghatee. Unsupervised feature selection based on adaptive similarity learning and subspace clustering. *Engineering Applications of Artificial Intelligence*, 95:103855, 2020.

Miao Qi, Ting Wang, Fucong Liu, Baoxue Zhang, Jianzhong Wang, and Yugen Yi. Unsupervised feature selection by regularized matrix factorization. *Neurocomputing*, 273:593–610, 2018.

James C Reed. *Chest Radiology: Patterns and Differential Diagnoses*. Elsevier Health Sciences, 2017.

Dongchun Ren, Wei Zhang, Ye Yuan, and Guoren Wang. Unsupervised active learning via subspace learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8332–8339, 2021.

Giorgio Roffo, Simone Melzi, Umberto Castellani, Alessandro Vinciarelli, and Marco Cristani. Infinite feature selection: A graph-based feature filtering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4396–4410, 2021.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Ronghua Shang, Kaiming Xu, Fanhua Shang, and Licheng Jiao. Sparse and low-redundant subspace learning-based dual-graph regularized robust feature selection. *Knowledge-Based Systems*, 187: 104830, 2020.

Ronghua Shang, Lujuan Wang, Fanhua Shang, Licheng Jiao, and Yangyang Li. Dual space latent representation learning for unsupervised feature selection. *Pattern Recognition*, 114:107873, 2021.

Chenyu Shao, Mulin Chen, Yuan Yuan, and Qi Wang. Projection concept factorization with self-representation for data clustering. *Neurocomputing*, 517:62–70, 2023.

Chao Sheng, Peng Song, Weijian Zhang, and Dongliang Chen. Dual-graph regularized subspace learning based feature selection. *Digital Signal Processing*, 117:103175, 2021.

Dan Shi, Lei Zhu, Jingjing Li, Zheng Zhang, and Xiaojun Chang. Unsupervised adaptive feature selection with binary hashing. *IEEE Transactions on Image Processing*, 32:838–853, 2023. doi: 10.1109/TIP.2023.3234497.

Saúl Solorio-Fernández, J Ariel Carrasco-Ochoa, and José Fco Martínez-Trinidad. A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53(2):907–948, 2020.

Yu Sun, Haicheng Li, Lei Zheng, Jinzhao Li, Yan Hong, Pengfei Liang, Lai-Yu Kwok, Yongchun Zuo, Wenyi Zhang, and Heping Zhang. iProbiotics: a machine learning platform for rapid identification of probiotic properties from whole-genome primary sequences. *Briefings in Bioinformatics*, 23(1): bbab477, 2022.

Chang Tang, Xinwang Liu, Miaomiao Li, Pichao Wang, Jiajia Chen, Lizhe Wang, and Wanqing Li. Robust unsupervised feature selection via dual self-representation and manifold regularization. *Knowledge-Based Systems*, 145:109–120, 2018.

Chang Tang, Meiru Bian, Xinwang Liu, Miaomiao Li, Hua Zhou, Pichao Wang, and Hailin Yin. Unsupervised feature selection via latent representation learning and manifold regularization. *Neural Networks*, 117:163–178, 2019.

Yongqiang Tang, Yuan Xie, and Wensheng Zhang. Affine subspace robust low-rank self-representation: From matrix to tensor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9357–9373, 2023.

Shiping Wang, Witold Pedrycz, Qingxin Zhu, and William Zhu. Subspace learning for unsupervised feature selection via matrix factorization. *Pattern Recognition*, 48(1):10–19, 2015a.

Shiping Wang, Witold Pedrycz, Qingxin Zhu, and William Zhu. Unsupervised feature selection via maximum projection and minimum redundancy. *Knowledge-Based Systems*, 75:19–29, 2015b.

Yu Wang, Liang Hu, Wanfu Gao, Xiaofeng Cao, and Yi Chang. Adans: Adaptive negative sampling for unsupervised graph representation learning. *Pattern Recognition*, 136:109266, 2023.

Wenhui Wu, Yujie Chen, Ran Wang, and Le Ou-Yang. Self-representative kernel concept factorization. *Knowledge-Based Systems*, 259:110051, 2023.

Lei Xu, Rong Wang, Feiping Nie, and Xuelong Li. Efficient top-k feature selection using coordinate descent method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10594–10601, 2023.

Jiancheng Yang, Rui Shi, and Bingbing Ni. MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, 2021.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Aihong Yuan, Mengbo You, Dongjian He, and Xuelong Li. Convex non-negative matrix factorization with adaptive graph for unsupervised feature selection. *IEEE TRANSACTIONS ON CYBERNETICS*, 52(6):5522–5534, 2022.

Yongshan Zhang, Xinxin Wang, Xinwei Jiang, and Yicong Zhou. Robust dual graph self-representation for unsupervised hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

Wei Zheng, Hui Yan, and Jian Yang. Robust unsupervised feature selection by nonnegative sparse subspace learning. *Neurocomputing*, 334:156–171, 2019.

A MOTIVATION

The main goal of the feature selection process is to opt for a subset of features with the purpose of diminishing noise and improving the efficiency and effectiveness of a learning task, such as classification and clustering. In essence, this process of selection aims to identify features that are important for the learning task while ensuring that the chosen set of features is non-redundant, thereby strengthening robustness. Within this framework, a crucial factor in feature selection is redundancy, which involves examining the similarity among features and assessing how effective it is to introduce a new feature into a designated feature set for data analysis.

In the field of linear algebra, a direct connection exists between redundancy and the concept of linear independence. Specifically, a set of vectors in a vector space is deemed linearly dependent if non-zero scalars can be identified in a way that their linear combination equals zero. This suggests that it is possible to express one of these vectors as a combination of the others. In contrast, a collection of vectors is deemed linearly independent if there is no linear dependence among them. Consequently, we can deduce from this principle that a set of vectors exhibiting linear dependence includes redundancy, enabling the removal of one vector from the initial set without altering the span of the remaining vectors. A significant benefit of eliminating linearly dependent features from the initial feature set is that it aids in fulfilling a crucial aim in feature selection, which is the reduction of redundancy. In the pursuit of this goal, the concept of a basis, a fundamental element in the theory of linear vector spaces, becomes crucial. Each basis for the feature space exhibits two essential traits. Firstly, it covers the entire feature space, and secondly, all features within a basis are linearly independent. These attributes positively and constructively influence the discriminative feature selection process. The initial property guarantees that a basis can represent the complete feature space through its elements, implying that the fundamental characteristics of the original features are present in the basis with a significantly reduced number of elements compared to the entire feature space. The second characteristic of a basis, which involves its elements being linearly independent, leads to a significant reduction in data redundancy. Exploiting the advantages of a basis in a vector space, the primary objective of this paper is to further investigate the impact of a basis for the feature space on tasks related to feature learning.

B RELATED WORK

Table 2 outlines the key characteristics of some unsupervised feature selection methods based on subspace learning or self-representation.

Table 2: A comparison of some related unsupervised feature selection methods.

Method	Subspace learning	Self-representation	Graph Regularization	Sparse Regularization	Orthogonality Constraint
MFFS Wang et al. (2015a)	✓	×	×	×	✓
MPMR Wang et al. (2015b)	✓	×	×	×	✓
RMFFS Qi et al. (2018)	✓	×	×	✓	×
DSRMR Tang et al. (2018)	×	✓	✓	✓	×
NSSLFS Zheng et al. (2019)	✓	×	×	✓	×
SCFS Parsa et al. (2020)	✓	✓	×	✓	✓
RNE Liu et al. (2020)	✓	×	×	×	✓
SLSDR Shang et al. (2020)	✓	×	✓	✓	✓
DGSLFS Sheng et al. (2021)	✓	×	✓	✓	✓
CNAFS Yuan et al. (2022)	✓	×	✓	✓	✓
SPLR Li et al. (2022)	✓	×	✓	✓	✓
RDMRS2FS Chen et al. (2022)	×	✓	✓	✓	×
GRSSLFS (Ours)	✓	✓	✓	✓	×

C THE PROOF FOR THEOREM 3.1

If the matrix $\mathbf{U}$ is defined as $\mathbf{U} = [\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_m}]$, where $\mathbf{u}_{i_j}$ (for $j = 1, \ldots, m$) is a vector whose i_j-element is 1 and other elements are 0, then it can be easily seen that $\mathbf{B} = \mathbf{XU}$. With this assumption, the distance-based subspace learning problem (6) will be of the following form:

$$\mathrm{dist}_{\mathrm{SL}}(\mathbf{X}, \mathbf{XU}) = \mathrm{dist}_{\mathrm{SL}}(\mathbf{X}, \mathbf{B}) = \min_{\mathbf{W} \in \mathbb{R}^{m \times n}} \|\mathbf{X} - \mathbf{BW}\|_F.$$

It turns out from the above problem that $\text{dist}_{\text{SL}}(\mathbf{X}, \mathbf{B}) \leq \|\mathbf{X} - \mathbf{BW}\|_F$, for each $\mathbf{W} \in \mathbb{R}^{m \times n}$. On the other hand, since $\mathbf{B}$ is a basis for the space generated by $\mathbf{X}$, there is a basis coefficient matrix $\mathbf{G} = [g_{ij}] \in \mathbb{R}^{m \times n}$ such that $\mathbf{X} = \mathbf{BG}$. Taking this observation into account and assuming $\mathbf{W} = \mathbf{G}$, it becomes evident that $0 \leq \text{dist}_{\text{SL}}(\mathbf{X}, \mathbf{B}) \leq \|\mathbf{X} - \mathbf{BG}\|_F = 0$, which completes the proof.

D The Proof for the Convergence Analysis

This section investigates the convergence of the iterative updating rules 10), (11), and (12, and shows that the objective function of the GRSSLFS method is non-increasing with respect to these updating rules.

The proof starts by assuming that the matrices $\mathbf{U} \geq 0$ and $\mathbf{V} \geq 0$ are fixed, and it is shown that the objective function of GRSSLFS monotonously decreases in terms of the other variable $\mathbf{G} \geq 0$. With this assumption, we define the following function:

$$Z(\mathbf{G}) = \|\mathbf{X} - \mathbf{BG}\|_F^2 + \alpha \text{Tr}\left(\mathbf{BGLG}^T\mathbf{B}^T\right) + \|\mathbf{B} - \mathbf{BGUV}\|_F^2.$$

Using simple calculations, the following can be deduced:

$$\begin{aligned} Z(\mathbf{G}) = &\ \text{Tr}(\mathbf{X}^T\mathbf{X}) - 2\text{Tr}(\mathbf{X}^T\mathbf{BG}) + \text{Tr}(\mathbf{G}^T\mathbf{B}^T\mathbf{BG}) \\ &+ \alpha \text{Tr}\left(\mathbf{BG}(\mathbf{P} - \mathbf{A})\mathbf{G}^T\mathbf{B}^T\right) + \text{Tr}(\mathbf{B}^T\mathbf{B}) \\ &- 2\text{Tr}(\mathbf{B}^T\mathbf{BGUV}) + \text{Tr}(\mathbf{V}^T\mathbf{U}^T\mathbf{G}^T\mathbf{B}^T\mathbf{BGUV}). \end{aligned}$$

Additionally, suppose that $\mathbf{G}' \geq 0$ is given. It can be easily shown that

$$\text{Tr}(\mathbf{G}^T\mathbf{MGN}) \leq \sum_{ij} \frac{(\mathbf{MG}'\mathbf{N})_{ij}\mathbf{G}_{ij}^2}{\mathbf{G}'_{ij}}, \tag{13}$$

for any matrices $\mathbf{M} \in \mathbb{R}^{m \times m}$ and $\mathbf{N} \in \mathbb{R}^{n \times n}$. Using the inequality (13) in conjunction with the inequality $c > 1 + \log(c)$, for any $c > 0$, we have

- $$-\text{Tr}(\mathbf{X}^T\mathbf{BG}) \leq -\sum_{l,r}\left((\mathbf{B}^T\mathbf{X})_{lr}\mathbf{G}'_{lr}\left(1 + \log\frac{\mathbf{G}_{lr}}{\mathbf{G}'_{lr}}\right)\right), \tag{14}$$
- $$\text{Tr}(\mathbf{G}^T\mathbf{B}^T\mathbf{BG}) \leq \sum_{l,r}(\mathbf{B}^T\mathbf{BG}')_{lr}\frac{\mathbf{G}_{lr}^2}{\mathbf{G}'_{lr}}, \tag{15}$$
- $$\text{Tr}(\mathbf{BGPG}^T\mathbf{B}^T) \leq \sum_{l,r}(\mathbf{B}^T\mathbf{BG}'\mathbf{P})_{lr}\frac{\mathbf{G}_{lr}^2}{\mathbf{G}'_{lr}}, \tag{16}$$
- $$-\text{Tr}(\mathbf{BGAG}^T\mathbf{B}^T) \leq -\sum_{l,r,q}\left((\mathbf{B}^T\mathbf{BG}'\mathbf{A})_{ql}\mathbf{G}'_{lr}\left(1 + \log\frac{\mathbf{G}_{ql}\mathbf{G}_{lr}}{\mathbf{G}'_{ql}\mathbf{G}'_{lr}}\right)\right), \tag{17}$$
- $$-\text{Tr}(\mathbf{B}^T\mathbf{BGUV}) \leq -\sum_{l,r}\left((\mathbf{B}^T\mathbf{BV}^T\mathbf{U}^T)_{lr}\mathbf{G}'_{lr}\left(1 + \log\frac{\mathbf{G}_{lr}}{\mathbf{G}'_{lr}}\right)\right), \tag{18}$$
- $$\text{Tr}(\mathbf{V}^T\mathbf{U}^T\mathbf{G}^T\mathbf{B}^T\mathbf{BGUV}) \leq \sum_{l,r}(\mathbf{B}^T\mathbf{BG}'\mathbf{UVV}^T\mathbf{U}^T)_{lr}\frac{\mathbf{G}_{lr}^2}{\mathbf{G}'_{lr}}). \tag{19}$$

Let us now define the following function:

$$\begin{aligned}\Omega(\mathbf{G},\mathbf{G}') &= \mathrm{Tr}(\mathbf{X}^T\mathbf{X}) - 2\sum_{l,r}\left((\mathbf{B}^T\mathbf{X})_{lr}\mathbf{G}'_{lr}\left(1+\log\frac{\mathbf{G}_{lr}}{\mathbf{G}'_{lr}}\right)\right)\\ &+ \sum_{l,r}(\mathbf{B}^T\mathbf{B}\mathbf{G}')_{lr}\frac{\mathbf{G}_{lr}^2}{\mathbf{G}'_{lr}} + \alpha\sum_{l,r}(\mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{P})_{lr}\frac{\mathbf{G}_{lr}^2}{\mathbf{G}'_{lr}}\\ &- \alpha\sum_{l,r,q}\left((\mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{A})_{ql}\mathbf{G}'_{lr}\left(1+\log\frac{\mathbf{G}_{ql}\mathbf{G}_{lr}}{\mathbf{G}'_{ql}\mathbf{G}'_{lr}}\right)\right)\\ &+ \mathrm{Tr}(\mathbf{B}^T\mathbf{B}) - 2\sum_{l,r}\left((\mathbf{B}^T\mathbf{B}\mathbf{V}^T\mathbf{U}^T)_{lr}\mathbf{G}'_{lr}\left(1+\log\frac{\mathbf{G}_{lr}}{\mathbf{G}'_{lr}}\right)\right)\\ &+ \sum_{l,r}(\mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T)_{lr}\frac{\mathbf{G}_{lr}^2}{\mathbf{G}'_{lr}}.\end{aligned}$$

Now, in view of the relations (14)-(19), it turns out that (1) $\Omega(\mathbf{G},\mathbf{G}) = Z(\mathbf{G})$ and (2) $Z(\mathbf{G}) \leq \Omega(\mathbf{G},\mathbf{G}')$, for any $\mathbf{G} \geq 0$. Taking these observations into account, it can be inferred that $\Omega(\mathbf{G},\mathbf{G})$ represents an auxiliary function for $Z(\mathbf{G})$ Lee & Seung (2001). Consequently, using the following relation

$$\mathbf{G}^* = \arg\min_{\mathbf{G}} \Omega(\mathbf{G},\mathbf{G}'), \tag{20}$$

the objective function of the GRSSLFS method monotonously decreases in terms of the variable $\mathbf{G} \geq 0$. In order to compute the minimization problem (20), the derivative of $\Omega(\mathbf{G},\mathbf{G}')$ in terms of $\mathbf{G}_{ij}$ leads us to the following relation:

$$\begin{aligned}\frac{\partial\Omega(\mathbf{G},\mathbf{G}')}{\partial\mathbf{G}_{ij}} &= -2(\mathbf{B}^T\mathbf{X})_{ij}\frac{\mathbf{G}'_{ij}}{\mathbf{G}_{ij}} + 2(\mathbf{B}^T\mathbf{B}\mathbf{G}')_{ij}\frac{\mathbf{G}_{ij}}{\mathbf{G}'_{ij}}\\ &+ 2\alpha(\mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{P})_{ij}\frac{\mathbf{G}_{ij}}{\mathbf{G}'_{ij}} - 2\alpha(\mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{A})_{ij}\frac{\mathbf{G}'_{ij}}{\mathbf{G}_{ij}}\\ &- 2(\mathbf{B}^T\mathbf{B}\mathbf{V}^T\mathbf{U}^T)_{ij}\frac{\mathbf{G}'_{ij}}{\mathbf{G}_{ij}}\\ &+ 2(\mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T)_{ij}\frac{\mathbf{G}_{ij}}{\mathbf{G}'_{ij}}.\end{aligned}$$

Thus, assuming that $\partial\Omega(\mathbf{G},\mathbf{G}')/\partial\mathbf{G}_{ij} = 0$, it follows that

$$\begin{aligned}&\left(\mathbf{B}^T\mathbf{X} + \alpha\mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{A} + \mathbf{B}^T\mathbf{B}\mathbf{V}^T\mathbf{U}^T\right)_{ij}(\mathbf{G}'_{ij})^2 =\\ &\left(\mathbf{B}^T\mathbf{B}\mathbf{G}' + \alpha\mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{P} + \mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T\right)_{ij}\mathbf{G}_{ij}^2,\end{aligned}$$

which leads to the following relation:

$$\mathbf{G}_{ij} \leftarrow \mathbf{G}'_{ij}\sqrt{\frac{(\mathbf{B}^T\mathbf{X} + \alpha\mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{A} + \mathbf{B}^T\mathbf{B}\mathbf{V}^T\mathbf{U}^T)_{ij}}{(\mathbf{B}^T\mathbf{B}\mathbf{G}' + \alpha\mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{P} + \mathbf{B}^T\mathbf{B}\mathbf{G}'\mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T)_{ij}}}. \tag{21}$$

In conclusion, solving the minimization problem (20) leads to a solution in the form shown in (21), which is exactly consistent with the relation (10) introduced as the update rule for variable $\mathbf{G}$. Here, the proof for the monotone property of the objective function of the GRSSLFS method in terms of $\mathbf{G}$ is completed. Similar to what has been explained above for the case of $\mathbf{G}$, it is possible to prove the monotone property of the objective function of GRSSLFS in terms of $\mathbf{U}$ and $\mathbf{V}$, which are omitted due to the similarity in the process of existing relations.

E DATASETS

Table 3 demonstrates the eight benchmark datasets, which are used to perform our experiments. We selected a mix of five different gene expression datasets and three facial image datasets to diversify our data samples and hence to better evaluate the performance of our proposed feature selection method. It should be noted that all of these datasets are accessible through the scikit-feature feature selection repository* Li et al. (2018) and the UCI Machine Learning Repository† Dua & Graff (2017).

Table 3: The information of eight datasets used in this study.

ID	Dataset	# of Samples	# of Features	# of Classes	Type
1	**CNS**	60	7129	2	Biological
2	**GLIOMA**	50	4434	4	Biological
3	**TOX_171**	171	5748	4	Biological
4	**SRBCT**	83	2308	4	Biological
5	**SMK**	187	19993	2	Biological
6	**ATT**	400	10304	40	Image
7	**ORL**	400	1024	40	Image
8	**warpAR10P**	130	2400	10	Image

F Comparison Methods

In this study, the following unsupervised feature selection methods are selected with their track records of performance in the literature to be compared with GRSSLFS. These methods are briefly explained below.

1. **Baseline**: All features are considered.
2. **Laplacian Score (LS)** He et al. (2005): The features selected by this method can best preserve the local manifold structure of the original data.
3. **Convex non-negative matrix Factorization with an Adaptive graph constraint Feature Selection (CNAFS)** Yuan et al. (2022): Through convex matrix factorization with adaptive graph constraint, it selects the best feature subset by considering the correlation between the data while preserving the local manifold structure.
4. **Orthogonal basis clustering and Local Structure Preserving(OCLSP)** Lin et al. (2022): An orthogonal basis clustering and an adaptive graph regularization are used to gain cluster separation while keeping the local information of data.
5. **Regularized Matrix Factorization Feature Selection (RMFFS)** Qi et al. (2018): This method modifies its objective function for feature selection by adding an inner product regularization to the matrix factorization.
6. **Robust Neighborhood Embedding (RNE)** Liu et al. (2020): In this method, features are selected using the neighborhood embedding and ℓ_1 norm as the loss function.
7. **Variance-Covariance Subspace Distance based Feature Selection (VCSDFS)** Karami et al. (2023): Variance-Covariance subspace distance is used to select the subset which contains the most representative features.
8. **Concrete Autoencoders (CAE)** Balın et al. (2019): CAE has an encoder-decoder architecture with a concrete selector layer as the encoder and a standard neural network as the decoder. It aims at selecting the most informative subset of global features, which are used to simultaneously train a neural network to reconstruct the input data.
9. **Coordinate Descent based Least Square Regression (CD-LSR)** Xu et al. (2023): This method leverages a coordinate descent-based optimization framework to solve the general $l_{2,0}$-norm constrained feature selection problem.

G Evaluation Metrics

In the experiments, the clustering performance of feature selection algorithms is evaluated using Accuracy (ACC) and Normalized Mutual Information (NMI) metrics Solorio-Fernández et al. (2020).

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

†https://archive.ics.uci.edu/ml/index.php

The higher ACC and NMI values are, the better the clustering performance is, which in turn results from better performance of the feature selection algorithm to select informative features.
The ACC is defined as:

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

Where p_i and q_i are, respectively, the clustering and the ground-truth labels of the sample x_i, and δ is the Kronecker delta function. In addition, map(.) denotes the best permutation mapping function that uses the Kuhn–Munkres algorithm to map each clustering label to the corresponding ground-truth label.
The NMI is defined as

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

Where P and Q are the set of centers of the predicted and the ground-truth clusters, and $MI(.)$ is the mutual information. Furthermore, $H(P)$ and $H(Q)$ denote the entropies of P and Q, respectively.

H EXPERIMENTAL SETTINGS

Our experiments are performed for various numbers of selected features in the range of $\{10, 20, \ldots, 100\}$, for all datasets. Furthermore, hyperparameter tuning is done to find the most optimum configuration of various feature selection models in our study. We fix $k = 5$ for the k-nearest neighbor algorithm and for all the datasets in LS, OCLSP and RNE methods. For CNAFS, the hyperparameters $\alpha, \beta, \gamma, \lambda$ and ϵ are tuned by searching from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ as mentioned in author's literature in lYuan et al. (2022). To implement OCLSP, Lin et al. (2022), the parameter α is fixed 10^4 and η, γ, β are tuned by selecting from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. For RMFFS, we followed the work in Qi et al. (2018) and searched the value of the parameter β in the range of $\{1, 10, \ldots, 10^{8}\}$. In VCSDFS, the values of the parameter ρ is selected from $\{10^{-6}, 10^{-5}, \ldots, 10^{6}\}$, Karami et al. (2023). We adopted the values for the parameters of RNE from Liu et al. (2020), and let the penalty coefficient α be 10^3. For the CAE method, we used the values in the original paper Balın et al. (2019). The performance of CD-LSR was assessed using the provided code based on the author's implementation as described in the original paper of the method Xu et al. (2023). We tuned the parameters α, β and γ of GRSSLFS method on $\{10^{-5}, \ldots, 10^{5}\}$. Finally, due to the random initialization in the k-means clustering algorithm, we repeated the clustering runs 20 times with different random starting points over which we reported the average and standard deviations of the ACC and NMI metrics. Finally, it should be noted that our tests were conducted utilizing Matlab 2018a on a personal computer equipped with 16GB of RAM and an Intel Core i5-4690 processor.

I EXPERIMENTAL RESULTS

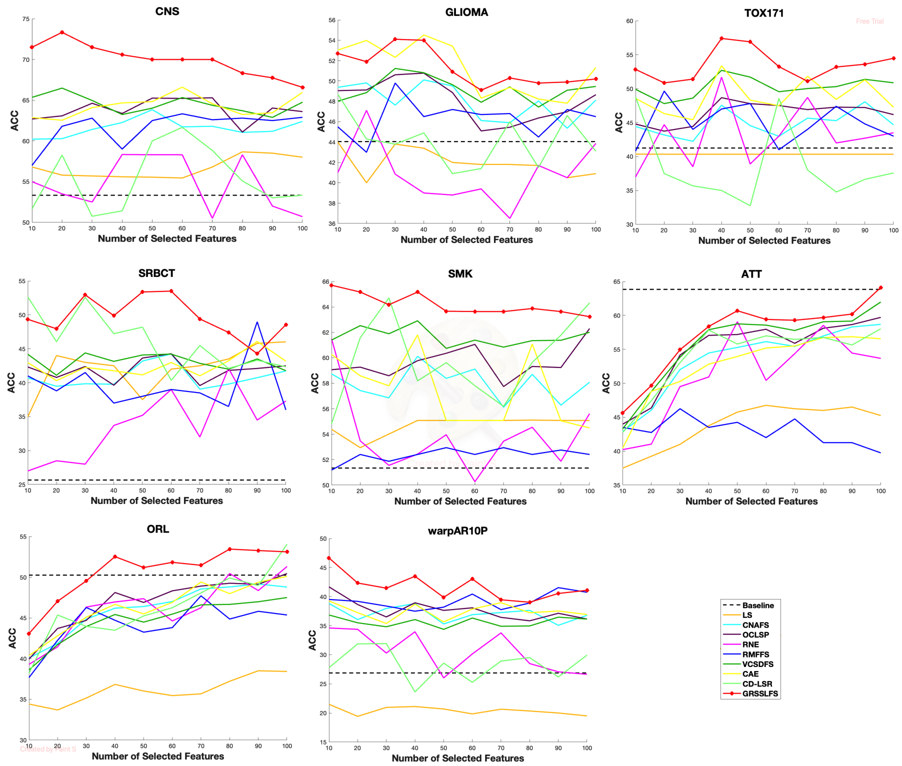
For more illustration of the reported clustering performance of all comparison methods, in Figure 3, we depict the values of ACC and NMI metrics for different numbers of selected features. To summarize the final clustering performance, the average value of ACC and NMI across all the datasets for each feature selection algorithm is also demonstrated as a bar plot in Figure 4. On this account, Figures 3 and 4 confirm the superiority of the proposed feature selection model in almost all cases.

I.1 CONVERGENCE TEST

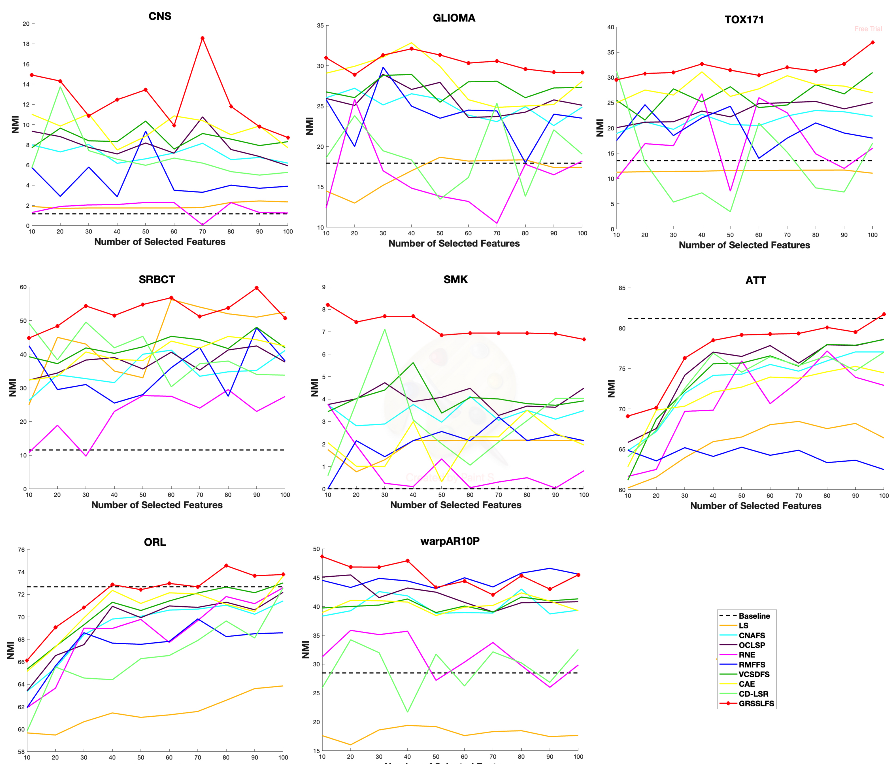
We set out to empirically evaluate Theorem 3.2, wherein we presented a theoretical proof for the monotonic decreasing behavior of the objective function of the GRSSLFS method. Figure 5 depicts the evolution of the objective function over the number of iterations for all datasets. One can clearly see that the values of the objective function across all datasets plummet quickly and the objective function converges quickly. This observation demonstrates the effectiveness of the alternative iterative algorithm that is proposed to solve the objective function of the GRSSLFS method.

I.2 RUNNING TIME COMPARISON

Figure 6 also demonstrates a comparison between the proposed method and all other unsupervised feature selection algorithms in terms of the running times (in seconds) for selecting 100 features from



(a) ACC



(b) NMI

Figure 3: Clustering ACC and NMI of nine algorithms on eight datasets as a function of the number of selected features.

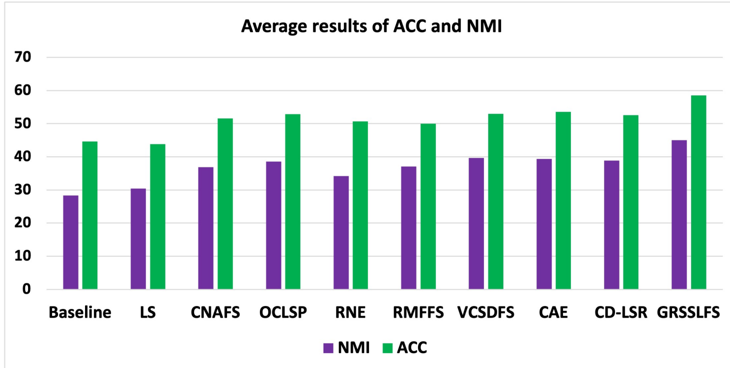


Figure 4: Average of the best ACC and NMI results over all datasets for each feature selection method.

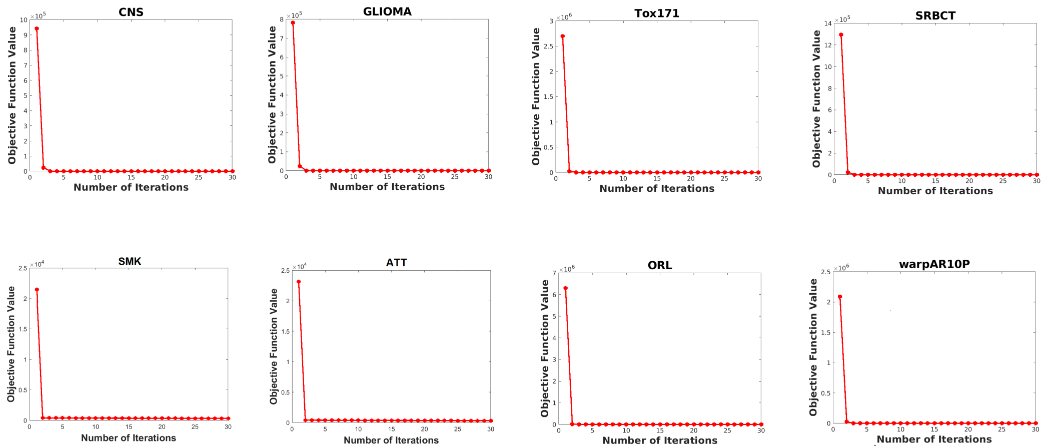


Figure 5: The convergence curves of the objective function of GRSSLFS on the eight datasets.

the TOX171 dataset. It is evident that CAE is the slowest method, followed by RNE and GRSSLFS methods, as the second and third slowest algorithms. The high running time of CAE is ascribed to the training neural networks in this deep, encoder-decoder feature selector Balın et al. (2019). The relative sluggishness of the RNE method can be attributed to its non-smooth nature, although it is convex Liu et al. (2020). To solve this issue, RNE employs the alternation direction method of multipliers (ADMM) to minimize its loss function. The proposed method is the next slow method due to the fact that the GRSSLFS method uses both subspace learning and self-representation, wherein multiple matrix multiplications are carried out. Consequently, as we demonstrated in Section 3.7, the time complexity of the GRSSLFS method is quadratic in terms of both the number of original features and the number of input data points. Furthermore, as we discussed in Section 3.5, an alternative approach is used to solve the loss function minimization problem of the proposed method.

I.3 Non-parametric Statistical Test

We demonstrated in previous sub-sections that the proposed method, GRSSLFS, has superior performance over the state-of-the-art unsupervised feature selection methods. Here, we employ the Friedman test as a non-parametric statistical test to evaluate the statistical significance of GRSSLFS's performance compared to other methods. The null hypothesis is then expressed as follows: there is no significant difference between any of the methods, and the selected features by these methods lead to the same clustering performance. The alternative hypothesis is that there are at least two methods whose selected features result in significant differences in the clustering performance.

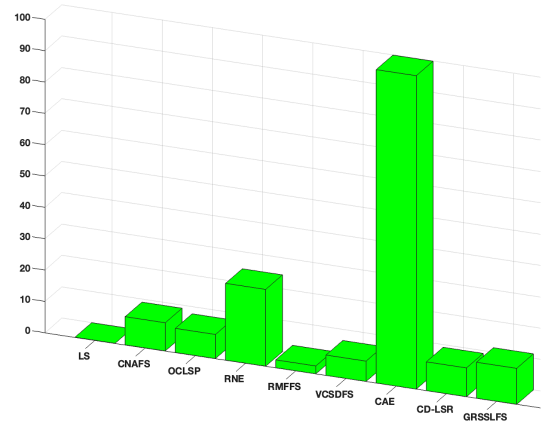


Figure 6: Running time (in seconds) for selecting 100 features from the TOX171 dataset.

We use the data presented in Table 1 to run the test. In this setting, the datasets, the feature selection methods, and the ACC or NMI values are considered as, respectively, the subjects, the treatments, and the measurements in the Friedman test terminology. The feature selection methods are then ranked to calculate the Friedman statistics. The average rankings over all datasets for each feature selection algorithm and two different measurements (i.e., ACC and NMI) are presented in Figure 7. The lower ranking means the corresponding measurement is higher.

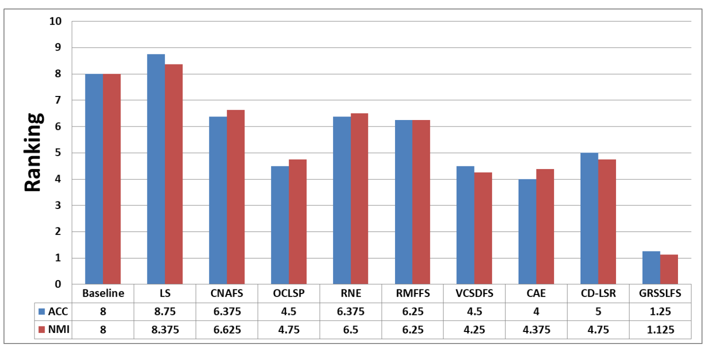


Figure 7: Average ranks obtained by the Friedman test for each method with respect to the ACC and NMI evaluation metrics on the datasets. Note that the lower rank of the evaluation metrics, the better the performance of the methods.

We obtained 36.190909 and 35.290909 for the Friedman statistics based on ACC and NMI measurements, respectively. Accordingly, the p-values corresponding to these Friedman statistics are 0.000053 and 0.000037, respectively. With the significance level of 5%, these two p-values lead us to reject the null hypothesis in the favor of the alternative hypothesis. With this description, it can be seen that there are at least two methods, VCSDFS, CAE and GRSSLFS, that select features with statistically significant differences in the clustering performance.

Following the Friedman test, we also carry out a post-hoc statistical test using the Holm's method (aka., Holm–Bonferroni method) Bolón-Canedo et al. (2014) to identify which pairs of methods result in a significant difference in the clustering performance. For this purpose, we select the proposed method, GRSSLFS, as the control method and compare it with all other 9 methods including the Baseline which are listed in the columns of Table 1. It means that we end up having 9 null hypotheses for each pair of GRSSLFS-X where X refers to those 9 other methods. Each null hypothesis states that there is no significant difference between GRSSLFS and the method "X" in terms of their feature selection effectiveness with performant downstream clustering. The significance level α is set to 0.05 for each null individual null hypothesis. The p-values of the nine null hypotheses are presented in

Table 4 based on ACC and NMI measurements, respectively. The corresponding Holm's p-values are also included in these tables.

ACC		
Method	p-value	Holm's p-value
LS	0.000001	0.005556
Baseline	0.000008	0.00625
CNAFS	0.000711	0.007143
RNE	0.000711	0.008333
RMFFS	0.000957	0.01
CD-LSR	0.013243	0.0125
OCLSP	0.031803	0.016667
VCSDFS	0.031803	0.025
CAE	0.06928	0.05

NMI		
Method	p-value	Holm's p-value
LS	0.000002	0.005556
Baseline	0.000006	0.00625
CNAFS	0.00028	0.007143
RNE	0.000384	0.008333
RMFFS	0.000711	0.01
OCLSP	0.016639	0.0125
CD-LSR	0.016639	0.016667
CAE	0.031803	0.025
VCSDFS	0.038989	0.055

Table 4: Post-hoc test using Holm's method to compare the effect of GRSSLFS and eight other feature selection methods and also the Baseline on the clustering ACC and NMI metrics. The significance level of $\alpha = 0.05$ is set for each individual pairwise null hypothesis. Holm's procedure rejects a null hypothesis when the p-value of the individual null hypothesis is less than or equal to Holm's p-value.

Holm's method rejects those hypotheses that their initial p-values are less than or equal to the calculated p-values by Holm's method. Thus, we infer from the results in Table 4 that we have enough evidence to reject the null hypothesis for the pairwise comparison between GRSSLFS and all methods except for OCLSP, CAE, VCSDFS, and CD-LSR. To be more precise, our proposed GRSSLFS method exhibits significant differences from all other methods, except for the aforementioned methods, in terms of the ACC and NMI measures. Nevertheless, as we observed in the results of Figure 3, GRSSLFS displayed a more consistent trend over a wide range of a number of selected features compared to all comparison methods.

I.4 Ablation Study

The effects of hyperparameters on the performance of our proposed method are investigated. The objective function of GRSSLFS in (9) has three adjustable parameters; α is for the feature graph regularization term in the self-representation framework and β and γ control the sparsity of the feature weight matrix $\mathbf{U}$ and representation matrix $\mathbf{V}$, respectively, in the subspace learning process. To isolate the impact of the hyperparameters, we set their values to zero according to the configuration shown in the first column of Table 5. We then choose one biological (GLIOMA) and one image (warpAR10P) dataset to run through our algorithm for feature selection and subsequent clustering.

The clustering metrics for the seven experiments are reported in Table 5 and are compared against the best results we obtained with non-zero hyperparameter values of our proposed method. It is evident that the inclusion of all three hyperparameters in our method results in higher clustering performance compared to cases with one or more hyperparameters excluded. Thus, it can be concluded that all parts of the objective function in (9) have a critical role in the feature selection process and can not be removed from this process.

Table 5: The outputs related to the ablation study.

Datasets	**GLIOMA**		**warpAR10P**	
Evaluation metrics	ACC± std	NMI± std	ACC± std	NMI± std
Main experiment	**54.10 ± 2.71(30)**	**32.09±2.14(40)**	**46.66 ± 3.51(10)**	**48.63 ± 2.83(10)**
$\alpha = \beta = \gamma = 0$	44.51±2.96(50)	21.38±4.67(50)	32.27±2.75(50)	34.01±3.18(50)
$\alpha = \beta = 0$	50.21±4.29(20)	27.09±3.09(70)	39.65±3.18(10)	42.59±4.01(30)
$\alpha = \gamma = 0$	48.30±3.13(10)	27.52±4.03(60)	38.15±3.94(30)	41.55±3.78(80)
$\beta = \gamma = 0$	46.60±2.16(10)	25.46±5.29(50)	40.84±3.06(10)	45.01±3.56(30)
$\alpha = 0$	50.00±3.94(40)	30.37±1.25(10)	43.34±4.13(20)	45.68±3.36(10)
$\beta = 0$	49.80±1.82(30)	29.89±2.28(30)	42.19±4.97(20)	44.04±3.58(20)
$\gamma = 0$	51.20±2.46(20)	28.97±4.18(70)	40.26±3.51(30)	43.99±2.19(10)

I.5 Additional Comparison method Results

In this subsection, the evaluation of our method is compared with the three well-known subspace leraning based feature selection methods:

1. **Matrix Factorization Feature Selection (MFFS)** Wang et al. (2015a): It is a matrix factorization-based method that is constructed from subspace learning with the orthogonal constraint.
2. **Maximum Projection and Minimum Redundancy (MPMR)** Wang et al. (2015b): This method selects features by projecting the original features into a feature subspace through a linear transformation with minimum reconstruction error and low redundancy in selected features.
3. **Subspace Clustering Unsupervised Feature Selection (SCFS)** Parsa et al. (2020): SCFS method employs the self-representation method and learns the clustering similarities by selecting discriminative features.

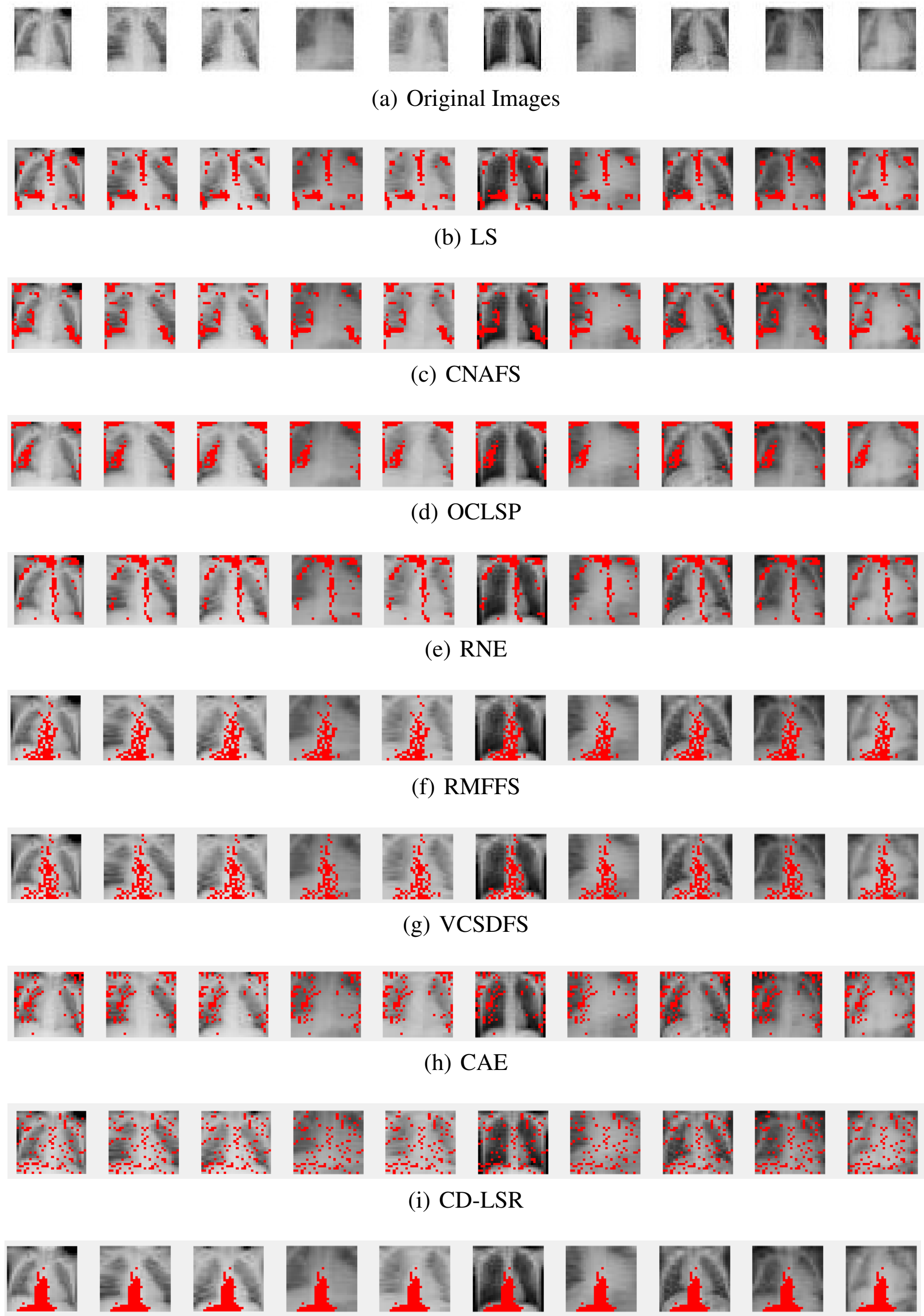
Table 6: Comparing the best values of the clustering ACC and NMI of nine algorithms on all datasets. The values in parentheses denote the selected number of features at which the ACC and NMI are reported.

ACC										
Datasets	**Baseline**	**LS**	**MFFS**	**MPMR**	**RNE**	**RMFFS**	**SCFS**	**CAE**	**CD-LSR**	**GRSSLFS**
CNS	53.33 ± 1.45	58.66 ± 1.02 (80)	61.75 ± 2.78 (50)	61.66 ± 0.00 (70)	58.33 ± 0.00 (40)	63.33 ± 0.00 (60)	63.54 ± 0.75 (10)	66.61±0.05(60)	61.67±1.83(60)	**73.33 ± 0.01 (20)**
GLIOMA	44.05 ± 0.00	44.00 ± 0.00 (10)	48.51 ± 2.66 (80)	46.40 ± 0.97 (30)	47.11 ± 2.19 (20)	49.80 ± 5.30 (30)	50.00 ± 0.00 (70)	**54.52±3.13(40)**	48.70±3.92(10)	54.10 ± 2.71(30)
TOX171	41.25 ± 0.72	40.35 ± 0.72 (10)	42.13 ± 0.13 (90)	41.49 ± 1.65 (60)	51.69 ± 0.48 (40)	49.64 ± 1.59 (20)	52.81 ± 0.07 (90)	53.36±1.24(40)	48.77±4.18(10)	**57.39 ± 1.27 (40)**
SRBCT	25.66 ± 2.82	46.02 ± 2.96 (100)	43.97 ± 2.15 (60)	37.77 ± 5.14 (90)	42.12 ± 3.37 (80)	48.97 ± 4.67 (90)	39.86 ± 1.76 (10)	46.08±4.18(90)	52.63±4.39(10)	**53.49 ± 3.11 (60)**
SMK	51.34 ± 0.56	55.11± 2.35 (80)	60.43 ± 4.88 (60)	60.36 ± 2.11 (100)	61.51 ± 0.34 (10)	52.94± 2.31 (50)	61.15 ± 1.33 (40)	61.80±2.06(40)	64.71±2.72(30)	**65.72 ± 2.59 (10)**
ATT	63.84 ± 4.03	46.75 ±3.87(60)	59.61 ±4.18(90)	58.62 ±2.83(90)	59.08 ± 3.12(50)	46.25±3.48(30)	60.01 ± 2.21 (90)	56.86±2.79(90)	58.03±3.33(100)	**64.11 6± 2.65(100)**
ORL	50.25 ± 2.91	38.51 ± 1.66(90)	49.78±3.14(80)	48.75 ±2.65(100)	51.31 ±3.01(100)	47.71±2.73(70)	40.59±2.11(100)	50.09±2.42(100)	**54.05±3.51(100)**	53.45 ± 3.13(80)
warpAR10P	26.88 ± 2.18	21.50 ±0.94(10)	41.92 ±3.46(30)	38.15 ±2.71(10)	34.62 ±2.45(10)	41.55±3.21(90)	27.31 ±1.48(90)	39.26±2.83(10)	31.94±2.78(30)	**46.66 ± 3.51(10)**

NMI										
Datasets	**Baseline**	**LS**	**MFFS**	**MPMR**	**RNE**	**RMFFS**	**SCFS**	**CAE**	**CD-LSR**	**GRSSLFS**
CNS	1.17 ± 0.00	2.44 ± 0.45 (90)	2.93 ± 0.45 (100)	1.62 ± 0.80 (30)	2.29 ± 0.00 (50)	9.35 ± 0.08 (50)	3.27 ± 0.00 (10)	11.05±0.41(30)	13.76±0.06(20)	**18.56 ± 0.00 (70)**
GLIOMA	17.94± 0.69	18.67± 0.29 (50)	23.69 ± 2.07 (80)	25.24 ± 2.08 (30)	25.84 ± 4.20 (20)	29.81± 1.46 (30)	28.27 ± 1.70 (70)	**32.83±1.83(40)**	25.35±1.68(70)	32.09 ± 2.14 (40)
TOX171	13.54± 0.23	11.68± 1.97 (90)	12.82 ± 1.52 (80)	16.22 ± 1.89 (60)	26.78 ± 0.08 (40)	24.60± 0.41 (20)	33.76 ± 0.54 (70)	31.12±0.31(40)	31.28±0.93(10)	**36.96 ± 0.60 (100)**
SRBCT	11.55 ± 3.47	56.23± 4.06 (60)	38.41 ± 7.38 (50)	24.27 ± 5.26 (40)	29.48 ± 3.52 (80)	47.82± 4.13 (90)	37.53 ± 2.81 (70)	45.28±3.59(80)	49.53±3.86(30)	**59.67 ± 4.18 (90)**
SMK	0.01 ± 0.00	2.17± 0.05 (80)	3.23 ± 0.27 (60)	3.73 ± 0.62 (60)	3.77 ± 0.08 (10)	3.21± 0.19 (70)	3.52 ± 0.22 (40)	3.51±0.09(80)	7.12±1.05(30)	**8.20 ± 0.53 (10)**
ATT	81.19±1.77	68.46 ± 1.37(70)	78.73± 2.11(90)	77.75 ±2.15(90)	77.17 ± 1.04(80)	65.27±2.03(50)	78.56 ±1.21(90)	75.27±2.24(90)	77.02±2.19(100)	**81.71± 1.31(100)**
ORL	72.68 ±1.27	63.85 ±1.23(100)	71.67 ±1.77(80)	70.64 ±1.25(100)	72.59±1.33(100)	69.72 ±1.69(70)	65.61 ±1.99(100)	73.64±1.41(100)	72.57±2.02(100)	**74.56 ± 1.72(80)**
warpAR10P	28.48 ± 3.31	19.39 ±1.84(40)	44.73 ±3.38(30)	48.12 ±2.58(20)	35.85±4.38(20)	46.60±2.94(90)	22.03 ±0.87(90)	42.35±2.23(80)	34.22±3.37(20)	**48.63 ± 2.83(10)**

J APPLICATION TO THE PNEUMONIAMNIST DATASET

This section presents the results of the selection of 100 features from the PneumoniaMNIST dataset, conducted by our proposed GRSSLFS method and other comparative methods.



(a) Original Images

(b) LS

(c) CNAFS

(d) OCLSP

(e) RNE

(f) RMFFS

(g) VCSDFS

(h) CAE

(i) CD-LSR

(j) Our proposed GRSSLFS method

Figure 8: The visualization of 100 selected features obtained by different feature selection methods on PneumoniaMNIST images.

K CODE AVAILABILITY

An implementation of Graph Regularized Self-Representation and Sparse Subspace Learning (GRSSLFS) is available in the following:

```
%------------ Inputs
%% X: Data matrix in R^(m*n), withe m samples and n features
%% B: Basis for the feature space
%% A: Similarity matrix in R^(n*n) associated with the features
%% P: Degree matrix
%% alpha,beta,gamma: Regularization parameters
%% k: The number of selected features
%% itermax: Maximum number of iterations
%% BB = (B')*B;
%% BX = (B')*X;
%------------ Output
%% Selected_feat: Selected features

%------------ The proposed GRSSLFS method:
function Selected_feat = GRSSLFS(X,BB,BX,P,A,alpha,beta,gamma,k,
    itermax)
%------------ Initializations
[m,n] = size(X);
G = rand(m,n);
U = rand(n,k);
V = rand(k,m);
onesm = ones(m,m);
%------------ Calculation of the matrix E used in the L2,1 norm
a = 4*max(diag(U*(U')),10^-9);
E  =  diag(sqrt(1./a));

for i  =  1:itermax
%------------ Update G
VU = (V')*(U');
BBG = BB*G;
NoG = BX+alpha*BBG*A+BB*VU;
DeG = BBG+alpha*BBG*P+BBG*U*V*VU;
re1 = rdivide(NoG,DeG);
G = G.*nthroot(re1,2);
%------------ Update U
GBB = (G')*BB;
NoU = GBB*(V');
DeU = GBB*G*U*V*(V')+beta*E*U;
re2 = rdivide(NoU,DeU);
U = U.*nthroot(re2,2);
%------------ Update E
a = 4*max(diag(U*(U')),10^-9);
E  =  diag(sqrt(1./a));
%------------ Update V
UGB = BB*G*U;
NoV = (UGB')+gamma*V;
DeV = (UGB')*G*U*V+gamma*V*onesm;
re3 = rdivide(NoV,DeV);
V = V.*nthroot(re3,2);
end
tempVector  =  sum(U.^2, 2);
[~, value]  =  sort(tempVector,'descend');
Selected_feat  =  value(1:k);
end
```

```

%------------ Constructing the Basis for Feature Space via the Variance information and the Basis Extension (VBE) method
function B = VBE(X,r)
%------------ Inputs
%% X: Data matrix
%% r: rank of X
%------------ Output
%%  B: Basis matrix
E = var(X);
[~,index] = sort(E,'descend');
A = index;
B(:,1) = X(:,A(1));
j = 2;
for i = 2:r
B(:,i) = X(:,A(j));
while rank(B)~ = i
j = j+1;
B(:,i) = X(:,A(j));
end
j = j+1;
end
end
```