Self-Representation Unsupervised Feature Selection based on Non-negative Matrix Factorization

Hossein Nasser Assadi[*] Faranges Kyanfar[†] Farid Saberi-Movahed[‡] Abbas Salemi[§]

Abstract

Non-negative matrix factorization (NMF) and self-representation are widely employed for dimensionality reduction and extracting intrinsic structures from high-dimensional data. However, integrating these techniques for effective unsupervised feature selection remains a complex challenge. In this article, we propose a novel method, Self-Representation Feature Selection based on Non-Negative Matrix Factorization (SRFSNMF), which bridges this gap. SRFSNMF uses the Gram matrix, built from the basis matrix of samples obtained through NMF, combined with a self-representation technique. This aims to capture the structural relationships between samples and uncover intrinsic data relationships, enhancing the selection of relevant features in complex datasets. To solve the SRFSNMF model, we develop an efficient iterative optimization algorithm with guaranteed convergence. Experimental results on multiple benchmark datasets demonstrate that SRFSNMF outperforms state-of-the-art methods, achieving superior effectiveness in unsupervised feature selection tasks. Additionally, we conduct a sensitivity analysis of the model's parameters and assess its robustness against noise, further validating the reliability and stability of our approach.

Keywords: Self-Representation, Non-negative Matrix Factorization, and Feature Selection

1 Introduction

In modern data science, the rapid increase in dimensionality poses significant challenges for traditional clustering and classification techniques [1]. Identifying meaningful patterns in high-dimensional data, especially in unsupervised settings, requires methods that can efficiently reduce dimensionality while preserving the underlying structure of the data [2]. Two such prominent approaches are self-representation and Non-negative Matrix Factorization (NMF) [3].

Self-representation leverages the inherent relationships within a dataset by expressing each feature as a linear combination of other features, capturing the datas internal structure and dependencies [4]. This approach has proven highly effective in unsupervised feature selection, particularly in cases where labeled data is unavailable [5]. By analyzing intrinsic characteristics, self-representation helps mitigate the curse of dimensionality, facilitating more effective clustering and classification tasks [6].

On the other hand, NMF has emerged as a powerful tool for dimensionality reduction and data segmentation [7]. By decomposing a non-negative data matrix into two lower-dimensional, non-negative matrices, NMF generates a parts-based representation of the data, making it highly effective for clustering [8]. However, one of the main limitations of NMF is its reliance on linear separability, which can lead to poor performance in datasets characterized by non-linear relationships.

In this work, we propose a novel approach that leverages the structural insights gained from NMF and combines them with a self-representation technique to enhance feature selection. Specifically, our method focuses on uncovering intrinsic relationships within high-dimensional data by modeling the interactions between data samples. By deriving the inner product of the learned sample representations, we can capture pairwise similarities, allowing us to understand the underlying structure of the data more effectively. This not only helps reduce dimensionality but also enhances the selection of the most informative features, particularly in complex datasets.

However, integrating NMF with self-representation for effective unsupervised feature selection remains a complex challenge. In this article, we propose a novel method, Self-Representation Feature Selection based on Non-Negative Matrix Factorization (SRFSNMF). Our approach employs a Gram matrix constructed from the basis matrix of samples obtained through NMF, combined with a self-representation technique. This methodology aims to capture the structural relationships between samples and uncover intrinsic data relationships, thereby enhancing the selection of relevant features in complex datasets. By emphasizing the inter-

[*] Department of Applied Mathematics, Shahid Bahonar University of Kerman, Kerman, Iran, hnaserasadi@math.uk.ac.ir

[†] Department of Applied Mathematics, Shahid Bahonar University of Kerman, Kerman, Iran, kyanfar@uk.ac.ir

[‡] Department of Applied Mathematics, Faculty of Sciences and Modern Technologies, Graduate University of Advanced Technology, Kerman, Iran, f.saberimovahed@kgut.ac.ir

[§] Department of Applied Mathematics, Shahid Bahonar University of Kerman, Kerman, Iran, Salemi@uk.ac.ir

actions among samples, SRFSNMF not only improves the accuracy of feature selection but also facilitates the identification of clusters and patterns within the data. Through empirical evaluations on benchmark datasets, we demonstrate that our proposed method significantly outperforms existing feature selection and clustering techniques.

In the following sections, we will provide a comprehensive overview of our research. Section 2 will cover the preliminaries, offering insights into NMF and self-representation techniques, as well as introducing our proposed method with a detailed formulation. Section 3 will explain the optimization processes and algorithms employed in our model. Section 4 will present the numerical results, demonstrating the performance and effectiveness of our method across various datasets and experiments to analyze the model's efficacy. Finally, Section 5 will conclude the article, summarizing our findings and highlighting the contributions and implications of our work in the context of unsupervised feature selection.

2 Proposed Method

In recent years, self-representation and matrix factorization have become key techniques for unsupervised feature selection, particularly when dealing with high-dimensional data. The challenge of high-dimensional data is to efficiently select the most informative features while maintaining the inherent structure of the data. Traditional dimensionality reduction techniques like NMF, offer robust methods for clustering and data representation [9]. However, these methods often face limitations in capturing complex data relationships, which can hinder their performance in feature selection tasks.

To address these issues, we propose an advanced self-representation approach integrated with the NMF framework. Our method combines the benefits of self-representation capturing the intrinsic relationships between features and samples with the power of NMF for clustering and feature selection. This novel approach aims to improve performance in high dimensional settings by simultaneously leveraging the structural relationships between features and samples.

In the following sections, we first outline the necessary notations and preliminaries, including NMF and self-representation. Then, we describe the details of our proposed method, leading to the final objective function that drives the optimization.

2.1 Preliminaries

2.1.1 Notation

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ represent the data matrix, where m is the number of samples and n is the number of features. The rows of $\mathbf{X}$ are denoted as $[\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_m]$, representing the samples, while the columns are denoted as $[\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n]$, representing the features. For any matrix $\mathbf{X}$, the i-th row is represented as $\mathbf{x}_i$, the j-th column as $\mathbf{f}_j$, and the (i, j)-th element as $\mathbf{x}_{ij}$. We denote $\mathbb{R}_+^{m \times n}$ as the set of non-negative matrices in $\mathbb{R}^{m \times n}$. The identity matrix in $\mathbb{R}^{m \times m}$ is represented as $\mathbf{I}$. The $\mathbf{L}_{2,1}$ norm of $\mathbf{X}$, denoted $\|\mathbf{X}\|_{2,1}$, is defined as the sum of the Euclidean norms of its rows, while the Frobenius norm $\|\mathbf{X}\|_F$ is the square root of the sum of the squared elements of $\mathbf{X}$.

2.1.2 Non-Negative Matrix Factorization

NMF is a popular method for dimensionality reduction, where the goal is to decompose a non-negative data matrix into two lower-dimensional non-negative matrices that capture the underlying structure of the data [10]. Given a data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, where m is the number of samples and n is the number of features, NMF seeks to find two matrices $\mathbf{V} \in \mathbb{R}^{m \times r}$ and $\mathbf{H} \in \mathbb{R}^{n \times r}$ such that:

$$\mathbf{X} \approx \mathbf{V}\mathbf{H}^T.$$

Here, r is the reduced rank, typically much smaller than n, which captures the intrinsic structure of the data. The objective of NMF is to minimize the reconstruction error between $\mathbf{X}$ and its approximation $\mathbf{V}\mathbf{H}^T$, which is formulated as:

$$\min_{\mathbf{V},\mathbf{H} \geq 0} \|\mathbf{X} - \mathbf{V}\mathbf{H}^T\|_F^2.$$

In this decomposition, $\mathbf{V}$ contains the basis vectors for the columns (features) of $\mathbf{X}$, while $\mathbf{H}$ represents the basis for the rows (samples). Each column of $\mathbf{H}$ corresponds to a latent feature in the reduced space, and the matrix $\mathbf{H}\mathbf{H}^T$ plays an important role in capturing the similarity between the samples. Specifically, the Gram matrix $\mathbf{H}\mathbf{H}^T$ provides pairwise inner products between the rows of $\mathbf{H}$, which reveals the relationships between different samples in the low-dimensional representation.

This similarity structure encoded in $\mathbf{H}\mathbf{H}^T$ helps to identify clusters or groups of similar samples, making it a powerful tool for understanding the underlying patterns in the data. Moreover, the matrix $\mathbf{H}\mathbf{H}^T$ also facilitates noise reduction and improves the accuracy of the approximation by focusing on the most significant dependencies between samples.

2.1.3 Self Representation

In this subsection, we explain the concept of self-representation in the context of feature selection. We assume that features in the dataset are not entirely independent, and each feature can be represented as a linear combination of the other features. Specifically, for a feature vector $\mathbf{f}_i$ in the data matrix $\mathbf{X}$, it can be expressed as:

$$\mathbf{f}_i = \sum_{j=1}^{n} z_{ji}\mathbf{f}_j.$$

Here, z_{ji} is an element of the self-representation matrix $\mathbf{Z}$, where each element represents the weight coefficient between the i-th feature $\mathbf{f}_i$ and the j-th feature $\mathbf{f}_j$. This formulation leverages the relationships between features to highlight dependencies and redundancies.

For a clearer understanding, the self-representation model for all features can be illustrated as follows:

$$\begin{aligned} \mathbf{f}_1 &\approx z_{11}\mathbf{f}_1 + z_{21}\mathbf{f}_2 + \cdots + z_{n1}\mathbf{f}_n \\ \mathbf{f}_2 &\approx z_{21}\mathbf{f}_1 + z_{22}\mathbf{f}_2 + \cdots + z_{n2}\mathbf{f}_n \\ &\vdots \\ \mathbf{f}_n &\approx z_{n1}\mathbf{f}_1 + z_{n2}\mathbf{f}_2 + \cdots + z_{nn}\mathbf{f}_n \end{aligned} \quad . \tag{1}$$

This formulation given in (1) shows that each feature is linearly reconstructed by the other features, reflecting the importance of each feature in representing the dataset. The i-th row of $\mathbf{Z}$, denoted as $\mathbf{z}_i$, influences the reconstruction of the corresponding feature $\mathbf{f}_i$. A higher norm of $\mathbf{z}_i$ indicates that feature $\mathbf{f}_i$ plays a more significant role in representing the dataset, thus signifying its importance. To formalize this self-representation model, we define the following objective function, which aims to minimize the reconstruction error of the feature representation:

$$\min_{\mathbf{Z}\geq 0} \|\mathbf{X} - \mathbf{X}\mathbf{Z}\|_F^2$$

This objective encourages a compact representation of features by using the self-representation matrix $\mathbf{Z}$ to reveal the underlying relationships between the features.

2.2 Objective Function

We propose a novel method that integrates NMF with self-representation to enhance unsupervised feature selection. NMF captures the structural patterns of high-dimensional data, while self-representation identifies relevant features by revealing dependencies between features.

In our approach, the self-representation matrix $\mathbf{Z} \in \mathbb{R}^{n\times n}$ encodes the relationships between features, where larger values in $\mathbf{Z}$ signify stronger dependencies. To ensure a low-rank structure in this feature relationship, we constrain $\mathbf{Z}$ as $\mathbf{Z} = \mathbf{H}\mathbf{H}^T$, where $\mathbf{H} \in \mathbb{R}^{n\times r}$ is a nonnegative matrix and $r \ll n$. This constraint ensures a symmetric and low-rank $\mathbf{Z}$, preserving the essential dependencies between features.

Matrix $\mathbf{H}$ serves as a critical component that links NMF and self-representation. In NMF, $\mathbf{H}$ represents the latent space for samples, and the product $\mathbf{H}\mathbf{H}^T$ forms a Gram matrix that encodes similarities between samples in a low-dimensional space. The symmetry of $\mathbf{H}\mathbf{H}^T$ guarantees mutual dependencies between features, which is essential for capturing both sample and feature relationships consistently. The objective function of our method is formulated as follows:

$$\min_{\mathbf{V},\mathbf{H}\geq 0} \|\mathbf{X} - \mathbf{V}\mathbf{H}^T\|_F^2 + \|\mathbf{X} - \mathbf{X}\mathbf{H}\mathbf{H}^T\|_F^2, \tag{2}$$

where $\mathbf{V} \in \mathbb{R}^{m\times r}$ represents the basis matrix for features, and $\mathbf{H}$ serves as the matrix that captures the latent structure of both samples and features. The first term minimizes the reconstruction error in the sample space using NMF, while the second term enforces the self-representation constraint, ensuring that feature dependencies are captured in a low-rank form through $\mathbf{H}\mathbf{H}^T$. In (2), it can be seen that both the feature space (via $\mathbf{V}\mathbf{H}^T$) and the sample space (via $\mathbf{H}\mathbf{H}^T$) are simultaneously modeled, leading to a more compact and interpretable feature selection process. The matrix $\mathbf{H}$ plays a dual role: reducing dimensionality and uncovering the most representative features while minimizing redundancy.

2.2.1 Sparsity

Incorporating sparsity into the self-representation framework enhances the ability to select the most relevant features. To achieve this, we impose an $L_{2,1}$-norm regularization on the matrix $\mathbf{H}$, which enforces row-wise sparsity in the factorized representation $\mathbf{Z} = \mathbf{H}\mathbf{H}^T$. The $L_{2,1}$-norm encourages only a few rows of $\mathbf{H}$ to have nonzero values, which effectively selects a subset of features for better interpretability and improved feature selection.

Let $\mathbf{H} \in \mathbb{R}^{n\times r}$, where each row $\mathbf{h}_i$ corresponds to a feature, and imposing sparsity ensures that only a small number of these features significantly contribute to the self-representation. The $L_{2,1}$-norm of $\mathbf{H}$ is defined as:

$$\|\mathbf{H}\|_{2,1} = \sum_{i=1}^{m} \|\mathbf{h}_i\|_2,$$

where $\|\mathbf{h}_i\|_2$ is the L_2-norm of the i-th row of $\mathbf{H}$. This term encourages sparsity by penalizing the sum of the norms of the rows of $\mathbf{H}$, ensuring that many of the rows become zero, thereby reducing the number of contributing features. These norms can also be expressed in terms of matrix trace functions as:

$$\|\mathbf{H}\|_{2,1} = \mathrm{trace}(\mathbf{H}^T\mathbf{G}\mathbf{H}),$$

where $\mathbf{G} = [\mathbf{G}_{ij}] \in \mathbb{R}^{m\times m}$ is a diagonal matrix with diagonal entries defined as:

$$\mathbf{G}_{ii} = \frac{1}{\max(2\|\mathbf{h}_i\|_2, \varepsilon)}, \quad i = 1, 2, \ldots, m, \tag{3}$$

where ε is a small constant to avoid division by zero.

The final objective function of our proposed method incorporates both the NMF-based sample representation and the sparse symmetric self-representation for features. It is formulated as follows:

$$\min_{\mathbf{V},\mathbf{H}\geq 0} \|\mathbf{X} - \mathbf{V}\mathbf{H}^T\|_F^2 + \|\mathbf{X} - \mathbf{X}\mathbf{H}\mathbf{H}^T\|_F^2 + \alpha\|\mathbf{H}\|_{2,1}. \tag{4}$$

By incorporating the sparsity constraint, our method not only selects key features but also ensures that the self-representation matrix $\mathbf{Z} = \mathbf{H}\mathbf{H}^T$ is low-rank, symmetric, and interpretable. The inclusion of the $L_{2,1}$-norm allows the model to focus on the most informative features, improving the performance of unsupervised tasks such as clustering and feature selection.

3 Optimization

To solve the above objective function, we propose an iterative optimization approach. The problem is non-convex due to the interaction between the self-representation and sparsity terms, so we apply a gradient-based algorithm to minimize the objective.

At each iteration t, the self-representation matrix $\mathbf{Z}$ is updated by solving:

$$\min_{\mathbf{V},\mathbf{H}\geq 0} \|\mathbf{X}-\mathbf{V}\mathbf{H}^T\|_F^2+\|\mathbf{X}-\mathbf{X}\mathbf{H}\mathbf{H}^T\|_F^2+\alpha\text{trace}(\mathbf{H}^T\mathbf{G}\mathbf{H}).$$

The optimization continues until the matrix $\mathbf{H}$ converges, meaning the change between iterations becomes smaller than a predefined threshold. The final solution provides the sparse self-representation matrix $\mathbf{H}$, from which the most informative and non-redundant features can be selected. We define the function $\mathbf{L}$ as:

$$\begin{aligned}\mathbf{L} = {} & \|\mathbf{X} - \mathbf{V}\mathbf{H}^T\|_F^2 + \|\mathbf{X} - \mathbf{X}\mathbf{H}\mathbf{H}^\mathbf{T}\|_F^2 + \alpha\text{trace}(\mathbf{H}^T\mathbf{G}\mathbf{H}) \\ & + \text{trace}(\boldsymbol{\Delta}^T\mathbf{H}),\end{aligned}$$

which can be expanded as:

$$\begin{aligned}\mathbf{L} = {} & \text{trace}(\mathbf{X}^T\mathbf{X}) - 2\text{trace}(\mathbf{H}^T\mathbf{X}^T\mathbf{V}) + \text{trace}(\mathbf{H}^T\mathbf{H}\mathbf{V}^T\mathbf{V}) \\ & + \text{trace}(\mathbf{X}^T\mathbf{X}) - 2\text{trace}(\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H}) \\ & + \text{trace}(\mathbf{H}^T\mathbf{H}\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H}) + \alpha\,\text{trace}(\mathbf{H}^T\mathbf{G}\mathbf{H}) \\ & + \text{trace}(\boldsymbol{\Delta}^T\mathbf{H}),\end{aligned}$$

where $\boldsymbol{\Delta}$ is a Lagrange multiplier matrix enforcing the non-negativity constraint on $\mathbf{H}$. To find the optimal $\mathbf{H}$, we take the gradient of $\mathbf{L}$ and set it to zero. The gradient is given by:

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{H}} = {} & -2\mathbf{X}^T\mathbf{V} + 2\mathbf{H}\mathbf{V}^T\mathbf{V} - 4\mathbf{X}^T\mathbf{X}\mathbf{H} + \mathbf{H}\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H} \\ & + \mathbf{X}^T\mathbf{X}\mathbf{H}\mathbf{H}^T\mathbf{H} + 2\alpha\mathbf{G}\mathbf{H} + \boldsymbol{\Delta}.\end{aligned}$$

By considering the Kuhn-Tucker conditions $\boldsymbol{\Delta}_{ij}\mathbf{H}_{ij} = 0$, and assuming that the gradient is zero, we update $\mathbf{H}_{ij}$ as follows:

$$\mathbf{H}_{ij} \leftarrow \mathbf{H}_{ij}\sqrt{\frac{\mathbf{P}_{ij}}{\mathbf{Q}_{ij}}}, \tag{5}$$

where

$$\begin{aligned}&\mathbf{P} = 2\mathbf{X}^T\mathbf{V} + 4\mathbf{X}^T\mathbf{X}\mathbf{H}, \\ &\mathbf{Q} = 2\mathbf{H}\mathbf{V}^T\mathbf{V} + \mathbf{H}\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H} + \mathbf{X}^T\mathbf{X}\mathbf{H}\mathbf{H}^T\mathbf{H} + 2\alpha\mathbf{G}\mathbf{H}.\end{aligned}$$

With similar way, the update rule of $\mathbf{V}$ obtain by:

$$\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij}\sqrt{\frac{(\mathbf{X}\mathbf{H})_{ij}}{(\mathbf{V}\mathbf{H}^T\mathbf{H})_{ij}}}. \tag{6}$$

3.1 Algorithm

In this subsection, we introduce the iterative updating algorithm designed for the proposed feature selection method. The algorithm begins with the initialization of key matrices, followed by iterative updates of the feature representation and selection matrices. Specifically, it alternates between updating the feature matrix $\mathbf{H}$, optimizing the data representation matrix $\mathbf{V}$, and adjusting the selection matrix $\mathbf{G}$, which is initialized as the identity matrix. The algorithm ensures convergence by iterating until a predefined maximum number of iterations is reached. In the final step, the algorithm identifies the most informative features by selecting those with the highest $\mathbf{L}_2$-norm values from the rows of matrix $\mathbf{H}$, outputting a set of k selected features for further analysis.

4 Numerical Results

In this section, we present the experimental evaluation of the proposed SRFSNMF method. We assess its performance across multiple datasets, comparing it to several baseline methods to ensure a comprehensive analysis. The experiments are designed to evaluate the clustering performance, feature selection capabilities, and robustness of SRFSNMF.

We first describe the datasets and baseline methods used for comparison. Next, we outline the experimental setup, including parameter settings and evaluation metrics. Finally, we provide a detailed analysis of the results, highlighting the advantages of SRFSNMF in terms of accuracy, feature selection, and resilience to noise.

In our experiments, the assessment of clustering performance involves the use of two widely-utilized evaluation metrics [12]: clustering accuracy (ACC) and normalized mutual information (NMI). Higher values for both ACC and NMI indicate improved performance.

Algorithm 1 Iterative updating algorithm for SRFS-NMF.

```
Require:
    $\mathbf{X} \in \mathbb{R}^{m\times n}$: The data matrix with $m$ data samples and $n$ features;
    $k$: the number of selected features;
    $\alpha$: the positive parameters;
    Iter_max: the maximum number of iterations.
Initialize the matrix $\mathbf{V} \in \mathbb{R}^{m\times r}$
Initialize the matrix $\mathbf{H} \in \mathbb{R}^{n\times r}$.
Set the matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ as the identity matrix.
while iteration $\leq$ Iter_max do
    Update $\mathbf{H}$ by (5).
    Fix $\mathbf{H}$ and update $\mathbf{V}$ by (6).
    Update the matrix $\mathbf{G}$ based on the rule (3).
end while
Find the $L_2$-norm for every row in the matrix $\mathbf{H}$, and arrange them in decreasing order. Then, choose $k$ features that correspond to the highest $k$ norms among the rows of $\mathbf{H}$.
Ensure: Choose a set of $k$ features as the output of the proposed feature selection method.
```

Suppose that $\mathbf{c}_i$ is the clustering label and $\mathbf{q}_i$ represent true label of data point $\mathbf{x}_i$. ACC is defined as follows:

$$\text{ACC} = \frac{\sum_{i=1}^{n} \delta(\mathbf{q}_i, map(\mathbf{c}_i))}{n},$$

where n is the total number of data, $\delta(.,.)$ is the delta function defined by:

$$\delta(\mathbf{x},\mathbf{y}) = \begin{cases} 1, & if\ \mathbf{x} = \mathbf{y}, \\ 0, & otherwise \end{cases}$$

the function $map(\mathbf{c}_i)$ represents the optimal mapping that permutes clustering labels to align with the true labels using the Kuhn-Munkres algorithm.

NMI is defined in the following manner:

$$\text{NMI} = \frac{\text{MI}(\mathbf{C},\mathbf{C}')}{\max(\text{H}(\mathbf{C}),\text{H}(\mathbf{C}'))},$$

where $\mathbf{C}$ and $\mathbf{C}'$ are clustering labels and the truth labels respectively. Furthermore, $\text{H}(\mathbf{C})$ and $\text{H}(\mathbf{C}')$ are the entropies of $\mathbf{C}$ and $\mathbf{C}'$, respectively. $\text{MI}(\mathbf{C},\mathbf{C}')$ is the information entropy between $\mathbf{C}$ and $\mathbf{C}'$:

$$\text{MI}(\mathbf{C},\mathbf{C}') = \sum_{\mathbf{c}_i\in\mathbf{C},\mathbf{c}'_j\in\mathbf{C}'} p(\mathbf{c}_i,\mathbf{c}'_j).\log_2 \frac{p(\mathbf{c}_i,\mathbf{c}'_j)}{p(\mathbf{c}_i)p(\mathbf{c}'_j)},$$

where $p(\mathbf{c}_i)$ and $p(\mathbf{c}'_j)$ denote the probabilities a sample belongs to the clusters $\mathbf{c}_i$ and $\mathbf{c}'_j$ respectively. $p(\mathbf{c}_i,\mathbf{c}'_j)$ is the joint probability that a sample belongs to the clusters $\mathbf{c}_i$ and $\mathbf{c}'_j$ simultaneously.

4.1 Datasets

In this study, we utilize a variety of datasets to evaluate the performance of the proposed method. The datasets span different domains, including face image data and biological data, providing a comprehensive evaluation of the method's applicability.

Table 1: An overview of the datasets utilized in this study, in which m refers to the number of data samples, n refers to the number of features, and c is the number of distinct classes.

ID	Dataset	m	n	c	Type of Data
1	ORL	400	1024	40	Face Image Data
2	Jaffe	213	676	10	Face Image Data
3	Orlraws10P	100	10304	10	Face Image Data
4	Yale	165	1024	15	Face Image Data
5	Prostate_GE	102	5966	2	Biological Data

The datasets used in this study are detailed as follows:

- **ORL**: This dataset consists of face images from 40 distinct individuals, with 10 different images per individual. Each image is resized to 32 × 32 pixels, resulting in 1024 features per image.
- **Jaffe**: The Jaffe dataset contains images of 10 different Japanese female subjects posing with different facial expressions. Each image has 676 features.
- **Orlraws10P**: This dataset comprises raw face images with 10304 features each, collected from 10 individuals with 10 images per person.
- **Yale**: The Yale face database includes 165 grayscale images of 15 individuals, with each image resized to 32 × 32 pixels, resulting in 1024 features per image.
- **Prostate_GE**: This dataset contains gene expression data from 102 prostate samples, with 5966 features each. The samples are categorized into two classes: tumor and normal.

The diversity of these datasets, covering both image and biological data, ensures that our method is robust and versatile across different types of data and applications.

4.2 Comparison methods

The effectiveness of our proposed method is evaluated by comparing it with several established unsupervised feature selection techniques. The following methods are included in our comparative analysis:

1. **Baseline**: This method uses all original features without any selection.
2. **LS** [11]: Selects features based on their variance while preserving local data structure using Laplacian score.
3. **MCFS** [12]: Selects an optimal subspace of features to best preserve multi-cluster structures within the data.
4. **RSR** [13]: Utilizes self-representation to model each feature as a linear combination of others, promoting sparsity using $\mathbf{L}_{2,1}$-norm.
5. **LS-CAE** [14]: Proposes an unsupervised feature selection approach using concrete layer mechanisms and Laplacian score, optimizing feature selection and reconstruction objectives.
6. **CD-LSR** [15]: Conducts feature selection based on $\mathbf{L}_{2,0}$-norm using simple least square regression, achieving efficient feature subset selection.

Each method offers distinct approaches to feature selection, focusing on various aspects such as variance, structure preservation, and sparsity enforcement. These comparative evaluations provide insights into the effectiveness of our proposed method across different datasets and scenarios.

4.3 Results and Analysis

The proposed method, SRFSNMF, is evaluated using clustering accuracy (ACC) and normalized mutual information (NMI) across multiple datasets. Tables 2 and 3 summarize the results, highlighting the best performance for each dataset in bold.

Table 2: ACC results for different datasets. Top-performing results are highlighted in bold.

Algorithms	Yale	Jaffe	Orlraws10P	ORL	Prostate_GE
Baseline	38.79	87.17	72.15	51.70	57.84
LS	38.90	88.26	67.00	38.78	60.68
MCFS	38.63	90.77	76.95	49.82	56.86
RSR	40.00	85.91	68.00	46.50	61.74
LS-CAE	42.24	90.14	74.00	56.75	61.76
CD-LSR	38.97	90.14	78.70	49.46	**63.72**
SRFSNMF	**42.43**	**92.49**	**81.00**	**58.01**	62.73

The results in Tables 2 and 3 indicate that SRFSNMF consistently outperforms the baseline and other comparative methods across most datasets in both ACC and NMI metrics.

- The proposed SRFSNMF method consistently outperforms the Baseline approach, highlighting that feature selection can improve clustering performance by reducing the impact of irrelevant and redundant features.

Table 3: NMI results for different datasets. Top-performing results are highlighted in bold.

Algorithms	Yale	Jaffe	Orlraws10P	ORL	Prostate_GE
Baseline	44.19	87.87	77.57	70.79	1.80
LS	43.51	90.11	73.27	61.58	4.51
MCFS	43.92	90.94	82.75	69.17	1.33
RSR	46.83	85.52	81.25	68.64	6.80
LS-CAE	48.32	91.10	80.35	74.99	7.54
CD-LSR	44.23	90.45	83.18	69.01	**7.93**
SRFSNMF	**49.70**	**92.02**	**84.47**	**75.05**	7.64

- SRFSNMF achieves higher accuracy and mutual information scores in most cases compared to other methods. However, it shows relatively lower performance on the Prostate_GE dataset, which may be attributed to the complexity of the biological data.
- With the exception of the Prostate_GE dataset, Figures 4 and 5 show that SRFSNMF surpasses all other unsupervised feature selection methods on the remaining datasets in terms of both ACC and NMI, indicating its robustness and effectiveness across diverse datasets.

Overall, the proposed SRFSNMF method demonstrates consistently strong performance across various datasets, indicating that it is an effective approach for unsupervised feature selection and clustering. The method effectively captures the inherent structure of the data, leading to higher clustering accuracy and meaningful clusters.

Figures 1 and 2 further illustrate the performance of SRFSNMF, showing the variation of ACC and NMI with different numbers of selected features across datasets. These charts highlight the stability of SRFSNMF compared to other unsupervised feature selection methods.

In conclusion, SRFSNMF proves to be a robust method, consistently yielding high ACC and NMI across diverse datasets, thus validating its effectiveness in unsupervised feature selection and clustering tasks.

4.4 Parameter Sensitivity Analysis

In order to evaluate the effect of the regularization parameter α on the clustering performance of the proposed method, a sensitivity analysis was conducted. This analysis aims to assess how variations in α impact the accuracy (ACC) and normalized mutual information (NMI) across different datasets.

Figure 3 presents 3D plots that illustrate the trends in ACC and NMI for the ORL, Jaffe, and Yale datasets. In these plots, the x-axis represents the parameter values of α, ranging from $\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 10^{2}, 10^{4}, 10^{6}, 10^{8}\}$, while the y-axis denotes the number of selected features, varying be-

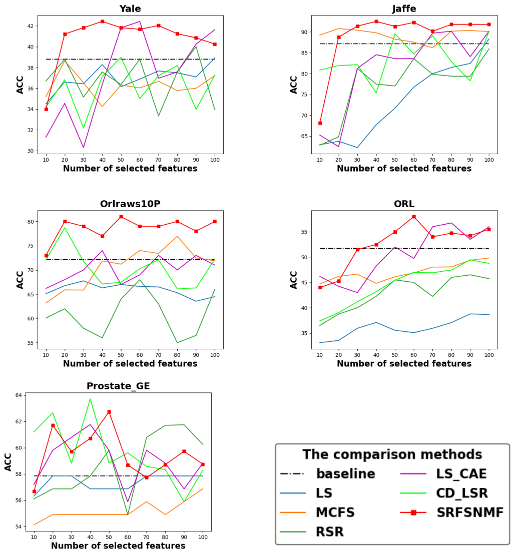

Figure 1: ACC of different methods on five datasets in terms of different numbers of selected features.

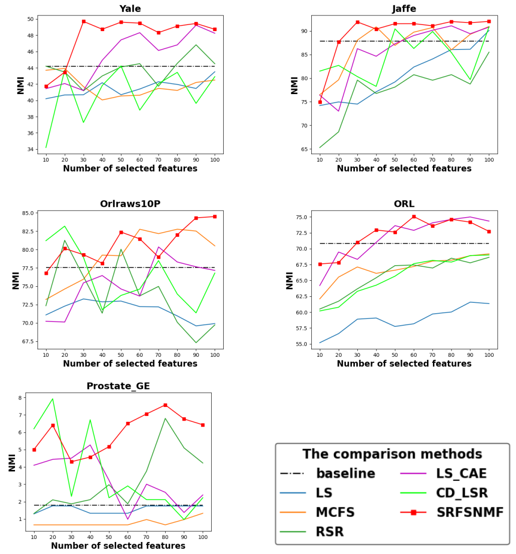

Figure 2: NMI of different methods on five datasets in terms of different numbers of selected features.

tween $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. The z-axis reflects the values of ACC and NMI.

From these plots, it is evident that the clustering performance shows sensitivity to changes in the regularization parameter. Specifically, for the ORL dataset, the variations in ACC and NMI with respect to α are relatively small across the different numbers of selected features, suggesting that the proposed SRFSNMF model is not highly sensitive to α within the tested range.

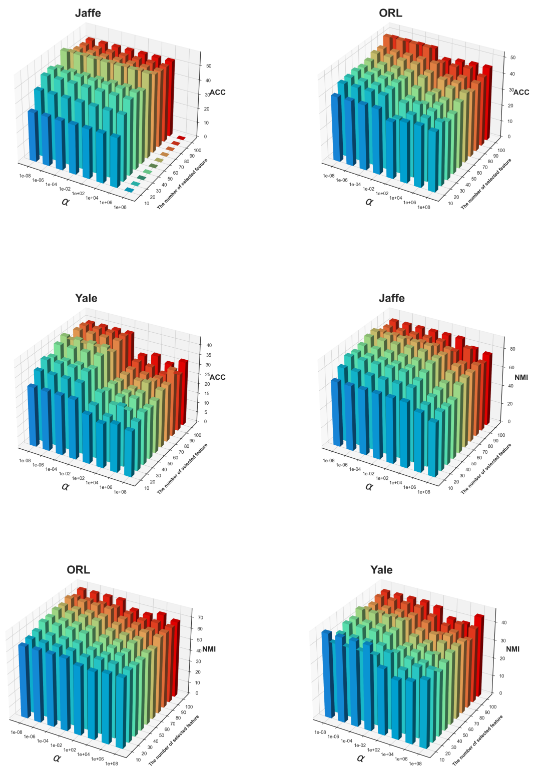

Figure 3: 3D plots showing the relationship between the number of selected features, parameter values (α), and ACC or NMI for the Jaffe, ORL, and Yale datasets. The x-axis represents the parameter values, the y-axis shows the number of selected features, and the z-axis indicates ACC or NMI values.

4.5 Robustness Against Noise

To test the resilience of the proposed method against different outlier intensities, three types of noise were introduced: Salt-and-Pepper Noise, Gaussian Noise, and Block Occlusion. These experiments simulate real-world conditions where images are corrupted by noise or occluded, offering insights into the robustness of the proposed method. The ORL dataset, containing facial images, was used for this evaluation.

Salt-and-Pepper Noise. This type of noise simulates pixel-level corruption by randomly flipping a percentage of pixels to either black or white. The Salt-and-Pepper noise density is varied as 0.05, 0.1, 0.15, and 0.2, corresponding to 5%, 10%, 15%, and 20% of corrupted pixels, respectively. As demonstrated in Figure 4, increasing the level of Salt-and-Pepper noise introduces progressively more random pixel disruptions, which affect the clarity of the image. Despite the presence of these noise artifacts, the proposed algorithm demonstrated strong robustness in maintaining performance as the noise intensity increased.

Gaussian Noise. Gaussian noise was applied to further challenge the method, with standard deviations (σ) set to 0.5, 1, 1.5, and 2. This introduces random fluctuations in pixel intensities across the entire image, simulating sensor noise or other types of uniform interference. The effect of Gaussian noise increases with larger σ values, resulting in noticeable degradation in image quality. As shown in Figure 5, the images become increasingly blurred as the noise intensity increases, but the method was able to maintain its stability even in the presence of heavy Gaussian noise.

Block Occlusion. In addition to pixel-level noise, block occlusion was introduced to simulate missing parts of the data. Square blocks of sizes 3×3, 4×4, 5×5, and 6×6 were randomly placed on the images to mimic occlusions. Larger block sizes led to more significant occlusion of the facial features, as shown in Figure 6. This experiment tested the models ability to handle partially missing information, a common issue in image datasets affected by occlusion or masking.

By applying these three types of noise and occlusion, we were able to examine the algorithms resilience in challenging, noisy environments. The results indicate that the method remains robust even as noise intensity and occlusion levels increase, highlighting its suitability for real-world applications where data imperfections are inevitable.

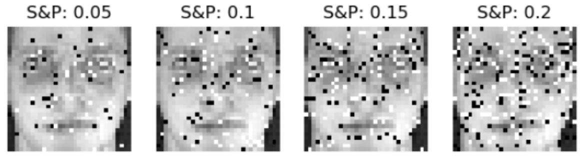

Figure 4: Results of the Salt & Pepper noise.

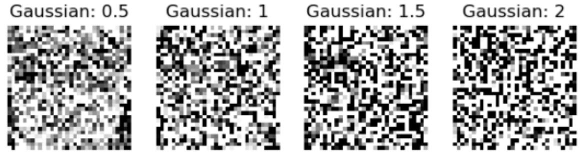

Figure 5: Results of the Gaussian noise.

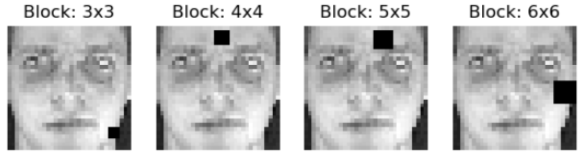

Figure 6: Results of the Occluded

5 Conclusion

In this article, we introduced a novel method, Self-Representation Feature Selection based on Non-Negative Matrix Factorization (SRFSNMF), to address the challenges of unsupervised feature selection in high-dimensional data. By leveraging the Gram matrix from NMF and combining it with self-representation techniques, we effectively captured the structural relationships between samples, leading to more meaningful and relevant feature selection. The proposed iterative optimization algorithm ensures efficient convergence, and experimental results on multiple benchmark datasets demonstrated that SRFSNMF significantly outperforms state-of-the-art methods in terms of feature selection performance. We also conducted a thorough sensitivity analysis, confirming the robustness of SRFSNMF across various parameter settings. Additionally, the method proved resilient to noisy data, further validating its stability and reliability. Looking ahead, future research could focus on strengthening the relationship between NMF and self-representation to achieve even more effective feature selection. Furthermore, exploring methods that simultaneously consider both the sample space and feature space could provide deeper insights into complex data structures and enhance the overall performance of unsupervised feature selection models.

References

[1] P. Tiwari, F. S. Movahed, S. Karami, F. Saberi-Movahed, J. Lehmann, and S. Vahdati A self-representation learning method for unsupervised feature selection using feature space basis. *Transactions on Machine Learning Research*, 2024.

[2] V. Jannesari, M. Keshvari, and K. Berahmand A novel nonnegative matrix factorization-based model for attributed graph clustering by incorporating complementary information. *Expert Systems with Applications*, 242:122799, 2024.

[3] C. Shao, M. Chen, Y. Yuan, and Q. Wang Projection concept factorization with self-representation for data clustering. *Neurocomputing*, 517:62–70, 2023.

[4] R. Chen Robust dual-graph regularized and minimum redundancy based on self-representation for semi-supervised feature selection. *Neurocomputing*, 490:104–123, 2022.

[5] J. Miao, Y. Ping, Z. Chen, X.-B. Jin, P. Li, and L. Niu Unsupervised feature selection by non-convex regularized self-representation. *Expert Systems with Applications*, 173:114643, 2021.

[6] C. Tang, X. Liu, M. Li, P. Wang, J. Chen, L. Wang, and W. Li Robust unsupervised feature selection via dual self-representation and manifold regularization. *Knowledge-Based Systems*, 145:109–120, 2018.

[7] Y.-T. Guo, Q.-Q. Li, and C.-S. Liang The rise of nonnegative matrix factorization: algorithms and applications. *Information Systems*, page 102379, 2024.

[8] Y. Dong, H. Che, M.-F. Leung, C. Liu, and Z. Yan Centric graph regularized log-norm sparse non-negative matrix factorization for multi-view clustering. *Signal Processing*, 217:109341, 2024.

[9] F. Yahaya, M. Puigt, G. Delmaire, and G. Roussel A framework for compressed weighted nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, pages 1–13, 2024.

[10] A. Hajiveiseh, S. A. Seyedi, and F. Akhlaghian Tab Deep asymmetric nonnegative matrix factorization for graph clustering. *Pattern Recognition*, 148:110179, 2024.

[11] X. He, D. Cai, and P. Niyogi Laplacian score for feature selection. In *Advances in Neural Information Processing Systems*, volume 18, pages 507–514, 2005.

[12] D. Cai, C. Zhang, and X. He Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 333–342, 2010.

[13] P. Zhu, W. Zuo, L. Zhang, Q. Hu, and S. C. K. Shiu Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446, 2015.

[14] U. Shaham, O. Lindenbaum, J. Svirsky, and Y. Kluger Deep unsupervised feature selection by discarding nuisance and correlated features. *Neural Networks*, 152:34–43, 2022.

[15] L. Xu, R. Wang, F. Nie, and X. Li Efficient top-k feature selection using coordinate descent method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, number 9, pages 10594–10601, 2023.