Supplementary material: Unsupervised feature selection by learning exponential weights

Chenchen Wang[a,b], Jun Wang[c], Zhichen Gu[a,b], Jin-Mao Wei[a,b], Jian Liu[a,b]

[a] *College of Computer Science, Nankai University, Tianjin, 300350, P. R. China.*
[b] *Institute of Big Data, Nankai University, Tianjin, 300350, P. R. China.*
[c] *School of Mathematics and Statistics Science, Ludong University, Yantai, 264025, P. R. China.*

Contents

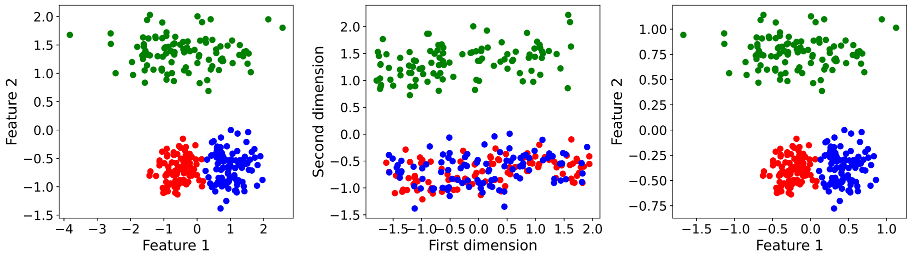
A. Visualization experiments on synthetic data set

Here we observe sample distributions and feature structures in different spaces on synthetic datasets. First, we give the data set description and experimental results of the synthetic data set. Table S.1 details the three categories derived from three Gaussian distributions with different means and variances. The first two features constitute the real sample distribution, and the last two noise-dimensional features are obtained from the uniform distribution.

Table S.1: Details description of synthetic data set

Class	First and second features	Third feature	Fourth feature
C1	100 samples are generate from the gaussian distribution with $\mu = [1, 1]$, $\sigma^2 = \begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}$	100 samples are generate from the uniform distribution over $[0, 5]$	100 samples generate from uniform distribution over $[1, 4]$
C2	100 samples are generate from the gaussian distribution with $\mu = [2, 1]$, $\sigma^2 = \begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}$	100 samples are generate from the uniform distribution over $[0, 5]$	100 samples are generate from the uniform distribution over $[1, 4]$
C3	100 samples are generate from the gaussian distribution with $\mu = [1, 3]$, $\sigma^2 = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.05 \end{bmatrix}$	100 samples are generate from the uniform distribution over $[0, 5]$	100 samples are generate from the uniform distribution over $[1, 4]$

The sample distribution and feature structure calculated in the low-dimensional space become distorted compared to those in the original feature subspace, as illustrated in Figure S.1. Figure S.1(b) depicts the distorted sample distribution in the 2-dimensional embedding space, contrasting with the real sample distribution shown in Figure S.1(a). Our model can learn a separated sample distribution in the feature subspace that is well suited for downstream learning, as demonstrated in Figure S.1(c).



(a) Real Distribution (b) Low-dimensional Space (c) Feature Subspace(Ours)

Figure S.1: Comparison of the real distribution, distribution learned in the low-dimensional embedding space, and distribution learned in the feature subspace. Note that the inconsistency of the data centers in Figure S.1 (a) and Table S.1 is due to the normalization of the data.

The sample similarity matrix produced in different spaces is shown in Figure S.2. The ideal similarity matrix is block-diagonal. In the original space, Figure S.2(a) displays the sample similarity structure calculated by the k-nearest neighbor (k-nn) graph with $k = 5$. Figure S.2(b) shows the sample similarity structure learned by SOGFS in a low-dimensional space. The sample similarity matrix learned by the proposed method LLSRFS in the feature subspace is shown in Figure S.2(c). We see that noise features impact the fixed similarity structure learned by the k-nn graph from Figure S.2(a). Due to the low-dimensional representation deforms the original sample distribution and feature structure, the similarity structure learned by SOGFS cannot correctly discriminate the three categories. LLSRFS learns the similarity structure in the feature subspace and derives the ideal block-diagonal while preserving the original sample distribution and feature structure.

The wine dataset contains three classes and 178 samples, which can be obtained from the UCI machine learning repository. We choose SOGFS as a representative approach for learning graph structures in the low-dimensional embedding space. We set the number of the low-dimensional embeddings of SOGFS to the number of categories and visualize its low-dimensional embedding representation in Figure 1(a). Figure 1(b) shows the sample distribution in the feature subspace learned by our method with the top three features.

B. Proof of Theorem 1

Theorem 1. *Let* $\mathbf{Q} = \mathbf{W}\mathbf{P}$ *and* $q = \frac{1}{p} - \frac{1}{2}$*, where* $0 < p \leq 1$*. Then, the problem in Eq. (11) is equivalent to the following sparse regression problem with* $\ell_{2,p}^2$*-norm:*

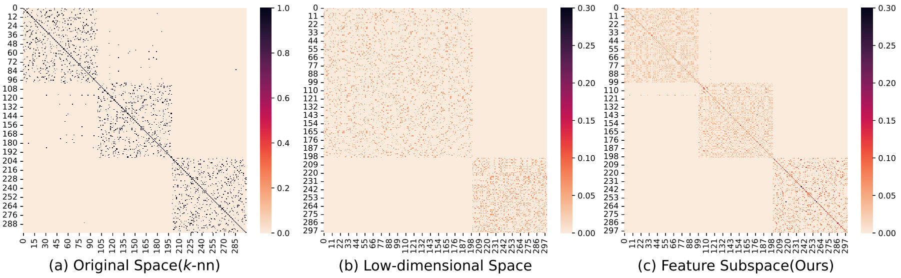
$$\min_{\mathbf{Q},\mathbf{b}} \|\mathbf{X}^T\mathbf{Q} + \mathbf{1}\mathbf{b}^T - \mathbf{F}\|_F^2 + \lambda_w \|\mathbf{Q}\|_{2,p}^2, \tag{S.1}$$

which is equivalent to the rescaled least squares regression method in [24]. The solution of problem (11) about $\mathbf{w}$ *is*

$$w_l = \frac{(\|\mathbf{Q}^l\|_2^2)^{\frac{1}{2q+1}}}{\sum_{l'}(\|\mathbf{Q}^{l'}\|_2^2)^{\frac{1}{2q+1}}}. \tag{S.2}$$

Proof. First, we give the general form of problem (11)

$$\begin{aligned} &\min_{\mathbf{w}} \|\mathbf{X}^T\mathbf{W}\mathbf{P} + \mathbf{1}\mathbf{b}^T - \mathbf{F}\|_F^2 + \lambda_w\|\mathbf{P}\|_F^2 \\ &s.t. \ \ \mathbf{w}^T\mathbf{1} = 1, \mathbf{w} > 0, \end{aligned} \tag{S.3}$$



(a) Original Space(*k*-nn) (b) Low-dimensional Space (c) Feature Subspace(Ours)

Figure S.2: Similarity structure of samples in different spaces

where $\mathbf{W} = diag(\mathbf{w})^q$. Let $\mathbf{Q} = diag(\mathbf{w})^q\mathbf{P}$, then

$$\begin{aligned}&\min_{\mathbf{w}}||\mathbf{X}^T\mathbf{W}\mathbf{P} + \mathbf{1}\mathbf{b}^T - \mathbf{F}||_F^2 + ||diag(\mathbf{w})^{-q}\mathbf{Q}||_F^2 \\ &= ||\mathbf{X}^T\mathbf{Q} + \mathbf{1}\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_w \sum_l \frac{||\mathbf{Q}^l||_2^2}{w_l^{2q}}.\end{aligned} \tag{S.4}$$

Then the Lagrangian function of problem (S.3) with respect to $\mathbf{w}$ is

$$\mathcal{L}(\mathbf{w}, \beta, \eta) = \sum_l \frac{||\mathbf{Q}^l||_2^2}{w_l^{2q}} + \beta(\sum_l w_l - 1) - \eta\mathbf{w}. \tag{S.5}$$

Setting the partial derivative of the above function with respect to w_l as 0, we have

$$\frac{\partial\mathcal{L}(\mathbf{w}, \beta, \eta)}{\partial w_l} = -2q\frac{||\mathbf{Q}^l||_2^2}{w_l^{2q+1}} + \beta = 0. \tag{S.6}$$

The optimal solution of w_l is

$$w_l = (\frac{2q||\mathbf{Q}^l||_2^2}{\beta})^{\frac{1}{2q+1}}. \tag{S.7}$$

Combined with $\sum_l w_l = 1$, we have

$$w_l = \frac{(||\mathbf{Q}^l||_2^2)^{\frac{1}{2q+1}}}{\sum_{l'}(||\mathbf{Q}^{l'}||_2^2)^{\frac{1}{2q+1}}}. \tag{S.8}$$

Then, we have

$$W_{ll} = (\frac{(||\mathbf{Q}^l||_2^2)^{\frac{1}{2q+1}}}{\sum_{l'}(||\mathbf{Q}^{l'}||_2^2)^{\frac{1}{2q+1}}})^q. \tag{S.9}$$

Let $p = \frac{2}{2q+1}$,

$$\begin{aligned}||\mathbf{W}^{-1}\mathbf{Q}||_F^2 &= \sum_l \frac{||\mathbf{Q}^l||_2^2}{w_l^{2q}} = \sum_l \frac{||\mathbf{Q}^l||_2^2}{(\frac{(||\mathbf{Q}^l||_2^2)^{\frac{1}{2q+1}}}{\sum_{l'}(||\mathbf{Q}^{l'}||_2^2)^{\frac{1}{2q+1}}})^{2q}} \\ &= \sum_l \frac{(\sum_{l'}(||\mathbf{Q}^{l'}||_2^2)^{\frac{1}{2q+1}})^{2q}}{||\mathbf{Q}^l||_2^{\frac{-2}{2q+1}}} \\ &= \sum_l (\sum_{l'}(||\mathbf{Q}^{l'}||_2^2)^{\frac{1}{2q+1}})^{2q}(||\mathbf{Q}^l||_2^{\frac{2}{2q+1}}) \\ &= (\sum_{l'}(||\mathbf{Q}^{l'}||_2^2)^{\frac{1}{2q+1}})^{2q} \sum_l (||\mathbf{Q}^l||_2^{\frac{2}{2q+1}}) \\ &= (\sum_l ||\mathbf{Q}^l||_2^{\frac{2}{2q+1}})^{2q+1} \\ &= (\sum_l ||\mathbf{Q}^l||_2^p)^{2/p} = ||\mathbf{Q}||_{2,p}^2\end{aligned} \tag{S.10}$$

Then, problem (S.3) is equivalent to problem (S.11) with $0 < p \leq 1$

$$\min_{\mathbf{Q}} ||\mathbf{X}^T\mathbf{Q} + \mathbf{1}\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_w ||\mathbf{Q}||_{2,p}^2. \tag{S.11}$$

□

C. Proof of Remark 1

Remark 1. The optimization problem (11) is approximately equivalent to the following sparse regression problem with $\ell_{2,p}^p$-norm,

$$\min_{\mathbf{Q},\mathbf{b}} ||\mathbf{X}^T\mathbf{Q} + \mathbf{1}\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_w ||\mathbf{Q}||_{2,p}^p, \tag{S.12}$$

where $0 < p \leq 2$.

Proof. Let $\mathbf{Q} = diag(\mathbf{w})\mathbf{P}, \mathbf{w}^T\mathbf{1} = 1, 0 \leq \mathbf{w} \leq 1$, then problem (S.12) is equivalent to

$$\min_{\mathbf{w}} ||\mathbf{X}^T diag(\mathbf{w})\mathbf{P} + \mathbf{1}\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_w ||diag(\mathbf{w})\mathbf{P}||_{2,p}^p. \tag{S.13}$$

Furthermore, we have

$$\min_{\mathbf{w}} ||\mathbf{X}^T diag(\mathbf{w})\mathbf{P} + \mathbf{1}\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_w \sum_l w_l^p ||\mathbf{P}^l||_2^p. \tag{S.14}$$

Since the weight of feature l is determined by w_l, $||\mathbf{P}^l||_2^p$ can be approximated by $||\mathbf{P}^l||_2^2$:

$$\min_{\mathbf{w}} ||\mathbf{X}^T diag(\mathbf{w})\mathbf{P} + \mathbf{1}\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_w ||diag(\mathbf{w})^{\frac{p}{2}}\mathbf{P}||_F^2 \tag{S.15}$$

Let $\mathbf{M} = diag(\mathbf{w})^{\frac{p}{2}}\mathbf{P}$, then we have

$$\min_{\mathbf{w}} ||\mathbf{X}^T diag(\mathbf{w})^{1-\frac{p}{2}}\mathbf{M} + \mathbf{1}\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_w||\mathbf{M}||_F^2 \tag{S.16}$$

Let $q = 1 - \frac{p}{2}$, then problem (S.16) is approximately equivalent to problem (11). □

D. Analysis of exponentially weighted sparse regression

In this section, we further verify the effectiveness of Exponentially Weighted Sparse Regression (EWSR) by comparing it with other sparse regression methods. We choose Linear Square Regression with $\ell_{2,p}$-norm (LSR-L2p) and SRLSR as comparison methods. SRLSR is equivalent to EWSR when $0.5 < q < 1$, so we use EWSR for simplicity. We vary the regularization parameters λ_w of LSR-L2p and EWSR from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. The parameters p and q are searched from $\{0.2, 0.4, 0.6, 0.8, 1\}$. The number of selected features ranges from 10 to 100. We use Linear-SVM and KNN classifiers (with $C = 1$ and $k = 5$ respectively) to evaluate the classification performance. We use 10-fold cross-validation to give the classification accuracy of each set of parameters. We report the best classification accuracy for feature subsets of different sizes, as shown in Table S.2.

Table S.2: Classification accuracy of Linear-SVM classifiers and KNN classifiers by different algorithms on the lung_discrete data set.

#Dim	SVM for LSR-L2p with different p					SVM for EWSR with different q				
	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
10	0.5911	0.6161	0.5911	0.6054	**0.6911**	0.6446	0.6446	0.6304	0.6304	0.6250
20	0.6929	0.6946	0.6500	0.7214	0.7446	0.7446	0.7571	**0.7714**	0.7571	**0.7714**
30	0.7714	0.7786	0.7321	0.8161	0.7768	0.7804	**0.8179**	0.7732	0.7643	0.7786
40	0.7661	0.8482	0.8357	0.8357	**0.8589**	0.8321	0.8304	0.8304	0.8446	0.8214
50	0.8357	0.8196	0.8036	0.8179	**0.8732**	0.8607	0.8464	**0.8732**	0.8357	0.8179
60	0.8732	0.8321	0.8054	0.8607	0.8875	**0.9018**	0.8875	0.8875	0.8732	0.8607
70	0.8589	0.8339	0.8464	0.8750	0.8589	**0.8875**	**0.8875**	0.8732	**0.8875**	**0.8875**
80	0.8446	**0.8750**	0.8357	0.8732	0.8589	0.8732	0.8732	0.8732	0.8732	0.8732
90	0.8625	0.8750	0.8625	0.8625	0.8625	0.8732	0.8732	**0.8875**	**0.8875**	**0.8875**
100	0.8875	0.8750	0.8589	0.8589	**0.8875**	**0.8875**	**0.8875**	**0.8875**	**0.8875**	**0.8875**
#Dim	KNN for LSR-L2p with different p					KNN for EWSR with different q				
	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
10	0.6536	0.6804	**0.7232**	0.6821	0.6518	0.6143	0.6268	0.6482	0.6589	0.6625
20	0.7607	0.7625	0.7464	0.7464	**0.7732**	0.7304	0.7446	0.7304	0.7464	0.7464
30	0.7768	0.7875	0.8018	**0.8161**	0.7875	0.7875	0.8179	0.8018	0.7875	0.8018
40	0.8321	0.8304	0.7875	0.8161	0.8018	0.8161	0.8161	0.8161	0.8161	**0.8446**
50	0.8304	0.8018	**0.8304**	**0.8304**	0.8161	0.8161	**0.8304**	0.8018	0.8018	0.8161
60	0.8464	**0.8464**	0.8214	0.8446	0.8304	0.8304	0.8304	0.8304	0.8214	**0.8464**
70	0.8321	0.8321	0.8304	0.8018	0.8304	0.8304	**0.8446**	0.8304	0.8196	0.8304
80	0.8321	0.8321	0.8321	0.8321	0.8321	0.8304	0.8304	0.8321	**0.8446**	0.8321
90	0.8196	0.8161	0.8161	0.8161	**0.8304**	**0.8304**	**0.8304**	**0.8304**	0.8161	0.8161
100	0.8304	0.8196	**0.8304**	0.8161	0.8161	0.8161	**0.8304**	0.8161	**0.8304**	**0.8304**

From Table S.2, we can draw the following conclusions and insights. 1) LSR-L2p performs better with larger values of p when the number of selected features is small. This contradicts the theoretical expectation that smaller p leads to a sparser mapping matrix and more concentrated feature weights. Therefore, the optimal feature subset cannot be directly obtained using the row

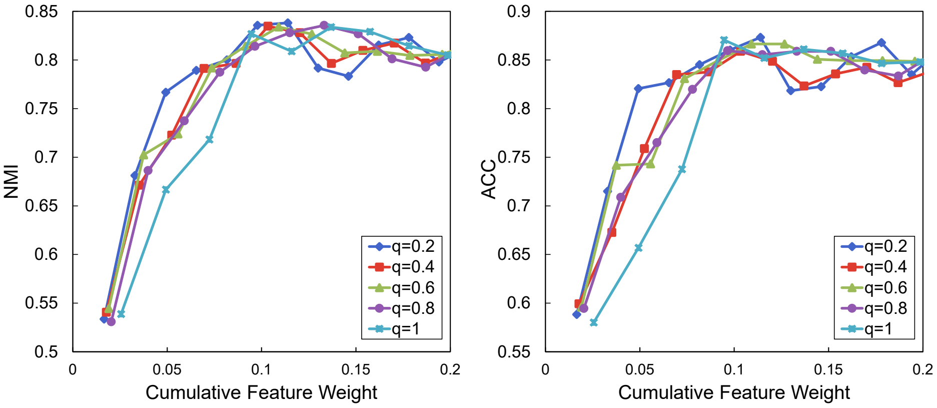


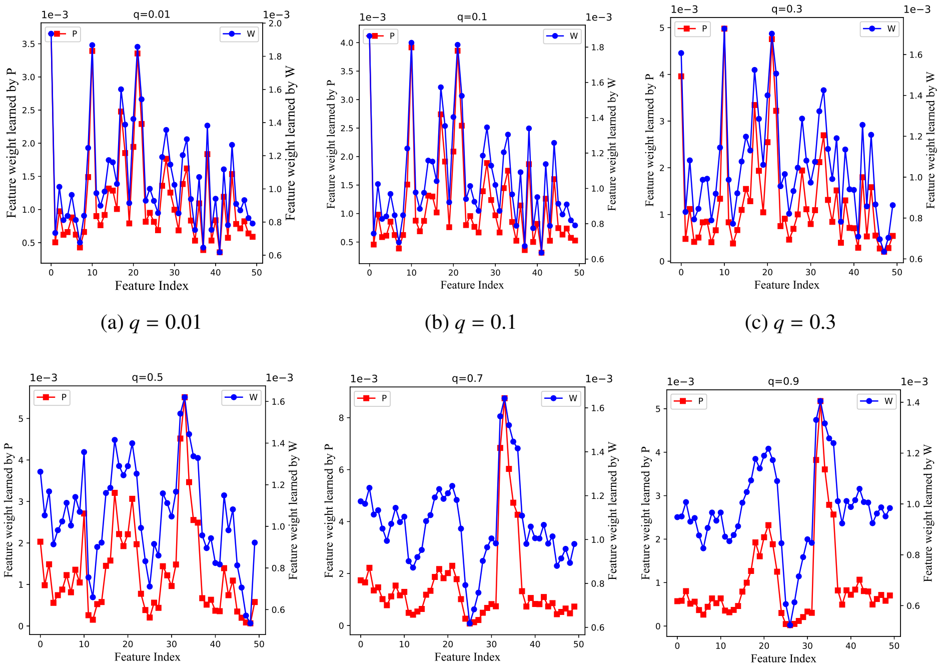
Figure S.3: NMI and ACC results of cumulative features on the lung_discrete data set

norm of the mapping matrix as a criterion. This phenomenon further confirms our argument that the feature weights calculated by the row norm of the mapping matrix are influenced by a few large-weight features.

2) Our method outperforms LSR-L2p in most cases under different numbers of features, demonstrating the effectiveness of our exponentially weighted sparse regression. Moreover, our method excels LSR-L2p when the number of selected features is large. For instance, our models with different q all achieve the best SVM classification accuracy with 100-dimensional features. This is because we use the exponent q to reduce the weight of large-weight features so that more features can share the weight. When more features are selected, our method can find better feature combinations. Besides, our method achieves an accuracy of 0.9018 with $q = 0.2$, $d = 50$ on SVM classification, while LSR-L2p only achieves an accuracy of 0.8875 with $p = 1$, $d = 100$.

Further, we investigated how feature weight distribution affects feature selection on the lung_discrete data set using EWSR with different q values. We defined cumulative feature weight as the sum of top-k feature weights, where k ranges from 5 to 45. We evaluated the feature subset using NMI and ACC metrics and presented the results in Figure S.3. We observe that for a given cumulative feature weight, NMI and ACC increase as q decreases. This is because smaller q values reduce the weight of large-weight features more, enabling us to select more diverse feature combinations.

Figure S.4 illustrates how the feature mapping matrix $\mathbf{P}$ and the feature selection matrix $\mathbf{W}$ are related in EWSR. We plot the first 50-dimensional feature weights of the Yale data set with different values of q. We have the following observations. When $q = 0.1$, $\mathbf{P}$ and $\mathbf{W}$ have very similar feature distributions. This is expected because when $q = 0$, our model reduces to least square regression. As q increases, the distribution gap between $\mathbf{P}$ and $\mathbf{W}$ also increases, which indicates that q affects the feature weight distribution and combination. The exponential weighting compresses the large weights, spreads the weights among small and medium-weight features, and enlarges the difference between small and medium weights. This enhances the role of small and medium-weight features in model learning and avoids the dominance of large-weight features.



(a) q = 0.01

(b) q = 0.1

(c) q = 0.3

(d) q = 0.5

(e) q = 0.7

(f) q = 0.9

Figure S.4: Feature weights learned by **P** and **W** on the Yale data set.

E. Ablation study

In this section, we present the results of the ablation experiments to verify the influence of each component in our model. Firstly, we remove the local structure learning component to ascertain the effect of exponentially weighted sparse regression. The optimization objection of exponentially weighted sparse regression for unsupervised feature selection (named as LLSRFS w/o LL) is

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{F},\mathbf{w}} ||\mathbf{H}\mathbf{X}^T\mathbf{P} - \mathbf{H}\mathbf{F}||_F^2 + \lambda_w \sum_l \frac{||\mathbf{p}_l||_2^2}{w_l^{2q}} + \lambda_f Tr(\mathbf{F}^T\mathbf{L_S}\mathbf{F}), \\ &s.t.\ \mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{w}^T\mathbf{1} = 1, \mathbf{w} > 0. \end{aligned} \tag{S.17}$$

The matrix **P** can be optimized using Eq. (27) and **F** is updated by problem (30). The optimization of **w** is carried out in Eq. (S.9).

Further, we only consider the local structure learning component by removing the exponentially weighted sparse regression to solve problem (16). However, solving **w** in this manner can only get one feature with one weight. To avoid the trivial solution, we use negative entropy $\sum_l w_l \log w_l$ as the regularization term of vector **w**. When w_l has equal values, $\sum_l w_l \log w_l$ reaches its minimum value. The optimization objective function of local structure learning for

Table S.3: NMI and ACC results with different q on the LUNG data set.

#Dim	NMI with different q						
	0.01	0.1	0.3	0.5	0.7	0.9	1
10	0.5616±0.0069	0.5932±0.0015	0.6846±0.0111	**0.6878±0.0070**	0.6120±0.0042	0.6851±0.0100	0.6870±0.0064
20	0.5766±0.0049	0.5919±0.0096	0.6471±0.0063	0.6598±0.0062	0.6646±0.0086	**0.6727±0.0000**	0.6548±0.0058
30	0.6660±0.0084	0.6591±0.0037	0.6646±0.0048	**0.6806±0.0024**	0.6650±0.0083	0.6599±0.0079	0.6737±0.0019
40	**0.8179±0.0130**	0.7895±0.0115	0.8101±0.0276	0.6663±0.0077	0.6866±0.0048	0.7563±0.0172	0.6807±0.0084
50	0.8042±0.0096	0.7843±0.0080	0.8086±0.0023	0.8111±0.0113	0.8110±0.0086	0.8079±0.0000	**0.8213±0.0000**
60	0.7862±0.0140	0.7749±0.0108	**0.8274±0.0074**	0.8267±0.0095	0.8072±0.0238	0.8086±0.0023	0.8144±0.0103
70	0.7962±0.0032	0.7980±0.0346	0.8164±0.0104	0.8207±0.0063	0.8202±0.0057	0.8376±0.0141	**0.8376±0.0097**
80	0.7831±0.0103	0.7860±0.0084	0.8052±0.0071	0.8095±0.0074	0.8179±0.0080	**0.8290±0.0186**	0.8247±0.0100
90	0.7713±0.0106	0.7863±0.0077	0.8049±0.0056	0.8132±0.0115	0.8152±0.0059	**0.8264±0.0075**	0.8145±0.0126
100	0.7832±0.0111	0.8018±0.0133	0.8102±0.0082	0.8133±0.0134	0.8135±0.0076	**0.8246±0.0080**	0.8155±0.0081
#Dim	ACC with different q						
10	0.6921±0.0108	0.7140±0.0070	0.8564±0.0061	0.8525±0.0043	0.7473±0.0150	**0.8635±0.0061**	0.8520±0.0040
20	0.6653±0.0144	0.6744±0.0286	**0.8453±0.0024**	0.7953±0.0167	0.7897±0.0023	0.7946±0.0517	0.8315±0.0076
30	0.7956±0.0059	0.7983±0.0036	0.8076±0.0294	**0.8372±0.0011**	0.8121±0.0189	0.8138±0.0240	0.8148±0.0042
40	**0.9530±0.0040**	0.9441±0.0023	0.9456±0.0260	0.8047±0.0112	0.8202±0.0099	0.9187±0.0128	0.8643±0.0111
50	0.9485±0.0033	0.9448±0.0030	0.9507±0.0000	0.9517±0.0040	0.9517±0.0030	0.9507±0.0000	**0.9557±0.0000**
60	0.9421±0.0034	0.9384±0.0025	0.9537±0.0024	**0.9571±0.0032**	0.9466±0.0018	0.9507±0.0000	0.9522±0.0032
70	0.9458±0.0016	0.9436±0.0025	0.9537±0.0036	0.9547±0.0020	0.9547±0.0020	0.9549±0.0039	**0.9552±0.0015**
80	0.9426±0.0036	0.9419±0.0030	0.9475±0.0023	0.9493±0.0023	0.9507±0.0000	**0.9520±0.0026**	0.9507±0.0000
90	0.9389±0.0045	0.9419±0.0025	0.9483±0.0025	**0.9517±0.0020**	0.9490±0.0023	0.9505±0.0011	0.9458±0.0038
100	0.9387±0.0040	0.9473±0.0050	0.9500±0.0028	0.9495±0.0026	0.9493±0.0035	**0.9512±0.0015**	0.9478±0.0036

unsupervised feature selection (named as LLSRFS w/o SR) is as follows:

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{S},\mathbf{F}} \sum_{ij}^{n}(\|\mathbf{x}_i-\mathbf{x}_j\|_w^2 S_{ij}+\alpha\mathbf{S}_{ij}^2)+\lambda_f Tr(\mathbf{F}^T\mathbf{L}_\mathbf{S}\mathbf{F})+\lambda_w\sum_l w_l\log w_l,\\ &s.t.\ \forall i, \mathbf{s}_i^T 1=1, 0\le\mathbf{s}_i\le 1, \mathbf{F}^T\mathbf{F}=\mathbf{I}, \mathbf{w}^T\mathbf{1}=1, \mathbf{w}>0. \end{aligned} \tag{S.18}$$

We use negtive entropy term to balance the appropriate weight between 0-1 coding and the same weight. For feature weight $\mathbf{w}$, we have

$$\begin{aligned} &\min_{\mathbf{w}} \sum_l w_l\delta_l+\lambda_w\sum_l w_l\log w_l\\ &s.t. \mathbf{w}^T\mathbf{1}=1, \mathbf{w}>0. \end{aligned} \tag{S.19}$$

The Lagrangian function of problem (S.19) is

$$\mathcal{L}(\mathbf{w},\beta)=\sum_l w_l\delta_l+\lambda_w\sum_l w_l\log w_l+\beta(\mathbf{w}^T\mathbf{1}-1). \tag{S.20}$$

Setting the partial derivative of the above function with respect to w_l as 0, we have

$$\frac{\partial\mathcal{L}}{\partial w_l}=\delta_l+\lambda_w(\log w_l+1)+\beta=0. \tag{S.21}$$

Thus, we have

$$w_l = exp(\frac{\delta_l + \beta}{\lambda_w} - 1). \tag{S.22}$$

Combined with $\sum_l w_l = 1$, we have

$$w_l = \frac{\exp(-\frac{\delta_l}{\lambda_w})}{\sum_{l'} \exp(-\frac{\delta_{l'}}{\lambda_w})}. \tag{S.23}$$

The optimal solution of $\mathbf{F}$ is formed by the r eigenvector corresponding to r smallest eigenvalues of $\mathbf{L_S}$, where r is the number of classes. When $\mathbf{F}, \mathbf{w}$ and $\mathbf{P}$ are fixed, the objective function about $\mathbf{S}$ is:

$$\begin{aligned} &\min_{\mathbf{s}_i} \sum_{ij}^{n} (\|\mathbf{x}_i - \mathbf{x}_j\|_w^2 S_{ij} + \alpha S_{ij}^2), \\ &s.t.\ \forall i, \mathbf{s}_i^T \mathbf{1} = 1, 0 \leq \mathbf{s}_i \leq 1. \end{aligned} \tag{S.24}$$

Let $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_w^2$, then we have

$$\begin{aligned} &\min_{\mathbf{s}_i} \sum_{ij}^{n} (d_{ij} S_{ij} + \alpha S_{ij}^2), \\ &s.t.\ \forall i, \mathbf{s}_i^T 1 = 1, 0 \leq \mathbf{s}_i \leq 1. \end{aligned} \tag{S.25}$$

The solution process of $\mathbf{s}_i$ and α is similar to that in the problem (23) and we omit the details.

For LLSRFS, LLSRFS w/o LL, and LLSRFS w/o SR, we searched q in the range $\{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$. For LLSRFS w/o LL, we searched λ_w and λ_f from the set $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. We searched parameter λ_w in LLSRFS w/o SR from $\{1, 10, 100, 1000\}$ to avoid numerical overflow. Figure S.5 displays the NMI and ACC results of LLSRFS w/o SR, LLSRFS w/o LL, and LLSRFS on the Urban, lung_discrete and Carcinoma data sets.

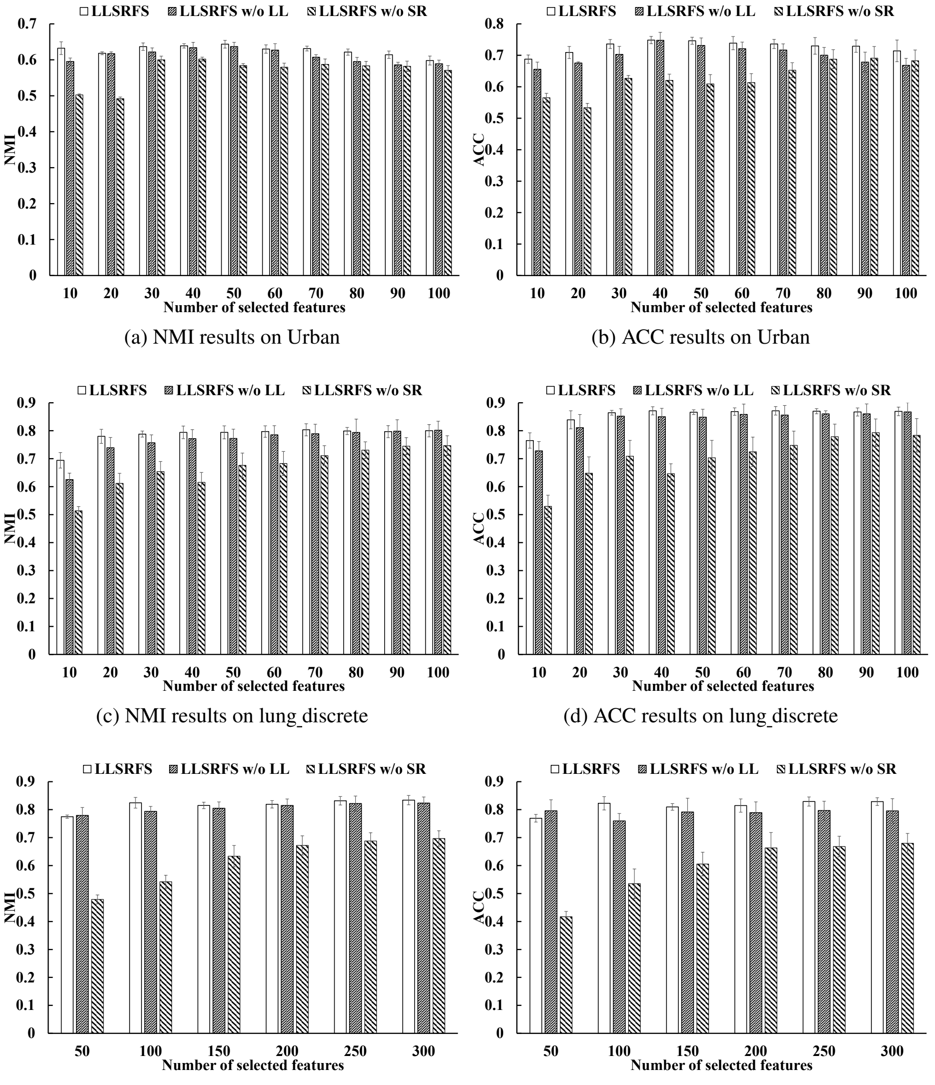
F. Convergence analysis

In Algorithm 1, we solve the problem (16) by decomposing it into four problems based on four variables and updating them iteratively. All sub-problems have convex objective losses and have the optimal solution theoretically. Here we give proofs of convergence for these three sub-problems.

Let $\{\mathbf{w}_l, \mathbf{S}_l, \mathbf{P}_l, \mathbf{F}_l\}$ be the solution for the l-th iteration, and $\mathcal{L}(\mathbf{w}_l, \mathbf{S}_l, \mathbf{P}_l, \mathbf{F}_l)$ be the objective function value of the l-th iteration. We first claim that the following inequalities hold:

$$\begin{aligned} \mathcal{L}(\mathbf{w}_l, \mathbf{S}_l, \mathbf{P}_l, \mathbf{F}_l) &\geq \mathcal{L}(\mathbf{w}_l, \mathbf{S}_{l+1}, \mathbf{P}_l, \mathbf{F}_l) \\ &\geq \mathcal{L}(\mathbf{w}_l, \mathbf{S}_{l+1}, \mathbf{P}_{l+1}, \mathbf{F}_l) \\ &\geq \mathcal{L}(\mathbf{w}_l, \mathbf{S}_{l+1}, \mathbf{P}_{l+1}, \mathbf{F}_{l+1}) \\ &\gtrapprox \mathcal{L}(\mathbf{w}_{l+1}, \mathbf{S}_{l+1}, \mathbf{P}_{l+1}, \mathbf{F}_{l+1}) \end{aligned} \tag{S.26}$$

First, $\mathbf{S}_{l+1} = \arg\min \mathcal{L}(\mathbf{w}_l, \mathbf{S}, \mathbf{P}_l, \mathbf{F}_l)$ and the subproblem (24) of $\mathbf{S}$ is convex. Therefore, we



(a) NMI results on Urban

(b) ACC results on Urban

(c) NMI results on lung_discrete

(d) ACC results on lung_discrete

(e) NMI results on Carcinoma

(f) ACC results on Carcinoma

Figure S.5: NMI and ACC results of different modules on the Urban, lung_discrete and Carcinoma data sets.

state

$$\mathcal{L}(\mathbf{w}_l, \mathbf{S}_l, \mathbf{P}_l, \mathbf{F}_l) \geq \mathcal{L}(\mathbf{w}_l, \mathbf{S}_{l+1}, \mathbf{P}_l, \mathbf{F}_l). \tag{S.27}$$

Second, the subproblem (25) of $\mathbf{P}$ is convex and it has the optimal solution. Then, we have

$$\mathcal{L}(\mathbf{w}_l, \mathbf{S}_{l+1}, \mathbf{P}_l, \mathbf{F}_l) \geq \mathcal{L}(\mathbf{w}_l, \mathbf{S}_{l+1}, \mathbf{P}_{l+1}, \mathbf{F}_l). \tag{S.28}$$

Third, $\mathbf{F}^{l+1} = \arg\min_{\mathbf{F},\mathbf{F}^T\mathbf{F}=\mathbf{I}} Tr(\mathbf{F}^T(\lambda_f \mathbf{L}_\mathbf{S} + \lambda_p \mathbf{\Delta})\mathbf{F})$ and the subproblem (30) of $\mathbf{F}$ is convex.

We obtain the following inequality

$$Tr((\mathbf{F}^l)^T(\lambda_f \mathbf{L_S} + \lambda_p \mathbf{\Delta})\mathbf{F}^l) \geq Tr((\mathbf{F}^{l+1})^T(\lambda_f \mathbf{L_S} + \lambda_p \mathbf{\Delta})\mathbf{F}^{l+1}). \tag{S.29}$$

Hence, we have the following inequality

$$\mathcal{L}(\mathbf{w}_l, \mathbf{S}_{l+1}, \mathbf{P}_{l+1}, \mathbf{F}_l) \geq \mathcal{L}(\mathbf{w}_l, \mathbf{S}_{l+1}, \mathbf{P}_{l+1}, \mathbf{F}_{l+1}). \tag{S.30}$$

We can explicitly give the closed-form solutions for these three sub-problems corresponding to **S**, **P**, and **F**, so their objective functions are decreasing in one iteration. However, problem (17) is challenging to solve directly due to its exponential term and equality constraint. Hence, we convert problem (17) to problem (19) and divide it into d independently convex suboptimization problems and solve them separately to obtain the near-optimal solution, where d is the number of features. Since these independent subproblems have closed-form solutions, the optimal solution of the problem (19) can be obtained approximately. Therefore, we give the approximate convergent form as follows

$$\mathcal{L}(\mathbf{w}_l, \mathbf{S}_{l+1}, \mathbf{P}_{l+1}, \mathbf{F}_{l+1}) \gtrapprox \mathcal{L}(\mathbf{w}_{l+1}, \mathbf{S}_{l+1}, \mathbf{P}_{l+1}, \mathbf{F}_{l+1}). \tag{S.31}$$

Finally, we have the following inequalities hold

$$\mathcal{L}(\mathbf{w}_l, \mathbf{S}_l, \mathbf{P}_l, \mathbf{F}_l) \gtrapprox \mathcal{L}(\mathbf{w}_{l+1}, \mathbf{S}_{l+1}, \mathbf{P}_{l+1}, \mathbf{F}_{l+1}). \tag{S.32}$$

Figures S.6 shows the convergence curves of Algorithm 1 on the Yale data sets under different q. The results show rapid convergence of the algorithm. The non-smooth curves are due to the approximate optimization of **w**, but this does not affect overall convergence.

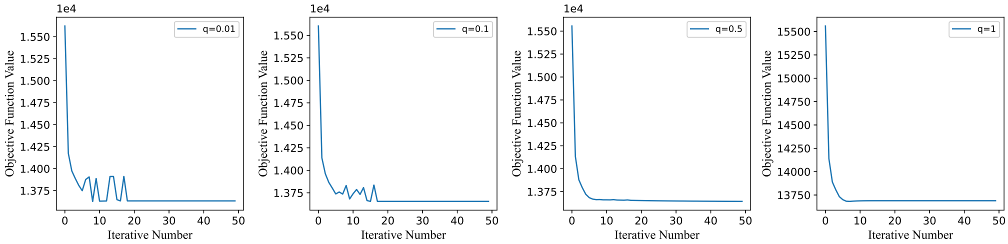


Figure S.6: Convergence speed of LLSRFS with different q on the Yale data set.

G. Time cost analysis

To validate the effectiveness of our method, we test the running time comparison between the comparative method and LLSRFS using public datasets. The experimental results are presented in Table S.4. On average, LLSRFS outperforms four comparison methods: UDFS, FSASL, DGUFS, and SOGFS. Compared with FSASL and SOGFS, which incorporate local structure learning, LLSRFS achieves a speed improvement of 2.27× times and 4.76× times, respectively. Furthermore, compared to UDFS, which preserves the fixed graph structure, LLSRFS demonstrates a substantial speed improvement of 7.43× times.

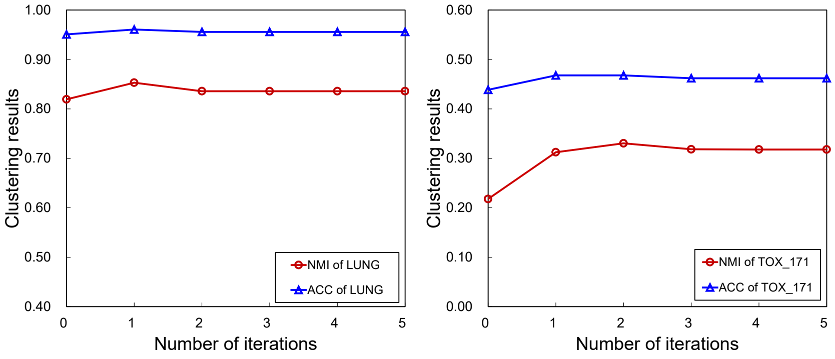
Table S.4: Experimental results on time cost (s).

Datasets	LapScore (ms)	MCFS	UDFS	NDFS	MFFS	FSASL	SOGFS	SPCAFS	SF2S	LLSRFS
Yale	1.84	0.24	12.13	3.92	0.44	6.88	17.34	0.17	0.42	2.13
LUNG	5.58	0.14	210.95	11.80	2.41	105.13	298.01	60.93	6.01	6.23
lung_discrete	2.29	0.17	0.58	0.26	0.06	0.82	1.79	0.02	0.08	0.27
COIL20	23.31	1.34	16.59	2.04	0.48	29.76	18.24	2.59	0.39	0.12
ORL	6.15	0.85	2.12	8.46	0.60	8.50	17.52	0.17	0.37	9.39
Carcinoma	14.25	0.70	6431.08	111.20	16.75	1706.39	4112.97	868.10	89.11	46.21
lymphoma	3.95	0.21	439.62	39.47	6.23	180.44	679.35	6.80	12.55	7.82
TOX_171	6.78	0.15	1887.09	38.92	6.60	480.57	585.50	19.90	19.17	20.76
Isolet	18.81	1.34	6.62	1.41	0.23	31.16	13.87	0.12	0.19	0.13
mfeat	68.28	1.41	17.62	2.07	0.30	51.08	22.24	0.09	0.23	322.00
smartphone	63.87	3.77	76.92	4.13	0.36	179.46	61.13	0.37	0.49	782.73
Urban	3.62	0.18	0.62	0.24	0.04	4.22	1.96	0.01	0.04	25.69
Average	18.23	0.87	758.49	18.66	2.87	232.03	485.83	79.94	10.75	101.96

H. Evaluation category discriminant matrix F

Here, we monitored the category information contained in the matrix **F** at each iteration of our algorithm. Specifically, we performed K-means clustering on the matrix **F** obtained at each iteration and analyzed the variation of the NMI and ACC performance as iterations progress, as shown in Figure S.7. The initial iteration, represented as iteration number 0 in the figure, corresponds to the clustering result of pseudo-labels learned from the initial category discriminant matrix.

Our findings indicate that the clustering quality improves after optimization, demonstrating that the optimized **F** can capture more reliable category discriminative information from the learned graph structure. For instance, the NMI and ACC results on the LUNG dataset initially stand at 0.8193 and 0.9507, respectively, and are improve to 0.8358 and 0.9557 at the 5th iteration. Similarly, the NMI performance shows an increase from 0.2178 to 0.3179 on the TOX_171 dataset and the ACC performance rises from 0.4386 to 0.462.

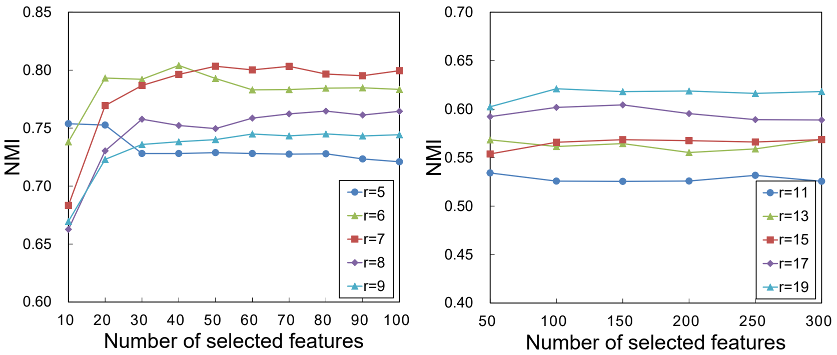


(a) LUNG (b) TOX_171

Figure S.7: NMI and ACC results for category discriminant matrix **F**.

I. Additional clustering experiments

In this evaluation, we investigate the effects of the number of clusters on the performance of LLSRFS. We conduct experiments on two datasets: lung_discrete and Yale. We systematically analyze a range of cluster numbers, specifically 5 to 9 for the lung_discrete dataset and 11 to 19 for the Yale dataset. These results are visually represented in Figure S.8. Optimal clustering solutions closely align with the actual number of underlying categories on the lung_discrete dataset for both clustering methods. In contrast, increasing the number of clusters beyond the default setting for the Yale dataset leads to noticeable improvements in clustering performance.

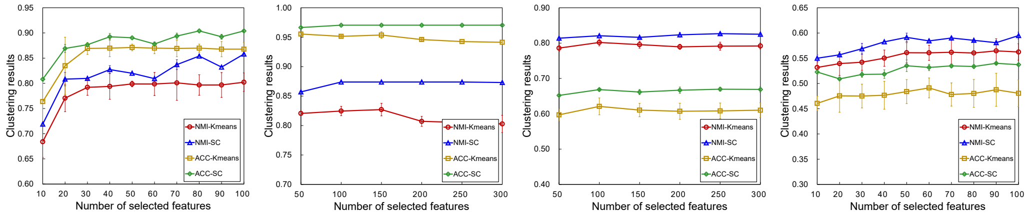


(a) lung_discrete

(b) Yale

Figure S.8: NMI results of LLSRFS under K-means clustering with different clusters.

We employ K-means clustering and Spectral Clustering techniques to comprehensively evaluate the algorithm's performance, as illustrated in Figure S.9. Spectral clustering outperforms K-means clustering in the case of the lung_discrete and LUNG datasets. Conversely, for the ORL and Yale datasets, K-means clustering yields better results. In conclusion, our comprehensive evaluation of the algorithm's performance, employing K-means and Spectral Clustering techniques, reveals distinct outcomes for different datasets.



(a) lung_discrete

(b) LUNG

(c) ORL

(d) Yale

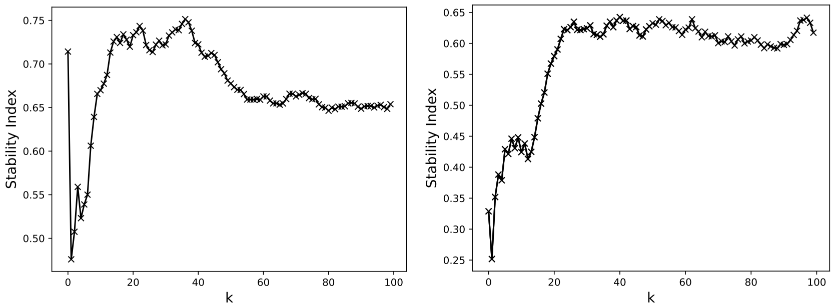
Figure S.9: NMI and ACC results of LLSRFS under Spectral Clustering (SC) and K-means clustering (Kmeans)

J. Stability of LLSRFS

Here, we verified the differences in the selected feature subsets under different parameters q. We choose the Consistency index as an indicator to measure feature stability [S1]. Figure S.10 shows the change curve of the stability index with the number of selected features. When the number of selected features is small, the consistency of different parameters q is low. This is due to the adjustment effect of the index q on features with large weights. As the number of feature subsets increases, the stability gradually increases.

K. Additional Feature visualization results

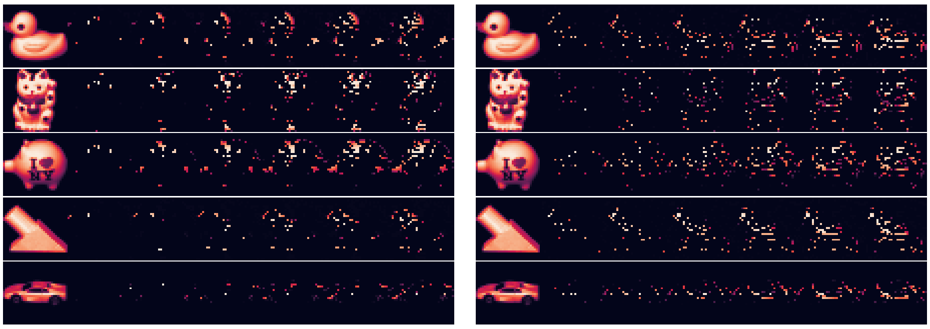
We provided additional quantitative analysis of feature visualization in this section. We selected NDFS as a representative for analysis.



(a) LUNG

(b) urban

Figure S.10: Stable index for a set of sequences of features with different q values.



(a) $F_L - F_N$

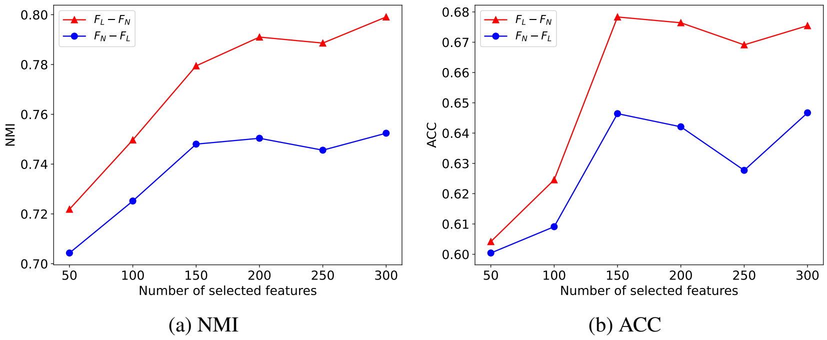
(b) $F_N - F_L$

Figure S.11: Visualization of feature differences. $F_L - F_N$ indicates unique features in LLSRFS, and $F_N - F_L$ indicates unique features in NDFS.

We collected the feature subsets respectively selected by LLSRFS and NDFS, denoted as F_L and F_N. To illustrate the disparity between F_L and F_N, we visualized the features in the sets $F_L - F_N$ and $F_N - F_L$ in Figure S.11 (a) and (b). In other words, the features in the former set are uniquely selected by LLSRFS, and those in the latter set are solely selected by NDFS. As depicted in Figure S.11(a), LLSRFS favors the features discriminative to some distinct regions, such as the head of the duck, the face of the cat, and the mouth and ears of the pig. In contrast, the features selected by NDFS pose attention to the regions less helpful for recognition, such as the bodies of pigs.

To quantitatively assess these distinctions, we conducted further evaluations using NMI and ACC as metrics, as depicted in Figure S.12. We implemented K-means clustering to categorize the samples in the subspaces constructed by the features in the sets $F_L - F_N$ and $F_N - F_L$, respectively. We repeated 20 times and recorded the mean results. The results in Figure S.12 demonstrate that the feature subspaces containing the unique features of LLSRFS exhibit superior clustering performance. These results further emphasize the advantages of LLSRFS.

We employed Fisher Score [S2] to assess the importance of the selected features in terms of class separability. A higher Fisher Score indicates better class separability of the feature subset. As depicted in Figure S.13(a), LLSRFS significantly outperforms NDFS under Fisher Score, revealing the strong discriminative capability of the features selected by LLSRFS. In addition, we

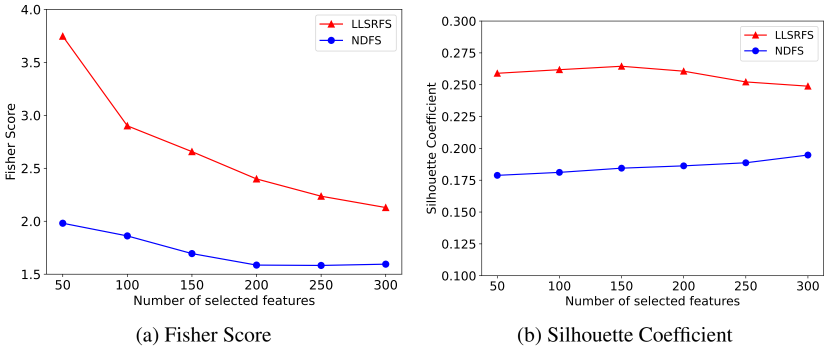


(a) NMI

(b) ACC

Figure S.12: NMI and ACC results of $F_L - F_N$ and $F_N - F_L$.

utilized Silhouette Coefficient [S3] with real labels to assess data dispersion in feature subspace. A higher Silhouette Coefficient indicates better separation of samples between the classes. As shown in Figure S.13(b), LLSRFS consistently outperforms NDFS under this metric, further emphasizing its superiority.



(a) Fisher Score

(b) Silhouette Coefficient

Figure S.13: Fisher Score and Silhouette Coefficient of the selected feature subsets.

References

[S1] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65.

[S2] Z. Zhao, L. Wang, H. Liu, J. Ye, On similarity preserving feature selection, IEEE Transactions on Knowledge and Data Engineering 25 (3) (2011) 619–632.

[S3] L. I. Kuncheva, A stability index for feature selection, in: Artificial Intelligence and Applications, AIAP'07, 2007, pp. 390–395.