Maximizing Schatten-p Norm Regularization Toward Balance

Fangfang Li[1], Quanxue Gao[1*], Yapeng Wang[2], Yu Duan[1], Yuzhuo Feng[3], Qin Li[4*]

[1] School of Telecommunications Engineering, Xidian University, Shaanxi 710071, China
[2] School of Mathematics and Statistics, Xidian University, Shaanxi 710071, China
[3] School of Mechano-Electronic Engineering, Xidian University, Shaanxi 710071, China
[4] School of Software Engineering, Shenzhen University of Information Technology, China
22011110201@stu.xidian.edu.cn, qxgao@xidian.edu.cn, ypwang0121@stu.xidian.edu.cn, duanyuee@gmail.com,
22049200389@stu.xidian.edu.cn, liqin@sziit.edu.cn

Abstract

The Schatten-p norm, as a class of structure-inducing norms based on singular values, has been widely used to enhance model low-rankness and representation capability due to its flexibility in structural modeling and favorable mathematical properties. However, its potential in cluster distribution modeling has long been overlooked. Therefore, we explore the potential of maximizing the Schatten-p norm as a regularization strategy specifically designed to achieve balanced clustering. This work is the first to investigate its effectiveness in promoting cluster balance. To be specific, maximizing Schatten-p norm effectively guides the assignment of data points, ensuring a more balanced distribution of samples across clusters. We have conducted an in-depth theoretical analysis and validated its effectiveness through extensive clustering experiments. Experimental results demonstrate that, compared to existing methods, this regularization term significantly improves clustering quality and obtain reasonable clustering.

Introduction

Regularization plays a crucial role in machine learning and pattern recognition. It not only effectively prevents model overfitting and improves generalization ability but also guides the model in learning more structured and interpretable parameters during optimization. Regularization techniques are widely applied across various tasks, including image processing(Gu et al. 2014), classification(Deng et al. 2024; Sun, Shen, and Sun 2023), regression(Xie et al. 2017; Wang et al. 2024), and clustering(Xia et al. 2023; Gao et al. 2025), to enhance model performance and robustness.

For example, Lasso regularization enables feature selection by forcing some feature weights to shrink to zero, thereby eliminating irrelevant variables and enhancing model interpretability. Ridge regularization prevents excessive parameter growth by imposing a Gaussian prior, improving model smoothness and stability. Elastic Net combines the advantages of L_1 and L_2 regularization, ensuring the model maintains stable generalization capability even in high-dimensional settings. The $L_{2,1}$ norm(Tang et al. 2018; Sun, Shen, and Sun 2023) is a widely used regularization technique that effectively leverages the intrinsic structure of data, making it a popular measurement standard for enhancing the robustness of feature extraction methods.

In matrix optimization problems, regularization is equally critical. The Frobenius norm is commonly used to measure the overall size of a matrix, and its regularization can help reduce the magnitude of parameters, preventing overfitting due to excessive model complexity. However, the Frobenius norm cannot directly control the rank of a matrix, making it difficult to enforce low-rank structures. In contrast, the nuclear norm, as a convex relaxation of matrix rank, effectively encourages low-rank structures, making it widely applicable in self-representation learning(Li et al. 2025b) and matrix completion(Yang, Li, and Wang 2022) tasks.

Additionally, minimizing the Schatten-p norm(Xie et al. 2016a; Gao et al. 2021; Xia et al. 2023) provides flexible control over singular value sparsity by adjusting the p value, offering diverse regularization strategies for different optimization problems. For instance, Xie et al. (Xie et al. 2016a) introduced weighted Schatten-p minimization to emphasize informative components, improving structural preservation, with applications such as hyperspectral image restoration (Xie et al. 2016b). Later, Zhang et al. (Zhang et al. 2019) integrated the Schatten-p norm into low-rank representation (LRR) models and proposed two improved variants tailored for subspace clustering, yielding better expressive power and efficiency. More recently, in (Zhang et al. 2024), they further refined the low-rank modeling process by introducing an optimal mean-removal constraint and factorized Schatten-p regularization, leading to more accurate rank approximation and improved clustering results.

However, existing studies have primarily focused on the role of the Schatten-p norm in enhancing model expressiveness and promoting low-rank representations, while its potential in characterizing the distribution of cluster data has largely been overlooked. In clustering tasks, achieving balanced cluster assignments is critical for producing reliable and interpretable results. In this work, we explore a novel perspective on Schatten-p norm regularization. Contrary to the common practice of minimizing the Schatten-p norm for low-rank modeling, we present a surprising and novel finding: maximizing the Schatten-p norm can effectively encourage balanced cluster.

*Corresponding authors.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The main contributions of this paper are summarized as follows:

- **Balance Regularization**: We surprisingly and innovatively find that **maximizing the Schatten-p norm** regularization term for $1 \leqslant p < 2$ can naturally lead to balanced sample distributions across clusters, without explicitly enforcing any balancing constraints.
- **Efficient Optimization**: We develop an efficient optimization strategy by applying a first-order Taylor expansion to handle the Schatten-p norm, making the problem more tractable.
- **Theory and Verification**: We provide theoretical analysis to prove the balance property of maximizing the Schatten-p norm, and perform clustering experiments on multiple datasets to validate its effectiveness in practice.

Related Work

Balance

Exclusive Lasso(Liu et al. 2017; Han, Liu, and Nie 2019; Chen et al. 2025) regularization term aims to measure the similarity between the columns of matrix $\mathbf{H}$ because $\mathbf{1}\mathbf{1}^T$ is a matrix of all ones, which weights the columns of H. The $tr(\mathbf{H}^T\mathbf{1}\mathbf{1}^T\mathbf{H})$ is essentially the sum of the weighted inner products between the columns of $\mathbf{H}$, thus encouraging similarity between them.

$$\Omega(\mathbf{H}) = tr(\mathbf{H}^T\mathbf{1}\mathbf{1}^T\mathbf{H}) \tag{1}$$

If the columns of $\mathbf{H}$ are nearly identical, or if the inner products between the columns are large, the inner product in $\mathbf{H}^T\mathbf{1}\mathbf{1}^T\mathbf{H}$ will be large, thus minimizing the trace. In this case, the model loses its ability to distinguish and classify effectively because each data point might be assigned the same probability of belonging to all classes.

Maximizing the Frobenius norm of the label matrix can be considered as a way to encourage cluster balance.

$$\Omega(\mathbf{H}) = \max_{\mathbf{H}\geq 0, \mathbf{H}\mathbf{1}=\mathbf{1}} \|\mathbf{H}\|_F^2 \tag{2}$$

However, when all elements in a certain column of $\mathbf{H}$ are 1, Eq. (2) reaches its maximum value. This indicates that all samples are assigned to a single cluster while others remain empty. This extreme assignment corresponds to a trivial clustering solution.

Schatten-p Norm

Since the rank of a matrix is a non-convex and non-differentiable function, directly optimizing rank-constrained problems is challenging. To address this issue, the nuclear norm is introduced as a convex relaxation of the rank function(Lu et al. 2016). It is defined as:

$$\min \|\mathbf{H}\|_* = \min \sum_i \sigma_i(\mathbf{H}) \tag{3}$$

where $\sigma_i(\mathbf{H})$ denote the i-th largest singular value of $\mathbf{H}$. By minimizing the sum of singular values, the nuclear norm encourages a low-rank structure.

The Schatten-p norm minimization ensures that the rank is more easily approximated to the rank of the target matrix(Li et al. 2025a).

$$\min \|\mathbf{H}\|_{sp}^p = \min (\sum_i \sigma_i^p(\mathbf{H})) \tag{4}$$

Different values of p provide different forms of low-rank constraints. For example, $p \to 0$ which is close to the rank of the matrix. The nuclear norm ($p = 1$) enforces more singular values to shrink towards zero , while the Frobenius norm ($p = 2$) applies a uniform constraint on all singular values. In low-rank optimization problems, the Schatten-p norm typically takes values in the range $0 < p \leq 1$, as it can more effectively promote sparsity and low-rank structures. Minimizing the Schatten-p norm has been widely used to enforce low-rank structures, but its potential for promoting class balance has not been explored in existing studies.

Methodology

Motivation

We explore the potential of maximizing the Schatten-p norm in achieving cluster balance. To support this, we introduce Theorems 1 and Theorems 2, which theoretically demonstrate the effectiveness of maximizing the Schatten-p in promoting balanced clustering.

Theorem 1 *Given matrix $\mathbf{H} \in \mathbb{R}^{N\times c}$, let $\sigma_j(\mathbf{H})$ denote the j-th largest singular value of $\mathbf{H}$, $\tau_j(\mathbf{H}^T\mathbf{H})$ denote the j-th largest eigenvalue of $\mathbf{H}^T\mathbf{H}$. $\delta_j = \tau_j(\mathbf{H}^T\mathbf{H})$, then*

$$\max_{\mathbf{H}\geq 0, \mathbf{H}\mathbf{1}=\mathbf{1}} \|\mathbf{H}\|_{sp}^p \Rightarrow \max_{\substack{\mathbf{H}\geq 0, \mathbf{H}\mathbf{1}=\mathbf{1}, \\ \delta_1=\ldots=\delta_c}} \|\mathbf{H}\|_F^2 \tag{5}$$

where $0 < p < 2$.

Proof 1 *The Schatten-p norm is defined as*

$$\|\mathbf{H}\|_{sp}^p = \sum_{j=1}^{c} \sigma_j^p(\mathbf{H}) = \sum_{j=1}^{c} (\tau_j(\mathbf{H}^T\mathbf{H}))^{\frac{p}{2}} = \sum_{j=1}^{c} {\delta_j}^{\frac{p}{2}} \tag{6}$$

*Let $\boldsymbol{\delta} = [\delta_1, \delta_2, \ldots, \delta_c]^T \in \mathbb{R}^{c\times 1}$, $\beta_1 = \ldots = \beta_c = \frac{1}{c}$. When $0 < p < 2$, $f(\delta_j) = {\delta_j}^{\frac{p}{2}}$ is a **concave function** with respest to δ_j, then according to Jensen inequality, we have*

$$f\left(\sum_{j=1}^{c} \beta_j \delta_j\right) \geqslant \sum_{j=1}^{c} \beta_j f(\delta_j) \tag{7}$$

Equality holds if and only if $\delta_1 = \delta_2 = \ldots = \delta_c$.

By simplifying the right side of the inequality(7), we obtain

$$\sum_{j=1}^{c} \beta_j f(\delta_j) = \frac{1}{c}\sum_{j=1}^{c} f(\delta_j) = \frac{1}{c}\|\mathbf{H}\|_{sp}^p \tag{8}$$

Similarly, by simplifying the left side of the inequality(7), we obtain

$$\sum_{j=1}^{c} \beta_j \delta_j = \sum_{j=1}^{c} \beta_j \tau_j(\mathbf{H}^T\mathbf{H}) = \frac{1}{c}\|\mathbf{H}\|_F^2 \tag{9}$$

Further,

$$f(\sum_{j=1}^{c}\beta_j\delta_j) = f(\frac{1}{c}\|\mathbf{H}\|_F^2) = (\frac{1}{c}\|\mathbf{H}\|_F^2)^{\frac{p}{2}} \quad (10)$$

Therefore, by combining Eq.(8) and Eq.(10), we can convert the inequality(7) into

$$c(\frac{1}{c}\|\mathbf{H}\|_F^2)^{\frac{p}{2}} \geqslant \|\mathbf{H}\|_{sp}^p \quad (11)$$

Equality holds if and only if $\delta_1 = \delta_2 = \ldots = \delta_c$. In order to find the maximum of $\|\mathbf{H}\|_{sp}^p$, we can translate to finding the maximum of $\|\mathbf{H}\|_F^2$ under the constraint $\delta_1 = \ldots = \delta_c$. □

Theorem 2 *Given $n_1 + n_2 + \ldots + n_c = N$, maximizing Eq.(12) ensures a balanced cluster distribution, i.e., $n_1 = n_2 = \ldots = n_c$.*

$$\max_{\mathbf{H}} \|\mathbf{H}\|_{sp}^p \quad s.t. \mathbf{H} \geq 0, \mathbf{H}\mathbf{1} = \mathbf{1} \quad (12)$$

where $0 < p < 2$.

Proof 2 *According to Theorem 1, we can transform Eq.(12) to into the following form under the constraint $\delta_1 = \ldots = \delta_c$.*

$$\begin{aligned} &\max_{\mathbf{H}} \|\mathbf{H}\|_F^2 = \max \sum_{ij} h_{ij}^2 = \max \sum_i \sum_j h_{ij}^2 \\ &s.t. \mathbf{H} \geq 0, \mathbf{H}\mathbf{1} = \mathbf{1} \end{aligned} \quad (13)$$

In Eq.(13), each row of $\mathbf{H}$ is independent, so for each row of $\mathbf{H}$, Eq.(13) becomes

$$\max \sum_{j=1}^{c} h_{ij}^2 \quad s.t. h_{ij} \geq 0, \sum_j h_{ij} = \mathbf{1}, \delta_1 = \ldots = \delta_c \quad (14)$$

The solution to the maximization problem (14) should be realized when $\mathbf{h}_i$ has only one element equal to 1 and the rest are 0, and the maximum value should be 1.

Thus, under the constraint $\delta_1 = \delta_2 = \ldots = \delta_c$, we can conclude that the problem $(\|\mathbf{H}\|_F^2)^{\frac{p}{2}}$ only reaches its maximum when $\mathbf{H}$ is a discrete matrix. In this case, $\mathbf{H}^T\mathbf{H}$ is a diagonal matrix, the value of the j-th diagonal element is the eigenvalue δ_j, which is also the sample number of the j-th cluster n_j.

Combined with Eq. (7), we have

$$f\left(\sum_{j=1}^{c}\frac{1}{c}\delta_j\right) = f(\frac{1}{c}\sum_{j=1}^{c}\delta_j) = f(\frac{1}{c}\sum_{j=1}^{c}n_j) = f(\frac{N}{c}) \quad (15)$$

So we know that for inequality (7), the equation holds if and only if $\delta_1 = \delta_2 = \ldots = \delta_c = n_1 = n_2 = \ldots = n_c = \frac{N}{c}$, then $\|\mathbf{H}\|_{sp}^p$ reaches its maximum. □

Objective

Graph-based clustering methods are capable of flexibly modeling similarity relationships between data points, making them well-suited for datasets with arbitrary distributions. However, conventional graph clustering often suffers from high computational complexity and memory overhead due to the need to construct full sample graphs, which limits their scalability. To address this, anchor graph structures have been introduced(Zhang, Nie, and Li 2023; Nie et al. 2024). By selecting a small number of anchor points to approximate the relationships among original samples, anchor graphs significantly reduce the cost of graph construction. Therefore, our model adopts an anchor graph-based clustering framework.

Given an anchor graph $\mathbf{S} \in \mathbb{R}^{N\times m}$, which captures the relationship between the data matrix $\mathbf{X} \in \mathbb{R}^{N\times d}$ and the anchor matrix (Xia et al. 2023), where N and $m \ll N$ represent the number of data samples and anchors respectively, and d is the feature dimension. We propose a probability-based anchor graph clustering method, as illustrated in Figure 1. Anchors are connected to samples via the anchor graph $\mathbf{S} \in \mathbb{R}^{N\times m}$, and label information is represented by $\mathbf{H} \in \mathbb{R}^{N\times c}$ for samples and $\mathbf{S}^T\mathbf{H}$ for anchors. During the optimization process, anchor labels provide guidance for the learning of sample labels, while updated sample labels in turn promote the adjustment of anchor labels. Based on this, our model is

$$\max_{\mathbf{H}} \left\|\mathbf{S}^T\mathbf{H}\right\|_{sp}^p \quad s.t. \mathbf{H} \geq 0, \mathbf{H}\mathbf{1} = \mathbf{1} \quad (16)$$

where $1 \leqslant p < 2$.

Optimization

$\|\mathbf{S}^T\mathbf{H}\|_{sp}^p$ with $1 \leqslant p < 2$ is a **convex function**, thus according to Theorem3, We relax model (16) with its first-order Taylor expansion as:

$$\max_{\mathbf{F},\mathbf{H}} tr(\mathbf{F}^T\mathbf{S}^T\mathbf{H}) \quad s.t. \mathbf{H} \geq 0, \mathbf{H}\mathbf{1} = \mathbf{1} \quad (17)$$

where $\mathbf{F} = \frac{\partial\left\|\mathbf{S}^T\mathbf{H}\right\|_{sp}^p}{\partial(\mathbf{S}^T\mathbf{H})}$.

Theorem 3 *For optimization problems:*

$$\max f(\mathbf{Z}) \quad (18)$$

where $f(\mathbf{Z})$ is differentiable function of matrix $\mathbf{Z}$. Relaxing $f(\mathbf{Z})$ using its first-order Taylor expansion at the current iteration point $\mathbf{Z}_k$, leading to the modified optimization problem:

$$\max \mathrm{tr}(\nabla f(\mathbf{Z}_k)^T\mathbf{Z}) \quad (19)$$

*then a necessary and sufficient condition for this relaxation to be valid is that $f(\mathbf{Z})$ is a **convex function**.*

Proof 3 *The first-order Taylor expansion of $f(\mathbf{Z})$ at $\mathbf{Z}_k$ is given by*

$$\begin{aligned} F(\mathbf{Z}, \mathbf{Z}_k) &= f(\mathbf{Z}_k) + <\nabla f(\mathbf{Z}_k), \mathbf{Z} - \mathbf{Z}_k> \\ &= f(\mathbf{Z}_k) + tr(\nabla f(\mathbf{Z}_k)^T(\mathbf{Z} - \mathbf{Z}_k)) \end{aligned} \quad (20)$$

Necessity: *When $f(\mathbf{Z})$ is to be approximated by its first-order Taylor expansion, it needs to satisfy the lower bound that the first-order Taylor expansion is $f(\mathbf{Z})$, i.e.*

$$f(\mathbf{Z}) \geq f(\mathbf{Z}_k) + tr(\nabla f(\mathbf{Z}_k)^T(\mathbf{Z} - \mathbf{Z}_k)) \quad (21)$$

*The inequality (21) is exactly **the first order condition for convex function**. Thus, $f(\mathbf{Z})$ must be a **convex function**.*

***Sufficiency**: If $f(\mathbf{Z})$ is a convex function, then for any $\mathbf{Z}$, by the properties of the convex function,*

$$f(\mathbf{Z}) \geq tr(\nabla f(\mathbf{Z}_k)^T \mathbf{Z}) + C \tag{22}$$

where $C = f(\mathbf{Z}_k) - tr(\nabla f(\mathbf{Z}_k)^T \mathbf{Z}_k)$ is a constant. Thus, we can replace $f(\mathbf{Z})$ with its lower bound $tr(\nabla f(\mathbf{Z}_k)^T \mathbf{Z}) + C$. Ignoring the constant term , the optimization problem reduces to

$$\max tr(\nabla f(\mathbf{Z}_k)^T \mathbf{Z}) \tag{23}$$

Therefore, a sufficient and necessary condition for approximating $f(\mathbf{Z})$ with a first-order Taylor expansion is that $f(\mathbf{Z})$ is a convex function. □

• **Update F , while keeping H**: To simplify optimization process, we set $\mathbf{Z} = \mathbf{S}^T\mathbf{H}$, $f(\mathbf{Z}) = \|\mathbf{Z}\|_{sp}^p$. We can directly get

$$\mathbf{F} = \frac{\partial \|\mathbf{Z}\|_{sp}^p}{\partial \mathbf{Z}} = \frac{\partial \left\|\mathbf{S}^T\mathbf{H}\right\|_{sp}^p}{\partial (\mathbf{S}^T\mathbf{H})} = p\mathbf{U}\mathbf{\Sigma}^{-1}|\mathbf{\Sigma}|^p\mathbf{V}^{\mathrm{T}} \tag{24}$$

where $\mathbf{Z} = \mathbf{S}^T\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, $\mathbf{\Sigma}^{-1}$is the Moore-Penrose pseudo-inverse of $\mathbf{\Sigma}$, $|\cdot|$ is the absolute value.

• **Update H , while keeping F**:

$$\underset{\mathbf{H}}{\operatorname{argmax}}\, tr(\mathbf{F}^T\mathbf{S}^T\mathbf{H}) = \underset{\mathbf{H}}{\operatorname{argmax}} \sum_{j=1}^{n} h^j a_j \tag{25}$$

where $\mathbf{H} \geq 0, \mathbf{H1} = \mathbf{1}$, each row $h^j \in \mathbb{R}^c$ of the matrix $\mathbf{H}$ represents a probability distribution, i.e., a non-negative vector whose entries sum to one. $\mathbf{A} = \mathbf{F}^\top\mathbf{S}^\top \in \mathbb{R}^{c\times N}$, where $a_j \in \mathbb{R}^c$ denotes the j-th column of matrix $\mathbf{A}$. This reformulation allows the original problem to be decomposed into n independent subproblems of the form:

$$\max_{h^j} h^j a_j, \quad \text{s.t.} \quad h_j \geq 0, \quad h^j\mathbf{1} = 1 \tag{26}$$

Each subproblem is a linear maximization over the probability simplex. The optimal solution is:

$$h^j = \mathbf{e}_b, \quad \text{where} \quad b = \arg\max_i (a_j)_i \tag{27}$$

i.e., h^j is a one-hot vector with 1 at the position corresponding to the maximum entry of a_j, and 0 elsewhere.

As a result, the optimal solution $\mathbf{H}$ is given by:

$$\mathbf{H}_{ij} = \begin{cases} 1 & \text{if } i = \underset{b}{\operatorname{argmax}}(\mathbf{F}^T\mathbf{S}^T)_{bj} \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

Convergence Analysis

We set $\mathbf{Z} = \mathbf{S}^T\mathbf{H}$, $\operatorname{argmax} f(\mathbf{Z}) = \operatorname{argmax} \|\mathbf{Z}\|_{sp}^p = \operatorname{argmin} -\|\mathbf{Z}\|_{sp}^p = \operatorname{argmin} g(\mathbf{Z})$, $1 \leqslant p < 2$. Since $g(\mathbf{Z}) = -\|\mathbf{Z}\|_{sp}^p$ is a concave function, we can approximate it at the point $\mathbf{Z}_k$ using its first-order Taylor expansion:

$$G(\mathbf{Z}, \mathbf{Z}_k) = g(\mathbf{Z}_k) + \mathrm{tr}(\nabla g(\mathbf{Z}_k)^T(\mathbf{Z} - \mathbf{Z}_k)) \tag{29}$$

Since $g(\mathbf{Z})$ is concave, it satisfies the first-order concavity condition:

$$g(\mathbf{Z}) \leq G(\mathbf{Z}, \mathbf{Z}_k) \tag{30}$$

Algorithm 1: solve problem (16)

1: **Input** anchor graph $\mathbf{S} \in \mathbb{R}^{N\times m}$, cluster number c, $1 \leqslant p < 2$.
2: **Initialize** label matrix $\mathbf{H} \in \mathbb{R}^{N\times c}$
3: **repeat**
4: update matrix $\mathbf{F}$ by Eq. (24);
5: update matrix $\mathbf{H}$ by Eq. (28);
6: **until** convergence
7: **Output** $\mathbf{H} \in \mathbb{R}^{N\times c}$

Let $\mathbf{Z}_{k+1}$ be the next iterate, we update $\mathbf{Z}$ as

$$\mathbf{Z}_{k+1} \in \arg\min G(\mathbf{Z}, \mathbf{Z}_k) \tag{31}$$

From the concavity condition, we obtain:

$$g(\mathbf{Z}_{k+1}) \leq G(\mathbf{Z}_{k+1}, \mathbf{Z}_k) \tag{32}$$

By the iterative update rule, our optimization procedure ensures that:

$$\begin{aligned} G(\mathbf{Z}_{k+1}, \mathbf{Z}_k) &\leq G(\mathbf{Z}_k, \mathbf{Z}_k) \\ &= g(\mathbf{Z}_k) \end{aligned} \tag{33}$$

Thus, combining the inequality (32) and inequality (33), we get:

$$g(\mathbf{Z}_{k+1}) \leq G(\mathbf{Z}_{k+1}, \mathbf{Z}_k) \leq g(\mathbf{Z}_k) \tag{34}$$

This confirms that the iteration sequence $\{g(\mathbf{Z}_k)\}_{k=1}^{\infty}$ is monotonically decreasing.

In practical optimization problem, the function $g(\mathbf{Z}) = -\|\mathbf{Z}\|_{sp}^p$ has a lower-bounded, applying the monotone bounded theorem, we conclude that the iteration sequence $\{g(\mathbf{Z}_k)\}_{k=1}^{\infty}$ converges to $\{g(\mathbf{Z}^*)\}$.

Experiments

Experimental Settings

Experiments are conducted on a Windows 10 desktop computer equipped with a 2.40 GHz Intel Xeon Gold 6240R CPU, 64 GB RAM, and MATLAB R2016a (64-bit).

Experimental on Benchmark Datasets

The performance of our method is evaluated on eight benchmark datasets, and the information of size is shown in Table 1. JAFFE (Lyons, Kamachi, and Gyoba 1998), ORL (Samaria and Harter 1994), MSRA25(Liu et al. 2011), FaceV5 [1], AR (Martinez and Benavente 1998), ISOLET[2], USPS (Hull 1994), and Pendigits [3]. In Table 1, we provide a detailed overview of these datasets, covering sample size, feature dimension, number of categories.

To comprehensively evaluate the effectiveness of the proposed anchor graph-based balanced clustering method, we compare it against a diverse set of representative clustering algorithms, including: Instance-based methods: **K-Means**(Hartigan and Wong 1979), **CDKM**(Nie et al.

[1]http://biometrics.idealtest.org/dbDetailForUser.do?id=9
[2]https://archive.ics.uci.edu/dataset/54/isolet
[3]https://odds.cs.stonybrook.edu/pendigits-dataset/

	JAFFE	ORL	MSRA25	Face-V5
Samples	213	400	1,799	2,500
Features	676	1,024	256	256
Classes	10	40	12	500
N_{entro}	0.9996	1	0.9949	0.4553
	AR	ISOLET	USPS	Pendigits
Samples	3,120	7,797	9,298	10,992
Features	2,000	617	256	16
Classes	120	26	10	10
N_{entro}	0.5189	1	0.9852	0.9997

Table 1: The information of datasets.

2022),**K-sum-x** (Pei et al. 2023) Density-based methods: **DBSCAN** (Ester et al. 1996) and **HDBSCAN** (Mcinnes, Healy, and Astels 2017). Graph-based clustering approaches: **K-sum** (Pei et al. 2023), **LCSOG** (Zhang, Nie, and Li 2023), **FCAG** (Nie et al. 2024) Balanced clustering algorithms: **RKM**(Lin, He, and Xiao 2019).

Results We use three commonly used clustering evaluation metrics to assess the performance: Accuracy (ACC), Normalized Mutual Information (NMI), Purity. All metrics range from 0 to 1, with higher values indicating better performance. After conducting experiments across eight datasets, we obtained the corresponding measurement results. These results are presented in a detailed comparison across Table 2. As shown in the table, our method consistently outperforms other approaches in terms of overall clustering performance. Traditional methods such as K-Means and DBSCAN often suffer from issues like cluster collapse or fragmentation, especially in complex or imbalanced data scenarios. In contrast, our method enhances the rationality of cluster assignment by maximizing the proposed schatten-p norm regularization term, which leads to better intra-cluster compactness, fewer misclassifications, and a more balanced distribution of samples across clusters. This ultimately improves the quality of clustering results.

And when comparing different algorithmic frameworks, we observe that anchor graph-based methods such as K-sum, LCSOG, and FCAG generally achieve better performance than instance-based methods like K-Means, RKM, CDKM, and K-sum-x. Our model also adopts an anchor graph structure. Although RKM takes cluster balance into consideration, the experimental results demonstrate that our model achieves significantly better balance and overall performance across multiple datasets. This confirms our maximized schatten-p norm regularization not only has favorable theoretical properties but also yields more natural and meaningful clustering results in practice.

Parameters setting and analysis To further investigate the impact of key hyperparameters, we conducted sensitivity analyses on the anchor rate m and the parameter p in the Schatten-p norm on the JAFFE, ORL, AR and ISOLET datasets.

Effect of anchor rater m: The anchor rate m determines the number of representative samples used to construct the anchor graph, which significantly affects both the graph quality and computational efficiency. As illustrated in Figure 1, we analyze the impact of varying the anchor rate from 0.1 to 1.0. The results show that optimal clustering performance is achieved when $m = 0.3$ for JAFFE, $m = 0.41$ for ORL, $m = 0.44$ for ISOLET, and $m = 0.69$ for AR. Moreover, our method consistently delivers stable and competitive performance across different datasets when the anchor rate varies within a reasonable range, indicating strong robustness to anchor selection.

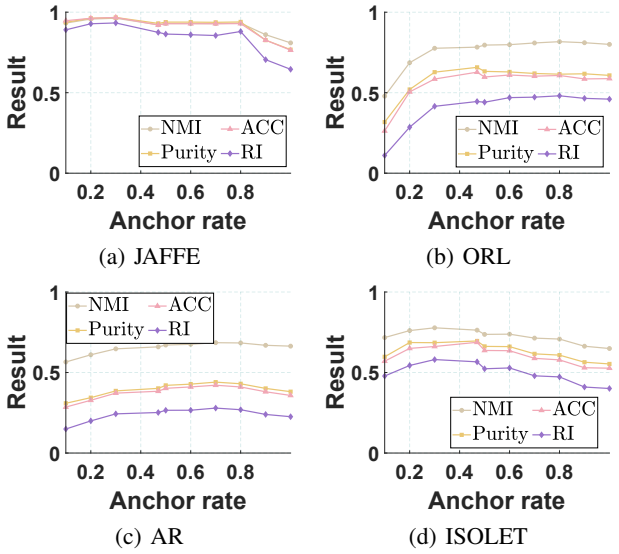


Figure 1: Clustering performance w.r.t. anchor rater m.

Effect of parameter p: The parameter p in the Schatten-p norm determines how the singular values of the matrix are regularized, thereby influencing the ability of the regularization term to promote class balance. We conducted a systematic evaluation of p within the range $[1, 1.9]$, and the results are shown in Figure 2. The experimental results show that the best clustering performance is achieved when $p = 1.3$ for the JAFFE dataset, $p = 1.2$ for ORL, $p = 1.6$ for ISOLET, and $p = 1$ for AR. Our method demonstrates strong robustness across different datasets and parameter settings, further validating the effectiveness and flexibility of the proposed regularization in promoting clustering performance and class balance.

Ablation Experiment

To further verify the effectiveness of the proposed maximized schatten-p norm regularization in achieving class balance, we perform a comparative experiment with a commonly used alternative method $\max_{\mathbf{H}\geq 0, \mathbf{H1}=\mathbf{1}} \left\|\mathbf{S}^T\mathbf{H}\right\|_F^2$.

Normalized entropy (Nentro)(Zhong and Ghosh 2003) is defined as follows:

$$N_{entro} = -\frac{1}{\log c}\sum_{h=1}^{c}\frac{n_h}{N}\log\frac{n_h}{N} \tag{35}$$

Datasets	JAFFE			ORL			MSRA25			Face-V5		
Methods	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
K-means	0.7085	0.8010	0.7455	0.5250	0.7381	0.5800	0.5097	0.5838	0.5451	0.7248	0.9285	0.7768
DBSCAN	0.8263	0.7871	0.8263	0.5000	0.7286	**0.6975**	0.5036	0.5878	0.6198	0.8720	0.9677	0.9480
HDBSCAN	0.8028	0.8236	0.8826	0.5625	0.7843	0.7625	0.5042	0.6246	0.6581	0.9376	0.9616	0.9384
RKM	0.8310	0.8159	0.8310	0.5000	0.7143	0.5200	0.5297	0.5410	0.5297	0.8140	0.9473	0.8124
CDKM	0.7451	0.8246	0.7812	0.5507	0.7529	0.6090	0.5313	0.5930	0.5605	0.8506	0.9639	0.8852
K-sum-x	0.8930	0.9013	0.8977	0.5877	0.7693	0.6060	0.5516	0.6191	0.5772	0.9638	0.9860	0.9662
K-sum	0.8789	0.8764	0.8789	0.6337	0.7940	0.6562	0.5665	0.5793	0.5762	0.9568	0.9860	0.9656
LCSOG	**0.9671**	**0.9623**	**0.9671**	0.6075	0.7839	0.6725	0.6142	0.6778	0.6142	0.9704	0.9879	0.9724
FCAG	0.9311	0.9354	0.9311	0.6142	0.7893	0.6492	0.6040	0.6638	0.6281	0.9339	0.9826	0.9428
OUR	**0.9671**	**0.9623**	**0.9671**	**0.6625**	**0.7999**	**0.6975**	**0.6543**	**0.6955**	**0.6609**	**0.9712**	**0.9917**	**0.9728**
Datasets	AR			ISOLET			USPS			Pendigits		
Methods	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
K-means	0.2514	0.5574	0.2749	0.5469	0.7154	0.5958	0.6458	0.6026	0.7129	0.6615	0.6629	0.6858
DBSCAN	0.2208	0.5455	**0.4619**	0.1681	0.2275	0.1825	0.3523	0.2504	0.3619	0.7492	0.7150	0.7667
HDBSCAN	0.2635	0.3675	0.3404	0.0826	0.0693	0.0872	0.3566	0.2977	0.3761	0.7076	0.7076	0.7807
RKM	0.2641	0.5752	0.3215	0.6299	0.7346	0.6387	0.6241	0.5748	0.7003	0.7296	0.6639	0.7296
CDKM	0.2653	0.5700	0.2862	0.5328	0.7159	0.5837	0.6526	0.6094	0.7237	0.7027	0.6697	0.7226
K-sum-x	0.2454	0.5676	0.3236	0.6094	0.7307	0.6254	0.6502	0.5853	0.7150	0.7768	0.7001	0.7768
K-sum	0.2970	0.5963	0.3686	0.6269	0.7347	0.6402	0.6802	0.6274	0.7486	0.7562	0.6743	0.7562
LCSOG	0.3423	0.4992	0.3817	0.5352	0.6678	0.5529	0.7448	**0.7778**	0.7922	0.8432	**0.7871**	0.8432
FCAG	0.3592	0.6656	0.3787	0.5833	0.7241	0.6078	0.7088	0.6626	0.7482	0.7652	0.7293	0.7817
OUR	**0.4212**	**0.6853**	0.4404	**0.6877**	**0.7631**	**0.6959**	**0.8329**	0.7504	**0.8329**	**0.8480**	0.7798	**0.8480**

Table 2: The clustering performances.

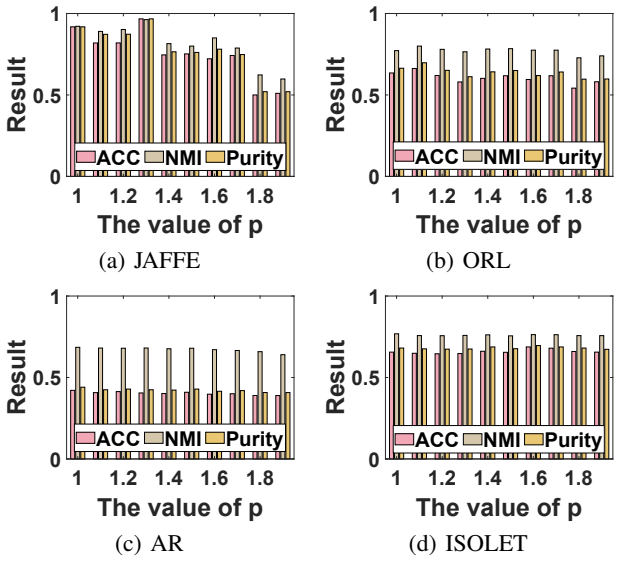


Figure 2: Clustering performance w.r.t. p.

where n_h denotes the size of the h-th cluster. An Nentro of 1 indicates perfectly balanced clusters and 0 indicates extremely unbalanced clusters.

As shown in Table 3, the experimental results clearly indicate that for the ORL and Pendigits balanced datasets, the Frobenius norm–based model suffers from poor clustering accuracy and severe imbalance in sample distribution, often resulting in extremely uneven cluster sizes. In contrast, our method, by maximizing the schatten-p norm regularization, significantly improves clustering performance. For the imbalanced dataset Face-V5, our model is still able to guide the data to cluster in a manner consistent with the distribution of the data itself, rather than forcing strictly balanced clustering. These results strongly demonstrate the crucial role of maximizing the schatten-p norm regularization, which can effectively preserve the original data distribution and can efficiently adapt to both balanced and imbalanced datasets.

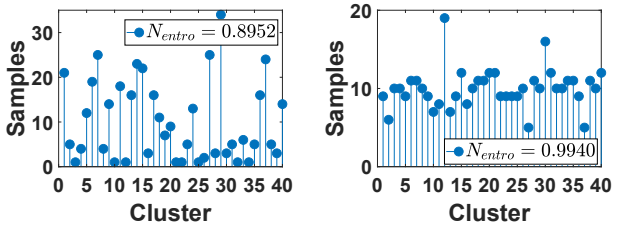
Datasets	Methods	ACC	NMI	Purity	N_{entro}
ORL	$\max_{\mathbf{H}\geq 0, \mathbf{H}\mathbf{1}=\mathbf{1}} \left\|\mathbf{S}^T\mathbf{H}\right\|_F^2$	0.5050	0.7054	0.5250	0.8952
	OUR	**0.6625**	**0.7999**	**0.6975**	**0.9940**
Face-V5	$\max_{\mathbf{H}\geq 0, \mathbf{H}\mathbf{1}=\mathbf{1}} \left\|\mathbf{S}^T\mathbf{H}\right\|_F^2$	0.6424	0.8843	0.6456	0.4534
	OUR	**0.9712**	**0.9917**	**0.9728**	**0.4554**
Pendigits	$\max_{\mathbf{H}\geq 0, \mathbf{H}\mathbf{1}=\mathbf{1}} \left\|\mathbf{S}^T\mathbf{H}\right\|_F^2$	0.6104	0.5464	0.6104	0.8021
	OUR	**0.8480**	**0.7798**	**0.8480**	**0.9921**

Table 3: The clustering results of ablation experiments.

Balance

To further evaluate the effectiveness of our proposed regularization term in promoting class balance, we conducted balance experiments on the ORL dataset and visualized the sample distribution per class using bar charts.

First, we compared the performance of the Frobenius norm regularization with our maximized schatten-p norm regularization, as shown in Figure 3. The method using the Frobenius norm results in several classes having only 1 sample, while a class contains up to 34 samples, indicating a serious class imbalance. In contrast, our method achieved a more balanced distribution, with most classes having around 10 samples, the lowest being 5 and the highest being 18, demonstrating significantly better class balance.



(a) $\max_{\mathbf{H}\geq 0,\mathbf{H1}=\mathbf{1}} \left\|\mathbf{S}^T\mathbf{H}\right\|_F^2$ (b) OUR

Figure 3: Distribution of Samples in clusters between (a) $\max_{\mathbf{H}\geq 0,\mathbf{H1}=\mathbf{1}} \left\|\mathbf{S}^T\mathbf{H}\right\|_F^2$ and (b) our $\max_{\mathbf{H}\geq 0,\mathbf{H1}=\mathbf{1}} \left\|\mathbf{S}^T\mathbf{H}\right\|_{sp}^p$.

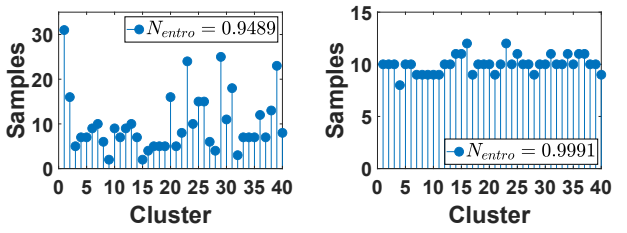
Second, we compare the standard K-Means with a regularization-enhanced version, namely K-Means-sp, which incorporates the maximization of the schatten-p norm. As shown in Table 4, K-Means-sp achieves not only better clustering accuracy but also improved class balance, as reflected by the N_{entro} metric. As shown in Figure 4, on the ORL dataset, the original K-Means (Figure 4(a)) showed notable imbalance, with class sizes ranging from as low as 2 to as high as 31. When our regularization was applied (Figure 4(b)), the distribution became much more uniform, with class sizes mostly around 10. This further confirms the effectiveness of our regularization in encouraging balanced clustering results.

Datasets	Methods	ACC	NMI	Purity	N_{entro}
ORL	K-Means	0.5250	0.7381	**0.5800**	0.9489
	K-Means-sp	**0.5450**	**0.7450**	0.5775	**0.9991**
Face-V5	K-Means	0.7248	0.9285	0.7768	0.4547
	K-Means-sp	**0.8804**	**0.9646**	**0.8904**	**0.4554**
Pendigits	K-Means	0.6615	0.6629	0.6858	0.9633
	K-Means-sp	**0.7291**	**0.6752**	**0.7291**	**0.9999**

Table 4: The results of K-Means and K-Means-sp.

Verification of Convergence

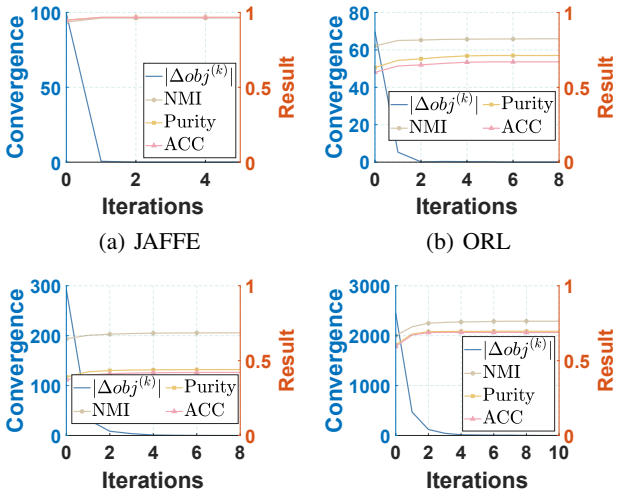
We also conducted convergence analysis on several datasets, including JAFFE, ORL, AR and ISOLET, and plotted the corresponding convergence curves, as shown in Figure 5. In the figure, the left side illustrates the convergence behavior of the regularization term during optimization, while the right vertical axis represents the trend of clustering performance. As observed, our method generally converges within



(a) $\min_{\mathbf{H,U}} \left\|\mathbf{X}-\mathbf{HU}\right\|_F^2$ (b) $\min_{\mathbf{H,U}} \left\|\mathbf{X}-\mathbf{HU}\right\|_F^2 - \lambda\left\|\mathbf{H}\right\|_{sp}^p$

Figure 4: Distribution of Samples in clusters between (a) K-Means and (b) K-Means-sp.

a dozen iterations, and the clustering performance reaches a high level early in the process and remains stable thereafter. These results demonstrate that the proposed model exhibits good convergence properties and optimization efficiency.



(a) JAFFE (b) ORL

(c) AR (d) ISOLET

Figure 5: Convergence curve.

Conclusion

In this work, we propose a maximizing Schatten-p norm regularization term with $1 \leqslant p < 2$ that significantly enhances cluster balance during the clustering process. Unlike traditional methods that minimize the Schatten-p norm, our method maximizes this regularization term to guide more reasonable data point assignments, resulting in improved clustering quality. Moreover, to ensure both representational power and computational efficiency, we employ an anchor graph structure, which approximates sample relationships using a small set of anchor points. This approach reduces computational cost while preserving essential structural information. Extensive experimental results demonstrate that the clustering framework combining the proposed regularization and the anchor graph mechanism exhibits superior performance and robustness across various datasets.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, Grant No. 62176203 and 62576263; the National Natural Science Foundation of China, Grant No. 625B2137; the Natural Science Basic Research Program of Shaanxi Province, Grant No. 2025JC-QYCX-051; the Fundamental Research Funds for the Central Universities and the Innovation Fund of Xidian University, Grant No. YJSJ25007.

References

Chen, H.; Zhang, R.; Wang, R.; and Nie, F. 2025. Fuzzy Min-Cut With Soft Balancing Effects. *IEEE Transactions on Fuzzy Systems*, 33(2): 767–778.

Deng, Y.-J.; Yang, M.-L.; Li, H.-C.; Long, C.-F.; Fang, K.; and Du, Q. 2024. Feature Dimensionality Reduction With L2,p-Norm-Based Robust Embedding Regression for Classification of Hyperspectral Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, 226–231.

Gao, Q.; Li, F.; Wang, Q.; Gao, X.; and Tao, D. 2025. Manifold Based Multi-View K-Means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 3175–3182.

Gao, Q.; Zhang, P.; Xia, W.; Xie, D.; Gao, X.; and Tao, D. 2021. Enhanced Tensor RPCA and its Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6): 2133–2140.

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Weighted Nuclear Norm Minimization with Application to Image Denoising. CVPR '14, 2862–2869. USA: IEEE Computer Society.

Han, J.; Liu, H.; and Nie, F. 2019. A Local and Global Discriminative Framework and Optimization for Balanced Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 30(10): 3059–3071.

Hartigan, J. A.; and Wong, M. A. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1): 100–108.

Hull, J. 1994. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5): 550–554.

Li, F.; Gao, Q.; Wang, Q.; Yang, M.; and Deng, C. 2025a. Tensorized Soft Label Learning Based on Orthogonal NMF. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7): 13219–13231.

Li, X.; Pan, Y. P.; Sun, Y.; Sun, Q.; Sun, Y.; W. Tsang, I.; and Ren, Z. 2025b. Incomplete Multi-view Clustering with Paired and Balanced Dynamic Anchor Learning. *IEEE Transactions on Multimedia*, 7087–7098.

Lin, W.; He, Z.; and Xiao, M. 2019. Balanced Clustering: A Uniform Model and Fast Algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 2987–2993.

Liu, H.; Han, J.; Nie, F.; and Li, X. 2017. Balanced clustering with least square regression. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 2231–2237.

Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; and Shum, H.-Y. 2011. Learning to Detect a Salient Object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2): 353–367.

Lu, C.; Tang, J.; Yan, S.; and Lin, Z. 2016. Nonconvex Nonsmooth Low Rank Minimization via Iteratively Reweighted Nuclear Norm. *IEEE Transactions on Image Processing*, 25(2): 829–839.

Lyons, M. J.; Kamachi, M. G.; and Gyoba, J. 1998. The Japanese Female Facial Expression (JAFFE) Dataset. In *third international conference on automatic face and gesture recognition*, 14–16.

Martinez, A.; and Benavente, R. 1998. *The AR Face Database: CVC Technical Report, 24*.

Mcinnes, L.; Healy, J.; and Astels, S. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11).

Nie, F.; Xue, J.; Wu, D.; Wang, R.; Li, H.; and Li, X. 2022. Coordinate Descent Method for k-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2371–2385.

Nie, F.; Xue, J.; Yu, W.; and Li, X. 2024. Fast Clustering With Anchor Guidance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 1898–1912.

Pei, S.; Chen, H.; Nie, F.; Wang, R.; and Li, X. 2023. Centerless Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 167–181.

Samaria, F.; and Harter, A. 1994. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, 138–142.

Sun, J.; Shen, X.; and Sun, Q. 2023. Efficient Feature Reconstruction via L2,1-Norm Regularization for Few-Shot Classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12): 7452–7465.

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

Wang, J.; Xie, F.; Nie, F.; and Li, X. 2024. Generalized and Robust Least Squares Regression. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5): 7006–7020.

Xia, W.; Gao, Q.; Wang, Q.; Gao, X.; Ding, C.; and Tao, D. 2023. Tensorized Bipartite Graph Learning for Multi-View Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 5187–5202.

Xie, J.; Yang, J.; Qian, J. J.; Tai, Y.; and Zhang, H. M. 2017. Robust Nuclear Norm-Based Matrix Regression With Applications to Robust Face Recognition. *IEEE Transactions on Image Processing*, 26(5): 2286–2295.

Xie, Y.; Gu, S.; Liu, Y.; Zuo, W.; Zhang, W.; and Zhang, L. 2016a. Weighted Schatten p-Norm Minimization for Image Denoising and Background Subtraction. *IEEE Transactions on Image Processing*, 25(10): 4842–4857.

Xie, Y.; Qu, Y.; Tao, D.; Wu, W.; Yuan, Q.; and Zhang, W. 2016b. Hyperspectral Image Restoration via Iteratively Regularized Weighted Schatten p-Norm Minimization. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8): 4642–4659.

Yang, M.; Li, Y.; and Wang, J. 2022. Feature and Nuclear Norm Minimization for Matrix Completion. *IEEE Transactions on Knowledge and Data Engineering*, 34(5): 2190–2199.

Zhang, H.; Nie, F.; and Li, X. 2023. Large-Scale Clustering With Structured Optimal Bipartite Graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9950–9963.

Zhang, H.; Yang, J.; Shang, F.; Gong, C.; and Zhang, Z. 2019. LRR for Subspace Segmentation via Tractable Schatten-p Norm Minimization and Factorization. *IEEE Transactions on Cybernetics*, 49(5): 1722–1734.

Zhang, H.; Zhao, J.; Zhang, B.; Gong, C.; Qian, J.; and Yang, J. 2024. Unified Framework for Faster Clustering via Joint Schatten p-Norm Factorization With Optimal Mean. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3): 3012–3026.

Zhong, S.; and Ghosh, J. 2003. Model-based Clustering with Soft Balancing. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, 459. USA: IEEE Computer Society.