Orthogonal Non-negative Tensor Factorization based Multi-view Clustering

Jing Li
Xidian University
Xi'an, Shaanxi, China
jinglxd@stu.xidian.edu.cn

Quanxue Gao *
Xidian University
Xi'an, Shaanxi, China
qxgao@xidian.edu.cn

Qianqian Wang
Xidian University
Xi'an, Shaanxi, China
qqwang@xidian.edu.cn

Ming Yang
Harbin Engineering University
Harbin, Heilongjiang, China
yangmingmath@gmail.com

Wei Xia
Xidian University
Xi'an, Shaanxi, China
xdweixia@gmail.com

A Proof of Convergence

A.1 Proof of the 1st part

Lemma 1 (Proposition 6.2 of [1]). *Suppose $F : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}$ is represented as $F(X) = f \circ \sigma(X)$, where $X \in \mathbb{R}^{n_1\times n_2}$ with SVD $X = U\mathrm{diag}(\sigma_1, \ldots, \sigma_n)V^{\mathrm{T}}$, $n = \min(n_1, n_2)$, and f is differentiable. The gradient of $F(X)$ at X is*

$$\frac{\partial F(X)}{\partial X} = U\mathrm{diag}(\theta)V^{\mathrm{T}}, \tag{1}$$

where $\theta = \frac{\partial f(y)}{\partial y}|_{y=\sigma(X)}$.

To minimize $\bar{\mathcal{Q}}^{(v)}$ at step $k+1$ in (21) in the main body, the optimal $\bar{\mathcal{Q}}_{k+1}^{(v)}$ needs to satisfy the first-order optimal condition

$$\bar{\mathcal{Q}}_{k+1}^{(v)} = \bar{\mathcal{H}}_k^{(v)} + \frac{\bar{\mathcal{Y}}_{1,k}}{\mu_k}.$$

By using the updating rule $\bar{\mathcal{Y}}_{1,k+1}^{(v)} = \bar{\mathcal{Y}}_{1,k}^{(v)} + \mu_k(\bar{\mathcal{H}}_k^{(v)} - \bar{\mathcal{Q}}_k^{(v)})$, we have

$$\frac{\bar{\mathcal{Y}}_{1,k+1}^{(v)}}{\mu_k} + (\bar{\mathcal{Q}}_k^{(v)} - \bar{\mathcal{Q}}_{k+1}^{(v)}) = 0.$$

According to our assumption $\lim_{k\to 0}\mu_k(\bar{\mathcal{Q}}_{k+1}^{(v)} - \bar{\mathcal{Q}}_k^{(v)}) = 0$, we know $\mathcal{Y}_{1,k+1}$ is bounded.

To minimize $\mathcal{J}$ at step $k+1$ in (24) in the main body, the optimal $\mathcal{J}_{k+1}$ needs to satisfy the first-order optimal condition

$$\lambda\nabla_{\mathcal{J}}\|\mathcal{J}_{k+1}\|_{\circledS p}^p + \rho_k(\mathcal{J}_{k+1} - \mathcal{H}_{k+1} - \frac{1}{\rho_k}\mathcal{Y}_{2,k}) = 0.$$

Recall that when $0 < p < 1$, in order to overcome the singularity of $(|\eta|^p)' = p\eta/|\eta|^{2-p}$ near $\eta = 0$, we consider for $0 < \epsilon \ll 1$ the approximation

$$\partial|\eta|^p \approx \frac{p\eta}{\max\{\epsilon^{2-p}, |\eta|^{2-p}\}}.$$

*Corresponding author.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Letting $\overline{\boldsymbol{\mathcal{J}}}^{(i)} = \overline{\boldsymbol{\mathcal{U}}}^{(i)} \mathrm{diag}\left(\sigma_j(\overline{\boldsymbol{\mathcal{J}}}^{(i)})\right) \overline{\boldsymbol{\mathcal{V}}}^{(i)\mathrm{H}}$, then it follows from Lemma 1 that

$$\frac{\partial \|\overline{\boldsymbol{\mathcal{J}}}^{(i)}\|_{\circledS p}^p}{\partial \overline{\boldsymbol{\mathcal{J}}}^{(i)}} = \overline{\boldsymbol{\mathcal{U}}}^{(i)} \mathrm{diag}\left(\frac{p\sigma_j(\overline{\boldsymbol{\mathcal{J}}}^{(i)})}{\max\{\epsilon^{2-p}, |\sigma_j(\overline{\boldsymbol{\mathcal{J}}}^{(i)})|^{2-p}\}}\right) \overline{\boldsymbol{\mathcal{V}}}^{(i)\mathrm{H}}.$$

And then one can obtain

$$\begin{aligned} &\frac{p\sigma_j(\overline{\boldsymbol{\mathcal{J}}}^{(i)})}{\max\{\epsilon^{2-p}, |\sigma_j(\overline{\boldsymbol{\mathcal{J}}}^{(i)})|^{2-p}\}} \leq \frac{p}{\epsilon^{1-p}} \\ &\Longrightarrow \left\| \frac{\partial \|\overline{\boldsymbol{\mathcal{J}}}^{(i)}\|_{\circledS p}^p}{\partial \overline{\boldsymbol{\mathcal{J}}}^{(i)}} \right\|_F^2 \leq \sum_{i=1}^n \frac{p^2}{\epsilon^{2(1-p)}}. \end{aligned}$$

So $\frac{\partial \|\overline{\boldsymbol{\mathcal{J}}}\|_{\circledS p}^p}{\partial \overline{\boldsymbol{\mathcal{J}}}}$ is bounded.

Let us denote $\widetilde{\mathbf{F}}_V = \frac{1}{\sqrt{V}}\mathbf{F}_V$, $\mathbf{F}_V$ is the discrete Fourier transform matrix of size $V \times V$, $\mathbf{F}_V^{\mathrm{H}}$ denotes its conjugate transpose. For $\boldsymbol{\mathcal{J}} = \overline{\boldsymbol{\mathcal{J}}} \times_3 \widetilde{\mathbf{F}}_V$ and using the chain rule in matrix calculus, one can obtain that

$$\nabla_{\boldsymbol{\mathcal{J}}} \|\boldsymbol{\mathcal{J}}\|_{\circledS p}^p = \frac{\partial \|\boldsymbol{\mathcal{J}}\|_{\circledS p}^p}{\partial \overline{\boldsymbol{\mathcal{J}}}} \times_3 \widetilde{\mathbf{F}}_V^{\mathrm{H}}$$

is bounded.

And it follows that

$$\begin{aligned} &\boldsymbol{\mathcal{Y}}_{1,k+1} = \boldsymbol{\mathcal{Y}}_{2,k} + \rho_k(\boldsymbol{\mathcal{H}}_{k+1} - \boldsymbol{\mathcal{J}}_{k+1}) \\ &\Longrightarrow \lambda \nabla_{\boldsymbol{\mathcal{J}}} \|\boldsymbol{\mathcal{J}}_{k+1}\|_{\circledS p}^p = \boldsymbol{\mathcal{Y}}_{2,k+1}, \end{aligned}$$

thus $\boldsymbol{\mathcal{Y}}_{2,k+1}$ appears to be bounded.

Moreover, by using the updating rule $\boldsymbol{\mathcal{Y}}_1 = \boldsymbol{\mathcal{Y}}_1 + \mu(\boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{Q}})$, $\boldsymbol{\mathcal{Y}}_2 = \boldsymbol{\mathcal{Y}}_2 + \rho(\boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{J}})$, we can deduce ($i = 1, 2$)

$$\begin{aligned} &\boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{Q}}_{k+1}, \boldsymbol{\mathcal{G}}_{k+1}, \boldsymbol{\mathcal{H}}_{k+1}, \boldsymbol{\mathcal{J}}_{k+1}; \boldsymbol{\mathcal{Y}}_{i,k}) \\ &\leq \boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{Q}}_k, \boldsymbol{\mathcal{G}}_k, \boldsymbol{\mathcal{H}}_k, \boldsymbol{\mathcal{J}}_k; \boldsymbol{\mathcal{Y}}_{i,k}) \\ &= \boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{Q}}_k, \boldsymbol{\mathcal{G}}_k, \boldsymbol{\mathcal{H}}_k, \boldsymbol{\mathcal{J}}_k; \boldsymbol{\mathcal{Y}}_{i,k-1}) \\ &+ \frac{\rho_k + \rho_{k-1}}{2\rho_{k-1}^2} \|\boldsymbol{\mathcal{Y}}_{2,k} - \boldsymbol{\mathcal{Y}}_{2,k-1}\|_F^2 + \frac{\|\boldsymbol{\mathcal{Y}}_{2,k}\|_F^2}{2\rho_k} - \frac{\|\boldsymbol{\mathcal{Y}}_{2,k-1}\|_F^2}{2\rho_{k-1}} \\ &+ \frac{\mu_k + \mu_{k-1}}{2\mu_{k-1}^2} \|\boldsymbol{\mathcal{Y}}_{1,k} - \boldsymbol{\mathcal{Y}}_{1,k-1}\|_F^2 + \frac{\|\boldsymbol{\mathcal{Y}}_{1,k}\|_F^2}{2\mu_k} - \frac{\|\boldsymbol{\mathcal{Y}}_{1,k-1}\|_F^2}{2\mu_{k-1}}. \end{aligned} \tag{2}$$

Thus, summing two sides of (2) from $k = 1$ to n, we have

$$\begin{aligned} &\boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{Q}}_{n+1}, \boldsymbol{\mathcal{G}}_{n+1}, \boldsymbol{\mathcal{H}}_{n+1}, \boldsymbol{\mathcal{J}}_{n+1}; \boldsymbol{\mathcal{Y}}_{i,n}) \\ \leq &\boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{Q}}_1, \boldsymbol{\mathcal{G}}_1, \boldsymbol{\mathcal{H}}_1, \boldsymbol{\mathcal{J}}_1; \boldsymbol{\mathcal{Y}}_{i,0})) \\ + &\frac{\|\boldsymbol{\mathcal{Y}}_{2,n}\|_F^2}{2\rho_n} - \frac{\|\boldsymbol{\mathcal{Y}}_{2,0}\|_F^2}{2\rho_0} + \sum_{k=1}^n \left(\frac{\rho_k + \rho_{k-1}}{2\rho_{k-1}^2} \|\boldsymbol{\mathcal{Y}}_{2,k} - \boldsymbol{\mathcal{Y}}_{2,k-1}\|_F^2 \right) \\ + &\frac{\|\boldsymbol{\mathcal{Y}}_{1,n}\|_F^2}{2\mu_n} - \frac{\|\boldsymbol{\mathcal{Y}}_{1,0}\|_F^2}{2\mu_0} + \sum_{k=1}^n \left(\frac{\mu_k + \mu_{k-1}}{2\mu_{k-1}^2} \|\boldsymbol{\mathcal{Y}}_{1,k} - \boldsymbol{\mathcal{Y}}_{1,k-1}\|_F^2 \right). \end{aligned} \tag{3}$$

Observe that

$$\sum_{k=1}^{\infty} \frac{\rho_k + \rho_{k-1}}{2\rho_{k-1}^2} < \infty, \sum_{k=1}^{\infty} \frac{\mu_k + \mu_{k-1}}{2\mu_{k-1}^2} < \infty,$$

we have the right-hand side of (3) is finite and thus $\mathcal{L}(\mathcal{Q}_{n+1}, \mathcal{G}_{n+1}, \mathcal{H}_{n+1}, \mathcal{J}_{n+1}; \mathcal{Y}_{i,n})$ is bounded. Notice from (7) in the main body

$$\begin{aligned}
&\mathcal{L}(\mathcal{Q}_{n+1}, \mathcal{G}_{n+1}, \mathcal{H}_{n+1}, \mathcal{J}_{n+1}; \mathcal{Y}_{i,n}) \\
&= \sum_{v=1}^{V} \left\| \bar{\mathcal{S}}^{(v)} - \bar{\mathcal{H}}_{n+1}^{(v)} (\bar{\mathcal{G}}_{n+1}^{(v)})^T \right\|_F^2 \\
&+ \lambda \|\mathcal{J}_{n+1}\|_{\circledS p}^p + \frac{\rho_n}{2} \|\mathcal{H}_{n+1} - \mathcal{J}_{n+1} + \frac{\mathcal{Y}_{2,n}}{\rho_n}\|_F^2 \\
&+ \frac{\mu_n}{2} \sum_{v=1}^{V} \|\bar{\mathcal{H}}_{n+1}^{(v)} - \bar{\mathcal{Q}}_{n+1}^{(v)} + \frac{\bar{\mathcal{Y}}_{1,n+1}^{(v)}}{\mu_n}\|_F^2,
\end{aligned} \tag{4}$$

and each term of (4) is nonnegative, following from the boundedness of $\mathcal{L}(\mathcal{Q}_{n+1}, \mathcal{G}_{n+1}, \mathcal{H}_{n+1}, \mathcal{J}_{n+1}; \mathcal{Y}_{i,n})$, we can deduce each term of (4) is bounded. And $\|\mathcal{J}_{n+1}\|_{\circledS p}^p$ being bounded implies that all singular values of $\mathcal{J}_{n+1}$ are bounded and hence $\|\mathcal{J}_{n+1}\|_F^2$ (the sum of squares of singular values) is bounded. Therefore, the sequence $\{\mathcal{J}_k\}$ is bounded.

Because

$$\mathcal{Y}_{1,k+1} = \mathcal{Y}_{1,k} + \mu_k(\mathcal{Q}_k - \mathcal{H}_k) \Longrightarrow \mathcal{H}_k = \mathcal{Q}_k + \frac{\mathcal{Y}_{1,k+1} - \mathcal{Y}_{1,k}}{\mu_k},$$

and in light of the boundedness of $\mathcal{Q}_k, \mathcal{Y}_{1,k}$, it is clear that $\mathcal{H}_k$ is also bounded.

And from (8) in the main body, it is evident that $\|\bar{\mathcal{G}}_k^{(v)}\|_F^2 \leq \|(\bar{\mathcal{S}}^{(v)})^{\mathrm{T}}\|_F^2 \|\bar{\mathcal{H}}_k^{(v)}\|_F^2$, so $\bar{\mathcal{G}}_k^{(v)}$ is also bounded. So $\mathcal{G}_k$ is bounded.

A.2 Proof of the 2nd part

From Weierstrass-Bolzano theorem, there exists at least one accumulation point of the sequence $\mathcal{P}_k$. We denote one of the points $\mathcal{P}^* = \{\mathcal{H}^*, \mathcal{Q}^*, \mathcal{G}^*, \mathcal{J}^*, \mathcal{Y}_{\mathbf{1}}^*, \mathcal{Y}_2^*\}$. Without loss of generality, we assume $\{\mathcal{P}_k\}_{k=1}^{+\infty}$ converge to P^*.

Note that from the updating rule for $\mathcal{Y}_1$, we have

$$\mathcal{Y}_{1,k+1} = \mathcal{Y}_{1,k} + \mu_k(\mathcal{H}_k - \mathcal{Q}_k) \Longrightarrow \mathcal{Q}^* = \mathcal{H}^*.$$

Note that from the updating rule for $\mathcal{Y}_2$, we have

$$\mathcal{Y}_{2,k+1} = \mathcal{Y}_{2,k} + \rho_k(\mathcal{H}_k - \mathcal{J}_k) \Longrightarrow \mathcal{J}^* = \mathcal{H}^*.$$

In the $\bar{\mathcal{G}}^{(v)}$-subproblem (8) in the main body, we have

$$\bar{\mathcal{G}}_k^{(v)} = (\bar{\mathcal{S}}^{(v)})^{\mathrm{T}} \bar{\mathcal{H}}_k^{(v)} \Longrightarrow \bar{\mathcal{G}}^{(v)*} = (\bar{\mathcal{S}}^{(v)})^{\mathrm{T}} \bar{\mathcal{H}}^{(v)*}.$$

In the $\mathcal{J}$-subproblem (24) in the main body, we have

$$\lambda \nabla_{\mathcal{J}} \|\mathcal{J}_{k+1}\|_{\circledS p}^p = \mathcal{Y}_{2,k} \Longrightarrow \mathcal{Y}_1^* = \lambda \nabla_{\mathcal{J}} \|\mathcal{J}^*\|_{\circledS p}^p.$$

Therefore, one can see that the sequences $\mathcal{H}^*, \mathcal{Q}^*, \mathcal{G}^*, \mathcal{J}^*, \mathcal{Y}_{\mathbf{1}}^*, \mathcal{Y}_2^*$ satisfy the KKT conditions of the Lagrange function (7) in the main body.

B Anchor Selection And Graph Construction

Inspired by [2], we adopt directly alternate sampling (DAS) to select anchors.

First of all, with the given data matrices $\{\mathbf{X}^{(v)}\}_{v=1}^V$, we concatenate the data matrix of each view along the feature dimension. The connected feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ can be represented as $\mathbf{X} = [\mathbf{X}^{(1)}; \mathbf{X}^{(2)}; \cdots; \mathbf{X}^{(v)}]$, where d is the sum of the number of features in each view. Let θ_i represent the i-th sample of the d-dimensional features, which can be calculated as

$$\theta_i = \sum_{j=1}^{dT} Tra(X_{ij}), \tag{5}$$

where $dT = \sum_{v=1}^{V} d_v$, and $Tra(\cdot)$ represents the transformation of the raw features. Specifically, if the features are negative, we process the features of each dimension by subtracting the minimum value in each dimension. Then we obtain the score vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_n] \in \mathbb{R}^n$. We choose the point where the maximum score is located as the anchor. The position of the largest score is

$$Index = \arg\max_i \theta_i. \tag{6}$$

Then the 1st anchor of the v-th view is $b_1^{(v)} = x_{Index}^{(v)}$.

After that, let θ_{Index} be the score of the anchor selected from the last round, then we normalize the score of each sample by:

$$\theta_i \leftarrow \frac{\theta_i}{\max \boldsymbol{\theta}}, (i = 1, 2, \cdots, n) \tag{7}$$

Then the score θ_i can be updated as

$$\theta_i \leftarrow \theta_i \times (1 - \theta_i). \tag{8}$$

Finally, we repeat (6) - (8) m times to select m anchors. After selecting m anchors, we construct an anchor graph of each view $\mathbf{S}^{(v)}$, in the same way, as [2].

C More Details of the Experiments

C.1 Experimental Configurations

The $Reuters$ and $NoisyMNIST$ are implemented on a standard Windows 10 Server with two Intel (R) Xeon (R) Gold 6230 CPUs 2.1 GHz and 128 GB RAM, MATLAB R2020a. The $MSRC$, $HandWritten4$, $Mnist4$ and AWA are implemented on a laptop computer with an Inter Core i5-8300H CPU and 16 GB RAM, using Matlab R2018b. Codes are available: https://github.com/xdjingli/Orth-NTF.

We repeated the all methods 20 times independently and showed the averages with the corresponding standard deviations. The specific hype-parameters on each dataset are as follows:

- MSRC: anchor rate = 0.7, p = 0.5, λ = 100.
- HandWritten4: anchor rate = 1.0, p = 0.1, λ = 1180.
- Mnist4: anchor rate = 0.6, p = 0.1, λ = 5000.
- AWA: anchor rate = 1.0, p = 0.5, λ = 1000.
- Reuters: anchor rate = 0.005 (anchor number = 100), p = 0.4, λ = 1209800.
- NoisyMnist: anchor rate = 0.03, p = 0.1, λ = 200000.

C.2 Impact for Parameters

In our proposed algorithm, the number of anchors, the value of p from the tensor Schatten p-norm, and the value of λ are variable parameters. In this section we take 4 datasets: MSRC, HandWritten4, Mnist4, and AWA as examples to analyze the effect of these variable parameters.

Effect of the number of anchors. We changed the anchor rate from 0.1 to 1.0 with step size 0.1. The changes of clustering results and algorithm running time along with the anchor rate were tested on MSRC, HandWritten4, Mnist4 and AWA, as shown in Fig 1 and Fig 2. When the anchor rate were 0.7, 1.0, 0.6 and 1.0, the best clustering results were obtained on MSRC, HandWritten4, Mnist4 and AWA, respectively. The time required for clustering is approximately linearly related to the increase of anchor rate.

Effect of the value of p. We set the value of p to be 0.1 to 1.0 with a step of 0.1. We obtained the results of ACC, NMI, and Purity in experiments with different values of p as shown in Fig 3. The best clustering results are obtained on MSRC, HandWritten4, Mnist4 and AWA when the values of p are $0.5,\ 0.2,\ 0.1$, and 0.5, respectively. It indicates that tensor Schatten pnorm can take advantage of the low-rank of views which helps mine the complementary information of different views. This helps get better clustering results.

Effect of the value of λ. To determine the value of λ, we initially approximate its range using the magnitude of the tensor Schatten p-norm regularization, followed by a more detailed fine-tuning within that range. The impact of varying parameter combinations on the method's performance can be seen in Fig 4. This figure highlights the clustering performance across different pairings of p and λ.

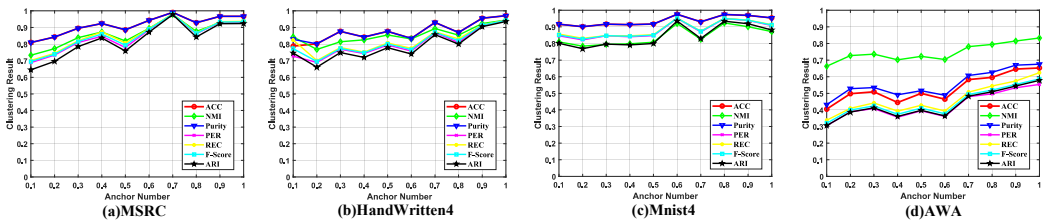

Figure 1: Clustering results with different anchor rate on MSRC, HandWritten4, Mnist4 and AWA.

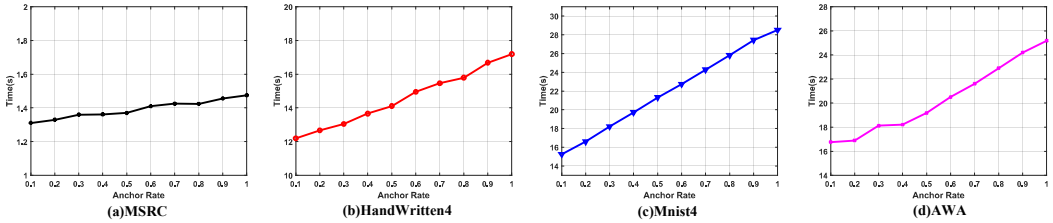

Figure 2: Time (sec.) with different number of anchors on MSRC, HandWritten4, Mnist4 and AWA.

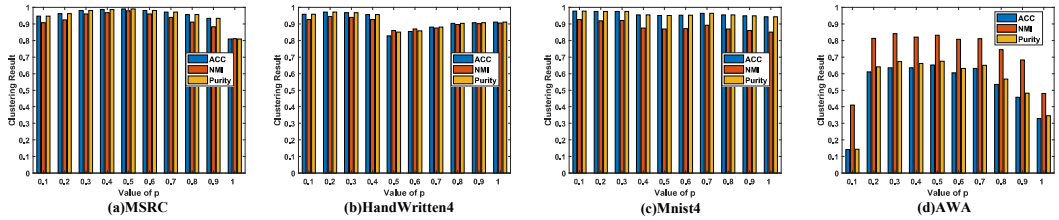

Figure 3: The influence of p on clustering results on MSRC, HandWritten4, Mnist4 and AWA.

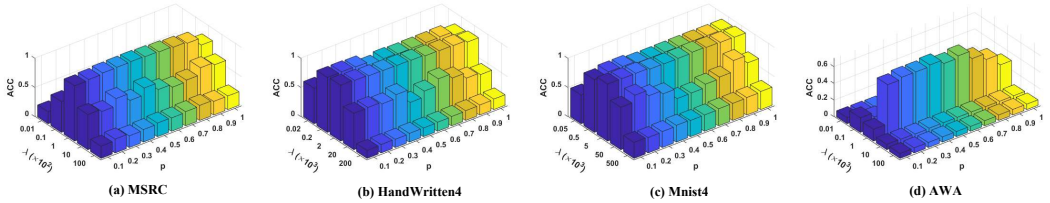

Figure 4: The influence of λ and p on clustering results on MSRC, HandWritten4, Mnist4 and AWA.

References

[1] Lewis, A. S. and Sendov, H. S. Nonsmooth analysis of singular values. part i: Theory. *Set-Valued Analysis*, 13(3):213–241, 2005.

[2] Li, X., Zhang, H., Wang, R., and Nie, F. Multiview clustering: A scalable and parameter-free bipartite graph fusion method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):330–344, 2022.