Supplementary Materials: Scalable Multi-view Unsupervised Feature Selection with Structure Learning and Fusion

Anonymous Authors

1 APPENDIX A. THE DETAILED OPTIMIZATION FOR α

By fixing other variables except $\boldsymbol{\alpha}$, we can optimize $\boldsymbol{\alpha}$ by rows:

$$\min_{\boldsymbol{\alpha}_i \mathbf{1}=1, \boldsymbol{\alpha}_i \geq 0} \|\mathbf{u}_i - \sum_{v=1}^{V} \alpha_i^v \widetilde{\mathbf{u}}_i^v\|_2^2. \tag{1}$$

Eq. (1) can be equally transformed into the following form:

$$\|\mathbf{u}_i - \sum_{v=1}^{V} \alpha_i^v \widetilde{\mathbf{u}}_i^v\|_2^2 = \|\sum_{v=1}^{V} \alpha_i^v \mathbf{u}_i - \sum_{v=1}^{V} \alpha_i^v \widetilde{\mathbf{u}}_i^v\|_2^2$$
$$\|\sum_{v=1}^{V} \alpha_i^v (\mathbf{u}_i - \widetilde{\mathbf{u}}_i^v)\|_2^2 = \|\boldsymbol{\alpha}_i \mathbf{D}_i\|_2^2 \tag{2}$$

where $\mathbf{D}_i = [\mathbf{d}_i^1; ...; \mathbf{d}_i^V] \in \mathbb{R}^{V \times c}$, and $\mathbf{d}_i^v = \mathbf{u}_i - \widetilde{\mathbf{u}}_i^v \in \mathbb{R}^{1 \times c}$. Therefore, $\boldsymbol{\alpha}_i$ can be solved as:

$$\min_{\boldsymbol{\alpha}_i \mathbf{1}=1, \boldsymbol{\alpha}_i \geq 0} \boldsymbol{\alpha}_i \mathbf{D}_i \mathbf{D}_i^T \boldsymbol{\alpha}_i^T. \tag{3}$$

Since $\mathbf{D}_i \mathbf{D}_i^T$ is semi-definite, Eq. (3) is a quadratic convex programming problem, which can be solved efficiently [2]. Specifically, Eq. (3) can be solved by tackling its counterpart:

$$\min_{\boldsymbol{\alpha}_i \geq 0, \boldsymbol{\alpha}_i \mathbf{1}=1, \boldsymbol{z}} \boldsymbol{\alpha}_i \mathbf{D}_i \mathbf{D}_i^T \boldsymbol{z}^T + \frac{\mu}{2} \left\|\boldsymbol{\alpha}_i - \boldsymbol{z} + \frac{\boldsymbol{\tau}}{\mu}\right\|_2^2, \tag{4}$$

where $\boldsymbol{z} \in \mathbb{R}^{1 \times V}$ denotes a slack variable, $\mu > 0$ is a penalty parameter, and $\boldsymbol{\tau} \in \mathbb{R}^{1 \times V}$ is a Lagrangian multiplier. Eq. (4) can be iteratively optimized by the augmented Lagrangian multiplier method. The solution steps are as follows:

Step 1. Update $\boldsymbol{z}$: When $\boldsymbol{\alpha}_i$ is fixed, Eq. (4) is an unconstrained optimization problem. By setting the derivative of Eq. (4) w.r.t. $\boldsymbol{z}$ to zero, we update $\boldsymbol{z}$ by:

$$\boldsymbol{z} = \boldsymbol{\alpha}_i - \frac{1}{\mu}\left(\boldsymbol{\alpha}_i \mathbf{D}_i \mathbf{D}_i^T - \boldsymbol{\tau}\right). \tag{5}$$

Step 2. Update $\boldsymbol{\alpha}_i$: When $\boldsymbol{z}$ is fixed with its current value of $\boldsymbol{z}$ (i.e., $\boldsymbol{z}^*$), $\boldsymbol{\alpha}_i$ can be updated by minimizing the following problem:

$$\min_{\boldsymbol{\alpha}_i \mathbf{1}=1, \boldsymbol{\alpha}_i \geq 0} \left\|\boldsymbol{\alpha}_i - \boldsymbol{z}^* + \frac{1}{\mu}(\boldsymbol{\tau} + \boldsymbol{z}^* \mathbf{D}_i \mathbf{D}_i^T)\right\|_2^2, \tag{6}$$

which can be solved with a closed-form solution [1].

Step 3. Update $\boldsymbol{\tau}$ and μ: In each iteration, we update the Lagrange multipliers $\boldsymbol{\tau}$ and the penalty parameter μ as follows:

$$\boldsymbol{\tau} = \boldsymbol{\tau} + \mu(\boldsymbol{\alpha}_i - \boldsymbol{z})$$
$$\mu = \rho\mu. \tag{7}$$

where ρ is a constant update rate. In this way, $\boldsymbol{\alpha}_i$ can be adaptively updated according to the aforementioned steps.

2 APPENDIX B. THE DETAILED OPTIMIZATION FOR S

By fixing other variables except S, we have the following problem:

$$\min_{\mathbf{S}\mathbf{1}=1, \mathbf{S} \geq 0} \lambda \|\mathbf{u}_i - \mathbf{u}_j\|_2^2 s_{ij} + \beta \|\mathbf{S}\|_F^2. \tag{8}$$

Noting that each row of S (i.e., $\mathbf{s}_i$) is uncorrelated with others, hence Eq. (8) can be optimized for each row independently as follows:

$$\min_{\mathbf{s}_i \mathbf{1}=1, \mathbf{s}_i \geq 0} \|\mathbf{s}_i + \frac{1}{2\beta_i} \mathbf{d}_i\|_2^2, \tag{9}$$

where $\mathbf{d}_i$ is a row vector with $d_{ij} = \lambda \|\mathbf{u}_i - \mathbf{u}_j\|_2^2$. The Lagrangian function of the above function is:

$$\mathcal{L}(\mathbf{s}_i, \theta_i, \boldsymbol{\zeta}_i) = \frac{1}{2}\|\mathbf{s}_i + \frac{1}{2\beta_i}\mathbf{d}_i\|_2^2 - \theta_i(\mathbf{s}_i \mathbf{1} - 1) - \mathbf{s}_i \boldsymbol{\zeta}_i,$$

where $\theta_i \in \mathbb{R}$ and $\boldsymbol{\zeta}_i \in \mathbb{R}^{n \times 1}$ are Lagrangian multipliers. According to the KKT condition, the optimal solution of $\mathbf{s}_i$ is:

$$s_{ij} = (-\frac{d_{ij}}{2\beta_i} + \theta_i^*)_+,$$

where θ_i^* denotes the optimal value equipped for the optimal solution of $\mathbf{s}_i$ and $(x)_+ = \max(x, 0)$. Since the local structure contains more useful and detailed information about the data compared to the global structure, it is suitable to construct a sparse graph to focus on a small number of neighbors [4]. With $\mathbf{d}_i$ being sorted from small to large (i.e., $\widetilde{\mathbf{d}}$), there is $s_{i1} \geq s_{i2} \geq \cdots \geq s_{in}$. Assuming that each sample has f-nearest neighbors (i.e., $\mathbf{s}_i$ has f nonzero elements), we derive:

$$\begin{cases} s_{i,f} > 0 \\ s_{i,f+1} = 0 \end{cases} \Longrightarrow \begin{cases} -\frac{\widetilde{d}_{i,f}}{2\beta_i} + \theta_i^* > 0 \\ -\frac{\widetilde{d}_{i,f+1}}{2\beta_i} + \theta_i^* \leq 0 \end{cases}.$$

Due to $\mathbf{s}_i \mathbf{1} = 1$, we obtain:

$$\sum_{j=1}^{f}(-\frac{\widetilde{d}_{ij}}{2\beta_i} + \theta_i^*) = 1 \Longrightarrow \theta_i^* = \frac{1}{f} + \frac{1}{2f\beta_i}\sum_{j=1}^{f}\widetilde{d}_{ij}.$$

Based on the above analysis, the inequality on β_i can be derived as:

$$\frac{f}{2}\widetilde{d}_{i,f} - \frac{1}{2}\sum_{j=1}^{f}\widetilde{d}_{ij} < \beta_i \leq \frac{f}{2}\widetilde{d}_{i,f+1} - \frac{1}{2}\sum_{j=1}^{f}\widetilde{d}_{ij}.$$

when $\beta_i = \frac{f}{2}\widetilde{d}_{i,f+1} - \frac{1}{2}\sum_{j=1}^{f}\widetilde{d}_{ij}$, it satisfies that $s_{i,f+1} = 0$ and $\mathbf{s}_i$ has f nonzero elements exactly. With the optimal β_i and θ_i^*, the solution of $\mathbf{s}_i$ is derived as:

$$s_{ij} = (\frac{\widetilde{d}_{i,f+1} - \widetilde{d}_{ij}}{f\widetilde{d}_{i,f+1} - \sum_{j=1}^{f}\widetilde{d}_{ij}})_+.$$

According to [3], $\beta = \sum_{i=1}^{n}\frac{f\widetilde{d}_{i,f+1} - \sum_{j=1}^{f}\widetilde{d}_{i,j}}{2n}$ is set to the mean of $\beta_1, \beta_2, \cdots, \beta_n$.

REFERENCES

[1] Jin Huang, Feiping Nie, and Heng Huang. 2015. A new simplex sparse learning model to measure data similarity for clustering. In *International Joint Conference on Artificial Intelligence*. 3569–3575.

[2] Bingbing Jiang, Chenglong Zhang, Yan Zhong, Yi Liu, Yingwei Zhang, Xingyu Wu, and Weiguo Sheng. 2023. Adaptive collaborative fusion for multi-view semi-supervised classification. *Information Fusion* 96 (2023), 37–50.

[3] Feiping Nie, Xiaoqian Wang, and Heng Huang. 2014. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*. 977–986.

[4] Hebing Nie, Qun Wu, Haifeng Zhao, Weiping Ding, and Muhammet Deveci. 2023. Flexible Adaptive Graph Embedding for semi-supervised dimension reduction. *Information Fusion* (2023), 101872.