Appendix of "Top-k Feature Selection Framework using Robust 0-1 Integer Programming"

Appendix A: Optimization of Sub-problem 1

In sub-problem 1, we fix $\mathbf{A}$ and $\mathbf{v}$ to optimize $\mathbf{Z}$ and $\mathbf{E}$. This sub-problem can be solved by the ADMM as described in [1]. The augmented Lagrangian function is given by:

$$\begin{aligned}\mathcal{L}_1(\mathbf{Z}, \mathbf{Q}, \mathbf{E}, \mathbf{Y}_1, \mathbf{Y}_2) =& \|\mathbf{Z}\|_1 + \lambda_E \|\mathbf{E}\|_1 + \lambda_Z \|\mathbf{Z} \odot \mathbf{\Theta}\|_1 \\ &+ \langle \mathbf{Y}_1, \mathbf{X} - \mathbf{X}\mathbf{Q} - \mathbf{E} \rangle + \langle \mathbf{Y}_2, \mathbf{Q} - \mathbf{Z} + \mathrm{diag}(\mathbf{Z}) \rangle \\ &+ \frac{\mu}{2} \left(\|\mathbf{X} - \mathbf{X}\mathbf{Q} - \mathbf{E}\|_F^2 + \|\mathbf{Q} - \mathbf{Z} + \mathrm{diag}(\mathbf{Z})\|_F^2 \right).\end{aligned}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are the Lagrange multipliers, $\mu > 0$ is an adaptive parameter, and $\langle \cdot, \cdot \rangle$ denotes the inner product.

When optimizing $\mathbf{Z}$, we solve the following sub-problem:

$$\mathbf{Z}^{(t+1)} = \arg\min_{\mathbf{Z}} \left\| (\mathbf{1}\mathbf{1}^T + \lambda_Z \mathbf{\Theta}) \odot \mathbf{Z} \right\|_1 + \frac{\mu^{(t)}}{2} \left\| \mathbf{Q}^{(t)} - \mathbf{Z} + \mathrm{diag}(\mathbf{Z}) + \frac{\mathbf{Y}_2^{(t)}}{\mu^{(t)}} \right\|_F^2 \tag{1}$$

where $\mathbf{1}$ indicates the vector whose entries are all 1s. The closed-form solution of Eq. (1) is:

$$\mathbf{Z}^{(t+1)} = \hat{\mathbf{Z}}^{(t+1)} - \mathrm{diag}(\mathbf{Z}^{(\hat{t+1})}). \tag{2}$$

where $\hat{\mathbf{Z}}^{(t+1)} = \mathcal{S}_{\frac{1}{\mu^{(t)}}(1+\lambda_Z \Theta_{ij})} \left(\mathbf{Q}^{(t)} + \frac{\mathbf{Y}_2^{(t)}}{\mu^{(t)}} \right)$, and $\mathcal{S}(\cdot)$ is the element-wise shrinkage thresholding operator.

When optimizing $\mathbf{Q}$, we minimize the following objective function:

$$\begin{aligned}\min_{\mathbf{Q}} \quad & \left\langle \mathbf{Y}_1^{(t)}, \mathbf{X} - \mathbf{X}\mathbf{Q} - \mathbf{E}^{(t)} \right\rangle + \left\langle \mathbf{Y}_2^{(t)}, \mathbf{Q} - \mathbf{Z}^{(t+1)} + \mathrm{diag}(\mathbf{Z}^{(t+1)}) \right\rangle \\ & + \frac{\mu^{(t)}}{2} \left(\|\mathbf{X} - \mathbf{X}\mathbf{Q} - \mathbf{E}^{(t)}\|_F^2 + \|\mathbf{Q} - \mathbf{Z}^{(t+1)} + \mathrm{diag}(\mathbf{Z}^{(t+1)})\|_F^2 \right). \end{aligned} \tag{3}$$

Eq. (3) can be solved as:

$$\mathbf{Q}^{(t+1)} = \left(\mathbf{X}^T\mathbf{X} + \mathbf{I}\right)^{-1} \left(\mathbf{X}^T \left(\mathbf{X} - \mathbf{E}^{(t)} + \frac{\mathbf{Y}_1^{(t)}}{\mu^{(t)}} \right) + \mathbf{Z}^{(t+1)} - \mathrm{diag}(\mathbf{Z}^{(t+1)}) \right) \tag{4}$$

When optimizing $\mathbf{E}$, we solve the following sub-problem:

$$\min_{\mathbf{E}} \quad \lambda_E \|\mathbf{E}\|_1 + \frac{\mu^{(t)}}{2} \left\| \mathbf{X} - \mathbf{X}\mathbf{Q}^{(t+1)} - \mathbf{E} + \frac{\mathbf{Y}_1^{(t)}}{\mu^{(t)}} \right\|_F^2 . \tag{5}$$

Similar to solving Eq.(1), we obtain its closed-form solution:

$$\mathbf{E}^{(t+1)} = \mathcal{S}_{\frac{\lambda_E}{\mu^{(t)}}} \left(\mathbf{X} - \mathbf{X}\mathbf{Q}^{(t+1)} + \frac{\mathbf{Y}_1^{(t)}}{\mu^{(t)}} \right). \tag{6}$$

At last, we update the Lagrange multipliers as follows:

$$\begin{aligned} & \mathbf{Y}_1^{(t+1)} = \mathbf{Y}_1^{(t)} + \mu^{(t)}(\mathbf{X} - \mathbf{X}\mathbf{Q}^{(t+1)} - \mathbf{E}^{(t+1)}). \\ & \mathbf{Y}_2^{(t+1)} = \mathbf{Y}_2^{(t)} + \mu^{(t)}(\mathbf{Q}^{(t+1)} - \mathbf{Z}^{(t+1)} + \mathrm{diag}(\mathbf{Z}^{(t+1)})). \\ & \mu^{(t+1)} = \rho\mu^{(t)}. \end{aligned} \tag{7}$$

The ADMM algorithm for solving this sub-problem is shown in Algorithm 1.

Algorithm 1 ADMM for solving Sub-problem 1

Input: $\mathbf{X}$ and $\mathbf{\Theta}$.

Output: $\mathbf{Z}$ and $\mathbf{E}$.

1: **while** not converge **do**
2: Update $\mathbf{Z}$ by Eq. (2).
3: Update $\mathbf{Q}$ by Eq. (4).
4: Update $\mathbf{E}$ by Eq. (6).
5: Update $\mathbf{Y}_1$, $\mathbf{Y}_2$ and μ by Eq. (7).
6: **end while**

Appendix B: Semi-supervised learning on W

In unsupervised learning, we use a data matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{x}_{n+1}, \cdots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$, and a few labels $\mathbf{Y}_l = [\mathbf{y}_1, \cdots, \mathbf{y}_n] \in \mathbb{R}^{C \times n}$, where $\mathbf{y}_j$ is the label of $\mathbf{x}_j$ and is a C-dimensional indicator vector (in which $y_{ij} = 1$ indicates that $\mathbf{x}_j$ belongs to the j-th class). $\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N$ are unlabeled data, whose labels are inferred from the labeled data. We define the label matrix of unlabeled data as $\mathbf{Y}_u \in \mathbb{R}^{C \times (N-n)}$.

When we obtain the soft data structure matrix $\mathbf{W}$, we construct the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with diagonal elements $D_{ii} = \sum_{j=1}^{N} W_{ij}$. Then we learn $\mathbf{Y}_u$ using the label propagation method, which solves the following problem:

$$\begin{aligned} \min_{\mathbf{Y}_u} \quad & tr\left([\mathbf{Y}_l, \mathbf{Y}_u]\mathbf{L}[\mathbf{Y}_l, \mathbf{Y}_u]^T\right) \\ s.t. \quad & \mathbf{Y}_u \in \mathcal{Y}. \end{aligned} \tag{8}$$

where $\mathcal{Y}$ is the space of label matrices, i.e., $\mathcal{Y} = \{\mathbf{Y}_u \in \{0,1\}^{C \times (N-n)} : \mathbf{Y}_u^T \mathbf{1}_{N-n} = \mathbf{1}_C, rank(\mathbf{Y}_u) = C\}$.

Eq. (8) can be solved approximately using label propagation approaches, e.g., the harmonic function approach [2]. Specifically, we first divide $\mathbf{L}$ into the following form:

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_{ll} & \mathbf{L}_{lu} \\ \mathbf{L}_{lu}^T & \mathbf{L}_{uu} \end{bmatrix} \tag{9}$$

where $\mathbf{L}_{ll} \in \mathbb{R}^{n \times n}$ and $\mathbf{L}_{uu} \in \mathbb{R}^{(N-n) \times (N-n)}$. Then we compute the harmonic solution:

$$\mathbf{Y}_u = \mathbf{Y}_l \mathbf{L}_{lu} \mathbf{L}_{uu}^{-1}. \tag{10}$$

At last, we discretize $\mathbf{Y}_u$ by setting the maximum value in each column as 1 and setting the others as 0.

References

[1] M. Fan, X. Chang, and D. Tao, “Structure regularized unsupervised discriminant feature analysis,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[2] X. Zhu, Z. Ghahramani, and J. D. Lafferty, “Semi-supervised learning using gaussian fields and harmonic functions,” in *Proceedings of the 20th International conference on Machine learning (ICML-03)*, 2003, pp. 912–919.