Supplementary Materials for Incomplete Multi-View Weak-Label Learning with Noisy Features and Imbalanced Labels

1 Related Works

Multi-View Learning. Multi-view learning leverages multiple data sources to improve performance by capturing view correlations [15]. In the *incomplete multi-view* setting, algorithms handle missing views using semi-supervised learning [12, 15] or contrastive learning [4]. Other methods project multi-view data into a low-dimensional subspace to uncover shared information [17, 6].

Weak-Label Learning. Weak-label learning has primarily focused on the single-view scenario. Approaches like LEML [16], SAFEML [10] and POLAR [11] address weak-label classification. Methods such as lrMMC [5] and McWL [9] use matrix completion for multi-view weak-label learning. MC [2] is the first work to complete missing features and labels, by promoting the low-rankness of the label matrix. MAXIDE [13] and COCO [14] recover low-rank label matrices by using features as side information.

Incomplete Multi-View Weak-Label Learning. There are limited studies on this setting. iMVWL [8] learns shared subspaces from incomplete views with weak labels. IMVL-IV [18] addresses multi-view multi-label learning with incomplete views and weak labels. Existing methods encourage the label matrix to be low-rank, disregarding label imbalance. NAIM^{3}L [3] considers high-rank multi-label structures, handling incomplete views and missing labels. However, these methods treat views equally and ignore label imbalance. In contrast, the proposed NAIL adapts weights for embedding incomplete views and weak labels, addressing label imbalance using focal loss while maintaining the low-rank principle.

2 Optimization

For convenience, we present the optimization problem of NAIL as follows:

$$\min_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta},\\ \mathbf{F},\{\mathbf{U}^v\}}} \sum_{v=1}^{m} \alpha_v^s ||\mathbf{O}_{\mathbf{X}}^v \odot (\mathbf{X}^v - \mathbf{F}\mathbf{U}^v)||_{2,1} + \lambda \sum_{(i,j)\in\mathbf{O}_Y} \mathrm{FL}(y_{ij}, \sigma(\mathbf{f}_{i:}^T \mathbf{u}_{:j})) \tag{A1}$$

$$+ \mu \sum_{v=1}^{m+1} \sum_{v'\neq v} \beta_{vv'} \mathrm{HSIC}(\mathbf{U}^{\mathrm{v}}, \mathbf{U}^{\mathrm{v}'}), \ \text{ s.t. } \sum \alpha_{\mathrm{v}} = 1, ||\boldsymbol{\beta}_{\mathrm{v}}||_2 = 1, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{F}, \{\mathbf{U}^{\mathrm{v}}\} \geq 0,$$

2.1 Optimization Algorithm

For simplicity, let $\mathcal{L}$ be the objective function in (A1). Obviously, $\mathcal{L}$ is convex w.r.t α, β, $\mathbf{F}$ and $\mathbf{U}^v$, respectively, that motivates us to develop an alternating optimization algorithm. In this work, the Gaussian kernel is used in HSIC to capture the non-linear correlations among $\{\mathbf{U}^v\}_{v=1}^{m+1}$. The algorithm repeats following steps until convergence.

2.1.1 Update F with fixed others.

When $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\{\mathbf{U}^v\}$ are fixed, the problem w.r.t. $\mathbf{F}$ is

$$\min_{\mathbf{F}\geq 0} \sum_{v=1}^{m} \alpha_v^s ||\mathbf{O}_{\mathbf{X}}^v \odot (\mathbf{X}^v - \mathbf{F}\mathbf{U}^v)||_{2,1} + \lambda \sum_{(i,j)\in\mathbf{O}_Y} \mathrm{FL}(y_{ij}, p_{ij}). \tag{A2}$$

We then propose to optimize (A2) by projected gradient descent (PGD) [1]:

$$\mathbf{F} \leftarrow \mathrm{Proj}(\mathbf{F} - \eta \frac{\partial \mathcal{L}(\mathbf{F})}{\partial \mathbf{F}}), \tag{A3}$$

where η is a learning rate. The projection function is defined as $\mathrm{Proj(a_{ij})} = \mathrm{a_{ij}}$ if $a_{ij} > 0$, and $\mathrm{Proj(a_{ij})} = 0$ otherwise. If the i-th sample is observed in the v-th view, i.e., $(\mathbf{O}_X^v)_i = 1$, the partial derivative of $\mathcal{L}(\mathbf{F})$ w.r.t. the i-th row $\mathbf{f}_{i:}$ of $\mathbf{F}$ is

$$\frac{\partial \mathcal{L}(\mathbf{F})}{\partial \mathbf{f}_{i:}} = -\sum_{v=1}^{m} \frac{\alpha_v^s \mathbf{U}^v (\mathbf{x}_i^v - (\mathbf{U}^v)^T \mathbf{f}_{i:})}{||((\mathbf{x}_i^v)^T - \mathbf{f}_{i:}^T \mathbf{U}^v)||_2} + \lambda \sum_{j\in(\mathbf{O}_Y)_{i:}} y_{ij}(1-q_{ij})^\gamma (\gamma q_{ij} log(q_{ij}) + q_{ij} - 1)\mathbf{u}_{:j}^v. \tag{A4}$$

2.1.2 Update $\mathbf{U}^v$ with fixed others.

With the others fixed, optimization for each $\mathbf{U}^v$ is independent. When $1 \leq v \leq m$, the optimization problem w.r.t. $\mathbf{U}^v$ is

$$\min_{\mathbf{U}^v\geq 0} \alpha_v^s ||\mathbf{O}_{\mathbf{X}}^v \odot (\mathbf{X}^v - \mathbf{F}\mathbf{U}^v)||_{2,1} + \mu \sum_{v'\neq v} \beta_{vv'} \mathrm{HSIC}(\mathbf{U}^{\mathrm{v}}, \mathbf{U}^{\mathrm{v}'}). \tag{A5}$$

The partial derivative of $\mathcal{L}(\mathbf{U}^v)$ w.r.t. $\mathbf{U}^v$ is

$$\frac{\partial \mathcal{L}(\mathbf{U}^v)}{\partial \mathbf{U}^v} = \sum_{i \in \mathbf{O}_X^v} - \frac{\alpha_v^s \mathbf{f}_{i:} (\mathbf{O}_X^v)_i ((\mathbf{x}_i^v)^T - (\mathbf{f}_{i:})^T \mathbf{U}^v)}{||(\mathbf{O}_X^v)_i ((\mathbf{x}_i^v)^T - (\mathbf{f}_{i:})^T \mathbf{U}^v)||_2} + \mu \sum_{v' \neq v} \beta_{vv'} \frac{\partial \mathrm{HSIC}}{\partial \mathbf{U}^{\mathbf{v}}}. \tag{A6}$$

When $v = m + 1$, the optimization problem w.r.t. $\mathbf{U}^v$ becomes

$$\min_{\mathbf{U}^v \geq 0} \sum_{(i,j) \in \mathbf{O}_Y} \mathrm{FL}(y_{ij}, p_{ij}) + \mu \sum_{v' \neq v} \beta_{vv'} \mathrm{HSIC}(\mathbf{U}^{\mathrm{v}}, \mathbf{U}^{\mathrm{v}'}), \tag{A7}$$

and the corresponding partial derivative is

$$\frac{\partial \mathcal{L}(\mathbf{U}^v)}{\partial \mathbf{u}_{:j}^v} = \lambda \sum_{(i,j) \in \mathbf{O}_Y} \mathbf{f}_{i:} y_{ij} (1 - q_{ij})^{\gamma} (\gamma q_{ij} log(q_{ij}) + q_{ij} - 1) + \mu \sum_{v' \neq v} \beta_{vv'} \frac{\partial \mathrm{HSIC}}{\partial \mathbf{u}_{:j}^v}. \tag{A8}$$

We then update $\mathbf{U}^v$ ($v \in [m+1]$) by PGD:

$$\mathbf{U}^v \leftarrow \mathrm{Proj}(\mathbf{U}^{\mathrm{v}} - \eta \frac{\partial \mathcal{L}(\mathbf{U}^{\mathrm{v}})}{\partial \mathbf{U}^{\mathrm{v}}}). \tag{A9}$$

2.1.3 Update α with fixed others.

With other variables fixed, the problem w.r.t. α^v is

$$\min_{\sum \alpha_v = 1, \boldsymbol{\alpha} \geq 0} \sum_{v=1}^{m} \alpha_v^s \mathbf{J}^v, \tag{A10}$$

where $\mathbf{J}^v = ||\mathbf{O}_{\mathbf{X}}^v \odot (\mathbf{X}^v - \mathbf{F}\mathbf{U}^v)||_{2,1}$. Based on its Lagrange function, we have

$$\alpha_v = \frac{(s\mathbf{J}^v)^{\frac{1}{1-s}}}{\sum_{v'=1}^{m} (s\mathbf{J}^{v'})^{\frac{1}{1-s}}}. \tag{A11}$$

2.1.4 Update β with fixed others.

When the others are fixed, the problem w.r.t. $\boldsymbol{\beta}$ becomes

$$\min_{||\boldsymbol{\beta}_v||_2 = 1, \boldsymbol{\beta} \geq 0} \sum_{v=1}^{m+1} \sum_{v' \neq v} \beta_{vv'} \mathrm{HSIC}(\mathbf{U}^{\mathrm{v}}, \mathbf{U}^{\mathrm{v}'}). \tag{A12}$$

According to Cauchy-Schwarz inequality [7], we have

$$\begin{aligned} \sum \beta_{vv'} \mathrm{HSIC}(\mathbf{U}^{\mathrm{v}}, \mathbf{U}^{\mathrm{v}'}) &\leq \sqrt{(\sum (\mathrm{HSIC}(\mathbf{U}^{\mathrm{v}}, \mathbf{U}^{\mathrm{v}'}))^2)(\sum \beta_{\mathrm{vv}'}^2)} \\ &= \sqrt{\sum (\mathrm{HSIC}(\mathbf{U}^{\mathrm{v}}, \mathbf{U}^{\mathrm{v}'}))^2}. \end{aligned} \tag{A13}$$

When the equality in (A13) holds, the closed-form solution of $\beta_{vv'}$ is obtained by

$$\beta_{vv'} = \frac{\mathrm{HSIC}(\mathbf{U}^{\mathrm{v}}, \mathbf{U}^{\mathrm{v}'})}{\sqrt{\sum_{v=1}^{m} (\mathrm{HSIC}(\mathbf{U}^{\mathrm{v}}, \mathbf{U}^{\mathrm{v}'}))^2}}. \tag{A14}$$

Table A1: Statistics of four multi-view multi-label datasets. #Samples is the number of samples; #Views is the number of views; #Features is the number of dimensions of multiple views; #Labels is the number of distinct labels; #Average is the average number of labels per sample.

Datasets	#Samples	#Views	#Features	#Labels	#Average	Domain
Corel5k	4999	6	100/512/1000/4096/4096/4096	260	3.397	image
Pascal07	9963	6	100/512/1000/4096/4096/4096	20	1.465	image
Yeast	2417	2	79/24	14	4.237	biology
Emotions	593	2	64/8	6	1.869	music

2.2 Complexity Analysis

The optimization procedure for solving the optimization problem in (A1) is outlined in Algorithm 1. In terms of computational complexity, updating β_v needs a cost of $\mathcal{O}(m^2k^2d)$, updating $\{\mathbf{U}^v\}_{v=1}^{m+1}$ costs $\mathcal{O}(mkd(n+k))$, and updating α_v and $\mathbf{F}$ cost $\mathcal{O}(mnkd)$, where $k << l$, and $d = min\{d_v, l\}$ $(v \in [m])$. Thus, the total computational complexity at each iteration is $\mathcal{O}(mkd(n+mk))$, which is linear w.r.t. the number of samples and features. In practice, the algorithm typically converges within 100 iterations.

Algorithm 1 The Algorithm of NAIL

```
Require: $\{\mathbf{X}^v\}_{v=1}^m$, $\mathbf{Y}$, $\{\mathbf{O}_\mathbf{X}^v\}_{v=1}^m$, $\mathbf{O}_\mathbf{Y}$, $\lambda$, $\mu$, $k$
Ensure: $\hat{\mathbf{Y}}$
while not converged do
    Update $\mathbf{F}$ according to (A3).
    Update $\{\mathbf{U}^v\}_{v=1}^{m+1}$ according to (A9).
    Update $\alpha$ according to (A11).
    Update $\beta$ according to (A14).
end while
$\hat{\mathbf{Y}} = \sigma(\mathbf{F}\mathbf{U}^{m+1})$.
```

3 More Experiment Results

3.1 Datasets

We conduct a comprehensive experimental study to evaluate the performance of the proposed NAIL on four widely used multi-view multi-label datasets. The statistics of used datasets are summarized in Table A1. Corel5k[1] and Pascal07[1] are image datasets, in which each sample is represented by six feature views. In the Yeast dataset[2], each gene is represented by a genetic expression and a phylogenetic profile. In the Emotions dataset[3], each music is represented by rhythmic and timbre feature views, and classified

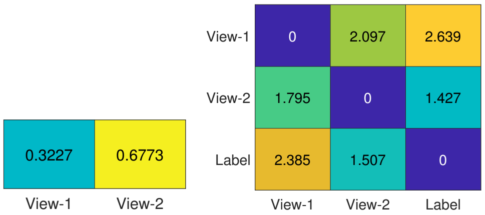
[1] http://lear.inrialpes.fr/people/guillaumin/data.php

[2] http://vlado.fmf.uni-lj.si/pub/networks/data/

[3] http://www.uco.es/kdis/mllresources

Table A2: Experimental results on four real-world datasets at $r\% = 50\%$ and $s\% = 50\%$. The best results are highlighted in boldface, and the second best results are underlined.

		lrMMC		McWL		iMVWL		NAIM^{3}L		**NAIL-L**		**NAIL**	
		Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	**Mean**	STD
Emotions	RS	0.5496	0.0063	0.6517	0.0080	0.6213	0.0136	0.6298	0.0018	**0.6890**	0.0319	0.6869	0.0191
	AUC	0.5373	0.1230	0.6517	0.0080	0.6213	0.0094	0.6792	0.0118	**0.6890**	0.0319	0.6869	0.0011
Yeast	RS	0.7811	0.0003	0.7894	0.0024	0.8072	0.0053	0.7888	0.0168	**0.8107**	0.0049	0.8073	0.0106
	AUC	0.7538	0.0002	0.7894	0.0024	0.8072	0.0042	0.7724	0.0143	**0.8107**	0.0195	0.8073	0.0106
Corel5k	RS	0.7815	0.0170	0.7279	0.0057	0.8473	0.0036	**0.8753**	0.0055	0.7805	0.0599	0.7864	0.0030
	AUC	0.7815	0.0040	0.7341	0.0056	0.8473	0.0147	**0.8753**	0.0164	0.7805	0.0599	0.7865	0.0058
Pascal07	RS	0.7091	0.0001	0.6476	0.0054	0.7609	0.0144	**0.7770**	0.0091	0.7499	0.0175	0.7434	0.0183
	AUC	0.6475	0.0015	0.6476	0.0054	**0.7609**	0.0004	0.7110	0.0199	0.7499	0.0100	0.7435	0.0028



(a) $\boldsymbol{\alpha}$

(b) $\boldsymbol{\beta}$

Figure A1: Heat maps of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of NAIL on the synthetic dataset at $r\% = 50\%$ and $s\% = 50\%$.

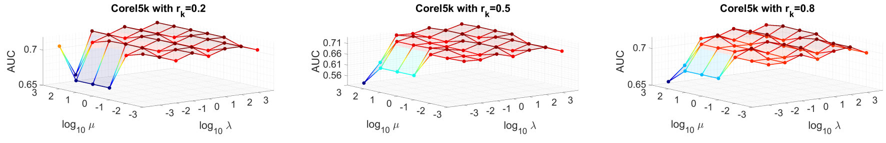
into emotions that it evokes.

3.2 Evaluation of Comparing Methods

Table A2 shows the experimental results of all comparing methods on four real-world datasets at $r\% = 50\%$ and $s\% = 50\%$. From Table A2, we can see that NAIL and NAIL-L outperform comparing methods in most of the cases. The performance superiority probably comes from their ability on handling noisy views and imbalanced labels, and decorrelating weight matrices for redundancy removal in an adaptive way. The incompleteness of multi-view data causes the performance degradation of lrMMC and McWL. As incomplete multi-view weak learning methods, iMVWL and NAIM^{3}L outperform lrMMC and McML in most cases, but perform worse than NAIL and NAIL-L. There are two possible reasons. One is that iMVWL assumes that the label matrix is low-rank, and the other is that both iMVWL and NAIM^{3}L treat multiple views equally. In contrast, NAIL and NAIL-L measure the importance of each view by adaptively choosing appropriate values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

3.3 Case Study on Auto-Weighting

To explore the effects of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we design a synthetic dataset based on the Emotions dataset. We replace the first feature view with random Gaussian noise and add a random portion of Gaussian noise to the second view. We report the learned values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ at $r\% = 50\%$ and $s\% = 50\%$ in Fig A1. As shown in Fig. A1(a), $\boldsymbol{\alpha}$ assigns larger



(a) (b) (c)

Figure A2: Hyperparameter sensitivity analysis of NAIL on the Corel5k dataset.

values to α_2 than to α_1, implying a lower reconstruction error for View-2. From Fig. A1(b), we can see that $\boldsymbol{\beta}$ assigns larger values to $\boldsymbol{\beta}_1 = [\beta_{12}, \beta_{21}, \beta_{13}, \beta_{31}]$, indicating a weaker correlation between the noisy View-1 and either View-2 or Label. Thus, the introduction of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ indeed helps to adaptively weight the importance of multiple views in both reconstruction and HSIC-based decorrelation.

3.4 Sensitivity Analysis

We analyze the sensitivity of NAIL w.r.t. its three hyperparameters λ, μ and k, where λ controls the penalty strength of focal loss, μ controls the penalty strength of auto-weighted HSIC and k controls the subspace dimension. The value of λ is selected from $\{10^i | i = -3, \ldots, 3\}$, and the value of r_k is selected from $\{0.2, 0.5, 0.8\}$. The results in terms of AUC on the Corel5k dataset at $r\% = 50\%$ and $s\% = 50\%$ are reported in Fig. A2. We can see that NAIL achieves relatively stable and good performance when $\lambda \approx 10^{-2}$, $\mu \approx 10^2$ and $r_k = 0.5$, and its performance decreases rapidly when $\mu < 10^{-2}$.

References

[1] Calamai, P.H., Moré, J.J.: Projected gradient methods for linearly constrained problems. Mathematical programming **39**(1), 93–116 (1987)

[2] Goldberg, A., Recht, B., Xu, J., Nowak, R., Zhu, J.: Transduction with matrix completion: Three birds with one stone. Advances in neural information processing systems **23**, 757–765 (2010)

[3] Li, X., Chen, S.: A concise yet effective model for non-aligned incomplete multi-view and missing multi-label learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)

[4] Lin, Y., Gou, Y., Liu, Z., Li, B., Lv, J., Peng, X.: Completer: Incomplete multi-view clustering via contrastive prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11174–11183 (2021)

[5] Liu, J., Wang, C., Gao, J., Han, J.: Multi-view clustering via joint nonnegative matrix factorization. In: Proceedings of the 2013 SIAM international conference on data mining. pp. 252–260. SIAM (2013)

[6] Liu, M., Luo, Y., Tao, D., Xu, C., Wen, Y.: Low-rank multi-view learning in matrix completion for multi-label image classification. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)

[7] Steele, J.M.: The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities. Cambridge University Press (2004)

[8] Tan, Q., Yu, G., Domeniconi, C., Wang, J., Zhang, Z.: Incomplete multi-view weak-label learning. In: IJCAI. pp. 2703–2709 (2018)

[9] Tan, Q., Yu, G., Domeniconi, C., Wang, J., Zhang, Z.: Multi-view weak-label learning based on matrix completion. In: Proceedings of the 2018 SIAM International Conference on Data Mining. pp. 450–458. SIAM (2018)

[10] Wei, T., Guo, L.Z., Li, Y.F., Gao, W.: Learning safe multi-label prediction for weakly labeled data. Machine Learning **107**(4), 703–725 (2018)

[11] Wei, T., Li, Y.F.: Does tail label help for large-scale multi-label learning? IEEE transactions on neural networks and learning systems **31**(7), 2315–2324 (2019)

[12] Xu, C., Tao, D., Xu, C.: Multi-view learning with incomplete views. IEEE Transactions on Image Processing **24**(12), 5812–5825 (2015)

[13] Xu, M., Jin, R., Zhou, Z.H.: Speedup matrix completion with side information: Application to multi-label learning. In: Advances in neural information processing systems. pp. 2301–2309 (2013)

[14] Xu, M., Niu, G., Han, B., Tsang, I.W., Zhou, Z.H., Sugiyama, M.: Matrix co-completion for multi-label classification with missing features and labels. arXiv preprint arXiv:1805.09156 (2018)

[15] Yin, Q., Wu, S., Wang, L.: Unified subspace learning for incomplete and unlabeled multi-view data. Pattern Recognition **67**, 313–327 (2017)

[16] Yu, H.F., Jain, P., Kar, P., Dhillon, I.: Large-scale multi-label learning with missing labels. In: International conference on machine learning. pp. 593–601. PMLR (2014)

[17] Zhang, W., Zhang, K., Gu, P., Xue, X.: Multi-view embedding learning for incompletely labeled data. In: Twenty-Third International Joint Conference on Artificial Intelligence (2013)

[18] Zhu, C., Miao, D., Zhou, R., Wei, L.: Improved multi-view multi-label learning with incomplete views and labels. In: 2019 International Conference on Data Mining Workshops (ICDMW). pp. 689–696. IEEE (2019)