Fast and Robust Multi-View Multi-Task Learning via Group Sparsity

Lu Sun[1] , **Canh Hao Nguyen**[1] , **Hiroshi Mamitsuka**[1,2]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan
[2]Department of Computer Science, Aalto University, Finland
{sunlu, canhhao, mami}@kuicr.kyoto-u.ac.jp

Abstract

Multi-view multi-task learning has recently attracted more and more attentions due to its dual-heterogeneity, i.e., each task has hetergeneous features from multiple views, and probably correlates with other tasks via common views. Existing methods usually suffer from three problems: 1) lack the ability to eliminate noisy features, 2) hold a strict assumption on view consistency and 3) ignore the possible existence of task-view outliers. To overcome these limitations, we propose a robust method with joint group-sparsity by decomposing feature parameters into a sum of two components, in which one saves relevant features (for Problem 1) and flexible view consistency (for Problem 2), while the other detects task-view outliers (for Problem 3). With a global convergence property, we develop a fast algorithm to solve the optimization problem in a linear time complexity w.r.t. the number of features and labeled samples. Extensive experiments on various synthetic and real-world datasets demonstrate its effectiveness.

1 Introduction

In recent years, many real-world applications, such as web page classification, bioinformatics analysis, semantic image annotation and product recommendation, usually exhibit *dual-heterogeneity* [He and Lawrence, 2011], i.e., each task has heterogeneous features from multiple views (*feature heterogeneity*), and multiple tasks are probably correlated via one or more common views (*task heterogeneity*). For example, in web page classification, each web page has at least two views: text and images, and multiple labels, such as politics, science and sports. In image annotation, each image has features extracted from multiple sources, such as color histogram, edge direction and wavelet texture, and is probably annotated with multiple objects, like cat, dog, lion, etc.

In order to handle these heterogeneous datasets, Multi-View Learning (MVL) and Multi-Task Learning (MTL) were proposed for feature heterogeneity and task heterogeneity, respectively. In order to improve the performance of a baseline learner, MVL aims to combine the information from multiple feature views, while MTL is proposed to learn multiple correlated tasks together. Experimental results in a variety of applications have shown the superiority of their performance against baseline learners. However no MVL method have considered heterogeneity of MTL and vice versa, limiting their application on the datasets with dual-heterogeneity.

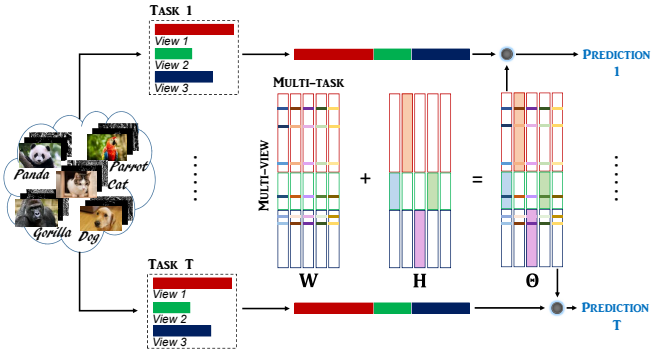


Figure 1: Illustration of the framework of AGILE. Its weight matrix $\mathbf{\Theta}$ is decomposed into two components, $\mathbf{W}$ and $\mathbf{H}$, where $\mathbf{W}$ saves task relationship and view consistency with *row-sparsity*, while $\mathbf{H}$ captures task outliers and inconsistent views with *group-sparsity*. In prediction, the output of each task is calculated by a linear model.

To overcome the limitation of MVL and MTL, recently Multi-View Multi-Task Learning (MVMTL) has been proposed by considering both view consistency and task relationship [He and Lawrence, 2011]. Compared with MVL and MTL methods, it achieved much success on various real-world problems with dual-heterogeneity [Zhang and Huan, 2012; Jin *et al.*, 2013; Lu *et al.*, 2017]. However, there are still three challenges in MVMTL. 1) Feature learning should be applied in the learning phase. Multi-view multi-task datasets usually emerge in high-dimensionality, where only a subset of features, i.e, discriminative features, are useful, and the noisy features might reduce the prediction performance. 2) Modeling flexible view consistency is recently demanded, as it is possible that different views provide complementary information, rather than strictly consistent with each other. One simple example is web page classification, where the visual content is complementary to the text content. 3) The existence of task outliers and inconsistent views should be considered. For example, each of texts and images has unique information in web page classification.

To cope with the aforementioned challenges, we propose a f**A**st and robust method with **G**roup spars**I**ty for multi-view multi-task **LE**arning (**AGILE**), to capture a more flexible relationship among both tasks and views, and meanwhile discard useless features and views. To handle the dual-heterogeneity in a robust manner, the weight matrix $\mathbf{\Theta}$ of AGILE is decomposed into a sum of two group-sparse components: $\mathbf{W}$ and $\mathbf{H}$. Specifically, for Challenges 1 and 2, $\mathbf{W}$ saves task relationship with selected task-common features by $\ell_{2,1}$-norm-based regularization, and models flexible view consistency by regularizing a new norm we call *group trace lasso*. For Challenge 3, $\mathbf{H}$ simultaneously captures task outliers and inconsistent views by penalizing a group lasso regularization. We illustrate the framework of AGILE in Fig. 1. To optimize the convex objective function of AGILE, we propose a fast optimization algorithm with global convergence, which has a linear time complexity w.r.t. the number of features and labeled samples. Empirical results on synthetic and real-world datasets demonstrate the superior performance of AGILE, compared with cutting-edge methods. The contributions of this work can be summarized into three-folds.

- We propose a robust MVMTL method, enabling to cope with dual-heterogeneity, select relevant common features, and capture task outliers and inconsistent views.
- To promote flexible view consistency, we propose the *group trace lasso* norm, which is regarded as a generalization of trace lasso [Grave *et al.*, 2011].
- We develop a fast optimization algorithm to solve the convex objective function, and show that it has a linear time complexity in problem size.

2 Related works

To address the problems with dual-heterogeneity, IteM^2 [He and Lawrence, 2011] was first proposed for MVMTL. IteM^2 constructs a bi-partite graph for each view, and projects any two tasks to a new reproducing kernel Hilbert space based on their common views. However, it is a transductive method, which is designed only for problems with non-negative features, and unable to generate predictive models for future testing samples. To overcome these limitations, inductive learning methods [Zhang and Huan, 2012; Jin *et al.*, 2013] were later developed based on linear models and co-regularization [Sindhwani *et al.*, 2005]. In [Zhang and Huan, 2012], regMVMT was proposed by minimizing the difference of predictive models for distinct tasks on the same view. A more general method, CSL-MVMT [Jin *et al.*, 2013] was developed by assuming that a low-dimensional subspace is shared among related tasks with common views. But the two methods hold a strong assumption on view consistency, i.e., predictive models from different views should be consistent with each other on the unlabeled data, which probably violates the problem setting of various real-world problems. Inspired by linear discriminant analysis, MAMUDA [Jin *et al.*, 2014] was proposed as a supervised feature extraction method to handle dual-heterogeneity via shared latent spaces among multiple views. Based on multilinear factorization machines, MFM was recently proposed in [Lu *et al.*, 2017] by learning both task-specific feature map and task-view shared multilinear structures. However, MFM is unable to directly capture inconsistent views and utilize a possible large amount of unlabeled data. In [Zhou *et al.*, 2018], MTMVL is treated as a multi-objective optimization problem by integrating relationships among tasks, views and samples.

Existing MVMTL methods typically build the models on original features, however, irrelevant features would negatively influence the prediction performance. Besides, they usually assume that predictive models from multiple views are consistent, and multiple tasks are correlated, which probably contradicts real-world applications.

3 The AGILE method

3.1 Preliminary

For MVMTL, suppose that we have T tasks, and each task has V views. For the t-th task, let $\mathbf{X}_t = [\mathbf{X}_t^1, ..., \mathbf{X}_t^V] \in \mathbb{R}^{n_t \times d}$ and $\mathbf{U}_t \in \mathbb{R}^{m_t \times d}$ be the labeled and unlabeled data matrices, respectively, where $\mathbf{X}_t^v \in \mathbb{R}^{n_t \times d_v}$ is the v-th view data in the t-th task, $\forall t, v$. Each data matrix $\mathbf{X}_t$ is associated with a target vector $\mathbf{y}_t \in \mathbb{R}^{n_t}$, $\forall t$. Without loss of generality, for an arbitrary matrix $\mathbf{A}$, $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_*$ denote its Frobenius norm and trace norm, respectively. Additionally, we define $\ell_{2,1}$-norm by $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{p} \|\mathbf{a}_{i\cdot}\|_2$, where $\|\mathbf{a}_{i\cdot}\|_2$ is the ℓ_2-norm of the i-th row $\mathbf{a}_{i\cdot}$ of $\mathbf{A}$.

For the t-th task of a MVMTL problem, we consider a linear predictive function defined by

$$\mathbf{y}_t \approx \frac{1}{V} \sum_{v=1}^{V} \mathbf{X}_t^v \boldsymbol{\theta}_t^v = \frac{1}{V} \mathbf{X}_t \boldsymbol{\theta}_t^\top, \tag{1}$$

where $\boldsymbol{\theta}_t = [\boldsymbol{\theta}_t^1; ...; \boldsymbol{\theta}_t^V] \in \mathbb{R}^d$ with $\boldsymbol{\theta}_t^v \in \mathbb{R}^{d_v}$ denoting the weight vector for the v-th view of the t-th task, $\forall t, v$. Eq. (1) actually averages the prediction results from all the V views in the t-th task, $\forall t$. For simplicity, we omit the intercept in the linear model by assuming that samples and targets have been centered in column-wise.

3.2 Formulation

For robust MVMTL, we decompose the weight matrix $\mathbf{\Theta} = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T] \in \mathbb{R}^{d \times T}$ into two components $\mathbf{W}$ and $\mathbf{H}$. It allows to model correlations among tasks and views by $\mathbf{W}$, and capture tasks/views that are specific to some views/tasks by $\mathbf{H}$, which cannot be fit into $\mathbf{W}$. Thus, based on (1) and empirical risk minimization with squared ℓ_2-norm loss, the optimization problem of AGILE is

$$\min_{\mathbf{\Theta}=\mathbf{W}+\mathbf{H}} \sum_{t=1}^{T} \frac{1}{2} \left\| \mathbf{y}_t - \frac{1}{V} \mathbf{X}_t \boldsymbol{\theta}_t \right\|_2^2 + R(\mathbf{W}) + R(\mathbf{H}), \tag{2}$$

where $R(\mathbf{W})$ and $R(\mathbf{H})$ denote the regularization terms on $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_T]$ and $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_T]$, respectively. $R(\mathbf{W})$ aims to cope with dual-heterogeneity, while $R(\mathbf{H})$ is designed to detect task outliers and inconsistent views.

For dual-heterogeneity, we decompose $R(\mathbf{W})$ into a sum of two parts, so as to capture both task relationship and view consistency. In terms of task relationship, we consider that multiple tasks are correlated based on an identical feature

subset, i.e., task-common features. In this respect, the $\ell_{2,1}$-norm [Liu *et al.*, 2009] is imposed on $\mathbf{W}$ to promote its row-sparsity (feature-wise group sparsity). Since the $\ell_{2,1}$-norm regularizer $\|\mathbf{W}\|_{2,1}$ encourages sparsity among features and non-sparsity among tasks, discriminative task-common features will be assigned larger values. In this way, irrelevant features are removed by relying on the complementary information from multiple views, and task relationship is modeled via the non-zero rows (common features) of $\mathbf{W}$.

In terms of view consistency, the independence assumption [Belkin *et al.*, 2006] of multiple views has been widely adopted by MVL methods [Blum and Mitchell, 1998; Sindhwani *et al.*, 2005], indicating that predictive models from multiple views can achieve mutual agreement on the unlabeled data. On some practical applications, however, it is possible that a few views become useless for certain tasks due to noise pollution, and the assumption is violated. To address this challenge, for the t-th task, we impose a novel *group trace lasso* norm $\|\mathbf{U}_t\mathcal{W}_t\|_*$, in which $\mathcal{W}_t$ is a special diagonal matrix whose diagonal elements are vectors $\{\mathbf{w}_t^v\}_v$,

$$\mathcal{W}_t = \begin{bmatrix} \mathbf{w}_t^1 & & \\ & \ddots & \\ & & \mathbf{w}_t^V \end{bmatrix} \in \mathbb{R}^{d\times V}, \tag{3}$$

with $\mathbf{w}_t^v \in \mathbb{R}^{d_v}$ denoting the v-th view weight vector of $\mathbf{w}_t$. In group trace lasso, $\mathbf{U}_t\mathcal{W}_t$ is a matrix containing the predictions from multiple views (each in a column) on the unlabeled data $\mathbf{U}_t$. Hence, its low-rankness imposes the relationships among views in the multi-view setting. If predictive models are restrictively consistent among multiple views, the rank of $\mathbf{U}_t\mathcal{W}_t$ equals to 1. In addition, as shown in Sec. 3.3, the regularizer promotes group-sparsity among views, so that consistent views will get larger values while useless views are discarded by $\mathbf{W}$. Therefore, we formulate $R(\mathbf{W})$ by

$$R(\mathbf{W}) \coloneqq \alpha\|\mathbf{W}\|_{2,1} + \beta\sum_{t=1}^{T}\|\mathbf{U}_t\mathcal{W}_t\|_*, \tag{4}$$

where α and β are positive hyperparameters. In this way, $\mathbf{W}$ models task relationship via common features, promotes flexible view consistency and discards useless views.

For robustness, we expect to capture both task outliers, that have specific supporting features, and inconsistent views, that are inconsistent to other views but provide complementary information. Thus, a group-sparsity inducing regularizer $R(\mathbf{H})$ is imposed on $\mathbf{H}$ by treating each task-view pair as a group. Specifically, for the v-th view, view-specific task outliers are captured by regularizing $\sum_{t=1}^{T}\|\mathbf{h}_t^v\|_2$, where $\mathbf{h}_t^v \in \mathbb{R}^{d_v}$ denotes the weight vector of the v-th view of $\mathbf{h}_t$. Similarly, for the t-th task, task-specific inconsistent views are detected by regularizing $\sum_{v=1}^{V}\|\mathbf{h}_t^v\|_2$. By summing over V views and T tasks on these two terms, we have

$$R(\mathbf{H}) \coloneqq \gamma\sum_{t=1}^{T}\sum_{v=1}^{V}\|\mathbf{h}_t^v\|_2 = \gamma\|\mathbf{H}\|_{G_1}, \tag{5}$$

where $\|\cdot\|_{G_1}$ denotes the Group ℓ_1-norm of groups $\{\mathbf{h}_t^v\}_{t,v}$, and γ is a positive hyperparameter. Once the t-th task is a outlier in the v-th view, or the v-th view is inconsistent in the t-th task, corresponding $\mathbf{h}_t^v$ would take arbitrary values, otherwise zero.

Based on (2), (4) and (5), we now have the optimization problem for the proposed AGILE method:

$$\min_{\mathbf{\Theta}=\mathbf{W}+\mathbf{H}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\theta}\|_2^2 + \alpha\|\mathbf{W}\|_{2,1} + \beta\|\mathbf{U}\mathcal{W}\|_* + \gamma\|\mathbf{H}\|_{G_1}, \tag{6}$$

where $\mathbf{X} \in \mathbb{R}^{\sum_t n_t \times dT}$, $\mathbf{U} \in \mathbb{R}^{\sum_t m_t \times dT}$ and $\mathcal{W} \in \mathbb{R}^{dT\times VT}$ are block diagonal matrices with the t-th blocks being $\frac{1}{V}\mathbf{X}_t$, $\mathbf{U}_t$ and $\mathcal{W}_t$, respectively, and $\mathbf{y} = [\mathbf{y}_1; ...; \mathbf{y}_T] \in \mathbb{R}^{\sum_t n_t}$. In (6), $\boldsymbol{\theta} \in \mathbb{R}^{dT}$ denotes a vectorization of the weight matrix $\mathbf{\Theta}$. Note that, the equation $\sum_{t=1}^{T}\|\mathbf{U}_t\mathcal{W}_t\|_* = \|\mathbf{U}\mathcal{W}\|_*$ holds based on Lemma 2.3 in [Recht *et al.*, 2010].

3.3 Remarks on group trace lasso

To model view consistency, some MVMTL methods [Zhang and Huan, 2012; Jin *et al.*, 2013] adopt *co-regularization* [Sindhwani *et al.*, 2005]. In contrast, AGILE applies a novel regularizer $\|\mathbf{U}\mathcal{W}\|_*$ in (6) for saving flexible consistency. Following lemmas[1] present its important properties.

Lemma 1. *For the t-th task, if predictions from multiple views are consistent, the following inequality holds,*

$$\|\mathbf{U}_t\mathcal{W}_t\|_* = \|\mathbf{U}_t\mathbf{w}_t\|_2 \le \|\mathbf{U}_t\|_{op}\|\mathbf{w}_t\|_2, \tag{7}$$

where $\|\cdot\|_{op}$ denotes operator norm. If prediction from multiple views are orthogonal, the following inequality holds,

$$\|\mathbf{U}_t\mathcal{W}_t\|_* = \|\mathbf{U}_t\mathbf{w}_t\|_{G_1} \le \sum_{v=1}^{V}\|\mathbf{U}_t^v\|_{op}\|\mathbf{w}_t^v\|_2, \tag{8}$$

where $\|\mathbf{U}_t\mathbf{w}_t\|_{G_1} = \sum_{v=1}^{V}\|\mathbf{U}_t^v\mathbf{w}_t^v\|_2$ is Group ℓ_1-norm.

Lemma 2. *For the t-th task, the following inequality holds,*

$$\|\mathbf{U}_t\mathbf{w}_t\|_2 \le \|\mathbf{U}_t\mathcal{W}_t\|_* \le \|\mathbf{U}_t\mathbf{w}_t\|_{G_1}. \tag{9}$$

Lemmas 1 and 2 tell us if predictions from two views are strongly consistent, the proposed norm behaves like ℓ_2-norm on these two predictive models; if predictions from two views are orthogonal, the norm equals to the group ℓ_1 norm, imposing group sparsity on $\{\mathbf{w}_t^v\}_{v=1}^V$. Thus, it enables to equally penalize the two consistent views, and discard the useless inconsistent views. In this sense, this method lies in between the usual strict assumption of "all views contain 'exactly' the same information" and feature concatenation, which imposes no relationship among the weights.

Note that the proposed norm is similar with the trace lasso [Grave *et al.*, 2011], since both impose a trace norm regularizer on the product of a data matrix and a diagonal matrix of model parameters. The difference exists in the diagonal matrix, where each diagonal element is a parameter in trace lasso, and a group (view) of parameters in the proposed norm. In other words, the proposed norm can be regarded as *group trace lasso*, as trace lasso is its special case once the number of parameters in each group is restricted to be one.

[1]The proofs are provided in supplementary materials: https://www.dropbox.com/s/5koj44hmx4qb504/AGILE_sup.pdf?dl=0.

4 Optimization algorithm

Since the objective function (6) is jointly convex w.r.t. $\mathbf{W}$ and $\mathbf{H}$, the global optimum is obtained by alternatively updating $\mathbf{W}$ and $\mathbf{H}$. The trace norm in (6) is non-smooth and involves the product of $\mathbf{U}$ and $\mathcal{W}$, thus it is impossible to directly employ the proximal gradient method. To circumvent this difficulty, we apply ADMM [Boyd *et al.*, 2011] to optimize (6) by introducing an auxiliary variable $\mathbf{P} = \mathbf{U}\mathcal{W}$, leading to the augmented Lagrangian defined as follows,

$$\mathcal{L}(\mathbf{\Theta}, \mathbf{P}, \mathbf{Q}) = \frac{1}{2}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right\|_2^2 + \alpha\left\|\mathbf{W}\right\|_{2,1} + \beta\left\|\mathbf{P}\right\|_* + \frac{1}{2}\left\|\mathbf{U}\mathcal{W} - \mathbf{P} + \mathbf{Q}\right\|_F^2 + \gamma\left\|\mathbf{H}\right\|_{G_1}, \quad (10)$$

where $\mathbf{Q}$ is the scaled dual variable. The algorithm on solving (10) repeats the following three steps until convergence:

(a) $\mathbf{\Theta}^* \leftarrow \min_{\mathbf{\Theta}} \frac{1}{2}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right\|_2^2 + \frac{1}{2}\left\|\mathbf{U}\mathcal{W} - \mathbf{P} + \mathbf{Q}\right\|_F^2 + \alpha\left\|\mathbf{W}\right\|_{2,1} + \gamma\left\|\mathbf{H}\right\|_{G_1}$,

(b) $\mathbf{P}^* \leftarrow \min_{\mathbf{P}} \frac{1}{2}\left\|\mathbf{U}\mathcal{W} - \mathbf{P} + \mathbf{Q}\right\|_F^2 + \beta\left\|\mathbf{P}\right\|_*$,

(c) $\mathbf{Q}^* \leftarrow \mathbf{Q} + \mathbf{U}\mathcal{W} - \mathbf{P}$.

For Problem (a), we propose to solve it by proximal method [Nesterov, 2013], which decomposes its objective into two components, smooth $f(\mathbf{\Theta})$ and non-smooth $g(\mathbf{\Theta})$:

$$f(\mathbf{\Theta}) = \frac{1}{2}\left\|\mathbf{y} - \mathbf{X}(\mathbf{w} + \mathbf{h})\right\|_2^2 + \frac{1}{2}\left\|\mathbf{U}\mathcal{W} - \mathbf{P} + \mathbf{Q}\right\|_2^2,$$
$$g(\mathbf{\Theta}) = \alpha\left\|\mathbf{W}\right\|_{2,1} + \gamma\left\|\mathbf{H}\right\|_{G_1}. \quad (11)$$

In the k-th iteration, the optimal solution can be calculated by the follow two proximal operators:

$$\min_{\mathbf{W}} \frac{1}{2\eta}\left\|\mathbf{W} - (\,\mathbf{W}^k - \eta\nabla_w f(\mathbf{\Theta}^k)\,)\right\|_2^2 + \alpha\left\|\mathbf{W}\right\|_{2,1}, \quad (12)$$

$$\min_{\mathbf{H}} \frac{1}{2\eta}\left\|\mathbf{H} - (\,\mathbf{H}^k - \eta\nabla_h f(\mathbf{\Theta}^k)\,)\right\|_2^2 + \gamma\left\|\mathbf{H}\right\|_{G_1}, \quad (13)$$

where $\nabla_w f$ and $\nabla_h f$ denote the derivatives of $f(\mathbf{\Theta})$ w.r.t $\mathbf{W}$ and $\mathbf{H}$, respectively, and η is a step size, which satisfies

$$f(\mathbf{\Theta}^{k+1}) \le f(\mathbf{\Theta}^k) + \langle\nabla_w f(\mathbf{\Theta}^k), \mathbf{W}^{k+1} - \mathbf{W}^k\rangle + \langle\nabla_h f(\mathbf{\Theta}^k), \mathbf{H}^{k+1} - \mathbf{H}^k\rangle + \frac{1}{2\eta}\left\|\mathbf{\Theta}^{k+1} - \mathbf{\Theta}^k\right\|_2^2. \quad (14)$$

The optimization problems (12) and (13) can be analytically solved by applying soft thresholding operations as summarized in [Liu *et al.*, 2009] and Lemma 3, respectively.

Lemma 3. *If $\mathbf{H}^*$ is the optimal solution of the optimization problem (12), its t-th column $\mathbf{h}_t^*$ is given by the proximal operator, $\mathcal{S}_{\eta\gamma}(\mathbf{r}_t) = [S_{\eta\gamma}(\mathbf{r}_t^1); ...; S_{\eta\gamma}(\mathbf{r}_t^V)]$, where*

$$S_{\eta\gamma}(\mathbf{r}_t^v) = \begin{cases} \mathbf{r}_t^v\left(1 - \frac{\eta\gamma}{\|\mathbf{r}_t^v\|_2}\right) & \|\mathbf{r}_t^v\|_2 > \eta\gamma \\ 0 & 0 \le \|\mathbf{r}_t^v\|_2 \le \eta\gamma, \end{cases} \quad (15)$$

and $\mathbf{r}_t$ denotes the t-th column of $\mathbf{H}^k - \eta\nabla_h f(\mathbf{\Theta}^k)$.

For Problem (b), it admits a closed-form solution by applying a matrix soft thresholding operation [Cai *et al.*, 2010]. It is worth noting that, based on Lemma 2.3 in [Recht *et al.*, 2010] and the fact that $\mathbf{U}\mathcal{W}$ is a diagonal block matrix, Problem (b) can be decoupled for T tasks and solved efficiently.

Table 1: The statistics of used four real-world dual-heterogeneous datasets. Here V, T and d denote the number of views, tasks and features, respectively, and n_p and n_n are the number of positive and negative samples in each task, respectively.

Dataset	V	T	d	n_p	n_n	Domain
FOX	2	4	3,458	178∼635	888∼1,345	Text categorization
Mirflickr	2	8	193	668∼7,849	3,413∼10,594	Image annotation
NUS-Object	5	7	634	964∼2,370	8,000∼9,406	Image annotation
NUS-Scene	5	15	629	1,039∼11,995	4,409∼15,365	Image annotation

4.1 Analysis on time complexity

In practice, we apply fast-ADMM [Goldstein *et al.*, 2014] and accelerated proximal method [Nesterov, 2013] to accelerate the optimization algorithm[2], leading to an optimal convergence rate of $\mathcal{O}(\frac{1}{k^2})$ with k being the number of iterations.

In time complexity, for updating $\mathbf{\Theta}$, gradient calculation ($\nabla_w f$ and $\nabla_h f$) and soft thresholding have time complexities of $\mathcal{O}(d\sum_t(n_t + Vm_t))$ and $\mathcal{O}(dT)$, respectively. On updating $\mathbf{P}$, matrix soft thresholding needs to conduct SVD on $\mathbf{U}\mathcal{W}$, leading to a time complexity of $\mathcal{O}(V\sum_t m_t^2)$, provided that $V \ll m_t, \forall t$. Therefore, the total time complexity of each iteration is $\mathcal{O}(d(\sum_t n_t + T) + V\sum_t m_t^2)$, in a linear complexity w.r.t. the number of features and labeled samples.

5 Experiments

5.1 Experimental setting

Synthetic data

For the generation of synthetic datasets, the number of tasks and views are set as $T = 12$ and $V = 6$, respectively, and each task has $n = 150$ labeled and $m = 150$ unlabeled data. The dimensionality d_v of each view is selected from 30 to 60 by step 6, with $d = \sum_v d_v = 300$. The data matrix $\mathbf{X}^1 \in \mathbb{R}^{n\times d_1}$ and the weight matrix $\mathbf{W}^1 \in \mathbb{R}^{d_1\times T}$ from the 1-st view are randomly sampled from normal distributions $\mathcal{N}(0, 25)$ and $\mathcal{N}(0, 16)$, respectively. In $\mathbf{W}^1$, top 30% of the features are treated as useful features by setting 70% of rows to $\mathbf{0}$. To keep view consistency, $\mathbf{X}^v \in \mathbb{R}^{d_v\times T}$ and $\mathbf{W}^v \in \mathbb{R}^{d_v\times T}$ ($v > 1$) are generated by $\mathbf{X}^v = \mathbf{X}^1\mathbf{P}^{v\top}$ and $\mathbf{W}^v = \mathbf{P}^v\mathbf{W}^1$, respectively, with constraints $\mathbf{P}^{v\top}\mathbf{P}^v = \mathbf{I}_{d_1}$ and $\mathbf{P}^v \in \mathbb{R}^{d_v\times d_1}$, whose columns are left-singular vectors of a matrix randomly sampled from uniform distribution $\mathcal{U}(0, 1)$. Another weight matrix $\mathbf{H} \in \mathbb{R}^{d\times T}$ is randomly sampled from $\mathcal{N}(0, 16)$, where a certain set of $\{\mathbf{h}_t^v\}_{t,v}$ is treated as outliers by assigning $\mathbf{0}$ to the rest. Finally, the target $\mathbf{y}_t$ is calculated by $\mathbf{y}_t = \frac{1}{V}\mathbf{X}_t(\mathbf{w}_t + \mathbf{h}_t) + \boldsymbol{\delta}_t$, with $\boldsymbol{\delta}_t \sim \mathcal{N}(0, 1)$ being stochastic noise.

Real-world data

We conduct experiments on four real-world heterogeneous datasets: FOX, Mirflickr, NUS-Scene and NUS-Object. The FOX dataset is extracted from FOX web news [Qian and Zhai, 2014], while Mirflickr, NUS-Scene and NUS-Object refer to image annotation problem [Huiskes and Lew, 2008; Chua *et al.*, 2009]. The statistics are summarized in Table 1, and more details are presented in supplementary materials.

[2]The procedure of the algorithm is provided in the supplement.

Comparing methods

We compare AGILE[3] with six cutting-edge methods: Elastic-Net [Zou and Hastie, 2005], rMTFL [Gong *et al.*, 2012], coMVL [Sindhwani *et al.*, 2005], IteM^2 [He and Lawrence, 2011], CSL-MVMT [Jin *et al.*, 2013] and MFM [Lu *et al.*, 2017]. As a generalization of ridge regression [Hoerl and Kennard, 1970] and lasso [Tibshirani, 1996], Elastic-Net is selected as a baseline method, while rMTFL and coMVL are selected as representative MTL and MVL methods, respectively. As state-of-the-art MVMTL methods, IteM^2, CSL-MVMT and MFM[4] are introduced in comparison, and the codes are provided by corresponding authors.

Configuration

For evaluation, in each task, we randomly select $a\%$, $a\%$, 20% and 20% of total samples as labeled training set, unlabeled training set, validation set and testing set, respectively, and a is selected from $\{10, 20, 30\}$. We repeat this process five times, and report average results with standard deviation. In parameter setting, the weight balancing ℓ_1 and ℓ_2 regularizers in Elastic-Net is selected from $\{0.2, 0.4, 0.6, 0.8, 1\}$. As recommended in original papers, the dimensionality of latent space in CSL-MVMT and MFM is set as 20. Values for other parameters are selected from $\{10^a \mid |a| \in [3]\}$. For each iterative algorithm, we terminate it once the relative change of its objective is below 10^{-5}, and set the maximum number of iterations as 500. The performances of comparing methods are evaluated by Area Under ROC-Curve (AUC) and Accuracy.

5.2 Experiments on synthetic data

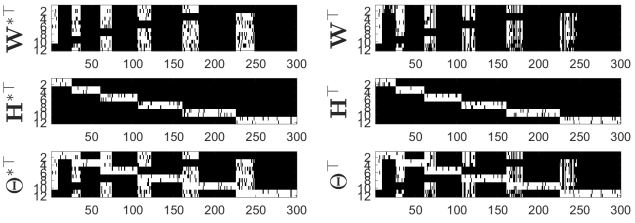
Illustration of weight matrix decomposition

Weight matrix decomposition of AGILE on one designed synthetic dataset is illustrated in Fig 2, where $\mathbf{\Theta}^* = \mathbf{W}^* + \mathbf{H}^*$ denotes the design truth model, and $\mathbf{\Theta} = \mathbf{W} + \mathbf{H}$ is learned by AGILE with the setting $\alpha = 10$, $\beta = 1$ and $\gamma = 36$. As shown in Fig. 2, AGILE successfully recovers the group-sparse pattern in $\mathbf{\Theta}^*$ by assigning group-sparsity to $\mathbf{W}$ and $\mathbf{H}$. To validate group trace lasso, for each task we select one view as the useless view by assigning noise and $\mathbf{0}$ to corresponding $\mathbf{X}_t^v$ and $\mathbf{w}_t^v$, respectively, producing a feature-wise and view-wise group-sparse $\mathbf{W}^*$ in Fig. 2(a). As shown in Fig. 2(b), $\mathbf{W}$ detects not only noise features (row-sparsity) but also useless views (group-sparsity).

Comparison of MVMTL methods

Next, we analyze the robustness of AGILE. In this experiment, we generate nine synthetic datasets by changing the percentage of task-view outliers in $\mathbf{H}$ from 10% to 90% by step 10%. Fig. 3 shows the comparison results of AGILE and three MVMTL methods in AUC and Accuracy, where AGILE consistently achieves superior performance.

To evaluate MVMTL methods in running time, we generate two sets of synthetic datasets by varying the number of training samples and features from 100 to 1000 by step 100, respectively. Fig. 4 shows results of AGILE and three MVMTL methods in running time on the two sets of datasets. AGILE and IteM^2 consume the least running time as increasing the data size. In contrast, the running time of MFM and CSL-MVMT significantly increase as the number of training samples and features increases, respectively.



(a) Designed model $\mathbf{\Theta}^*$ (b) Learned model $\mathbf{\Theta}$

Figure 2: Illustration of weight matrix decomposition of AGILE on the synthetic dataset. **Left**: designed model $\mathbf{\Theta}^* = \mathbf{W}^* + \mathbf{H}^*$; **Right**: learned model $\mathbf{\Theta} = \mathbf{W} + \mathbf{H}$ by AGILE. White (black) color indicates non-zero (zero) values in magnitude.

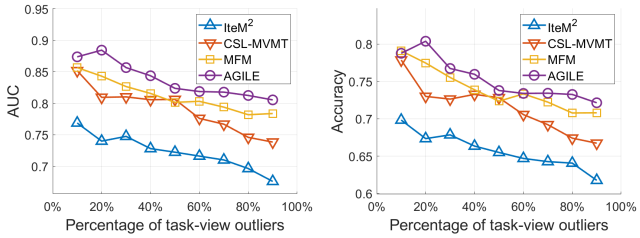


Figure 3: Robustness analysis of AGILE and three MVMTL methods on nine synthetic datasets, which are generated by varying the percentage of task-view outliers from 10% to 90% by step 10%.

5.3 Experiments on real-world data

Evaluation of comparing methods

To evaluate the performances of comparing methods, an experiment on four real-world heterogeneous datasets is conducted. The ratio n/N of labeled training samples is selected from the set $\{10\%, 20\%, 30\%\}$. Experimental results are reported in Table 2, where the best performance is highlighted in boldface. In Table 2, as n/N increases, the performances of comparing methods increase. AGILE obtains the best performance in 75% of total cases. This performance superiority probably comes from AGILE's ability on capturing task-view outliers and removing useless features. Another MVMTL method, CSL-MVMT, performs the second best among MVMTL methods, and competes with comparing methods except AGILE and rMTFL, indicating the importance on modeling dual-heterogeneity in real-world datasets. By handling only task-heterogeneity, rMTFL performs better or comparable compared with the baseline Elastic-Net, especially on FOX, indicating that there probably is a strong task-heterogeneity in the dataset. IteM^2 performs worst on the used datasets, and one possible explanation is that it is originally designed for problems with non-negative features.

[3]We provide the MATLAB code of AGILE at: https://www.dropbox.com/s/f6pvpy6umagwa1b/AGILE.zip?dl=0

[4]MFM-F-S in [Lu *et al.*, 2017] is used here for its best performance among the variants of MFM.

Table 2: Experimental results on real-world datasets by selecting the percentage n/N of labeled data from $\{10\%, 20\%, 30\%\}$.

		AUC							Accuracy						
Dataset	n/N	Elastic-Net	rMTFL	coMVL	IteM^2	CSL-MVMT	MFM	AGILE	Elastic-Net	rMTFL	coMVL	IteM^2	CSL-MVMT	MFM	AGILE
FOX	10%	.967±.002	.976±.001	.965±.002	.814±.004	.972±.005	.923±.012	**.985±.002**	.874±.004	.946±.001	.825±.003	.855±.002	.942±.002	.763±.020	**.956±.002**
	20%	.969±.002	.978±.001	.967±.003	.808±.009	.973±.003	.934±.011	**.988±.002**	.892±.003	.951±.001	.850±.004	.856±.006	.944±.002	.770±.019	**.963±.003**
	30%	.971±.001	.980±.002	.970±.001	.816±.006	.976±.003	.929±.010	**.991±.001**	.898±.003	.951±.003	.856±.003	.855±.002	.948±.004	.736±.013	**.969±.001**
Mirflickr	10%	.573±.002	.621±.001	.618±.002	.519±.002	.633±.003	.618±.003	**.645±.002**	.731±.001	.759±.001	.759±.001	.704±.001	.754±.004	**.766±.004**	.728±.006
	20%	.579±.002	.630±.001	.629±.002	.518±.001	.633±.002	.615±.002	**.635±.008**	.741±.006	.761±.001	.761±.001	.705±.000	.755±.004	**.767±.004**	.717±.012
	30%	**.655±.002**	.630±.002	.629±.001	.510±.002	.636±.004	.620±.006	.640±.002	**.796±.001**	.763±.001	.764±.001	.701±.002	.757±.005	.770±.003	.710±.027
NUS-Object	10%	.851±.003	.853±.005	.877±.001	.565±.004	.858±.001	.836±.002	**.884±.001**	.874±.002	.875±.003	.882±.001	.746±.001	.847±.003	.857±.001	**.888±.000**
	20%	.857±.001	.862±.004	.853±.002	.564±.003	.860±.001	.848±.001	**.871±.006**	.878±.001	.880±.001	.877±.001	.745±.000	.850±.002	.866±.002	**.889±.002**
	30%	.864±.003	.865±.002	.860±.002	.566±.003	.862±.002	.856±.003	**.874±.003**	.881±.001	.881±.000	.881±.001	.746±.001	.848±.001	.871±.001	**.890±.001**
NUS-Scene	10%	.744±.010	.724±.006	**.777±.001**	.646±.002	.756±.001	.744±.005	.761±.005	.838±.003	.833±.002	**.848±.000**	.712±.001	.839±.003	.820±.003	.822±.006
	20%	.747±.001	.753±.003	.745±.001	.676±.009	.760±.007	.745±.001	**.761±.010**	.841±.001	**.843±.000**	.840±.000	.739±.006	.842±.007	.840±.006	**.843±.009**
	30%	.769±.001	.770±.001	.767±.001	.696±.012	.769±.003	.767±.005	**.772±.007**	.842±.001	.844±.002	.843±.001	.743±.007	.844±.005	.844±.008	**.845±.005**

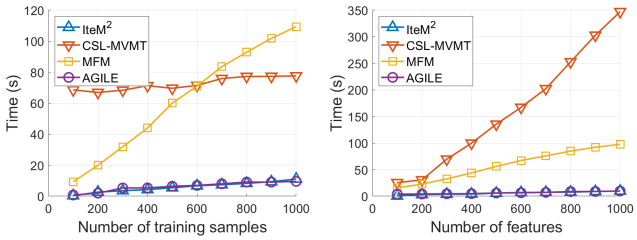


Figure 4: Comparison of methods in running time on two sets of synthetic datasets, which are generated by varying the numbers of training samples (**Left**) and features (**Right**) from 100 to 1000.

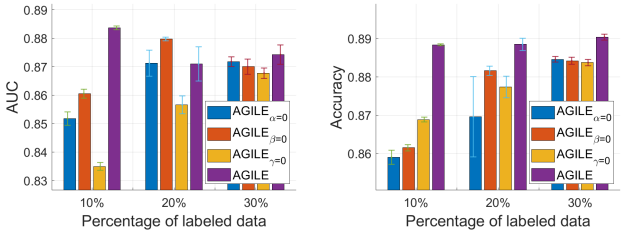


Figure 5: Analysis on the effect of regularizers in AGILE on NUS-Object. AGILE$_{\alpha=0}$, AGILE$_{\beta=0}$ and AGILE$_{\gamma=0}$ denote three variants of AGILE by setting $\alpha = 0$, $\beta = 0$ and $\gamma = 0$, respectively.

Analysis on the effect of regularizations

To evaluate the effect of regularizations used in AGILE, an experiment is performed by assigning 0 to each of three parameters, respectively, and evaluation results on NUS-Object are shown in Fig. 5. As shown in Fig. 5, setting $\gamma = 0$ results in the largest performance loss in four out of six cases, demonstrating the importance of capturing task-view outliers on improving performance. Although setting $\beta = 0$ gives the least performance loss in half of cases, there is still a significant loss compared with the original AGILE.

Hyperparameter sensitivity analysis

The sensitivity of AGILE in three regularization parameters α, β and γ is investigated on NUS-Object. Specifically, α controls the row-sparsity of $\mathbf{W}$, β measures the degree of view consistency and γ controls the group-sparsity of $\mathbf{H}$. Values of the parameters are selected from the set

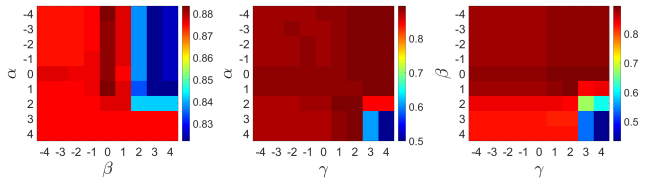


Figure 6: Sensitivity analysis of α, β and γ in AUC on NUS-Object. Parameter values are shown in the logarithmic scale.

$\{10^a \mid |a| \in [4]\}$. Fig. 6 shows the experimental results in AUC with $n/N = 30\%$. The subfigure on α and β is shown by fixing $\gamma = 1$, and the similar setting is used for the other two subfigures. As shown in Fig. 6, AGILE achieves its best performance on NUS-Object with $\alpha \leq 10^1$, $\beta = 10^0$ and $10^1 \leq \gamma \leq 10^2$. Generally, it is recommended to assign smaller values to α and β, and a relatively larger value to γ.

6 Conclusion

In this paper, we propose a fast and robust MVMTL method, AGILE, by decomposing the weight matrix into two components, and adopting a joint regularization to promote group-sparsity. To select relevant features and model flexible view consistency, joint group-sparsity is imposed on the first component by $\ell_{2,1}$-norm and group trace lasso. To detect task outliers and inconsistent views, the second component is regularized by group ℓ_1-norm to encourage its group sparsity. Thanks to the convexity of the objective function, a fast iterative algorithm is developed to solve it with global convergence. Experiments on both synthetic and real-world datasets demonstrate its effectiveness. It shows that in real MVMTL setting, data exhibit all the problems, like existence of noisy features, useless views and task-view outliers, and our method is flexible enough to handle it.

Acknowledgements

L. S. has been supported in part by JST ACCEL (grant number JPMJAC1503). C. H. N. has been supported in part by MEXT Kakenhi 18K11434. H. M. has been supported in part by JST ACCEL (grant number JPMJAC1503), MEXT Kakenhi (grant numbers 16H02868 and 19H04169), FiDiPro by Tekes (currently Business Finland) and AIPSE program by Academy of Finland.

References

[Belkin *et al.*, 2006] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[Boyd *et al.*, 2011] S Boyd, N Parikh, B Peleato E Chu, and J Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[Cai *et al.*, 2010] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., 2009.

[Goldstein *et al.*, 2014] Tom Goldstein, Brendan O'Donoghue, Simon Setzer, and Richard Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

[Gong *et al.*, 2012] Pinghua Gong, Jieping Ye, and Changshui Zhang. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 895–903. ACM, 2012.

[Grave *et al.*, 2011] Edouard Grave, Guillaume R Obozinski, and Francis R Bach. Trace lasso: a trace norm regularization for correlated designs. In *Advances in Neural Information Processing Systems*, pages 2187–2195, 2011.

[He and Lawrence, 2011] Jingrui He and Rick Lawrence. A graph-based framework for multi-task multi-view learning. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 25–32, 2011.

[Hoerl and Kennard, 1970] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[Huiskes and Lew, 2008] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[Jin *et al.*, 2013] Xin Jin, Fuzhen Zhuang, Shuhui Wang, Qing He, and Zhongzhi Shi. Shared Structure Learning for Multiple Tasks with Multiple Views. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 353–368, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[Jin *et al.*, 2014] Xin Jin, Fuzhen Zhuang, Hui Xiong, Changying Du, Ping Luo, and Qing He. Multi-task Multi-view Learning for Heterogeneous Tasks. *Proc.CIKM*, 3:441–450, 2014.

[Liu *et al.*, 2009] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 339–348, Arlington, Virginia, United States, 2009. AUAI Press.

[Lu *et al.*, 2017] Chun-Ta Lu, Lifang He, Weixiang Shao, Bokai Cao, and Philip S. Yu. Multilinear Factorization Machines for Multi-Task Multi-View Learning. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining - WSDM '17*, pages 701–709, 2017.

[Nesterov, 2013] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[Qian and Zhai, 2014] Mingjie Qian and Chengxiang Zhai. Unsupervised feature selection for multi-view clustering on text-image web news data. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1963–1966. ACM, 2014.

[Recht *et al.*, 2010] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[Sindhwani *et al.*, 2005] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views*, volume 2005, pages 74–79. Citeseer, 2005.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Zhang and Huan, 2012] Jintao Zhang and Jun Huan. Inductive multi-task learning with multiple view data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, page 543, 2012.

[Zhou *et al.*, 2018] D. Zhou, J. Wang, B. Jiang, H. Guo, and Y. Li. Multi-task multi-view learning based on cooperative multi-objective optimization. *IEEE Access*, 6:19465–19477, 2018.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.