# Heterogeneous Representation Learning with Structured Sparsity Regularization

Pei Yang
*Arizona State University*
*Tempe, AZ 85281, USA*
*Email: cs.pyang@gmail.com*

Jingrui He
*Arizona State University*
*Tempe, AZ 85281, USA*
*Email: jingrui.he@gmail.com*

*Abstract*—**Motivated by real applications, heterogeneous learning has emerged as an important research area, which aims to model the co-existence of multiple types of heterogeneity. In this paper, we propose a HEterogeneous REpresentation learning model with structured Sparsity regularization (HERES) to learn from multiple types of heterogeneity. HERES aims to leverage two kinds of information to build a robust learning system. One is the rich correlations among heterogeneous data such as task relatedness, view consistency, and label correlation. The other is the prior knowledge of the data in the form of, e.g., the soft-clustering of the tasks. HERES is a generic framework for heterogeneous learning, which integrates multi-task, multi-view, and multi-label learning into a principled framework based on representation learning. The objective of HERES is to minimize the reconstruction loss of using the factor matrices to recover the input matrix for heterogeneous data, regularized by the structured sparsity constraint. The resulting optimization problem is challenging due to the non-smoothness and non-separability of structured sparsity. We develop an iterative updating method to solve the problem. Furthermore, we prove that the reformulation of structured sparsity is separable, which leads to a family of efficient and scalable algorithms for solving structured sparsity penalized problems. The experimental results in comparison with state-of-the-art methods demonstrate the effectiveness of the proposed approach.**

## 1. Introduction

In many data mining applications, heterogeneity is an intrinsic property of the data, which has fueled the research on heterogeneous learning [15], [36], [37], [44] in recent years. In this paper, we focus on the co-existence of triple types of heterogeneity such as task, view, and label heterogeneity. Take image annotation as an example. The images collected from different websites or using different devices follow different feature distributions, corresponding to task heterogeneity. The images are represented by different types of features, such as global features, grid-based features, bag of visual words, corresponding to view heterogeneity. Also, each image is usually associated with multiple classes, corresponding to label heterogeneity. On the other hand, the data may exhibit the soft-clustering property from different perspectives. For example, the tasks may be naturally grouped into clusters with overlap. Here, the goal is to build a heterogeneous representation learning system to maximally leverage the correlations among heterogeneous data, while imposing the structured sparsity regularization to encode our prior knowledge in the model.

The major challenges arising from these problems are two-fold. One is how to model the rich correlations among the heterogeneous data such as task relatedness [5], view consistency [10], and label correlation [46], in a principled framework [37]. The other is how to solve the structured sparsity regularized problem, which is challenging due to the properties of non-smoothness and non-separability.

In this paper, we propose a HEterogeneous REpresentation learning model with structured Sparsity regularization (HERES). HERES is a generic approach for heterogeneous learning from multiple types of heterogeneity, which incorporates the task relatedness, view consistency, and label correlation into a principled framework. Specifically, we study the heterogeneous learning problem from matrix factorization perspective, which decomposes the data matrix into basis matrix and encoding matrix. HERES models the task relatedness by requiring different tasks to share a common basis matrix, the view consistency (or label correlation) by requiring different views (or labels) to share a common encoding matrix. The objective of HERES is to minimize the reconstruction loss of using both the basis matrices and the encoding matrices to recover the heterogeneous data, penalized by the structured sparsity. The structured sparsity regularization allows us to encode our prior knowledge of the data (e.g., the soft-clustering of tasks) into the model. It requires the similar tasks to behave similarly in selecting the informative latent features, while truncating the irrelevant ones, which enhances the robustness and improves the generalization performance of the model.

The main obstacle to solve this problem is the non-smoothness and non-separability of the structured sparsity. To tackle this problem, we first transform the nonsmooth objective with structured sparsity into a smooth one by making use of an auxiliary function. Then, we prove that the reformulated problem is separable. Thanks to this appealing property, we are able to split the structured sparsity penalized problem into multiple independent sub-problems, which can be solved in parallel. The sub-problems enjoy the nice properties of convexity and having analytical solutions.

Also, it is worth mentioning that the proposed structured sparsity subsumes various sparsity regularizations, such as task-specific sparsity, task-common sparsity, their combinations, etc. The main contributions of this paper can be summarized as follows:

- We propose a novel heterogeneous representation learning model with structured sparsity to learn from complex heterogeneity by leveraging the correlations and prior knowledge of heterogeneous data.
- HERES generalizes heterogeneous learning with single or dual heterogeneity by integrating multi-task, multi-view, and multi-label learning into a principled framework based on representation learning.
- We prove that the reformulation of structured sparsity is separable, which leads to a family of efficient and scalable algorithms for solving the structured sparsity regularized problems.
- Experimental results on various data sets show the effectiveness of the proposed approach.

The rest of the paper is organized as follows. After the review of the related work in Section 2, we present the proposed model in Section 3, and discuss some of its special cases in Section 4. Section 5 shows the experimental results. Finally, we conclude the paper in Section 6.

## 2. Related Work

In this section, we review the related work on both structured sparsity regularization and heterogeneous learning.

### 2.1. Structured Sparsity

Imposing sparsity regularizations such as Lasso [32] and group Lasso [43] on learning model usually leads to better performance and model interpretability. While Lasso [32] using the $\ell_1$ penalty results in sparse models, it can not leverage any prior information such as the natural grouping of features. When features are partitioned into groups, group Lasso [43] leads to the selection of groups of features. However, the non-overlapping group structure in group Lasso limits its applicability.

Some recent work has studied the structured sparsity problem. Most of them are based on first-order or second-order optimization algorithms. A straightforward solution is to duplicate features that belong to more than one group, and apply forward-backward splitting algorithms on the expanded space [16]. However, it is limited by its scalability and the need to maintain the data consistency. A primal-dual algorithm for overlapping group sparse regularization was proposed in [26] with no need of data duplication, which is based on proximal methods. A network flow algorithms for structured sparsity was introduced in [24]. It shows that the proximal operator associated with the structured norm can be computed by solving a quadratic min-cost flow problem. In [28], an alternation direction method was developed for structured sparsity penalized problem, which is based on the augmented Lagrangian method. The Structured-Lasso (SLasso) presented in [17] applied an active set algorithm to solve the optimization problem. In [2], the proximal gradient method was adopted to solve the overlapping group Lasso, and a fixed point method was developed to compute the proximal operator. A smoothing proximal gradient method [7] reformulated the structured sparse regression problems by using the Nesterov's smoothing technique, and solved the approximation problem by proximal gradient method. The FoGLasso method [42] solved the overlapping group Lasso penalized problem via the accelerated gradient descent method using the proximal operator as a key building block. The structural graphical Lasso approach [39] established a bridge between the computation of the proximal operator associated with a structural regularization and the derivation of a screening rule for structural graphical Lasso.

### 2.2. Heterogeneous Learning

Heterogeneous learning aims to leverage different types of heterogeneity such as task, view, and label heterogeneity, to improve the learning performance.

Multi-task learning seeks to learn the relatedness among multiple tasks to improve the learning performance for each task. Different assumptions on task relatedness lead to different regularizations imposed on the model. Multi-task feature learning [1], [23] assumed that multiple related tasks share a low-dimensional representation. Clustered multi-task learning [47] assumed that multiple tasks follow a clustered structure. Trace-norm regularized methods constrained the models of different tasks to share a low-dimensional subspace [1], [19]. Robust multi-task learning aimed to identify irrelevant tasks by decomposing the model into a shared feature structure that captures task relatedness, and a group-sparse structure that detects outliers [11]. Sparsity regularizations are widely used in multi-task learning, such as $\ell_{2,1}$-norm regularized method [1], group Lasso regularized method [40], sparse group Lasso regularized methods [48], tree-guided group Lasso regularized method [20], tree-guided fused lasso [14], elastic net regularized method [12], etc.

In multi-label learning, each instance is associated with a set of labels. The key issue for multi-label learning is how to exploit the correlations among multiple labels. Multi-label learning methods can be classified into three categories [46]: first-order method such as ML-kNN [45], second-order approach such as Rank-SVM [9], and high-order methods. High-order methods have become the mainstream of multi-label learning due to their strong correlation-modeling capabilities. To name a few, subspace learning approach LS-ML [18] assumed that a common subspace is shared among multiple labels; trace-norm regularized method LEML [41] modeled a multi-label learning framework with rank constraints; multi-label feature learning method [6] applied $\ell_{2,1}$-norm regularization to the objective function to make the classifier robust for outliers; sparse local embeddings method SLEEC [3] learned a small ensemble of local distance preserving embeddings, and used $\ell_1$ regularization to obtain sparse embeddings. Some other methods such as TRAM [21] worked in a transductive way by leveraging the information from unlabeled data to estimate the optimal label concept compositions.

The goal of multi-view learning is to leverage the complementary information among different views to improve the performance [4]. Multi-view learning methods can be divided into two groups: co-regularization algorithms such as SVM-2K [10] and CoMR [30], and canonical-correlation analysis (CCA) based algorithms [31]. Some recent work aimed to learn the subspaces from multi-view data, such as the MISL algorithm [35] which discovered a latent intact representation of the data by using Cauchy loss to measure the reconstruction cost; the MSL [34] model which tried to jointly recover the corresponding latent representation and reconstruction model; the subspace representation learning method [13] which formulated the unsupervised multi-view clustering as a joint optimization problem with a common subspace representation matrix and a group sparsity inducing norm. Both of [13], [34] used $\ell_{2,1}$ norm regularization for the subspace representation learning.

Most recently, heterogeneous learning from multiple types of heterogeneity became to receive much attentions, such as multi-task multi-view learning which modeled task relatedness in the presence of multiple views [15], [36], [44], multi-view multi-label learning which modeled both the view consistency and the label correlation [38], and heterogeneous learning from triple types of heterogeneity [37] including task, view, and label heterogeneity.

In this paper, we focus on heterogeneous learning from triple types of heterogeneity by leveraging both the correlations among heterogeneous data and the prior knowledge on the data, which is formulated as a structured sparsity regularized representation learning problem. These make our work distinctive from the previous approaches.

## 3. The Proposed HERES Model

We first introduce the proposed HERES model. Then, an efficient algorithm is developed to solve the optimization problem.

### 3.1. Objective

Suppose we are given multi-task multi-view multi-label data. $T$ and $V$ are the numbers of tasks and views, respectively. Let $X_{ij} \in \mathcal{R}^{d_j \times n_i}$ be the feature-instance matrix for the data of $i$-th task and $j$-th view, $Y_i \in \mathcal{R}^{m \times n_i}$ the label-instance matrix for the data of $i$-th task, where $n_i$ is the number of instances in $i$-th task, $d_j$ is the number of features in $j$-th view, and $m$ is the number of labels.

The $i$-th row and $j$-th column vectors of a matrix $W$ are represented by $W_{i:}$ and $W_{:j}$, respectively. $diag(v)$ returns a diagonal matrix with elements of vector $v$ on the main diagonal. $\|W\|_F$ is the Frobenius norm of matrix $W$. The $\ell_{2,1}$ norm of a matrix $W$ is defined as $\|W\|_{2,1} = \sum_i \|W_{i:}\|_2 = \sum_i \sqrt{\sum_j W_{ij}^2}$.

We study the heterogeneous learning problem from matrix factorization perspective. We try to reconstruct the feature-instance matrix and label-instance matrix by letting $X_{ij} \approx B_j \phi_i$ and $Y_i \approx U\phi_i$ simultaneously, where $1 \leq i \leq T$ and $1 \leq j \leq V$. Here, $\phi_i \in \mathcal{R}^{p \times n_i}$ is the encoding matrix, where $p$ is the dimensionality of the latent space. $B_j \in \mathcal{R}^{d_j \times p}$ is the basis matrix for the feature-instance data. $U \in \mathcal{R}^{m \times p}$ is the basis matrix for label-instance data.

The main idea of the proposed HERES model is to learn a robust representation from the heterogeneous data. First, HERES integrates multi-task, multi-view, and multi-label learning into a principled framework based on representation learning. We model the *task relatedness* by requiring different tasks to share a common basis matrix $B_j$ in the $j$-view. The *label correlation* is encoded into the common basis matrix $U$ shared across multiple tasks. The *view consistency* is captured in the common encoding matrix $\phi_i$ shared across multiple views for the $i$-task. Also, the decompositions of the feature-instance and the label-instance matrices in the $i$-task share the common encoding matrix $\phi_i$. Second, HERES leverages the prior knowledge on the soft-clustering of the tasks to build a robust system by imposing the *structured sparsity* regularization on the model. Let $G$ be the number of task clusters. Denote the set of task index in the $k^{th}$ cluster by $g\left(k\right)$, and the cluster set by $\Omega = \{g(k) | 1 \leq k \leq G\}$. Let $\phi_{(k)}$ be the block matrix corresponding to the $k^{th}$ cluster, which is a concatenated matrix of all the encoding matrices $\phi_i (1 \leq i \leq T)$ if $i \in g(k)$. For example, $\|[\phi_1, \cdots, \phi_T]\|_{2,1}$ puts all tasks into one cluster and encourages them to share a common sparsity structure. The soft-clustering of tasks is given as the prior knowledge.

The objective of HERES is to minimize the reconstruction loss resulting from using both the basis matrices and encoding matrices to recover the heterogeneous data, while imposing the structured sparsity constraints on the groups of the encoding matrices, i.e.,

$$\begin{aligned}\min_{U}\min_{\{B_j\}}\min_{\{\phi_i\}} & \sum_{i=1}^{T}\sum_{j=1}^{V}\|X_{ij} - B_j\phi_i\|_F^2 + \lambda^2\|Y_i - U\phi_i\|_F^2 \\ & + \sum_{k=1}^{G}\alpha_k\left\|\phi_{(k)}\right\|_{2,1}\end{aligned} \tag{1}$$

where $\lambda$ and $\alpha_k$ $(1 \leq k \leq G)$ are non-negative parameters to control the importance of the empirical loss and the structured sparsity, respectively. The $l_{2,1}$ norm $\left\|\phi_{(k)}\right\|_{2,1}$ encourages certain rows of $\phi_{(k)}$ to become sparse, hence reducing the dimensionality of the learned representations for the $k^{th}$ task cluster.

The intuition behind the HERES model is as follows. First, $\phi_i$ can be viewed as the new representations of the instances of $i$-th task in the latent space, which acts as a bridge to connect the feature spaces between different views, as well as to connect the feature space in each view with the label space. Taking webpage classification as an example, the words (one view) on the webpage, the hyperlinks (another view) pointing to the webpage, and categories (labels) of webpage could be linked by the latent topics of the webpage. Second, $B_j$ can be viewed as the new representations of the features of $j$-th view in the latent space, which builds the connection among tasks. Likewise, $U$ can be viewed as the new representations of the labels in the latent space, which correlates the multiple labels of different tasks. Furthermore, HERES introduces the structured sparsity regularization to encourage the similar tasks to behave similarly in selecting the most informative bases, while truncating the irrelevant ones.

By letting $X_i = \begin{bmatrix} X_{i1} \\ \vdots \\ X_{iV} \\ \lambda Y_i \end{bmatrix}, B = \begin{bmatrix} B_1 \\ \vdots \\ B_V \\ \lambda U \end{bmatrix}, \phi = \{\phi_i\}$, Eq. 1 can be transformed into a compact form:

$$\min_B \min_\phi \sum_{i=1}^{T} \|X_i - B\phi_i\|_F^2 + \sum_{k=1}^{G} \alpha_k \left\|\phi_{(k)}\right\|_{2,1} \tag{2}$$

To avoid degeneracy in representation learning, we add the smoothness regularization on the basis matrix $B$. Then, the overall objective of HERES is as follow:

$$\min_B \min_\phi \sum_{i=1}^{T} \|X_i - B\phi_i\|_F^2 + \sum_{k=1}^{G} \alpha_k \left\|\phi_{(k)}\right\|_{2,1} + \beta \|B\|_F^2 \tag{3}$$

where $\beta$ is a non-negative parameter to control the smoothness of the model.

The major advantages of the proposed HERES model are two-fold. First, it is a generic framework for learning complex heterogeneity. HERES is widely applicable to heterogeneous learning with single or multiple types of heterogeneity. Second, the introduced structured sparsity allows for the modeling of soft-clustering of tasks. It accommodates multiple sparsity regularizations. For example, we can simultaneously impose multiple types of sparsity constraints on the model, such as task-specific sparsity, task-common sparsity, their combinations, etc. Task-specific sparsity $\|\phi_i\|_{2,1}$ is used to capture the task-dependent structures, while task-common sparsity $\|[\phi_1, \cdots, \phi_T]\|_{2,1}$ is used to capture the task-independent structures. Some case studies will be introduced in the next section.

### 3.2. Optimization

Solving the optimization problem in HERES is challenging due to the non-smoothness and non-separability of structured sparsity. Although the overlapping sparsity constraints are difficult to separate in its original form, we find that the reformulation of the problem is separable. Next, we first transform the non-smooth problem into a smooth one by making use of an auxiliary function. Then, we prove the separability of the new formulation of structured sparsity, which leads to an efficient and scalable algorithm to solve the structured sparsity penalized problem.

**Theorem 1.** *[Reformulation of Objective] The objective function in Eq. 3 with respect to $\phi$ is non-increasing under the update*

$$\phi^{(t+1)} = \arg\min_\phi \sum_{i=1}^{T} \|X_i - B\phi_i\|_F^2 + \sum_{k=1}^{G} \alpha_k tr \left[\phi_{(k)}^T D_{(k)}^{(t)} \phi_{(k)}\right] \tag{4}$$

*where $t$ is the iteration index, and*

$$D_{(k)}^{(t)} = diag\left(\frac{1}{2\left\|\left[\phi_{(k)}^{(t)}\right]_{1:}\right\|_2}, \cdots, \frac{1}{2\left\|\left[\phi_{(k)}^{(t)}\right]_{p:}\right\|_2}\right). \tag{5}$$

*Proof.* We make use of an auxiliary function to derive the updating rule for $\phi$. Let

$$\begin{aligned} F(\phi) &= \sum_{i=1}^{T} \|X_i - B\phi_i\|_F^2 + \sum_{k=1}^{G} \alpha_k \left\|\phi_{(k)}\right\|_{2,1} \\ &= \sum_{i=1}^{T} \|X_i - B\phi_i\|_F^2 + \sum_{k=1}^{G} \alpha_k \sum_{r=1}^{p} \left\|[\phi_{(k)}]_{r:}\right\|_2 \end{aligned}$$

Then, we define a new function,

$$\begin{aligned} J\left(\phi, \phi^{(t)}\right) = \sum_{i=1}^{T} \|X_i - B\phi_i\|_F^2 + \\ \sum_{k=1}^{G} \alpha_k \sum_{r=1}^{p} \frac{\|[\phi_{(k)}]_{r:}\|_2^2 + \|[\phi_{(k)}^{(t)}]_{r:}\|_2^2}{2\|[\phi_{(k)}^{(t)}]_{r:}\|_2} \end{aligned}$$

where $\phi^{(t)}$ is the value of $\phi$ at iteration $t$. Note that $J\left(\phi, \phi^{(t)}\right)$ is an auxiliary function of $F(\phi)$ due to the facts that $J(\phi, \phi) = F(\phi)$ and $J\left(\phi, \phi^{(t)}\right) \geq F(\phi)$. The latter follows from $a^2 + b^2 \geq 2ab$ for any scalars $a$ and $b$.

Since $J\left(\phi, \phi^{(t)}\right)$ is an auxiliary function of $F(\phi)$, $F(\phi)$ is non-increasing under the update

$$\phi^{(t+1)} = \arg\min_\phi J\left(\phi, \phi^{(t)}\right).$$

Similar to [27], we can obtain the derivative of $J\left(\phi, \phi^{(t)}\right)$:

$$\begin{aligned} &\frac{\partial}{\partial \phi} J\left(\phi, \phi^{(t)}\right) \\ =&\frac{\partial}{\partial \phi} \left[\sum_{i=1}^{T} \|X_i - B\phi_i\|_F^2 + \sum_{k=1}^{G} \alpha_k \sum_{r=1}^{p} \frac{\|[\phi_{(k)}]_{r:}\|_2^2}{2\|[\phi_{(k)}^{(t)}]_{r:}\|_2}\right] \\ =&\frac{\partial}{\partial \phi} \left[\sum_{i=1}^{T} \|X_i - B\phi_i\|_F^2 + \sum_{k=1}^{G} \alpha_k tr \left(\phi_{(k)}^T D_{(k)}^{(t)} \phi_{(k)}\right)\right] \end{aligned}$$

Since $F\left(\phi^{(t)}\right) = J\left(\phi^{(t)}, \phi^{(t)}\right) \geq \min_\phi J\left(\phi, \phi^{(t)}\right) = J\left(\phi^{(t+1)}, \phi^{(t)}\right) \geq F\left(\phi^{(t+1)}\right)$, the objective function $F(\phi)$ is non-increasing under the above update. □

Theorem 2 shows that the reformulation of structured sparsity is separable. The intuition of separability here is that the optimization problem with the overlapping structured sparsity can be split into multiple independent sub-problems, which could be solved in parallel.

**Theorem 2.** *[Separability of Structured Sparsity] The reformulation of structured sparsity is separable:*

$$\begin{aligned}&\sum_{k=1}^{G}\alpha_k tr\left[\phi_{(k)}^T D_{(k)}^{(t)}\phi_{(k)}\right]\\=&\sum_{i=1}^{T} tr\left[\phi_i^T\left(\sum_{1\le k\le G, i\in g(k)}\alpha_k D_{(k)}^{(t)}\right)\phi_i\right]\end{aligned}\tag{6}$$

*Proof.*

$$\begin{aligned}&\sum_{k=1}^{G}\alpha_k tr\left[\phi_{(k)}^T D_{(k)}^{(t)}\phi_{(k)}\right]\\=&\sum_{k=1}^{G}\alpha_k tr\left[\sum_{r=1}^{p}\left[D_{(k)}^{(t)}\right]_{rr}\cdot\left[\phi_{(k)}\right]_{r:}^T\left[\phi_{(k)}\right]_{r:}\right]\\=&\sum_{k=1}^{G}\alpha_k\sum_{r=1}^{p}\left[D_{(k)}^{(t)}\right]_{rr}\cdot\left\|\left[\phi_{(k)}\right]_{r:}\right\|_2^2\\=&\sum_{k=1}^{G}\alpha_k\sum_{r=1}^{p}\left[D_{(k)}^{(t)}\right]_{rr}\cdot\sum_{i\in g(k)}\left\|[\phi_i]_{r:}\right\|_2^2\\=&\sum_{k=1}^{G}\alpha_k\sum_{i\in g(k)}\sum_{r=1}^{p}\left[D_{(k)}^{(t)}\right]_{rr}\cdot\left\|[\phi_i]_{r:}\right\|_2^2\\=&\sum_{k=1}^{G}\alpha_k\sum_{i\in g(k)} tr\left(\phi_i^T D_{(k)}^{(t)}\phi_i\right)\\=&\sum_{i=1}^{T} tr\left[\phi_i^T\left(\sum_{1\le k\le G, i\in g(k)}\alpha_k D_{(k)}^{(t)}\right)\phi_i\right]\end{aligned}$$

which completes the proof. □

According to Theorem 1 and Theorem 2, the objective function in Eq. 3 is then transformed into:

$$\min_B\min_\phi\sum_{i=1}^{T}\|X_i-B\phi_i\|_F^2+tr\left[\phi_i^T D_i^{(t)}\phi_i\right]+\beta\|B\|_F^2\tag{7}$$

where

$$D_i^{(t)}=\sum_{1\le k\le G, i\in g(k)}\alpha_k D_{(k)}^{(t)}\tag{8}$$

By now the optimization problem in Eq. 7 with respect to $\phi$ can be split into $T$ independent sub-problems, each of which is related to $\phi_i$ $(1\le i\le T)$, respectively. This appealing property makes it possible to design an efficient and scalable algorithm to solve the problem in parallel.

We reach the final solution of HERES. The objective function in Eq. 7 is an unconstrained quadratic optimization with respect to $B$ and $\phi_i$, respectively. Thus, we can obtain the analytical solutions as shown in Theorem 3.

**Theorem 3.** *[Optimums] The optimal solutions for the objective function in Eq. 7 are as follows,*

$$B=\left(\sum\nolimits_{i=1}^{T}X_i\phi_i^T\right)\left(\beta I+\sum\nolimits_{i=1}^{T}\phi_i\phi_i^T\right)^{-1}\tag{9}$$

$$\phi_i=\left(B^TB+D_i^{(t)}\right)^{-1}B^TX_i\tag{10}$$

*where $1\le i\le T$.*

*Proof.* First, fix $\phi$ and optimize $B$. The zero gradient condition of Eq. 7 with respect to $B$ gives

$$\begin{aligned}&\beta B+\sum\nolimits_{i=1}^{T}\left(B\phi_i\phi_i^T-X_i\phi_i^T\right)=0\\\Rightarrow&B=\left(\sum\nolimits_{i=1}^{T}X_i\phi_i^T\right)\left(\beta I+\sum\nolimits_{i=1}^{T}\phi_i\phi_i^T\right)^{-1}\end{aligned}$$

Second, fix $B$ and optimize $\phi$. The zero gradient condition of Eq. 7 with respect to $\phi_i$ gives

$$\begin{aligned}&B^TB\phi_i-B^TX_i+D_i^{(t)}\phi_i=0\\\Rightarrow&\phi_i=\left(B^TB+D_i^{(t)}\right)^{-1}B^TX_i\end{aligned}$$

which completes the proof. □

We summarize the above procedures into the HERES algorithm as shown in Algorithm 1. We make use of block coordinate descent method [33] to iteratively solve the problem. It updates $D_{(k)}$, $B$ and $\phi_i$ respectively in an iterative way until convergence. Theorem 4 demonstrates the convergence of the proposed HERES algorithm.

**Theorem 4.** *[Convergence] The proposed HERES algorithm is guaranteed to converge to the local optimum.*

*Proof.* Since the objective in Eq. 7 is block-wise convex with respect to $B$ and $\phi_i(1\le i\le T)$, according to the property of block coordinate descent method (Lemma 3.1 in [33]), the proposed HERES algorithm based on block coordinate descent will converge to the local optimum. □

**Algorithm 1** HERES Algorithm

```
Input: Feature-instance matrices $X_{ij}$ and label-instance
    matrices $Y_i$ where $1 \le i \le T$ and $\le j \le V$, dimension
    $p$, task clusters $\Omega$, parameters $\alpha_k\ (1 \le k \le G), \beta, \lambda$.
Output: Predicted label-instance matrices.
1: Set $t = 0$;
2: Initialize $\phi_i^{(t)}$ by using traditional clustering algorithm
   such as $K$-means for each task, where $1 \le i \le T$;
3: repeat
4:    Update $D_{(k)}^{(t)}$ by Eq. 5 for each task cluster, where
      $1 \le k \le G$;
5:    Update $B^{(t+1)}$ by Eq. 9;
6:    Update $\phi_i^{(t+1)}$ by Eq. 10 for each task, where $1 \le$
      $i \le T$;
7:    Set $t \leftarrow t + 1$;
8: until converged
9: return Predicted label-instance matrices $F_i = U\phi_i$
   where $1 \le i \le T$.
```

Regarding the algorithm complexity of HERES, the most time or space consuming steps are to update $B$ (step 5) and $\phi_i$ (step 6). According to Eq. 9 and Eq. 10, the size of the involved matrices in inversion is $p\times p$, where $p$ is the dimensionality of the latent space. Note that we usually have $p\ll\max(n_i,d_j)$, where $n_i$ is the number of instances in $i$-th task, and $d_j$ is the number of feature in the $j$-th view. Therefore, the proposed HERES algorithm is scalable to the problem size.

## 4. Case Study

In this section, we introduce some special cases of the proposed model. Based on the prior knowledge of the data, the structured sparsity can be instantiated as various sparsity regularizations.

It is well known that Lasso [32] encourages element-wise sparsity; group Lasso [43] encourages group-wise sparsity; while sparse group Lasso [29] encourages both group-wise and element-wise sparsity. Accordingly, in heterogeneous learning, we may hope to achieve task-level sparsity, task-cluster-level sparsity, their combinations, etc. All these sparsity regularizations can be accommodated in the proposed model. Note that the structured sparsity introduced here focuses on $\ell_{2,1}$ norm, which is different from Lasso, group Lasso, and sparse group Lasso that are based on $\ell_1/\ell_2$ norm. Also, we study the structured sparsity problem in heterogeneous representation learning, which is more challenging since we need to simultaneously learn both the basis matrix and the encoding matrix.

The following corollaries show some special cases of the proposed model, which are based on the relationships between the tasks and task clusters. In all the special cases, the optimal solutions for $B$ (Eq. 9) and $\phi_i$ $(1 \leq i \leq T)$ (Eq. 10) in HERES remain unchanged. The only difference is the update of $D_i^{(t)}$ in Eq. 8, which can be further reduced to specific forms. The proofs of the corollaries are omitted for brevity.

**Corollary 1.** *(Task-specific Sparsity): Suppose there are $T$ task clusters and each task corresponds one cluster, i.e., the cluster set $\Omega = \{\{1\}, \cdots, \{T\}\}$, the objective function in Eq. 3 is rewritten into*

$$\min_{B,\phi} \sum\nolimits_{i=1}^{T} \left(\|X_i - B\phi_i\|_F^2 + \alpha_i \|\phi_i\|_{2,1}\right) + \beta \|B\|_F^2$$

*then, we obtain the diagonal matrix $D_i^{(t)}$ whose $(j,j)(1 \leq j \leq p)$ element is $\left[D_i^{(t)}\right]_{jj} = \frac{\alpha_i}{2\left\|\left[\phi_i^{(t)}\right]_{j:}\right\|_2}$.*

**Corollary 2.** *(Task-common Sparsity): Suppose all the tasks are in one cluster, i.e., the cluster set $\Omega = \{\{1, \cdots, T\}\}$, the objective in Eq. 3 is rewritten into*

$$\min_{B,\phi} \sum\nolimits_{i=1}^{T} \|X_i - B\phi_i\|_F^2 + \alpha_a \|\phi_a\|_{2,1} + \beta \|B\|_F^2$$

*where $\phi_a = [\phi_1, \cdots, \phi_T]$ and $\alpha_a$ is the parameter, then, we obtain the diagonal matrix $D_i^{(t)}$ whose $(j,j)(1 \leq j \leq p)$ element is $\left[D_i^{(t)}\right]_{jj} = \frac{\alpha_a}{2\left\|\left[\phi_a^{(t)}\right]_{j:}\right\|_2}$.*

**Corollary 3.** *(Task-specific and Task-common Sparsity): Suppose cluster set $\Omega = \{\{1\}, \cdots, \{T\}, \{1, \cdots, T\}\}$ contains $T+1$ clusters, the objective in Eq. 3 is rewritten into*

$$\min_{B,\phi} \sum_{i=1}^{T} \left(\|X_i - B\phi_i\|_F^2 + \alpha_i \|\phi_i\|_{2,1}\right) + \alpha_a \|\phi_a\|_{2,1} + \beta \|B\|_F^2$$

*then, we obtain the diagonal matrix $D_i^{(t)}$ whose $(j,j)(1 \leq j \leq p)$ element is $\left[D_i^{(t)}\right]_{jj} = \frac{\alpha_i}{2\left\|\left[\phi_i^{(t)}\right]_{j:}\right\|_2} + \frac{\alpha_a}{2\left\|\left[\phi_a^{(t)}\right]_{j:}\right\|_2}$.*

Note that in Corollary 3, each task belongs to two clusters, i.e., $i \in \{i\}$ and $i \in \{1, \cdots, T\}$. In other words, the cluster $\{1, \cdots, T\}$ overlaps with each of the remaining $T$ clusters. It facilitates the joint modeling of task-specific sparsity and task-common sparsity.

It is worth pointing out that the scope of the proposed structured sparsity is beyond these special cases. It allows for flexible modeling of soft-clustering of tasks, i.e., each task can belong to any clusters.

## 5. Experiments

We implement the model introduced in Corollary 3, which employs the structured sparsity to capture both task-dependent and task-independent relatedness.

### 5.1. Data Sets and Setup

We evaluate the algorithms on three benchmark data sets including two text data and one image data. All of them are available online[1].

Reuters Corpus Volume I (RCV1V2) data set [22] is widely used for the evaluation of heterogeneous learning algorithm. RCV1V2 contains about 800,000 newswire stories. There are three category sets of data: topics, industry codes, and regions. Each of these category sets has a hierarchical structures. It is common to use four subsets of this data, each containing 6000 instances on average and with a total number of 101 class labels.

EUR-Lex [25] data set contains nearly 20,000 text documents about European Union official laws, different kinds of treaties and agreements, parliamentary journals. The documents are organized in a hierarchical structures according to three different schemas: subject matter, directory codes, and EUROVOC. There are 412 labels in total.

NUS-WIDE [2] is a real-world web image data set [8]. It consists of more than 269,000 images with over 5,000 user-provided tags, and ground-truth of 81 concepts with a hierarchical structures. The images are represented by different types of visual features such as 64-D color histogram in LAB color space, 144-D color correlogram in HSV color space, 73-D edge distribution histogram, and 500-D bag of visual words. The light version of NUS-WIDE is used in our experiments.

In these data sets, the label refers to the multiple categories each instance belongs to. The task refers to classifying the instances belonging to different sub-categories, which follow different data distributions [15], [37]. For the NUS-WIDE data, the view refers to different types of visual feature. For either RCV1V2 or EUR-Lex data set, similar to [37], the data are represented by two views: one corresponds to the TF-IDF features; and the other corresponds to the latent topics obtained by applying probabilistic latent semantic analysis[3] on the term frequency.

Table 1 shows the properties of different data sets. Label cardinality refers to the average number of labels

[1] http://mulan.sourceforge.net/datasets-mlc.html

[2] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

[3] http://lear.inrialpes.fr/people/verbeek/code

TABLE 1: Statistics of different data sets.

| Data set | Instances | Features | Labels | Cardinality | Density | Diversity |
|---|---|---|---|---|---|---|
| RCV1V2_1 | 6000 | 47236 | 101 | 2.880 | 0.029 | 1028 |
| RCV1V2_2 | 6000 | 47236 | 101 | 2.634 | 0.026 | 954 |
| RCV1V2_3 | 6000 | 47229 | 101 | 2.614 | 0.026 | 939 |
| RCV1V2_4 | 6000 | 47236 | 101 | 2.484 | 0.025 | 816 |
| EUR-Lex | 19348 | 5000 | 412 | 1.292 | 0.003 | 1615 |
| NUS-WIDE | 55615 | 708 | 81 | 1.869 | 0.023 | 18430 |

per instance. Accordingly, label density normalizes label cardinality by the the number of labels. Label diversity is the number of distinct label combinations observed in the data set [46].

## 5.2. Evaluation Metrics and Comparison Methods

We use $F_1$-score, accuracy and Hamming loss [46] as the evaluation metrics to examine the performance of all comparison algorithms on the test data.

$F_1$-score is the ratio of predicted correct labels to the mean of the numbers of actual labels and predicted labels, averaged over all instances. $F_1$-score is the harmonic mean of precision and recall. Accuracy refers to the ratio of predicted correct labels to the intersection between actual labels and predicted labels, averaged over all instances. Hamming Loss is the sum of the prediction error (an incorrect label is predicted) and the missing error (a relevant label not predicted), normalized over total number of labels and total number of instances. Note that the larger the values of $F_1$-score and accuracy, or the smaller the value of Hamming loss, the better the performance of the learning algorithm.

In this work, we focus on improving the performance of multi-label learning by leveraging the multiple types of heterogeneity. We compare our method with the heterogeneous learning method HiMLS proposed recently in [37]. Also, HERES is compared with different types of multi-label learning [4] methods including: graph-based multi-label approach ML-kNN [45], multi-label method based on subspace learning LS-ML [18], and transductive multi-label learning approach TRAM [21].

Note that both HERES and HiMLS [37] are capable of learning from triple types of heterogeneity, while the other methods are limited to a single type of heterogeneity, i.e., label heterogeneity. Therefore, for those methods dealing with label heterogeneity only, their input is the combined data, where the instances of all the tasks are pooled together, and the features from each view are concatenated into one single view.

We repeat the experiments ten times for each data set and report the average performance and the standard deviation. The parameters are tuned for each algorithm using cross-validation on the training data.

## 5.3. Performance Study

Figures 1-3 (in next page) show the performance in terms of $F_1$-score, accuracy, and Hamming loss respectively for the comparison methods on the six data sets. First of all, we observe that the results are basically consistent across different metrics, i.e., the larger the $F_1$-score, the larger the accuracy, the smaller the Hamming loss.

The results show that both the proposed method HERES and HiMLS [37] perform better than the other algorithms in most cases. It demonstrates that learning performance can be improved by leveraging the rich correlations among heterogeneous data, such as the task relatedness, view consistency, and label correlation. The other methods including ML-kNN [45], LS-ML [18], and TRAM [21] only consider the label correlation, while ignoring the other types of heterogeneity. The results also suggest that treating the instances in different tasks discriminatively is usually better than just pooling them into one single task. Analogously, treating different views in a complementary way is usually better than just concatenating all the features together.

HERES performs the best among all the algorithms. Although both HERES and HiMLS [37] take advantage of triple heterogeneity, the major competency of HERES over HiMLS lies in the proposed structured sparsity regularization imposed on the model. Structured sparsity helps pick out the informative latent features, while truncating the irrelevant ones, which leads to a better representation of the heterogeneous data in the new latent space. This automatic feature learning procedure improves the generalization performance of the model. Furthermore, structured sparsity regularization provides a flexible way to encode the prior knowledge on the data into the model. In this experiment, we use task-specific sparsity to select the latent features related to each individual task, while using the task-common sparsity to pick out the latent features shared across multiple tasks. As a consequence, it gains the performance promotion by modeling the task relatedness in a more refined way.

In summary, heterogeneous learning could greatly benefit from leveraging both the rich correlations among heterogeneous data including the task relatedness, view consistency, label correlation, as well as the prior knowledge about the data such as the soft-clustering of tasks.

## 5.4. Influence of Parameters

It is interesting to study how the dimensionality $p$ of the latent space influences the performance of HERES. Taking RCV1V2_1 data as an example, we vary $p$ from 100 to 8000, and observe the change of performance. The result is shown in Figure 4. We can see that HERES performs better when $p \geq 2000$ than $p < 2000$. A larger $p$ indicate a potential better representation capability of the model. Then, the sparsity constraint imposed on the model helps select the most informative latent features, leading to a better performance.

[4] Due to space limit, we omit the comparison results with multi-task or multi-view learning, as their performance is not as good as HERES.

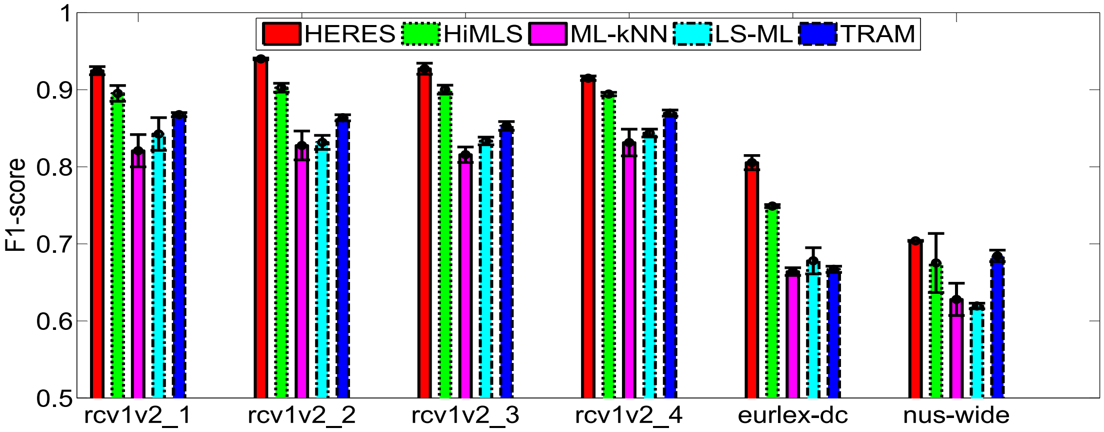

Figure 1: $F_1$-score of different algorithms on the six data sets.

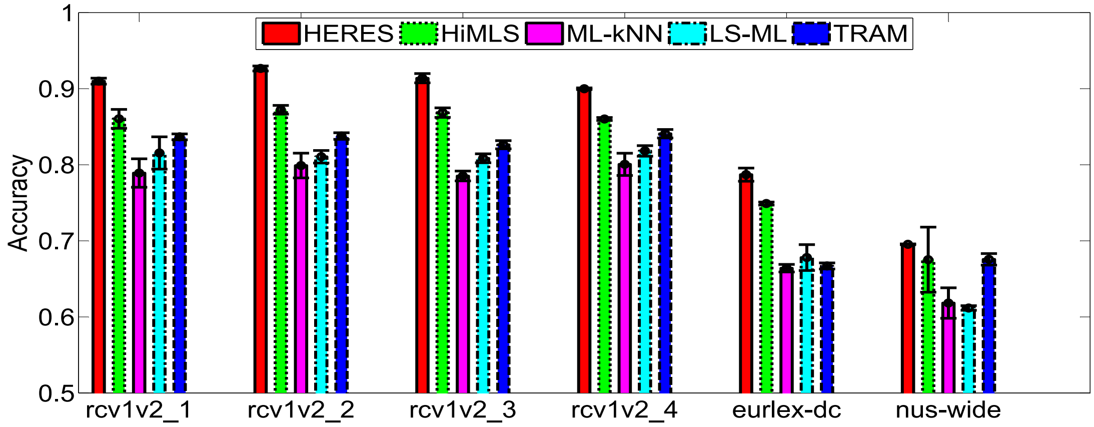

Figure 2: Accuracy of different algorithms on the six data sets.

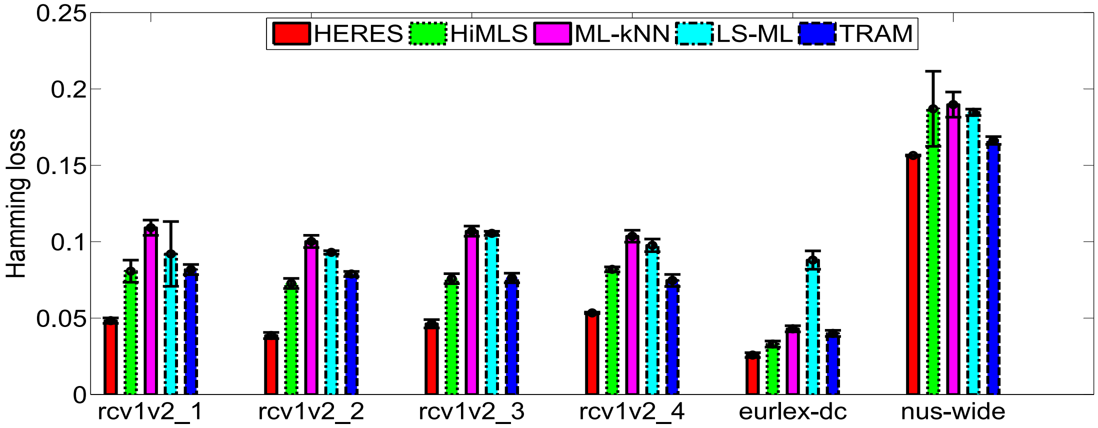

Figure 3: Hamming loss of different algorithms on the six data sets.

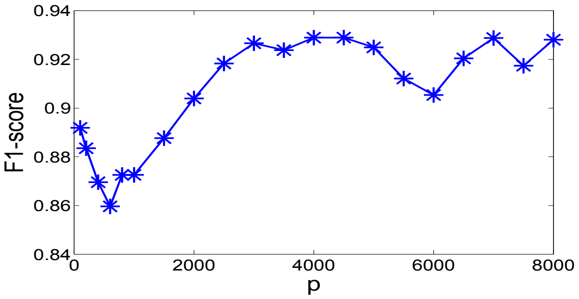

Figure 4: $F_1$-score varies with dimension $p$.

Next, we study the performance of HERES varying with the parameter $\alpha_k$, which is used to control the importance of the structured sparsity. Note that in our experiments, the cluster set $\Omega = \{\{1\}, \cdots, \{T\}, \{1, \cdots, T\}\}$ contains $T+1$ clusters (please refer to Corollary 3 for details). Therefore, we empirically set $[\alpha_1, \cdots, \alpha_T, \alpha_a] = [\alpha, \cdots, \alpha, T\alpha]$. The performance of HERES with respect to $\alpha$ on RCV1V2_1 data is shown in Figure 5. First of all, the performance is poor when $\alpha = 0$. It suggests that the structured sparsity constraint is indispensable to the proposed model. Then, the performance increases significantly as $\alpha$ increases, and reaches to the peak at $\alpha = 1/16$. But a large $\alpha$ (e.g., $\alpha > 16$) may hurt the performance.

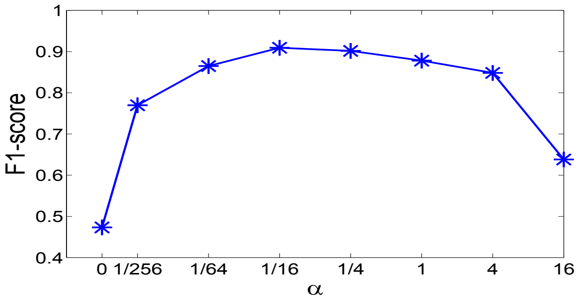

Figure 5: $F_1$-score varies with $\alpha$ ($log_4$ scale).

Figure 6 shows the performance of HERES varying with $\beta$ on RCV1V2_1 data, which controls the importance of smoothness regularization. When $\beta = 0$, it means that no smoothness constraint is imposed on the model, resulting in poor performance. The results show that HERES is robust over a wide range of $\beta$ values.

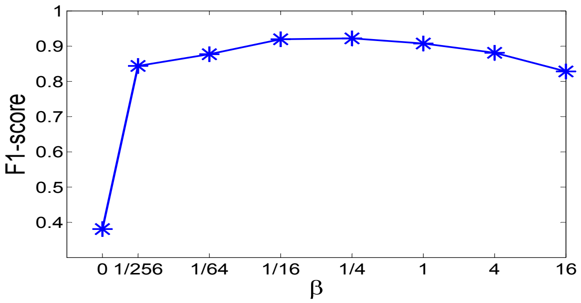

Figure 6: $F_1$-score varies with $\beta$ ($log_4$ scale).

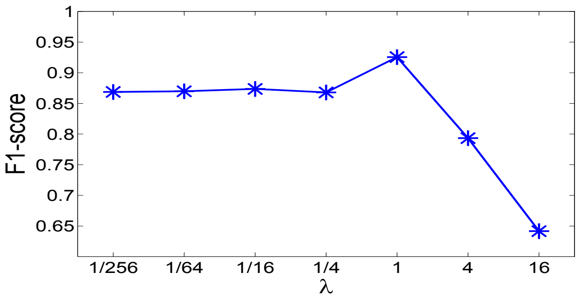

Figure 7: $F_1$-score varies with $\lambda$ ($log_4$ scale).

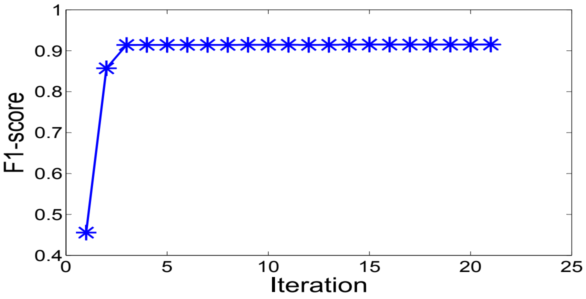

Figure 8: $F_1$-score varies with each iteration.

Figure 7 shows the performance sensitivity with respect to $\lambda$, which is used to balance the contribution from empirical loss. The performance curve is relatively flat when $\lambda \leq 1$. But a large $\lambda$ (e.g., $\lambda = 16$) may hurt the performance, suggesting that too much weight is placed on the empirical loss.

We empirically study the convergence of HERES on the RCV1V2_1 data set. The result is shown in Figure 8. The $F_1$-score increases rapidly, and becomes stable after a few iterations. The results verify the fast convergence property of the proposed algorithm.

## 6. Conclusion

In this paper, we propose a heterogeneous representation learning model with structured sparsity regularization. HERES incorporates multiple types of correlations among heterogeneous data into a representation learning framework. The separability of the reformulated problem leads to an efficient and scalable algorithm to solve the structured sparsity regularized problem. The effectiveness of the proposed method is validated on various data sets in comparison with different heterogeneous learning methods.

## Acknowledgments

This work is supported by the NSF research grant IIS-1552654, ONR Research grant N00014-15-1-2821, IBM Faculty Award, and NSFC research grant 61473123. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the governments.

# References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, pages 41–48, 2006.

[2] A. Argyriou, C. A. Micchelli, M. Pontil, L. Shen, and Y. Xu. Efficient first order methods for linear composite regularizers. *CoRR*, abs/1104.1436, 2011.

[3] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain. Sparse local embeddings for extreme multi-label classification. In *NIPS*, pages 730–738, 2015.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100, 1998.

[5] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[6] X. Chang, F. Nie, Y. Yang, and H. Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, pages 1171–1177, 2014.

[7] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing. Smoothing proximal gradient method for general structured sparse learning. In *UAI*, pages 105–114, 2011.

[8] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In *CIVR*, 2009.

[9] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *NIPS*, pages 681–687, 2001.

[10] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmák. Two view learning: SVM-2K, theory and practice. In *NIPS*, 2005.

[11] P. Gong, J. Ye, and C. Zhang. Robust multi-task feature learning. In *KDD*, pages 895–903, 2012.

[12] P. Gong, J. Zhou, W. Fan, and J. Ye. Efficient multi-task feature learning with calibration. In *KDD*, pages 761–770, 2014.

[13] Y. Guo. Convex subspace representation learning from multi-view data. In *AAAI*, 2013.

[14] L. Han and Y. Zhang. Learning tree structure in multi-task learning. In *KDD*, pages 397–406, 2015.

[15] J. He and R. Lawrence. A graph-based framework for multi-task multi-view learning. In *ICML*, pages 25–32, 2011.

[16] L. Jacob, G. Obozinski, and J. Vert. Group Lasso with overlap and graph Lasso. In *ICML*, pages 433–440, 2009.

[17] R. Jenatton, J. Audibert, and F. R. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

[18] S. Ji, L. Tang, S. Yu, and J. Ye. Extracting shared subspace for multi-label classification. In *KDD*, pages 381–389, 2008.

[19] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *ICML*, pages 457–464, 2009.

[20] S. Kim and E. P. Xing. Tree-guided group Lasso for multi-task regression with structured sparsity. In *ICML*, pages 543–550, 2010.

[21] X. Kong, M. K. Ng, and Z.-H. Zhou. Transductive multilabel learning via label set propagation. *IEEE Trans. Knowl. Data Eng. (TKDE)*, pages 704–719, 2013.

[22] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research (JMLR)*, 5:361–397, 2004.

[23] Y. Li, X. Tian, T. Liu, and D. Tao. Multi-task model and feature joint learning. In *IJCAI*, pages 3643–3649, 2015.

[24] J. Mairal, R. Jenatton, G. Obozinski, and F. R. Bach. Network flow algorithms for structured sparsity. In *NIPS*, pages 1558–1566, 2010.

[25] E. L. Mencía and J. Fürnkranz. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *ECML-PKDD*, pages 126–135, 2008.

[26] S. Mosci, S. Villa, A. Verri, and L. Rosasco. A primal-dual algorithm for group sparse regularization with overlapping groups. In *NIPS*, pages 2604–2612, 2010.

[27] F. Nie, H. Huang, X. Cai, and C. H. Q. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[28] Z. T. Qin and D. Goldfarb. Structured sparsity via alternating direction methods. *Journal of Machine Learning Research*, 13:1435–1468, 2012.

[29] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231, 2013.

[30] V. Sindhwani and D. S. Rosenberg. An RKHS for multi-view learning and manifold co-regularization. In *ICML*, pages 976–983, 2008.

[31] K. Sridharan and S. M. Kakade. An information theoretic framework for multi-view learning. In *COLT*, pages 403–414, 2008.

[32] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[33] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

[34] M. White, Y. Yu, X. Zhang, and D. Schuurmans. Convex multi-view subspace learning. In *NIPS*, pages 1682–1690, 2012.

[35] C. Xu, D. Tao, and C. Xu. Multi-view intact space learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(12):2531–2544, 2015.

[36] H. Yang and J. He. Learning with dual heterogeneity: A nonparametric bayes model. In *KDD*, pages 582–590, 2014.

[37] P. Yang and J. He. Model multiple heterogeneity via hierarchical multi-latent space learning. In *KDD*, pages 1375–1384, 2015.

[38] P. Yang, J. He, H. Yang, and H. Fu. Learning from label and feature heterogeneity. In *ICDM*, pages 1079–1084, 2014.

[39] S. Yang, Q. Sun, S. Ji, P. Wonka, I. Davidson, and J. Ye. Structural graphical Lasso for learning mouse brain connectivity. In *KDD*, pages 1385–1394, 2015.

[40] X. Yang, S. Kim, and E. P. Xing. Heterogeneous multitask learning with joint sparsity constraints. In *NIPS*, pages 2151–2159, 2009.

[41] H.-F. Yu, P. Jain, P. Kar, and I. S. Dhillon. Large-scale multi-label learning with missing labels. In *ICML*, pages 593–601, 2014.

[42] L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group Lasso. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(9):2104–2116, 2013.

[43] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[44] J. Zhang and J. Huan. Inductive multi-task learning with multiple view data. In *KDD*, pages 543–551, 2012.

[45] M.-L. Zhang and Z.-H. Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, pages 2038–2048, 2007.

[46] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[47] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *NIPS*, pages 702–710, 2011.

[48] J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via fused sparse group Lasso. In *KDD*, pages 1095–1103, 2012.