Heterogeneous multitask learning with joint sparsity constraints

Xiaolin Yang
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
xyang@stat.cmu.edu

Seyoung Kim
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
sssykim@cs.cmu.edu

Eric P. Xing
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
epxing@cs.cmu.edu

Abstract

Multitask learning addresses the problem of learning related tasks whose information on parameters is assumed to be shared with each other. Previous approaches usually deal with homogeneous tasks such as a set of regression tasks only or a set of classification tasks only. In this paper, we consider the problem of learning multiple related tasks, where tasks consist of predicting both continuous and discrete outputs from a common set of input variables that lie in a high-dimensional space. All of the tasks are related in the sense that they share the same set of relevant input variables, but the amount of influence of each input on different outputs may vary. We formulate this problem as a combination of linear regressions and logistic regressions, and model the joint sparsity as L_1/L_∞ or L_1/L_2 norm of the model parameters. Among several possible applications, our approach addresses an important open problem in genetic association mapping, where the goal is to discover genetic markers that influence multiple correlated traits jointly. In our experiments, we demonstrate our method in the setting of association mapping, using simulated and asthma datasets, and show that our method can effectively recover the relevant inputs with respect to all of the tasks.

1 Introduction

In multitask learning, we are interested in learning a set of related models for predicting multiple (possibly) related outputs (i.e., tasks) given a set of input variables [3]. In most applications, the multiple tasks share a common input space, but have different functional mappings to different output variables corresponding to different tasks. When the tasks and their corresponding models are believed to be related, it is desirable to learn all of the models jointly rather than treating each task as independent of each other and fitting each model separately. Such methods that allow us to borrow information across tasks can potentially increase the predictive power.

Depending on the type of information shared among the tasks, many different algorithms have been proposed. Hierarchical Bayesian modeling has been introduced when the parameter values themselves are thought to be similar across tasks [1, 13]. A probabilistic method for modeling the latent structure shared across multiple tasks has been proposed [15]. On the other hand, when the input lies in high-dimensional space and the goal is to recover the shared sparsity structure, a regularized regression method has been used [9].

Traditional multitask learning only considers homogeneous tasks such as regression and classification individually. The lasso using L_1 regularization has been a popular approach for variable selection when the problem involves a single task or a set of independent tasks of linear regression [10, 11]. One of the successful extensions of the standard lasso is group lasso with an L_1/L_2 penalty [14], or more generally an L_1/L_q-regularized regression with $q > 0$ [16]. When the

L_1/L_q penalty is used in a multivariate regression for a single task, it makes use of information on the groupings of input variables, and applies the L_1 penalty over the L_q norm of the regression coefficients for each group of inputs when the variables in that group are believed to be significant as a whole. We can use simliar regularization in logistic regression and multinomial logistic regression for classification tasks[8, 12]. In the specific multitask regression problem, in order to recover these common supports to the multiple responses, we assume the common supports across all the tasks in one dimension as a group and the group variable selection method could be applied to solve this problem. Since the problem formulated in this way has a convex objective function, many of the algorithms developed for a general convex optimization problem can be used for optimization. For example, an interior point method and preconditioned conjugate gradient has been used to solve large scale L_1 regularized linear regression and logistic regression [7]. In [5, 11], coordinate descent method was used in solving L_1 regularized linear regression and generalized linear models, where the soft thresholding operator gives a closed form solution for each coordiate in each iteration.

In this paper, we consider an interesting and not uncommon scenario of multitask learning, where the tasks are *heterogeneous* and bear *union support*. That is, each task can be either a multivariate regression or classification, with the inputs lying in a very high-dimensional space, but only a small number of input variables (i.e., predictors) are relevant to each of the output variables (i.e., responses). Furthermore, we assume that all of the related tasks possibly share common relevant predictors with varying amount of influence on each task. We propose to use an L_1/L_q regularization ($q > 0$) when the tasks consist of both regression and classification problems. Assuming a linear regression for continuous-valued output and a logistic regression for discrete-valued output with dummy variables for multiple categories, an L_1/L_q penalty can be used to learn both types of tasks jointly for a union support recovery. We consider particular cases of L_1/L_q regularizations with $q = 2$ and $q = \infty$.

Our work is primarily motivated by the genetic association based on the genotype data for genetic markers called single nucleotide polymorphisms (SNPs) and phenotype data such as disease status, clinical traits, and microarray data collected over a large set of individuals. The goal in this type of study is to identify the SNPs (or inputs) that explain the variation in the phenotypes (or outputs), and at the same time to reduce false positives. Many clinical traits for a given disease are highly correlated, and genes in the same biological pathway are co-expressed in microarray data. Since the input often involves millions of SNPs and the association signal of causal (or relevant) SNPs tends to be very weak, it is greatly beneficial to combine information across multiple related phenotypes to increase the power for variable selection among predictors. Over the recent years, researchers started recognizing the importance of the joint analysis of multiple correlated phenotypes [4, 17], but there has been a lack of statistical tools to systematically perform such analysis. In our previous work [6], we developed a regularized regression method, called a graph-guided fused lasso, for multitask regression problem that takes advantage of the graph structure over tasks to encourage a selection of common inputs across highly correlated traits in the graph. While this method was only applied to the restricted case of continuous-valued outputs, the set of clinical traits related to a disease often contain both continuous- and discrete-valued traits. As we demonstrate in our experiments, the L_1/L_q regularization for the joint regression and classification can successfully handle this situation.

The paper is organized as follows. In Section 2 we introduce the notation and the basic formulation for joint regression and classification problem and we describe the L_1/L_∞ and L_1/L_2 regularized regression for heterogeneous multitask learning in the joint regression and classification setting. In Section 3 we formulate the parameter estimation as a convex optimization problem, and solve it using an interior-point method. Section 4 presents experimental results and analysis on simulated and asthma datasets. In Section 5, we conclude with a brief discussion of future work.

2 Joint Multitask Linear Regression and Multinomial Logistic Regression

Suppose we have K response variables of which K_r are continuous and K_c are discrete and $K = K_r + K_c$. They shared a common P dimensional input variables. Then for the K_r continuous variables we can model them as a regression problem where

$$Y_k = \beta_{k0}^{(r)} + X\boldsymbol{\beta}_k^{(r)} + \epsilon, \quad k = 1, ..., K_r,$$

where $\boldsymbol{\beta}_k^{(r)} = (\beta_{k1}^{(r)}, \ldots, \beta_{kP}^{(r)})'$ represents a vector of P regression coefficients for the kth regression task, with the superscript (r) indicating that this is a parameter for regression, $\beta_{k0}^{(r)}$ is the intercept, and ϵ is the residual.

The loss for the K_r regression tasks is

$$L_r = \sum_{k=1}^{K_r} (\mathbf{y}_k - \mathbf{1}\beta_{k0}^{(r)} - \mathbf{X}\boldsymbol{\beta}_k^{(r)})' \cdot (\mathbf{y}_k - \mathbf{1}\beta_{k0}^{(r)} - \mathbf{X}\boldsymbol{\beta}_k^{(r)}), \tag{1}$$

For the tasks with discrete-valued output, we set up a multinomial logistic regression for each of the K_c tasks, assuming that the kth task has M_k categories:

$$\begin{aligned} P(Y_k = m|\mathbf{X} = \mathbf{x}) &= \frac{\exp(\beta_{k0}^{(c)} + \mathbf{x}\boldsymbol{\beta}_{km}^{(c)})}{1 + \sum_{l=1}^{M_k-1} \exp(\beta_{k0}^{(c)} + \mathbf{x}\boldsymbol{\beta}_{kl}^{(c)})}, \text{for } m = 1, \ldots, M_k - 1, \\ P(Y_k = M_k|\mathbf{X} = \mathbf{x}) &= \frac{1}{1 + \sum_{l=1}^{M_k-1} \exp(\beta_{k0}^{(c)} + \mathbf{x}\boldsymbol{\beta}_{kl}^{(c)})}, \end{aligned} \tag{2}$$

where $\boldsymbol{\beta}_{km}^{(c)} = (\beta_{km1}^{(c)}, \ldots, \beta_{kmP}^{(c)})'$, $m = 1, \ldots, (M_k - 1)$, is the parameter vector for the mth category of the kth classification task, and $\beta_{k0}^{(c)}$ is the intercept.

Assuming that the measurements for the K_c output variables are collected for the same set of N samples as in the regression tasks, we expand each output data y_{ki} for the kth task of the ith sample into a set of M_k binary variables $\boldsymbol{y}'_{ki} = (y_{k1i}, \ldots, y_{k(M_k)i})$, each y_{kmi}, $m = 1, \ldots, M_k$, takes value 1 if the ith sample for the kth classification task belongs to the mth category and value 0 otherwise, and thus $\sum_m y_{kmi} = 1$. Using the observations for the output variable in this representation and the shared input data $\mathbf{X}$, one can estimate the parameters $\boldsymbol{\beta}_{km}^{(c)}$'s by minimizing the negative log-likelihood given as below:

$$L_c = -\sum_{i=1}^{N} \sum_{k=1}^{K_c} \left(\sum_{m=1}^{M_k-1} y_{kmi}(\beta_{k0}^{(c)} + \sum_{j=1}^{P} x_{ij}\beta_{kmj}^{(c)}) - \log\left(1 + \sum_{m=1}^{M_k-1} \exp(\beta_{k0}^{(c)} + \sum_{j=1}^{P} x_{ij}\beta_{kmj}^{(c)})\right)\right). \tag{3}$$

In a heterogeneous multitask learning problem, we form a global objective function by combining the two empirical loss functions in Equations (1) and (3):

$$L = L_r + L_c. \tag{4}$$

This is equivalent to estimating the $\boldsymbol{\beta}_k^{(r)}$'s and $\boldsymbol{\beta}_{km}^{(c)}$'s independently for each of the K tasks, assuming that there are no shared patterns in the way that each of the K output variables is dependent on the input variables. Our goal is to increase the performance of variable selection and prediction power by allowing the sharing of information among the heterogeneous tasks.

3 Heterogeneous Multitask Learning with Joint Sparse Feature Selection

Consider the case where the related tasks share the same sparsity pattern such that they have a common set of relevant input variables for both the regression and classification tasks and the amount of influence of the relevant input variables on the output may vary across the tasks, we introduce an L_1/L_qregularization to the problem of the heterogeneous multitask learning in Equation (4) as below:

$$L = L_r + L_c + \lambda P_q, \tag{5}$$

where P_q is the group penalty to the sum of linear regression loss and logistic loss and λ is a regularization parameter which determines the sparsity level and could be chosen by cross validation. We consider two extreme cases of the L_1/L_q penalty for group variable selection in our problem which are sup norm and L_2 norm across different tasks at one dimension.

$$P_\infty = \left(\sum_{j=1}^{P} \max_{k,m} \left(|\beta_{kj}^{(r)}|,\ |\beta_{kmj}^{(c)}| \right) \right) \text{ or } P_2 = \left(\sum_{j=1}^{P} |\boldsymbol{\beta}_j^{(r)}, \boldsymbol{\beta}_j^{(c)}|_{L_2} \right) \tag{6}$$

$\boldsymbol{\beta}_j^{(r)}, \boldsymbol{\beta}_j^{(c)}$ are vector of parameters over all regression and classification tasks for the jth dimension. Here the sup norm and L_2 norm over the parameters of that dimension across different tasks which can regulate the joint sparsity among tasks and the multiple tasks could be related during the optimization instead of optimizing each objective function individually.

4 Optimization Method

Different methods such as gradient descent, steepest descent, Newton's method and Quasi-Newton method can be used to solve the problem in Equation (5). Although second order methods have a fast convergence fast near the global minimum for convex objective functions, they involve computing a hessian matrix and inverting it which may be infeasible in a high-dimensional setting. The coordinate-descent method iteratively updates each element of the parameter vector one at a time, using a closed-form update equation given all of the other elements. However, since it is a first-order method, the speed of convergence becomes slow as the number of tasks and dimension increase. In [7], the truncated Newton's method that uses a preconditionor and solves the linear system instead of inverting the hessian matrix has been proposed as a fast optimization method for a very large scale problem. We adopt this interior-point method because the linear regression loss and logistic regression loss have different forms, and it is more intuitive to see how the two heterogeneous tasks affect each other by optimizing their loss functions directly.

In this section, we discuss the case of the L_1/L_∞ penalty since the same optimization method can be easily extended to handle the L_1/L_2 penalty. First, we re-write the problem of minimizing Equation (5) with the nondifferentiable L_1/L_∞ penalty in Equation (6) as

$$\begin{aligned} &\text{minimize } L_r + L_c + \lambda \sum_{j=1}^{P} u_j \\ &\text{subject to } \max_{k,m} \left(|\beta_{kj}^{(r)}|,\ |\beta_{kmj}^{(c)}| \right) < u_j, \text{ for } j = 1, \ldots, P,\ k = 1, \ldots, K_r + K_c. \end{aligned} \tag{7}$$

Further re-writing the constraints in the above problem, we obtain $2 \cdot P \cdot (K_r + \sum_{k=1}^{K_c}(M_k - 1))$ inequality constraints as follows:

$$\begin{aligned} -u_j < \beta_{kj}^{(r)} < u_j, \quad &\text{for} \quad k = 1, \ldots, K_r,\ j = 1, \ldots, P, \\ -u_j < \beta_{kmj}^{(c)} < u_j, \quad &\text{for} \quad k = 1, \ldots, K_c,\ j = 1, \ldots, P,\ m = 1, \ldots, M_k - 1. \end{aligned}$$

Using the barrier method [2], we re-formulate the objective function in Equation (7) into an unconstrained problem given as

$$\begin{aligned} L_{\text{Barrier}} = L_r + L_c + \lambda \sum_{j=1}^{P} u_j + \sum_{k=1}^{K_r} \sum_{j=1}^{P} \left(I_-(-\beta_{kj}^{(c)} - u_j) + I_-(\beta_{kj}^{(c)} - u_j) \right) \\ + \sum_{k=1}^{K_c} \sum_{m=1}^{M_k - 1} \sum_{j=1}^{P} I_-(-\beta_{kmj}^{(c)} - u_j) + I_-(\beta_{kmj}^{(c)} - u_j), \end{aligned}$$

where

$$I_-(x) = \begin{cases} 0 & x \leq 0 \\ \infty & x > 0 \end{cases}.$$

Then, we apply the log barrier function $I_-(f(x)) = -(1/t)\log(-f(x))$, where t is an additional parameter that determines the accuracy of the approximation.

Based on the above conversion, the total number of parameters needed to be estimated is $(K_r + \sum_{k=1}^{K_c}(M_k - 1))(P+1) + P$. The goal of the optimization algorithm is to find the direction towards the optimal solution which can be computed by Newton method,

$$H \begin{bmatrix} \Delta\boldsymbol{\beta} \\ \Delta\boldsymbol{u} \end{bmatrix} = -\boldsymbol{g}$$

Where $\Delta\boldsymbol{\beta}$ and $\Delta\boldsymbol{u}$ are the searching directions of the coeffients and bounding parameters. H is the hessian matrix and $\boldsymbol{g}$ is the gradient evaluated at the current parameters.

The gradient vector include three parts, regression tasks, classification tasks and additional regularization parameters. $g = [\boldsymbol{g}^{(r)}, \boldsymbol{g}^{(c)}, \boldsymbol{g}^{(u)}]^T$ where $\boldsymbol{g}^{(r)}$ has K_r components, $\boldsymbol{g}^{(c)}$ has K_c components and each component has $M_k - 1$ subcomponents which represent the gradients of the function corresponding each class of the classification tasks.

$$H = \begin{bmatrix} R & 0 & D^{(r)} \\ 0 & L & D^{(c)} \\ D^{(r)} & D^{(c)} & F \end{bmatrix}$$

Where R and L are second derivatives of the parameters β for regression tasks in the form of, $R = \nabla^2 L_r + \nabla^2 P_g|_{\partial\beta^{(r)}\partial\beta^{(r)}}$, $L = \nabla^2 L_c + \nabla^2 P_g|_{\partial\beta^{(c)}\partial\beta^{(c)}}$, $D = \nabla^2 P_g|_{\partial\beta\partial u}$ and $F = D^{(r)} + D^{(c)}$.

The process of contructing and inverting Hessian matrix is the most time consuming part in the whole algorithm. But we can use a preconditionor $diag(H)$ and apply the preconditioned conjugate gradient algorithm to compute the searching direction which could make the algorithm more scalable and faster. The optimization algorithm is given by:
Let $\boldsymbol{\Theta}$ denote the set of parameters $\boldsymbol{\beta}_k^{(r)}$'s and $\boldsymbol{\beta}_{km}^{(c)}$'s. Given a strictly feasible $\boldsymbol{\Theta}$, $t = t^{(0)} > 0$, $\mu > 1$, and tolerance $\epsilon > 0$, we iterate the following steps until convergence.

Step 1 Compute $\boldsymbol{\Theta}^*(t)$ by minimizing L_{Barrier}, starting at $\boldsymbol{\Theta}$.

Step 2 Update: $\boldsymbol{\Theta} := \boldsymbol{\Theta}^*(t)$

Step 3 Stopping criterion: quit if $m/t < \epsilon$ where m is the number of constraint functions.

Step 4 Increase t: $t := t\mu$

In **Step 1**, we use the Newton method to minimize L_{Barrier} at t. In each iteration, we increase t in **Step 4**, so that we have a more accurate approximation of $I_-(u)$ through $I_-(f(x)) = -(1/t)\log(-f(x))$.

5 Experiment Result

We demonstrate our methods for heterogeneous multitask learning with L_1/L_∞ and L_1/L_2 regularizations on simulated and asthma datasets, and compare their performances with those from solving two types of multitask-learning problems for regressions and classifications separately. We show our method on both simulated and asthma datasets.

5.1 Simulation Study

In the context of genetic association analysis, we simulate the input and output data with known model parameters as follows. We start from the 120 haplotypes of chromosome 7 from the population of European ancestry in HapMap data, and randomly mate the haplotypes to generate genotype data for 500 individuals. We randomly select 50 SNPs across the chromosome as inputs. In order to simulate the paramters $\boldsymbol{\beta}_k^{(r)}$'s and $\boldsymbol{\beta}_{km}^{(c)}$'s, we assume 6 regression tasks and 1 classification task one with 5 categories and choose five common SNPs from the total of 50 SNPs as relevant covariates across all of the tasks. We fill the non-zero entries in the regression coefficients $\boldsymbol{\beta}_k^{(r)}$'s with values uniformly distributed in the interval (a, b) with $5 \leq a, b \leq 10$, and the non-zero entries in the logistic regression parameters $\boldsymbol{\beta}_{km}^{(c)}$'s such that the five categories are separated in the output space. Given these inputs and the model parameters, we generate the output values using the noise for regression tasks distributed as $N(0, \sigma^2_{\text{sim}})$. In the classification task, we expand the single output into five dummy variables representing different categories which take values of 0 or 1 depending on which category each sample belongs to. We repeat this whole process of simulating inputs and outputs to obtain 50 datasets, and report the results averaged over these datasets.

The regularization paths of the different multitask-learning methods with an L_1/L_∞ regularization obtained from a single simulated dataset are shown in Figure 1. The results from learning all of the

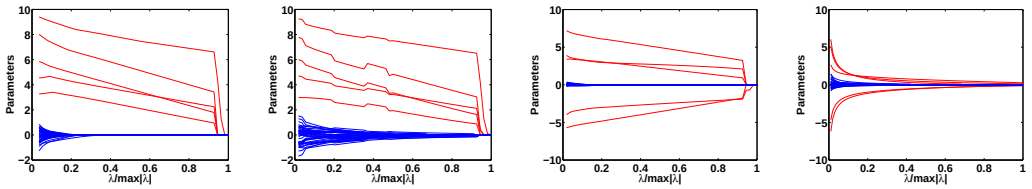


Figure 1: The regularization path for L_1/L_∞-regularized methods. (a) Regression parameters estimated from the joint regression-classification method, (b) regression parameters estimated from regression tasks only, (c) logistic-regression parameters estimated from the joint regression-classification method, and (d) logistic-regression parameters estimated from classification tasks only. The blue curves show parameters for irrelevant inputs, and the red curves for relevant inputs.

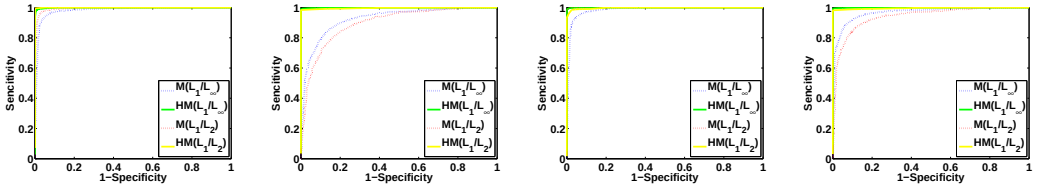
tasks jointly are shown in Figures 1(a) and 1(c) for regression and classification tasks, respectively, whereas the results from learning the sets of regression and classification tasks separately are shown in Figures 1(b) and 1(d). The red curves indicate the parameters for true relevant inputs, and the blue curves for true irrelevant inputs. We find that when learning both types of tasks jointly, the parameters of the irrelevant inputs are more reliably set to zero along the regularization path than learning the two types of tasks separately.

In order to evaluate the performance of the methods, we use two criteria of sensitivity/specificity plotted as receiver operating characteristic (ROC) curves and prediction error on test data. To obtain ROC curves, we estimate the parameters, sort the input-output pairs according to the magnitude of the estimated $\boldsymbol{\beta}_{kj}^{(r)}$'s and $\boldsymbol{\beta}_{kmj}^{(c)}$'s, and compare the sorted list with the list of true correlated input-output pairs.

We vary the sample size to $N = 100$ and 200, and show the ROC curves for detecting true relevant inputs using different methods in Figure 2. We use $\sigma_{\text{sim}} = 1$ to generate noise in the regression tasks. Results for the regression and classification tasks with $N = 100$ are shown in Figures 2(a) and (b), respectively, and similarly, the results with $N = 200$ in Figures 2(c) and (d). The "M" and "HM" represent multitask learning and heterogeneous multitask learning methods respectively. The results with L_1/L_∞ penalty are shown with color blue and green to compare the homogeneous and heterogeneous method. Red and yellow are results using the L_1/L_2 penalty. Although the performance of learning the two types of tasks separately improves with a larger sample size, the joint estimation performs significantly better for both sample sizes. A similar trend can be seen in the prediction errors for the same simulated dataset in Figure 3.

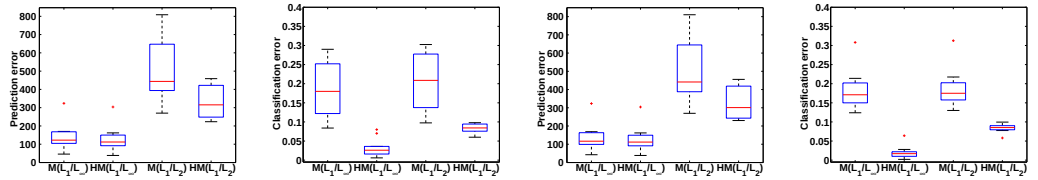
In order to see how different signal-to-noise ratios affect the performance, we vary the noise level to $\sigma_{\text{sim}}^2 = 5$ and $\sigma_{\text{sim}}^2 = 8$, and plot the ROC curves averaged over 50 datasets with a sample size $N = 300$ in Figure 4. Our results show that for both of the signal-to-noise ratios, learning regression and classification tasks jointly improves the performance significantly. The same observation can be made from the prediction errors in Figure 5. We can see that the L_1/L_2 method tends to improve the variable selection, but the tradeoff is that the prediction error will be high when the noise level is low. While L_1/L_∞ has a good balance between the variable selection accuracy and prediction error at a lower noise level. As the noise increases, the L_1/L_2 outperforms L_1/L_∞ in both variable selection and prediction accuracy.

From the above analysis, we can see that the heterogeneous model can significantly improve the performance of parameter estimation, variable selection and prediction accuracy. Different regularization method shows different levels of improvement in the above three evaluation criteria. The reason for the difference between L_1/L_∞ and L_1/L_2 is that the L_2 norm is greater than L_∞, so it tends to make the relevant groups of variables more significant than the L_∞ norm. The disadvantage is that the larger the penalty is, the estimation is more penalized. This is the reason you see the prediction error is very high for the L_1/L_2 method. The smaller penalty of L_1/L_∞ penalty makes the relevant covariates less significant among all the covariates. So the performance of variable selection and prediction decrease under the high noise level. But if there are both correlated regression and classification tasks, the heterogeneous model can improve the performance of all tasks.



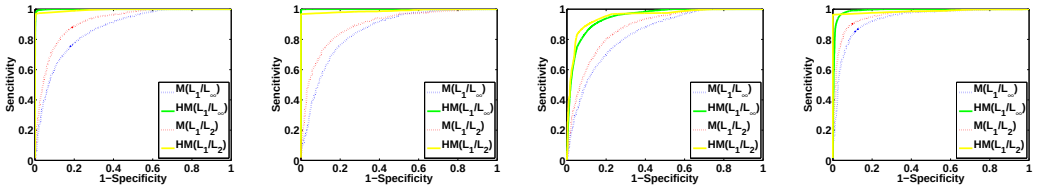
(a) (b) (c) (d)

Figure 2: ROC curves for detecting true relevant input variables when the sample size N varies. (a) Regression tasks with $N = 100$, (b) classification tasks with $N = 100$, (c) regression tasks with $N = 200$, and (d) classification tasks with $N = 200$. Noise level $N(0,1)$ was used. The joint regression-classification methods achieve nearly perfect accuracy, and their ROC curves are completely aligned with the axes.



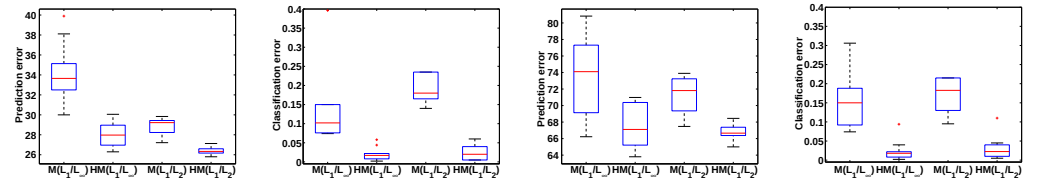
(a) (b) (c) (d)

Figure 3: Prediction errors when the sample size N varies. (a) Regression tasks with N=100, (b) classification tasks $N = 100$, (c) regression tasks $N = 200$, and (d) classification tasks $N = 200$. Noise level $N(0,1)$ was used.



(a) (b) (c) (d)

Figure 4: ROC curves for detecting true relevant input variables when the noise level varies. (a) Regression tasks with noise level $N(0, 5)$, (b) classification tasks with noise level $N(0, 5)$, (c) regression tasks with noise level $N(0, 8)$, and (d) classification tasks with noise level $N(0, 8)$. Sample size N=300 was used.

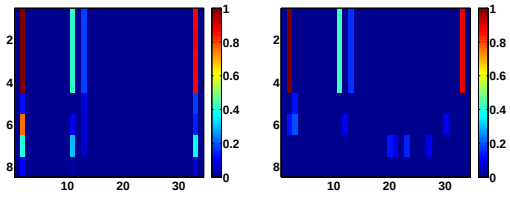


(a) (b) (c) (d)

Figure 5: Prediction errors when the noise level varies. (a) Regression tasks with noise level $N(0, 5^2)$, (b) classification tasks with noise level $N(0, 5^2)$, (c) regression tasks with noise level $N(0, 8^2)$, and (d) classofication tasks with noise level $N(0, 8^2)$. Sample size N=300 was used.

5.2 ASTHMA DATASET

We apply our method to the asthma dataset with 34 SNPs in the IL4R gene of chromosome 11 and five asthma-related clinical traits collected over the 613 patients. The set of traits include four continuous-valued traits related to lung physiology such as baseline predrug FEV1, maximum



(a) (b)

Figure 6: Results from the analysis of asthma dataset for discovery of causal SNPs for the correlated phenotypes. (a) Joint regression-classification method, and (b) separate analysis of multitask regressions and multitask classifications.

FEV1, baseline predrug FVC, and maximum FVC as well as a single discrete-valued trait with five categories. The goal of this analysis is to discover whether any of the SNPs (inputs) are influencing each of the asthma-related traits (outputs). We use 400 samples as training data, and the remaining 213 samples as test data. We fit the joint regression-classification method with L_1/L_∞ and L_1/L_2 regularizations, and compare the results from fitting L_1/L_∞ and L_1/L_2 regularized methods only for the regression tasks or only for the classification task. We show the estimated parameters for the joint learning in Figure 6(a) and the separate learning in Figure 6(b), where the first four rows correspond to the four regression tasks, the next four rows are parameters for the four dummy variables of the classification task, and the columns represent SNPs. We can see that the heterogeneous multitask-learning method encourages to find common causual SNPs for the multiclass classification task and the regression tasks.

6 CONCLUSIONS

In this paper, we proposed a method for a recovery of union support in heterogeneous multitask learning, where the set of tasks consists of both regressions and classifications. In our experiments with simulated and asthma datasets, we demonstrated that using L_1/L_2 or L_1/L_∞ regularizations in the joint regression-classification problem improves the performance in identifying the input variables that are commonly relavant to multiple tasks.

The union support recovery as was presented in this paper is concerned with finding inputs that influence at least one task. In the real-world problem of association mapping, there is a clustering structure such as co-regulated genes, and it would be interesting to discover SNPs that are causal to at least one of the genes within the subgroup rather than all of the genes. In addition, SNPs in a region of chromosome are often correlated with each other because of the non-random recombination process during inheritance, and this correlation structure, called linkage disequilibrium, has been actively investigated. A promising future direction is to model this complex correlation pattern in both the input and output space within our framework.

Acknowledgments LS is supported by a Ray and Stephenie Lane Research Fellowship. EPX is supported by grant ONR N000140910758, NSF DBI-0640543, NSF DBI-0546594, NSF IIS-0713379 and an Alfred P. Sloan Research Fellowship. We also thank Grace Tzu-Wei Huang for helpful discussions.

References

[1] B. Bakker and T. Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, 2003.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[4] V. Emilsson, G. Thorleifsson, B. Zhang, A.S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G.B. Walters, S. Gunnarsdottir, et al. Variations in dna elucidate molecular networks that cause disease. *Nature*, 452(27):423–28, 2008.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. Technical Report 703, Department of Statistics, Stanford University, 2009.

[6] S. Kim and E. P. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genetics*, 5(8):e1000587, 2009.

[7] K. Koh, S. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, 8(8):1519–1555, 2007.

[8] Sara van de Geer Lukas Meier and Peter Buhlmann. The group lasso for logistic regression. *J. R. Statist. Soc. B*, 78(1):53–71, 2008.

[9] G. Obozinski, M.J. Wainwright, and M.J. Jordan. High-dimensional union support recovery in multivariate regression. In *Advances in Neural Information Processing Systems 21*, 2008.

[10] M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for l_1 regularization: a comparative study and two new approaches. In *Proceedings of the European Conference on Machine Learning*, 2007.

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

[12] Bernd Fischer Volker Roth. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *the 25th international conference on Machine learning*, 2009.

[13] K. Yu, V. Tresp, and A. Schwaighofer. Learning gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

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

[15] J. Zhang, Z. Ghahramani, and Y. Yang. Flexible latent variable models for multi-task learning. *Machine Learning*, 73(3):221–242, 2008.

[16] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. Technical Report 703, Department of Statistics, University of California, Berkeley, 2008.

[17] J. Zhu, B. Zhang, E.N. Smith, B. Drees, R.B. Brem, L. Kruglyak, R.E. Bumgarner, and E.E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40:854–61, 2008.