A Two-Graph Guided Multi-task Lasso Approach for eQTL Mapping

Xiaohui Chen[1,3,*], **Xinghua Shi**[2,*], **Xing Xu**[1], **Zhiyong Wang**[1],
Ryan Mills[2], **Charles Lee**[2,†], **Jinbo Xu**[1,†]
1. Toyota Technological Institute at Chicago
2. Brigham Women's Hosptial and Harvard Medical School
3. The University of British Columbia

Abstract

Learning a small number of genetic variants associated with multiple complex genetic traits is of practical importance and remains challenging due to the high-dimensional nature of data. In this paper, we proposed a two-graph guided multi-task Lasso to address this issue with an emphasis on estimating *subnetwork-to-subnetwork* associations in expression quantitative trait loci (eQTL) mapping. The proposed model can learn such *subnetwork-to-subnetwork* associations and therefore can be seen as a generalization of several state-of-the-art multi-task feature selection methods. Additionally, this model has a nice property of allowing flexible structured sparsity on both feature and label domains. Simulation study shows the improved performance of our model and a human eQTL data set is analyzed to further demonstrate the applications of the model.

1 Introduction

Recent advances in biotechnologies, including next generation DNA and RNA sequencing, have resulted in the generation of a large amount of genomic and transcriptomic data. One active research area of integrating these genomic and transcriptomic datasets is to identify expression quantitative trait loci (eQTLs) through eQTL mapping. eQTL mapping seeks for a set of statistically significant associations between genetic variants and gene expressions. The challenge of eQTL mapping lies in the fact that there are a large number of genetic variants and gene expression traits, and hence the search space for potential eQTLs is vast.

A widely used approach to detect eQTL associations is to calculate pair-wise correlations of the genotypes or intensities of genetic variants with the expression profiles of genes in the neighborhood of these variants [1, 2]. However, this approach assumes that genetic variants are independent and gene expressions are not correlated. This assumption will inevitably miss many complex yet observed cases where multiple genetic variants jointly affect the co-expressions of multiple genes, i.e., *subnetwork-to-subnetwork* associations. Moreover, given the large number of tests performed using such pair-wise correlation analysis, this approach is subject to the burden of multiple test correction which may introduce false positive associations.

In eQTL analysis, both the feature matrix and the label matrix are usually high-dimensional, with the number of features (i.e. genetic variants) and the number of labels (i.e. gene expressions) significantly larger than the number of samples. Therefore, the problem of eQTL mapping can be formed to a classical feature selection problem and Lasso-based methods have therefore proposed. Kim and Xing [3] presented a graph-weighted fused Lasso approach to estimate genetic variants (e.g. SNPs) that perturb a subset of highly correlated traits. Tree-guided group Lasso [4] was formulated for multi-task regression that utilizes structured sparsity to learn the associations between genetic variants and groups of co-

[*] These authors contribute equally.
[†] To whom correspondence should be addressed.

Appearing in Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

expressed genes. Another Lasso approach, adaptive multi-task Lasso, was proposed to detect eQTL associations and this model considers the correlation among traits while incorporating the priors on SNPs such as regulatory features for these SNPs [5].

Nevertheless, none of these existing methods consider a more general question: how multiple genetic variants in a biological process or pathway, by forming a subnetwork, jointly affect a subnetwork of multiple correlated traits; see Fig. 1(a) for an illustration. These subnetworks under investigation can overlap and their sizes may vary. In this paper, we formulate this *subnetwork-to-subnetwork* association problem into a two-graph guided multi-task Lasso model to capture the observation that multiple genetic variants jointly affect correlated traits. Moreover, the proposed model allows overlapped subnetworks in associations. This novel model has flexible structured sparsity as illustrated in Fig. 1(b). On the one hand, the model can induce sparsity on the association coefficients; on the other hand, it can bias the learned sparsity pattern to the prior networks in both label and feature domains. Therefore, our proposed model can be viewed as a generalization of several state-of-the-art multi-task feature selection methods [3, 4, 5, 6] by utilizing prior information on both feature and label graphs.

The rest of the paper is organized as follows. In Section 2, the two-graph guided multi-task lasso model is formulated and introduced. We then present a coordinate-descent algorithm to obtain the numeric estimates of the model in Section 3. In Section 4, asymptotic properties of the proposed model are studied. Simulations are carried out in Section 5 to show that our model outperforms several other multi-task sparse learning models. A real eQTL data is further analyzed as an example of the applications of our model and results are presented in Section 6. The paper is concluded in Section 7.

2 Model

Suppose K traits or labels are collected for n subjects and we denote these measurements by $Y_{n\times K}$. We further assume that each trait is potentially associated with J genetic variants or features. Specifically, the association model considered here is the following multiple-input-multiple-output (MIMO) linear system

$$Y = XB + E, \tag{1}$$

where $B_{J\times K} = \{\mathbf{b}^1, \cdots, \mathbf{b}^K\}$ is the association coefficient matrix denoting the connection strengths between traits and genetic variants and E is a Gaussian white-noise term with constant variance σ^2. Here, X is an $n \times J$ matrix, where each row contains quantitative measurements for the J genetic variants and each column contains n observations for one genetic variant. For high-dimensional problems where K and J are large, (1) is well-posed only if certain regularization is introduced. In association studies, *sparsity* is a reasonable assumption since we expect only a small fraction of genetic variants are associated with gene expressions. Thus, the association study is reduced to a classical feature selection problem. To this end, a standard multi-task lasso proposed in [5]

$$\text{minimize}_B \qquad \sum_{k=1}^{K} \|\mathbf{y}^k - X\mathbf{b}^k\|_2^2 + \lambda \sum_{j=1}^{J} \delta_j \|\mathbf{b}_j\|_2, \tag{2}$$

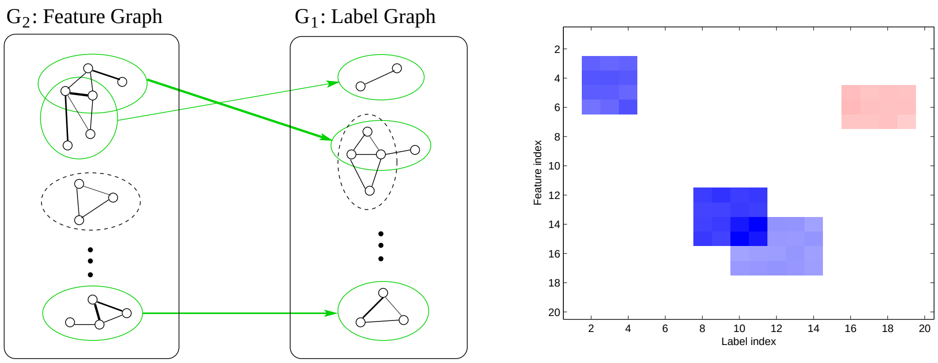
where $\mathbf{b}^k$ is the k-column of B representing the association coefficients of all genetic variants to the k-th trait and $\mathbf{b}_j$ is the j-th row of B meaning the association strengths of the j-th genetic variant to all traits.

Several extensions of the multi-task lasso model (2) have been proposed in literature. For instance, the graph guided multi-task lasso [3, 6] was designed in the following way in association studies. Let $G = (V, E)$ be a graph where V is the set of vertices and E is the set of edges; then the graph-guided multi-task lasso is defined as the solution of

$$\begin{aligned} \text{minimize}_B \qquad & \sum_{k=1}^{K} \|\mathbf{y}^k - X\mathbf{b}^k\|_2^2 + \lambda \|B\|_1 \\ & + \gamma \sum_{e_{m,l}\in E} w(e_{m,l}) \sum_{j=1}^{J} |b_{jm} - \text{sign}(r_{m,l}) b_{jl}|, \end{aligned} \tag{3}$$

where $w(e_{m,l})$ is a weight assigned to the edge $e_{m,l}$ in graph E and $r_{m,l}$ is the correlation between $\mathbf{y}^m$ and $\mathbf{y}^l$. Such a graph guided multi-task model (3) can learn the associations between one particular genetic variant and a group of traits. Alternatively, Kim and Xing [4] leveraged the idea that a co-expressed set of genes should share a larger common set of genetic variants and thus proposed another one-to-many association model, namely the tree guided lasso where the tree structure can be user-specified or a hierarchical clustering tree on labels. Note that the tree guided lasso is an extension of a group-lasso model.

In this paper, we consider a more general framework that subsumes all aforementioned models as



(a)

(b)

Figure 1: Illustrations of subnetwork-subnetwork associations and structured sparsity for modeling these associations. (a) Subnetwork-subnetwork associations. Ellipses represent the subnetworks of feature and label graphs, with green ones highlighting associated subnetworks. Green lines between feature and label subnetworks represent identified associations. (b) Structured sparsity for modeling subnetwork-subnetwork associations. Blue and red blocks represent positive and negative associations respectively.

special cases of the new model. We propose a multi-task lasso model to learn eQTL mapping by incorporating structural information on both genetic variants and labels. Specifically, let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two graphs defined on labels and features, respectively. Then the two-graph guided multi-task lasso is defined as

$$\begin{aligned} &\text{minimize}_B \qquad \|Y - XB\|_F^2 + \lambda\|B\|_1 \\ &+ \gamma_1 \times pen_1(E_1, B) + \gamma_2 \times pen_2(E_2, B), \end{aligned} \tag{4}$$

where pen_1 and pen_2 are two penalty functions measuring the discrepancy between the prior label and feature graphs and the association pattern. Here, we simultaneously consider two symmetric penalty functions on features and labels. In particular, as in [3], we design the penalty functions on the label and feature graphs as in the following form

$$\begin{aligned} pen_1(E_1, B) &= \sum_{e_{m,l} \in E_1} w(e_{m,l}) \sum_{j=1}^{J} |b_{jm} - \text{sign}(r_{m,l}) b_{jl}| \\ pen_2(E_2, B) &= \sum_{e_{f,g} \in E_2} w(e_{f,g}) \sum_{k=1}^{K} |b_{fk} - \text{sign}(r_{f,g}) b_{gk}|, \end{aligned} \tag{5}$$

where the weight $w(\cdot)$ in our simulation is simply chosen as the absolute value of correlation. The penalty above is closely related to that in fused lasso [17]. In addition, our proposed model can be viewed as a generalization of the fused lasso model in the sense that fusion is dictated by the topology of input graphs, rather than physical proximity. Other penalty functions are also possible according to different problem settings. For the penalty functions in (5), the optimization problem (4) can be efficiently solved by a coordinate-descent algorithm as in [3] where the objective function of (4) is transferred into an equivalent differentiable function.

3 Algorithm

The objective function in (4) is non-differentiable and its optimization is achieved by transforming it to a series of smooth functions that can be efficiently minimized by the coordinate-descent algorithm [3]. Specifically, our algorithm works as follows. First, we consider the following constrained ridge-type optimization

$$\begin{aligned} &\text{minimize}_{B, d_{jk}, d1_{jml}, d2_{kfg}} \qquad \|Y - XB\|_F^2 \\ &+ \lambda \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{b_{jk}^2}{d_{jk}} \\ &+ \gamma_1 \sum_{e_{m,l} \in E_1} w^2(e_{m,l}) \sum_{j=1}^{J} \frac{(b_{jm} - \text{sign}(r_{m,l}) b_{jl})^2}{d1_{jml}} \\ &+ \gamma_2 \sum_{e_{f,g} \in E_2} w^2(e_{f,g}) \sum_{k=1}^{K} \frac{(b_{fk} - \text{sign}(r_{f,g}) b_{gk})^2}{d2_{kfg}}, \end{aligned} \tag{6}$$

subject to

$$\sum_{j,k} d_{jk} = 1, \qquad \sum_{e_{m,l} \in E_1, j} d1_{jml} = 1,$$

$$\sum_{e_{f,g} \in E_2, k} d2_{kfg} = 1, \qquad d_{jk}, d1_{jml}, d2_{kfg} \geq 0.$$

This can be analytically solved via its Lagrangian form. For an initial value of B, we optimize (6) over $d_{jk}, d1_{jml}, d2_{kfg}$ by setting their corresponding derivatives to zeros; hence we obtain

$$d_{jk} = \frac{|b_{jk}|}{\sum_{j',k'} |b_{j'k'}|}, \tag{7}$$

$$d1_{jml} = \frac{w(e_{m,l})|b_{jm} - \text{sign}(r_{m,l})b_{jl}|}{\sum_{e_{m',l'}\in E_1, j'} w(e_{m',l'})|b_{j'm'} - \text{sign}(r_{m',l'})b_{j'l'}|}, \tag{8}$$

$$d2_{kfg} = \frac{w(e_{f,g})|b_{fk} - \text{sign}(r_{f,g})b_{gk}|}{\sum_{e_{f',g'}\in E_2, k'} w(e_{f',g'})|b_{f'k'} - \text{sign}(r_{f',g'})b_{g'k'}|}. \tag{9}$$

Then conditioning on the current estimate of $d_{jk}, d1_{jml}, d2_{kfg}$, we optimize over B. The solution of this minimization can be found as

$$\begin{aligned} b_{jk} =& \Big\{ \sum_{i=1}^{n} x_{ij}(y_{ik} - \sum_{j'\neq j} x_{ij'} b_{j'k}) \\ &+ \gamma_1 \sum_{e_{m,k}\in E_1} w^2(e_{m,k}) \frac{b_{jm}\text{sign}(r_{m,k})}{d1_{jmk}} \\ &+ \gamma_1 \sum_{e_{k,l}\in E_1} w^2(e_{k,l}) \frac{b_{jl}\text{sign}(r_{k,l})}{d1_{jkl}} \\ &+ \gamma_2 \sum_{e_{f,j}\in E_2} w^2(e_{f,j}) \frac{b_{fk}\text{sign}(r_{f,j})}{d2_{kfj}} \\ &+ \gamma_2 \sum_{e_{j,g}\in E_2} w^2(e_{j,g}) \frac{b_{gk}\text{sign}(r_{j,g})}{d2_{kjg}} \Big\} \\ & \Big/ \Big\{ \sum_{i=1}^{n} x_{ij}^2 + \frac{\lambda}{d_{jk}} \\ &+ \gamma_1 \sum_{e_{m,k}\in E_1} \frac{w^2(e_{m,k})}{d1_{jmk}} + \gamma_1 \sum_{e_{k,l}\in E_1} \frac{w^2(e_{k,l})}{d1_{jkl}} \\ &+ \gamma_2 \sum_{e_{f,j}\in E_2} \frac{w^2(e_{f,j})}{d2_{kfj}} + \gamma_2 \sum_{e_{j,g}\in E_2} \frac{w^2(e_{j,g})}{d2_{kjg}} \Big\}. \end{aligned}$$

These two steps alternate until $\left\| B^{(t+1)} - B^{(t)} \right\|_1 \leq \varepsilon$ for some small $\varepsilon > 0$.

We remark that the coordinate-descent algorithm can be seen as a concrete algorithm in the majorization-minimization (MM) paradigm proposed by [11]. Indeed, we observe that, (4) is equivalent to a slightly modified Lagrangian version by squaring each penalty terms

$$\begin{aligned} \text{minimize}_B \quad & \|Y - XB\|_F^2 + \lambda \|B\|_1^2 \\ &+ \gamma_1 \left(\sum_{e_{m,l}\in E_1} w(e_{m,l}) \sum_{j=1}^{J} |b_{jm} - \text{sign}(r_{m,l})b_{jl}| \right)^2 \\ &+ \gamma_2 \left(\sum_{e_{f,g}\in E_2} w(e_{f,g}) \sum_{k=1}^{K} |b_{fk} - \text{sign}(r_{f,g})b_{gk}| \right)^2 . \end{aligned} \tag{10}$$

Note that, for any $\mathbf{w} = (w_i)$ such that $\|\mathbf{w}\|_1 = 1$ and $w_i \geq 0$,

$$\begin{aligned} \|\mathbf{b}\|_1 = \sum_i |b_i| &= \sum_i \sqrt{w_i} \frac{|b_i|}{\sqrt{w_i}} \\ &\leq (\sum_i w_i)^{1/2} (\sum_i b_i^2/w_i)^{1/2} = (\sum_i b_i^2/w_i)^{1/2}, \end{aligned} \tag{11}$$

where we use the Cauchy-Schwarz inequality. The chain inequalities in (11) holds trivially when some elements of $\mathbf{w}$ are zeros (i.e. sparsity) since the RHS equals to ∞. This implies that (6) is an upper envelop function of (4) over an arbitrary B. Moreover, the equality of (11) is attained when $\mathbf{w} = \mathbf{b}/\|\mathbf{b}\|_1$. Therefore, it follows that the update equations, (7),(8),(9) for $d, d1, d2$ respectively, are direct consequences of the monotonic descent property of the MM algorithm.

Tuning parameters $\lambda, \gamma_1, \gamma_2$ are determined by K-fold cross-validations (CVs). Since an exhaustive search of the optimal triplet on a three-dimensional lattice is computationally infeasible for large-scale multi-task learning problems such as the eQTL mapping with a large number of genetic variants and genes, we adopt a gradient-descent approach proposed in [3] to iteratively update $(\lambda, \gamma_1, \gamma_2)$. Particularly, three line searches in the descent direction of minimizing the current CV error are sequentially applied to each component in $(\lambda, \gamma_1, \gamma_2)$ while holding the other two components. The coordinate gradients for the three components are approximated by their finite differences. Therefore, the tuning procedure contains those alternating steps, where each step corresponds to learn a multi-task Lasso proposed in this paper.

4 Asymptotic Properties

In this section, we present the asymptotic properties of the proposed model, where the sample size is large enough. The number of genetic variants and traits are assumed to be fixed and we allow the number of observations or the sample size $n \to \infty$. We also allow that λ, γ_1, and γ_2 depend on n; however, we shall suppress this implicit dependency in the following notation. We now establish the asymptotic normality of the proposed two-graph guided multi-task lasso estimator.

Theorem 4.1. *Assume $n^{-1}X^TX \to C$ for some positive definite matrix C, $\lambda/\sqrt{n} \to \lambda_0$, $\gamma_i/\sqrt{n} \to \gamma_{0i}$ for $i = 1, 2$, $r_{m,l} \xrightarrow{P} c_{m,l}$ for all $m, l \in \{1, \cdots, K\}$, and $r'_{f,g} \xrightarrow{P} c'_{f,g}$ for all $f, g \in \{1, \cdots, J\}$. Let $\hat{B}_n$ be the two-graph guided multi-task lasso estimator in (4). Then we have*

$$\sqrt{n}(\hat{B}_n - B) \xrightarrow{d} \arg\min(V), \tag{12}$$

with $V : \mathbb{R}^{p \times p} \to \mathbb{R}$ as a random function defined by

$$\begin{aligned}
V(U) &= -2\sum_{k=1}^{K} \mathbf{u}^{k^T}\mathbf{w}_k + \sum_{k=1}^{K} \mathbf{u}^{k^T} C \mathbf{u}^k \\
&+ \lambda_0 \sum_{j=1}^{J}\sum_{k=1}^{K} [u_{jk}\mathit{sign}(b_{jk})\mathbb{I}(b_{jk} \neq 0) + |u_{jk}|\mathbb{I}(b_{jk} = 0)] \\
&+ \gamma_{01} \sum_{e_{m,l} \in E_1} w(e_{m,l}) \sum_{j=1}^{J} \Big[(u_{jm} - \mathit{sign}(c_{m,l})u_{jl}) \\
&\quad \times \mathit{sign}(b_{jm} - \mathit{sign}(c_{m,l})b_{jl})\mathbb{I}(b_{jm} \neq \mathit{sign}(c_{m,l})b_{jl})\Big] \\
&+ \gamma_{01} \sum_{e_{m,l} \in E_1} w(e_{m,l}) \sum_{j=1}^{J} \Big[|u_{jm} - \mathit{sign}(c_{m,l})u_{jl}| \\
&\quad \times \mathbb{I}(b_{jm} = \mathit{sign}(c_{m,l})b_{jl})\Big] \\
&+ \gamma_{02} \sum_{e_{f,g} \in E_2} w(e_{f,g}) \sum_{k=1}^{K} \Big[(u_{fk} - \mathit{sign}(c'_{f,g})u_{gk}) \\
&\quad \times \mathit{sign}(b_{fk} - \mathit{sign}(c'_{f,g})b_{gk})\mathbb{I}(b_{fk} \neq \mathit{sign}(c'_{f,g})b_{gk})\Big] \\
&+ \gamma_{02} \sum_{e_{f,g} \in E_2} w(e_{f,g}) \sum_{k=1}^{K} \Big[|u_{fk} - \mathit{sign}(c'_{f,g})u_{gk}| \\
&\quad \times \mathbb{I}(b_{fk} = \mathit{sign}(c'_{f,g})b_{gk})\Big] \qquad (13)
\end{aligned}$$

and $\mathbf{w}_k \sim N(\mathbf{0}, \sigma^2 C)$ are i.i.d.random J-dimensional vectors.

The theorem proof is relegated to the appendix and it is an application of standard epi-convergence in finite dimensions as used in [7].

An immediate consequence of Theorem 4.1 is that when $\lambda, \gamma_1, \gamma_2$ all grow at appropriate rates as the sample size increases, the proposed estimator converges in probability to the true coefficient matrix. Hence we have

Corollary 4.2. *If $\max\{\lambda, \gamma_1, \gamma_2\} = o(\sqrt{n})$, then $\hat{B}_n$ is a $\sqrt{n}$-consistent estimator of B.*

5 Simulation Study

In this section, we carry out simulations to demonstrate the performance of the proposed two-graph guided multi-task Lasso (MTLasso 2G). We compare our model, including a special case of our model, the feature-graph guided multi-task Lasso (MTLasso FG) which only contains the second penalty term in (5), with state-of-the-art feature selection models including: Lasso, multi-task/group lasso (MTLasso) and label-graph guided multi-task Lasso (MTLasso LG) [6, 3]. We remark that all of these models are special cases of our proposed model by fixating particular penalty weight to zero. For example, setting $\gamma_1 = \gamma_2 = 0$ yields the Lasso model. Hence, our model is in fact a more flexible and general framework for multi-task feature selection.

5.1 Data Generation

We study the performances of various models with a range of setups, each of which is represented by (K, J, n). We are particularly interested in the scenarios where $J \gg n$ and $K \gg 1$. We set the three parameters (K, J, n) as $K \in \{10, 50\}$, $J \in \{100, 200, 300, 400, 500\}$, $n = 50$ and consider all the 10 possible setups based on the combinations of these parameters. For each setup, we generate 50 data sets and compare the averaged performances of those aforementioned models on the generated data.

The simulation data is generated by considering high correlations among genetic variants as seen in real data. We first randomly sample feature and label subnetworks as groups. The number of groups sampled at features and labels should be equal so that a series of one-to-one mappings between label and feature groups can be determined by group ranks. The association matrix B is set to be binary, where

$$b_{jk} = \begin{cases} 1, & k \in g_1^t \text{ and } j \in g_2^t \\ 0, & \text{otherwise} \end{cases}.$$

g_1^t represents the t-th group sampled from label domain and g_2^t is the t-th group sampled from features. Then, the input feature matrix X is a randomly generated matrix multiplied by a covariate matrix to make sure the features in the same group in G_2 have relatively high correlations. The input label matrix Y is the product of X and B with an additive Gaussian noise as in (1). We remark that the data simulation process is independent of our model. Instead, the simulation data is generated considering highly correlated genetic variants as seen in the real data.

5.2 Results

We compare the performance of various models with the guidance of label and/or feature graphs, if applicable. Our goal here is evaluate the capability of those models to correctly identify the association patterns subject to erroneous or noisy prior graphs, which we define as *subnetwork pruning*. Here, the correlation graphs on labels (with cutoff 0.4) and features (with cutoff 0.6) are used and the weights are the absolute values of correlations. We consider the performance measures in terms of the (vectorized) ℓ^1-norm errors and the areas under the precision-recall curve (AUCs). The precision-recall curve is calculated by varying the cutoff parameter of the thresholding procedure on the estimated coefficients from all models in a post-hoc manner. Therefore, this AUC measures the averaged performance of the *optimal* model for detecting the sparse association pattern. Our results on the estimation error and AUC are shown in Fig. 2. Several observations can be drawn here.

First, in terms of ℓ^1 estimation error, the proposed two-graph guided multi-task Lasso (MTLasso 2G) and feature-graph guided multi-task Lasso (MTLasso FG) uniformly out-perform the label-graph guided multi-task Lasso (MTLasso LG), standard multi-task lasso (MTLasso), and the Lasso models. Note that MTLasso errors are too large to plot and thus skipped from Fig. 2(a) and Fig. 2(b). This result is in accordance with our intuition that, when feature graph is larger than label graph (note that we have $K < J$ in our setup), more salient information will be incorporated into the model and improve the performance. The comparison between Fig. 2(a) and Fig. 2(b) confirms our observation since the difference between MTLasso 2G/MTLasso FG and MTLasso LG becomes smaller as K increases.

Second, from the plots of the PR-AUC Fig. 2(c) and Fig. 2(d), we observe that the MTLasso 2G model is the best among others, which suggests that MTLasso 2G be more robust than MTLasso FG and MTLasso LG. Therefore, the proposed MTLasso 2G can be seen as a balanced feature selection method where it combines the advantages of MTLasso FG and MTLasso LG and improves upon both of them.

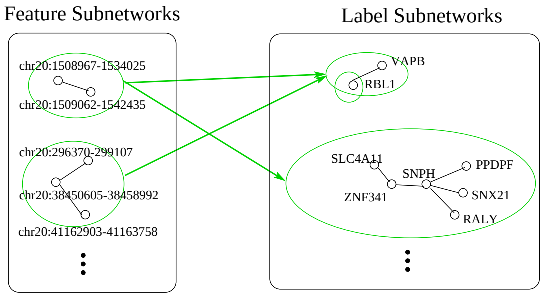


Figure 4: Examples of subnetwork-subnetwork associations on a human eQTL dataset.

Third, we also examine the Mathews correlation coefficients (MCCs) of those models in Fig. 3 by fixing a small threshold at $\sqrt{\log(JK)/(NK)}$. MCC measures the "correlation" between the estimated binary pattern and the true pattern. The same conclusions as the previous two observations can be made as seen from Fig. 3(a,b).

Finally, we give an example of the actually estimated patterns of B by those models, as illustrated in Fig. 3(c). We can see that the MTLasso 2G achieves the best performance since the other models have more false negatives.

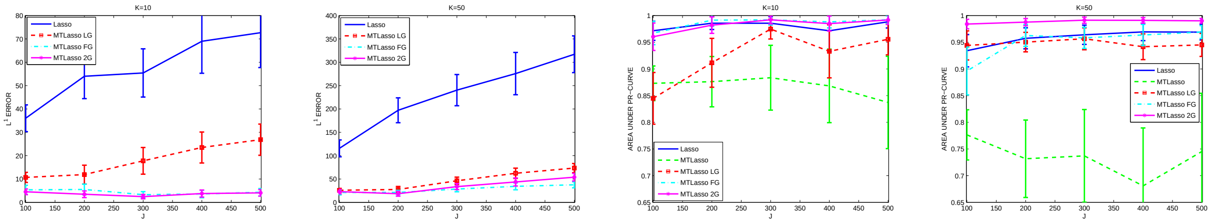
6 Human eQTL Dataset

To demonstrate an application of our proposed model on real datasets, we apply our model to a human eQTL data set.

6.1 Data

The human eQTL data set we utilize here includes a set of genetic variants in the form of copy number variants (CNVs) from the latest data release of the 1000 Genomes Project [8, 12], and gene expression profiles from the RNA sequencing data [16]. As an example, we pick chromosome 20, with 139 genotyped CNVs and 379 genes with expression data in 51 samples from Yoruba in Ibadan Nigeria(YRI). Our model can be run on any single chromosome and the computational cost is reasonable.

Previously, co-expression networks have been constructed by thresholding gene correlations from expression profiles [14]. We used a similar strategy to build a co-expression network on the labels of our data. To capture the relationships among the genetic variants, we used the same strategy to construct a genetic network on the features of our data. More specifically, both feature and label graphs are constructed using



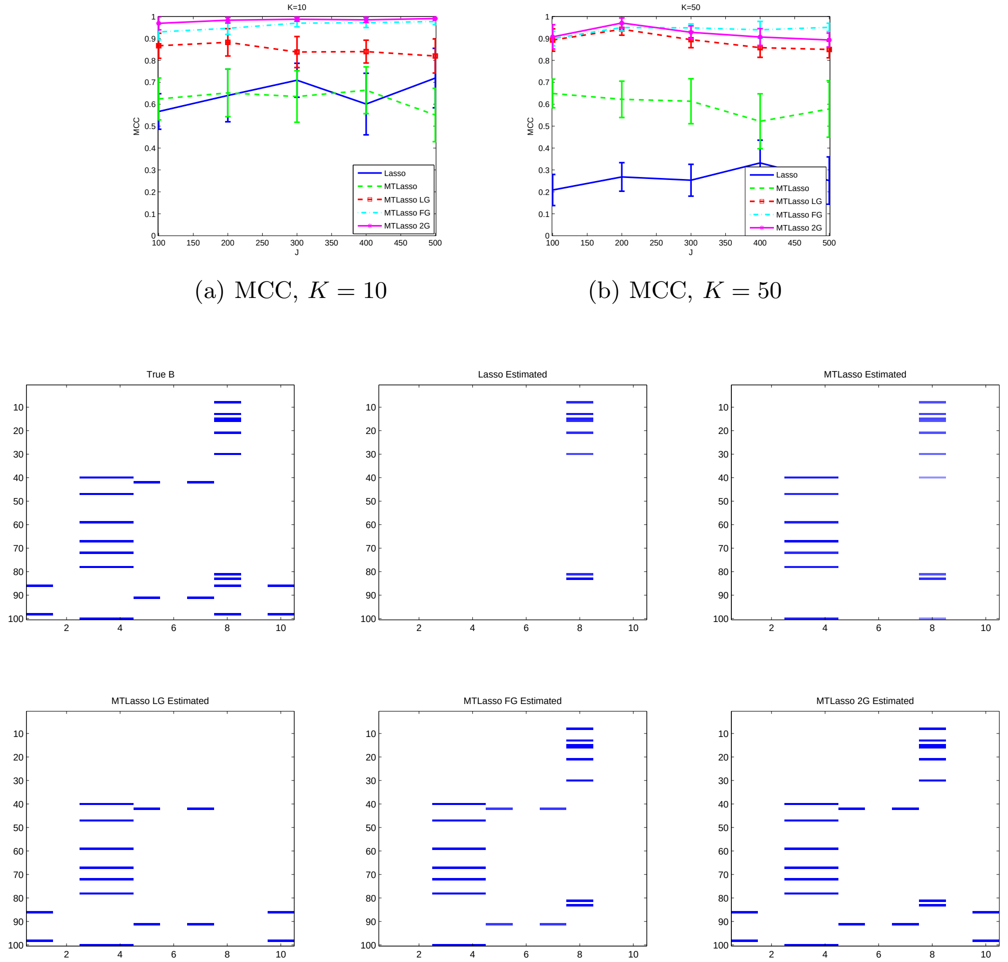
(a) ℓ_1 Error, $K = 10$

(b) ℓ_1 Error, $K = 50$

(c) AUC, $K = 10$

(d) AUC, $K = 50$

Figure 2: Averaged ℓ^1 errors and areas under the precision-recall curve (AUCs) of Lasso, multi-task lasso (MTLasso), label-graph guided MTLasso (MTLasso LG), feature-graph guided MTLasso (MTLasso FG), and two-graph guided MTLasso (MTLasso 2G), with $n = 50$, $K = 10, 50$, and $J = 100, 200, 300, 400, 500$. The distance between upper bar and lower bar equals to standard deviation.



(a) MCC, $K = 10$

(b) MCC, $K = 50$

(c) An example estimation result of $K = 10$, $J = 100$

Figure 3: The Mathews correlation coefficients (MCCs) of various Lasso methods and an example of the estimated patterns of B by these models, the threshold is set as $\sqrt{\log(JK)/(NK)}$.

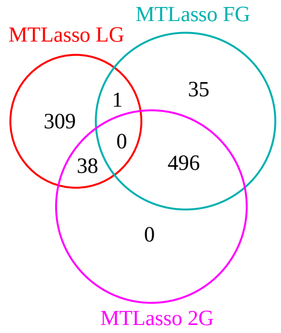


Figure 5: Venn diagram of associations found by label-graph guided MTLasso (MTLasso LG), feature-graph guided MTLasso (MTLasso FG), and two-graph guided MTLasso (MTLasso 2G).

$$E_{uv} = \begin{cases} corr(u,v), & u,v \in G_i, u \neq v, |corr(u,v)| \geq \theta \\ 0, & \text{otherwise} \end{cases} .$$

where u and v are two different nodes (features or labels) and the edge between them E_{uv} has the weight of the correlation between them if the absolute value of their correlation is above a cutoff θ. By setting θ = 0.4, a gene co-expression network is constructed as a label graph G_1 and a feature graph on CNVs G_2 is constructed to capture the correlations among CNVs. This step generates a feature graph with 144 edges and a label graph with 957 edges.

6.2 Results

Applying the proposed two-graph guided multi-task model to analyze the human eQTL dataset, we identify 532 novel associations between 49 CNVs and the expression of 190 genes in a *subnetwork-to-subnetwork* fashion. Fig. 4 demonstrates three examples of the subnetwork-to-subnetwork associations in our results. In comparison to the models using only label (MTLasso LG) or feature graph (MTLasso FG) on the same dataset, our two-graph model MTLasso 2G (Fig. 5) selects those associations with support from the other two models and thus might remove some false positives.

Many of the genes whose expression profiles are affected by CNVs in our results are disease associated and/or have important biological functions. For example, the first example of *subnetwork-to-subnetwork* associations in Fig. 4 shows a scenario that two CNVs namely “chr20:1508967-1534025“ and “chr20:1509062-1542435“, are jointly associated with the expression of two genes in a co-expression subnetwork, *RBL1* and *VAPB*. Both genes have been identified as disease-associated through genome-wide association studies. Particularly, *RBL1* has been extensively studied as a cancer gene which is correlated to lung cancers [13]. The mutation of *VAPB* has been shown to be associated with myotrophic lateral sclerosis [15], breast cancer [10], and many other common diseases like hypertension, coronary artery disease and diabetes [9]. The observation, that the two CNVs are jointly associated with the expressions of these two genes simultaneously, might provide biological insights into the mechanism of disease manifestation of these genes and genetic variants.

7 Conclusions

In this paper, we propose a novel two-graph guided multi-task lasso model that takes advantage of the prior structures of features and labels for *subnetwork-subnetwork* associtions in eQTL mapping. This new model is a generalization form of previously proposed lasso models and thus subsumes those models as special cases. Additionally, the model is flexible with different types of features and labels and is applicable (but not limited) to eQTL mapping. For instance, our model can be applied to identify a full panel of genetic variants (e.g. SNPs, small insertions and deletions, and CNVs) that affect diverse traits such as gene expression and epigenetic profiles. Simulation study shows the nice performance of our model and real data analysis provides an example of its applications in eQTL mapping.

We remark that the feature and label graphs imposed on our model are flexible as well. We show that both graphs are constructed from examining the correlations among genetic variants and among co-expressed genes respectively. Nonetheless, other biological networks can be overlaid on the features or the labels. For example, we can use a protein-protein interaction network on the genetic variants as a feature graph to capture the interactions among variants, and utilize a regulatory network on the genes as a label graph. As long as these graphs provide reasonable structures underlying the data, our model can leverage structural priors to identify novel *subnetwork-to-subnetwork* associations.

Acknowledgements

This work was supported by a T32 training grant from Brigham and Women’s Hospital and Harvard Medical School (T32GM007748) to XS, NIH grants HG005209, HG004221, and CA111560 to CL, NSF DBI-0960390 and NIH R01GM089753 grants to JX.

References

[1] BE Stranger, MS Forrest, M Dunning, CE Ingle, C Beazley, N Thorne, R Redon, CP Bird, A de Grassi, C Lee, C Tyler-Smith, N Carter, SW Scherer, S Tavar, P Deloukas, ME Hurles, and ET Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–53, 2007.

[2] SB Montgomery, M Sammeth, M Gutierrez-Arcelus, RP Lach, C Ingle, J Nisbett, R Guigo, and ET Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Science*, 464(7289):773–7, 2010.

[3] S Kim and EP Xing. Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network. *PLoS Genetics*, 5(8), 2009.

[4] S Kim and EP Xing. Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity. *The 27th International Conference on Machine Learning (ICML)*, 2010.

[5] S Lee, J Zhu, and EP Xing. Adaptive Multi-Task Lasso: with Application to eQTL Detection. *The 24th Annual Conference on Neural Information Processing Systems (NIPS)*, 2010.

[6] X Chen, S Kim, Q Lin, JG Carbonell, and EP Xing. Graph-Structured Multi-task Regression and an Efficient Optimization Method for General Fused Lasso. *Preprints, arXiv*, 2010.

[7] X Chen, ZJ Wang, and MJ McKeown. Asymptotic Analysis of Robust LASSOs in the Presence of Noise with Large Variance. *IEEE Transactions on Information Theory*, 56(10):5131–5149, 2010.

[8] The 1000 Genomes Project Consortium. A Map of Human Genome Variation from Population Scale Sequencing. *Nature*, 467(7319):1061–1073, 2010.

[9] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 2007.

[10] DJ Hunter, P Kraft, KB Jacobs, DG Cox, M Yeager, SE Hankinson, S Wacholder, Z Wang, R Welch, A Hutchinson, J Wang, K Yu, N Chatterjee, N Orr, WC Willett, GA Colditz, RG Ziegler, CD Berg, SS Buys, CA McCarty, HS Feigelson, EE Calle, MJ Thun, RB Hayes, M Tucker, DS Gerhard, JF Jr Fraumeni, RN Hoover, G Thomas, and SJ Chanock. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39(7):870–4, 2007.

[11] K Lange, DR Hunter, and I Yang. Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

[12] R Mills, K Walter, C Stewart, R Handsaker, K Chen, C Alkan, A Abyzov, S Yoon, K Ye, R Cheetham, A Chinwalla, D Conrad, Y Fu, F Grubert, I Hajirasouliha, F Hormozdiari, L Iakoucheva, Z Iqbal, S Kang, J Kidd, M Konkel, J Korn, E Khurana, D Kural, H Lam, J Leng, R Li, Y Li, C Lin, R Luo, X Mu, J Nemesh, H Peckham, T Rausch, A Scally, X Shi, M Stromberg, A Stutz, A Urban, J Walker, J Wu, Y Zhang, Z Zhang, M Batzer, L Ding, G Marth, G McVean, J Sebat, M Snyder, J Wang, K Ye, E Eicher, M Gerstein, M Hurles, C Lee, S McCarroll, J Korbel, and the 1000 Genomes Project. Mapping Copy Number Variation by Population-Scale Genome Sequencing. *Nature*, 470(7332):59–65, 2011.

[13] S Modi, A Kubo, H Oie, AB Coxon, A Rehmatulla, and FJ Kaye. Protein expression of the RB-related gene family and SV40 large T antigen in mesothelioma and lung cancer. *Oncogene*, 19(40):4632–9, 2000.

[14] RR Nayak, M Kearns, RS Spielman, VG Cheung. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Research*, 19(11): 1953–1962, 2009.

[15] AL Nishimura, M Mitne-Neto, HC Silva, A Richieri-Costa, S Middleton, D Cascio, F Kok, JR Oliveira, T Gillingwater, J Webb, P Skehel, and M Zatz. A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. *Nature*, 447(7145):661–78, 2007.

[16] JK Pickrell, JC Marioni, AA Pai, JF Degner, BE Engelhardt, E Nkadori, JB Veyrieras, M Stephens, Y Gilad, and JK Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Science*, 464(7289):768–72, 2010.

[17] R Tibshirani, M Saunders, S Rosset, J Zhu, and K Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society*, 91–108, 2005.

Appendix: proofs

Proof of Theorem 4.1

Proof. Let $\hat{B}_n$ minimize $f(B)$ where f is objective function defined in (4); then it follows that $\sqrt{n}(\hat{B}_n - B)$ minimizes V_n, where

$$\begin{aligned} V_n(U) = & \sum_{k=1}^{K}\sum_{i=1}^{n}\left[\left(e_{ik} - \frac{\mathbf{u}^{k^T}\mathbf{x}_i}{\sqrt{n}}\right)^2 - e_{ik}^2\right] \\ & + \lambda \sum_{j=1}^{J}\sum_{k=1}^{K}\left[\left|b_{jk} + \frac{u_{jk}}{\sqrt{n}}\right| - |b_{jk}|\right] \\ & + \gamma_1 \sum_{e_{m,l}\in E_1} w(e_{m,l}) \sum_{j=1}^{J}\Big[|b_{jm} - \text{sign}(r_{m,l})b_{jl} \\ & + \frac{u_{jm} - \text{sign}(r_{m,l})u_{jl}}{\sqrt{n}}| - |b_{jm} - \text{sign}(r_{m,l})b_{jl}|\Big] \\ & + \gamma_2 \sum_{e_{f,g}\in E_2} w(e_{f,g}) \sum_{k=1}^{K}\Big[|b_{fk} - \text{sign}(r'_{f,g})b_{gk} \\ & + \frac{u_{fk} - \text{sign}(r'_{f,g})u_{gk}}{\sqrt{n}}| - |b_{fk} - \text{sign}(r'_{f,g})b_{gk}|\Big] \end{aligned}$$

with $U = (\mathbf{u}^1, \cdots, \mathbf{u}^K)$. We now consider each terms in V_n. Since K is fixed and $n \to \infty$, by the central limit theorem (CLT) and the assumptions that $\{e_{ik}\}$ are independent and $n^{-1}X^TX \to C$, we observe that

$$\begin{aligned} & \sum_{k=1}^{K}\sum_{i=1}^{n}\left[\left(e_{ik} - \frac{\mathbf{u}^{k^T}\mathbf{x}_i}{\sqrt{n}}\right)^2 - e_{ik}^2\right] \\ = & \sum_{k=1}^{K}\left[-2\mathbf{u}^{k^T} n^{-1/2}\sum_{i=1}^{n} e_{ik}\mathbf{x}_i + \mathbf{u}^{k^T} n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\mathbf{u}^k\right] \\ \xrightarrow{d} & -2\sum_{k=1}^{K}\mathbf{u}^{k^T}\mathbf{w}_k + \sum_{k=1}^{K}\mathbf{u}^{k^T}C\mathbf{u}^k, \end{aligned}$$

where $\mathbf{w}_k \overset{\text{iid}}{\sim} N(\mathbf{0}, \sigma^2 C)$. For the second term, it is obvious that

$$\begin{aligned} & \lambda \sum_{j=1}^{J}\sum_{k=1}^{K}\left[\left|b_{jk} + \frac{u_{jk}}{\sqrt{n}}\right| - |b_{jk}|\right] \\ \to & \lambda_0 \sum_{j=1}^{J}\sum_{k=1}^{K}\left[u_{jk}\text{sign}(b_{jk})\mathbb{I}(b_{jk} \neq 0) + |u_{jk}|\mathbb{I}(b_{jk} = 0)\right]. \end{aligned}$$

For the third term, we similarly obtain that

$$\begin{aligned} & \gamma_1 \sum_{e_{m,l}\in E_1} w(e_{m,l}) \sum_{j=1}^{J}\Big[|b_{jm} - \text{sign}(r_{m,l})b_{jl} \\ & + \frac{u_{jm} - \text{sign}(r_{m,l})u_{jl}}{\sqrt{n}}| - |b_{jm} - \text{sign}(r_{m,l})b_{jl}|\Big] \\ \xrightarrow{P} & \gamma_{01} \sum_{e_{m,l}\in E_1} w(e_{m,l}) \sum_{j=1}^{J}\Big[(u_{jm} - \text{sign}(c_{m,l})u_{jl}) \\ & \times \text{sign}(b_{jm} - \text{sign}(c_{m,l})b_{jl})\mathbb{I}(b_{jm} \neq \text{sign}(c_{m,l})b_{jl})\Big] \\ & + \gamma_{01} \sum_{e_{m,l}\in E_1} w(e_{m,l}) \sum_{j=1}^{J}\Big[|u_{jm} - \text{sign}(c_{m,l})u_{jl}| \\ & \times \mathbb{I}(b_{jm} = \text{sign}(c_{m,l})b_{jl})\Big]. \end{aligned}$$

The last term has a similar limit as in the third term. Combining all terms together and applying Slutsky's lemma, we therefore deduce that $V_n(U) \xrightarrow{d} V(U)$, where $V(U)$ is defined in (13). Since V_n is convex and V has a unique minimum, the theorem follows immediately from the finite-dimensional epi-convergence result as in [7]; that is, $\arg\min(V_n) = \sqrt{n}(\hat{B}_n - B) \xrightarrow{d} \arg\min(V)$. □

Proof of Corollary 4.2

Proof. Condition $\max\{\lambda, \gamma_1, \gamma_2\} = o(\sqrt{n})$ implies that $\lambda_0 = \gamma_{01} = \gamma_{02} = 0$ and therefore

$$V(U) = -2\sum_{k=1}^{K}\mathbf{u}^{k^T}\mathbf{w}_k + \sum_{k=1}^{K}\mathbf{u}^{k^T}C\mathbf{u}^k.$$

The last expression is separable in $\{\mathbf{u}^k\}$ and hence we can consider the minimizer for each column of U. Then it is obvious that $C^{-1}\mathbf{w}_k$ minimizes

$$-2\mathbf{u}^{k^T}\mathbf{w}_k + \mathbf{u}^{k^T}C\mathbf{u}^k.$$

The proof is complete since $C^{-1}\mathbf{w}_k \overset{\text{iid}}{\sim} N(\mathbf{0}, \sigma^2 C^{-1})$. □