10-708 PGM Project Final Report
Learning SNP-Gene Network Using Mixed Graphical Model

Hyun Ah Song HYUNAHS@ANDREW.CMU.EDU
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213

Ji Oh Yoo JIOHY@ANDREW.CMU.EDU
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

Learning expression quantitative trait locus (eQTL) mapping between single-nucleotide polymorphisms (SNPs) and expression rates of related proteins is important in discovering key factors of the diseases. Most of the previous approaches require prior knowledge in gene expression rates or enforce unnatural assumptions in the representations of SNPs. In this project, we propose a conditional mixed graphical model for SNPs-gene network learning that resolves these limitations. Our approach models SNPs as discrete values to relax the unnatural assumption and does not require prior knowledge of the output structure for learning SNPs-gene network. Experimental results on synthetic data show that our model outperforms previously proposed methods in prediction error and edge recovery rate. Experimental results on Human Liver Cohort dataset show that our model does not demonstrate improved performance compared to the previous methods due to the relatively high sample complexity. More in-depth analysis on the behavior of our model is left as further work.

1. Introduction

Recent progress on large scale genome-wide studies show that many human common diseases, including diabetes, asthma, and cancer, consist of complex reactions of proteins, which are controlled by regulatory networks and are expressed from highly correlated genetic variations (Basso et al., 2005; Chen et al., 2008). Thus, rather than simple correlation studies between a few clinical phenotypes and genetic variations of interest, discovering expression quantitative trait locus (eQTL) mapping between the multiple genetic variations and multiple disease-related proteins' expression rates is necessary to analyze the occurrence mechanism of the diseases and the functional roles of those proteins in the mechanism. This requires the joint learning of the genetic variations and the expression of the related proteins rather than combining the results from the independent study on each genome or phenome.

Multiple regression studies have been conducted to identify these mappings between single-nucleotide polymorphisms (SNPs) and related gene expression rates, and to develop more efficient and accurate methods for discovery of these mappings (Kim & Xing, 2010; Sohn & Kim, 2012). However, all of the previous methods enforce unnatural assumption on the representation of SNPs. Although SNPs information is represented as discrete states of 0, 1, or 2, based on the degree of deviation from major alleles, previous methods convert these discrete values into continuous values, and treat them as if relationship between the deviation and the output is linear.

In this project, we adopt conditional undirected mixed graphical model (Lee & Hastie, 2013) for SNP-gene network learning to relax the unnatural assumptions on the representation of SNPs and lessen requirements for learning in the previous methods. We adopt discrete representations of SNPs as it is without making assumptions on the linear relationship between the genetic deviation and gene expression rates. We generate synthetic dataset that reflects our hypothesis on the nature of the SNPs-gene expression data, and analyze the performance of our proposed model in prediction error and edge recovery rate. We conduct experiment on real-world dataset to convince whether our hypothesis is correct and investigate the behavior of our proposed model.

The organization of the report is as follows. In section 2, we discuss previous works in more detail. We briefly in-

Proceedings of the 31st International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

troduce previous methods and address limitations of them. In section 3, we introduce our proposed model. We explain how we reformulate the existing model to fit our problem, and the optimization method we used. In section 4, we discuss our experimental results on synthetic dataset and a real-world dataset. Finally, we close our report in section 5.

2. Background & Related Work

One of the most commonly used methods in estimation of SNPs-gene network is the multiple-output extension of single-output regression method. Given SNPs information as inputs, multiple-output regression tries to solve the regression problem for each output simultaneously while sharing the coefficient matrix for all the tasks. It can incorporate L-1 penalty (lasso) (Tibshirani, 1996) for sparse model. The regression problem can be formulated as in Eq. 1

$$\arg\min_{\mathbf{B}} \mathrm{tr}((\mathbf{Y}-\mathbf{XB})(\mathbf{Y}-\mathbf{XB})^T) + \lambda\|\mathbf{B}\|_1, \quad (1)$$

where $\mathbf{X} = (\mathbf{x_1}, ..., \mathbf{x_N})^T$, $\mathbf{Y} = (\mathbf{y_1}, ..., \mathbf{y_N})^T$ are input and output data, and $\mathbf{x_i} = (x_{i1}, ..., x_{iJ})^T \in \mathbb{R}^J$, and $\mathbf{y_i} = (y_{i1}, ..., y_{iK})^T \in \mathbb{R}^K$. The output is estimated by regression: $\mathbf{y_i} = \mathbf{B}\mathbf{x_i} + \varepsilon_\mathbf{i}$, where $\varepsilon \sim \mathcal{N}(0, \mathbf{\Psi})$.

By enforcing sparsity in $\mathbf{B}$, it is able to take into account of the problem setting that we have $J >> N$, where the number of SNPs is larger than the number of samples. Also, using L-1/L-2 mixed norm penalty can also be used for joint feature selection (Obozinski et al., 2008).

Despite the simple formulation and optimization method for multiple-output regression, it does not capture the relatedness among the output variables in the data. To address this problem of learning structured sparsity and structure in output variables, Graph-guided Fused Lasso has been developed with fusion penalty according to pre-computed degree of relatedness among the output (Kim & Xing, 2009). However, it requires prior knowledge of the data, which can differ by each problem setting and its problem domain.

Rothman and coworkers (Rothman et al., 2010) introduced Multivariate Regression with Covariance Estimation (MRCE) that aims to learn both multivariate regression parameters and the correlation of outputs. By regularizing regression parameter and correlation matrix separately, MRCE solves bi-convex problem, which may not lead to the global optimum. Although MRCE is favorable in a way that it learns the output structure, MRCE does not force structured sparsity when learning regression coefficients, which is an important property when dealing with SNPs-gene data.

Sohn and coworkers (Sohn & Kim, 2012) introduced Sparse Gaussian Graphical Model that brings together the advantages of the two previous works by extending the problem of learning standard regression model into learning a sparse conditional Gaussian graphical model. It jointly models the input and output as a multivariate Gaussian distribution and learns the conditional model of output given input as in Eq. 2.

$$\begin{bmatrix}\mathbf{x_i}\\ \mathbf{y_i}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}_J\\ \mathbf{0}_K\end{bmatrix}, \Sigma\right)$$
$$\mathbf{\Sigma}^{-1} = \begin{pmatrix}\mathbf{\Theta}_{xx} & \mathbf{\Theta}_{xy}\\ \mathbf{\Theta}_{xy}^T & \mathbf{\Theta}_{yy}\end{pmatrix} \quad (2)$$

Then, the parameters $\mathbf{\Theta}$ and $\mathbf{\Phi}$ in Eq. 1 can be reconstructed by $\mathbf{B} = \mathbf{\Sigma}_{xy}^T\mathbf{\Sigma}_{xx}^{-1} = -\mathbf{\Theta}_{yy}^{-1}\mathbf{\Theta}_{xy}^T$, and $\mathbf{\Psi} = \mathbf{\Theta}_{yy}^{-1}$.

Although this method by Sohn suits the needs for the SNPs-gene network estimation problem as it learns both structured sparsity in $\mathbf{\Theta}$ and output structure in $\mathbf{\Phi}$, and formulates the problem using the standard regression problem, the underlying assumption is that both input and output data constitute of continuous variables. This assumption is from that the degree of genetic variation, the deviation from major alleles, can be measured by the number of minor alleles in a person's genotype, and the relationship between the deviation and the output is linear. However, the phenotypes from genes with 0 or 1 minor alleles can be drastically different from those of 2 minor alleles in many cases including examples from Mendel's Law. Thus, it is reasonable approach to model SNPs as categorical, discrete values to relax this assumption in the previously developed regression models.

Mixed graphical models refer to graphical models that contains both discrete and continuous variables. Lauritzen and coworkers (Lauritzen & Wermuth, 1989) combined two commonly used graphical models so that it can handle data consisting of both continuous and discrete variables as: $p(y|x) = \mathcal{N}(\mu(x), \Sigma(x))$. Despite of the novelty and specific applications of the mixed graphical model, it has not been used as widely due to its complexity in learning the parameters, and lack of intuition when reformulated to adopt various cases of data composition.

Lee and coworkers (Lee & Hastie, 2013) have proposed a special case of the mixed graphical model that resolves the problem of the complexity in parameter learning and non-intuitive reformulation of the model by making several assumptions. The conditional distribution of a discrete variable reduces to multi-class logistic regression, and that of a continuous variable to Gaussian linear regression. By sharing common covariance matrix and additive mean property, the newly proposed algorithm scales up more efficiently.

To our knowledge, there has not been any study that learns the SNPs-gene network using mixed graphical models. In this project, we reformulate mixed graphical model to conditional form and apply it to the domain of SNPs-gene network learning.

3. Methods

Our goal is to improve the method of learning SNPs-gene network by adopting the conditional undirected mixed graphical model in which we can embed both discrete variables for SNPs and continuous variables for gene expression rates.

3.1. Mixed Graphical Model

Lee and coworkers (Lee & Hastie, 2013) proposed a mixed graphical model of discrete and continuous variables, where the joint distribution is expressed as:

$$p(x,y;\Theta) \propto \exp\Big(\sum_{k=1}^{q}\sum_{l=1}^{q}\phi_{kl}(x_k,x_l) + \sum_{k=1}^{p}\sum_{s=1}^{q}\rho_{ks}(x_k)y_s + \sum_{s=1}^{p}\alpha_s y_s + \sum_{s=1}^{p}\sum_{t=1}^{q} -\frac{1}{2}\beta_{st}y_s y_t\Big) \tag{3}$$

where $x_1, ...x_p$ are discrete variables representing the occurrences for each SNP and $y_1, ..., y_q$ are continuous variables representing the gene expression rate of each gene. The parameters are $\Theta = [\{\phi_{kl}\}, \{\rho_k\}, \{\alpha_s\}, \{\beta_{st}\}]$, where ϕ_{kl}, ρ_k, α_s, β_{st} are the discrete-discrete edge potential, continuous-discrete edge potential, continuous node potential, and continuous-continuous edge potential, respectively.

By making further assumptions on mixed graphical models originally introduced in (Lauritzen & Wermuth, 1989), the mixed graphical model 3 proposed by Lee (Lee & Hastie, 2013) yields intuitive results when reformulated into various forms of distributions. The conditional distribution of continuous variables and discrete variables given the rest reduces to Gaussian linear regression, and multi-class logistic regression, respectively. Furthermore, when the data consists of continuous and discrete variables only, the model reduces to multivariate Gaussian model, and pairwise discrete Markov random field, respectively.

3.2. Conditional Mixed Graphical Model

We are interested in learning the SNPs-gene network structure, and the conditional distribution of the gene expression, given observed SNPs. Therefore, we only need to focus on learning the conditional distribution of the network rather than learning the full joint distribution.

The conditional distribution of continuous variables given discrete variables of the mixed graphical model in (Lee & Hastie, 2013) reduces to the form of multivariate Gaussian distribution.

$$p(y|x) = \mathcal{N}(B^{-1}\gamma(x), B^{-1}) \tag{4}$$

$$\{\gamma(x)\}_s = \alpha_s + \sum_k \rho_{ks}(x_k) \tag{5}$$

$$p(x) \propto \exp\Big(\sum_k\sum_l \phi_{kl}(x_k,x_l) + \frac{1}{2}\gamma(x))^\intercal B^{-1}\gamma(x)\Big) \tag{6}$$

where B is a symmetric, positive definite inverse covariance matrix $B = \{\beta_{st}\}$ that is shared across the Gaussian distributions.

From the property of multivariate Gaussian distribution, the log likelihood of the parameters Θ given discrete variables y can be expressed as:

$$\log p(y|x;\Theta) = -\frac{1}{2}\log|B^{-1}| - \frac{k}{2}(2\pi) - \frac{1}{2}(y - B^{-1}\gamma(x)^\intercal B(y - B^{-1}\gamma(x)) \tag{7}$$

$$l_p(\Theta) = -\frac{1}{2}tr\Big((Y - B^{-1}\Gamma(X))^\intercal B(Y - B^{-1}\Gamma(X))\Big) - \frac{N}{2}\log|B^{-1}| + \lambda_1\|\{\beta_{st}\}\|_1 + \lambda_2\|\{\rho_{ks}\}\|_2 \tag{8}$$

The parameter Θ^* that minimizes this L-1 penalized log-likelihood will give us a sparse and consistent estimator, and by comparing the parameters with results in multi-task regression settings, we can analyze the existence of direct and indirect relationship between SNPs and gene expression rates.

For the parameter estimation, we minimize the negative log pseudo-likelihood expressed in product of each variable given the rest as follows:

$$\tilde{l}(\Theta|y) = -\sum_{s=1}^{p}\log p(y_s|y_{\setminus s}, x;\Theta), \tag{9}$$

where from 3, $p(y_s|y_{\setminus s}, x;\Theta)$ can be reformulated as follows:

$$p(y_s|y_{\setminus s}, x;\Theta) = \frac{\sqrt{\beta_{ss}}}{\sqrt{2\pi}}\exp\left(\frac{-\beta_{ss}}{2}\left(y_s - \frac{\alpha_s + \sum_j \rho_{sj}(y_j) - \sum_{t=s}\beta_{st}y_t}{\beta_{ss}}\right)\right)^2 \tag{10}$$

3.3. Optimization

For the optimization, proximal Newton method (Schmidt, 2010; Schmidt et al., 2011) is used. Proximal Netwon optimization method breaks down the optimization problem into $f(x) + g(x)$, where $f(x)$ is smooth and convex function, and $g(x)$ is a convex function, but not necessarily smooth. In our problem, functions $f(x)$ and $g(x)$ correspond to the negative log-likelihood and penalty terms, respectively.

Proximal Newton method is an extension of proximal gradient method that considers the second order information of $f(x)$ in addition to the first order information. The proximal gradient seeks next point by computing $x_{k+1} = prox_t(x_k - t \bigtriangledown (x_k))$, where $prox_t(x) = \text{argmin}_u(\frac{1}{2t}\|x-u\|^2 + g(u))$, and t is chosen by Armijo line search algorithm. In proximal Newton method, $Hprox$ is used, where the $prox$ operator is extended to $\|\cdot\|_H$, where $H = \bigtriangledown^2 f(x_k)$. For each iteration, proximal Newton method finds the next point that minimizes below function.

$$\bigtriangledown f(x_k)^T(u - x_k) + \frac{1}{2t}(u + x_k)^T H(u - w_k) + g(u)$$
$$= Hprox_t(x_k - tH^{-1} \bigtriangledown f(x_k)). \quad (11)$$

Previous studies on proximal Newton method (Lee et al., 2012) shows that it converges faster especially when n is large, with better accuracy compared to the proximal gradient method.

For the implementation, H is approximated using BFGS. We used PNOPT package (Lee et al., 2012) for implementation of proximal Newton optimization algorithm.

4. Experiments

4.1. Experiments on Synthetic Data

Before delving into real-world dataset, we run experiments on synthetic data to verify our assumptions on the behavior of our proposed model.

4.1.1. Synthetic Data Generation

We create a simple ground truth model with 7 discrete variables (input) and 5 continuous variables (output). We design the discrete variables to have three possible states to reflect the nature of SNPs. We construct a chain structure for the discrete variables, where each node is linked to the following node. For two consecutive nodes that are connected by an edge, with probability of 0.5, we create a state-to-state interactions as following with

$$0.5 * \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

The continuous variables are designed to have symmetric interactions between the nodes, where we generate random values for node-to-node interaction in range of [-1, 1], and remove the edge with probability of 0.7 maintaining the positive semi-definiteness of the matrix. For the input-output interaction, we randomly create edges with probability of 0.2. In order to reflect our assumption on the nature of SNPs-gene data, and therefore verify that our proposed model excels in performance under this assumption, we design the interactions between discrete and continuous variables that are linked via an edge to be increasing in non-linear manner with respect to the index of the states; we define the interactions between three states of the discrete variable and the continuous variable to be in non-linear relationship as follows: 1 to 1.2 to 6.

Total 2000 samples are sampled from the true model and used for estimating the edges between the discrete variables and the continuous variables. We analyze the performance based on the two measures: average prediction error (average mean squared error of predicted output) and accuracy of edge recovery (true positive and true negative out of total number of edges). We repeat the experiment for 10 times by randomly splitting 2000 samples into training set (80%) and test set (20%), and average the prediction error. To find the optimal hyper-parameters, we use 3-fold cross-validation based on the prediction errors.

4.1.2. Prediction Error

The prediction errors of proposed method, sparse CGGM, lasso, and multi-task lasso are shown in Table 1. Our proposed method achieves the lowest prediction error, followed by multi-task Lasso, Lasso, and SCGGM, where SCGGM achieves the worst prediction with the largest variation. We can clearly see that multi-task lasso and lasso perform better prediction than SCGGM. The lasso family is expected to learn non-linear input-output relationships by making approximations to linear weights across the states via interpolations. Compared to lasso family, learning of SCGGM model takes more intricate procedures, which may have resulted in more confusion in incorporating non-linearity into the parameter learning. From this experimental result, we can conclude that under given assumptions on the data, where non-linear relationships between the states of input and output variables are expected, our proposed model takes advantage of the model structure, and conducts parameter learning in more desired manner.

Method	Prediction Error
MG	$\mathbf{1.4798 \pm 0.1811}$
SCGGM	2.1308 ± 0.5252
Lasso	1.8394 ± 0.2107
MTLasso	1.7950 ± 0.1625

Table 1. Prediction error on synthetic dataset with methods MG (mixed graphical model), SCGGM (sparse CGGM), Lasso, MT-Lasso (multi-task Lasso).

4.1.3. Edge Recovery

Learning the underlying structure of the dataset that can reveal SNPs-gene expression rates or gene expression rates themselves is of great interest as achieving better prediction errors. Graph-base models, sparse CGGM and mixed graphical model, are considered more powerful and meaningful as they can capture the conditional independence relationships among the variables, compared to multiple-output regression models.

Using the same settings for the data generation, we assess how well our model and SCGGM learn the true data structure. We analyze the dependency of the structure learning on the sample size, by computing the accuracy of the edge recovery rate of the input-output network with varying number of samples used for model learning. We increase the number of sample size by 250 interval until 2000. The experimental results are averaged over five repeated trials. (Both our proposed method and SCGGM do not learn sparse-enough parameters for input-output network structure, so we threshold the recovered parameters over certain values (1 for proposed method, and 0.001 for SCGGM) to finally attain meaningful network structures.)

Figure 1 shows the edge recovery rate of our proposed method (Mixed Graphical model - MG), and SCGGM. We see that for the sample size of 250, SCGGM achieves slightly better edge recovery. As we increase the number of sample size, our proposed method shows improvement in edge recovery rate, and converges to 1 with 750 samples. On the other hand, SCGGM does not show any improvement over incremental sample size. The interpretation of the result is straightforward. Since our proposed method is capable of managing the non-linearity in the data structure, it shows expected learning curve based on the sample size. For SCGGM, since the structure learning is not being done as desired, increasing the sample size does not resolve the intrinsic problem of ill-learning of the model structure.

In Figure 2, images of input-output, and output-output structures learned by proposed method and SCGGM are displayed with ground-truth. Each row shows the images of ground-truth, outcomes of proposed method, and SCGGM when the sample size is 2000, respectively. First

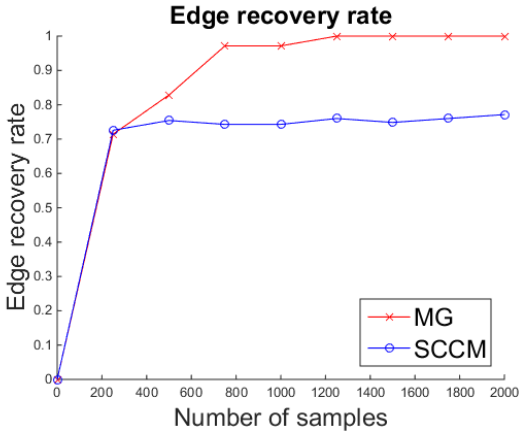

Figure 1. Edge recovery rate of MG and SCGGM on the synthetic dataset.

two columns are outcomes of input-output network structures, and the third column is for output-output network structure. Each of the first two columns shows raw parameters, and binarized parameters over threshold.

We see that even when sufficient number of samples are provided for model learning, while proposed method learns and recovers perfect ground-truth input-output network structure, SCGGM fails to recover the true network structure. For the network structure of output-output relationships, both proposed method and SCGGM seem to recover almost perfect structure. This result is understandable; while it is relatively easier for SCGGM to learn output-output network structure, it is harder for it to learn input-output network structure due to improper underlying assumptions on the data structure.

From this experimental result, we can conclude that our proposed model excels in structure recovery when the assumptions on the data are met.

4.2. Experiments on Human Liver Cohort Data

For learning the eQTL mapping between SNPs and the expression rates of the related genes, we use the data from Human Liver Cohort(HLC) study (Schadt et al., 2008) [1]. Out of 427 patients with more than 40,000 expression rates and more than 700,000 SNPs in the given data, we first extract 178 samples that contain both information on genotype and expression rates. To focus on learning of the structure of the model rather than handling the huge dimension of this dataset, we focus on smaller subset of genes and expression rates as in the work of Sohn and coworker (Sohn & Kim, 2012). We consider only expression rates whose variance is greater than 0.05 and SNPs on chromosome 6 with

[1]The dataset is available at https://www.synapse.org/#!Synapse:syn4499

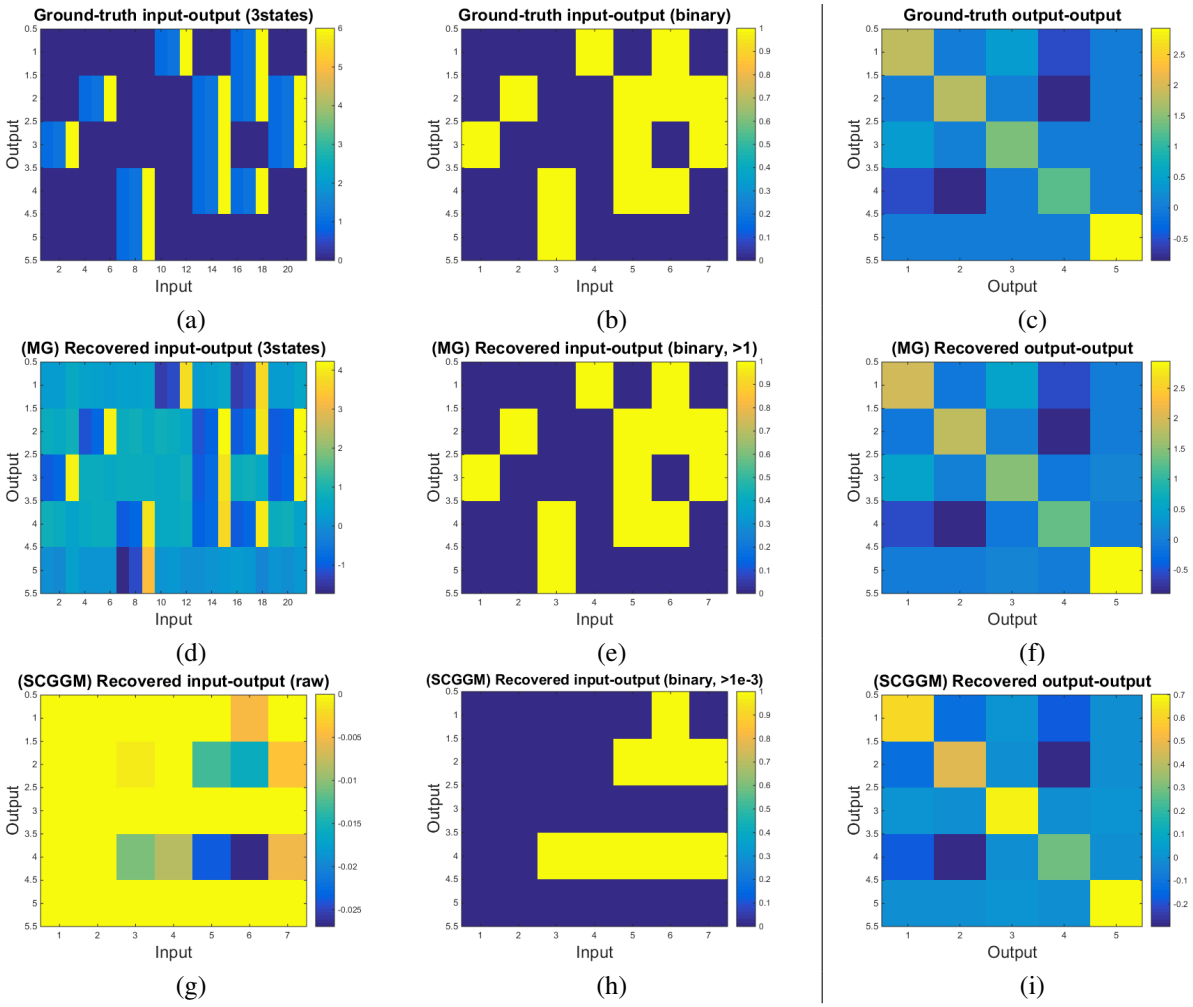

Figure 2. Ground-Truth (GT) and recovered 'input-output' and 'output-output' interactions of Mixed Graphical model (MG), and sparse CGGM (SCGGM). (a) GT - input-output (3 states), (b) GT - input-output (binary), (c) GT - output-output, (d) MG - input-output (3 states), (e) MG - input-output (binary), (f) MG - output-output, (g) SCGGM - input-output (3 states), (h) SCGGM - input-output (binary), (i) SCGGM - output-output

Method	Prediction Error
MG	0.0239
SCGGM	0.0246
Lasso	1.6323×10^{-8}
MTLasso	3.1753×10^{-7}

Table 2. Average prediction error on Human Liver Cohort data testset from Conditional Mixed Graphical Model (MG), Sparse Conditional Gaussian Graphical Model (SCGGM), Lasso, Multi-task Lasso with joint feature selection (MTLasso). The number of SNPs selected is 937 and the number of proteins for expression rates is 100.

Method	Prediction Error
MG	0.0938
SCGGM	0.0471
Lasso	0.0255
MTLasso	0.0217

Table 3. Average prediction error on selected SNPs and genes from Human Liver Cohort data from Conditional Mixed Graphical Model (MG), Sparse Conditional Gaussian Graphical Model (SCGGM), Lasso, Multi-task Lasso with joint feature selection (MTLasso). The number of SNPs selected is 36 and the number of proteins for expression rates is 9.

minor allele frequency greater than 0.01 and pair-wise correlation less than 0.1 to select genes that are not too biased to have major allele and not too related each other according to dbSNP[2]. For a single nucleotide, where major allele is X and minor is Y, SNPs are encoded to categorical variables as XX = 0, XY = 1, YY = 2. SNPs are considered to be discrete variables with three states $\{0, 1, 2\}$ and expression rates are to be continuous variables in our model. We use 143 samples for training and the remaining 35 samples for testing. For hyperparameter selection, we use 3-fold cross-validation on mean squared error.

The prediction error on our model and other baseline models are shown in Table 2. Our method shows lower average prediction error than SCGGM, and from this, our assumption on genetic variation can be confirmed: the phenotypes of 0 or 1 minor alleles can be drastically different from those with 2 minor alleles, and we need discrete models for SNPs. But the traditional multi-task regression methods (Lasso, MTLasso) shows errors much closer to zero. We find the explanation of this discrepancy in error from the sparsity in the parameters in each learned model. The learned coefficient matrix $\mathbf{B}$ of Lasso and MTLasso has all non-zero entries, but the input-output mapping learned from MG and SCGGM has very high sparsity. Thus, we see that the Lasso and MTLasso methods make a trade-off for better prediction errors with far less interpretability, and SCGGM and MG converge to a state with very sparse model, only using output-output relationships for better prediction error. We also conjecture that the relatively high sample complexity for our model might be an obstacle for better structure learning. We wanted to further analyze the performance of our method varying settings including hyperparameter candidates and metrics for cross-validation steps, but the learning of our model takes about a week and could not analyze other behaviors on this dataset.

To analyze the behaviors of our model further, we conduct the experiment on selected SNPs and gene expression rates for smaller dimensions. We strengthen the threshold for the selection of SNPs and selected fewer SNPs for our study. We select 36 SNPs with pair-wise correlation is less than 0.01 instead of 0.1. For gene expression rates, we focus on 9 HLA Class II genes following the findings from the previous study (Burton et al., 2007). The number of samples for training and test and the model selection method are the same as the previous experiment.

The prediction error on our model and other baseline models on smaller dataset are shown in Table 3. The prediction errors of SCGGM and our model are worse than that of Lasso, and MTLasso. The sparsity patterns in the coefficient matrix learned from Lasso and MTLasso, and input-output relationship learned from SCGGM and our model are the same as previous experiment.

5. Conclusion

In this project, we adopted conditional mixed graphical model for a better generative model in learning the eQTL mapping in SNPs-gene networks. Our assumption in the SNPs-gene network is that phenotypes of 0 or 1 minor alleles can be drastically different from those of 2 minor alleles. We reflected our assumption on the nature of the problem domain into our proposed model structure. Our proposed model has two advantages: 1) the relaxation of unnatural assumption in encoding of SNPs in sparse CGGM and 2) the unnecessity of prior knowledge among correlated gene expression rates in each problem settings. In synthetic experiments, we confirmed that our model outperforms both in prediction error and edge recovery rate when given the data following our assumption. In experiments on Human Liver Cohort dataset, it was hard to decide whether our model performs better than the other baseline models including sparse CGGM and multiple-output regression models, and we conjecture that this is due to relatively high sample complexity of our model. The future work includes developing a faster optimization method than PNOPT and incorporating more refined form of reg-

[2] http://www.ncbi.nlm.nih.gov/projects/SNP/

ularization for the discovery of other types of structured sparsity patterns in the network.

References

Basso, Katia, Margolin, Adam A, Stolovitzky, Gustavo, Klein, Ulf, Dalla-Favera, Riccardo, and Califano, Andrea. Reverse engineering of regulatory networks in human b cells. *Nature genetics*, 37(4):382–390, 2005.

Burton, Paul R, Clayton, David G, Cardon, Lon R, Craddock, Nick, Deloukas, Panos, Duncanson, Audrey, Kwiatkowski, Dominic P, McCarthy, Mark I, Ouwehand, Willem H, Samani, Nilesh J, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

Chen, Yanqing, Zhu, Jun, Lum, Pek Yee, Yang, Xia, Pinto, Shirly, MacNeil, Douglas J, Zhang, Chunsheng, Lamb, John, Edwards, Stephen, Sieberts, Solveig K, et al. Variations in dna elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435, 2008.

Kim, Seyoung and Xing, Eric P. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS genetics*, 5(8):e1000587, 2009.

Kim, Seyoung and Xing, Eric P. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 543–550, 2010.

Lauritzen, Steffen L and Wermuth, Nanny. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, pp. 31–57, 1989.

Lee, Jason and Hastie, Trevor. Structure learning of mixed graphical models. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pp. 388–396, 2013.

Lee, Jason D, Sun, Yuekai, and Saunders, Michael A. Proximal newton-type methods for minimizing convex objective functions in composite form. 2012.

Obozinski, Guillaume R, Wainwright, Martin J, and Jordan, Michael I. High-dimensional support union recovery in multivariate regression. In *Advances in Neural Information Processing Systems*, pp. 1217–1224, 2008.

Rothman, Adam J, Levina, Elizaveta, and Zhu, Ji. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4): 947–962, 2010.

Schadt, Eric E, Molony, Cliona, Chudin, Eugene, Hao, Ke, Yang, Xia, Lum, Pek Y, Kasarskis, Andrew, Zhang, Bin, Wang, Susanna, Suver, Christine, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS biology*, 6(5):e107, 2008.

Schmidt, Mark. *Graphical model structure learning with l1-regularization*. PhD thesis, UNIVERSITY OF BRITISH COLUMBIA (Vancouver, 2010.

Schmidt, Mark, Kim, Dongmin, and Sra, Suvrit. Projected newton-type methods in machine learning. 2011.

Sohn, Kyung-Ah and Kim, Seyoung. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1081–1089, 2012.

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