Gene-gene interaction analysis incorporating network information via a structured Bayesian approach

Xing Qin¹, Shuangge Ma², and Mengyun Wu¹

¹ School of Statistics and Management, Shanghai University of Finance and Economics
² Department of Biostatistics, Yale School of Public Health
email: wu.mengyun@mail.shufe.edu.cn

Abstract

Increasing evidence has shown that gene-gene interactions have important effects on biological processes of human diseases. Due to the high dimensionality of genetic measurements, existing interaction analysis methods usually suffer from a lack of sufficient information and are still unsatisfactory. Biological networks have been massively accumulated, allowing researchers to identify biomarkers from a system perspective by utilizing network selection (consisting of functionally related biomarkers) as well as network structures. In the main-effect analysis, network information has been widely incorporated, leading to biologically more meaningful and more accurate estimates. However, there is still a big gap in the context of interaction analysis. In this study, we develop a novel structured Bayesian interaction analysis approach, effectively incorporating the network information. This study is among the first to identify gene-gene interactions with the assistance of network selection for phenotype prediction, while simultaneously accommodating the underlying network structures. It innovatively respects the multiple hierarchies among main effects, interactions, and networks. Bayesian method is adopted, which has been shown to have multiple advantages over some other techniques. An efficient variational inference algorithm is developed to explore the posterior distribution. Extensive simulation studies demonstrate the practical superiority of the proposed approach. The analysis of TCGA data on melanoma and lung cancer leads to biologically sensible findings with satisfactory prediction accuracy and selection stability.

Keywords: Assistance of network selection; Gene-gene interaction; Network structured analysis.

1 Introduction

Gene-gene interactions have significant importance for the basis of human diseases beyond main genetic effects [Cordell 2009, Mackay 2014]. Due to the higher dimensionality, lower signal-to-noise ratio, and other reasons, there are more challenges in the analysis of interactions compared to main effects. We refer to Upton et al. (2016), Wu et al. (2019), and the references therein for
more discussions. In recent interaction analysis research, the “main effects-interactions” hierarchy is generally employed to improve both estimation and interpretation (Hao and Zhang, 2017). Specifically, an interaction can be identified only when one of its main effects (weak hierarchy) or both (strong hierarchy) are also identified. A number of statistical methods have been developed to identify important interactions and enforce this hierarchy. Among the available techniques, penalization has drawn much attention. Published works include the Lasso for hierarchical interaction (Bien et al., 2013), interaction learning via a hierarchical group-lasso regularization (Lim and Hastie, 2015), penalized tensor regression (Wu et al., 2018), and quadratic regression under the marginality principle (Hao, Feng).

Despite the vast literature on penalization and some other methods, there are very few Bayesian methods for hierarchical interaction analysis. Limited existing studies include Liu et al. (2015), which proposes a Bayesian hierarchical mixture model for interaction analysis and incorporates the natural hierarchical structure using the conditional prior probability technique. As another example, Kim et al. (2018) develops a Bayesian interaction analysis method with a hierarchical prior that fully considers the hierarchy constraint and controls the degree of sparsity simultaneously. There are also a few recent Bayesian methodological developments without enforcing hierarchy conditions, including Ren et al. (2020) and Ferrari and Dunson (2020).

With the high dimensions of genetic measurements but still limited sample sizes, the existing interaction analysis usually suffers from a lack of sufficient information and leads to unsatisfactory results. To bring in “additional information”, in main-effect analysis, a promising direction is to incorporate biological network information, which can be roughly classified into two strategies. The first strategy has been developed to take advantage of the assistance of network selection, where the “main effects-networks” hierarchy is usually enforced. That is, a main effect can be included in the model only when at least one of its involved networks is also included, and vice versa. Examples include Breheny (2015), which achieves a bi-level selection using the group exponential Lasso. In
addition, Xu and Ghosh (2015) proposes the Bayesian sparse group selection with spike and slab priors to select variables both at the network level and also within a network. Complementary to the first strategy, the second strategy has been developed to incorporate network structures. A representative technique is the network regularization based on the graph Laplacian matrix. Examples include penalization methods with the Laplacian-based penalty (Li and Li, 2008; Gao et al., 2019), and Bayesian methods with the Laplacian Gaussian prior (Cai et al., 2020). To take advantages of both strategies, multiple Bayesian methods have been developed to utilize network selection and also effectively account for underlying network structures (Zhe et al., 2013; Zhao, Kang, and Yu, 2014; Peterson, Stingo, and Vannucci, 2016). The aforementioned and other studies have shown that integrating network information can improve not only biological interpretability but also identification and predictive performance. However, most existing methods have been designed for main-effect analysis, and methodological developments in the context of interaction analysis are still very limited.

In this study, we propose a new structured Bayesian interaction analysis approach. This study is the first to conduct gene-gene interaction analysis with the assistance of network selection and simultaneously accommodate network structures. The most significant advancement is that both the “main effects-interactions” and “main effects/interactions-networks” hierarchy conditions are effectively respected, which is much more challenging than in the existing interaction analysis or network selection-assisted main effect analysis that enforce only one hierarchy. Furthermore, the underlying network structures are explored in the analysis of not only main effects but also interactions, making this study a big step forward from the existing main effect structured analysis. The proposed approach is based on the Bayesian method, which has been shown to have multiple advantages over some other techniques, such as penalization (Narisetty et al., 2019). Different from most published Bayesian interaction studies based on the Markov Chain Monte Carlo (MCMC) inference technique, we take advantage of the hybrid model integrating conditional and generative
components, and develop a more efficient variational inference algorithm. This is especially desirable with the extremely high dimensions in gene-gene interaction analysis. Overall, this study can provide a useful new venue for genetic interaction analysis.

2 Methods

Consider $K$ networks, which have been constructed using the existing biological network information. For the $k$th network, an adjacency matrix $E_k$ is used to describe the network structure among $p_k$ genetic factors (genes, SNPs, etc.), where $E_k(j,l) = 1$ if the $j$th and $l$th factors are connected in the network and $E_k(j,l) = 0$ otherwise. Suppose that we have $n$ i.i.d. subjects. For the $i$th subject, denote $y_i$ as the response, $x_i(1) = (x_{i1}^{(1)}, \cdots, x_{ip_k}^{(1)})$ as the $p$-dimensional vector of genetic measurements in the $k$th network with $\sum_{k=1}^{K} p_k = p$, and $x_i = (x_i^{(1)}, \cdots, x_i^{(K)})$. Note that if a genetic factor is involved in multiple networks, the corresponding measurement is duplicated in these networks.

2.1 Model

We consider the most popular continuous response, and the proposed approach can be extended to other responses. Specifically, consider the Gaussian distribution with the joint effects of all genetic factors and their interactions incorporating the network information:

$$
p(y_i | x_i, w, \tau) = \mathcal{N} \left( y_i \left| \sum_{k=1}^{K} \sum_{j=1}^{p_k} \left( x_{ij}^{(k)} w_{kj}^{(1)} + \sum_{k' = k}^{K} \sum_{k' \neq k, l=1}^{p_k'} x_{ij}^{(k)} x_{il}^{(k')} w_{kk',jl}^{(2)} \right) \right|, \tau^{-1} \right),
$$  

(1)

where $w_{kj}^{(1)}$’s and $w_{kk',jl}^{(2)}$’s are the regression coefficients for the main effects and interactions within the $k$th network ($k' = k$) or between the $k$th and $k'$th networks ($k' \neq k$), respectively, $w$ is the vector composed of all $w_{kj}^{(1)}$’s and $w_{kk',jl}^{(2)}$’s, and $\tau$ is the precision parameter. For $\tau$, a conjugate prior Gamma($\tau | g, h$) is assumed with $g$ and $h$ being the shape and scale parameters.
To identify important main effects and interactions, we consider the spike and slab priors:

\[
p ( \mathbf{w}_k^{(1)} | \beta_k^{(1)} ) = \prod_{j=1}^{p_k} \mathcal{N} \left( \mathbf{w}_{kj}^{(1)} | 0, s_1 \right)^{\beta_k^{(1)}} \mathcal{N} \left( \mathbf{w}_{kj}^{(1)} | 0, s_2 \right)^{1-\beta_k^{(1)}},
\]

and

\[
p ( \mathbf{w}_{kk'}^{(2)} | \beta_{kk'}^{(2)} ) = \prod_{j=1}^{p_k} \prod_{k' \neq k, j+1} \mathcal{N} \left( \mathbf{w}_{kk',jl}^{(2)} | 0, s_1 \right)^{\beta_{kk',jl}^{(2)}} \mathcal{N} \left( \mathbf{w}_{kk',jl}^{(2)} | 0, s_2 \right)^{1-\beta_{kk',jl}^{(2)}},
\]

where \( \mathbf{w}_k^{(1)} \) and \( \mathbf{w}_{kk'}^{(2)} \) are the vectors composed of \( w_{kj}^{(1)} \)'s and \( w_{kk',jl}^{(2)} \)'s, \( s_1 \) and \( s_2 \) are two parameters with \( s_1 > s_2 > 0 \) and \( s_2 \) being very small, \( \beta_{kj}^{(1)} \)'s are the main-effect-selection indicators with \( \beta_{kj}^{(1)} = 1 \) if the \( j \)th main effect in the \( k \)th network is selected and 0 otherwise, and \( \beta_{kk',jl}^{(2)} \)'s are the interaction-selection indicators defined similarly as \( \beta_{kj}^{(1)} \)'s.

To accommodate the “main effects-interactions” hierarchy, we introduce a latent vector \( \tilde{\mathbf{w}}_{kk'}^{(2)} \) for interactions with two priors. The first one is based on the main-effect-selection indicators:

\[
p \left( \tilde{\mathbf{w}}_{kk'}^{(2)} | \beta_k^{(1)} , \beta_{kk'}^{(2)} \right) = \prod_{j=1}^{p_k} \prod_{k' \neq k, j+1} \mathcal{N} \left( \tilde{\mathbf{w}}_{kk',jl}^{(2)} | 0, s_1 \right)^{\beta_{k,j}^{(1)}} \beta_{kk',jl}^{(2)} \mathcal{N} \left( \tilde{\mathbf{w}}_{kk',jl}^{(2)} | 0, s_2 \right)^{1-\beta_{k,j}^{(1)} - \beta_{kk',jl}^{(2)}}, \tag{4}
\]

and the second one is based on the interaction effect vector \( \mathbf{w}_{kk'}^{(2)} \):

\[
p \left( \tilde{\mathbf{w}}_{kk'}^{(2)} | \mathbf{w}_{kk'}^{(2)} \right) = 1_{\{ \mathbf{w}_{kk'}^{(2)} = \tilde{\mathbf{w}}_{kk'}^{(2)} \}}, \tag{5}
\]

where \( 1_{\{ \cdot \}} \) is an indicator function.

Denote \( \mathbf{w}_k = \left( \mathbf{w}_k^{(1)} \right)^T, \left( \mathbf{w}_k^{(2)} \right)^T \) as the effect of the \( k \)th network. To utilize network selection and accommodate the network structures, a corresponding latent vector \( \tilde{\mathbf{w}}_k = \left( \tilde{\mathbf{w}}_k^{(1)} \right)^T, \left( \tilde{\mathbf{w}}_k^{(2)} \right)^T \) is introduced with prior:

\[
p \left( \tilde{\mathbf{w}}_k | \alpha_k \right) = \mathcal{N} \left( \tilde{\mathbf{w}}_k | 0, \mathbf{L}_k + \epsilon \mathbf{I} \right)^{-\alpha_k} \mathcal{N} \left( \tilde{\mathbf{w}}_k | 0, \mathbf{I} \right)^{1-\alpha_k} = p (0 | \tilde{\mathbf{w}}_k, \alpha_k), \tag{6}
\]

where \( \alpha_k \) is the network-selection indicator, \( \mathbf{L}_k = \left( \mathbf{L}_k^{(1)} \right) \) with \( \mathbf{L}_k^{(1)} \) and \( \mathbf{L}_k^{(2)} \) describing the network structures of the main effects and interactions, respectively, \( \mathbf{I} \) is an identity matrix, and \( \epsilon \) is a small constant (\( \epsilon = 10^{-6} \) in our numerical studies) to make \( \mathbf{L}_k + \epsilon \mathbf{I} \) strictly
positive-definite. Specifically, consider the Laplacian matrix $L_k^{(1)} = I - D_k^{-1/2} E_k D_k^{-1/2}$ with $D_k = \text{diag}(\sum_{l=1}^{p_k} E_k(1, l), \cdots, \sum_{l=1}^{p_k} E_k(p_k, l))$, and $L_k^{(2)}$ is a block-diagonal matrix with each block $L_{k,j}^{(2)}$, $j = 1, \cdots, p_k - 1$ being the Laplacian matrix for \{ $x_{ij}^{(k)}, x_{i(j+1)}^{(k)}, \cdots, x_{ip_k}^{(k)}$ \} defined in a similar manner as that for \{ $x_{i(j+1)}^{(k)}, \cdots, x_{ip_k}^{(k)}$ \}. To further accommodate the “main effects/interactions-networks” hierarchy, the other prior is assumed for $\bar{w}_k$:

$$p(\bar{w}_k|w_k) = 1\{\bar{w}_k = w_k\}.$$ (7)

Denote $\xi = (\xi_j)_{(K+p+(p-1)p/2)\times1}$ as the vector composed of all selection indicators $\alpha_k$'s, $\beta_{kj}^{(1)}$'s, and $\beta_{kk',jl}^{(2)}$'s. For each $\xi_j$, we assume a Bernoulli prior $p(\xi_j|\psi_j) = \psi_j (1 - \psi_j)^{-\xi_j}$, where $\psi_j$ determines the expected selected proportion. A Beta prior $\text{Beta}(\psi_j|a, b)$ is further assumed for $\psi_j$ with $a$ and $b$ being the two parameters.

Thus, the parameters’ posterior distribution is:

$$p(w, \bar{w}, \xi, \psi, \tau|y, X) = \frac{1}{C} \prod_{i=1}^{n} p(y_i|x_i, w, \tau) \prod_{k=1}^{K} \left\{ p\left(w_k^{(1)}|\beta_k^{(1)}\right) \prod_{k' = 1}^{K} p\left(w_{kk'}^{(2)}|\beta_{kk'}^{(2)}\right) p\left(0|w_{kk'}, \beta_k^{(1)}, \beta_{k'}^{(1)}\right) \right\} ^{K+p+(p-1)p/2} \prod_{j=1}^{K} p(\xi_j|\psi_j) \text{Beta}(\psi_j|a, b) \Gamma(\tau|g, h),$$ (8)

where $y$ and $X$ are the vector and matrix composed of $y_i$'s and $x_i$'s, respectively, and $C$ is the normalization constant. The graphical representation of (8) is given in Figure 1. Denote $E(\xi_j)$ as the posterior expectation of the selection indicator $\xi_j$ under (8). Following Narisetty et al. (2019), we adopt the thresholding approach, and the main effects (interactions, networks) with $E(\xi_j)$'s larger than 0.5 are identified as important.

Rationale In (1), the main effects, and their interactions within and across networks are jointly analyzed, where the identification is achieved using the spike and slab priors (2) and (3). Priors (4) and (5) are developed to accommodate the strong “main effects-interactions” hierarchy. Specifically, if an interaction is selected with $\beta_{kk',jl}^{(2)} = 1$, then $\bar{w}_{kk',jl}^{(2)} = w_{kk',jl}^{(2)} \neq 0$, leading to $\beta_{kj}^{(1)} \beta_{k'l}^{(1)} = 1$ (i.e.
\( \beta_{kj}^{(1)} = \beta_{k'l'}^{(1)} = 1 \) with a higher probability. Prior \( p(\tilde{w}_k|\alpha_k) \) is introduced in (6) to select networks and also accommodate the network structures. Specifically, when the kth network is selected \( (\alpha_k = 1) \), the precision matrix for \( \tilde{w}_k^{(1)} \) is related to the Laplacian matrix \( L_k^{(1)} \), where the jth and lth main genetic factors are conditionally dependent if \( E_k(j, l) = 1 \). Therefore, (6) promotes the effects of connected main factors over the kth network to be similar. With \( L_k^{(2)} \), for each main factor, we define the network structure of its interactions using that of the interacted main factors, following the literature [Wu et al., 2020]. Thus, similar to \( \tilde{w}_k^{(1)} \), elements of \( \tilde{w}_k^{(2)} \) are promoted to be similar if the corresponding interactions are connected. The “main effects/interactions-networks” hierarchy is achieved using prior (7). Specifically, if \( \alpha_k = 0 \), we have \( w_k = \tilde{w}_k \approx 0 \), leading to all \( \beta_{kj}^{(1)} \)'s and \( \beta_{kk,jl}^{(2)} \)'s being zero with a higher probability. Moreover, if at least one of \( \beta_{kj}^{(1)} \)'s and \( \beta_{kk,jl}^{(2)} \)'s is nonzero, \( \alpha_k \) is also nonzero with a higher probability. We rewrite priors (4) and (6) as the generative models with observation vector \( 0 \). As such, the posterior distribution (8) can be formulated as a hybrid Bayesian model which includes tractable partition functions and effectively approximated with the variational inference. The priors for other parameters are set to be conjugate to facilitate posterior inference.

2.2 Computation

We develop an effective variational inference algorithm. Compared to MCMC techniques, variational approximation is computationally more efficient and more feasible with high dimensional parameters. Specifically, we consider minimizing the Kullback-Leibler (KL) divergence between the exact and approximate posterior distributions:

\[
\mathrm{KL}(Q(\Omega)||p(\Omega|y, X)) = \int Q(\Omega) \log \left[ \frac{Q(\Omega)}{p(\Omega|y, X)} \right] \, d\Omega, 
\]

where \( Q(\Omega) = Q(w)Q(\xi)Q(\psi)Q(\tau) \) is a candidate approximate distribution of our true posterior distribution (8), and \( \Omega \) represents all latent variables. Note that with \( p(\tilde{w}_k|w_k) \) and \( p\left( \tilde{w}_k^{(2)}|w_k^{(2)} \right) \) being the indicator functions, there is no need to include the separate distributions \( Q(\tilde{w}) \) and
Figure 1: Graphical representation of the proposed hybrid Bayesian approach.
\(Q(\omega)\) in the approximate distribution. After some derivations, we obtain the optimal variational distribution \(Q(\Omega)\) as follows (we refer to the Web Appendix A for details),

\[
Q(\omega) = \prod_{k=1}^{K} \prod_{j=1}^{p_k} \mathcal{N}\left(\omega_{kj}^{(1)} \mid m_{kj}^{(1)}, \left(\sigma_{kj}^{(1)}\right)^2\right) \prod_{k' = k}^{K} \prod_{k' = k+1}^{K} \prod_{j = j+1}^{j'} \mathcal{N}\left(\omega_{kk'j'l}^{(2)} \mid m_{kk'j'l}^{(2)}, \left(\sigma_{kk'j'l}\right)^2\right),
\]

\[
Q(\xi) = \prod_{j=1}^{K+p+p(p-1)/2} \gamma_j^{\xi_j} (1 - \gamma_j)^{1-\xi_j},
\]

\[
Q(\psi) \propto \prod_{j=1}^{K+p+p(p-1)/2} (\psi_j^{\tilde{a}_j - 1} (1 - \psi_j)^{\tilde{b}_j - 1},
\]

\[
Q(\tau) = \text{Gamma}(\tau \mid \tilde{g}, \tilde{h}),
\]

where \(\left(\left(m_{kj}^{(1)}, \sigma_{kj}^{(1)}, m_{kk'j'l}^{(2)}, \sigma_{kk'j'l}\right), \left(\tilde{a}_j, \tilde{b}_j\right), \left(\tilde{g}, \tilde{h}\right)\right)\) are the corresponding estimated values of the parameters for the Gaussian, Beta, and Gamma distributions, respectively, and \(\gamma_j\) is the expectation of \(\xi_j\) under the approximate distribution \(Q(\Omega)\). The proposed algorithm updates the estimators iteratively until convergence (Algorithm 1 in the Web Appendix A) and adopts the final estimated value of \(\gamma_j\) as the estimator of \(E(\xi_j)\).

To proceed with this algorithm, we need to set values for the latent parameters. Following the literature [Zhe et al., 2013], we consider an uninformative diffuse Gamma prior with \(g = h = 10^{-12}\) and a uniform Beta prior with \(a = b = 1\). The proposed model involves two tuning parameters \(s_1\) and \(s_2\). Our numerical investigation suggests that the value of \(s_1\) is not very important when it is in a sensible range. To reduce computational cost, we fix \(s_1 = 1\) in our numerical studies. The value of \(s_2\) is selected using the Bayesian information criterion (BIC). The proposed algorithm is computationally feasible. Take a simulated dataset with \(p = 1,000\) and \(n = 300\) as an example. With fixed tuning parameter, the proposed analysis takes about half a minute using a laptop with standard configurations. To facilitate data analysis, we have developed R package \(JNNI\) implementing the proposed approach, which is publicly available at https://github.com/mengyunwu2020/JNNI and can be installed with devtools.
3 Simulation

We perform simulations to evaluate performance of the proposed approach under the following settings. (a) \( n = 300 \) and \( p = 1,000 \). Thus, there are a total of 1,000 candidate main effects and 499,500 interactions. (b) Consider two settings for the number of networks with \( K = 100 \) and 50. (c) We follow the network construction procedure used in Zhao and Shojaie (2016). Specifically, for the \( k \)th network \( (k = 1, \cdots , K) \), set \( p_k = \frac{p}{K} \), generate one transcription factor (TF) \( x_{TF} \) from \( \mathcal{N}(0,1) \), and then generate the rest \( p_k - 1 \) genetic factors from \( \mathcal{N}(\rho x_{TF},1-\rho^2) \) with a parameter \( \rho \). Consider \( \rho = 0.2, 0.4, \) and \( 0.6 \), representing weak, moderate, and strong dependence between the TF and its target factors in each network. Genetic factors with nonzero correlations are connected in the network. (d) There are three important networks, where 18 main genetic effects and 17 interactions have nonzero coefficients. Both the “main effects/interactions-networks” and “main effects-interactions” hierarchies are satisfied. Nonzero signals of the important TFs are generated from Uniform(0.8,1.2), and the other important main effects and interactions have relatively weaker signals with a ratio \( r \) of that of the corresponding important TF. Consider \( r = 1/\sqrt{5} \) and \( 1/\sqrt{12} \). Four specific settings S1-S4 for the important variables are considered. Under setting S1, all signals are positive. Setting S2 is the same as S1, except that the signals in the second network and those between the first and second networks are negative. Under setting S3, within each network, the signals can be either positive or negative. Under setting S4, the important interactions only involve the none-TF main effects with weaker signals. We refer to the Web Appendix B for more details. (e) For the response, we generate \( y_i \) from the Gaussian distribution (1) with variance 1. There are 48 scenarios, comprehensively covering a wide spectrum with different levels of correlations within networks and signals associated with the response, and different patterns of networks and associations.

In addition to the proposed approach, six alternatives are conducted. (a) glinternet, which learns linear interaction model based on the hierarchical group-Lasso regularization and is imple-
mented using the R package \textit{glinternet} \cite{lim2015}. (b) Lasso, which applies the Lasso penalization to both main effects and all pairwise interactions directly and is realized using the R package \textit{glmnet}. (c) iFORM, which identifies interactions in a greedy forward fashion while maintaining the hierarchical structure \cite{hao2014}. (d) HierNet, which is Lasso for hierarchical interactions by adding a set of convex constraints and is realized using the R package \textit{HierNet} \cite{breheny2015}. (e) Grace, which applies the graph-constrained estimation method developed by Li and Li \cite{li2008} to both main effects and all pairwise interactions. (f) GEL, which achieves a bi-level variable selection for groups and individual predictors (main effects and interactions) in those groups \cite{breheny2015}. Among these methods, glinternet and iFORM respect the strong “main effects-interactions” hierarchy. We consider HierNet with the weak hierarchy, as the counterpart with strong hierarchy is not computationally feasible in large-scale simulations. Lasso, Grace, and GEL are originally developed for main-effect analysis, and we extend them for interaction analysis by modeling additional all pairwise interactions, without enforcing the “main effects-interactions” hierarchy. Both Grace and GEL incorporate the network information, where Grace accommodates the underlying network structures, and GEL achieves the joint selection of interactions and networks.

To evaluate identification performance, we compute the numbers of true positives and false positives for main effects (M:TP and M:FP) and interactions (I:TP and I:FP), respectively. For the proposed approach and GEL, we also consider the true positives and false positives (N:TP and N:FP) for identifying networks. Estimation performance is assessed using the root sum of squared errors (RSSE) defined as $||\hat{w}_M - w_0^M||_2$ and $||\hat{w}_I - w_0^I||_2$ for main effects and interactions, where $(\hat{w}_M, \hat{w}_I)$ and $(w_0^M, w_0^I)$ are the estimated and true values of coefficients. For prediction evaluation, we adopt the prediction median-squared error (PMSE) based on independent testing data with 100 subjects.

Under each scenario, we simulate 100 replications. The summary results under the scenarios
with $\rho = 0.2$ and $r = 1/\sqrt{5}$ are presented in Table 1 ($K = 100$) and Table 2 ($K = 50$). The rest of the results are shown in the Web Appendix B.

It is observed that across the whole spectrum of simulation, the proposed approach has superior or similar performance compared to the alternatives with respect to both selection and prediction accuracy. It is able to identify the majority of true positives, while having much fewer false positives than most alternatives. For instance, under the scenario with setting S4 in Table 1, the proposed approach has $(M:TP, M:FP, I:TP, I:FP) = (17.72, 1.40, 14.02, 5.14)$, compared to $(12.06, 1.66, 3.48, 0.88)$ for glinternet, $(5.56, 0.08, 2.02, 79.42)$ for Lasso, $(14.56, 41.84, 6.98, 36.02)$ for iFORM, $(7.56, 0.16, 1.80, 2.72)$ for HierNet, $(10.32, 0.30, 5.74, 23.40)$ for Grace, and $(17.86, 12.40, 138.74)$ for GEL. Under the scenarios in Table 2 with larger network sizes ($p_k = 20$), the advantages of the proposed approach become more prominent, especially under setting S4, where the important interactions have main effects with weaker signals. Specifically, the proposed approach has $(M:TP, M:FP, I:TP, I:FP) = (16.38, 1.40, 11.56, 5.52)$, compared to $(11.14, 1.34, 2.74, 0.86)$ for glinternet, $(5.56, 0.14, 1.96, 81.34)$ for Lasso, $(13.96, 43.82, 5.82, 35.68)$ for iFORM, $(7.74, 0.24, 1.76, 2.70)$ for HierNet, $(9.46, 0.50, 5.46, 27.02)$ for Grace, and $(9.84, 25.50, 5.86, 331.56)$ for GEL. The proposed approach also performs well in terms of estimation. For example, under setting S1 in Table 1, the proposed approach has $(M:RSSE, I:RSSE) = (0.31, 0.46)$, compared to $(1.27, 1.36)$ for glinternet, $(1.76, 1.52)$ for Lasso, $(1.26, 1.16)$ for iFORM, $(1.50, 1.48)$ for HierNet, $(1.45, 1.30)$ for Grace, and $(0.71, 1.96)$ for GEL. In addition, higher prediction accuracy of the proposed approach is observed. For example, under setting S2 in Table 2, the PMSEs are 0.85 (proposed), 2.02 (glinternet), 2.89 (Lasso), 2.13 (iFORM), 2.40 (HierNet), 2.79 (Grace), and 25.42 (GEL), respectively. Furthermore, we note that the proposed approach identifies all important networks correctly with $N:FP=0$ under all scenarios. In contrast, GEL cannot effectively select important networks (especially under setting S4 with $N:TP=2.47$ and $N:FP=0.06$) and often misidentifies networks (details omitted). The glinternet approach generally has the second best performance, and under some scenarios with
higher within network correlations ($\rho = 0.6$) and simpler signal patterns (S1 and S2), it behaves competitively in main-effect identification. However, the proposed approach can keep its superiority in interaction identification, estimation, and prediction. With a lower signal level ($r = 1/\sqrt{12}$), performance of all methods decay as expected, however, the proposed approach is again observed to perform favourably.

To mimic the scenarios under which a genetic factor (i.e. a gene) is involved in multiple networks, we conduct additional simulations with $\rho = 0.2$, $r = 1/\sqrt{5}$, and $K = 100$. Specifically, among the 1,000 genetic factors, there are 100 each of which is involved in 2 to 6 networks. Summary results are provided in Web Table 11 (Web Appendix B). Similar conclusions can be drawn that the proposed approach has significant advantages over the alternatives.

4 Data Analysis

We analyze The Cancer Genome Atlas (TCGA) data on cutaneous melanoma (SKCM) and lung adenocarcinoma (LUAD) to identify important interactions (main effects, and networks) associated with phenotypes/outcomes. As one of the largest cancer genetics program, TCGA contains unique and valuable information. In this study, we consider mRNA gene expression measurements which are downloaded from the TCGA Provisional using the R package *cgdsr*. Networks are constructed using the information from KEGG, which is a popular choice in recent network analysis studies (Zhou and Zheng 2013; Zhe et al. 2013; Gao et al. 2019). Specifically, we follow Gao et al. (2019) and obtain the network structures from KEGG pathway database using the R package *KEGGgraph*, where each pathway is presented as a network with nodes being molecules (protein, compound, etc) and edges representing relation types between the nodes, e.g. activation or phosphorylation (Zhang and Wiemann 2009). Here, we set $E_k(j, l) = E_k(l, j) = 1$ if the $j$th and $l$th genes are connected in the pathway and 0 otherwise.
Table 1: Simulation results under the scenarios with $\rho = 0.2$, $r = 1/\sqrt{5}$, and $K = 100$. In each cell, mean (SD) based on 100 replicates.

| Approach   | M:TP    | M:FP    | M:RSSE | I:TP    | I:FP    | I:RSSE | PMSE |
|------------|---------|---------|--------|---------|---------|--------|------|
| S1         |         |         |        |         |         |        |      |
| proposed   | 17.82(0.39) | 0.16(0.51) | 0.31(0.12) | 16.34(0.82) | 3.36(2.25) | 0.46(0.15) | 0.57(0.19) |
| glinternet | 13.46(2.91) | 3.14(3.78) | 1.27(0.17) | 8.36(3.21) | 2.16(3.01) | 1.36(0.12) | 1.95(0.46) |
| Lasso      | 3.88(1.19) | 0.00(0.00) | 1.76(0.12) | 3.08(2.05) | 3.12(3.00) | 1.52(0.06) | 2.87(0.59) |
| iFORM      | 16.84(1.90) | 37.28(4.05) | 1.26(0.31) | 14.18(4.01) | 29.68(3.83) | 1.16(0.35) | 1.82(0.92) |
| HierNet    | 8.20(3.36) | 0.50(0.97) | 1.50(0.18) | 4.14(2.65) | 5.00(5.18) | 1.48(0.09) | 2.34(0.66) |
| Grace      | 10.30(2.13) | 0.14(0.35) | 1.45(0.09) | 9.60(2.34) | 22.54(14.49) | 1.30(0.09) | 2.18(0.53) |
| GEL        | 18.00(0.00) | 12.44(1.98) | 0.71(0.14) | 13.30(0.74) | 140.14(7.24) | 1.96(0.24) | 2.30(0.85) |
| S2         |         |         |        |         |         |        |      |
| proposed   | 17.24(0.89) | 0.18(0.52) | 0.40(0.16) | 15.62(1.24) | 3.92(2.35) | 0.59(0.20) | 0.66(0.22) |
| glinternet | 12.16(3.81) | 2.34(2.39) | 1.31(0.19) | 7.72(3.84) | 1.50(1.82) | 1.44(0.13) | 2.14(0.59) |
| Lasso      | 4.84(2.12) | 0.24(0.62) | 1.63(0.12) | 3.10(2.48) | 92.94(140.21) | 1.67(0.09) | 2.98(0.72) |
| iFORM      | 16.08(2.54) | 39.54(4.87) | 1.36(0.40) | 13.50(4.32) | 29.22(3.74) | 1.24(0.38) | 2.09(1.12) |
| HierNet    | 7.00(3.16) | 0.36(0.78) | 1.50(0.14) | 4.16(2.66) | 4.20(4.66) | 1.53(0.10) | 2.52(0.62) |
| Grace      | 6.04(1.55) | 0.50(0.86) | 1.73(0.05) | 4.00(1.29) | 25.82(12.04) | 1.56(0.05) | 2.94(0.74) |
| GEL        | 18.00(0.00) | 12.40(1.98) | 0.72(0.14) | 13.32(0.68) | 140.58(7.26) | 2.00(0.26) | 2.19(0.80) |
| S3         |         |         |        |         |         |        |      |
| proposed   | 17.74(0.49) | 0.14(0.50) | 0.36(0.13) | 14.06(1.41) | 3.90(2.87) | 0.76(0.16) | 0.73(0.22) |
| glinternet | 11.54(3.31) | 2.34(2.45) | 1.47(0.17) | 4.44(2.87) | 1.12(1.38) | 1.49(0.09) | 2.23(0.58) |
| Lasso      | 3.46(1.59) | 0.00(0.00) | 1.90(0.11) | 1.64(1.76) | 2.24(2.89) | 1.56(0.04) | 2.93(0.69) |
| iFORM      | 16.54(2.37) | 38.28(5.02) | 1.34(0.39) | 13.10(4.99) | 30.64(4.23) | 1.22(0.40) | 2.09(1.24) |
| HierNet    | 7.70(2.74) | 0.24(0.43) | 1.65(0.15) | 2.86(1.90) | 3.54(2.82) | 1.51(0.07) | 2.51(0.70) |
| Grace      | 9.48(1.59) | 0.58(0.86) | 1.53(0.10) | 7.80(1.84) | 33.76(17.50) | 1.38(0.07) | 2.17(0.49) |
| GEL        | 18.00(0.00) | 12.62(2.40) | 0.74(0.14) | 12.24(0.43) | 142.96(8.39) | 2.04(0.23) | 2.32(0.73) |
| S4         |         |         |        |         |         |        |      |
| proposed   | 17.72(0.54) | 1.40(0.67) | 0.37(0.11) | 14.02(1.61) | 5.14(2.47) | 0.76(0.19) | 0.75(0.20) |
| glinternet | 12.06(3.42) | 1.66(1.57) | 1.34(0.16) | 3.48(2.27) | 0.88(1.04) | 1.52(0.06) | 2.25(0.66) |
| Lasso      | 5.56(2.84) | 0.08(0.34) | 1.65(0.14) | 2.02(2.33) | 79.42(131.40) | 1.62(0.08) | 2.89(0.71) |
| iFORM      | 14.56(2.54) | 41.84(4.00) | 1.78(0.41) | 6.98(4.70) | 36.02(4.39) | 1.80(0.39) | 3.38(1.48) |
| HierNet    | 7.56(2.96) | 0.16(0.37) | 1.54(0.14) | 1.80(1.82) | 2.72(3.10) | 1.56(0.05) | 2.64(0.77) |
| Grace      | 10.32(2.49) | 0.30(0.65) | 1.44(0.11) | 5.74(2.56) | 23.40(12.56) | 1.48(0.07) | 2.34(0.73) |
| GEL        | 17.86(0.99) | 12.42(2.99) | 0.78(0.16) | 12.40(1.41) | 138.74(21.81) | 2.09(0.27) | 2.36(0.84) |
Table 2: Simulation results under the scenarios with $\rho = 0.2, r = 1/\sqrt{5}$, and $K = 50$. In each cell, mean (SD) based on 100 replicates

| Approach   | S1 | M:TP | M:FP | M:RSSE | I:TP | I:FP | I:RSSE | PMSE |
|------------|----|------|------|--------|------|------|--------|------|
| proposed   |    | 16.70(1.16) | 0.70(1.33) | 0.54(0.20) | 14.88(1.61) | 4.82(4.84) | 0.65(0.20) | 0.74(0.21) |
| glinternet |    | 12.66(3.37) | 2.80(2.60) | 1.31(0.18) | 7.76(3.68) | 1.80(2.27) | 1.39(0.12) | 2.05(0.74) |
| Lasso      |    | 3.80(1.20) | 0.02(0.14) | 1.76(0.13) | 3.06(1.98) | 3.70(3.10) | 1.53(0.06) | 2.85(0.84) |
| iFORM      |    | 16.22(2.01) | 38.30(4.13) | 1.37(0.36) | 13.46(3.70) | 30.16(3.20) | 1.23(0.34) | 1.96(1.02) |
| HierNet    |    | 8.30(3.01) | 0.62(1.14) | 1.51(0.15) | 4.38(2.71) | 5.04(5.52) | 1.48(0.09) | 2.42(0.68) |
| Grace      |    | 9.72(2.18) | 0.64(0.75) | 1.47(0.09) | 9.38(2.35) | 27.96(15.38) | 1.33(0.10) | 2.15(0.53) |
| GEL        |    | 9.24(3.02) | 24.36(7.46) | 2.50(0.49) | 5.62(2.52) | 318.30(81.05) | 6.67(1.57) | 21.26(12.99) |

| Approach   | S2 | M:TP | M:FP | M:RSSE | I:TP | I:FP | I:RSSE | PMSE |
|------------|----|------|------|--------|------|------|--------|------|
| proposed   |    | 15.58(1.14) | 0.40(0.64) | 0.65(0.14) | 14.84(1.18) | 3.30(2.32) | 0.69(0.16) | 0.85(0.25) |
| glinternet |    | 12.44(4.01) | 3.68(4.01) | 1.25(0.21) | 9.02(4.37) | 2.24(2.54) | 1.38(0.17) | 2.02(0.63) |
| Lasso      |    | 5.20(2.54) | 0.22(0.68) | 1.62(0.11) | 3.32(2.58) | 86.80(135.53) | 1.65(0.10) | 2.89(0.74) |
| iFORM      |    | 15.88(2.82) | 38.86(4.34) | 1.40(0.43) | 13.34(4.54) | 29.86(4.48) | 1.26(0.45) | 2.13(1.20) |
| HierNet    |    | 7.82(3.47) | 0.62(1.16) | 1.46(0.17) | 5.46(3.45) | 4.58(4.45) | 1.49(0.13) | 2.40(0.70) |
| Grace      |    | 5.96(1.63) | 0.54(0.91) | 1.72(0.05) | 4.06(1.35) | 28.46(14.27) | 1.54(0.07) | 2.79(0.66) |
| GEL        |    | 9.88(3.02) | 24.16(7.43) | 2.56(0.72) | 7.48(2.38) | 319.42(88.63) | 7.29(3.31) | 25.42(29.24) |

| Approach   | S3 | M:TP | M:FP | M:RSSE | I:TP | I:FP | I:RSSE | PMSE |
|------------|----|------|------|--------|------|------|--------|------|
| proposed   |    | 16.38(1.37) | 1.36(1.48) | 0.62(0.19) | 11.92(1.68) | 6.10(4.09) | 0.96(0.14) | 0.93(0.30) |
| glinternet |    | 9.94(4.21) | 2.18(3.15) | 1.55(0.22) | 3.66(3.32) | 0.86(1.53) | 1.51(0.10) | 2.39(0.61) |
| Lasso      |    | 3.24(1.71) | 0.02(0.14) | 1.93(0.13) | 1.32(1.53) | 1.70(2.76) | 1.57(0.04) | 3.16(0.75) |
| iFORM      |    | 16.50(2.05) | 38.52(4.39) | 1.33(0.37) | 12.94(4.42) | 30.00(4.86) | 1.24(0.38) | 1.86(0.87) |
| HierNet    |    | 6.50(2.97) | 0.26(0.63) | 1.71(0.17) | 2.48(1.98) | 3.48(3.65) | 1.52(0.07) | 2.70(0.66) |
| Grace      |    | 9.12(1.77) | 0.96(1.05) | 1.55(0.11) | 7.64(2.07) | 40.62(16.53) | 1.39(0.08) | 2.24(0.53) |
| GEL        |    | 7.56(2.66) | 22.84(8.85) | 2.41(0.42) | 3.84(2.46) | 292.16(87.24) | 6.31(1.49) | 18.77(12.56) |

| Approach   | S4 | M:TP | M:FP | M:RSSE | I:TP | I:FP | I:RSSE | PMSE |
|------------|----|------|------|--------|------|------|--------|------|
| proposed   |    | 16.38(1.28) | 1.40(1.18) | 0.62(0.18) | 11.56(2.34) | 5.52(3.12) | 0.99(0.20) | 0.99(0.26) |
| glinternet |    | 11.14(3.62) | 1.34(1.71) | 1.37(0.16) | 2.74(2.47) | 0.86(1.18) | 1.54(0.06) | 2.13(0.61) |
| Lasso      |    | 5.56(2.31) | 0.14(0.35) | 1.62(0.11) | 1.96(2.23) | 81.34(133.92) | 1.62(0.08) | 2.77(0.65) |
| iFORM      |    | 13.96(2.74) | 43.82(4.67) | 1.93(0.45) | 5.82(4.77) | 35.68(4.90) | 1.85(0.38) | 3.61(1.47) |
| HierNet    |    | 7.74(2.64) | 0.24(0.48) | 1.52(0.12) | 1.76(1.71) | 2.70(2.99) | 1.56(0.04) | 2.49(0.62) |
| Grace      |    | 9.46(2.06) | 0.50(0.68) | 1.46(0.09) | 5.46(2.92) | 27.02(14.57) | 1.50(0.07) | 2.26(0.61) |
| GEL        |    | 9.84(2.91) | 25.50(7.32) | 2.40(0.60) | 5.86(2.73) | 331.56(83.89) | 6.46(1.89) | 19.81(13.80) |
4.1 Cutaneous melanoma (SKCM) data

The response of interest is the (log-transformed) Breslow’s thickness, which is a measure of melanoma growth and has been widely used in the assessment of melanoma. Data are available on 361 subjects and 19,904 gene expression measurements. Although the proposed approach is potentially applicable to a large number of genes, with the consideration that the number of cancer-related genes is not large, as well as to improve stability, a simple marginal screening is conducted. Specifically, the top 2,000 genes with the smallest p-values computed from a marginal linear model are selected. Among them, we construct 173 networks with the sizes from 1 to 50, containing 1,505 genes in total and 578 distinct genes after removing duplicates.

15 distinct main effects and 19 distinct interactions are identified by the proposed approach (19 main effects and 31 interactions before removing duplicates). The identified genes, their interactions, as well as networks are showed in Figure 2, where two genes are connected if the corresponding interaction is also selected. The detailed estimation results are provided in Web Table 12. Literature search suggests that the identified genes may be of great significance. For example, it has been found that high expression of gene PMM2 is associated with poor prognosis in melanoma. Gene FBP1, which is involved in three identified networks, has been shown to be significantly down-regulated in human melanoma cells. The expression of gene PCK2 has been found to be down-regulated in melanoma regenerative cells and closely related to the survival of tumor patients. Published studies have reported that gene PFKFB4, a known regulator of glycolysis, displays an unconventional role in melanoma cell migration and has increased expression levels in several human tumors including cutaneous melanoma. The simultaneous inactivation of genes HK1 and HK2 has been demonstrated to be sufficient to decrease the proliferation and viability of melanoma. In addition, gene PMM1 has been identified in published studies to be regulated in human melanoma and melanoma-associated pathways.

Furthermore, the proposed approach identifies seven networks, all of which are metabolics re-
lated and have important biological implications. For example, in a recent study, Citrate cycle (TCA cycle) has been suggested to be significantly down-regulated, while Galactose metabolism is up-regulated in tumor formation and progression. Other interesting networks have been linked to the development, progression, and outcome of melanoma. For instance, Fatty Acid metabolism has been shown to be essential for cancer cell proliferation. Glycolysis has been confirmed to play a significant role in developing metabolic symbiosis in metastatic melanoma progression. In addition, the Pentose phosphate has been found to be critical for cancer cell survival and ribonucleotide, as well as lipid biosynthesis.

Beyond the proposed approach, we also conduct analysis using the alternatives. The summary comparison results are presented in Table 3, where the numbers of main effects and interactions identified by different approaches, their overlaps, and RV coefficients are provided. Here, RV coefficient is a measure that describes the similarity of two matrices, and a larger value indicates a higher similarity. Different approaches are observed to identify quite different sets of main effects and interactions, and have moderate similarity as suggested by the RV coefficients. In particular, Lasso, Grace, and GEL, without accounting for the “main effects-interactions” hierarchy, identify a larger number of interactions than main effects. The other four approaches with the “main effects-interactions” hierarchy, including the proposed one, select a moderate number of main effects and interactions.

We further use a resampling approach to examine prediction performance and selection stability. The subjects are randomly partitioned into a training and a testing set. The mean PMSEs for the testing subjects over 100 resamplings are 0.57 (proposed), 0.60 (glinternet), 0.60 (Lasso), 1.16 (iFORM), 0.57 (HierNet), 0.64 (Grace), and 6.32 (GEL), suggesting the satisfactory prediction accuracy of the proposed approach. To evaluate selection stability, for each of the aforementioned important main effects and interactions, we compute its observed occurrence index (OOI), which is the selected frequency in 100 resamplings. The proposed approach has the mean OOI value of
Figure 2: Analysis of the TCGA SKCM data using the proposed approach: identified main genetic effects, interactions, and networks. Different colors represent different networks, and two genes are connected if the corresponding interaction is also selected.

0.95, compared to 0.07 (glinternet), 0.28 (Lasso), 0.15 (iFORM), 0.62 (HierNet), 0.78 (Grace), and 0.50 (GEL). With the joint network selection and incorporated network structures, the proposed approach has significant improvement in selection stability. The prediction and stability analysis provides a certain degree of confidence to the proposed identification analysis.

4.2 Lung adenocarcinoma (LUAD) data

The response of interest is the reference value for the pre-bronchodilator forced expiratory volume in one second in percent (FEV1). It is a major indicator of pulmonary function impairment. Data are available on 232 subjects and 18,325 gene expression measurements. We conduct a similar prescreening, and 181 networks, containing 1,360 genes in total and 499 distinct genes are found.

With the proposed approach, 22 main effects and 21 interactions (28 main effects and 34 interactions before removing duplicates) are identified and presented in Figure 3. The detailed estimation
results are provided in Web Table 13. Independent evidences of their biological implications have been reported in the literature. For example, the ALDH2 locus has been associated with a higher risk of lung cancer among light smokers. Activated ACLY has been suggested as a negative prognostic factor in lung adenocarcinomas. Significantly higher ACSS2 expression has been observed in a substantial number of lung tumor samples. Published studies have demonstrated that late-stage LUAD patients have higher expression levels of HK2 and GBE1 than early-stage ones. Over-expression of PGAM1 has been observed in multiple human cancer types including lung cancer. In addition, up-regulation of GMDS has been found to be critical for cell proliferation and survival in LUAD and serves as a potential biomarker for LUAD diagnosis and treatment. ACAAI has been confirmed to be down-regulated in lung cancer at both mRNA and protein levels and has favorable effects on the overall survival of LUAD patients.

In addition, the proposed approach identifies nine networks, which have been shown to have possible associations with lung cancer. For instance, it has been suggested that the most important treatment for lung cancer may be through the inhibition of fatty acid synthase in fatty acid metabolism, which has been reported to be over-expressed and over-activated in certain human cancers including lung cancer. In addition, several genes involved in the galactose metabolism, fructose and mannose metabolism, and pentose phosphate have been reported to be related to the pathogenesis of lung cancer. Ascorbate and aldarate metabolism has been regarded as critical in LUAD.

We also conduct analysis using the alternatives and summarize the comparison results in Table 3. Similar to that for SKCM data, different approaches lead to identification results with low levels of overlapping. Prediction performance and selection stability are further examined based on 100 resamplings. The mean (PMSE, OOI) values are (0.05, 0.93) for the proposed approach, (0.05, 0.68) for glinternet, (0.06, 0.33) for Lasso, (0.22, 0.12) for iFORM, (0.04, 0.08) for HierNet, (0.04, 0.19) for Grace, and (0.73, 0.13) for GEL. The proposed approach again has competitive prediction
Discussion

In the study of complex diseases, gene-gene interaction analysis has attracted extended attention. Recently, biological networks have been accumulated, containing information on functionally related genetic groups and within-group structures. Thus, incorporating network information can potentially lead to a deeper biological understanding of phenotypes from a system perspective. In this study, we have developed the gene-gene interaction analysis, where the network information is incorporated. It advances from the existing interaction analyses by taking advantage of the assistance of network selection, where not only the “main effects-interactions” hierarchy but also the “main effects/interactions-networks” hierarchy have been respected. In addition, the graph
Table 3: Data analysis: numbers of main effects and interactions (diagonal elements) identified by different approaches and their overlaps and RV coefficients (off-diagonal elements).

| SKCM       | proposed | glinternet | Lasso | iFORM | HierNet | Grace | GEL |
|------------|----------|------------|-------|-------|---------|-------|-----|
| Main       | 15       | 15         | 15    | 1     | 1       | 1     | 1   |
| Lasso      | 15       | 1(0.22)    | 0(0.02) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| iFORM      | 1        | 1(0.12)    | 0(0.03) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| HierNet    | 51       | 8(0.72)    | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| Grace      | 51       | 1(0.33)    | 1(0.59) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| GEL        | 1        | 0(0.13)    |         |        |         |       | 28  |
| Interaction| 19       | 6(0.59)    | 1(0.08) | 1(0.40) | 0(0.00) | 0(0.00) | 0(0.00) |
| Lasso      | 16       | 8(0.71)    | 1(0.40) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| iFORM      | 20       | 1(0.16)    | 0(0.01) | 1(0.10) | 0(0.00) | 0(0.00) | 0(0.00) |
| HierNet    | 44       | 20(0.84)   | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| Grace      | 2        | 0(0.00)    |         |        |         |       | 16  |
| GEL        | 16       | 1(0.03)    |         |        |         |       | 363 |

| LUAD       | proposed | glinternet | Lasso | iFORM | HierNet | Grace | GEL |
|------------|----------|------------|-------|-------|---------|-------|-----|
| Main       | 22       | 16         | 16    | 4     | 4       | 4     | 4   |
| Lasso      | 22       | 1(0.38)    | 0(0.31) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| iFORM      | 16       | 1(0.49)    | 8(0.71) | 13(0.78) | 0(0.00) | 0(0.00) | 0(0.00) |
| HierNet    | 4        | 2(0.57)    | 1(0.54) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| Grace      | 43       | 20(0.84)   | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| GEL        | 64       | 0(0.00)    |         |        |         |       | 0   |
| Interaction| 21       | 12         | 5(0.34) | 0(0.00) | 1(0.20) | 0(0.00) | 0(0.00) |
| Lasso      | 21       | 0(0.03)    | 0(0.09) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| iFORM      | 253      | 4(0.12)    | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) | 0(0.00) |
| HierNet    | 53       | 1(0.02)    |         |        |         |       | 17  |
| Grace      | 13       | 0(0.00)    |         |        |         |       | 1   |
| GEL        | 17       | 1(0.03)    |         |        |         |       | 269 |
Laplacian Gaussian prior has been adopted to accommodate the underlying network structures. As demonstrated in Cai et al. (2020), under certain regularization conditions, the graph Laplacian Gaussian prior has posterior consistency with a diverging number of nodes and edges in networks. The proposed approach may enjoy very broad applicability, where the networks can be potentially sparse or dense. The spike and slab priors for the regression coefficients and conjugate priors for the other parameters have been adopted, which offers the advantage of analytical simplification (Narisetty et al., 2014, 2019). Since results may be sensitive to the choice of hyperparameters $\psi_j$'s, we have introduced a Beta prior on $\psi_j$ to improve stability. The proposed approach can be formulated as a hybrid Bayesian model, with a solid statistical foundation and the potential to be effectively realized using the variational inference. This is significantly advanced from the published Bayesian interaction analysis that usually adopts MCMC techniques, which have very low computational efficiency. Compared to some other variable selection techniques, such as penalization, Bayesian methods have demonstrated superiority in multiple aspects, such as providing readily available uncertainty estimates and a more informative approach to model selection (Sara et al., 2019; Narisetty et al., 2019). Extensive simulation studies have been conducted, suggesting the practical superiority of the proposed approach in identification, estimation, and prediction. Two TCGA cancer studies have been used to illustrate application, leading to biologically sensible findings with satisfactory prediction accuracy and selection stability.

This study has focused on continuous response and assumed the Gaussian distribution. It can be of interest to extend the proposed approach to handle categorical and censored outcomes. For example, a data augmentation approach based on a probit model (categorical outcome) or an accelerated failure time model (censored outcome) can be potentially adopted (Stingo et al., 2011). However, our preliminary investigation suggests that this extension is expected to be nontrivial and warrants a separate work. The strong “main effects-interactions” hierarchy has been explored in this study, which is popular in recent interaction analysis (Lim and Hastie, 2015; Hao, Feng).
Modification can be potentially conducted in prior (4) to respect the weak hierarchy. The duplication strategy has been adopted to accommodate overlappings in networks due to its simplicity. We acknowledge that when networks have a high overlapping level, the proposed analysis may not be stable. However, in practical data analysis with a moderate overlapping level, such as the SKCM data where some genes are involved in 30 to 50 networks, the proposed approach has been shown to be still validity. More prudent strategies are deferred to further investigation. In data analysis, we have utilized the KEGG information to construct networks. Other information sources, such as Gene Ontology terms and protein-protein interaction networks, can also be adopted.

References

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. Annals of Statistics, 41(3), 1111-1141.

P. Breheny. The group exponential lasso for bi-level variable selection. Biometrics, 71(3):731–740, 2015.

Q. Cai, J. Kang, and T. Yu. Bayesian network marker selection via the thresholded graph Laplacian Gaussian prior. Bayesian Analysis, 15(1):79–102, 2020.

Heather J. Cordell. Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics 10(6):392–404, 2009.

Federico Ferrari and David B Dunson. Bayesian factor analysis for inference on interactions. Journal of the American Statistical Association page DOI: 10.1080/01621459.2020.1745813, 2020.

B. Gao, X. Liu, H. Li, and Y. Cui. Integrative analysis of genetical genomics data incorporating network structures. Biometrics 75(4):1063–1075, 2019.
N. Hao, Y. Feng, and H. H. Zhang. Model selection for high dimensional quadratic regression via regularization. Journal of the American Statistical Association 113(522):615–625, 2018.

N. Hao and H. H. Zhang. Interaction screening for ultrahigh-dimensional data. Journal of the American Statistical Association 109(507):1285–1301, 2014.

N. Hao and H. H. Zhang. A note on high-dimensional linear regression with interactions. The American Statistician 71(4):291–297, 2017.

J. Kim, J. Lim, Y. Kim, and W. Jang. Bayesian variable selection with strong heredity constraints. Journal of the Korean Statistical Society 47(3):314–329, 2018.

C. Li and H. Li. Variable selection and regression analysis for graph-structured covariates with an application to genomics. The Annals of Applied Statistics 4(3):1498–1516, 2010.

M. Lim and T. Hastie. Learning interactions via hierarchical group-lasso regularization. Journal of Computational and Graphical Statistics 24(3):627–654, 2015.

C. Liu, J. Ma, and C. I. Amos. Bayesian variable selection for hierarchical gene-environment and gene-gene interactions. Human Genetics 134(1):23–36, 2015.

T. F. Mackay. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nature Reviews Genetics 15(1):22–33, 2014.

N. N. Narisetty, J. Shen, and X. He. Skinny Gibbs: a consistent and scalable Gibbs sampler for model selection. Journal of the American Statistical Association 114(527):1205–1217, 2019.

Naveen N Narisetty and Xuming He. Bayesian variable selection with shrinking and diffusing priors. Annals of Statistics 42(2):789–817, 2014.

J. Ren, F. Zhou, X. Li, Q. Chen, H. Zhang, S. Ma, Y. Jiang, and C. Wu. Semiparametric Bayesian variable selection for gene-environment interactions. Statistics in Medicine 39(5):617–638, 2020.
F. C. Stingo, Y. A. Chen, M. G. Tadesse, and M. Vannucci. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. The Annals of Applied Statistics 5(3):1978–2002, 2011.

A. Upton, O. Trelles, Cornejo Garcia Jose, and J. Perkins. Review: High-performance computing to detect epistasis in genome scale data sets. Briefings in Bioinformatics 17(3):368–379, 2016.

Sara Van Erp, Daniel L Oberski, and Joris Mulder. Shrinkage priors for Bayesian penalized regression. Journal of Mathematical Psychology 89:31–50, 2019.

M. Wu, J. Huang, and S. Ma. Identifying gene-gene interactions using penalized tensor regression. Statistics in Medicine 37(4):598–610, 2018.

M. Wu and S. Ma. Robust genetic interaction analysis. Briefings in Bioinformatics 20(2):624–637, 2019.

M. Wu, Q. Zhang, and S. Ma. Structured gene-environment interaction analysis. Biometrics 76(1):23–35, 2020.

Xiaofan Xu and Malay Ghosh. Bayesian variable selection and estimation for group lasso. Bayesian Analysis 10(4):909–936, 2015.

Jitao David Zhang and Stefan Wiemann. KEGGgraph: a graph approach to KEGG pathway in R and bioconductor. Bioinformatics 25(11):1470–1471, 2009.

Sen Zhao and Ali Shojaie. A significance test for graph-constrained estimation. Biometrics 72(2):484–493, 2016.

S. Zhe, S. A. Naqvi, Y. Yang, and Y. Qi. Joint network and node selection for pathway-based genomic data analysis. Bioinformatics 29(16):1987–1996, 2013.
Hui Zhou and Tian Zheng. Bayesian hierarchical graph-structured model for pathway analysis using gene expression data. Statistical Applications in Genetics & Molecular Biology 12(3):393–412, 2013.