Mixture Network Regularized Generalized Linear Model with Feature Selection

Kaiqiao Li[1], Xuefeng Wang[2], Pei Fen Kuan[1, *]

1 Stony Brook University
2 Moffitt Cancer Center

*** peifen.kuan@stonybrook.edu**

Abstract

High dimensional genomics data in biomedical sciences is an invaluable resource for constructing statistical prediction models. With the increasing knowledge of gene networks and pathways, this information can be utilized in the statistical models to improve prediction accuracy and enhance model interpretability. However, in some scenarios the network structure may only be partially known or inaccurately specified. Thus, the performance of statistical models incorporating such network structure may be compromised. In this paper, we proposed a weighted sparse network learning method by optimally combining a data driven network with sparsity property to a known or partially known prior network to address this issue. We showed that our proposed model attained the oracle property which aims to improve the accuracy of parameter estimation and achieved a parsimonious model in high dimensional setting for different outcomes including continuous, binary and survival data in extensive simulations studies. Case studies on ovarian cancer proteomics and melanoma gene expression further demonstrated that our proposed model achieved good operating characteristics in predicting response to chemotherapy and survival risk. An R package `glmaag` implemented our method is available on the Comprehensive R Archive Network (CRAN).

Introduction

The rapid advancement in high throughput genomics profiling has revolutionized biomedical research towards personalized medicine for treating and preventing various diseases including cancer. Several consortia have been established as part of the collaborative efforts to decipher the molecular mechanisms underlying these diseases, for example the Cancer Genome Atlas project have enabled researchers to access the rich cancer genomics database. Together with the rapid development of machine learning and artificial intelligence, these databases have been utilized extensively for improving computational and statistical model building and predictions.

One key attribute of these dataset is the high dimensionality, i.e., $p \gg n$ in which the number of candidate features/predictors is much larger than the sample size. For instance, in a typical DNA methylation data, several hundred-thousands of CpGs are interrogated. Regularization framework has emerged as an attractive alternative to address the limitations of classical feature selection method in generalized linear models (GLM) including computational efficiency and multicollinearity issues. For instance,

GLM regularization with l_1 penalty (least absolute shrinkage and selection operator (LASSO)) [28, 29] allows for simultaneous variable selection to prevent overfitting, whereas [37] showed that combining l_1 with l_2 penalty (elastic net (EN)) not only provides variable selection property but also robustness on correlated features (group property). [6] and [8] argued that a good feature selection procedure should have the oracle property which includes feature selection accuracy and asymptotically unbiased parameters estimation. Thus, [36] and [38] proposed adaptive LASSO and adaptive EN that have oracle property and can be optimized efficiently.

The above-mentioned methods have been shown to achieve positive performance in prediction models in which no prior knowledge is available. However, the abundance of genomics research has enable biological knowledge associated with the diseases to be inferred from gene regulatory networks and pathways. Several well-known databases of gene regulatory networks include the KEGG: Kyoto Encyclopedia of Genes and Genomes (`https://www.genome.jp/kegg/`) ([11]) and the Reactome Pathways (`https://reactome.org/`). If the network structure of the data is known in advance, one can potentially improve the model prediction and interpretability by incorporating the prior network information. One possible extension is to replace the l_2 penalty with a quadratic penalty that utilizes the unsigned or signed adaptive Laplacian matrix of the network structure ([14, 15]), which yields better performance in both prediction and variable selection. This framework has been applied to both the classification ([25]) and survival ([24]) outcomes. On the other hand, [33] adapted the l_1 penalty with unsigned network penalty to achieve the oracle property in Gaussian regression framework.

Although the above-mentioned public regulatory network databases are invaluable prior knowledge, one limitation is that most of the known networks only show the connectivity but without information on the strengths of the connectivity . The strengths of the connectivity are important factors which may influence the group property generated by the prediction model. Conversely, one may also encounter a dataset that only has unknown or partially known network structure. In this scenario, one can still apply the graph based method by estimating the network empirically from the data, e.g., the neighborhood selection method ([19]) to learn the connectivity among the candidate features and use the reliability score provided by reference gene association (RGA) ([31] as the strengths of connectivity.

Another challenge in regularization framework is to correctly tune the penalty parameters. A common approach is via the cross validation, which is straightforward to be applied in regularization framework. However, the cross validation approach has been shown to have the tendency to overfit the data when the number of features are relatively large compared to the sample size ([32]). An alternative approach is via the stability selection method developed based on the consistency of variable selection across multiple subsamples ([20]), and this method has been shown to perform well in graph-based models ([17]).

In this paper, we addressed the above-mentioned limitations of existing network/graph-based prediction models by proposing a mixture network prediction framework that combines two candidate networks (usually one being the fixed network obtained from gene regulatory network database while the other one is estimated from the data). To this end, we adapted the l_1 penalty in order to achieve the oracle property. In addition, to attain a robust variable selection accuracy, we implemented the stability selection tuning method for parameters tuning and compared this approach to the cross validation method. We developed our proposed framework for various outcomes including continuous, binary and survival data.

This paper is organized as follows. The description of our proposed method and the corresponding model fitting algorithm are provided in Section 2. The Monte Carlo simulations and case study are provided in Section 3 and 4, respectively. We conclude

with a discussion in Section 5.

Methodology

Network Regularized Regression

We start our exposition by reviewing the method associated with the (partial) log likelihood $l(\beta)$ of generalized linear model (GLM) for continuous, binary and survival outcomes:

$$l(\beta) \propto \begin{cases} -\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - x_i^T\beta\right)^2 & \text{Gaussian} \\ \sum_{i=1}^{n}\left[y_i\left(\beta_0 + x_i^T\beta\right) - \ln\left(1+\exp\left(\beta_0 + x_i^T\beta\right)\right)\right] & \text{Logistic} \\ \sum_{i=1}^{n}\delta_i\left[x_i^T\beta - \ln\left(\sum_{j\in R_i}\exp\left(x_j^T\beta\right)\right)\right] & \text{Cox} \end{cases}$$

where $Y = (y_1, \cdots, y_n)$ is the outcome vector, $X = \left(x_1^T, \cdots, x_p^T\right)$ is the predictors matrix, δ_i is the event indicator for right censored variable, and $R_i = \{j | t_j \geq t_i\}$ is the risk set of subject i. None of these GLM models can be optimized in high dimensional ($p \gg n$) case. One approach to circumvent this challenge is to solve the maximum penalized (partial) likelihood estimator (MPLE). We proposed a network LASSO with l_1 adaptive weights (abbreviated as AAG), in which the MPLE in primal form is given as below:

$$\max\left\{l(\beta)\right\}, \text{ subject to } s_1\sum_{i=1}^{p} w_i|\beta_i| + s_2|\beta|^T L|\beta| \leq t \tag{1}$$

where $w \succeq 0$ is the weight vector for l_1 penalty, L is the normalized Laplacian matrix, and $s_1 \geq 0$, $s_2 \geq 0$ and $t > 0$ are tuning parameters. To estimate the sign adapter for the network penalty we can fit the GLM model without penalty or ridge GLM model with l_2 penalty (denoted $\tilde{\beta}_a$), and use the estimated signs of $\tilde{\beta}_a$ as the sign estimate ([24]). Therefore we have $|\beta|^T L|\beta| \approx \beta^T \text{diag}\left\{\text{sgn}\left(\tilde{\beta}_a\right)\right\} L \text{diag}\left\{\text{sgn}\left(\tilde{\beta}_a\right)\right\}\beta \triangleq \beta^T\hat{L}\beta$ and use $\hat{L}$ as the signed network to be used. Next, to estimate the weight vector w we can estimate the coefficients with 1 where $s_1 = 0$,i.e., no l_1 penalty (denoted $\tilde{\beta}_b$). The l_1 adapted weights w can be estimated by $\hat{w} = \frac{1}{\tilde{\beta}_b}$ ([36] and [38]). The normalized Laplacian matrix L is given by

$$L_{ij} = \begin{cases} 1 & i = j \\ -\omega_{ij}/\sqrt{\xi_i\xi_j} & i \neq j,\ (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

where E is the connectivity set, ξ is the degree of the node and ω is the strength (can be either positive or negative) of the connectivity which can be estimated from the reference gene association network utilizing the reliability score of Pearson correlation ([31]). The reliability score of feature i and j denoted as R_{ij} is given by $R_{ij} = \frac{1}{r_{ij}r_{ji}}$ where r_{ij} is the ranking of correlation between feature i and j among all the correlations of feature i to others.

We require L to be positive definite if X^TX is not invertible. If L is an identity matrix, it reduces to adaptive elastic network model. This indicates that adaptive elastic net model ([38]) is a special case of our AAG model when there is no connection in the network (i.e., independent structure). To solve equation 1, we consider optimizing the objective function

$$\max\left\{\frac{1}{n}l(\beta) - \lambda_1\sum_{i=1}^{p} w_i|\beta_i| - \frac{\lambda_2}{2}|\beta|^T L|\beta|\right\} \tag{2}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$.

To solve equation 2 when $\lambda_1 > 0$, we implemented the proximal Newton based coordinate ascent algorithm derived by [9] and [22] with adaptive l_1 penalty. For Gaussian regression we have the coordinate-wise update given by

$$\tilde{\beta}_j^{(t+1)} \leftarrow \frac{S\left(n^{-1}\sum x_{ij}\left(y_i - \tilde{\beta}_0 - \sum_{k\neq j} x_{ik}\tilde{\beta}_k^{(t)}\right) - \lambda_2 \sum_{k\neq j} \hat{L}_{jk}\tilde{\beta}_k^{(t)},\ \lambda_1 w_j\right)}{n^{-1}\sum_{i=1}^n x_{ij} + \lambda_2 \hat{L}_{jj}}$$

where $S(a,b) = \text{sign}(a)(|a|-b)_+$ is the soft-thresholding operator ([5]).

For logistic and Cox regression, we require a quadratic approximation of the (partial) log likelihood using secondary Taylor expansion given by

$$\tilde{l}(\beta) = \frac{1}{2}(z - \beta_0\mathbf{1} - X\beta)^T l_\eta^{''}(z - \beta_0\mathbf{1} - X\beta) + C\left(\eta, \tilde{\beta}\right)$$

where $z = \eta - \left(l_\eta^{''}\right)^{-1} l^{'}$, $\eta = \tilde{\beta}_0\mathbf{1} + X\tilde{\beta}$, $C\left(\eta, \tilde{\beta}\right)$ is a term which does not depend on β, $\tilde{\beta}$ is the working update of β, and $\beta_0 = 0$ for Cox model. For Cox model we only need to calculate the diagonal entries of $l_\eta^{''}$ and fix all the off-diagonal entries to be zeros to speed up the computation, based on the argument provided by [22] where the off-diagonal entries of $l_\eta^{''}$ are small compared to the diagonal entries. For logistic model $l_\eta^{''}$ is already in a diagonal matrix form. Therefore, let u be the diagonal elements of $l_\eta^{''}$, we have

$$l_Q(\beta) = -\frac{1}{2}\sum_{i=1}^n u_i\left(z_i - \beta_0 - x_i^T\beta\right)^2$$

Note that the Gaussian model is a special case in which $u_i = 1$ and $z_i = y_i$. The coordinate-wise update step for logistic model is given by

$$\tilde{\beta}_j^{(t+1)} \leftarrow \frac{S(A-B,\ \lambda_1 w_j)}{n^{-1}\sum_{i=1}^n u_i^{(t)} x_{ij}^2 + \lambda_2 \hat{L}_{jj}}. \tag{3}$$

$$A = n^{-1}\sum_{i=1}^n u_i^{(t)} x_{ij}\left(z_i^{(t)} - \tilde{\beta}_0^{(t)} - \sum_{k\neq j} x_{ik}\tilde{\beta}_k^{(t)}\right)$$

$$B = \lambda_2 \sum_{k\neq j} \hat{L}_{jk}\tilde{\beta}_k^{(t)}$$

The working update for logistic model is given by

$$z_i^{(t)} = \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i - \mu_i^{(t)}}{\mu_i^{(t)}\left(1-\mu_i^{(t)}\right)}$$

$$u_i^{(t)} = \mu_i^{(t)}\left(1-\mu_i^{(t)}\right)$$

$$\mu_i^{(t)} = \frac{1}{1+\exp\left(-X\tilde{\beta}^{(t)}\right)}.$$

For Cox model, we used Breslow's ([2]) method to handle tied survival time. The working update is given by

$$z_i^{(t)} = \eta_i^{(t)} + \frac{1}{u_i^{(t)}}\left[\delta_i - \sum_{j\in C_i} \frac{d_i e^{\eta_i^{(t)}}}{\sum_{k\in R_j} e^{\eta_k^{(t)}}}\right]$$

$$u_i^{(t)} = \sum_{j \in C_i} d_i \frac{e^{\eta_i^{(t)}} \left[\sum_{k \in R_j} e^{\eta_k^{(t)}} - e^{\eta_i^{(t)}}\right]}{\left[\sum_{k \in R_j} e^{\eta_k^{(t)}}\right]^2}$$

$$\eta^{(t)} = X\tilde{\beta}^{(t)}$$

where R_j is the set of k for the jth sample with $t_k \geq t_j$, $C_i = \{j | t_j \leq t_i\}$ is the set of indices j for the ith sample with $t_j \leq t_i$ and d_i is the number of tied samples in survival time for the ith sample.

Mixture Network Tuning

In real data analysis, obtaining the correctly specified complete network structure could be infeasible for model fitting. In addition, in scenarios where the network structure is known, the strengths/weight of the connectivity might not be available. To circumvent these issues, we proposed a mixture network method that combines a pre-specified network L_1 and a data driven network L_2 in the following penalized likelihood framework:

$$\max \left\{ \frac{1}{n} l(\beta) - \lambda_1 \sum_{i=1}^{p} w_i |\beta_i| - \frac{\lambda_2}{2} |\beta|^T (cL_1 + (1-c) L_2) |\beta| \right\}$$

where $0 \leq c \leq 1$ is the network weight. If L_1 and L_2 are both positive definite, the final mixture network $L = cL_1 + (1-c) L_2$ is also positive definite, thus the consistency property still holds. To obtain the network weight c we recommend fixing $\lambda_1 = 0$ when tuning the weight between networks and only search for the combination of $\lambda_2 \times c$ for computational efficiency. As suggested by [4] we searched λ_2 over $\left\{0.01 \cdot 2^0, 0.01 \cdot 2^1, \cdots, 0.01 \cdot 2^7\right\}$. To tune the parameter c we recommend searching over the set $\{0, 0.1, 0.2, \cdots, 1\}$. We tuned the two networks via cross validation method and chose the value of c that optimized the cross validation performance. Upon identifying the optimal c we fixed the final mixed network with $\hat{L} = cL_1 + (1-c) L_2$ when tuning λ_1 and λ_2.

To estimate a data-driven network L_2, we obtained the connectivity using the R package `huge` ([35]) with penalized neighborhood selection method ([19]) tuned by rotation information criterion (RIC). However, this method does not provide the strength of connectivity. Therefore, we estimated the strengths/weights using the reliability score provided by the reference gene association network ([31]).

Parameter Tuning

We compared two frameworks for tuning λ_1 and λ_2. The first is the cross validation (CV) framework, where we performed the CV via deviance ($\hat{l}(\text{full}) - \hat{l}(\text{train})$) or robust measure including negative mean absolute error (MAE) for Gaussian model, area under the receiver operating characteristic curve (AUC) for logistic model and concordance index (C) for Cox model. For Gaussian model, the deviance measure is equivalent to negative mean squared error (MSE). One can either use the maximum (max) rule, i.e., obtaining (λ_1, λ_2) that maximizes the CV measure or the one standard error (1se) rule, i.e., obtaining (λ_1, λ_2) that results in the most parsimonious model within one standard error of the CV measures. We also imposed a $p/2$ constraint to the number of variables to improve computational speed.

Although CV is a convenient framework and has been shown to achieve good performance in low dimensional data, it may result in overfitting in high dimensional case ([32]). An alternative approach is via the stability selection (SS) proposed by [20]

which measures the feature selection stability across subsampling replicates, and has been shown to be robust in graphical models ([17]).

Suppose we randomly draw K samples (usually $K = 100$) with $\lfloor n/2 \rfloor$ or $\lfloor 10\sqrt{n} \rfloor$ observations depending on the sample size as suggested by [20] and [17], the selection probability of feature j is given by

$$\hat{P}^{\lambda_1,\lambda_2}(j) = \frac{1}{K}\sum_{k=1}^{K} I\left(\hat{\beta}_j^{\lambda_1,\lambda_2}(S_k)\right)$$

where S_k is the kth subsample. The selection variance is given by

$$\hat{\text{var}}^{\lambda_1,\lambda_2}(j) = \hat{P}^{\lambda_1,\lambda_2}(j)\left[1 - \hat{P}^{\lambda_1,\lambda_2}(j)\right].$$

A stable method should have a low selection variance, thus the instability score across all features is defined as

$$\hat{IS}(\lambda_1,\lambda_2) = \frac{2}{p}\sum_{j=1}^{p} \hat{\text{var}}^{\lambda_1,\lambda_2}(j).$$

To make the score comparable across different λ_2's, we consider a monotone transformation of the instability score given by

$$\bar{IS}(\lambda_1,\lambda_2) = \sup_{x \geq \lambda_1} \hat{IS}(x,\lambda_2)$$

such that the instability path decreases with increasing λ_1 for each fixed λ_2. By combining the instability score together, we find the maximum score that is lower than a specific cutoff, usually 0.15 and use the corresponding (λ_1, λ_2) as the selected tuning parameter.

Tuning λ_1 and λ_2 usually works iteratively by searching λ_1 for each fixed λ_2 until all possible values of $\lambda_1 \times \lambda_2$ have been considered. According to the strong rules for discarding predictors ([30]), it is not necessarily to consider all predictors for every λ_2. In particular, we can discard predictors that are not likely to be retained in the model under the Karush-Kuhn-Tucker (KKT) condition. We applied the strong rules in our model to improve computational speed.

Theoretical Properties

Group Effect

We showed how the network penalty adjusts for the multicollinearity issues by proving the group effect. Without loss of generality, we assumed that the response vector y for Gaussian models and predictor matrix X have been standardized. We assume that feature i and j are linked and only linked to each other and that the sign of estimation is correct. Assume further that the sample correlation of X_i and X_j is ρ_{ij}, the sign is consistent with the coefficient, the l_1 penalty weight for feature i is w_i and the strength of connectivity for feature i and j is ω_{ij} and $0 \leq \omega_{ij} \leq 1$. We have

$$\left|\left|\hat{\beta}_i\right| - \left|\hat{\beta}_j\right|\right| \leq \frac{\sqrt{2\left(1 - |\rho_{ij}|\right)} + \lambda_1 |w_i - w_j|}{\lambda_2\left(1 + \omega_{ij}\right)}.$$

Oracle Property

We provided the theoretical proof for the oracle property on our proposed method to ensure that the model is robust with respect to variable selection and coefficient estimation. Gaussian and logistic models are in the exponential family whereas the Cox proportional hazard model is not, thus the oracle property for Cox model is different from Gaussian and logistic models. We proved the oracle property for Cox model and exponential family GLM (not limited to Gaussian and logistic models) separately in the following subsections.

0.0.1 Generalized Linear Model in Exponential Family

For GLM in exponential family, e.g., Gaussian and logistic models, the likelihood function can be written as $l\left(Y|X,\theta\right)=h\left(Y\right)\exp\left(Y^{T}\theta-\phi\left(\theta\right)\right)$ where $\theta=X\beta$ and β is the true coefficient vector. We denote the maximum penalized likelihood estimation $\hat{\beta}$ as

$$\hat{\beta}^{(n)}=\text{argmax}_{\beta}\left\{\frac{1}{n}\left[Y^{T}X\beta-\phi\left(X\beta\right)\right]-\lambda_{1}^{(n)}\sum_{i=1}^{p}\hat{w}_{i}\left|\beta_{i}\right|-\frac{\lambda_{2}^{(n)}}{2}\beta^{T}L\beta\right\},$$

where $\hat{w}=\frac{1}{|\tilde{\beta}|}$ where $\tilde{\beta}$ is a root-n-consistent estimator of β such as the OLS estimator and $r>0$. Let $A_{n}^{*}=\left\{i|\hat{\beta}_{i}^{(n)}\neq0\right\}$ and A denote the selected features and true predictor set, respectively. In our case

$$\phi\left(X\beta\right)=\begin{cases}\frac{1}{2}\beta^{T}X^{T}X\beta & \text{Gaussian}\\ \mathbf{1}^{T}\ln\left(\mathbf{1}+\exp\left(X\beta\right)\right) & \text{Logistic}\end{cases}$$

Suppose that $\sqrt{n}\lambda_{1}^{(n)}\rightarrow0$, $n\lambda_{1}^{(n)}\rightarrow+\infty$, $\sqrt{n}\lambda_{2}^{(n)}\rightarrow0$ and $\Lambda_{\max}\left(L\right)\leq\lambda_{L}<+\infty$, where $\Lambda_{\max}\left(\cdot\right)$ represents the largest eigenvalue of a given matrix. Given the two regularity conditions:

1. Fisher information matrix $I\left(\beta\right)=E\left[\phi^{''}\left(x\beta\right)X^{T}X\right]$ and $n\lambda_{2}^{(n)}L$ are finite and positive definite.
2. There exists a sufficiently large open set O where $\beta\in O$ and $\forall B\in O$ we have $\left|\phi^{'''}\left(XB\right)\right|\leq M\left(X\right)<+\infty$ and $E\left[M\left(X\right)\left|x_{i}x_{j}x_{k}\right.\right]<+\infty$ for any $1\leq i,j,k\leq p$,

we have the following two properties:

1. Variable selection consistency: $\lim_{n}P\left(A_{n}^{*}=A\right)=1$.
2. Asymptotic normality: $\sqrt{n}\left(\hat{\beta}_{A}^{(n)}-\beta_{A}\right)\xrightarrow{d}N\left(\mathbf{0},I_{A}^{-1}\right)$.

0.0.2 Cox's Proportional Hazards Model

The Cox model is not within exponential family, thus the proof of the oracle property requires some modifications as shown in [7]. Define the at-risk and counting process as $N_{i}\left(t\right)=\delta_{i}I\left(\tilde{T}_{i}\leq t\right)$ and $Y_{i}\left(t\right)=I\left(\tilde{T}_{i}\geq t\right)$ where $\tilde{T}_{i}=\min\left(T_{i},C_{i}\right)$ (T_{i} is the failure time and C_{i} is the censoring time for the ith subject), and the Fisher information matrix as

$$I\left(\beta\right)=\int_{0}^{1}v\left(\beta,t\right)s^{(0)}\left(\beta,t\right)h_{0}\left(t\right)\mathrm{d}t$$

where

$$v(\beta,t)=\frac{s^{(2)}(\beta,t)}{s^{(0)}(\beta,t)}-\left(\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right)\left(\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right)^{T},$$

$s^{(k)}(\beta,t)=E\left[x(t)^{\otimes k}Y(t)\exp\left(x(t)^{T}\beta\right)\right]$, $k=0,1,2$ and $h_0(t)$ is the baseline hazards function. Here we assume all the regularity conditions (A-D) in [**?**] hold. We assume that A is the true predictor set and A^C the complement set. Given that $\sqrt{n}\lambda_1^{(n)}\to 0$, $n\lambda_1^{(n)}\to+\infty$, $\sqrt{n}\lambda_2^{(n)}\to 0$ and $\Lambda_{\max}(L)\leq M<+\infty$, the root-$n$ consistent estimator $\hat{\beta}^{(n)}$ satisfies the following conditions:

1. Sparsity: $\hat{\beta}_{A^C}^{(n)}=\mathbf{0}$
2. Asymptotic normality: $\sqrt{n}\left(\hat{\beta}_A^{(n)}-\beta_A\right)\xrightarrow{d}N\left(\mathbf{0},I_A^{-1}\right)$.

Monte Carlo Simulations

We conducted a Monte Carlo study to evaluate the performance of our proposed model. We considered two network structures namely (1) the autoregressive (AR) structure where each feature is connected and only connected to its neighbor, and (2) the HUB structure where the features formed groups with one dominant feature within each group.

In our simulation, we generated $p=200$ features and $n=500$ samples in which 100 samples were used as training data and the remaining 400 samples were set aside as test data. The features were generated from a multivariate Gaussian distribution with mean zero and diagonal covariance one. We assigned three twenty-feature groups with absolute coefficients 0.5, 1 and 2 and random signs for the noninformative features (i.e., those with zero coefficients).

$$\beta=\left(\underbrace{\pm 2,\cdots,\pm 2}_{20},\underbrace{\pm 1,\cdots,\pm 1}_{20},\underbrace{\pm 0.5,\cdots,\pm 0.5}_{20},\underbrace{0,\cdots,0}_{140}\right)^{T}$$

For Gaussian models, we generated Gaussian noise with mean zero and standard error $\|\beta\|_2/2$. For logistic models, we generated the outcome variable from the Bernoulli distribution with probability of the success as the logistic score of the predictors. For Cox models, we generated Weibull baseline hazards with shape parameter 5, scale parameter 2 and censoring time following a uniform distribution $U(2,15)$ which leads to a censoring rate of approximately 30%.

Cross Validation with $p/2$ Constraints

We compared our proposed model to the elastic net model (implemented in the R package `glmnet`) and network-LASSO regression without the l_1 adaptive weights (implemented in the R package `glmgraph`). Since `glmgraph` is not implemented for Cox model, we wrote our own codes for fitting the Cox models in network-LASSO regression without the l_1 adaptive weights. To assess the effect of network misspecification, we considered the scenarios where we used (1) the correct network (cor), (2) the incorrect network (AR misspecified as HUB and vice versa) (wr), and (3) the estimated network (est). The signs of network were estimated empirically. We compared this signed network model to our proposed mixture network model that combined (1) a correct network with an incorrect network, (2) a correct network with an estimated network, and (3) an incorrect network with an estimated network. The

results of the AR structure as the true network are shown in Tables 1. The tuning parameters were chosen via cross validation with one standard error rule and the number of parameters were constrained to be at most $p/2$.

Table 1. Model Comparison with AR Structure as True Network

	Gaussian			
Method	MAE	MSE	Pearson	Spearman
EN	7.283 (0.740)	84.340 (17.447)	0.839 (0.033)	0.824 (0.036)
Graph_cor	6.343 (0.849)	64.578 (22.617)	0.886 (0.049)	0.875 (0.049)
Graph_wr	7.006 (0.682)	77.896 (14.883)	0.859 (0.026)	0.845 (0.028)
Graph_est	6.828 (0.698)	74.217 (15.408)	0.864 (0.028)	0.851 (0.032)
AAG_cor	5.788 (0.543)	53.025 (9.909)	0.896 (0.023)	0.884 (0.026)
AAG_wr	6.685 (0.472)	70.723 (9.827)	0.859 (0.022)	0.846 (0.025)
AAG_est	6.372 (0.525)	64.188 (10.550)	0.873 (0.023)	0.861 (0.026)
MixAAG_corwr	5.770 (0.531)	52.756 (9.899)	0.896 (0.022)	0.886 (0.025)
MixAAG_corest	5.794 (0.522)	53.180 (9.497)	0.896 (0.021)	0.885 (0.025)
MixAAG_wrest	6.557 (0.516)	68.010 (10.817)	0.865 (0.024)	0.852 (0.028)

	Logistic			
Method	AUC	ACC	MCC	Biserial
EN	0.839 (0.059)	0.746 (0.049)	0.501 (0.097)	0.716 (0.114)
Graph_cor	0.906 (0.041)	0.812 (0.047)	0.630 (0.090)	0.867 (0.087)
Graph_wr	0.893 (0.031)	0.795 (0.037)	0.599 (0.069)	0.834 (0.063)
Graph_est	0.877 (0.040)	0.777 (0.041)	0.564 (0.077)	0.803 (0.082)
AAG_cor	0.931 (0.024)	0.833 (0.033)	0.676 (0.060)	0.939 (0.054)
AAG_wr	0.889 (0.028)	0.792 (0.031)	0.591 (0.059)	0.839 (0.063)
AAG_est	0.891 (0.031)	0.793 (0.033)	0.596 (0.061)	0.844 (0.071)
MixAAG_corwr	0.931 (0.027)	0.834 (0.035)	0.677 (0.066)	0.937 (0.063)
MixAAG_corest	0.930 (0.025)	0.831 (0.034)	0.672 (0.061)	0.936 (0.058)
MixAAG_wrest	0.892 (0.033)	0.791 (0.040)	0.594 (0.072)	0.846 (0.073)

	Cox
Method	C
EN	0.858 (0.031)
Graph_cor	0.911 (0.022)
Graph_wr	0.876 (0.027)
Graph_est	0.879 (0.027)
AAG_cor	0.931 (0.018)
AAG_wr	0.869 (0.022)
AAG_est	0.900 (0.018)
MixAAG_corwr	0.929 (0.019)
MixAAG_corest	0.929 (0.018)
MixAAG_wrest	0.900 (0.019)

In the results, EN represents elastic (`glmnet`) method, Graph represents network LASSO without l_1 adaptive weights, AAG represents our proposed network LASSO with l_1 adaptive weights and MixAAG represents our proposed network LASSO with l_1 adaptive weights and mixture network. For Gaussian model, we compared mean absolute error (MAE), mean squared error (MAE), Pearson and Spearman correlation. For logistic model, we compared the area under the receiver operating characteristic curve (AUC) calculated via [21] method, accuracy (ACC), Matthews correlation coefficient (MCC) and biserial correlation. For Cox model, we compared the concordance index (C). We reported the mean and standard deviations of these metrics across 100 replicates.

From the simulation results, our proposed method with l_1 adaptive weights yields better performance compared to elastic net and network-LASSO without l_1 adaptive weights. For both the AR and HUB structures, incorporating the correctly network yields significantly better results compared to the case where network is misspecified as expected. On the other hand, the network mixture approach (i.e., mixing an incorrect network with an estimated data driven network) yields better performance compared to a model with a wrong network or elastic net model.

Cross validation vs stability selection

In practice we constrain the number of selected features to be no more than $p/2$ similar to the default method of R package `glmgraph` to speed up computation. However, this constraint may not be desirable if the true number of informative features is greater than $p/2$. An alternative approach is the stability selection (SS) method as described earlier. In this subsection we compared the variable selection accuracy between cross validation without $p/2$ constraint and the stability selection method. We reported the MCC of the estimate coefficients, and Sensitivity (Sn) for large, medium, and small effect sizes and Specificity (Sp) averaged over 100 replications. The results for the AR structure are shown in Table 2.

Table 2. Cross Validation vs Stability Selection with AR Structure on Variable Selection Accuracy

Gaussian					
Method	MCC	Sn_large	Sn_medium	Sn_small	Sp
CV_corwr	0.637 (0.075)	0.974 (0.045)	0.737 (0.150)	0.364 (0.183)	0.902 (0.138)
CV_corest	0.638 (0.082)	0.973 (0.040)	0.748 (0.136)	0.362 (0.161)	0.916 (0.051)
CV_wrest	0.604 (0.063)	0.926 (0.074)	0.587 (0.142)	0.228 (0.163)	0.936 (0.139)
SS_corwr	0.614 (0.066)	0.984 (0.034)	0.804 (0.126)	0.469 (0.186)	0.868 (0.042)
SS_corest	0.613 (0.074)	0.982 (0.038)	0.799 (0.136)	0.481 (0.181)	0.867 (0.041)
SS_wrest	0.597 (0.069)	0.974 (0.042)	0.735 (0.123)	0.359 (0.159)	0.893 (0.044)
Logistic					
Method	MCC	Sn_large	Sn_medium	Sn_small	Sp
CV_corwr	0.482 (0.102)	0.976 (0.056)	0.778 (0.212)	0.492 (0.280)	0.718 (0.216)
CV_corest	0.482 (0.101)	0.975 (0.048)	0.775 (0.206)	0.498 (0.281)	0.725 (0.200)
CV_wrest	0.421 (0.081)	0.844 (0.135)	0.443 (0.223)	0.258 (0.196)	0.855 (0.153)
SS_corwr	0.536 (0.086)	0.957 (0.054)	0.638 (0.143)	0.319 (0.162)	0.883 (0.030)
SS_corest	0.537 (0.083)	0.959 (0.057)	0.643 (0.127)	0.314 (0.156)	0.883 (0.029)
SS_wrest	0.432 (0.086)	0.895 (0.075)	0.496 (0.128)	0.277 (0.129)	0.860 (0.035)
Cox					
Method	MCC	Sn_large	Sn_medium	Sn_small	Sp
CV_corwr	0.705 (0.094)	0.979 (0.034)	0.848 (0.100)	0.539 (0.165)	0.910 (0.062)
CV_corest	0.709 (0.089)	0.981 (0.034)	0.847 (0.102)	0.536 (0.168)	0.914 (0.049)
CV_wrest	0.618 (0.078)	0.946 (0.054)	0.703 (0.134)	0.352 (0.143)	0.916 (0.057)
SS_corwr	0.656 (0.077)	0.976 (0.038)	0.862 (0.101)	0.590 (0.167)	0.866 (0.032)
SS_corest	0.659 (0.072)	0.979 (0.037)	0.858 (0.103)	0.588 (0.159)	0.868 (0.034)
SS_wrest	0.583 (0.073)	0.956 (0.045)	0.761 (0.114)	0.427 (0.128)	0.871 (0.027)

From Table 2, the logistic model fitted via cross validation without $p/2$ constraint yields lower MCC and Sp compared to the stability selection. For Gaussian and Cox model the cross validation and stability selection have similar performance. The cross validation approach is also more computationally efficient (e.g., for Gaussian model, with 100 samples and 20 features, five-fold cross validation took 0.05s while 100

Table 3. Log Platinum Free Interval in Proteomic Ovarian Cancer Prediction

Method	MAE	MSE	Pearson	Spearman	# features
CV_EN	0.947	1.439	0.461	0.438	1026
CV_MixAAG	0.897	1.344	0.464	0.473	146
SS_MixAAG	0.904	1.414	0.388	0.384	193

replicated stability selection took 4.80s).

Case Study

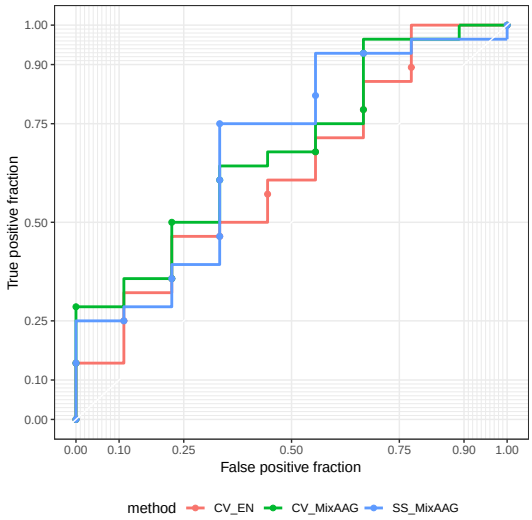
Platinum Therapy in Ovarian Cancer

We applied our proposed method to ovarian cancer proteomics dataset from the Cancer Genome Atlas (TCGA) generated by the Johns Hopkins University (JHU) and Pacific Northwest National Laboratory (PNNL) ([34]). The dataset was downloaded from the National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC) (`https://cptac-data-portal.georgetown.edu/cptacPublic/`). Ovarian cancer is one of the most lethal gynecologic malignancy which is difficult to be detected early. Most ovarian cancer cases are only detected in late stage and treated with chemotherapy using a platinum compound drug. Unfortunately, the platinum therapy is not effective for all patients as some patients develop resistance to the treatment. To improve ovarian cancer survival, it is important to predict whether a patient will response to the treatment. In this case study, we used the proteins as candidate features to predict two types of outcome measurements, namely the platinum free interval, a continuous measurement for the treatment sensitivity and platinum status (sensitive versus resistant by dichotomizing platinum free interval, i.e., platinum free interval greater than 6 months was marked as sensitive). Our sample size consisted of 95 sensitive patients and 34 resistant patients. We used ComBat ([10] and [12]). Since the missing values in mass spectrometry proteomics data can be attributed to detection limit ([27]), we imputed the missing values by the minimum value of each protein divided by $\sqrt{2}$ ([23]). The set of features with more than 20% missing were removed from our study. The pre-processed and normalized dataset consisted of 6451 candidate features/proteins for our subsequent analysis. Among the 129 samples, we randomly assigned 92 (71.3%) samples to form the training set and the remaining 37 (28.7%) as test set. We pre-screened the features using feature-wise Gaussian and logistic regression on log platinum free interval and platinum status, respectively in training data and retained the candidate features with p-values ≤ 0.15. 1026 and 881 features were retained for log platinum free interval and platinum status, respectively. To obtain a prior network structure, we downloaded the protein-protein interaction network (protein links within human sapiens) from the STRING database ([26]) and combined with the Laplacian matrix estimated from the training data. The signs and strengths of connectivity of the network were estimated using the method described in Sections 2.1 and 2.2. For platinum status prediction, a cutoff value that maximized the Youden's index in training data was used to compute the accuracy (ACC), Matthew's correlation coefficient (MCC), Youden index (J), Sensitivity (Sn) and Specificity (Sp) in test data. The results are shown in Tables 3 and 4.

Both the predictions of log platinum free interval and platinum status showed that elastic net (EN) method tend to overfit the data since the model chose $\alpha = 0$ (ridge regression) as the optimal value, thus all the features were retained. On the other hand, our proposed mixture network method selected a smaller number of features and achieved better prediction performance. The results also indicated that the ovarian

Table 4. Platinum Status in Proteomic Ovarian Cancer Prediction

Method	AUC	ACC	MCC	J	Sn	Sp	Biserial	# features
CV_EN	0.623	0.622	0.111	0.123	0.679	0.444	0.279	881
CV_MixAAG	0.683	0.514	0.183	0.206	0.429	0.778	0.376	394
SS_MixAAG	0.690	0.486	0.083	0.095	0.429	0.667	0.399	171



The ROC curves for platinum status prediction in test data.

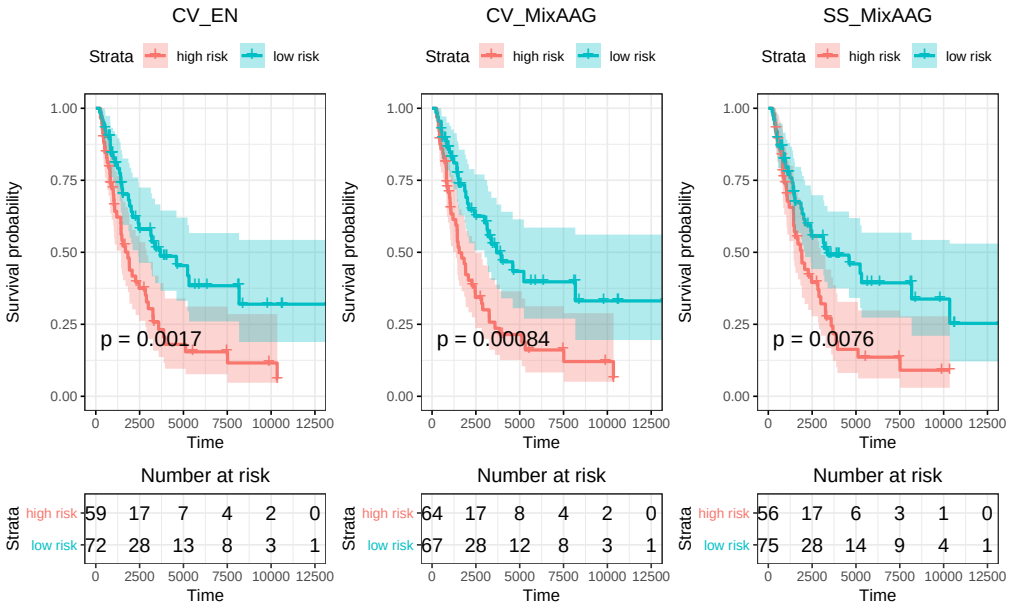
cancer proteomics dataset has an inherent network structure, thus out proposed method is suitable for modeling this type of data.

Survival Time in Skin Cutaneous Melanoma

Skin cutaneous melanoma (SKCM) is an aggressive malignancy that arises from uncontrolled melanocytic proliferation. Gene expression has been shown to be a promising biomarker for predicting survival in SKCM ([1], [3] and [18]). We applied our proposed method to the Cancer Genome Atlas (TCGA) SKCM gene expression data generated using the RNA-Seq platform. Gene expression values were summarized using RSEM ([13]) and normalized via the quantile normalization procedure ([16]). Our data consisted of the overall survival time of 436 patients with 217 events. We first pre-screened the candidate features, i.e., genes by individual Cox regression and retained 864 features with p-values ≤ 0.15. We randomly divided the data into training (305 patients) and test (131 patients) sets. The network structure was based the melanoma pathway from KEGG, combined with an estimated network in which the signs and strengths were estimated via the method described in Section 2.1 and 2.2. We compared the results of our proposed methods to elastic net models. The results are shown in Figure 1. We trained the models and obtained the optimal cutoff value for the log rank test. We used the selected cutoff in the test data to divide the patients into high and low risk groups, and evaluated the prediction via the Kaplan-Meier curves and log rank tests. We reported the concordance index (C) in test data and the number of features selected in the training data.

The results showed that our proposed methods with cross validation performed the best (best C index and lowest number of retained features). Our proposed method via stability selection had comparable performance to elastic net method via cross validation.

Figure 1. Survival Risk in Gene Expression SKCM Prediction



Method	C	# features
CV_EN	0.616	214
CV_MixAAG	0.626	79
SS_MixAAG	0.603	208

Kaplan Meier curves and log rank tests.

Conclusion

Incorporating network structure in the prediction model has been shown to be important in high dimensional genomics studies for accurate feature selection and model interpretability. In this paper we proposed a mixture network regularized generalized linear model which allows us to optimally combine a prior network and a data driven network. This is particularly useful in the scenarios in which the prior network is not known with certainty. Our model safeguards against incorporating an incorrect prior network by allowing an optimally mixed network structure in the model.

Our simulation studies showed that the proposed l_1 adaptive method yields higher prediction and feature selection accuracies across different scenarios. We also found that cross validation may not be the best approach for feature selection in high dimensional data, especially for binary classification. An alternative strategy is the stability selection method which was shown to yield better performance than cross validation in such scenarios, though it requires a much higher computational cost. Based on our simulation results, we suggested using the stability selection method for parameter tuning in binary classification problem, whereas cross validation is often sufficient for Gaussian and Cox outcomes.

An interesting future work includes replacing the l_1 penalty with a grouped LASSO penalty to allow for group-wise instead of feature-wise selection. However, the challenge would be to ensure that the group structure inferred from the group LASSO penalty is consistent with the group structure from the data driven network. One possibility is to define the grouped LASSO penalty after obtaining the network mixture within an iterative framework. Another future research is to apply the AAG method to other exponential family, for example, the Poisson and negative binomial regression for modeling count data outcomes. Our proposed model `glmaag` is available on the

Comprehensive R Archive Network (CRAN).

Acknowledgments

This work was supported in part by the CDC/NIOSH award U01 OH011478.

References

1. M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536, 2000.

2. N. E. Breslow. Discussion of professor cox's paper. *J Royal Stat Soc B*, 34:216–217, 1972.

3. K. M. Carr, M. Bittner, and J. M. Trent. Gene-expression profiling in human cutaneous melanoma. *Oncogene*, 22(20):3076, 2003.

4. L. Chen, H. Liu, J.-P. A. Kocher, H. Li, and J. Chen. glmgraph: an r package for variable selection and predictive modeling of structured genomic data. *Bioinformatics*, 31(24):3991–3993, 2015.

5. D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

6. J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

7. J. Fan, R. Li, et al. Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99, 2002.

8. J. Fan, H. Peng, et al. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

9. J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

10. W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

11. M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

12. J. T. Leek, W. E. Johnson, H. S. Parker, E. J. Fertig, A. E. Jaffe, J. D. Storey, Y. Zhang, and L. C. Torres. *sva: Surrogate Variable Analysis*, 2018. R package version 3.28.0.

13. B. Li and C. N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

14. C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

15. 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, 2010.

16. P. Li, Y. Piao, H. S. Shon, and K. H. Ryu. Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data. *BMC bioinformatics*, 16(1):347, 2015.

17. H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems*, pages 1432–1440, 2010.

18. S. Mandruzzato, A. Callegaro, G. Turcatel, S. Francescato, M. C. Montesco, V. Chiarion-Sileni, S. Mocellin, C. R. Rossi, S. Bicciato, E. Wang, et al. A gene expression signature associated with survival in metastatic melanoma. *Journal of translational medicine*, 4(1):50, 2006.

19. N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

20. N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

21. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):77, 2011.

22. N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.

23. P. A. Succop, S. Clark, M. Chen, and W. Galke. Imputation of data values that are less than a detection limit. *Journal of occupational and environmental hygiene*, 1(7):436–441, 2004.

24. H. Sun, W. Lin, R. Feng, and H. Li. Network-regularized high-dimensional cox regression for analysis of genomic data. *Statistica Sinica*, 24(3):1433, 2014.

25. H. Sun and S. Wang. Penalized logistic regression for high-dimensional dna methylation data with case-control studies. *Bioinformatics*, 28(10):1368–1375, 2012.

26. D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2014.

27. S. L. Taylor, G. S. Leiserowitz, and K. Kim. Accounting for undetected compounds in statistical analyses of mass spectrometry 'omic studies. *Statistical applications in genetics and molecular biology*, 12(6):703–722, 2013.

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

29. R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

30. R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

31. D. Ucar, I. Neuhaus, P. Ross-MacDonald, C. Tilford, S. Parthasarathy, N. Siemers, and R.-R. Ji. Construction of a reference gene association network from multiple profiling data: application to data analysis. *Bioinformatics*, 23(20):2716–2724, 2007.

32. L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

33. H. Yang and D. Yi. Studies of the adaptive network-constrained linear regression and its application. *Computational Statistics & Data Analysis*, 92:40–52, 2015.

34. H. Zhang, T. Liu, Z. Zhang, S. H. Payne, B. Zhang, J. E. McDermott, J.-Y. Zhou, V. A. Petyuk, L. Chen, D. Ray, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*, 166(3):755–765, 2016.

35. T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research*, 13(Apr):1059–1062, 2012.

36. H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

37. H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

38. H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733, 2009.