The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1250–1270
https://doi.org/10.1214/17-AOAS1097

NETWORK-BASED FEATURE SCREENING WITH APPLICATIONS TO GENOME DATA[1]

BY MENGYUN WU*, LIPING ZHU† AND XINGDONG FENG*

*Shanghai University of Finance and Economics** *and*
Renmin University of China†

Modern biological techniques have led to various types of data, which are often used to identify important biomarkers for certain diseases with appropriate statistical methods, such as feature screening. Model-free feature screening has been extensively studied in the literature, and it is effective to select useful predictors for ultra-high dimensional data. These existing screening procedures are conducted based on certain marginal correlations between predictors and a response variable, therefore network structures connecting the predictors are usually ignored. Google's PageRank algorithm has achieved remarkable success. We adopt its spirit to adjust original screening approaches by incorporating the network information. We can then significantly improve the performance of those screening methods in choosing useful biomarkers, which is demonstrated in an intensive simulation study. A couple of real genome datasets along with a biological network are further analyzed by comparing results on both accuracy of predicting responses and stability of identifying biomarkers.

1. Introduction. The explosion of high-throughput profiling technologies has generated an unprecedented amount of high-dimensional data for bioinformatics and biomedical research, such as gene expression data, single nucleotide polymorphisms (SNPs) and DNA methylation. These molecular characterizations have led to the proposal of diagnostic and prognostic biomarkers for varieties of cancer types, which play an important role in revealing the mechanisms of disease pathogenesis, selection of new therapeutic approaches and the prediction of later clinical benefit [Martinezledesma, Verhaak and Trevino (2015), Fan, Han and Liu (2014)].

The procedure of identifying these disease-related biomarkers can be regarded as a statistical problem of variable selection or feature screening. The variable selection methods have already detected some effective biomarkers to predict diseases or personalize treatments in the clinic, such as the LASSO [Tibshirani

Received January 2017; revised May 2017.

[1]Supported in part by National Youth Top-notch Talent Support Program, National Natural Science Foundation of China (11571218, 61402276, 11371236, 11422107 and 11690012), the State Key Program in the Major Research Plan of National Science Foundation of China (91546202) and Program for Innovative Research Team of SUFE.

Key words and phrases. Correlation, feature screening, model-free, network, ultra-high dimension, variable selection.

(1996)], the elastic net [Zou and Hastie (2005)] and the smoothly clipped absolute deviation (SCAD) [Fan and Li (2001)]. Recently, network information is taken into account in the penalties under the linear regression [Li and Li (2008), Pan, Xie and Shen (2010), Yu and Liu (2016)].

The model-free feature screening approaches are popular in identifying useful biomarkers for ultrahigh-dimensional data, where less evidence is present for identifying models and the computational expediency becomes challenging for current variable selection approaches in practice. These screening methods are mainly constructed based on the marginal correlation measures and often possess the sure screening property [Fan and Lv (2008)], where all truly important predictors can be selected with probability tending to one as the sample size diverges to infinity. These model-free methods include sure independence screening (SIS) [Fan and Lv (2008)], sure independent ranking and screening (SIRS) [Zhu et al. (2011)], and feature screening via distance correlation (DC-SIS) [Li, Zhong and Zhu (2012)], which often focus on single markers that can discriminate patients with different clinical characteristics, such as disease status.

Despite the success of these methods in prediction, the typical low reproducibility of these selected signatures combined with the difficulty to achieve a clear biological interpretation remains obvious obstacle for the application in clinical diagnosis [Cun and Fröhlich (2012)]. The reproducibility here refers to that the identified variable subset should perform consistently across similar studies in different labs or at different time, which is just one sufficient property of real biomarkers [He and Yu (2010)]. One good indicator of biomarker reproducibility is the stability of feature screening results with respect to sampling variations [He and Yu (2010)]. Good stability of feature selection is equally as important as good prediction performance in biomarker discovery and has gained increasing attention recently [Chuang et al. (2006), Shi, Yi and Ma (2015), Hawrylycz et al. (2015)]. Due to similar discriminatory power of some different biomarkers and unstable algorithms across samples, it is common that the biomarkers of the same disease identified by traditional methods in different studies hardly have overlaps.

Diseases' complexity brings another issue on feature screening methods. Diseases such as asthma, diabetes or obesity, often involve altered interactions between thousands of genes who may act together in various signaling and regulatory pathways and protein complexes [Leiserson et al. (2015), Gustafsson et al. (2014)]. Therefore, the impact of a specific genetic abnormality can spread along these interactions and alter the activity of the connected gene products [Barabási, Gulbahce and Loscalzo (2011)]. The structures of biological networks constructed based on these interactions can be traced back for a better understanding of biological processes and further used to identify human disease-associated genes [Vidal, Cusick and Barabasi (2011)]. For example, the disease-genes have been shown to be often characterized by large degrees in biological networks, and more likely to interact with other disease-genes [Taylor et al. (2009)]. Previous feature screening

methods may overlook these low discriminative biomarkers who are interacting with some differentially expressed genes, or those who play a critical role in the molecular mechanism of a complex biological network by interconnecting many other genes.

Motivated by the challenges posed by the instability and complex interactions in high-dimensional genome data, we propose a network-based feature screening method by adopting the spirit of the Google's PageRank algorithm and the Markov chain theory, both can be integrated with those commonly used model-free feature screening approaches. The network information is embedded in the feature screening by constructing a variant of "Google's matrix" [Langville and Meyer (2012)], which makes the proposed method less sensitive to the training samples. We further study its theoretical properties based on the Markov chain theories. Our numerical studies show that the integration of high-dimensional data and network structures can improve both performance of prediction and reproducibility of feature screening. The proposed procedure can detect various types of important predictors, including those covariates with weak marginal effects but active in interacting with other predictors. The experimental results on two publicly available cancer datasets attest to the competitive predictive performance with the state-of-art algorithms and the robustness against the inclusion or exclusion of some patients. More detailed analysis of one pancreatic adenocarcinoma dataset indicates that many biomarkers identified by our method have been verified to be related with pancreatic adenocarcinoma in biochemical or biomedical research, including some genes with weak marginal effects who actually play a vital role in the biological network. The KEGG pathway analysis further manifests that the proposed screening procedure tends to identify sets of biomarkers with the significant biological and functional correlations.

2. Method.

2.1. *Network-based feature screening.* Suppose we have p predictors to be taken into account. Let $\mathbf{c} = (c_1, \ldots, c_p)^T$ be the ranking measurement in traditional feature screening, such as the absolute value of Pearson correlation [Fan and Lv (2008)] and distance correlation [Li, Zhong and Zhu (2012)]. An adjacency matrix $\mathbf{A} = (A_{jl})_{p\times p}$ is used to represent a network structure, where $A_{jl} = 1$ if there exists an edge between nodes j and l in the network and $A_{jl} = 0$ otherwise, and all diagonal elements of $\mathbf{A}$ are set to be one to account for self-loops. For an undirected graph, the adjacency matrix $\mathbf{A}$ is symmetric. The degree of node j is defined as $D_j = \sum_{l=1}^{p} A_{jl}$.

We first construct the following matrix that includes both the original ranking measures and the network information:

$$\mathbf{S} = (1-d)\mathbf{1}_p \mathbf{c}^{*T} + d\mathbf{A}^*, \tag{2.1}$$

where $\mathbf{1}_p$ is the p-dimensional column vector of ones, $\mathbf{c}^* = \mathbf{c}/\mathbf{1}_p^T\mathbf{c}$ is the normalized vector of marginal ranking measures, $\mathbf{A}^*$ is the normalized counterpart of the adjacency matrix $\mathbf{A}$ with the jlth element $A^*_{jl} = A_{jl}/D_j$ and d is a weight lying between zero and one which balances the original ranking measures and the network information. Since each row sum of both matrices $\mathbf{1}_p\mathbf{c}^{*T}$ and $\mathbf{A}^*$ equals to one, it ensures that $\mathbf{S}$ is also a stochastic matrix.

For a variable connected with many unimportant variables, the spurious correlations due to noise may accumulate through the network. Thus, some variables may be incorrectly chosen as active with the proposed method. Hereby, we use an adaptive ranking measure $\mathbf{c}^*(r)$ in constructing the matrix $\mathbf{S}$ given in (2.1), where r refers to the size of the variable set in which we believe all active predictors have been included. To be specific, the original ranking values of top r predictors are kept intact, and others are set to be a very small number much less than the smallest marginal ranking measure. We denote the vector of ranking measures as $\mathbf{c}(r)$. Finally, the vector of the normalized ranking measures $\mathbf{c}^*(r) = \mathbf{c}(r)/\mathbf{1}_p^T\mathbf{c}(r)$ is used to construct the adaptive matrix $\mathbf{S}(r)$ similarly as given in (2.1). We will ignore the parameter r in the following context for simple notation if no confusion exists.

According to the Markov chain theory, as long as the matrix $\mathbf{S}$ is stochastic, irreducible and aperiodic, there will be a unique vector of positive numbers $\mathbf{v} = (v_1, \ldots, v_p)^T$ that satisfies the following stationary equation:

$$\mathbf{v}^T = \mathbf{v}^T\mathbf{S}, \tag{2.2}$$

subject to $\mathbf{1}_p^T\mathbf{v} = 1$, of which the proof can be found in Robert and Casella (1999). Since the constant d lies in the interval $(0, 1)$, every element of the matrix $\mathbf{S}$ will be positive, where the non-zero diagonal values create aperiodicity and the non-zero off-diagonal elements result in irreducibility. In our proposed method, the stationary probabilities $v_1, \ldots, v_p$ serve as the network-based ranking measures, which actually is the eigenvector of the matrix $\mathbf{S}$ corresponding to its largest eigenvalue that happens to be one. As self-loops are considered, we have $v_j = c^*_j$ based on equation (2.2) if the jth predictor is a singleton in the network.

In particular, for any starting vector $\mathbf{v}^{(0)}$ satisfying the restriction $\mathbf{1}_p^T\mathbf{v}^{(0)} = 1$, the following power iteration algorithm will converge to the desired eigenvector because the matrix $\mathbf{S}$ is stochastic, that is,

$$\mathbf{v}^{(k+1)T} = \mathbf{v}^{(k)T}\mathbf{S}. \tag{2.3}$$

This algorithm is computationally efficient even as we deal with a large set of predictors.

2.2. *Some properties.* It is straightforward from (2.1) and (2.2) to obtain the following equation:

$$\mathbf{v} = (1-d)\big(\mathbf{I}_p - d\mathbf{A}^{*T}\big)^{-1}\mathbf{c}^*, \tag{2.4}$$

where $\mathbf{I}_p$ is a $p \times p$ identity matrix. The matrix $(\mathbf{I}_p - d\mathbf{A}^{*T})^{-1}$ plays a crucial role in adjusting the original ranking measure. The matrix $\mathbf{I}_p - d\mathbf{A}^{*T}$ is nonsingular and the norm of its inverse matrix satisfies [Langville and Meyer (2012), Section 7.1]

$$\left\|\left(\mathbf{I}_p - d\mathbf{A}^{*T}\right)^{-1}\right\|_\infty = (1-d)^{-1}, \tag{2.5}$$

where $\|\cdot\|_\infty$ is the norm refers to the largest absolute row sum of a matrix. We shall avoid a large constant d that is close to one, which will make the network-based ranking measure extremely sensitive to the choice of both the weight constant d and the normalized adjacency matrix $\mathbf{A}^*$ [Langville and Meyer (2012), Chapter 6].

With the Taylor expansion, we further obtain from (2.4) that

$$\mathbf{v} = (1-d)\left\{\mathbf{I}_p + \sum_{k=1}^{\infty} d^k \left(\mathbf{A}^{*T}\right)^k\right\}\mathbf{c}^*. \tag{2.6}$$

Thus, for a specific predictor X, its correlation with the response variable will be adjusted by its connected explanatory variables, and their contributions to the correlation will diminish quickly with respect to the distance to the considered variable X along the network.

We obtain an algebraic inequality between the original and the network-based ranking measures in the following theorem, which is useful for us to analyze the asymptotic property of the proposed screening statistic.

THEOREM 1. *For two nonnegative vectors* $\mathbf{c}_1$ *and* $\mathbf{c}_2$ *satisfying the constraint* $\mathbf{1}_p^T\mathbf{c}_1 = \mathbf{1}_p^T\mathbf{c}_2 = 1$, *we have*

$$\|\mathbf{v}_1 - \mathbf{v}_2\|_\infty \le \|\mathbf{c}_1 - \mathbf{c}_2\|_\infty,$$

where $\mathbf{v}_i = (1-d)(\mathbf{I}_p - d\mathbf{W}^T)^{-1}\mathbf{c}_i$, *and* $\mathbf{W}$ *is a* $p \times p$ *stochastic matrix which is irreducible and aperiodic. Furthermore, all elements of both* $\mathbf{v}_1$ *and* $\mathbf{v}_2$ *are nonnegative.*

If we replace the vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ with the sample and the population marginal ranking measures, respectively, it is obvious from Theorem 1 that the resulting distance is greater than that of the corresponding network-adjusted measures. The sure screening property of marginal sample measures have been well studied [Fan and Lv (2008), Li, Zhong and Zhu (2012), Zhu et al. (2011)]. Hence, if the network-based population measure can accurately identify the set of active predictors, we shall expect the corresponding sample measure will capture these variables with high probability.

2.3. *Choice of tuning parameters.* There are two tuning parameters in the proposed method of Section 2.1, the weight d which balances the original adaptive ranking measure $\mathbf{c}(r)$ and the network information $\mathbf{A}$, and the integer r refers to the size of the variable set including all active predictors. We adopt a criterion based on the random decoupling strategy as suggested by Barut, Fan and Verhasselt (2016) to choose these tuning parameters, which determines the number of potentially active variables under a certain level of false positives.

We first generate the response Y^* by randomly permuting the original response Y, while keeping the design matrix $\mathbf{X}$ intact. Since the randomly permuted response Y^* is independent of the design matrix, the estimated marginal correlations based on decoupled data represent the level of spurious correlations under the null model that no predictors are useful [Barut, Fan and Verhasselt (2016)].

Given certain constants d and r, let

$$V^*(d,r) = \max_{1\le j\le p} \bigl|v_j^*(d,r)\bigr|,$$

where $v_j^*(d,r)$ is the network-based ranking measure for the decoupled data $\mathbf{X}$ and Y^* as computed in (2.2). In order to control the proportion of false positives, $V^*(d,r)$ is used to determine a threshold value by the following two-stage procedure. First, the random permutation is repeated K times, resulting in a set of values $\{V_1^*(d,r),\dots,V_K^*(d,r)\}$. Second, the maximum value of the set $\widetilde{V}^*(d,r) = \max\{V_1^*(d,r),\dots,V_K^*(d,r)\}$ is used as the final threshold. Thus the set of chosen variables under a certain level of false positives is

$$\widetilde{\mathcal{M}}_{d,r} = \bigl\{j : v_j(d,r) \ge \widetilde{V}^*(d,r)\bigr\},$$

where $v_j(d,r)$ is the network-based ranking measure based on the original data. The following criterion is then considered to choose the values for both d and r:

$$(\hat{d},\hat{r}) = \arg\max_{d,r} |\widetilde{\mathcal{M}}_{d,r}|,$$

where $|\widetilde{\mathcal{M}}_{d,r}|$ refers to the cardinality of the set $\widetilde{\mathcal{M}}_{d,r}$. In our analysis, we randomly permute the original data five times as also used by Barut, Fan and Verhasselt (2016). This criterion has been adopted with the following considerations. Feature screening usually serves as a preliminary reduction step, and is often followed by a conventional feature selection for further refinement. Thus, it is more important for screening to retain all the truly active predictors opposed to feature selection, which focuses on achieving a high true positive and a low false positive simultaneously. The above criterion can lead to a model which includes as many useful predictors as possible, which is consistent with the goal of screening.

3. Simulation studies. In this section, we assess the performance of the proposed method of Section 2.1 with the simulated data. We consider two existing popular feature screening approaches, SIS [Fan and Lv (2008)] and DC-SIS [Li,

Zhong and Zhu (2012)] for the network-based screening implementation, denoted as SIS-Network and DC-SIS-Network, respectively. Besides the above four methods, we also consider the following three alternatives:

- VS-Network: a variable selection method with the network-constrained penalty proposed by Li and Li (2008), which incorporates information encoded by the known biological networks into the variable selection procedure;
- HOLP: a screening technique based on high-dimensional ordinary least squares projection, where the sure screening property holds without the restrictive marginal correlation assumption [Wang and Leng (2016)];
- Screen3S: A data-driven conditional screening algorithm with three steps, which enjoys the sure screening property under weaker assumptions on the model [Hong, Wang and He (2016)].

3.1. *Artificial networks*. Simulation settings are summarized in Table S.1 of Supplementary materials [Wu, Zhu and Feng (2018)]. We generate the datasets with $p = 5000$ and $n = 100$. Two types of regulatory networks are considered in this study. First, the degrees of the network are generated randomly from power-law or exponential distributions, which are the two most common structures of modeling biological networks [Barabasi and Oltvai (2004)]. A power-law degree distribution implies that the network has few highly connected nodes which are also known as hubs. By contrast, an exponential distribution indicates that the system has a characteristic degree and there are no highly connected hubs. The histograms of the variable degrees are plotted in Figure A.1 of Appendix. Second, we set the first variable X_1 as important, and it regulates six active variables, $X_2, \ldots, X_7$. The variable X_7 is connected with another inactive variable X_8. Third, for the remaining 4992 inactive variables, we use the R package "igraph"to randomly generate the edges following the corresponding degree distribution. Thus, there are one important variable with degree seven, five important variables with degree two and one important variable with degree three.

We consider two models for generating datasets with discrete responses in the simulation study:

(I) linear model (model I): $Y = \text{logit}(\mathbf{X}^T\boldsymbol{\beta} + \varepsilon)$,
(II) exponential model (model II): $Y = \text{logit}(\exp(\mathbf{X}^T\boldsymbol{\beta}/2) + \varepsilon)$,

where $\text{logit}(\cdot)$ is the logistic function, $\varepsilon \sim N(0, \sigma^2)$ with $\sigma^2 = \boldsymbol{\beta}^T\boldsymbol{\beta}/r_\sigma$. The parameter r_σ refers to the signal-noise ratio of the model, which is set to eight or 16 in our study. The predictors $\mathbf{X}$ are simulated from a multivariate normal distribution with marginal means 0 and marginal variances 1. Two correlation structures $\Sigma = (\sigma_{jl})_{p\times p}$ are considered to represent different networks. The first correlation matrix Σ_1 corresponds to a "correlation" network with $\sigma_{jl} = 0.5$ if nodes j and l are connected by an edge, otherwise $\sigma_{jl} = 0$. The second correlation matrix Σ_2 corresponds to a "concentration" network with $\Sigma_2 = \Sigma_1^{-1}$.

For each model, four settings of the coefficient vector $\boldsymbol{\beta}$ are considered:

(i)
$$\boldsymbol{\beta} = \left(4, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, \ldots, 0\right)^T;$$

(ii)
$$\boldsymbol{\beta} = \left(4, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}, 0, \ldots, 0\right)^T;$$

(iii)
$$\boldsymbol{\beta} = \left(\frac{1}{4}, 2, 2, 2, 2, 2, 2, 0, \ldots, 0\right)^T;$$

(iv)
$$\boldsymbol{\beta} = \left(\frac{1}{4}, 2, 2, 2, -2, -2, -2, 0, \ldots, 0\right)^T.$$

For the first two cases, the first variable X_1 has a strong effect on the response, and six other connected predictors have weak effects. Under settings (iii) and (iv), the first variable's effect is weak, and six other connected variables carry strong effects. Furthermore, under settings (ii) and (iv), some of the variables are negatively related with the response Y.

Simulation suggests that the proposed method is computationally feasible. The computation of original marginal correlations for 5000 predictors can be accomplished within five seconds using a laptop with standard configurations. The power iteration for calculating network-based measures can converge within 100 overall iterations and be accomplished within one minute. Although a number of permutations need to be computed for the choice of tuning parameters, as they can be analyzed in a highly parallel manner, the overall computational cost is still affordable. To facilitate data analysis and applications beyond this study, we have developed R code and made it publicly available at http://bb.shufe.edu.cn/bbcswebdav/users/2011000070/Codes/Net_FS_FUN.R.

We assess the performance through the following criteria which have been adopted in the literature [Li, Zhong and Zhu (2012), Zhu et al. (2011), among others]:

(a) $\mathcal{S}$: the minimum number of predictors required to ensure the inclusion of all the truly active predictors. We report the median and the average absolute deviation of the numbers out of 500 replications.

(b) $\mathcal{P}_a$: the proportion of active predictors that are selected under a given model size in 500 replications. For cases (i) and (iii), seven important predictors will be divided into two categories: variable X_1 that owns relatively high degrees in the network and variables $X_2, \ldots, X_7$. For cases (ii) and (iv), they will be divided

into three categories: variable X_1, variables X_2, X_3, X_4 which are positively correlated with the response and variables X_5, X_6, X_7 which are negatively correlated with the response.

(c) TPR and FPR: the true and false positive rates for identification performance when the model size is determined using minimum prediction error criterion via cross-validation.

(d) AUC: the area under an ROC curve for prediction performance, when the model size is determined using minimum prediction error criterion via cross-validation.

For the first criterion $\mathcal{S}$, a better feature screening approach should lead to a closer value to the true number of active variables, which is seven in this simulation study. For the second criterion $\mathcal{P}_a$, a better feature screening approach should give a higher proportion of truly important variables given a model size. The numbers of predictors considered in the model are chosen as $[n/\log(n)]$ and $[2n/\log(n)]$, respectively, where $[a]$ denotes the integer part of a. For the third and the fourth criteria, a linear logistic regression model is always considered to further refine the variable selection after feature screening, and used to predict responses. We adopt 5-fold cross-validation to find the model with the minimum prediction error. A better approach should have larger TPR and smaller FPR. An independent test dataset with sample size 100 is generated in the same manner. Then prediction performance is quantified using AUC based on predicted probability, of which a larger value indicates better prediction.

3.1.1. *Variable selection and prediction.* In Tables 1, 2 and 3, we report the simulation results of the above five criteria for models I–II with the power-law degree distribution, the signal-noise ratio $r_\sigma = 16$ and the correlation matrix Σ_1 (Scenario 1). According to the result based on the criterion $\mathcal{S}$ in Table 1, the performance of seven approaches are comparable under model I with the coefficient setting (iii) except HOLP, where the model is linear and all active predictors are positively correlated with the response Y. HOLP is slightly inferior to other methods as it is based on a linear model for continuous response. Moreover, the result of the criterion $\mathcal{P}_a$ in Table 2 implies that the six methods can identify active variables almost perfectly. Under model I with settings (i), (ii) and (iv), where the predictors with low degrees in the network carry either weak or negative effects on the response, it is observed from Table 1 that both SIS-Network and DC-SIS-Network outperform their original counterparts significantly. The result of criterion $\mathcal{P}_a$ reported in Table 2 indicates that these two network-based feature screening procedures can identify most of the active predictors given model sizes $[n/\log(n)]$ and $[2n/\log(n)]$. Although in Table 2 with setting (iv), for variables X_2–X_4, the proposed methods have $\mathcal{P}_a$ slightly smaller than their original counterparts, they can more accurately identify the first variable X_1 with weak effect, which is of greater concern.

TABLE 1
The median of the estimated size $\mathcal{S}$ out of 500 *replications for Scenario* 1, *including the average absolute deviation within parentheses*

	(i)	(ii)	(iii)	(iv)
Model I				
VS-Network	39.4 (28.4)	126.8 (87.8)	8.5 (1.5)	587.1 (28.9)
HOLP	539.5 (408.5)	1078.5 (734.5)	174.0 (128.5)	3638.0 (816.5)
Screen3S	20.0 (12.0)	74.0 (63.0)	8.0 (0.6)	2477.0 (1011.5)
SIS	20.0 (12.0)	86.5 (72.5)	8.1 (0.6)	2701.0 (918.0)
DC-SIS	27.5 (18.5)	122.5 (103.5)	8.0 (1.0)	3080.0 (1023.5)
SIS-Network	10.5 (2.5)	10.0 (2.0)	8.0 (1.0)	103.0 (93.0)
DC-SIS-Network	11.0 (4.0)	10.0 (2.0)	8.0 (1.0)	103.0 (92.0)
Model II				
VS-Network	374.4 (100.4)	430.0 (104.0)	320.2 (112.0)	563.6 (15.5)
HOLP	3831.5 (823.0)	4233.5 (538.0)	3857.5 (804.0)	4398.0 (397.5)
Screen3S	3453.0 (1032.5)	3871.5 (743.5)	3664.0 (1045.5)	3945.0 (660.5)
SIS	1427.5 (938.0)	2255.5 (1212.5)	629.0 (483.0)	4023.5 (580.5)
DC-SIS	1230.0 (811.5)	1828.5 (1040.5)	506.0 (384.5)	3982.0 (599.0)
SIS-Network	15.0 (6.0)	16.0 (7.0)	23.0 (14.0)	250.0 (109.5)
DC-SIS-Network	12.0 (3.0)	14.0 (4.0)	23.0 (14.0)	143.5 (117.0)

Under model II, the advantages of the DC-SIS-Network are more prominent as the nonlinear relationship is present between the response variable and the predictors. In general, the network information is quite useful for screening those active predictors in this study as implied in Tables 1 and 2.

When model size is determined using minimum prediction error criterion (Table 3), the proposed methods have similar or better performance than the alternatives. Although different predictors are used with different methods, the values of AUC are similar when the true model is included in the family specified by the working model. As the true model is model II, the misspecified working model leads to relatively poor prediction for all methods. However, the proposed method has achieved slightly better prediction.

We also examine the performance of the proposed method under the scenarios of which the predictors are generated with a "concentration" network. The results are summarized in Tables S.2–S.4 and S.17–S.19 of the Supplementary materials [Wu, Zhu and Feng (2018)]. All methods perform slightly worse under the "concentration" network, but we observe similar favorable performance of the proposed method.

Additional simulation results with different network structures, signal-noise-ratios and types of responses are reported in the Supplementary materials [Wu, Zhu and Feng (2018)], and similar conclusions can be drawn.

TABLE 2
The proportion $\mathcal{P}_a$ of selected truly active predictor with a given number of predictors ($[\frac{n}{\log(n)}] = 22, [\frac{2n}{\log(n)}] = 44$) *in* 500 *replications for Scenario* 1

		(i)		(ii)			(iii)		(iv)		
		X_1	X_{2-7}	X_1	X_{2-4}	X_{5-7}	X_1	X_{2-7}	X_1	X_{2-4}	X_{5-7}
Model I											
VS-Network	$\frac{n}{\log(n)}$	1.00	0.91	1.00	0.91	0.58	1.00	0.99	0.00	0.30	0.22
	$\frac{2n}{\log(n)}$	1.00	0.95	1.00	0.95	0.67	1.00	1.00	0.02	0.50	0.35
HOLP	$\frac{n}{\log(n)}$	1.00	0.38	1.00	0.40	0.21	0.98	0.61	0.00	0.17	0.13
	$\frac{2n}{\log(n)}$	1.00	0.52	1.00	0.49	0.30	1.00	0.72	0.00	0.23	0.18
Screen3S	$\frac{n}{\log(n)}$	1.00	0.89	1.00	0.88	0.64	1.00	0.98	0.07	0.48	0.30
	$\frac{2n}{\log(n)}$	1.00	0.95	1.00	0.92	0.74	1.00	0.99	0.09	0.59	0.41
SIS	$\frac{n}{\log(n)}$	1.00	0.89	1.00	0.88	0.64	1.00	0.99	0.01	0.47	0.26
	$\frac{2n}{\log(n)}$	1.00	0.95	1.00	0.92	0.73	1.00	0.99	0.01	0.57	0.37
DC-SIS	$\frac{n}{\log(n)}$	1.00	0.86	1.00	0.86	0.61	1.00	0.97	0.00	0.45	0.23
	$\frac{2n}{\log(n)}$	1.00	0.92	1.00	0.90	0.70	1.00	0.99	0.01	0.56	0.33
SIS-Network	$\frac{n}{\log(n)}$	1.00	0.94	1.00	0.97	0.89	1.00	0.99	0.82	0.42	0.21
	$\frac{2n}{\log(n)}$	1.00	0.98	1.00	0.99	0.94	1.00	1.00	0.84	0.47	0.31
DC-SIS-Network	$\frac{n}{\log(n)}$	1.00	0.92	1.00	0.98	0.92	1.00	0.98	0.81	0.44	0.22
	$\frac{2n}{\log(n)}$	1.00	0.98	1.00	0.99	0.96	1.00	0.99	0.84	0.47	0.34
Model II											
VS-Network	$\frac{n}{\log(n)}$	0.94	0.27	0.94	0.28	0.10	0.82	0.32	0.01	0.09	0.03
	$\frac{2n}{\log(n)}$	0.98	0.39	0.98	0.42	0.21	0.90	0.53	0.03	0.14	0.10
HOLP	$\frac{n}{\log(n)}$	0.18	0.04	0.14	0.03	0.02	0.10	0.05	0.00	0.01	0.02
	$\frac{2n}{\log(n)}$	0.28	0.06	0.20	0.06	0.03	0.16	0.08	0.01	0.03	0.04
Screen3S	$\frac{n}{\log(n)}$	0.95	0.23	0.94	0.23	0.13	0.82	0.39	0.01	0.09	0.07
	$\frac{2n}{\log(n)}$	0.96	0.31	0.97	0.31	0.19	0.89	0.48	0.01	0.14	0.10
SIS	$\frac{n}{\log(n)}$	0.95	0.23	0.94	0.23	0.13	0.82	0.39	0.01	0.10	0.07
	$\frac{2n}{\log(n)}$	0.96	0.31	0.97	0.32	0.19	0.90	0.49	0.01	0.14	0.09
DC-SIS	$\frac{n}{\log(n)}$	0.98	0.22	0.97	0.21	0.10	0.86	0.38	0.01	0.08	0.05
	$\frac{2n}{\log(n)}$	1.00	0.31	0.99	0.30	0.16	0.92	0.48	0.01	0.14	0.09
SIS-Network	$\frac{n}{\log(n)}$	0.93	0.60	0.89	0.59	0.56	0.94	0.58	0.18	0.11	0.16
	$\frac{2n}{\log(n)}$	0.93	0.66	0.93	0.64	0.63	0.94	0.64	0.22	0.15	0.18
DC-SIS-Network	$\frac{n}{\log(n)}$	0.94	0.65	0.94	0.66	0.60	0.94	0.57	0.19	0.12	0.18
	$\frac{2n}{\log(n)}$	0.95	0.71	0.96	0.71	0.69	0.95	0.64	0.23	0.21	0.19

TABLE 3
The means of TPR, FDR, AUC out of 500 *replications for Scenario* 1, *including the average absolute deviation within parentheses*

	TPR	**FDR**	**AUC**	**TPR**	**FDR**	**AUC**
Model I		(i)			(ii)	
VS-Network	0.81 (0.10)	0.25 (0.09)	0.92 (0.02)	0.74 (0.12)	0.29 (0.29)	0.89 (0.05)
HOLP	0.36 (0.07)	0.68 (0.11)	0.91 (0.02)	0.30 (0.13)	0.73 (0.12)	0.88 (0.03)
Screen3S	0.76 (0.09)	0.28 (0.14)	0.92 (0.02)	0.66 (0.09)	0.40 (0.12)	0.88 (0.03)
SIS	0.76 (0.09)	0.28 (0.14)	0.92 (0.02)	0.66 (0.09)	0.40 (0.12)	0.88 (0.03)
DC-SIS	0.75 (0.11)	0.32 (0.11)	0.91 (0.02)	0.62 (0.09)	0.44 (0.13)	0.88 (0.03)
SIS-Network	0.82 (0.10)	0.25 (0.11)	0.92 (0.02)	0.78 (0.08)	0.24 (0.09)	0.89 (0.03)
DC-SIS-Network	0.82 (0.08)	0.26 (0.14)	0.92 (0.02)	0.78 (0.21)	0.26 (0.09)	0.89 (0.03)
		(iii)			(iv)	
VS-Network	0.95 (0.05)	0.17 (0.07)	0.95 (0.02)	0.41 (0.11)	0.89 (0.07)	0.59 (0.05)
HOLP	0.53 (0.10)	0.53 (0.10)	0.88 (0.03)	0.07 (0.07)	0.94 (0.06)	0.56 (0.03)
Screen3S	0.95 (0.05)	0.14 (0.11)	0.95 (0.02)	0.25 (0.11)	0.79 (0.09)	0.60 (0.05)
SIS	0.95 (0.05)	0.14 (0.11)	0.95 (0.02)	0.22 (0.08)	0.81 (0.08)	0.59 (0.05)
DC-SIS	0.94 (0.06)	0.16 (0.09)	0.94 (0.02)	0.21 (0.08)	0.83 (0.07)	0.59 (0.05)
SIS-Network	0.95 (0.06)	0.14 (0.11)	0.95 (0.02)	0.44 (0.19)	0.75 (0.15)	0.63 (0.09)
DC-SIS-Network	0.95 (0.07)	0.14 (0.08)	0.95 (0.02)	0.42 (0.17)	0.76 (0.14)	0.63 (0.09)
Model II		(i)			(ii)	
VS-Network	0.45 (0.16)	0.84 (0.10)	0.59 (0.08)	0.37 (0.08)	0.87 (0.07)	0.59 (0.07)
HOLP	0.03 (0.03)	0.98 (0.02)	0.57 (0.04)	0.03 (0.03)	0.97 (0.03)	0.56 (0.04)
Screen3S	0.25 (0.10)	0.79 (0.09)	0.63 (0.06)	0.22 (0.08)	0.81 (0.08)	0.61 (0.05)
SIS	0.25 (0.10)	0.79 (0.09)	0.62 (0.05)	0.22 (0.07)	0.81 (0.08)	0.61 (0.05)
DC-SIS	0.25 (0.11)	0.78 (0.09)	0.63 (0.05)	0.22 (0.07)	0.81 (0.08)	0.62 (0.05)
SIS-Network	0.51 (0.28)	0.65 (0.22)	0.65 (0.06)	0.43 (0.24)	0.68 (0.22)	0.64 (0.06)
DC-SIS-Network	0.54 (0.29)	0.63 (0.23)	0.66 (0.06)	0.47 (0.29)	0.65 (0.22)	0.65 (0.06)
		(iii)			(iv)	
VS-Network	0.51 (0.20)	0.79 (0.13)	0.60 (0.04)	0.14 (0.14)	0.97 (0.02)	0.51 (0.04)
HOLP	0.03 (0.03)	0.98 (0.02)	0.56 (0.04)	0.01 (0.01)	0.99 (0.01)	0.55 (0.03)
Screen3S	0.34 (0.09)	0.70 (0.13)	0.62 (0.06)	0.03 (0.03)	0.97 (0.03)	0.55 (0.03)
SIS	0.34 (0.09)	0.70 (0.13)	0.62 (0.06)	0.03 (0.03)	0.97 (0.03)	0.55 (0.03)
DC-SIS	0.34 (0.20)	0.70 (0.15)	0.62 (0.06)	0.03 (0.03)	0.98 (0.02)	0.54 (0.03)
SIS-Network	0.61 (0.29)	0.64 (0.24)	0.65 (0.06)	0.24 (0.04)	0.97 (0.03)	0.57 (0.03)
DC-SIS-Network	0.63 (0.29)	0.65 (0.22)	0.66 (0.06)	0.31 (0.05)	0.96 (0.04)	0.60 (0.03)

3.1.2. *Sensitivity analysis*. We further consider the sensitivity of the proposed method to the network choice. We employ two network structures with different parameters. The first network (N1) is constructed based on the estimated correlation matrix. A cutoff value $\tilde{\eta}$ is adopted to generate the sparse adjacency matrix $\mathbf{A}$ as most genes only interact with a few genes, where $A_{jl} = 1$ if the absolute Pearson correlation between jth and lth variables is larger than $\tilde{\eta}$ and 0 otherwise. We use $\tilde{\eta} = 0.3, 0.4$ and 0.5 to represent different interaction criteria. The second net-

work (N2) is generated based on the true network used in this simulation study, but of which $\tilde{\tau}$ percent interactions are misspecified. We adopt $\tilde{\tau} = 20$ and $\tilde{\tau} = 40$ to consider different misspecified levels. In this study, there are six important edges, so the probability of at least one of them is misspecified will be around 0.738 and 0.953, respectively.

The summary results of $\mathcal{S}$ based on different prior networks are reported in Tables S.29 and S.30 of the Supplementary materials [Wu, Zhu and Feng (2018)]. It is observed that the methods based on the estimated correlation matrix outperform the corresponding marginal methods (SIS or DC-SIS), but are inferior to the proposed methods incorporating the true or the lightly misspecified network structure. Hence, we can definitely benefit from the prior network with the proposed method if a large proportion of information provided by this network is correct.

3.2. *Network constructed based on a real dataset*. In the previous simulation study, the "gene" measurements and the "biological" networks have been generated from parametric distributions, which may be overly simplified compared to what is practically observed. To tackle this problem, we further conduct a simulation study based on the the real dataset GSE71729, which will be fully examined in the following section. Specifically, we use the observed gene expression and the protein-protein interaction (PPI) network in H.sapiens from the High-quality INTeractomes (HINT) database (http://hint.yulab.org/, version: 06 /03 /2013) [Das and Yu (2012)]. There are seven active predictors of which the coefficients are generated under the same settings in Section 3.1. We randomly select one gene which is connected with six other genes and regard these seven genes as active predictors. Both models I and II are employed to simulate the response. The results are reported in Tables S.31–S.33 of the Supplementary materials [Wu, Zhu and Feng (2018)]. With more predictors together with a complicated correlation structure, it is more challenging to screen important features than the previous study. However, the proposed method still reaches favorable performance, especially for scenarios with settings (i) and (ii).

4. Real data analysis. In this section, we analyze two real microarray gene expression datasets with discrete responses based on the methods considered in the previous section.

4.1. *Data*. The datasets GSE50493 and GSE71729 are publicly available in the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). GSE50493 includes the gene expression of 72 samples from two classes: the melanoma brain metastases (M-BM, 29 samples) and the melanoma extracranial metastases (M-EM, 43 samples) [Chen et al. (2014)]. A total of 47,323 measurements are available on these samples. Identification of molecular differences between these two types of tumors would be useful in the development of organ-specific therapeutic approaches. GSE71729 includes 145 primary and

61 metastatic pancreatic ductal adenocarcinoma (PDAC) tumors, 17 cell lines, 46 pancreas and 88 distant site adjacent normal samples [Moffitt et al. (2015)]. Among them, we select 145 primary PDAC samples (P-PDAC) and 46 pancreas normal samples (N-PDAC) with 19,749 measurements of gene expression for analysis.

The network information on genes are obtained based on a high-quality PPI network. In this PPI network, there are 18,864 pairwise interactions among 6342 genes, where the remaining unpaired genes are treated as disconnected nodes. All datasets used in this study have been already published and require no ethics approval.

4.2. *Prediction performance*. In the real data analysis, we use a two-stage procedure to predict responses, including a feature screening at the first stage and a regular logistic regression analysis for discrete responses at the second.

We assess the prediction accuracy of different methods based on a 10-fold cross-validation (CV) and a 5-fold CV, which are repeated 50 times on each dataset, respectively. For both SIS-Network and DC-SIS-Network, the parameters d and r are chosen with the method stated in Section 2.3, which is summarized in Table S.34 of the Supplementary materials [Wu, Zhu and Feng (2018)]. Furthermore, we use the 5-fold CV to determine the number of predictors retained in models with feature screening methods under the constraint that the number is smaller than the sample size. We then assess the prediction performance of each method based on their AUC values.

In Table 4, we report the mean AUC and the corresponding average absolute deviation. Those three methods incorporating network information (VS-Network, SIS-Network and DC-SIS-Network) achieve comparable performance in prediction, and outperform their competitors HOLP, Screen3S, SIS and DC-SIS.

4.3. *Stability of gene selection processes*. With a wealth of public molecular datasets for the same disease, it is common that the biomarker signatures from different studies have few overlaps, which leads to difficulties in biological explains. Therefore, in addition to prediction capabilities, the stability of the screening procedure is also crucial. The stability of a gene selection process is evaluated by the following Jaccard coefficient in two settings, the 10-fold CV and the 5-fold CV:

$$\mathrm{JC} = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|}, \tag{4.1}$$

where G_1 and G_2 are the two considered gene sets, and $|G|$ refers to the cardinality of the set G. The value of Jaccard coefficient will achieve one if the compared sets are just the same. According to the mechanism of CV, the training sets are generated with $8/9 \approx 89\%$ overlap for 10-fold CV and 75% overlap for 5-fold CV.

TABLE 4
Prediction performance and stability of different methods in terms of AUC and Jaccard coefficient, respectively

Methods	GSE50493		GSE71729	
	10-fold CV	**5-fold CV**	**10-fold CV**	**5-fold CV**
AUC:				
VS-Network	0.689 (0.040)	0.623 (0.052)	0.883 (0.009)	0.881 (0.008)
HOLP	0.665 (0.052)	0.612 (0.051)	0.855 (0.021)	0.845 (0.015)
Screen3S	0.674 (0.042)	0.607 (0.049)	0.861 (0.016)	0.862 (0.022)
SIS	0.664 (0.042)	0.596 (0.047)	0.850 (0.013)	0.852 (0.015)
DC-SIS	0.667 (0.055)	0.606 (0.052)	0.851 (0.016)	0.853 (0.020)
SIS-Network	0.681 (0.043)	0.614 (0.045)	0.886 (0.020)	0.880 (0.011)
DC-SIS-Network	0.686 (0.050)	0.619 (0.047)	0.882 (0.019)	0.879 (0.014)
Jaccard coefficient:				
VS-Network	0.245 (0.010)	0.101 (0.010)	0.436 (0.166)	0.290 (0.023)
HOLP	0.192 (0.053)	0.038 (0.021)	0.497 (0.246)	0.302 (0.093)
Screen3S	0.214 (0.045)	0.057 (0.013)	0.512 (0.186)	0.314 (0.103)
SIS	0.202 (0.063)	0.084 (0.061)	0.527 (0.226)	0.347 (0.113)
DC-SIS	0.167 (0.041)	0.093 (0.064)	0.490 (0.198)	0.303 (0.117)
SIS-Network	0.543 (0.112)	0.275 (0.095)	0.651 (0.229)	0.524 (0.114)
DC-SIS-Network	0.592 (0.141)	0.437 (0.103)	0.732 (0.217)	0.690 (0.135)

The mean measure obtained for each method on the two datasets over 50 times 10-fold cross validation or 5-fold cross validation. Average absolute deviation is shown within parentheses.

In Table 4, we have reported the mean Jaccard coefficients of the genes chosen by seven models out of 50 replications, respectively. It is observed that both SIS-Network and DC-SIS-Network have manifest superiority in terms of screening stability. For example, for the dataset GSE50493 with 5-fold CV, the methods SIS-Network and DC-SIS-Network have average JC $= 0.275$ and 0.437, respectively, compared to 0.101 (VS-Network), 0.038 (HOLP), 0.057 (Screen3S), 0.084 (SIS) and 0.093 (DC-SIS). Even when we consider the 10-fold CV where there is 89% overlap in the training samples, the four screening methods without the network information and VS-Network identify very few same genes. The combined information from both the biological networks and the gene expressions are quite useful for improving the stability of retrieving important biomarkers.

4.4. *Biomarker identification*. Compared to SIS-Network, DC-SIS-Network achieves better overall performance as demonstrated in previous two sections. It gets competitive performance in prediction and higher gene selection stability. Thus, we further analyze the biomarkers chosen by DC-SIS-Network for the dataset GSE71729 as pancreatic ductal adenocarcinoma remains a lethal disease with a 5-year survival of 4% [Moffitt et al. (2015)]. The details of the top 100 genes

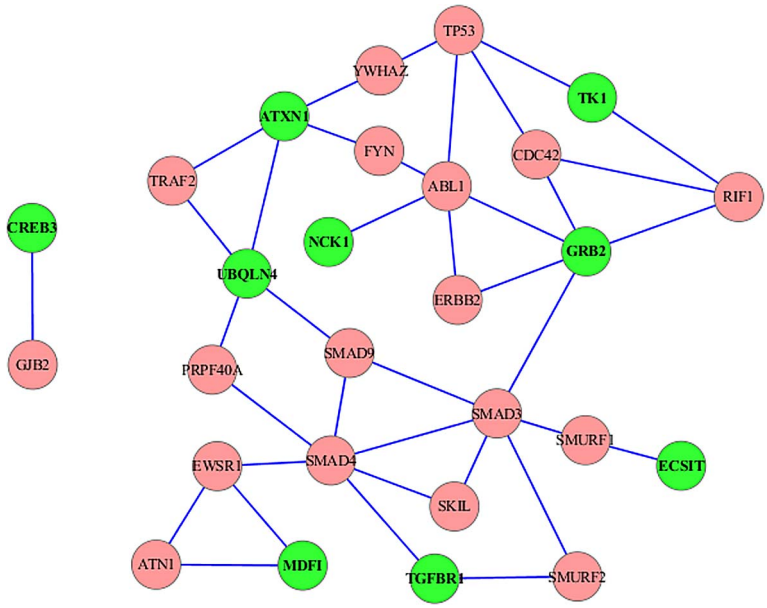

FIG. 1. *The subnetworks of the top* 100 *genes identified by DC-SIS-Network for dataset GSE*71729. *Nodes represent human genes, and they are connected by a link if they belong to the PPI network. Each gene is labeled by its gene symbol. The genes that are not connected with any other top* 100 *genes are not displayed. The top* 10 *genes selected by DC-SIS-Network are highlighted in bold, but the* 4*th gene MAPK*6 *is not shown since it is not connected with any other top* 100 *genes.*

chosen by DC-SIS-Network are presented in Table S.35 of the Supplementary materials [Wu, Zhu and Feng (2018)], and the degrees of these genes and their ranks provided by DC-SIS are also reported. The top 10 genes identified by DC-SIS are highlighted in bold. It is shown that most of the active predictors with strong marginal effects can still be selected by the network-based screening procedure.

Figure 1 shows the subnetwork structures of these top 100 genes, where the genes that are not connected with any other top 100 genes are not displayed. The top 10 genes selected by DC-SIS-Network are marked in bold in Figure 1, where the 4th gene MAPK6 is omitted since it is not connected with any other top 100 genes. The results indicate that those biomarkers (e.g., GRB2, CREB3, UBQLN4, ATXN1 and TK1) identified by DC-SIS-Network often play important roles in the gene network, which are often overlooked by DC-SIS, and they interact with some discriminative or even non-discriminative genes to form a collective biological function.

In addition, we conduct enrichment analysis to examine the functional and biological relationships of the selected genes based on the KEGG pathway, which is implemented using DAVID 6.8 [Huang, Sherman and Lempicki (2009a, 2009b)]. The pathway enrichment analysis is used for the identification of pathways that are significantly over-represented in a given gene set, which may suggest possible

functional characteristics of the given set. The pathways shared by the considered gene set are compared to the background distribution. Specifically, a p-value of a certain pathway is calculated using the hypergeometric distribution, given the proportion of genes in the whole genome that are annotated to that pathway. The pathway with smaller p-value is more significantly associated with the group of genes. There are 34 out of the top 100 genes involved in the 27 pathways with the p-values less than 0.05, while only 13 out of the top 100 genes detected by DC-SIS have significantly p-values with seven pathways. We summarize the results of the KEGG pathway terms in Table S.36 of the Supplementary materials [Wu, Zhu and Feng (2018)], where the pathways that are also significant for DC-SIS are highlighted in bold. For example, the pathway Pancreatic cancer is identified by DC-SIS-Network with p-value 6.41×10^{-5}, while it is not significant for DC-SIS. This pathway includes some genes that process the normal duct epithelium to pancreatic ductal adenocarcinoma in different stages, such as p53 and SMAD4 [Hruban et al. (2000)]. Transforming growth factor TGF-beta signaling pathway (p-value: 2.18×10^{-4}) has been shown to act both as a tumor suppressor and as a tumor promoter in pancreatic cancer [Javle et al. (2014)], depending on tumor stage and cellular context, which can be utilized in targeted therapy clinical trials of pancreatic cancer. Many detected pathways based on DC-SIS-network are also found to be related with pancreatic cancer by other studies, such as Adherens junction (p-value: 9.81×10^{-5}) and MAPK signaling pathway (p-value: 0.0262) [Campagna et al. (2008)]. The KEGG pathway analysis imply that the proposed network-based feature screening tends to select the genes in the same pathway as a whole, which is very meaningful since the complex diseases are multifactorial biological events manifested through simultaneous changes in expressions of many genes and proteins.

Furthermore, the disease-gene association is analyzed based on the Genetic Association Database (GAD) by DAVID 6.8 [Huang, Sherman and Lempicki (2009a, 2009b)], where the p-value of these top 100 genes identified by DC-SIS-Network is 0.034 for the pancreatic adenocarcinoma. However, the similar disease-gene association analysis for those top 100 genes selected by DC-SIS indicates that these genes do not have a significant association with the pancreatic adenocarcinoma. Most of biomarkers selected by the proposed procedure are considered to have diagnostic values for the pancreatic adenocarcinoma in the existing literature. For example, GRB2 (the rank based on DC-SIS is 8253) was coimmunoprecipitated with EGFR after EGF stimulation, and the expression of EGFR and its ligand on pancreatic adenocarcinoma have been shown to be associated with tumor aggressiveness [Huang et al. (2003)]. MDFI (the rank based on DC-SIS is 4684) is a tumor suppressor gene that has been found to be epigenetically modified in pancreatic adenocarcinoma [Jagirdar et al. (2013)]. In addition, mutations of the TGFBR1 (the rank based on DC-SIS is 3234) are found in some of patients with pancreatic adenocarcinoma [Wong and Lemoine (2009)]. SMAD4 (the rank based on DC-SIS is 8993) is genetically inactivated in about 55% of all pancreatic

adenocarcinomas, which is shown to influence the prognosis after surgical resection for invasive pancreatic adenocarcinoma [Tascilar et al. (2001)]. TP53 gene (the rank based on DC-SIS is 5754) is mutated in 75% of sporadic pancreatic adenocarcinomas and is regarded as a potential molecular marker of pancreatic adenocarcinomas [Brune et al. (2008)].

5. Discussion. Gene networks are consistently developed with gradual understanding of complex relationships among genes. In previous studies, it is common to identify biomarkers by measuring the marginal correlation between each predictor (gene) and a response variable (disease or survival time), and the network information on these predictors are usually ignored. This often brings up some confusing screening or variable selection results in similar biological studies. By incorporating the gene network information, we are able to make use of prior knowledge on gene relationships, and use the network-based correlation measures to select important groups of genes. In practice, the computation of the network-based measures is effective even for large-scale datasets, which can not only lead to good response prediction but also provide stable biomarker selection.

Further biomarker analysis of the dataset GSE71729 indicates that the proposed method tends to choose those genes in the same pathway as a whole, which have been shown active in the development of the pancreatic adenocarcinoma in some biological studies. As the network information is available, we advocate using the network-based screening approaches so that we can capture more complicated interactions between genes since the development of one disease is usually derived from a complex biological process and it is not controlled by some genes individually.

This study can be potentially extended in multiple aspects. First, we have focused on data with a discrete or continuous response variable without censoring. Since the research on prognosis outcomes with censoring survival time is also very important, the incompletely observed data should be taken into account in the development of network-based screening methods. Second, a undirected network has been adopted for its lucid interpretation and widespread existence. With minor modifications, the proposed approach can be extended to account for weighted or directed networks.

APPENDIX

A.1. Proof of Theorem 1.

$$\begin{aligned}\|\mathbf{v}_1 - \mathbf{v}_2\|_\infty &= (1-d)\big\|\big(I - d\mathbf{W}^T\big)^{-1}(\mathbf{c}_1 - \mathbf{c}_2)\big\|_\infty \\ &\le (1-d)\big\|\big(I - d\mathbf{W}^T\big)^{-1}\big\|_\infty \|\mathbf{c}_1 - \mathbf{c}_2\|_\infty \\ &= \|\mathbf{c}_1 - \mathbf{c}_2\|_\infty.\end{aligned}$$

The last equality follows from (2.5).

Moreover, it follows from (2.6) that all the elements of vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are nonnegative.

A.2. The histogram of node degrees.

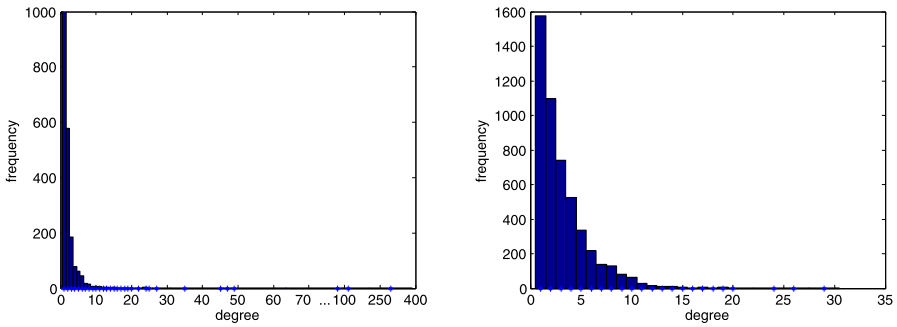

FIG. A.1. *The histogram of degrees of the variables in the simulated network* **A**. (a) *Pow-law distribution*, *the maximum degree is* 271; (b) *Exponential distribution*, *the maximum degree is* 29.

A.3. Some additional numerical studies. We also consider the following models for generating continuous data in our simulation study:

(III) linear model (model III): $Y = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon$,
(IV) exponential model (model IV): $Y = \exp(\mathbf{X}^T\boldsymbol{\beta}/2) + \varepsilon$,

where $\varepsilon \sim N(0, \sigma^2)$ with $\sigma^2 = \boldsymbol{\beta}^T\boldsymbol{\beta}/r_\sigma$. The prior network and predictor X are generated in the same manner as Section 3. A linear model is adopted for determining the final model size to obtain the values of TPR and FPR. Different from AUC for the discrete response. The results with different network structures, signal-noise-ratios and correlation matrices are reported in the Supplementary materials [Wu, Zhu and Feng (2018)].

SUPPLEMENTARY MATERIAL

Some additional tables (DOI: 10.1214/17-AOAS1097SUPP; .pdf). The Supplementary Materials includes some additional simulation results with different network structures, signal-noise-ratios and types of responses, the top 100 biomarkers for the dataset GSE71729 identified by DC-SIS-Network and the corresponding KEGG pathway analysis results.

REFERENCES

BARABÁSI, A.-L., GULBAHCE, N. and LOSCALZO, J. (2011). Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12** 56–68.

BARABASI, A. L. and OLTVAI, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5** 101–113.

BARUT, E., FAN, J. and VERHASSELT, A. (2016). Conditional sure independence screening. *J. Amer. Statist. Assoc.* **111** 1266–1277. MR3561948

BRUNE, K., HONG, S.-M., LI, A. et al. (2008). Genetic and epigenetic alterations of familial pancreatic cancers. *Cancer Epidemiol. Biomark. Prev.* **17** 3536–3542.

CAMPAGNA, D., COPE, L., LAKKUR, S. S., HENDERSON, C., LAHERU, D., IACOBUZIO-DONAHUE, C. A. et al. (2008). Gene expression profiles associated with advanced pancreatic cancer. *Int. J. Clin. Exp. Pathol.* **1** 32–43.

CHEN, G., CHAKRAVARTI, N., AARDALEN, K. et al. (2014). Molecular profiling of patient-matched brain and extracranial melanoma metastases implicates the PI3K pathway as a therapeutic target. *Clin. Cancer Res.* **20** 5537–5546.

CHUANG, H., LEE, E., LIU, Y. T. et al. (2006). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3** 140.

CUN, Y. and FRÖHLICH, H. (2012). Biomarker gene signature discovery integrating network knowledge. *Biol.* **1** 5–17.

DAS, J. and YU, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6** 92.

FAN, J., HAN, F. and LIU, H. (2014). Challenges of Big Data analysis. *Nat. Sci. Rev.* **1** 293–314.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322

GUSTAFSSON, M., NESTOR, C. E., ZHANG, H. et al. (2014). Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Gen. Med.* **6** 1–11.

HAWRYLYCZ, M., MILLER, J. A., MENON, V., FENG, D., DOLBEARE, T., GUILLOZET-BONGAARTS, A. L., JEGGA, A. G., ARONOW, B. J., LEE, C.-K., BERNARD, A., GLASSER, M. F., DIERKER, D. L., MENCHE, J., SZAFER, A., COLLMAN, F., GRANGE, P., BERMAN, K. A., MIHALAS, S., YAO, Z., STEWART, L., BARABÁSI, A.-L., SCHULKIN, J., PHILLIPS, J., NG, L., DANG, C., HAYNOR, D. R., JONES, A., ESSEN, D. C. V., KOCH, C. and LEIN, E. (2015). Canonical genetic signatures of the adult human brain. *Nat. Neurosci.* **18** 1832–1844.

HE, Z. and YU, W. (2010). Stable feature selection for biomarker discovery. *Comput. Biol. Chem.* **34** 215–225.

HONG, H. G., WANG, L. and HE, X. (2016). A data-driven approach to conditional screening of high-dimensional variables. *Statistics* **5** 200–212.

HRUBAN, R. H., GOGGINS, M., PARSONS, J. and KERN, S. E. (2000). Progression model for pancreatic cancer. *Clin. Cancer Res.* **6** 2969–2972.

HUANG, D. W., SHERMAN, B. T. and LEMPICKI, R. A. (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4** 44–57.

HUANG, D. W., SHERMAN, B. T. and LEMPICKI, R. A. (2009b). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37** 1–13.

HUANG, Z.-Q., BUCHSBAUM, D. J., RAISCH, K. P., BONNER, J. A., BLAND, K. I. and VICKERS, S. M. (2003). Differential responses by pancreatic carcinoma cell lines to prolonged exposure to Erbitux (IMC-C225) anti-EGFR antibody. *J. Surg. Res.* **111** 274–283.

JAGIRDAR, R., SOLENOV, E. I., HATZOGLOU, C., MOLYVDAS, P.-A., GOURGOULIANIS, K. I. and ZAROGIANNIS, S. G. (2013). Gene expression profile of aquaporin 1 and associated interactors in malignant pleural mesothelioma. *Genetics* **517** 99–105.

JAVLE, M., LI, Y., TAN, D., DONG, X., CHANG, P., KAR, S. and LI, D. (2014). Biomarkers of TGF-β signaling pathway and prognosis of pancreatic cancer. *PLoS ONE* **9** e85942.

LANGVILLE, A. N. and MEYER, C. D. (2012). *Google's PageRank and Beyond*: *The Science of Search Engine Rankings*. Princeton Univ. Press, Princeton, NJ. MR3052718

LEISERSON, M. D., VANDIN, F., WU, H. T. et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47** 106–114.

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

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139.

MARTINEZLEDESMA, E., VERHAAK, R. G. and TREVINO, V. (2015). Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm. *Sci. Rep.* **5**.

MOFFITT, R. A., MARAYATI, R., FLATE, E. L. et al. (2015). Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **47** 1168–1178.

PAN, W., XIE, B. and SHEN, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics* **66** 474–484. MR2758827

ROBERT, C. P. and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York. MR1707311

SHI, X., YI, H. and MA, S. (2015). Measures for the degree of overlap of gene signatures and applications to TCGA. *Brief. Bioinform.* **16** 266–272.

TASCILAR, M., SKINNER, H. G., ROSTY, C. et al. (2001). The SMAD4 protein and prognosis of pancreatic ductal adenocarcinoma. *Clin. Cancer Res.* **7** 4115–4121.

TAYLOR, I. W., LINDING, R., WARDEFARLEY, D. et al. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **27** 199–204.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VIDAL, M., CUSICK, M. E. and BARABASI, A. L. (2011). Interactome networks and human disease: Cell. *Cell* **144** 986–998.

WANG, X. and LENG, C. (2016). High dimensional ordinary least squares projection for screening variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 589–611.

WONG, H. H. and LEMOINE, N. R. (2009). Pancreatic cancer: Molecular pathogenesis and new therapeutic targets. *Nat. Rev. Gastroenterol. Hepatol.* **6** 412–422.

WU, M., ZHU, L. and FENG, X. (2018). Supplement to "Network-based feature screening with applications to genome data." DOI:10.1214/17-AOAS1097SUPP.

YU, G. and LIU, Y. (2016). Sparse regression incorporating graphical structure among predictors. *J. Amer. Statist. Assoc.* **111** 707–720. MR3538699

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via elastic net. *J. R. Stat. Soc. Ser. B* **67** 301–320.

M. WU
X. FENG
SCHOOL OF STATISTICS AND MANAGEMENT
INSTITUTE OF DATA SCIENCE AND STATISTICS
SHANGHAI UNIVERSITY OF FINANCE
AND ECONOMICS
777 GUODING ROAD
SHANGHAI 200433
CHINA
E-MAIL: wu.mengyun@mail.sufe.edu.cn
feng.xingdong@mail.sufe.edu.cn

L. ZHU
RESEARCH CENTER OF APPLIED STATISTICS
INSTITUTE OF STATISTICS AND BIG DATA
RENMIN UNIVERSITY OF CHINA
59 ZHONGGUANCUN AVENUE, HAIDIAN DISTRICT
BEIJING 100872
CHINA
E-MAIL: zhu.liping@ruc.edu.cn