Gene Network-based Cancer Prognosis Analysis with Sparse Boosting

Shuangge Ma*[1], Yuan Huang[2], Jian Huang[3], Kuangnan Fang[4]

[1] School of Public Health, Yale University, New Haven, CT, USA
[2] Department of Statistics, Penn State University, University Park, PA, USA
[3] Departments of Statistics and Actuarial Science, and Biostatistics, University of Iowa, Iowa City, IA, USA
[4] Department of Statistics, Xiamen University, Xiamen, China

Email: Shuangge Ma*- shuangge.ma@yale.edu; Yuan Huang - yuanhuang.stat@psu.edu; Jian Huang - jian-huang@uiowa.edu; Kuangnan Fang - ruiqwy@163.com;

*Corresponding author

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 414

Abstract

Background: High-throughput gene profiling studies have been extensively conducted, searching for markers associated with cancer development and progression. In this study, we analyze cancer prognosis studies where the response variables are censored survival. With gene expression data, we adopt the weighted co-expression network to describe the interplay among genes. In network-based analysis, nodes represent genes. There are subsets of nodes, called modules, that are tightly connected to each other. Genes within the same modules tend to have coregulated biological functions. For cancer prognosis data with gene expression measurements, our goal is to identify cancer markers, while properly accounting for the network module structure.

Results: A two-step sparse boosting-based approach, called NSBoost (Network-based Sparse Boosting), is proposed for marker selection. In the first step for each module separately, we adopt a sparse boosting approach for within-module marker selection and construct module-level "super markers". In the second step, we use the super markers to represent effects of all genes within the same modules and conduct module-level selection using a sparse boosting approach. Simulation study shows that NSBoost can more accurately identify markers than alternative approaches. In the analysis of breast cancer and lymphoma prognosis studies, NSBoost outperforms alternatives including boosting and penalization approaches by identifying a smaller number of genes/modules and having better prediction performance.

Conclusions: Network provides an effective way of describing the interplay among genes. Accommodating the network structure using NSBoost may improve cancer marker selection.

Background

High-throughput gene expression profiling studies have been extensively conducted, searching for markers associated with the development and progression of cancer. In this study, we analyze cancer prognosis studies, where the outcome variables are progression-free, overall, or other types of survival. In many existing analyses of cancer gene expression data, it has been assumed that genes have interchangeable effects [1]. Biomedical studies have shown that there exists inherent coordination among genes and, essentially, all biological functions of living cells are carried out through the coordinated effects of multiple genes. There are multiple ways of describing the interplay among genes. The most popular ways are

perhaps gene pathway and network [2]. Compared with pathway analysis, network analysis sometimes may be more informative as it describes not only whether two genes are "connected" but also the strength of connection. In addition, network analysis may analyze all genes, whereas pathway analysis often focuses on annotated genes only. On the negative side, unlike with pathways, research that targets at linking specific network structures to biological functions remains scarse. In the literature, there is no definitive evidence on the relative performance of pathway and network analyses. In this article, we focus on network-based analysis and refer to other studies for discussions and comparisons of pathway- and network-based analyses.

In network analysis, nodes represent genes. Nodes are connected if the corresponding genes have coregulated biological functions or correlated expressions. Multiple approaches have been proposed to compute the connectedness measure between two genes. See for example [3, 4] and references therein. Published studies have suggested that the connectedness measure may have important implications. For example, hub genes, which are genes "well connected" with a large number of genes, tend to have more important biological functions. There are subsets of nodes, called modules, that are tightly connected to each other. Genes within the same modules tend to have coordinated biological functions, whereas genes in different modules tend to have different, unrelated biological functions.

Statistical methods that can accommodate the high dimensionality of high-throughput cancer data can be roughly classified as dimension reduction or variable selection methods. Both families of methods have been employed in network-based analysis. Dimension reduction methods, for example principal component analysis-based methods, search for linear combinations of all genes or all genes within the same modules as cancer markers. See for example [5] and references therein. Such methods may have satisfactory prediction performance but often suffer a lack of interpretability. In addition, they contradict the fact that not all genes are involved in cancer pathogenesis. Variable selection methods search for a subset of profiled genes as markers and may be more interpretable [6]. In this article, we focus on the development of a network-based variable selection method and refer to other publications for comprehensive discussions and comparisons of dimension reduction and variable selection methods.

For cancer prognosis studies with gene expression measurements, we adopt the weighted co-expression network [7] to describe the interplay among genes. We develop NSBoost (Network-based Sparse Boosting), a two-step sparse boosting based method, for cancer marker selection. The proposed method may advance from existing methods by explicitly accounting for the module structure of network in marker selection and

hence can be more informative. Another advantage of the proposed method is that it is relatively "independent" of the network construction procedure and thus is applicable to multiple types of networks.

Methods

Construction of weighted coexpression network

There are multiple ways of building gene networks. Examples include Boolean network, Bayesian network, use of continuous model, and others. To the best of our knowledge, in the literature there is a lack of definitive evidence on the relative performance of different network construction methods. In this study, we adopt WGCNA [7], which is built on the understanding that the coordinated co-expressions of genes encode interacting proteins with closely related biological functions and cellular processes. Detailed investigation of WGCNA has been conducted by Dr. Steve Horvath and his group at UCLA. Their studies suggest that modules in the weighted co-expression network have important biological implications. Genes with a higher connectivity are more likely to be involved in important molecular processes. In addition, incorporating connectivity in the detection of differentially expressed genes can lead to significantly improved reproducibility.

Construction of the weighted co-expression network is computationally simple, and a user-friendly R package is available for such a purpose [8]. In addition, WGCNA is completely inferred from gene expression measurements of a single study and hence does not demand a large amount of biological experiments. On the negative side, it is built on the estimated covariance matrix. In cancer gene expression studies, with the sample size significantly smaller than the number of genes, the uniform consistency of the covariance matrix estimation is debatable. Thus, unlike some other ways of describing gene interplay (for example, pathways), the weighted co-expression network structure may vary considerably in studies with comparable setup. For integrity of this study, we describe the WGCNA algorithm below and refer to [7] for more details.

1. Assume that there are d genes. For genes k and j $(= 1 \dots d)$, compute $cor(k, j)$, the Pearson correlation coefficient of their expressions. Compute the similarity measure $S(k, j) = |cor(k, j)|$;

2. Compute the adjacency function $a_{k,j} = S^b(k, j)$, where the adjacency parameter b is chosen using the scale-free topology criterion. In our data analysis, we find that $b = 6$, which has been suggested in quite a few published studies, lead to satisfactory results;

3. For gene k, compute its connectivity $C_k = \sum_u a_{k,u}$;

4. For gene k $(= 1 \dots d)$, compute the topological overlap based dissimilarity measure $d_{k,j} = 1 - \omega_{k,j}$ where $\omega_{k,j} = (l_{k,j} + a_{k,j})/(min(C_k, C_j) + 1 - a_{k,j})$ and $l_{k,j} = \sum_u a_{k,u} a_{j,u}$. Define the dissimilarity matrix D, whose (k, j)th element is $d_{k,j}$;

5. Identify network modules using matrix D and the hierarchical clustering approach. Apply the dynamic tree cut approach [9] to cut the clustering tree (dendrogram), and identify the resulting branches as modules.

Denote M as the number of modules constructed using the above algorithm and $S(m)$ as the size of module $m(= 1, \dots, M)$.

Statistical modeling

Let T_i be the logarithm of survival time and X_i be the d-dimensional gene expressions for the ith subject in a random sample of size n. The AFT (accelerated failure time) model assumes

$$T_i = \alpha + X_i'\beta + \epsilon_i, \ i = 1, \dots, n, \tag{1}$$

where α is the intercept, $\beta \in \mathrm{R}^d$ is the unknown regression coefficient, and ϵ_i is the random error with an unknown distribution. Under right censoring, one observation consists of (Y_i, δ_i, X_i), where $Y_i = \min\{T_i, C_i\}$, C_i is the logarithm of censoring time, and $\delta_i = 1\{T_i \leq C_i\}$ is the censoring indicator.

The AFT model provides a flexible alternative to the Cox proportional hazards model [10]. It assumes a linear function for the log-transformed survival time and thus may provide a more straightforward description of the gene effects on survival than alternatives (for example Cox model, which describes the survival hazard). There are multiple approaches for estimating the AFT model with an unspecified error distribution. Examples include the Buckley-James estimator which adjusts censored observations using the Kaplan-Meier estimator and the rank based estimator which is motivated by the score function of the partial likelihood function. With high-dimensional gene expression data, those approaches suffer a prohibitively high computational cost.

A computationally more feasible approach is the weighted least squares approach [11]. Denote $\hat{F}_n$ as the Kaplan-Meier estimator of F, the distribution function of T. It can be computed as $\hat{F}_n(y) = \sum_{i=1}^n w_i 1\{Y_{(i)} \leq y\}$. Here w_is are the jumps in the Kaplan-Meier estimator computed as $w_1 = \frac{\delta_{(1)}}{n}$ and $w_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}$, $i = 2, \dots, n$. w_is have also been referred to as the Kaplan-Meier

weights in [11]. Here $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the order statistics of Y_i's, $\delta_{(1)}, \ldots, \delta_{(n)}$ are the associated censoring indicators, and $X_{(1)}, \ldots, X_{(n)}$ are the associated gene expressions. The weighted least squares loss function is

$$\frac{1}{2}\sum_{i=1}^{n} w_i(Y_{(i)} - \alpha - X'_{(i)}\beta)^2.$$

We center $X_{(i)}$ and $Y_{(i)}$ using their corresponding w_i-weighted means, respectively. Let $\bar{X}_w = \sum_{i=1}^n w_i X_{(i)} / \sum_{i=1}^n w_i$ and $\bar{Y}_w = \sum_{i=1}^n w_i Y_{(i)} / \sum_{i=1}^n w_i$. Denote $X^*_{(i)} = w_i^{1/2}(X_{(i)} - \bar{X}_w)$ and $Y^*_{(i)} = w_i^{1/2}(Y_{(i)} - \bar{Y}_w)$. We can then rewrite the weighted least squares loss function as

$$l(\beta) = \frac{1}{2}\sum_{i=1}^{n}(Y^*_{(i)} - X^{*\prime}_{(i)}\beta)^2. \tag{2}$$

The simple form of this loss function makes it computationally affordable and thus suitable for high-dimensional gene expression data.

Network-based gene selection

The proposed approach belongs to the family of boosting approaches. Boosting assembles a strong learner using a set of weak learners [12]. It is appropriate for cancer genomic data analysis as individual genes usually have weak effects, but combined together, they may have strong effects.

Algorithm

We first rearrange gene expressions so that $\beta = (\beta^{1\prime}, \ldots, \beta^{M\prime})'$, where β^m is the length $S(m)$ vector of regression coefficients for all genes within module m. Denote β^m_j as the jth component of β^m and $X^{*m}_{(i)}$ as the components of $X^*_{(i)}$ that correspond to β^m.

Step I: Within-module boosting

For $m = 1, \ldots, M$, consider the objective function $\frac{1}{2}\sum_{i=1}^n (Y^*_{(i)} - X^{*m\prime}_{(i)}\beta^m)^2$, which is $l(\beta)$ evaluated only on genes within the mth module. This is equivalent to the objective function obtained from fitting an AFT model using only the mth module.

(a) Initialization. Set $k = 0$ and $\beta^{m[k]} = 0$ (component-wise).

(b) Fit and update. $k = k + 1$.

Compute $\hat{s} = argmin_{1\leq s\leq S(m)} argmin_\gamma \frac{1}{2}\sum_{i=1}^n (Y^*_{(i)} - X^{*m\prime}_{(i)}\beta^{m[k-1]} - \gamma X^{*m}_{(i),s})^2 +$
$\log(n)\sum_{1\leq s\leq S(m)} I(\beta_s^{m[k-1]} + \gamma \neq 0)$;

Compute $\hat{\gamma} = argmin_{\gamma} \sum_{i=1}^{n} \frac{1}{2}(Y^*_{(i)} - X^{*m\prime}_{(i)}\beta^{m[k-1]} - \gamma X^{*m}_{(i),\hat{s}})^2$;

Update $\beta_s^{m[k]} = \beta_s^{m[k-1]}$ for $s \neq \hat{s}$ and $\beta_{\hat{s}}^{m[k]} = \beta_{\hat{s}}^{m[k-1]} + \nu\hat{\gamma}$, where ν is the step size. As suggested in [13] and references therein, the choice of ν is not critical as long as it is small. In our numerical study, we set $\nu = 0.1$;

(c) Iteration. Repeat Step (b) for K iterations;

(d) Stopping. At iteration k, compute the objective function $F(k) = \sum_{i=1}^{n} \frac{1}{2}(Y^*_{(i)} - X^{*m\prime}_{(i)}\beta^{m[k]})^2 + \log(n) \sum_{1 \leq s \leq S(m)} I(\beta_s^{m[k]} \neq 0)$. Estimate the stopping iteration by $\tilde{k} = argmin_{1 \leq k \leq K} F(k)$. For subject i, define its "super marker" for module m as $Z^m_{(i)} = X^{*m\prime}\beta^{m[\tilde{k}]}_{(i)}$;

Step II: Module-wise boosting

Consider the objective function $\frac{1}{2}\sum_{i=1}^{n}(Y^*_{(i)} - Z'_{(i)}\tau)^2$, where $Z_{(i)} = (Z^1_{(i)}, \ldots, Z^M_{(i)})'$ and $\tau = (\tau_1, \ldots, \tau_M)$ is the unknown regression coefficient. That is, in the least squares objective function (2), we use the super markers, which can represent the effects of all genes within the same modules, to replace the original gene expressions.

(a) Initialization. Set $k = 0$ and $\tau^{[k]} = 0$ (component-wise).

(b) Fit and update. $k = k + 1$.

Compute

$\hat{s} = argmin_{1 \leq s \leq M} argmin_{\gamma} \sum_{i=1}^{n} \frac{1}{2}(Y^*_{(i)} - Z'_{(i)}\tau^{[k-1]} - \gamma Z_{(i),s})^2 + \log(n) \sum_{1 \leq s \leq M} I(\tau_s^{[k-1]} + \gamma \neq 0)$;

Compute $\hat{\gamma} = argmin_{\gamma} \sum_{i=1}^{n} \frac{1}{2}(Y^*_{(i)} - Z'_{(i)}\tau^{[k-1]} - \gamma Z_{(i),\hat{s}})^2$;

Update $\tau_s^{[k]} = \tau_s^{[k-1]}$ for $s \neq \hat{s}$ and $\tau_{\hat{s}}^{[k]} = \tau_{\hat{s}}^{[k-1]} + \nu\hat{\gamma}$, where $\nu = 0.1$ is the step size;

(c) Iteration. Repeat Step (b) for K iterations;

(d) Stopping. At iteration k, compute the objective function $F(k) = \sum_{i=1}^{n} \frac{1}{2}(Y^*_{(i)} - Z'_{(i)}\tau^{[k]})^2 + \log(n) \sum_{1 \leq s \leq M} I(\tau_s^{[k]} \neq 0)$. Estimate the stopping iteration by $\hat{k} = argmin_{1 \leq k \leq K} F(k)$.

$\sum_{m=1}^{M} \tau_m^{\hat{k}} Z^m_{(i)} = \sum_{m=1}^{M} \tau_m^{\hat{k}} \{X^{*m\prime}_{(i)}\beta^{m[\tilde{k}]}\}$ is the resulted strong learner for $Y^*_{(i)}$. Genes and modules with nonzero regression coefficients in the strong learner are identified as associated with cancer.

Rationale

With NSBoost, marker selection is achieved in two steps. This basic strategy is similar to that in [14]. In the first step, each module is analyzed separately. Genes within different modules tend to have different biological functions. Thus, it is sensible to analyze each module separately in the sense that different biological functionalities should be considered separately. On the other hand, genes within the same modules tend to have correlated but never identical biological functions. Thus, with the within-module selection, we search for genes that are associated with cancer within a group of functionally related genes. This step of selection can not only remove noises but also lead to the construction of super marker, which is a linear combination of selected genes and can represent effects of all genes within the same module. The introduction of super marker shares a similar spirit with [5]. In the second step, we consider the joint effects of all super markers and hence all modules. When a large number of genes are profiled, it is not reasonable to assume that all modules are cancer-associated. It is necessary to conduct the second step of selection and discriminate cancer-associated modules from noises. Thus, with the proposed approach, we may identify which modules are cancer-associated as well as which genes are cancer-associated within selected modules. Considering the important biological implications of modules, this approach may be more informative than alternatives that ignore the module structure. Another advantage of the proposed approach is its computational affordability. In the within-module boosting, the number of genes per module can be much smaller than the total number of genes. In addition, this step can be carried out in a parallel manner. Thus, the first step of boosting has computational cost much smaller than ordinary boosting with all genes. With WGCNA, the number of modules and hence super markers is usually not large – numerical studies in [5, 6] suggesting less than 20. Thus, the computational cost of the second step of boosting is almost negligible.

In both steps, marker selection is achieved using a sparse boosting approach. In high-dimensional data analysis, boosting may be preferred because of its low computational cost, flexibility, and satisfactory empirical performance. With ordinary boosting, when the stopping rule is properly chosen, the resulted strong learners may enjoy a certain degree of sparsity, and so marker selection can be achieved. This can be seen from [15] and followup studies as well as our numerical study. However, recent studies [13, 16] suggest that with high dimensional data, ordinary boosting may not be "sparse enough". That is, it may identify a considerable number of false positives. The sparse boosting approach adopted here has been motivated by [13]. In particular, the objective function used for boosting and stopping has two terms. The

first term measures goodness-of-fit, and the second term measures model complexity. We adopt the BIC criterion for model complexity measure. As a comparison, ordinary boosting only considers goodness-of-fit in boosting, which may introduce noisy variables that happen to be able to slightly improve the goodness-of-fit. The addition of model complexity measure in sparse boosting may lead to sparser models and hence reduce the number of false positives. On the negative side, sparse boosting can be computationally more expensive than ordinary boosting as the model complexity measure and hence the whole objective function are not differentiable and cannot be minimized using gradient-based approaches. The sparse boosting approach adopted in this study differs from those in [13, 16]. Particularly, previous studies focus on continuous and categorical data, whereas we analyze censored survival data, which can be more complicated. The adopted BIC criterion has been more commonly adopted as a model complexity measure than the MDL (minimum description length). In addition, the two-step boosting procedure can effectively accommodate the module structure.

Parameter path

Parameter path, which is the graphical presentation of the estimates as a function of number of iterations, may provide further insights into NSBoost. Consider the simulation setting corresponding to row 1 of Table 1. For a better view, we simplify the simulation setting and consider 4 modules with 4 genes per module. The first 2 modules are cancer-associated, within which there are 2 cancer-associated genes. Thus, among the 16 simulated genes, 4 are associated with cancer. For comparison, we also study NBoost (details described in the Results section). For a randomly generated dataset, the parameter paths are shown in Figures 1 (NSBoost) and 2 (NBoost), respectively.

Within each module, the parameter paths of NSBoost are similar to those of other regularized variable selection approaches [12]. By considering model complexity in boosting, the NSBoost parameter paths are "smoother" than their NBoost counterparts. NBoost does not consider model complexity in boosting and thus may suffer a risk of false positives. For example in the top right panel, NBoost has one false positive while NSBoost does not. Our limited numerical study suggests that, in the within-module boosting step, NSBoost may identify "signals" even with purely noisy modules. Thus, the module-level boosting is needed, which can effectively remove noisy modules as a whole (see the bottom panel). With a combination of the two boosting steps, NSBoost can be sparser than NBoost at both within-module level and module level.

Results

Simulation

We conduct simulation to better understand properties of the proposed approach. In each simulated dataset, there are 100 subjects. We simulate 50 gene clusters with 20 genes in each cluster. Gene expressions have marginally standard normal distributions. Genes within different clusters have independent expressions. For genes within the same clusters, their expressions have the following correlation structures: (1) auto-regressive correlation, where expressions of genes j and k have correlation coefficient $\rho^{|j-k|}$. $\rho = 0.3$ or 0.7, corresponding to weak and strong correlations; (2) Banded correlation, where expressions of genes j and k have correlation coefficient $max(0, 1 - |j - k| \times \rho)$. $\rho = 0.2$ or 0.33; (3) Compound symmetry, where expressions of genes j and k have correlation coefficient ρ when $j \neq k$. $\rho = 0.3$ or 0.7. Within each of the first 4 clusters, the first 5 genes are associated with survival. There are thus a total of 20 cancer-associated genes, and the rest are noises. For cancer-associated genes, we generate their regression coefficients from $Unif[0.5, 1.5]$. Thus, some genes have large effects, and others have moderate to small effects. We generate the logarithm of survival time from the AFT model. The logarithm of censoring time is independently generated from a normal distribution. We adjust the censoring distribution parameters so that the censoring rate is about 40%. The simulation settings mimic the real data setting considered in this study, where the interplay among genes can be described using the network module structure. Genes within the same modules tend to have correlated expressions, whereas genes within different modules tend to have weakly correlated or uncorrelated expressions. Among a large number modules, only a few are associated with survival. Within those important modules, some genes are cancer-associated and others are noises.

To better gauge performance of the proposed approach, we also consider the following alternatives: (1) Enet (elastic net) [17], which is a penalization approach and has been extensively used in the analysis of gene expression data; (2) Boost, which is the ordinary boosting approach and takes the goodness-of-fit as the only criterion for choosing weaker learners. A BIC-type criterion similar to that with NSBoost is adopted for stopping; (3) SBoost, which is a sparse boosting approach and considers the goodness-of-fit and model complexity measured using the BIC criterion in boosting and stopping. The above three approaches ignore the network structure and treat all gene effects as interchangeable. In addition, we also consider (4) NBoost, which is a network-based boosting approach and has a two-step algorithm similar to that with the proposed approach. The difference is that in boosting, only the goodness-of-fit is considered

when choosing weaker learners. We are aware that there are a large number of approaches that can be used to analyze the simulated data. The above four approaches are chosen for comparison, as the Enet is one of the most extensively used penalization approaches and particularly includes Lasso and ridge penalization as special cases; and Boost, SBoost and NBoost have a boosting framework closest to that of NSBoost.

Summary statistics based on 200 replicates are presented in Table 1. Enet and Boost can identify all the true positives. However, under some scenarios, they may identify a considerable number of false positives. SBoost, which considers model complexity in boosting but ignores the network structure, is "overly sparse" in our simulation by having a considerable number of false negatives. Without accounting for model complexity in boosting, NBoost identifies a large number of false positives. Under all simulated scenarios, NSBoost is capable of identifying the majority or all of the true positives while having a small number of false positives. We have also experienced with a few other simulation settings and reached similar conclusions.

Data analysis

We collect four cancer prognosis studies with gene expression measurements. Brief descriptions are provided in Table 2 and below. We refer to the original publications for more details.

D1. Breast cancer is the second leading cause of cancer death among women in the United States. Despite major progress in breast cancer treatment, the ability to predict metastasis of the tumor still remains limited. Huang et al. [18] reported a study investigating metastatic states and relapses in breast cancer patients. Affymetrix genechips were used for the profiling of 71 samples.

D2. Sorlie et al. [19] conducted a gene expression profiling study, investigating whether it was feasible to classify breast carcinomas based on the gene expression patterns. cDNA Profiling of a total of 85 samples was conducted, showing that breast cancer could be classified into a basal epithelial-like group, an ERBB2-overexpressing group, and a normal breast-like group. Among the 85 samples, 58 had survival information available and will be analyzed in this study.

D3. Mantle cell lymphoma (MCL) accounts for $\sim 8\%$ of all NHLs (non-Hodgkin lymphoma). Rosenwald et al. [20] reported a gene expression study of MCL survival. Among 101 untreated patients with no history of previous lymphoma, 92 were classified as having MCL based on morphologic and

immunophenotypic criteria. Survival times of 64 patients were available, and the rest were censored. The median survival time was 2.8 years. Lymphochip DNA microarrays were used to quantify mRNA expressions in the lymphoma samples. Gene expression data on 8,810 cDNA elements were available.

D4. Diffuse large B-cell lymphoma (DLBCL) is a cancer of the B-cell. It accounts for $\sim 40\%$ of all NHL cases. A DLBCL gene expression study was reported in [21]. This study retrospectively collected tumor biopsy specimens and clinical data for 240 patients with untreated DLBCL. The median follow up was 2.8 years, with 138 observed deaths. Lymphochip cDNA microarray was used to measure the expressions of 7,399 genes.

Among the four studies, one used Affymetrix and three used cDNA for profiling. We process the datasets as follows. We conduct normalization using the lowess approach for cDNA data and the robust multi-array (RMA) approach for Affymetrix data. Missing measurements are imputed using the K-nearest neighbors approach. Affymetrix gene expression measurements are log2 transformed. We select the 500 genes with the largest variances for downstream analysis. Here the prescreening may serve multiple purposes. First in cancer gene expression studies, usually genes with higher variations are of more interest. Second, it is expected that the number of cancer prognosis-associated genes is far smaller than 500. Prescreening may remove genes that are highly unlikely to be cancer-associated and significantly reduce computational cost. More importantly, as discussed above, WGCNA involves estimating the covariance matrix. Prescreening may significantly reduce the dimensionality of this matrix and improve estimation accuracy. With selected genes, we normalize their expressions to have zero median and unit variance.

With datasets D1-D4, the WGCNA approach constructs 4, 5, 6 and 6 modules, respectively. For dataset D4, we show in Figure 3 the details on module construction. Results for other datasets are available from the authors. We apply the NSBoost as well as the four alternative approaches discussed in the above section. Analysis results are presented in Table 3. More details on the identified genes are available from the authors. Table 3 shows that NSBoost identifies a small number of genes as cancer prognosis markers. By conducting the module-level sparse boosting and hence encouraging sparsity at the module level, NSBoost identifies the smallest number of modules, which may lead to more focused hypotheses for downstream analysis. Genes identified by NSBoost differ significantly from those identified using Enet, Boost, and SBoost. For example for dataset D1, the numbers of overlapped genes are 4, 5 and 3, respectively. The sets of genes identified by NBoost and NSBoost are more similar, which is as expected, as

the two approaches both use boosting for marker selection and account for the module structure. For example for dataset D1, the number of overlapped genes is 23. With our limited understanding of cancer genomics, we are unable to draw conclusions on which sets of identified markers are "the most meaningful". As an indirect evaluation, we examine the prediction performance of different approaches, which proceeds as follows: (1) Randomly split the data into a training set and a testing set with sizes 3:1; (2) Analyze the training data and identify markers. We note that a natural byproduct of the proposed approach is a prediction model; (3) Make prediction for subjects in the testing set. The predictive model can lead to a predicted risk score $X'\beta$ for each subject. Dichotomize the risk scores at median and create two risk groups. Compute the logrank statistic, which measures the survival difference between the two groups; (4) To avoid an extreme partition, repeat Steps (1)-(3) 100 times, and compute the average logrank statistic. Table 3 shows that with the four analyzed datasets, NSBoost has the largest logrank statistcis and hence best performance in separating subjects into groups with different survival risks. The superior prediction performance may provide an indirect support to the marker selection validity of NSBoost.

Discussion and Conclusions

In cancer genomic studies, an important goal is to identify markers associated with prognosis. There exists inherent coordination among genes, and network provides an effective way of describing such coordination. In this study, we adopt the weighted co-expression network and develop a two-step sparse boosting-based approach to account for the network structure in cancer marker selection. The proposed approach is intuitively reasonable. Simulation and data analysis show its satisfactory performance.

In the literature, multiple ways of describing the interplay among genes have been developed. To the best of our knowledge, there is a lack of consensus on the most effective way of describing genes' interplay or the optimal network construction. Our analysis shows that with WGCNA, the proposed NSBoost may improve cancer marker selection. As the focus is on the development of NSBoost, a more comprehensive examination of its performance under different networks is beyond our scope. We adopt the AFT model to describe gene effects on survival. Compared with alternatives such as the Cox model, this model may have more lucid interpretation and significantly lower computational cost. Model diagnostics is not conducted as there is a lack of existing diagnostics tools for high-dimensional survival data. The satisfactory prediction performance may partly support the validity of this model. NSBoost can effectively account for the "module-gene" two level hierarchical structure, which is not the full information contained in the network.

WGCNA and other networks contain other information, for example the connectedness measure between two genes within the same modules. It may be possible to extend the proposed approach and accommodate the connectedness measure in marker selection. However as discussed above, with $n << d$, uniform estimation consistency of $\begin{pmatrix} d \\ 2 \end{pmatrix}$ connectedness measures is questionable. In contrast, the module structure can be much more reliable. Thus, we focus on the module structure in our research. The simulation settings considered in this study are simpler than what's encountered in practical data analysis. We intentionally choose such settings as they may actually favor simple approaches such as Enet and Boost. In data analysis, we conclude that NSBoost may be preferred as it identifies a smaller number of modules and genes and has superior prediction performance. Analysis of independent validation studies may be needed to fully confirm performance of NSBoost and identified markers.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors were involved in the study design and writing. SM and YH conducted numerical studies. All authors read and approved the final manuscript.

Acknowledgements

This study has been supported by NSF grant DMS-0904181, NIH grants LM009754, CA120988 and CA142774, and Research of Longitudinal Data Analysis Methodology and Its Application (2009JJD910002) from Key Research Institute of Humanities and Social Sciences Program and Fujian Social Science Fund (2011C042), China.

References

1. Knudsen S: *Cancer Diagnostics with DNA Microarray*. Hoboken, NJ: Wiley 2006.
2. Casci T: **Gene networks: Meaningful connections**. *Nature Reviews Genetics* 2010, **11**:172–173.
3. Langfelder P, Horvath S: **Eigengene networks for studying the relationships between co-expression modules**. *BMC System Biology* 2007, **1**:54.
4. Saris CG, Horvath S, van Vught PW, van Es MA, Blauw HM, Fuller TF, Langfelder P, DeYoung J, Wokke JH, Veldink JH, van den Berg LH, Ophoff RA: **Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients**. *BMC Genomics* 2009, **10**:405.
5. Ma S, Kosorok MR, Huang J, Dai Y: **Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis**. *BMC Medical Genomics* 2011, **4**:5.

6. Ma S, Shi M, Li Y, Yi D, Shia BC: **Incorporating gene co-expression network in identification of cancer prognosis markers**. *BMC Bioinformatics* 2010, **11**:271.

7. **WGCNA** [http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/].

8. **R Package WGCNA** [http://cran.r-project.org/web/packages/WGCNA/index.html].

9. Langfelder P, Zhang B, Horvath S: **Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R**. *Bioinformatics* 2008, **24**:719–720.

10. Wei LJ: **The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis**. *Statistics in Medicine* 1992, **11**:1871–1879.

11. Stute W: **Consistent estimation under random censorship when covariables are available**. *Journal of Multivariate Analysis* 1993, **45**:89–103.

12. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning*. Springer 2009.

13. Buhlmann P, Yu B: **Sparse boosting**. *Journal of Machine Learning Research* 2006, **7**:1001–1024.

14. Ma S, Song X, Huang J: **Supervised group Lasso with applications to microarray data analysis**. *BMC Bioinformatics* 2007, **8**:60.

15. Dettling M, Buhlmann P: **Boosting for tumor classification with gene expression data**. *Bioinformatics* 2003, **19**:1061–1069.

16. Huang Y, Huang J, Shia BC, Ma S: **Identification of cancer genomic markers via integrative sparse boosting**. *Biostatistics* 2011. in press.

17. Zou H, Hastie T: **Regularization and variable selection via the elastic net**. *JRSSB* 2005, **67**:301–320.

18. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT: **Gene expression predictors of breast cancer outcomes**. *Lancet* 2003, **361**:1590–1596.

19. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Rijn van de M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *PNAS* 2001, **98**:10869–10874.

20. Rosenwald A, Wright G, Wiestner A, Chan WC, Connors JM, Campo E, Gascoyne RD, Grogan TM, Muller-Hermelink HK, Smeland EB, Chiorazzi M, Giltnane JM, Hurt EM, Zhao H, Averett L, Henrickson S, Yang L, Powell J, Wilson WH, Jaffe ES, Simon R, Klausner RD, Montserrat E, Bosch F, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Fisher RI, Miller TP, LeBlanc M, Ott G, Kvaloy S, Holte H, Delabie J, Staudt LM: **The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma**. *Cancer Cell* 2003, **3**:185–197.

21. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink, K H, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, Lopez-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM: **The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma**. *The New England Journal of Medicine* 2002, **346**:1937–1947.

Tables

Table 1 - Simulation study: median number of genes (T) and true positives (TP) identified over 200 replicates. Correlation structure: auto-regressive (auto), banded, and compound symmetry (comp).

		Enet		Boost		SBoost		NBoost		NSBoost	
Structure	ρ	T	TP	T	TP	T	TP	T	TP	T	TP
auto	0.3	30	20	38	20	30	18	103	16	23	20
	0.7	23	20	33	20	23	16	91	13	23	20
banded	0.2	22	20	34	20	22	15	80	12	21	18
	0.33	26	20	34	20	24	16	90	14	24	20
comp	0.3	33	20	49	20	27	15	102	14	27	20
	0.7	36	20	47	20	18	12	76	10	20	17

Table 2 - Description of datasets. Gene/Sample: number of genes/subjects profiled.

Data	Disease	Platform	Gene	Sample
D1: Huang et al. (2003)	Breast cancer	Affymetrix	12,625	71
D2: Sorlie et al. (2001)	Breast cancer	cDNA	8,102	58
D3: Rosenwald et al. (2003)	MCL	cDNA	8,810	92
D4: Rosendwald et al. (2002)	DLBCL	cDNA	7,399	240

Table 3 - Data analysis results. Gene: number of genes identified; Overlap: number of overlapped genes with NSBoost; Module: number of modules identified; logrank: prediction logrank statistic.

		D1	D2	D3	D4
Enet	gene	29	39	82	60
	overlap	2	3	6	0
	module	4	3	5	2
	logrank	0.089	8.931	3.405	5.629
Boost	gene	70	74	17	12
	overlap	5	4	1	0
	module	4	4	3	4
	logrank	1.704	2.478	1.642	7.976
SBoost	gene	31	26	22	12
	overlap	3	1	1	0
	module	3	2	4	2
	logrank	0.063	0.128	5.961	6.662
NBoost	gene	102	91	44	35
	overlap	23	21	13	14
	module	3	2	5	2
	logrank	0.266	0.318	8.996	17.015
NSBoost	gene	31	30	21	22
	module	2	1	1	1
	logrank	2.863	11.504	15.613	18.937

Figures

Figure 1 — Parameter path of NSBoost: estimates as a function of number of iterations. The upper four panels correspond to four modules in Step 1 of boosting. The lower panel corresponds to four super markers in Step 2 of boosting. Different colors correspond to different modules.

Figure 2 — Parameter path of NBoost: estimates as a function of number of iterations. The upper four panels correspond to four modules in Step 1 of boosting. The lower panel corresponds to four super markers in Step 2 of boosting. Different colors correspond to different modules.

Figure 3 — Analysis of data D4: network modules constructed using WGCNA.