Published in final edited form as:

J Am Stat Assoc. 2016 ; 111(516): 1427–1439. doi:10.1080/01621459.2016.1164051.

Hierarchical Feature Selection Incorporating Known and Novel Biological Information: Identifying Genomic Features Related to Prostate Cancer Recurrence

Yize Zhao,

Postdoctoral Fellow, Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709

Matthias Chung,

Assistant Professor, Department of Mathematics, Virginia Tech, Blacksburg, VA 24061

Brent A. Johnson,

Associate Professor, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642

Carlos S. Moreno, and

Associate Professor, Department of Pathology and Laboratory Medicine

Qi Long*

Associate Professor, Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322

Abstract

Our work is motivated by a prostate cancer study aimed at identifying mRNA and miRNA biomarkers that are predictive of cancer recurrence after prostatectomy. It has been shown in the literature that incorporating known biological information on pathway memberships and interactions among biomarkers improves feature selection of high-dimensional biomarkers in relation to disease risk. Biological information is often represented by graphs or networks, in which biomarkers are represented by nodes and interactions among them are represented by edges; however, biological information is often not fully known. For example, the role of microRNAs (miRNAs) in regulating gene expression is not fully understood and the miRNA regulatory network is not fully established, in which case new strategies are needed for feature selection. To this end, we treat unknown biological information as missing data (i.e., missing edges in graphs), different from commonly encountered missing data problems where variable values are missing. We propose a new concept of imputing unknown biological information based on observed data and define the imputed information as the novel biological information. In addition, we propose a hierarchical group penalty to encourage sparsity and feature selection at both the pathway level and the within-pathway level, which, combined with the imputation step, allows for incorporation

Correspondence should be addressed to qlong@emory.edu.

Supplementary Materials

The supplementary materials contain a) exploratory analyses of the prostate cancer data for assessing modeling assumptions and b) the detailed computational implementation of the L-BFGS algorithm in Section 2.5. The proposed method is implemented in Matlab and is available upon request.

of known and novel biological information. While it is applicable to general regression settings, we develop and investigate the proposed approach in the context of semiparametric accelerated failure time models motivated by our data example. Data application and simulation studies show that incorporation of novel biological information improves performance in risk prediction and feature selection and the proposed penalty outperforms the extensions of several existing penalties.

Keywords

Hierarchical Feature selection; Imputation; Missing data; Regularization; Semiparametric AFT model

1 Introduction

In many biomedical studies, it is of great interest to conduct feature selection of high-dimensional data such as genomic biomarkers in relation to disease risk and prognosis while also incorporating biological information on pathway memberships and interactions among biomarkers, which has been shown in previous works (Li and Li, 2008; Pan et al., 2010) to improve performance in feature selection and risk prediction. Biological information is often represented by graphs or networks, in which biomarkers are represented by nodes and interactions among them are represented by edges; e.g. Table 1 and Figure 1 in Kanehisa et al. (2002). However, interactions among the biomarkers of interest are often not fully known, in which case new strategies are needed to identify biomarkers that are predictive of disease risk and prognosis.

1.1 Motivating Data

Our work is motivated by a recent study of genomic biomarkers in relation to risk for prostate cancer recurrence after prostatectomy (Long et al., 2011). One of the important challenges in current prostate cancer research is to develop effective methods such as biomarker assays to determine whether a patient is likely to progress to aggressive, metastatic disease or have cancer recurrence following a treatment such as surgery in order to aid clinicians in deciding on the appropriate course of treatment, e.g., whether immediate adjuvant therapy is warranted after prostatectomy. In particular, biomarker assays that could predict progression, metastasis and recurrence for prostate cancer patients would be of great utility in aiding clinical management of this large patient population (Nakagawa et al., 2008; Cuzick et al., 2011). This study includes a training set of 73 patients and an independent testing set of 40 patients with each patient undergoing radical prostatectomy after a diagnosis of prostate cancer. RNA samples were isolated from formalin-fixed paraffin-embedded (FFPE) radical prostatectomy specimens that were obtained from these patients. Using these RNA samples, customized cDNA-mediated Annealing, Selection, extension and Ligation (DASL) assay (Abramovitz et al., 2008) was conducted to measure expression profiles of 522 prostate cancer related genes that were selected from literature and the miRNA Expression Profiling Panel (Illumina, Inc.) designed for the DASL assay was used to determine expression of 403 human miRNAs that were present on the miRNA microarrays used for both the training data and the testing data. Four clinical variables prostate-specific antigen (PSA) levels, histological grade (Gleason Score), tumor stage (T-

stage) and age as well as the long-term clinical outcome on cancer recurrence were collected for all patients. The primary goal of this study is to identify mRNA and miRNA biomarkers that are predictive of prostate cancer recurrence after prostatectomy. Earlier analyses of the data (Long et al., 2011) used the l1-regularized (lasso) Cox proportional hazards model (Tibshirani, 1997) and identified two prediction models – one using only mRNA biomarkers and the other using both mRNA and miRNA biomarkers – for prostate cancer recurrence, which did not incorporate biological information or conduct hierarchical feature selection.

To fix ideas, for subject i ($i=1, \dots, n$), we denote by Y_i the time to prostate cancer recurrence (i.e., the endpoint of interest) which is subject to censoring, by $\mathbf{X}_i=(X_1, \dots, X_p)^T$ the p clinical variable. We also denote by $\mathbf{Z}_i=(Z_i^{(1)}, \dots, Z_i^{(d)})^T$ the set of d mRNA biomarkers, and by $\mathbf{W}_i=(W_i^{(1)}, \dots, W_i^{(d')})^T$ the set of d' miRNA biomarkers with $d > n$ and $d' > n$. Let $\mathbf{Z}=(\mathbf{Z}_1, \dots, \mathbf{Z}_n)^T$ and $\mathbf{W}=(\mathbf{W}_1, \dots, \mathbf{W}_n)^T$ represent the collection of mRNA and miRNA expression data, respectively, for all n subjects. Our goal is to identify genomic biomarkers in $\mathbf{Z}_i$ and $\mathbf{W}_i$ that are predictive of Y_i , while incorporating known and novel biological information on $\mathbf{Z}_i$ and $\mathbf{W}_i$. As discussed in Section 1.2, there are some important differences between information on $\mathbf{Z}_i$ (mRNA) and information on $\mathbf{W}_i$ (miRNA), which needs to be accounted for as well.

1.2 Incorporating Biological Information

Biomedical researchers have made enormous strides in discovering biological functions of genes in particular for those that are involved in complex diseases such as cancer; in addition, genes have been shown to function through networks or pathways. Pathway analysis, including topology-based methods (Shojaie and Michailidis, 2009, 2010) and integration analysis of multi-type data sources (Jauhiainen et al., 2012), has received substantial interest and played a vital role in the enrichment of underlying biological knowledge; Khatri et al. (2012) provides a thorough review of the current approaches for pathway analysis and their limitations. In addition, major efforts have been made to structure and store the resulting biological information on pathways or networks in multiple public or commercial databases, including Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), Gene Ontology (GO) (Ashburner et al., 2000), BioCarta (www.biocarta.com), and Cell Signaling Technology (CST) Pathway (www.cellsignal.com). Since such prior biological information sheds important insights on regulatory relationships between genes and/or gene products that are associated with the genesis or progression of a disease, it is very valuable to incorporate such information into analysis of gene expression data in relation to disease risk; for example, Costello et al. (2014) showed that incorporation of biological pathway information improved drug sensitivity prediction. More recently, non-coding RNAs (ncRNAs) – functional transcripts that do not code for proteins – have been identified as important regulators of gene expression in many eukaryotes. Of all ncRNAs, microRNAs (miRNAs) are the most widely studied and characterized with more than 600 human miRNAs identified; it has been suggested that miRNAs regulate more than 30% of the human genes (Lewis et al., 2005). For example, miRNAs have been shown to down-regulate gene expressions in various cancers, suggesting a role for miRNAs in tumorigenesis (Bagga et al., 2005; Pillai et al., 2005; Catto et al., 2011), and they can act as both tumor

suppressor genes and oncogenes. Consequently, it has been recognized that combining information on ncRNAs such as miRNAs with other types of biomarkers may improve cancer risk assessment, detection, and prognosis (Tricoli and Jacobson, 2007; Slack and Weidhaas, 2008; Schoof et al., 2012). However, unlike the prior-knowledge pathways for protein-coding genes such as KEGG pathways that can be readily obtained from existing databases, the scope of miRNA's roles in regulating gene expression is not fully understood and the complete picture of the miRNA regulatory network is still being established (Thermann and Hentze, 2007). Based on specific sequence and structural features, several algorithms have been developed to predict gene sequences that are potential regulatory targets of miRNAs, e.g., miRanda (Betel et al., 2008), PITA (Kertesz et al., 2007), TargetScan (Grimson et al., 2007), and PicTar (Krek et al., 2005). However, the vast majority of these predicted targets have not been experimentally validated and may be false positives. As a result, new strategies are needed to incorporate miRNAs and their regulatory network information in modeling cancer risk and prognosis.

Specifically for our data example, we denote by $\mathcal{G}_1$ the graph on the known biological pathways for $\mathbf{Z}_j$ (mRNA), which can be obtained from the KEGG pathway database. $\mathcal{G}_1$ includes two levels of biological information: 1) whether a gene belongs to pathway g ($g = 1, \dots, G$) which is known as information on pathway (or group) memberships, and 2) whether two genes interact with one another in one pathway which is known as information on gene-gene interactions, represented by edges connecting two genes in the graph. Thus, the biological information in $\mathcal{G}_1$ is inherently hierarchical. In addition, we denote by $\mathcal{G}_2$ the graph for the miRNA regulatory network (the biological information for $\mathbf{W}_j$), where an edge between miRNA j and mRNA k indicates that miRNA j regulates the expression of gene k . Since $\mathcal{G}_2$ is not well-established, we partition $\mathcal{G}_2$ into two components ($\mathcal{G}_2 = \mathcal{G}_2^o \cup \mathcal{G}_2^m$), the experimentally validated component denoted by $\mathcal{G}_2^o$ and the unknown component denoted by $\mathcal{G}_2^m$. In all graphs ($\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$), $j \sim k$ denotes that biomarkers (nodes) j and k are neighbors, i.e., connected by an edge representing either a coordinating or regulatory relationship.

We propose to obtain $\mathcal{G}_2^o$ from an existing database on validated miRNA targets such as TarBase (Vergoulis et al., 2012) and treat $\mathcal{G}_2^m$ as missing data. Of note, different from commonly encountered missing data problems, the missing data here are the biological information in the network structure (i.e., edges in a graph) instead of variable values. Thus, we propose a new concept of multiply imputing the missing biological information (i.e., the missing edges in the graph $\mathcal{G}_2$) under the assumption of missing at random (MAR). Multiple imputation (MI) (Rubin, 1987; Little and Rubin, 2002), arguably the most widely used technique for handling missing data, has been investigated extensively and Harel and Zhou (2007) provides a nice review of theory, implementation and software for MI. One advantage of handling missing data via MI is that the imputation step and the subsequent analysis step are separate and can employ different techniques. While MI can be conducted through either Bayesian or frequentist's approaches (Little and Rubin, 2002), it is known to be inherently Bayesian and Rubin recommends that imputations be created through a Bayesian process (Rubin, 1987; Harel and Zhou, 2007). In the setting of our interest, we propose to impute the missing edges in the graph based on observed data while also borrowing prior information from existing databases on predicted miRNA targets, noting that none of the existing MI

methods are directly applicable in this setting. We propose to impute $\mathcal{G}_2^m$ by adapting a directed graphical model proposed by Stingo et al. (2010) for inferring the miRNA regulatory network. The imputed $\mathcal{G}_2^m$ is considered *the novel biological information* and the imputation process is essentially an integrated de novo discovery of the miRNA regulatory network.

1.3 Network-based Regularized Regression

To deal with high-dimensional data such as our data example (i.e., $d > n$ and $d' > n$), penalized regression has been a popular approach and various types of penalties have been developed over the years including Lasso (Tibshirani, 1996), Smoothly Clipped Absolute Deviation (Fan and Li, 2001), Least Angle Regression (Efron et al., 2004), elastic net (Zou and Hastie, 2005) and adaptive lasso (Zou, 2006). More recently Dicker et al. (2012) and Shen et al. (2012) investigated variable selection methods for ultra high-dimensional data and Loh and Wainwright (2011) investigated variable selection in the presence of noisy and missing data. While these methods successfully realize simultaneous variable selection and parameter estimation, they are not well-suited for analysis of genomic data as they fail to incorporate biological information among genes, or as in our problem, among mRNAs and miRNAs. Alternatively, regularization methods such as the network-constrained regularization (Li and Li, 2008) and the grouped penalty (Pan et al., 2010) have been developed to utilize known network information. In addition, a number of regularization methods including fused lasso (Tibshirani et al., 2005), group lasso (Yuan and Lin, 2006) and its extension sparse group lasso (Simon et al., 2013) as well as other penalties for grouped variables (Wang et al., 2009; Wu and Wang, 2013) were not developed specifically for analysis of genomic data but, nonetheless, can be used to incorporate biological information.

However, none of these existing regularization methods can be used to incorporate both levels of biological information on pathway memberships and on gene-gene and gene-miRNA interactions; fused lasso, grouped penalty and network-constrained regularization can only incorporate information on gene-gene and gene-miRNA interactions, and penalties for grouped variables (Yuan and Lin, 2006; Wang et al., 2009; Wu and Wang, 2013) can only incorporate information on pathway (or group) memberships. Since the biological information contained in $\mathcal{G}_1$ is inherently hierarchical, it is likely advantageous to account for both types of information in analysis. To this end, we propose a novel penalty, hierarchical grouped penalty (HGP), to induce sparsity at both the pathway level and the within-pathway level and, more importantly, allow for incorporation of biological information at both levels as well. HGP, to the best of our knowledge, has not been investigated in the literature.

During the peer-review process, an anonymous reviewer brought to our attention another paper by Chekouo et al. (2015), which addresses a similar problem. Our work was developed independently of Chekouo et al. (2015) and differs in several important respects. First, their approach does not allow for hierarchical feature selection. Second, their approach builds on a parametric AFT model with normal errors, which is less robust than the semiparametric AFT model in our approach. Third, while they use the gene pathway

information for inferring the miRNA regulatory network, they incorporate only the miRNA regulatory network information in variable selection for the survival outcome model, whereas we use both the gene pathway information and the miRNA regulatory network information in variable selection for the survival outcome model. Lastly, we formalize the concept of “imputing” the component of the miRNA regulatory network that is not validated, whereas they do not make such distinction between putative and validated miRNA targets.

The remainder of this article is organized as follows. In Section 2, we first extend several existing regularization methods to the setting of interest and define the novel regularization method, hierarchical grouped penalty. We then define the new concept of multiply imputing the unknown components of the miRNA regulatory network and subsequently incorporating the resulting novel biological information into feature selection. In Section 3, we analyze the motivating data example using the proposed approach combined with our proposed penalty and the extensions of several existing penalties, with or without incorporating the novel biological information. In Section 4, we conduct simulation studies to evaluate the performance of the proposed penalty in comparison with the existing penalties. We conclude with some discussion remarks in Section 5.

2 Methodology

Our proposed approach is applicable to general regression problems that use a loss function (or an objective function) to obtain parameter estimates. To avoid overly restrictive modeling assumptions, we adopt the semiparametric accelerated failure time (AFT) model (Cox and Oakes, 1984; Kalbfleisch and Prentice, 2002) for our data analysis.

2.1 Regularized Estimation for Semiparametric AFT Models

In the analysis of censored data, the AFT model is a powerful tool due to its direct physical interpretation (Reid, 1994) and may be preferable to the Cox proportional hazards model (Cox, 1972) when risk prediction is the main objective. In addition, unlike parametric AFT models, semiparametric AFT models combined with rank-based estimation (Tsiatis, 1990; Ying, 1993) avoid strong parametric assumptions and are robust to model mis-specifications, a desirable property for use in practice.

Let T_i denote the log-transformed time to prostate cancer recurrence (i.e., $T_i = \log(Y_i)$) which is subject to censoring. Let $\tilde{T}_i = \min(T_i, C_i)$ denote the observed log-transformed follow-up time, C_i denote the log-transformed censoring time, and $\delta_i = I(T_i \leq C_i)$ denote the indicator for whether the survival time is observed. It follows that the observed data are $\{(\tilde{T}_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i), i=1, \dots, n\}$. Throughout, we assume that C_i is independent of T_i conditional on the set of observed variables. The semiparametric AFT model can be written as follows

$$T_i = \alpha^T \mathbf{X}_i + \vartheta^T \mathbf{Z}_i + \varphi^T \mathbf{W}_i + \varepsilon_i, \quad (1)$$

where the random errors (ε_i) are independently and identically distributed (i.i.d.) and follow some completely unspecified distribution F_ε . This represents another major difference of our approach, in particular, compared to the Bayesian approach for hierarchical feature selection by Stingo et al. (2011) which assumes that ε_i follows a distribution of some known parametric form, say, normal distribution.

Given the semiparametric AFT model (1), our goal is to estimate $\alpha=(\alpha_1, \dots, \alpha_p)^T$, $\vartheta=(\vartheta_1, \dots, \vartheta_d)^T$ and $\phi=(\phi_1, \dots, \phi_d)^T$ in (1) and conduct feature selection among mRNA ($\mathbf{Z}_i$) and miRNA ($\mathbf{W}_i$) biomarkers while incorporating the known biological information ($\mathcal{G}_1$ and $\mathcal{G}_2^o$) and the novel information on the miRNA regulatory network ($\mathcal{G}_2^m$) in the case of $d > n$ and $d' > n$.

If both $\mathcal{G}_1$ and $\mathcal{G}_2$ are fully known, we propose to estimate the three sets of coefficients (α , ϑ and ϕ) in Model (1) using the general approach of regularization

$$(\hat{\alpha}, \hat{\vartheta}, \hat{\phi}) = \min_{\alpha, \vartheta, \phi} \mathcal{L}_n(\alpha, \vartheta, \phi; \tilde{T}_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) + J(\vartheta, \phi), \quad (2)$$

where $\mathcal{L}_n(\alpha, \vartheta, \phi; \tilde{T}_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$ is the loss function based on the observed data and $J(\vartheta, \phi)$ is the regularization term that allows for feature selection in high-dimensional biomarkers ($\mathbf{Z}$ and $\mathbf{W}$) while also incorporating biological information in $\mathcal{G}_1$ and $\mathcal{G}_2$. We perform rank-based inference in the AFT model by adopting Gehan's loss function (Gehan, 1965; Jin et al., 2003)

$$\mathcal{L}_n(\alpha, \vartheta, \phi; \tilde{T}_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \delta_i (e_i - e_j)_-, \quad (3)$$

where $e_i = \tilde{T}_i - \alpha^T \mathbf{X}_i - \vartheta^T \mathbf{Z}_i - \phi^T \mathbf{W}_i$ and $c_- = \max(0, -c)$. This may be regarded as the weighted average of pairwise differences in the residuals and the censored data analogue of Wilcoxon's loss.

As to the choice of $J(\vartheta, \phi)$, we first consider extensions of several existing penalties to our setting, namely,

$$\text{sparse group lasso (sGL): } \lambda_1 \sum_{g=1}^G \|\phi_g\|_{\Omega_g} + \lambda_2 \{\|\vartheta\|_1 + \|\phi\|_1\}, \quad (4)$$

$$\text{generalized fused lasso (FL): } \lambda_1 (\|\vartheta\|_1 + \|\phi\|_1) + \lambda_2 \left(\sum_{j \sim k \in \mathcal{G}_1} |\vartheta_j - \vartheta_k| + \sum_{j \sim k \in \mathcal{G}_2} |\vartheta_j - \phi_k| \right), \quad (5)$$

$$\text{network- constrained regularization (NC): } \lambda_1 (\|\boldsymbol{\vartheta}\|_1 + \|\boldsymbol{\varphi}\|_1) + \lambda_2 \left\{ \sum_{j \sim k \in \mathcal{G}_1} (\vartheta_j - \vartheta_k)^2 + \sum_{j \sim k \in \mathcal{G}_2} (\vartheta_j - \varphi_k)^2 \right\}, \quad (6)$$

$$\text{grouped penalty (GP): } \lambda_1 \left\{ \sum_{j \sim k \in \mathcal{G}_1} (|\vartheta_j|^r + |\vartheta_k|^r)^{1/r} + \sum_{j \sim k \in \mathcal{G}_2} (|\vartheta_j|^r + |\varphi_k|^r)^{1/r} \right\}, \quad (7)$$

where λ_1 and λ_2 are tuning parameters, $\|\cdot\|_1$ is the l_1 norm, $\boldsymbol{\phi}_g$ is the set of regression coefficients for biomarkers in group g ($g=1, \dots, G$), $\|\boldsymbol{\phi}_g\|_{\Omega_g} = (\boldsymbol{\phi}_g^T \boldsymbol{\Omega}_g \boldsymbol{\phi}_g)^{1/2}$ with $\boldsymbol{\Omega}_g$ as the weight matrix, and $r > 1$. Group is defined based on both $\mathcal{G}_1$ and $\mathcal{G}_2$; each group either includes a pathway in $\mathcal{G}_1$ and miRNAs that regulate the genes in this pathway or includes an mRNA/miRNA that is not connected to any other mRNA or miRNA in $\mathcal{G}_1$ and $\mathcal{G}_2$. $\boldsymbol{\phi}_g$'s may overlap with one another since one gene may be part of multiple pathways. It is straightforward to also consider the weighted versions of these penalties as appropriate that assign biologically meaningful weights to each edge ($j \sim k$) or each node in $\mathcal{G}_1$ and $\mathcal{G}_2$ (Li and Li, 2008; Pan et al., 2010). In addition, the generalized fused lasso defined in (5) is different from the original fused lasso (Tibshirani et al., 2005) in that the neighbors are defined based on $\mathcal{G}$ not a linear ordering of variables. Of note, sGL allows for the incorporation of the biological information on pathway memberships in $\mathcal{G}_1$ through its first term; FL, NC, and GP allow for the incorporation of the biological information on gene-gene and gene-miRNA interactions in both $\mathcal{G}_1$ and $\mathcal{G}_2$ through the term that involves $j \sim k$. However, none of them allow for the incorporation of both types of biological information.

2.2 Hierarchical Grouped Penalty

As noted in Section 1.2, it is desirable to incorporate the biological information in both $\mathcal{G}_1$ and $\mathcal{G}_2$, assuming that both are fully observed. To this end, we propose a new penalty, hierarchical grouped penalty (HGP), as follows

$$\lambda_1 \sum_{g=1}^G \|\boldsymbol{\phi}_g\|_{\Omega_g} + \lambda_2 \left\{ \sum_{j \sim k \in \mathcal{G}_1} (|\vartheta_j|^r + |\vartheta_k|^r)^{1/r} + \sum_{j \sim k \in \mathcal{G}_2} (|\vartheta_j|^r + |\varphi_k|^r)^{1/r} \right\}, \quad r > 1 \quad (8)$$

where λ_1 and λ_2 are tuning parameters and $\boldsymbol{\phi}_g$ is defined in (4). As an alternative, we can also assign two different tuning parameters to the second and the third terms in (8), allowing for different regularization effects from $\mathcal{G}_1$ and $\mathcal{G}_2$; this approach, however, is considerably more expensive in terms of computation and is not considered in the current work. Noticing that in (8) the first term is analogous to the group lasso and the second and the third terms are essentially a L_r -norm of vector $(\boldsymbol{\vartheta}_j, \boldsymbol{\vartheta}_k)'$ and $(\boldsymbol{\vartheta}_j, \boldsymbol{\phi}_k)'$ respectively, it is straightforward to show that the HGP (8) is convex for any real $r > 1$.

The first term in HGP (8), similar to the group lasso penalty (Yuan and Lin, 2006), can be used to incorporate the information on pathway memberships in $\mathcal{G}_1$ and encourage feature selection and sparsity at the general group level. In addition, the specification of Ω_g can be used to incorporate the strength of each gene-gene interaction or gene-miRNA regulation, e.g., using their correlations. The second and third terms in HGP (8), similar to the group penalty (Pan et al., 2010), can be used to incorporate the information on gene-gene and gene-miRNA interactions and allow for feature selection at the within-pathway (or biomarker) level. Specifically, the second term accounts for interactions between mRNAs and the third term accounts for interactions between mRNAs and miRNAs. Compared with other penalties, HGP achieves hierarchical selection and allows for incorporation of two levels of biological information on pathway memberships and on gene-gene and gene-miRNA interactions, likely leading to better performance. For example, in settings where two pathways overlap and one pathway is associated with the outcome whereas the other is not, HGP is expected to perform better as a result of its hierarchical nature.

2.3 Imputation Model for Unknown $\mathcal{G}_2^m$

As noted in Section 1.2, $\mathcal{G}_2$ is not fully observed in practice and it contains the known component $\mathcal{G}_2^o$ and the unknown component $\mathcal{G}_2^m$. We handle this problem from the perspective of handling missing data: consider $\mathcal{G}_2^m$ as the missing component of the data and conduct multiple imputation. Again, the missing data here are the network structure (i.e., edges in a graph) not variable values and none of the existing MI methods are directly applicable in this setting. Our key idea is to replace the missing component ($\mathcal{G}_2^m$) with a set of “plausible” networks using their predictive distributions conditional on the observed data ($\mathbf{Z}$ and $\mathbf{W}$) and the observed component ($\mathcal{G}_2^o$), resulting in multiply imputed data sets. Specifically, under the missing at random (MAR) assumption (Little and Rubin, 2002; Rubin, 1987), we have in our setting

$$f(\mathcal{G}_2^m | \mathbf{Z}, \mathbf{W}, \mathcal{G}_2^o) = \int f(\mathcal{G}_2^m | \mathbf{Z}, \mathbf{W}, \mathcal{G}_2^o, \boldsymbol{\beta}) f(\boldsymbol{\beta} | \mathbf{Z}, \mathbf{W}, \mathcal{G}_2^o) d\boldsymbol{\beta}; \quad (9)$$

$$f(\boldsymbol{\beta} | \mathbf{Z}, \mathbf{W}, \mathcal{G}_2^o) \propto f(\boldsymbol{\beta}) \int f(\mathcal{G}_2^m, \mathcal{G}_2^o, \mathbf{Z}, \mathbf{W} | \boldsymbol{\beta}) d\nu(\mathcal{G}_2^m), \quad (10)$$

where $\boldsymbol{\beta}$ is the unknown parameter associated with the distribution of $\mathcal{G}_2^m$ and $f(\boldsymbol{\beta})$ is the prior for $\boldsymbol{\beta}$. Based on (9) and (10), multiple imputation of $\mathcal{G}_2^m$ can be implemented by repeating the following sequential procedure H times: first obtains a posterior sample for $\boldsymbol{\beta}$ denoted by $\hat{\boldsymbol{\beta}}$, and then impute $\mathcal{G}_2^m$ by a random draw from the conditional predictive distribution $f(\mathcal{G}_2^m | \mathbf{Z}, \mathbf{W}, \mathcal{G}_2^o, \hat{\boldsymbol{\beta}})$. Our imputation approach is adapted from the directed graphical modeling approach proposed by Stingo et al. (2010) and we refer readers to their work for more details regarding the model formulation and prior specifications.

Let $\mathbf{Z}^{(j)} = (Z_1^{(j)}, \dots, Z_n^{(j)})^T$ denote the gene expression data for gene $j=1, \dots, d$. Since miRNAs are known to down-regulate expressions of their target genes, we use the following imputation model for the missing component of the miRNA regulatory network based on the observed mRNA and miRNA expression data,

$$\mathbf{Z}^{(j)} = -\mathbf{W}\boldsymbol{\beta}^{(j)} + \boldsymbol{\varepsilon}^{(j)} \quad j=1, \dots, d, \quad (11)$$

where $\boldsymbol{\beta}^{(j)} = (\beta_1^{(j)}, \dots, \beta_{d'}^{(j)})^T$ are a set of unknown non-negative parameters with each element $-\beta_k^{(j)}$ representing the regulatory effect of miRNA k on gene j . The random error $\boldsymbol{\varepsilon}^{(j)}$ is assumed to be independently distributed and follows $N(0, \sigma_j^2 \mathbf{I}_n)$ where $\mathbf{I}_n$ is the identity matrix of dimension $n \times n$.

Based on Model (11), the unknown component of the miRNA regulatory network $\mathcal{G}_2^m$ is contained in a matrix $\boldsymbol{\Psi}_{d \times d'}$ with the (j, k) -th element defined as $\Psi_{j,k} = I[\beta_k^{(j)} \neq 0]$, where $I(\mathcal{A}) = 1$ if event $\mathcal{A}$ is true and 0 otherwise. In other words, imputing $\mathcal{G}_2^m$ is equivalent to imputing $\boldsymbol{\Psi}_{d \times d'}$. $\boldsymbol{\Psi}_{d \times d'}$ may also contain information in the validated component $\mathcal{G}_2^o$. It follows that imputing $\mathcal{G}_2^m$ becomes a problem of variable selection for Model (11). With the goal to obtain the predictive posterior distribution of $\boldsymbol{\Psi}_{d \times d'}$ for the imputation procedure in (9), we adopt the Bayesian approach proposed by Stingo et al. (2010); this approach has a natural advantage of producing valid posterior samples through Markov chain Monte Carlo (MCMC), noting that imputation is inherently Bayesian. Following Stingo et al. (2010), given the latent indicator $\boldsymbol{\Psi}_{j,k}$, we impose a mixture prior distribution for $\beta_k^{(j)}$,

$$p(\beta_k^{(j)} | \sigma_j^2, \Psi_{j,k}) \sim \Psi_{j,k} \text{Ga}(1, \kappa \sigma_j^2) + (1 - \Psi_{j,k}) \delta(0), \quad (12)$$

and a hyper-prior $\sigma_j^2 \sim \text{IG}(\frac{\alpha + d'}{2}, \gamma)$, for $j=1, \dots, d$ and $k=1, \dots, d'$, where $\delta(0)$ is a point mass at 0, $\text{Ga}(a, b)$ and $\text{IG}(a, b)$ denote the Gamma and inverse Gamma distribution with shape parameter a and scale parameter b , respectively, and κ, α , and γ are parameters that are pre-specified. Given prior (12), the variable selection procedure is implemented through the posterior inference of the binary indicators $\Psi_{j,k}$. Again, following Stingo et al. (2010), we incorporate known biological information on putative miRNA targets into the prior for $\boldsymbol{\Psi}$ to guide the posterior inference for $\boldsymbol{\Psi}$, or equivalently, guide the imputation for $\mathcal{G}_2^m$. For example, we can use information from the existing databases such as miRanda (Betel et al., 2008), PITA (Kertesz et al., 2007), TargetScan (Grimson et al., 2007), and PicTar (Krek et al., 2005) to specify the prior for $\boldsymbol{\Psi}$ as $\text{logit}(p(\Psi_{j,k} = 1)) = \eta + \boldsymbol{\tau}^T \mathbf{s}_{j,k}$ with hyper-parameters $(\eta, \boldsymbol{\tau})$ where $\mathbf{s}_{j,k} = (s_{j,k}^1, s_{j,k}^2, s_{j,k}^3, s_{j,k}^4)$ are the genetic scores from the four databases. $\boldsymbol{\tau}$ is assigned a hyper-prior, $\text{Ga}(\mathbf{a}, \mathbf{b}_{\boldsymbol{\tau}})$. The posterior inference for Model (11) results in the

posterior samples for Ψ and hence the posterior samples for $\mathcal{G}_2^m$, which are used to impute $\mathcal{G}_2^m$ and are considered the *novel biological information*.

2.4 Feature Selection in the Presence of Unknown $\mathcal{G}_2^m$

In the presence of unknown $\mathcal{G}_2^m$, we propose a two-stage approach for feature selection in Model (1). The approach treats $\mathcal{G}_2^m$ as missing data and uses Model (11) as the imputation model for $\mathcal{G}_2^m$. Similar to Ψ , let $\mathcal{G}_2$, $\mathcal{G}_2^o$, and $\mathcal{G}_2^m$ also denote the corresponding $d \times d'$ matrices that represent the edge information in the graph $\mathcal{G}_2$, $\mathcal{G}_2^o$, and $\mathcal{G}_2^m$; specifically, the (j, k) -th component of the matrix is 1 if $j \sim k$ in the graph and 0 otherwise.

In the first stage, we conduct multiple imputation for $\mathcal{G}_2^m$ and hence $\mathcal{G}_2$ following the strategy in Section 2.3. Specifically, we generate a set of posterior samples for Ψ , namely, $\{\Psi^{(1)}, \dots, \Psi^{(H)}\}$, by randomly drawing H times from its posterior distribution.

Subsequently, we generate multiply imputed $\mathcal{G}_2^m$ as follows: the (j, k) -th component of the h -

th imputed $\mathcal{G}_2^m$ is $\hat{\mathcal{G}}_2^{m(h)}(j, k) = \max \left\{ \hat{\Psi}^{(h)}(j, k) - \mathcal{G}_2^o(j, k), \emptyset \right\}$ ($h=1, \dots, H$). It follows that the (j, k) -th component of the h -th imputed $\mathcal{G}_2$ is

$\hat{\mathcal{G}}_2^{(h)}(j, k) = \max \left\{ \hat{\Psi}^{(h)}(j, k), \mathcal{G}_2^o(j, k) \right\}$. After multiple imputation, $\hat{\mathcal{G}}_2^{(h)}$ contains both the

known biological information in $\mathcal{G}_2^o$ and the novel biological information in $\hat{\mathcal{G}}_2^{m(h)}$. We denote by $\mathcal{F} = (\mathcal{O}^{(1)}, \dots, \mathcal{O}^{(H)})$ the collection of H imputed data sets where

$\hat{\mathcal{O}}^{(h)} = \left\{ (\tilde{T}_i, \delta_i, \mathbf{Z}_i, \mathbf{W}_i)_{i=1}^n, \mathcal{G}_1, \mathcal{G}_2^o, \hat{\mathcal{G}}_2^{m(h)} \right\}$. Again, it is widely accepted in the missing data literature and in practice to conduct MI using a Bayesian modeling approach, e.g., the mice package in R (van Buuren and Groothuis-Oudshoorn, 2011), and then analyze the imputed data using a frequentist approach.

In the second stage, we apply the regularization methods proposed in Sections 2.1 and 2.2 to fit Model (1) using each imputed data set $\mathcal{O}^{(h)}$ in $\mathcal{F}$ and obtain the regularized parameter estimates, denoted by $\hat{\boldsymbol{\theta}}^{(h)}$ and $\hat{\boldsymbol{\phi}}^{(h)}$. Similar to the standard multiple imputation, the final results need to be summarized from the empirical distributions of parameter estimates, namely, $\{\hat{\boldsymbol{\theta}}^{(h)}, h=1, \dots, H\}$ and $\{\hat{\boldsymbol{\phi}}^{(h)}, h=1, \dots, H\}$. However, we cannot directly apply Rubin's rule (Little and Rubin, 2002; Rubin, 1987) in our setting. In particular, the mean of

the parameter estimates over H imputed data sets (e.g., $\frac{1}{H} \sum_h \hat{\boldsymbol{\theta}}^{(h)}$) does not automatically lead to variable selection except for the parameters whose estimates are 0 in all H imputed data sets. As a result, we first calculate the proportion of each parameter being selected in the H imputed data as follows

$$\pi(\vartheta_j) = \frac{1}{H} \sum_{h=1}^H I(\hat{\vartheta}_j^{(h)} \neq 0), \quad j=1, \dots, d, \quad (13)$$

$$\pi(\varphi_j) = \frac{1}{H} \sum_{h=1}^H I(\hat{\varphi}_j^{(h)} \neq 0), j=1, \dots, d'. \quad (14)$$

Given $\pi(\vartheta_j)$ and $\pi(\phi_j)$, we define the final set of selected variables using a cutoff π_0 as follows,

$$S_{\pi_0} = \{\vartheta_j : \pi(\vartheta_j) > \pi_0; \phi_j : \pi(\phi_j) > \pi_0\}. \quad (15)$$

where $|S_{\pi_0}|$ denotes the cardinality of the set S_{π_0} and π_0 is a user-defined parameter that is pre-specified, say at 0.5. After variable selection, we consider two approaches for parameter estimation. In the first approach, we fit the unregularized AFT model in (1) using the set of selected variables S_{π_0} to obtain the final parameter estimates $\hat{\vartheta}$ and $\hat{\phi}$, where $\hat{\vartheta}_j = 0$ if $\vartheta_j \notin S_{\pi_0}$ and $\hat{\phi}_j = 0$ if $\phi_j \notin S_{\pi_0}$. Alternatively, we can modify Rubin's rule to obtain the final estimates using $\{\hat{\vartheta}^{(h)}, h=1, \dots, H\}$ and $\{\hat{\phi}^{(h)}, h=1, \dots, H\}$ only for parameters in S_{π_0} .

Specifically, $\hat{\vartheta}_j = \frac{1}{H\pi(\vartheta_j)} \sum_h \hat{\vartheta}_j^{(h)}$ if $\vartheta_j \in S_{\pi_0}$, and $\hat{\vartheta}_j = 0$ if $\vartheta_j \notin S_{\pi_0}$; $\hat{\phi}_j = \frac{1}{H\pi(\phi_j)} \sum_h \hat{\phi}_j^{(h)}$ if $\phi_j \in S_{\pi_0}$, and $\hat{\phi}_j = 0$ if $\phi_j \notin S_{\pi_0}$. In cases where $|S_{\pi_0}|$ is greater than n or the number of uncensored observations, the first approach cannot be used but the second approach remains applicable.

2.5 Numerical Optimization

With the observed and novel information in $\mathcal{G}_1$ and $\mathcal{G}_2$, the optimization problem (2) for given tuning parameter values is convex under all penalties of interest including HGP (8), guaranteeing a unique solution. However, it involves a non-smooth loss function (3) and non-smooth regularization terms such as (8); it is not trivial to obtain $(\hat{\alpha}, \hat{\vartheta}, \hat{\phi})$ in (2) and most existing algorithms for least squares problems cannot be easily extended to this case. We apply a numerical algorithm that was recently developed for efficient computation of regularized rank-based estimation for AFT models (Chung et al., 2013). In essence, we optimize an asymptotically equivalent smooth approximation to the Gehan loss function (3) and use the technique from Huber (1964) to handle a non-smooth term in $\mathcal{J}(\vartheta, \phi)$ which is common in applied mathematics; for smoothened rank-based estimators in the AFT model, see also Heller (2007) and Johnson and Strawderman (2009). Subsequently it optimizes the smoothed objective function using a limited memory Broyden-Fletcher-Goldfarb-Shanno method (Nocedal and Wright, 2006). One advantage of this approach is that it can be easily adapted to other loss functions and penalties so that we can avoid writing a specialized algorithm for each loss function and penalty; cf. Table 1 in Chung et al. (2013). This approach has been shown to achieve good performance with considerably lower computational cost than other alternative algorithms for problems that are similar to the one of our interest (Chung et al., 2013). The details of the computational implementation are included in the supplementary materials.

To select the tuning parameters (λ_1, λ_2) , we can use the cross validation (CV) or generalized cross validation (GCV) approaches (Tibshirani, 1996). V -fold CV chooses the optimal values of tuning parameters that maximize the criterion based on the Gehan loss function (3). GCV chooses the optimal values of tuning parameters that maximize the criterion, $\mathcal{L}_n(\mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\phi}) / (1 - d_\lambda/n)^2$, where d_λ is the number of nonzero regression coefficient estimates. We use CV in our numerical studies because in our experience it has better performance in small samples and when $p > n$.

2.6 Operating Characteristics

In regularized least squares problems, the penalties (4)–(8) possess one or both of the two properties that are conducive to incorporate biological information in feature selection: the grouped variable selection and the grouping effects on regression coefficients; cf. Sections 2.4 and 2.5 in Pan et al. (2010). The grouped variable selection property encourages sparsity at the general group level, in particular pathway level, in the case of sGL (Simon et al., 2013; Yuan and Lin, 2006) or encourages the regression coefficient estimates for two neighboring nodes in $\mathcal{G}$ to be 0 simultaneously in the case of GP and NC (Pan et al., 2010; Li and Li, 2008). The grouping effects property encourages nonzero regression coefficient estimates for two neighboring nodes in $\mathcal{G}$ to be closer to each other regardless of their correlation in the case of FL (Tibshirani et al., 2005) or as their correlation increases to 1 in the case of GP and NC; cf. Corollary 2 in (Pan et al., 2010). Most notably, the proposed penalty HGP, including a group lasso term and a grouped penalty term, is the only penalty that encourages sparsity at the pathway level as well as for two neighboring nodes within $\mathcal{G}_1$ and $\mathcal{G}_2$ and it also encourages the grouping effects for two neighboring nodes in $\mathcal{G}_1$ when their correlation is close to 1. Of note, miRNAs are known to down-regulate gene expressions and the correlation between two neighboring nodes (i.e., an miRNA and its mRNA target) in $\mathcal{G}_2$ is negative. As a result, while HGP does encourage the regression coefficient estimates for two neighboring nodes (i.e., an miRNA and its mRNA target) in $\mathcal{G}_2$ to be 0 simultaneously, HGP does not encourage the non-zero coefficients for these neighboring nodes in $\mathcal{G}_2$ to be similar to each other, a biologically meaningful and very attractive feature.

In the absence of missing biological information in $\mathcal{G}_2$, the regularized semiparametric AFT models in Section 2.1 are expected to inherit the aforementioned two properties by following similar arguments in Cai et al. (2009), which showed the “oracle” property (Zou, 2006) of the adaptive lasso in the regularized Gehan loss function (2) based on the properties of U-Processes, and arguments in Pan et al. (2010), Li and Li (2008), and Yuan and Lin (2006). While the theoretical results in the presence of missing biological information in $\mathcal{G}_2$ are beyond the scope of this article and are a topic for future research, the proposed approach is shown in our numerical studies to outperform the approaches that ignore missing biological information or use the existing penalties.

In addition, the proposed approach in Section 2.4, particularly, the component for summarizing the results on feature selection across multiply imputed data sets, is similar in spirit to stability selection (Meinshausen and Bühlmann, 2010) and Bootstrap Inference for Network CONstruction (BINCO) (Li et al., 2013), which are developed for variable selection

and/or network construction using fully observed data. These methods use resampling/subsampling techniques and calculate selection frequencies across resampled/subsampled data sets that are similar to $\pi(\vartheta_j)$ and $\pi(\phi_j)$ in (13) and (14); they are less sensitive to the amount of regularization and, more importantly, offer principled cutoff rules for selection frequencies. Another direction for future research is to combine these techniques with our approach for feature selection in the presence of missing biological information.

3 Application to the Prostate Cancer Data

We apply the proposed approach to the motivating prostate cancer data. We conduct variable selection and parameter estimation using the training set of 73 patients and then test prediction performance using the independent set of 40 patients. Following the previous work, four clinical variables, T-stage, PSA, Gleason score and age, are also included as predictors. As a screening step to remove noise variables, we first conduct univariate regression of each genomic biomarker in relation to the time to cancer recurrence (Y) and select genomic biomarkers with a false discovery rate of < 0.5 for subsequent analyses. After the screening step, the training data set consists of 179 mRNA biomarkers and 75 miRNA biomarkers. Exploratory analyses are conducted to assess modeling assumptions and the results are included in the supplementary materials. The gene pathway information for $\mathcal{G}_1$ is obtained from the KEGG database. For the miRNA regulatory network $\mathcal{G}_2$, the experimentally validated component $\mathcal{G}_2^o$ is obtained from the TarBase database with the total number of edges 70, and the prior information used to guide the imputation of $\mathcal{G}_2^m$ is obtained from four databases, namely miRanda, PITA, TargetScan, and PicTar. To fit the imputation model (11), our prior specifications closely follow the recommendations in Section 5.1 of Stingo et al. (2010), e.g., $\kappa=0.7$, $\gamma=10$, and $\alpha=3$ in prior (12), $\eta=-3$, $a_\tau=1.5$ and $b_\tau=0.2$. Briefly, some hyper-parameter values such as α , a_τ and b_τ are chosen so that the resulting priors are fairly vague. Other hyper-parameters such as η are set to biologically meaningful values; plus the results are not very sensitive to their values, consistent with the findings in Stingo et al. (2010). We use $H=100$ imputations and for each imputation an MCMC chain of 100,000 iterations is used with the first 20,000 as burn-in. The average number of edges in $\mathcal{G}_2^m$ over all the imputations is 251.

After imputation, Model (1) is used to analyze each imputed data set. In addition to the proposed penalty HGP, we also use the extensions of the existing penalties described in Section 2.1, namely, fused lasso (FL), network-constrained regularization (NC), sparse group lasso (sGL) and grouped penalty (GP). Following the suggestions in Pan et al. (2010), we set $r=8$ in the data analysis for both HGP and GP and we set Ω_g to the identity matrix in both the data analysis and the simulations. Ten-fold CV is conducted to select optimal values of the tuning parameters for all penalties. The final set of the important variables is determined by their proportions of being selected across all imputed data sets with a cutoff value of $\pi_0=0.5$ and hence is denoted by $S_{0.5}$. Estimates of parameters in $S_{0.5}$ are obtained using the modified Rubin's rule, noting that $|S_{0.5}|$ is larger than the sample size for GP and HGP.

After obtaining the parameter estimates ($\hat{\alpha}$, $\hat{\vartheta}$ and $\hat{\phi}$) based on the training data, we calculate risk scores for subjects in the testing data set as follows $\hat{\alpha}^T \mathbf{X}_i + \hat{\vartheta}^T \mathbf{Z}_i + \hat{\phi}^T \mathbf{W}_i$. Using the risk

scores and $(\tilde{T}_i, \delta_i)$ in the testing data, we calculate the concordance statistic (c statistic) for survival data (Harrell, 2001; Steyerberg et al., 2010) to evaluate and compare prediction performance. The results from the proposed approach combined with different penalties are summarized in Table 1.

Table 1 shows that the proposed penalty (HGP) leads to the highest c statistic and hence the best prediction performance in the testing data set; though the differences are not statistical significant, likely due to the small sample size in the testing data. The number of selected variables varies across different penalties. The penalty GP selects the largest number of important variables, whereas sGL and FL are the most conservative ones, selecting only 13 biomarkers. A closer examination of our results further reveals that the sets of the selected variables using different penalties overlap considerably with one another as shown in Table 2. For example, of the biomarkers selected by HGP, 11, 20, 27 and 9 are also identified as important features by FL, NC GP and sGL, respectively, representing 84.6%, 76.9%, 24.1% and 69.2% of their respective sets of selected variables. The overlapping proportions between GP and the other penalties are 100.0%, 80.8%, 76.9% and 93.1% for FL, NC, sGL and HGP, respectively. As a comparison, we also calculate the c statistic for the two prediction models from the aforementioned earlier analyses of the data (Long et al., 2011) using lasso Cox PH models: one includes mRNA biomarkers and the other includes both mRNA and miRNA biomarkers. With a c statistic of 0.520 for both models, their prediction performance is considerably worse than the model using our approach combined with the proposed HGP.

The proposed approach using HGP identifies three pathways with at least half of their genes being selected, all of which have documented roles in cancer, and most in prostate cancer. For example, one selected pathway is known as “pathways in cancer” (hsa05200). For the genes within this pathway, C-terminal binding protein 1 (CtBP1) is a transcriptional corepressor of the androgen receptor that is known to play a role in multiple cancers. In prostate cancer, CtBP1 affects proliferation, invasion, metabolism, growth, and metastasis (Moiola et al., 2014; Wang et al., 2012). Vascular endothelial growth factor A (VEGFA) is the critical signaling ligand for angiogenesis, or growth of new blood vessels, that is highly expressed in prostate cancer (Woollard et al., 2013). Of the miRNA biomarkers selected, several have been associated with prostate cancer, including miR-29, which suppresses EMT (Ru et al., 2012), and miR149, which is down regulated in prostate cancer (Schaefer et al., 2010). In addition, several of these miRNAs are responsive to insulin signaling (Granjon et al., 2009), which is important in prostate cancer and is associated with increased risk of recurrence (Lehrer et al., 2002).

For the purpose of comparisons, we also analyze the data using two other approaches. First, we conduct feature selection after imputing $\mathcal{G}_2^m$ only once using the posterior mode of Ψ in Model (11), which is known as the “plug-in” approach. Second, we conduct a naive analysis by removing the missing component $\mathcal{G}_2^m$ from the model fitting process; in other words, $\mathcal{G}_2 = \mathcal{G}_2^o$. Analysis is again conducted for all the five penalties and the results are summarized in Table 3. A comparison between the two approaches in Table 3 clearly shows that incorporating the imputed miRNA regulatory network improves prediction performance, evidenced by considerably larger c statistics. In addition, incorporating the novel

information in $\mathcal{G}_2^m$ seems to result in removal of additional noise variables when compared to only using the experimentally validated information in $\mathcal{G}_2^o$. As for comparisons among different penalties, HGP still achieves the best performance among all penalties in both cases when $\mathcal{G}_2^m$ is imputed and used in feature selection and the naive approach, further lending support to the superiority of the proposed penalty.

Lastly, a comparison between Table 1 and Table 3 indicates that the proposed approach using multiply imputed $\mathcal{G}_2^m$ is more conservative than the plug-in approach. As for the c statistic, combined with the same penalty function, the multiple imputation approach always results in a larger c statistic compared to the plug-in approach, though such a difference is not statistically significant, again likely due to the small sample size in the testing data. In addition, the proposed approach using multiple imputation accounts for the uncertainty of imputation and is more likely to remove false positives than the plug-in approach which uses a single imputation and fails to account for such uncertainty; as a result, the multiple imputation approach leads to a better prediction performance than the plug-in approach. The limitation of the multiple imputation approach is that it incurs considerably greater computational cost.

4 Simulations

We further demonstrate in simulation studies the performance of the proposed penalty in comparison with other penalties and the advantage of incorporating the novel biological information. We consider six settings with a sample size of $n = 100$ and these settings are different in terms of the graph structure $\mathcal{G}$, how $(\mathbf{Z}, \mathbf{W})$ is generated, and the values of $(\boldsymbol{\vartheta}, \boldsymbol{\phi})$.

In the first setting, we generate $(\mathbf{Z}, \mathbf{W})$ from a multivariate normal distribution with mean zero and covariance matrix Σ with its (j, k) -th element σ_{jk} . We specify σ_{jk} as follows: $\sigma_{jk} = 0.5$ if $j \sim k$ is a gene-gene connection; $\sigma_{jk} = -0.7$ if $j \sim k$ is a gene-miRNA connection; and $\sigma_{jk} \sim \text{Unif}(-0.2, 0.2)$ for the other cases. The same specification of Σ is used in the other settings as well. We consider both $p = 63$ and $p = 252$. Figure 1 presents the network structure when $p = 63$, including 40 genes within 5 separate pathways and 23 miRNAs with 15 of them regulating genes and 8 of them isolated. In the case of $p = 252$, we simply duplicate the network in Figure 1 four times. We further assume that the first three subnetworks (pathways and the connected miRNAs) are associated with the clinical outcome T , whereas the remaining components are not associated with T . To mimic the real data, we add sparsity within some of the important pathways and set the regression coefficients for the 24 genes $\boldsymbol{\vartheta}(1 : 24)$ and the 10 connected miRNAs $\boldsymbol{\phi}(1 : 10)$ in the three important pathways to

$$(\boldsymbol{\vartheta}(1:24), \boldsymbol{\phi}(1:10)) = (0, 0, 0, \frac{2}{\sqrt{8}}, 0, 0, \frac{2}{\sqrt{8}}, \frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}}, \frac{2}{\sqrt{8}}, 0, -\frac{2}{\sqrt{8}}, \frac{2}{\sqrt{8}}, 0, 0, \frac{2}{\sqrt{8}}, \frac{2}{\sqrt{8}}, \frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}}, \frac{2}{\sqrt{8}}, \frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}}, 0, 0, -\frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}}, 0, -\frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}}, -\frac{2}{\sqrt{8}})^T,$$

which is similar to the simulation settings in Li and Li (2008) and Pan et al. (2010). The regression coefficients for the genes and miRNAs in the unimportant subnetwork and isolated miRNAs – the remaining components of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ – are set to 0. We generate $\boldsymbol{v} \sim \text{Unif}(0, 1)$ and the censoring time as $C = \boldsymbol{\theta}^T \mathbf{Z} + \boldsymbol{\phi}^T \mathbf{W} + \boldsymbol{v}$, resulting in a censoring rate up to 35%.

In the second and third settings, with the other configurations the same as the first setting, we specify the regression coefficients in different ways. Let ζ and ζ' denote the index sets of the important mRNA and miRNA, respectively. In the second setting, while $\zeta = (1 : 24)$ and $\zeta' = (1 : 10)$ are the same as those in setting 1, we generate $(\boldsymbol{\theta}(\zeta), \boldsymbol{\phi}(\zeta'))$ from standard normal distributions; this is the case where the regression coefficient values are not constrained by the network structure. In the third setting, we choose ζ and ζ' by randomly selecting 24 mRNAs from all the mRNAs and 10 miRNAs from all the miRNAs, and then generate $(\boldsymbol{\theta}(\zeta), \boldsymbol{\phi}(\zeta'))$ from standard normal distributions. This is the case where the important mRNAs and miRNAs are randomly selected without following the network structure.

As a comparison, settings 4–6 have the same configurations as settings 1–3 respectively except that we modify the network structure $\mathcal{G}$ by allowing overlapping between different pathways. Specifically, based on Figure 1, we replace gene 31 in the unimportant pathway 4 with gene 9 from the important pathway 2, making gene 9 the shared gene between the second and fourth pathways. To keep the total number of genes constant, we insert another unimportant gene to pathway 4 which is connected to gene 9. Similarly, we replace gene 36 in the unimportant pathway 5 with gene 17 in the important pathway 3, making gene 17 the shared gene between the third and fifth pathways. One additional unimportant gene connected to gene 17 is added to pathways 5 to keep the total number of genes unchanged.

In each setting, we further consider three scenarios for the miRNA regulatory network $\mathcal{G}_2$. Scenario 1 is the gold standard (GS) where the fully known $\mathcal{G}_2$ is used in the analysis with different penalties and is not applicable in practice. The second and third scenarios mimics real data in which $\mathcal{G}_2$ is partially missing; specifically, the observed component $\mathcal{G}_2^o$ includes the edges between miRNA 48 and gene 22 and between miRNA 55 and gene 35 with the edges in $\mathcal{G}_2$ are missing. We conduct the naive analysis in scenario 2 (Naive) and adopt the proposed multiple imputation approach with $H = 50$ in scenario 3 (MI).

For each setting under both $p = 63$ and $p = 252$, we generate 50 Monte Carlo (MC) data sets. In each MC data set, ten-fold CV is used to select optimal tuning parameter values. To evaluate performance in prediction, we report the sum of squared errors (SSE) defined as $((\boldsymbol{\theta}^T, \boldsymbol{\phi}^T) - (\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\phi}}^T))((\boldsymbol{\theta}, \boldsymbol{\phi}) - (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}))$ where $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ denotes parameter estimates. To evaluate performance in variable selection, we report specificity and sensitivity. The results for the simulation studies are summarized in Table 4 for $p = 63$ and Table 5 for $p = 252$.

We first compare different penalties. For all the settings and scenarios under both “ $p < n$ ” and “ $p > n$ ” cases, our proposed penalty HGP achieves the best or close to the best performance in terms of feature selection and risk prediction. Specifically, in setting 1 which is constructed as a natural extension of simulations in Li and Li (2008) and Pan et al. (2010)

with added sparsity within pathways, HGP achieves the lowest SSE with high sensitivity and specificity. In setting 2 where the important variables are the same as in setting 1 but their coefficient values are not constrained by the network structure, HGP again achieves the overall best performance and its superiority over FL, NC and GP in particular becomes more pronounced as p increases. In setting 3 where the important variables are randomly selected and not constrained by the network structure, the performance of all methods deteriorates compared to settings 1 and 2; the superior performance of HGP in this setting demonstrates its robustness relative to the other methods when the network structure is not informative for variable selection. Similarly, when pathways are overlapped in settings 4 to 6, the proposed approach remains as the top performer with the lowest SSE as well as high sensitivity and specificity. In addition, the impact of p on HGP is in general more moderate compared to the other methods. As for comparisons among the other methods, sGP has a lower SSE and better risk prediction performance than FL, NC and GP in most settings, illustrating the necessity of the group level regularization. The results of FL, NC and GP are somewhat mixed across different settings with $p = 63$. But as p increased, while the performance of FL and NC is still comparable, GP suffers from considerably larger SSE and lower sensitivity. In terms of comparison among different scenarios (namely, GS, Naive, MI), Table 4 and Table 5 show a general pattern that the MI approach outperforms the naive approach in terms of risk prediction and feature selection across all the settings. In addition, the GS approach tends to outperform the MI approach with a couple exceptions. In summary, our simulation results suggest the superiority of conducting selection hierarchically and reinforce the importance of incorporating both known and novel biological information.

5 Discussion

In this article, we present a new approach for feature selection of mRNA and miRNA biomarkers in relation to cancer recurrence while incorporating known and novel biological information, which has not been investigated in the literature. Specifically, our proposed approach incorporates two levels of biological information on pathway memberships and on gene-gene and gene-miRNA interactions and facilitates hierarchical feature selection of genomic biomarkers at both the general group level and the biomarker level in relation to cancer recurrence that is subject to censoring. In addition, our approach incorporates the biological information on mRNA biomarkers and the information on the miRNA regulatory network, where the missing component of the miRNA regulatory network is imputed from the observed data and the imputation is guided by existing knowledge on putative miRNA targets. Our data analysis and simulation studies show the importance and value of incorporating novel miRNA regulatory network and the superiority of the proposed penalty compared to the extensions of several existing penalties in terms of feature selection and risk prediction.

While our method is proposed for analysis of genomic data, it can be readily applied to incorporate other types of biological information that can be similarly represented by a graph; cf. Table 1 in Kanehisa et al. (2002). As for the incorporation of novel biological information, the imputation model in Section 2.3 is implemented based on the approach by Stingo et al. (2010). One strength of this approach is that it obtains valid posterior samples from a predictive posterior distribution, which is a good fit for the objective of multiple

imputation. However, this modeling approach is not essential to our proposed approach and we can replace it with an alternative imputation approach; for example, we can use regularized multivariate regression (Lutz and Buhlmann, 2006; Peng et al., 2010) for Model (11). As noted previously, while we present the proposed approach in the context of semiparametric AFT models for survival endpoints, the proposed approach is general and can be readily extended to general regression problems where a loss function or objective function is used to obtain parameter estimates.

One limitation of the proposed penalty is its computational cost. In the real data analysis, the computing time for one data set is 32.47 seconds for FL, 26.81 seconds for NC, 93.99 seconds for GP, 2562.71 seconds for sGL and 1845.03 seconds for our proposed HGP (3.4 GHz CPU, 8GB Memory, Windows System). To address this issue, more efficient computational algorithms are likely needed, if we want to extend our approach to the case of ultra high-dimensional data such as imaging data. In addition, in our data application we implement a straightforward screening procedure, which can be replaced by more sophisticated screening procedures; for example, Song et al. (2014) recently proposed a screening procedure with the sure screening property which is appropriate for AFT models. Assessment of screening procedures in the settings considered in this work is a subject for future research. The MAR assumption underlying the proposed imputation approach is another potential limitation, since it may not hold in biological settings such as ours. Since the MAR assumption cannot be verified based on observed data, another topic for future research is to investigate sensitivity analysis under deviations from the MAR assumption in these biological settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by NIH grants R03CA173770, R03CA183006, and R21GM107683. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors thank the editor, an associate editor and three referees for comments.

References

- Abramovitz M, Ordanic-Kodani M, Wang Y, Li Z, Catzavelos C, Bouzyk M, Sledge G, Moreno C, Leyland-Jones B. Optimization of RNA extraction from FFPE tissues for expression profiling in the DASL assay. *Biotechniques*. 2008; 44:417–23. [PubMed: 18361796]
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*. 2000; 25:25–29. [PubMed: 10802651]
- Bagga S, Bracht J, Hunter S, Massirer K, Holtz J, Eachus R, Pasquinelli AE. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell*. 2005; 122:553. [PubMed: 16122423]
- Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets and expression. *Nucleic Acids Research*. 2008; 36:D149–D153. [PubMed: 18158296]
- Cai T, Huang J, Tian L. Regularized estimation for the accelerated failure time model. *Biometrics*. 2009; 65:394–404. [PubMed: 18573133]

- Catto JW, Alcaraz A, Bjartell AS, De Vere White R, Evans CP, Fussel S, Hamdy FC, Kallioniemi O, Mengual L, Schlomm T, et al. microRNA in prostate, bladder, and kidney cancer: a systematic review. *European Urology*. 2011; 59:671–681. [PubMed: 21296484]
- Chekouo T, Stingo FC, Doecke JD, Do K-A. miRNA–target gene regulatory networks: A Bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics*. 2015:428–438. [PubMed: 25639276]
- Chung M, Long Q, Johnson B. A tutorial on rank-based coefficient estimation for censored data in small-and large-scale problems. *Statistics and Computing*. 2013; 23(5):601–614. [PubMed: 23956500]
- Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M, Hintsanen P, Khan SA, Mpindi J-P, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*. 2014; 32:1202–1212.
- Cox D. Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society Series B*. 1972; 34:187–202.
- Cox, D., Oakes, D. *Analysis of Survival Data*. Chapman and Hall; London: 1984.
- Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, Mesher D, Speights V, Stankiewicz E, Foster CS, et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. *The Lancet Oncology*. 2011; 12:245–255. [PubMed: 21310658]
- Dicker L, Huang B, Lin X. Variable selection and estimation with the seamless-L0 penalty. *Statistica Sinica*. 2012; 23:929–962.
- Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *The Annals of statistics*. 2004; 32:407–499.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*. 1965; 52:203–223. [PubMed: 14341275]
- Granjon A, Gustin M-P, Rieusset J, Lefai E, Meugnier E, Güller I, Cerutti C, Paultre C, Disse E, Rabasa-Lhoret R, et al. The microRNA Signature in Response to Insulin Reveals Its Implication in the Transcriptional Action of Insulin in Human Skeletal Muscle and the Role of a Sterol Regulatory Element–Binding Protein1c/Myocyte Enhancer Factor 2C Pathway. *Diabetes*. 2009; 58:2555–2564. [PubMed: 19720801]
- Grimson A, Farh KK-H, Johnston WK, Garrett-Engle P, Lim LP, Bartel DP. microRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*. 2007; 27:91–105. [PubMed: 17612493]
- Harel O, Zhou X-H. Multiple imputation-Review of theory, implementation and software. *Statistics in Medicine*. 2007; 26:3055–3057.
- Harrell, FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer; 2001.
- Heller G. Smoothed rank regression with censored data. *Journal of the American Statistical Association*. 2007; 102
- Huber PJ. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*. 1964; 35:73–101.
- Jauhiainen A, Nerman O, Michailidis G, Jörnsten R. Transcriptional and metabolic data integration and modeling for identification of active pathways. *Biostatistics*. 2012; 13:748–761. [PubMed: 22699861]
- Jin Z, Lin D, Wei L, Ying Z. Rank-based inference for the accelerated failure time model. *Biometrika*. 2003; 90:341–353.
- Johnson LM, Strawderman RL. Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data. *Biometrika*. 2009; 96:577–590. [PubMed: 23049117]
- Kalbfleisch, J., Prentice, R. *The Statistical Analysis of Failure Time Data*. John Wiley; New York: 2002.

- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. 2000; 28:27–30. [PubMed: 10592173]
- Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Research*. 2002; 30:42–46. [PubMed: 11752249]
- Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. *Nature genetics*. 2007; 39:1278–1284. [PubMed: 17893677]
- Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*. 2012; 8:e1002375. [PubMed: 22383865]
- Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, et al. Combinatorial microRNA target predictions. *Nature genetics*. 2005; 37:495–500. [PubMed: 15806104]
- Lehrer S, Diamond EJ, Stagger S, Stone NN, Stock RG. Increased serum insulin associated with increased risk of prostate cancer recurrence. *The Prostate*. 2002; 50:1–3. [PubMed: 11757030]
- Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. 2005; 120:15–20. [PubMed: 15652477]
- Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*. 2008; 24:1175–1182. [PubMed: 18310618]
- Li S, Hsu L, Peng J, Wang P. Bootstrap inference for network construction with an application to a breast cancer microarray study. *The Annals of Applied Statistics*. 2013; 7:391–417. [PubMed: 24563684]
- Little, R., Rubin, D. *Statistical Analysis with Missing Data*. John Wiley; New York: 2002.
- Loh P-L, Wainwright MJ. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in Neural Information Processing Systems*. 2011:2726–2734.
- Long Q, Johnson BA, Osunkoya AO, Lai Y-H, Zhou W, Abramovitz M, Xia M, Bouzyk MB, Nam RK, Sugar L, et al. Protein-coding and microRNA biomarkers of recurrence of prostate cancer following radical prostatectomy. *The American Journal of Pathology*. 2011; 179:46–54. [PubMed: 21703393]
- Lutz RW, Buhlmann P. Boosting for high-multivariate responses in high-dimensional linear regression. *Statistica Sinica*. 2006; 16:471.
- Meinshausen N, Buhlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010; 72:417–473.
- Moiola CP, De Luca P, Zalazar F, Cotignola J, Rodriguez-Segui S, Gardner K, Meiss R, Vallecorsa P, Pignataro O, Mazza O, et al. Prostate tumor growth is impaired by CtBP1 depletion in high fat diet fed mice. *Clinical Cancer Research*. 2014 clincanres–0322.
- Nakagawa T, Kollmeyer T, Morlan B, Anderson S, Bergstralh E, Davis B, Asmann Y, Klee G, Ballman K, Jenkins R. A Tissue Biomarker Panel Predicting Systemic Progression after PSA Recurrence Post-Definitive Prostate Cancer Therapy. *PLoS One*. 2008; 3(5):e2318. [PubMed: 18846227]
- Nocedal, J., Wright, S. *Numerical Optimization*, Second Edition. Springer; New York: 2006.
- Pan W, Xie B, Shen X. Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*. 2010; 66:474–484. [PubMed: 19645699]
- Peng J, Zhu J, Bergamaschi A, Han W, Noh D-Y, Pollack JR, Wang P. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*. 2010; 4:53–77. [PubMed: 24489618]
- Pillai RS, Bhattacharyya SN, Artus CG, Zoller T, Cougot N, Basyuk E, Bertrand E, Filipowicz W. Inhibition of translational initiation by Let-7 microRNA in human cells. *Science*. 2005; 309:1573–1576. [PubMed: 16081698]
- Reid N. A conversation with Sir David Cox. *Statistical Science*. 1994; 9:439–455.
- Ru P, Steele R, Newhall P, Phillips NJ, Toth K, Ray RB. miRNA-29b suppresses prostate cancer metastasis by regulating epithelial–mesenchymal transition signaling. *Molecular Cancer Therapeutics*. 2012; 11:1166–1173. [PubMed: 22402125]
- Rubin, D. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 1987.

- Schaefer A, Jung M, Mollenkopf H-J, Wagner I, Stephan C, Jentzmik F, Miller K, Lein M, Kristiansen G, Jung K. Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma. *International journal of cancer*. 2010; 126:1166–1176. [PubMed: 19676045]
- Schoof CRG, da Silva Botelho EL, Izzotti A, dos Reis Vasques L. MicroRNAs in cancer treatment and prognosis. *American Journal of Cancer Research*. 2012; 2:414–433. [PubMed: 22860232]
- Shen X, Pan W, Zhu Y. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*. 2012; 107:223–232. [PubMed: 22736876]
- Shojaie A, Michailidis G. Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology*. 2009; 16:407–426. [PubMed: 19254181]
- Shojaie A, Michailidis G. Network enrichment analysis in complex experiments. *Statistical applications in genetics and molecular biology*. 2010; 9
- Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group lasso. *Journal of Computational and Graphical Statistics*. 2013; 22:231–245.
- Slack FJ, Weidhaas JB. MicroRNA in Cancer Prognosis. *New England Journal of Medicine*. 2008; 359:2720–2722. [PubMed: 19092157]
- Song R, Lu W, Ma S, Jeng XJ. Censored rank independence screening for high-dimensional survival data. *Biometrika*. 2014; 101:799–814. [PubMed: 25663709]
- Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass)*. 2010; 21:128.
- Stingo FC, Chen YA, Tadesse MG, Vannucci M. Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*. 2011; 5:1978–2002. [PubMed: 23667412]
- Stingo FC, Chen YA, Vannucci M, Barrier M, Mirkes PE. A Bayesian graphical modeling approach to microRNA regulatory network inference. *The Annals of Applied Statistics*. 2010; 4:2024–2048. [PubMed: 23946863]
- Thermann R, Hentze MW. Drosophila miR2 induces pseudo-polysomes and inhibits translation initiation. *Nature*. 2007; 447:875–878. [PubMed: 17507927]
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996:267–288.
- Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in Medicine*. 1997; 16:385–395. [PubMed: 9044528]
- Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005; 67:91–108.
- Tricoli JV, Jacobson JW. MicroRNA: potential for cancer detection, diagnosis, and prognosis. *Cancer research*. 2007; 67:4553–4555. [PubMed: 17510380]
- Tsiatis A. Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics*. 1990; 18:354–372.
- van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011; 45(3):1–67.
- Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Research*. 2012; 40:D222–D2290. [PubMed: 22135297]
- Wang R, Asangani IA, Chakravarthi BV, Ateeq B, Lonigro RJ, Cao Q, Ram-Shankar M, Camacho DF, McGregor N, Schumann TE, et al. Role of transcriptional corepressor CtBP1 in prostate cancer progression. *Neoplasia*. 2012; 14:905–IN8. [PubMed: 23097625]
- Wang S, Nan B, Zhu N, Zhu J. Hierarchically penalized Cox regression with grouped variables. *Biometrika*. 2009; 96:307–322.
- Woollard DJ, Opeskin K, Coso S, Wu D, Baldwin ME, Williams ED. Differential expression of VEGF ligands and receptors in prostate cancer. *The Prostate*. 2013; 73:563–572. [PubMed: 23038639]
- Wu TT, Wang S. Doubly regularized Cox regression for high-dimensional survival data with group structures. *Statistics and Its Interface*. 2013; 6:175–186.

- Ying Z. A Large Sample Study of Rank Estimation for Censored Regression Data. *Annals of Statistics*. 1993; 21:76–99.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2006; 68:49–67.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005; 67:301–320.

Figure 1. Network structure for the simulation studies with $k = 5$. Circles represent mRNAs and triangles represent miRNAs. The important pathways and miRNAs are highlighted in filled circles and triangles.

Table 1

Analysis results for the prostate cancer data using the proposed multiple imputation approach combined with different penalties. S_{π_0} is defined in (15) and $|S_{0.5}|$ denotes the size (cardinality) of the set $S_{0.5}$ and hence the number of selected variables. P-value is based on a comparison between each penalty and the HGP.

Penalty	$ S_{0.5} $	c statistic	P-value
FL	13	0.618	0.163
NC	26	0.645	0.212
GP	122	0.553	0.437
sGL	13	0.632	0.196
HGP	29	0.664	–

Table 2

Overlapping of the selected biomarkers between different penalties for the prostate cancer data from the proposed approach with multiple imputation.*

	FL	NC	GP	sGL	HGP
FL	13	11	13	9	11
NC	-	26	21	11	20
GP	-	-	122	10	27
sGL	-	-	-	13	9
HGP	-	-	-	-	29

*The diagonal cells provide the number of the selected biomarkers for each penalty and the o_i-diagonal cells provide the numbers of overlapping selected biomarkers between two penalties.

Table 3

Analysis results for the prostate cancer data using the “plug-in” approach and the naive approach. S is the estimated active set and $|S|$ denotes the size (cardinality) of the set S and hence the number of selected variables. P-value is based on a comparison between each approach and the corresponding multiple imputation approach in Table 1.

Penalty	Approach	$ S $	c statistic	P-value
FL	plug-in	18	0.601	0.285
	naive	19	0.571	0.409
NC	plug-in	27	0.614	0.382
	naive	71	0.604	0.381
GP	plug-in	113	0.487	0.240
	naive	219	0.383	0.669
sGL	plug-in	18	0.612	0.286
	naive	18	0.596	0.330
HGP	plug-in	28	0.650	0.107
	naive	44	0.629	0.153

Table 4

Simulation results under low-dimensional case ($p = 63$). SSE, average sum of squared errors; Sens, sensitivity of variable selection; Spec, specificity of variable selection. The Monte Carlo standard deviation for each performance metric is included in the parentheses.

Setting	Method	GS				Naive				MI			
		SSE	Sens	Spec	SSE	Sens	Spec	SSE	Sens	Spec	SSE	Sens	Spec
1	FL	4.04 (1.30)	0.86 (0.08)	0.87 (0.06)	3.66 (1.24)	0.88 (0.05)	0.85 (0.06)	5.11 (1.42)	0.74 (0.11)	0.90 (0.05)			
	NC	3.45 (0.55)	0.90 (0.05)	0.71 (0.09)	2.98 (0.73)	0.94 (0.04)	0.78 (0.08)	2.67 (0.51)	0.95 (0.04)	0.74 (0.07)			
	GP	3.17 (1.64)	0.96 (0.06)	0.78 (0.06)	7.76 (2.9)	0.85 (0.85)	0.65 (0.06)	2.56 (0.88)	0.96 (0.04)	0.86 (0.06)			
	sGL	2.39 (0.86)	0.95 (0.05)	0.92 (0.05)	3.55 (0.87)	0.87 (0.07)	0.90 (0.05)	2.73 (0.59)	0.95 (0.04)	0.85 (0.06)			
	HGP	1.11 (0.39)	0.99 (0.01)	0.93 (0.05)	1.75 (0.60)	0.98 (0.03)	0.90 (0.06)	1.46 (0.46)	0.99 (0.02)	0.90 (0.07)			
2	FL	10.87 (4.15)	0.61 (0.14)	0.79 (0.07)	8.07 (2.40)	0.61 (0.09)	0.78 (0.07)	13.51 (5.66)	0.45 (0.19)	0.87 (0.06)			
	NC	8.36 (1.69)	0.65 (0.05)	0.61 (0.08)	9.01 (2.04)	0.77 (0.07)	0.62 (0.09)	5.70 (1.20)	0.73 (0.06)	0.65 (0.07)			
	GP	6.14 (2.09)	0.76 (0.06)	0.68 (0.06)	14.24 (3.37)	0.66 (0.08)	0.54 (0.06)	4.91 (1.41)	0.72 (0.05)	0.77 (0.06)			
	sGL	3.89 (1.63)	0.70 (0.06)	0.86 (0.06)	6.79 (1.78)	0.62 (0.07)	0.85 (0.07)	5.26 (1.47)	0.67 (0.05)	0.81 (0.07)			
	HGP	1.63 (0.60)	0.83 (0.06)	0.88 (0.04)	2.85 (0.96)	0.79 (0.07)	0.84 (0.06)	2.93 (0.90)	0.75 (0.06)	0.83 (0.06)			
3	FL	11.30 (2.93)	0.75 (0.07)	0.58 (0.10)	10.76 (2.73)	0.76 (0.08)	0.55 (0.09)	14.21 (2.93)	0.70 (0.09)	0.62 (0.09)			
	NC	9.78 (2.22)	0.79 (0.07)	0.55 (0.09)	9.21 (2.29)	0.79 (0.06)	0.54 (0.09)	9.47 (1.84)	0.80 (0.06)	0.54 (0.08)			
	GP	9.50 (3.32)	0.78 (0.08)	0.59 (0.09)	11.12 (3.82)	0.77 (0.08)	0.55 (0.08)	11.28 (2.96)	0.75 (0.08)	0.63 (0.09)			
	sGL	13.88 (2.98)	0.69 (0.12)	0.69 (0.10)	15.99 (2.50)	0.60 (0.12)	0.75 (0.10)	13.08 (2.61)	0.72 (0.11)	0.67 (0.08)			
	HGP	6.77 (1.58)	0.86 (0.04)	0.62 (0.09)	8.98 (2.84)	0.80 (0.09)	0.68 (0.08)	7.96 (2.41)	0.82 (0.09)	0.64 (0.10)			
4	FL	5.52 (1.47)	0.81 (0.10)	0.81 (0.07)	5.22 (1.67)	0.84 (0.10)	0.79 (0.06)	6.53 (1.83)	0.64 (0.20)	0.88 (0.07)			
	NC	3.68 (0.66)	0.88 (0.05)	0.73 (0.09)	3.25 (0.64)	0.92 (0.04)	0.76 (0.07)	2.81 (0.54)	0.94 (0.05)	0.76 (0.06)			
	GP	3.81 (1.62)	0.94 (0.09)	0.75 (0.05)	8.32 (2.60)	0.82 (0.11)	0.62 (0.07)	3.32 (1.08)	0.94 (0.06)	0.81 (0.06)			
	sGL	2.72 (0.89)	0.94 (0.06)	0.87 (0.07)	4.08 (1.02)	0.85 (0.08)	0.86 (0.05)	3.00 (0.90)	0.94 (0.06)	0.84 (0.06)			
	HGP	1.83 (0.49)	0.98 (0.03)	0.85 (0.07)	3.10 (0.69)	0.93 (0.06)	0.83 (0.06)	2.03 (0.50)	0.99 (0.02)	0.81 (0.07)			
5	FL	5.28 (1.28)	0.79 (0.11)	0.82 (0.07)	5.09 (1.38)	0.83 (0.09)	0.79 (0.07)	6.54 (1.40)	0.63 (0.15)	0.87 (0.07)			
	NC	3.51 (0.54)	0.89 (0.05)	0.72 (0.08)	3.27 (0.58)	0.93 (0.03)	0.76 (0.07)	2.72 (0.55)	0.95 (0.05)	0.74 (0.07)			
	GP	3.76 (1.31)	0.93 (0.06)	0.76 (0.06)	8.64 (3.09)	0.80 (0.12)	0.63 (0.06)	3.14 (0.89)	0.95 (0.05)	0.81 (0.06)			
	sGL	2.59 (0.70)	0.95 (0.05)	0.87 (0.06)	3.84 (0.75)	0.86 (0.07)	0.88 (0.06)	2.87 (0.69)	0.95 (0.05)	0.84 (0.06)			

Setting	Method	GS			Naive			MI		
		SSE	Sens	Spec	SSE	Sens	Spec	SSE	Sens	Spec
		(0.40)	(0.03)	(0.06)	(0.55)	(0.04)	(0.07)	(0.46)	(0.03)	(0.08)
9	HGP	1.81	0.98	0.85	3.09	0.93	0.83	1.96	0.99	0.81
		(2.82)	(0.06)	(0.10)	(2.23)	(0.05)	(0.10)	(3.42)	(0.08)	(0.09)
	FL	12.29	0.70	0.64	11.05	0.77	0.58	13.80	0.69	0.66
		(2.28)	(0.05)	(0.09)	(2.49)	(0.06)	(0.09)	(2.25)	(0.06)	(0.09)
	NC	13.20	0.74	0.56	12.58	0.76	0.54	12.53	0.75	0.56
		(2.61)	(0.05)	(0.08)	(3.01)	(0.05)	(0.07)	(2.74)	(0.07)	(0.08)
	GP	10.96	0.78	0.60	12.75	0.78	0.64	10.41	0.75	0.65
		(2.75)	(0.07)	(0.09)	(3.94)	(0.11)	(0.10)	(2.27)	(0.06)	(0.08)
	sGL	10.81	0.74	0.63	14.68	0.62	0.71	10.46	0.75	0.62
		(1.81)	(0.05)	(0.09)	(2.04)	(0.05)	(0.07)	(2.09)	(0.05)	(0.09)
	HGP	8.10	0.84	0.62	9.17	0.80	0.63	8.66	0.82	0.61

Table 5

Simulation results under high-dimensional case ($p = 252$). SSE, average sum of squared errors; Sens, sensitivity of variable selection; Spec, specificity of variable selection. The Monte Carlo standard deviation for each performance metric is included in the parentheses.

Setting	Method	GS				Naive				MI			
		SSE	Sens	Spec	SSE	Sens	Spec	SSE	Sens	Spec	SSE	Sens	Spec
1	FL	5.57 (1.69)	0.78 (0.14)	0.93 (0.02)	5.32 (1.69)	0.82 (0.11)	0.92 (0.03)	5.56 (1.78)	0.78 (0.12)	0.93 (0.02)			
	NC	5.61 (0.93)	0.80 (0.07)	0.88 (0.05)	4.82 (1.00)	0.87 (0.07)	0.88 (0.04)	4.56 (0.99)	0.87 (0.07)	0.89 (0.04)			
	GP	12.17 (4.94)	0.70 (0.23)	0.83 (0.03)	82.75 (37.04)	0.28 (0.05)	0.69 (0.02)	19.01 (5.06)	0.46 (0.18)	0.80 (0.02)			
	sGL	2.04 (0.74)	0.97 (0.04)	0.97 (0.02)	4.09 (0.79)	0.84 (0.07)	0.95 (0.02)	3.59 (0.76)	0.87 (0.06)	0.96 (0.02)			
	HGP	1.54 (0.61)	1.00 (0.01)	0.94 (0.03)	3.47 (1.55)	0.94 (0.06)	0.91 (0.03)	2.63 (0.79)	0.97 (0.04)	0.93 (0.03)			
2	FL	13.78 (4.29)	0.53 (0.13)	0.90 (0.03)	10.56 (2.77)	0.56 (0.09)	0.91 (0.03)	11.82 (2.83)	0.54 (0.09)	0.92 (0.02)			
	NC	13.55 (2.55)	0.52 (0.07)	0.87 (0.03)	11.41 (2.80)	0.63 (0.07)	0.87 (0.04)	10.45 (1.95)	0.61 (0.07)	0.88 (0.02)			
	GP	18.08 (7.11)	0.6 (0.13)	0.81 (0.03)	148.78 (45.38)	0.31 (0.07)	0.67 (0.02)	30.01 (11.53)	0.43 (0.14)	0.77 (0.03)			
	sGL	2.95 (0.89)	0.73 (0.04)	0.97 (0.01)	7.36 (1.71)	0.59 (0.07)	0.95 (0.02)	6.27 (1.62)	0.61 (0.05)	0.95 (0.02)			
	HGP	2.33 (0.87)	0.83 (0.06)	0.92 (0.03)	5.65 (2.65)	0.76 (0.07)	0.89 (0.04)	3.82 (1.12)	0.78 (0.06)	0.91 (0.04)			
3	FL	18.74 (3.34)	0.61 (0.12)	0.86 (0.04)	19.14 (3.56)	0.58 (0.11)	0.85 (0.05)	20.06 (3.04)	0.79 (0.11)	0.84 (0.02)			
	NC	18.66 (1.94)	0.59 (0.08)	0.84 (0.06)	18.28 (2.50)	0.60 (0.08)	0.82 (0.06)	18.24 (1.82)	0.85 (0.10)	0.81 (0.03)			
	GP	33.42 (9.61)	0.51 (0.12)	0.80 (0.04)	202.89 (60.1)	0.33 (0.05)	0.66 (0.02)	46.50 (14.09)	0.66 (0.19)	0.71 (0.04)			
	sGL	15.87 (2.29)	0.63 (0.10)	0.91 (0.03)	18.24 (2.62)	0.53 (0.10)	0.90 (0.03)	16.92 (2.66)	0.85 (0.13)	0.87 (0.02)			
	HGP	13.34 (2.71)	0.75 (0.08)	0.88 (0.04)	16.02 (4.32)	0.68 (0.11)	0.83 (0.04)	12.46 (3.71)	0.84 (0.08)	0.87 (0.04)			
4	FL	7.05 (1.60)	0.71 (0.15)	0.92 (0.03)	6.89 (1.39)	0.77 (0.11)	0.9 (0.03)	7.17 (1.27)	0.68 (0.13)	0.93 (0.02)			
	NC	5.77 (1.07)	0.79 (0.09)	0.89 (0.05)	5.18 (0.86)	0.85 (0.06)	0.87 (0.05)	4.81 (0.82)	0.85 (0.07)	0.90 (0.03)			
	GP	13.34 (4.81)	0.64 (0.22)	0.83 (0.03)	73.94 (26.03)	0.31 (0.07)	0.67 (0.02)	21.91 (3.73)	0.40 (0.13)	0.80 (0.02)			
	sGL	3.00 (0.66)	0.93 (0.06)	0.95 (0.02)	4.80 (0.96)	0.80 (0.08)	0.95 (0.02)	4.40 (0.90)	0.83 (0.08)	0.95 (0.01)			
	HGP	2.40 (0.70)	0.98 (0.03)	0.92 (0.04)	3.74 (1.26)	0.93 (0.06)	0.90 (0.03)	3.56 (0.82)	0.93 (0.04)	0.92 (0.02)			
5	FL	6.97 (1.47)	0.71 (0.14)	0.92 (0.02)	6.85 (1.42)	0.74 (0.13)	0.91 (0.03)	7.06 (1.61)	0.69 (0.14)	0.93 (0.02)			
	NC	5.68 (1.07)	0.80 (0.08)	0.88 (0.04)	5.13 (0.88)	0.87 (0.06)	0.87 (0.05)	4.74 (0.85)	0.87 (0.06)	0.90 (0.03)			
	GP	14.22 (4.33)	0.61 (0.18)	0.84 (0.03)	79.69 (24.93)	0.31 (0.06)	0.67 (0.02)	21.88 (5.44)	0.43 (0.08)	0.78 (0.04)			
	sGL	2.80 (0.70)	0.93 (0.06)	0.95 (0.01)	4.73 (0.93)	0.81 (0.07)	0.95 (0.02)	4.51 (0.80)	0.82 (0.07)	0.95 (0.02)			

Setting	Method	GS			Naive			MI		
		SSE	Sens	Spec	SSE	Sens	Spec	SSE	Sens	Spec
6	HGP	2.48 (0.78)	0.98 (0.03)	0.92 (0.04)	4.01 (1.73)	0.93 (0.06)	0.90 (0.04)	3.64 (0.71)	0.93 (0.06)	0.92 (0.02)
	FL	19.21 (2.91)	0.58 (0.09)	0.87 (0.04)	20.55 (3.79)	0.59 (0.09)	0.85 (0.05)	20.21 (3.46)	0.85 (0.12)	0.83 (0.03)
	NC	20.51 (2.34)	0.56 (0.10)	0.85 (0.05)	20.57 (2.68)	0.59 (0.09)	0.85 (0.05)	19.77 (2.11)	0.86 (0.13)	0.81 (0.03)
	GP	34.36 (8.27)	0.48 (0.12)	0.81 (0.04)	220.92 (83.16)	0.35 (0.05)	0.66 (0.02)	49.16 (10.62)	0.52 (0.17)	0.74 (0.02)
	sGL	13.97 (2.48)	0.66 (0.09)	0.90 (0.04)	17.51 (2.98)	0.53 (0.09)	0.92 (0.03)	16.50 (3.52)	0.87 (0.14)	0.86 (0.03)
	HGP	12.83 (2.90)	0.74 (0.08)	0.88 (0.03)	16.21 (5.61)	0.63 (0.14)	0.87 (0.05)	11.88 (3.95)	0.84 (0.04)	0.88 (0.06)