Statistica Sinica (2013): Preprint

NETWORK-REGULARIZED HIGH-DIMENSIONAL COX REGRESSION FOR ANALYSIS OF GENOMIC DATA

Hokeun Sun[1], Wei Lin[2], Rui Feng[2] and Hongzhe Li[2]

[1] *Columbia University and* [2] *University of Pennsylvania*

Abstract: We consider estimation and variable selection in high-dimensional Cox regression when a prior knowledge of the relationships among the covariates, described by a network or graph, is available. A limitation of the existing methodology for survival analysis with high-dimensional genomic data is that a wealth of structural information about many biological processes, such as regulatory networks and pathways, has often been ignored. In order to incorporate such prior network information into the analysis of genomic data, we propose a network-based regularization method for high-dimensional Cox regression, by using an ℓ_1-penalty to induce sparsity of the regression coefficients and a quadratic Laplacian penalty to encourage smoothness between the coefficients of neighboring variables on a given network. The proposed method is implemented by an efficient coordinate descent algorithm. In the setting where the dimensionality p may grow exponentially fast with the sample size n, we establish model selection consistency and estimation bounds for the proposed estimators. The theoretical results provide insights into the gain from taking into account the network structural information. Extensive simulation studies indicate that our method outperforms Lasso and elastic net in terms of variable selection accuracy and stability. We apply our method to a breast cancer gene expression study and identify several biologically plausible subnetworks and pathways that are associated with breast cancer distant metastasis.

Key words and phrases: Laplacian penalty, network analysis, regularization, sparsity, survival data, variable selection, weak oracle property.

1. Introduction

With advances in high-throughput technology, gene expression profiling is extensively used to discover new markers, pathways, and new therapeutic targets. This technique measures the expression levels of tens of thousands of genes. In cancer genomics, gene expression levels provide important molecular signatures for cancers, which in turn can be very predictive for cancer recurrence or survival.

To link high-dimensional genomic data to censored survival outcomes, Cox's proportional hazards model (Cox (1972)) is most commonly used, which specifies that the hazard function of a failure time T conditional on a p-dimensional vector of genomic measurements $\mathbf{X}$ takes the form

$$\lambda(t \mid \mathbf{X}) = \lambda_0(t) \exp(\boldsymbol{\beta}_0^T \mathbf{X}), \tag{1.1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}_0$ is a p-vector of regression coefficients. A key feature of genomic data is that the dimensionality p may be much larger than the sample size n, so that traditional methodology cannot be directly applied. To make inferences for the high-dimensional Cox model (1.1), a variety of regularization approaches have been proposed. Of particular interest is the Lasso method (Tibshirani (1996, 1997); Gui and Li (2005)), which can perform estimation and variable selection simultaneously by shrinking some estimates to exactly zero. Alternative methods that exploit sparsity include the SCAD (Fan and Li (2001, 2002)), adaptive Lasso (Zou (2006); Zhang and Lu (2007)), and Dantzig selector (Candes and Tao (2007); Antoniadis, Fryzlewicz and Letué (2010)), among others. All these methods can lead to parsimonious models, which are crucial for achieving good prediction performance and easy interpretation with high dimensionality.

Although the Lasso-type regularization methods have been demonstrated to be very useful in high-dimensional failure time regression, two major drawbacks remain. First, in the linear regression context, the Lasso has been shown to be model selection consistent only under the irrepresentable condition (Zhao and Yu (2006)), which is quite stringent and may not be satisfied in high dimensions because of multicollinearity. Recent developments have also confirmed that similar restrictions exist for survival models (Bradic, Fan and Jiang (2011); Lin and Lv (2013)). Second, these procedures lack a built-in mechanism to incorporate prior structural information about the covariates, which is often available in scientific applications. For instance, in genomic studies, a wealth of knowledge about genes that are functionally similar or belong to the same pathways has accumulated over the years and can be obtained through several publicly available databases. It is expected that taking into account such biological knowledge should help to identify important genes that are functionally related and produce more reliable and biologically more interpretable results.

Several efforts have been made to overcome these drawbacks. The elastic net (Zou and Hastie (2005)) has been applied to high-dimensional Cox regression (Engler and Li (2009); Wu (2012)) to achieve some grouping effects. This method, still, does not utilize any prior information on the graphical structure among the covariates. Wang et al. (2009) proposed hierarchically penalized Cox regression when the variables can be naturally grouped. However, their method is not intended for incorporating any graphical or network structure and, more importantly, their penalty function is nonconvex, which may be a potential issue for efficient computation.

The complexity of genomic data and the aforementioned considerations have motivated us to propose in this paper a network-based regularization method for high-dimensional Cox regression. We aim to incorporate prior gene regulatory network information, as represented by an undirected graph, into the analysis of genomic data and censored survival outcomes. Specifically, our method uses an ℓ_1-penalty to enforce sparsity of the regression coefficients and a quadratic Laplacian penalty to encourage smoothness between the coefficients of neighboring variables on a given network. The resulting optimization problem is convex and allows for an efficient implementation by coordinate descent optimization. Our method extends the work of Li and Li (2010), where only linear regression models were considered. The extension, however, is nontrivial in that new techniques are required for theoretical development under the Cox model. Owing to the semiparametric nature of survival models, high-dimensional analysis of regularization methods for survival data is much more challenging than for (generalized) linear models, and results of this kind are very rare. In fact, even in the special case of ℓ_1-penalized Cox regression, our theoretical results are novel and substantially different from the few available in the literature (e.g., Bradic, Fan and Jiang (2011); Huang et al. (2013); Kong and Nan (2012)). Moreover, our theoretical results provide new insights into the gain from taking into account the covariate graphical structure information. We demonstrate through extensive simulation studies and a real data example that our method outperforms Lasso and elastic net, which do not utilize any prior network information, in terms of variable selection and biological interpretability.

The rest of this paper is organized as follows. In Section 2, we introduce a network-based regularization method for high-dimensional Cox regression and describe a coordinate descent algorithm for implementation. We provide in Section 3 theoretical results in the setting where the dimensionality p may grow exponentially fast with the sample size n, and discuss their consequences and implications. Simulation studies and real data analysis are presented in Sections 4 and 5, respectively. We conclude with a brief discussion in Section 6. Proofs and additional simulation results are relegated to the Appendix and Supplementary Material.

2. Methodology

2.1. Network-regularized Cox regression

We begin by introducing some notation. Let T be the failure time and C the censoring time. Denote by $\widetilde{T} = T \wedge C$ the censored failure time and $\Delta = I(T \leq C)$ the failure indicator, where $I(\cdot)$ is the indicator function. Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a vector of covariates and assume that T and C are conditionally independent given $\mathbf{X}$. The observed data consist of the triples $(\widetilde{T}_i, \Delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, which are independent copies of $(\widetilde{T}, \Delta, \mathbf{X})$. Moreover, we assume that the relationships among the covariates are specified by a network (weighted graph) $G = (V, E, W)$, where $V = \{1, \ldots, p\}$ is the set of vertices corresponding to the p covariates, an element (i, j) in the edge set $E \subset V \times V$ indicates a link between vertices i and j, and $W = (w_{ij})$, $(i, j) \in E$ is the set of weights associated with the edges. For simplicity, we assume that G contains no loops or multiple edges. In practice, the weight of an edge can be used to measure the strength or uncertainty of the link between two vertices. For instance, in a gene regulatory network constructed from data, the weight may indicate the probability that two genes are functionally related. Further, denote by $d_i = \sum_{j:\,(i,j)\in E} w_{ij}$ the degree of vertex i and define the normalized Laplacian matrix $\mathbf{L} = (l_{ij})$ of the graph G by

$$l_{ij} = \begin{cases} 1, & \text{if } i = j \text{ and } d_i \neq 0, \\ -w_{ij}/\sqrt{d_i d_j}, & \text{if } (i, j) \in E, \\ 0, & \text{otherwise.} \end{cases}$$

In the low-dimensional setting, estimation of $\boldsymbol{\beta}_0$ in model (1.1) is based on maximizing the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}^T \mathbf{X}_i)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}^T \mathbf{X}_j)} \right\}^{\Delta_i},$$

where R_i is the index set for the subjects that are at risk just before time $\widetilde{T}_i$. In the high-dimensional setting where the dimensionality p is comparable to or much larger than the sample size n, however, some form of regularization is required. We assume that $\boldsymbol{\beta}_0$ is sparse in the sense that only a small portion of the components of $\boldsymbol{\beta}_0$ are nonzero. We are interested in identifying the nonzero components of $\boldsymbol{\beta}_0$ as well as accurate estimation and prediction.

In the context of linear regression, to obtain a sparse estimate that approximately retains the structure of a given network, Li and Li (2010) introduced a network-constrained penalty,

$$\begin{aligned} p(\boldsymbol{\beta}; \lambda_1, \lambda_2) &= \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \boldsymbol{\beta}^T \mathbf{L} \boldsymbol{\beta} \\ &= \lambda_1 \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{2} \sum_{(i,j) \in E} w_{ij} \left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_j}{\sqrt{d_j}} \right)^2, \end{aligned} \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\lambda_1, \lambda_2 \geq 0$ are two regularization parameters. The penalty (2.1) consists of two parts. The first term is an ℓ_1 part that penalizes the regression coefficients individually and is the key to achieving sparsity and performing variable selection. The second term is a quadratic Laplacian penalty that penalizes on the differences of scaled coefficients between neighboring variables on a given network, thus promoting local smoothness over the network and encouraging simultaneous selection of related variables. The scaling of coefficients by the (square root of) degrees is preferable for two reasons. First, the penalty on each linked pair suggests that the scaling should allow variables with a larger degree to achieve a more dramatic effect. This is often desirable in practice; for example, in genomic studies, genes that are highly connected to others, such as the hub genes, are believed to play a fundamental role in biological processes (Lehner et al. (2006)). Second, in addition to the bias caused by the ℓ_1 part, the quadratic penalty induces extra estimation bias and, without scaling or normalization, a highly connected variable would have been overpenalized and

hence subject to unendurable bias. In fact, the normalized Laplacian matrix has eigenvalues between 0 and 2 (Chung (1997)), leading to a numerically more stable procedure.

Using the penalty specified by (2.1), we propose to estimate $\boldsymbol{\beta}_0$ in the high-dimensional model (1.1) by the penalized partial likelihood estimator

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{-\frac{1}{n}\ell(\boldsymbol{\beta}) + p(\boldsymbol{\beta};\lambda_1,\lambda_2)\right\}, \tag{2.2}$$

where $\ell(\boldsymbol{\beta})$ is the log partial likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\Delta_i\left[\boldsymbol{\beta}^T\mathbf{X}_i - \log\left\{\sum_{j\in R_i}\exp(\boldsymbol{\beta}^T\mathbf{X}_j)\right\}\right]. \tag{2.3}$$

2.2. Accounting for different signs of coefficients

As pointed out by Li and Li (2010), the penalty (2.1) may not perform well when two neighboring variables have opposite signs of regression coefficients, which is reasonable in, e.g., network-based analysis of gene expression data. To address this issue, they proposed a modified version of (2.1),

$$\begin{aligned} p^*(\boldsymbol{\beta};\widetilde{\boldsymbol{\beta}},\lambda_1,\lambda_2) &= \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2}\boldsymbol{\beta}^T\widetilde{\mathbf{L}}\boldsymbol{\beta} \\ &= \lambda_1\sum_{j=1}^{p}|\beta_j| + \frac{\lambda_2}{2}\sum_{(i,j)\in E}w_{ij}\left(\frac{\operatorname{sgn}(\tilde{\beta}_i)\beta_i}{\sqrt{d_i}} - \frac{\operatorname{sgn}(\tilde{\beta}_j)\beta_j}{\sqrt{d_j}}\right)^2, \end{aligned} \tag{2.4}$$

where $\widetilde{\mathbf{L}} = (\tilde{l}_{ij}) = \mathbf{S}^T\mathbf{L}\mathbf{S}$ with $\mathbf{S} = \operatorname{diag}(\operatorname{sgn}(\tilde{\beta}_1),\ldots,\operatorname{sgn}(\tilde{\beta}_p))$ and $\widetilde{\boldsymbol{\beta}} = (\tilde{\beta}_1,\ldots,\tilde{\beta}_p)$ is obtained from a preliminary regression analysis.

Here we motive the penalty (2.4) from another point of view. To account for regression coefficients with opposite signs, it is natural to consider the penalty

$$\begin{aligned} p^{**}(\boldsymbol{\beta};\lambda_1,\lambda_2) &= \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2}|\boldsymbol{\beta}|^T\mathbf{L}|\boldsymbol{\beta}| \\ &= \lambda_1\sum_{j=1}^{p}|\beta_j| + \frac{\lambda_2}{2}\sum_{(i,j)\in E}w_{ij}\left(\frac{|\beta_i|}{\sqrt{d_i}} - \frac{|\beta_j|}{\sqrt{d_j}}\right)^2, \end{aligned} \tag{2.5}$$

where $|\boldsymbol{\beta}| = (|\beta_1|,\ldots,|\beta_p|)^T$. Similar to the penalty (2.1), it explicitly uses the Laplacian matrix as a differential operator, distinguishing them from other network-based penalties such as that considered in Pan, Xie and Shen (2010). To

emphasize this unique feature, we refer to penalties (2.1) and (2.5) as the *Laplacian net* and *absolute Laplacian net*, respectively. Note however that the latter penalty is in general nonconvex, posing challenges for efficient implementation and theoretical analysis. In a similar spirit to the idea of Zou and Li (2008), we propose to use the approximation

$$|\beta_j| \approx |\tilde{\beta}_j| + \text{sgn}(\tilde{\beta}_j)(\beta_j - \tilde{\beta}_j) = \text{sgn}(\tilde{\beta}_j)\beta_j \qquad \text{for} \quad \beta_j \approx \tilde{\beta}_j,$$

in the second term of (2.5), which gives rise to (2.4). Therefore, the penalty (2.4) can be viewed as an adaptive, convex approximation to (2.5) and should inherit the performance of the latter provided that a reasonably good initial estimate $\widetilde{\boldsymbol{\beta}}$ can be obtained. We call the penalty (2.4) the *adaptive Laplacian net*.

We now propose to estimate $\boldsymbol{\beta}_0$ by the adaptively penalized partial likelihood estimator

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left\{-\frac{1}{n}\ell(\boldsymbol{\beta}) + p^*(\boldsymbol{\beta}; \widetilde{\boldsymbol{\beta}}, \lambda_1, \lambda_2)\right\}, \tag{2.6}$$

where $\ell(\boldsymbol{\beta})$ and $p^*(\boldsymbol{\beta}; \widetilde{\boldsymbol{\beta}}, \lambda_1, \lambda_2)$ are defined in (2.3) and (2.4), respectively. Since an ordinary least squares estimator does not perform well or can even fail when p grows fast with n, whereas the Lasso and elastic net produce sparse estimates that may prevent many edges on a given network from being active, we recommend that the initial estimate $\widetilde{\boldsymbol{\beta}}$ be computed from a ridge regression for model (1.1),

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left\{-\frac{1}{n}\ell(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{n}\beta_j^2\right\},$$

where $\lambda \geq 0$ is a regularization parameter. The ridge method does not shrink any coefficient to exactly zero and thus help to preserve and utilize all the information contained in the network. We will demonstrate in our simulation studies and real data analysis that this modified approach can effectively adapt to the different signs of the coefficients and yield very encouraging results. Note that the optimization problem (2.2) is a special case of (2.6) with $\text{sgn}(\tilde{\beta}_i) = \text{sgn}(\tilde{\beta}_j) \neq 0$ for all $(i, j) \in E$; hence, to avoid redundancy, we will present implementation details and theoretical properties only for the latter.

2.3. Implementation

Since the objective function in (2.6) is convex, the optimization problem can be solved by many commonly used algorithms for convex optimization. We now

describe an implementation by coordinate descent, a method that is especially appealing for large-scale sparse problems (Friedman et al. (2007); Wu and Lange (2008)). We adapt the coordinate descent algorithm to network-regularized high-dimensional Cox regression, which turns out to be very efficient.

Denote $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^T = (\boldsymbol{\beta}^T\mathbf{X}_1, \ldots, \boldsymbol{\beta}^T\mathbf{X}_n)^T$. Following Simon et al. (2011), we first approximate $\ell(\boldsymbol{\beta})$ by

$$\tilde{\ell}(\boldsymbol{\beta}; \boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{n} u_i(\boldsymbol{\gamma})\big(y_i(\boldsymbol{\gamma}) - \boldsymbol{\beta}^T\mathbf{X}_i\big)^2,$$

where $u_i(\boldsymbol{\gamma}) = \partial^2\ell(\boldsymbol{\beta})/\partial\gamma_i^2$ and $y_i(\boldsymbol{\gamma}) = \boldsymbol{\gamma}_i - (\partial\ell(\boldsymbol{\beta})/\partial\gamma_i)/u_i(\boldsymbol{\gamma})$. A simple calculation as in Li and Li (2010) yields that the univariate optimization problem

$$\hat{\beta}_j = \arg\min_{\beta_j\in\mathbb{R}}\left\{-\frac{1}{n}\tilde{\ell}(\boldsymbol{\beta}; \boldsymbol{\gamma}) + p^*(\boldsymbol{\beta}; \widetilde{\boldsymbol{\beta}}, \lambda_1, \lambda_2)\right\}$$

has the exact solution

$$\hat{\beta}_j = \frac{\operatorname{sgn}(z_j)(|z_j| - \lambda_1)_+}{n^{-1}\sum_{i=1}^{n} u_i(\boldsymbol{\gamma})X_{ij}^2 + \lambda_2\tilde{l}_{jj}}, \tag{2.7}$$

where

$$z_j = \frac{1}{n}\sum_{i=1}^{n} u_i(\boldsymbol{\gamma})X_{ij}\left(y_i(\boldsymbol{\gamma}) - \sum_{k\neq j}\beta_k X_{ik}\right) - \lambda_2\sum_{k\neq j}\tilde{l}_{jk}\beta_k$$

and X_{ij} is the jth component of $\mathbf{X}_i$. We then obtain the following algorithm for computing the solution to the optimization problem (2.6) for a given pair of regularization parameters (λ_1, λ_2):

Step 1. Initialize $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ and $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\beta}}^T\mathbf{X}_1, \ldots, \widehat{\boldsymbol{\beta}}^T\mathbf{X}_n)^T$.

Step 2. Compute $u_i(\widehat{\boldsymbol{\gamma}})$ and $y_i(\widehat{\boldsymbol{\gamma}})$ for $i = 1, \ldots, n$.

Step 3. Update $\hat{\beta}_j$ by (2.7) cyclically for $j = 1, \ldots, p$ until convergence.

Step 4. Update $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\beta}}^T\mathbf{X}_1, \ldots, \widehat{\boldsymbol{\beta}}^T\mathbf{X}_n)^T$ and repeat Steps 2 and 3 until convergence.

To select the tuning parameters λ_1 and λ_2 in the above algorithm, it is convenient to reparameterize them as $\lambda_1 = \lambda a$ and $\lambda_2 = \lambda(1-a)$, where $\lambda \geq 0$ and $0 \leq a \leq 1$. We first set a to a sufficiently fine grid of values on $[0, 1]$. For each fixed a, set $\lambda_{\max} = (na)^{-1}\max_j\sum_{i=1}^{n} u_i(\mathbf{0})X_{ij}y_i(\mathbf{0})$, which ensures that $\widehat{\boldsymbol{\beta}} = \mathbf{0}$, and let $\lambda_{\min} = \varepsilon\lambda_{\max}$ for some small $\varepsilon \in (0, 1)$. We then compute the solution

path for a decreasing sequence of λ from $\lambda_{\max}$ to $\lambda_{\min}$, at each step using the solution from the previous position as a warm start. Finally, we use K-fold cross-validation to choose the optimal pair (λ, a) that minimizes the cross-validation error

$$\mathrm{CV}(\lambda, a) = -\frac{1}{n}\sum_{k=1}^{K}\Big\{\ell\big(\widehat{\boldsymbol{\beta}}^{(-k)}(\lambda, a)\big) - \ell^{(-k)}\big(\widehat{\boldsymbol{\beta}}^{(-k)}(\lambda, a)\big)\Big\},$$

where $\widehat{\boldsymbol{\beta}}^{(-k)}(\lambda, a)$ is the estimate obtained from excluding the kth part of the data with a given pair of values of (λ, a), and $\ell^{(-k)}(\cdot)$ is the log partial likelihood without the kth part of the data.

3. Theoretical Properties

To state the theoretical properties of the proposed estimators, we adopt the usual counting process notation. For subject i, denote by $N_i(t) = I(\widetilde{T}_i \le t, \Delta_i = 1)$ the counting process for the observed failure and $Y_i(t) = I(\widetilde{T}_i \ge t)$ the at-risk indicator, and denote by $N(t)$ and $Y(t)$ the generic processes. For notational convenience, we write $\mathbf{v}^{\otimes 0} = 1$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$, and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$, for any vector $\mathbf{v}$. Define $\mathbf{S}^{(k)}(\boldsymbol{\beta}, t) = n^{-1}\sum_{j=1}^{n} Y_j(t)\mathbf{X}_j^{\otimes k}\exp(\boldsymbol{\beta}^T\mathbf{X}_j)$, $\mathbf{s}^{(k)}(\boldsymbol{\beta}, t) = E\{Y(t)\mathbf{X}^{\otimes k}\exp(\boldsymbol{\beta}^T\mathbf{X})\}$, $k = 0, 1, 2$, $\overline{\mathbf{X}}(\boldsymbol{\beta}, t) = \mathbf{S}^{(1)}(\boldsymbol{\beta}, t)/S^{(0)}(\boldsymbol{\beta}, t)$, and $\mathbf{e}(\boldsymbol{\beta}, t) = \mathbf{s}^{(1)}(\boldsymbol{\beta}, t)/s^{(0)}(\boldsymbol{\beta}, t)$. Using the counting process notation, the partial likelihood score function can be written as

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\mathbf{X}_i - \overline{\mathbf{X}}(\boldsymbol{\beta}, t)\}\, dN_i(t),$$

where τ is the maximum follow-up time. The performance of the penalized partial likelihood estimators depends critically on the covariance structure reflected by the empirical information matrix

$$\mathcal{I}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{\frac{\mathbf{S}^{(2)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} - \mathbf{S}^{(1)}(\boldsymbol{\beta}, t)^{\otimes 2}\right\} dN_i(t)$$

and its population counterpart

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \int_0^{\tau}\left\{\frac{\mathbf{s}^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \mathbf{s}^{(1)}(\boldsymbol{\beta}, t)^{\otimes 2}\right\} s^{(0)}(\boldsymbol{\beta}, t)\lambda_0(t)\, dt.$$

Also, denote the *augmented* empirical and population information matrices by $\mathcal{I}^*(\boldsymbol{\beta}, \lambda_2) = \mathcal{I}(\boldsymbol{\beta}) + \lambda_2\widetilde{\mathbf{L}}$ and $\boldsymbol{\Sigma}^*(\boldsymbol{\beta}, \lambda_2) = \boldsymbol{\Sigma}(\boldsymbol{\beta}) + \lambda_2\widetilde{\mathbf{L}}$, respectively. Note that $\widetilde{\mathbf{L}}$,

and hence $\boldsymbol{\Sigma}^*(\boldsymbol{\beta}, \lambda_2)$, depends on the initial estimator $\widetilde{\boldsymbol{\beta}}$ through the signs of the coefficients in $\widetilde{\boldsymbol{\beta}}$.

Further, define the *active set* $A = \{j\colon \beta_{0j} \neq 0\}$ and estimated active set $\widehat{A} = \{j\colon \hat{\beta}_j \neq 0\}$, where β_{0j} and $\hat{\beta}_j$ are the jth components of $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$, respectively. Let $s = |A|$ be the number of nonzero coefficients in $\boldsymbol{\beta}_0$, and denote the complement of a set B by B^c. We use sets to index vectors and matrices; for example, $\boldsymbol{\beta}_{0A}$ is the subvector formed by β_{0j} with $j \in A$, and $\boldsymbol{\Sigma}^*_{A^cA}(\boldsymbol{\beta}, \lambda_2)$ is the submatrix formed by the (i, j)th entries of $\boldsymbol{\Sigma}^*(\boldsymbol{\beta}, \lambda_2)$ with $i \in A^c$ and $j \in A$. Finally, let d be a *signal threshold* such that $\min_{j\in A} |\beta_{0j}| \geq d$ and let $\mathcal{B}_0$ be the hypercube $\{\boldsymbol{\beta} \in \mathbb{R}^p\colon \|\boldsymbol{\beta}_A - \boldsymbol{\beta}_{0A}\|_\infty \leq d, \boldsymbol{\beta}_{A^c} = \mathbf{0}\}$, where $\|\cdot\|_\infty$ is the supremum norm. Note that all quantities we have defined so far can depend on the sample size n, and in particular, we allow the dimensions s and p to grow with n.

We need to impose the following conditions:

(C1) $\int_0^\tau \lambda_0(t)\, dt < \infty$ and $P\{Y(\tau) = 1\} > 0$.

(C2) The covariates X_j, $j = 1, \ldots, p$, are bounded and there exists a constant $M > 0$ such that $\sum_{j\in A} |X_j| \leq M$.

(C3) There exists a constant $C_{\min} > 0$ such that

$$\inf_{\boldsymbol{\beta}\in\mathcal{B}_0} \Lambda_{\min}\big(\boldsymbol{\Sigma}^*_{AA}(\boldsymbol{\beta}, \lambda_2)\big) \geq C_{\min},$$

where $\Lambda_{\min}(\cdot)$ denotes the minimum eigenvalue.

(C4) There exists a constant $\alpha \in (0, 1]$ such that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}_0} \|\boldsymbol{\Sigma}^*_{A^cA}(\boldsymbol{\beta}, \lambda_2)\boldsymbol{\Sigma}^*_{AA}(\boldsymbol{\beta}, \lambda_2)^{-1}\|_\infty \leq 1 - \alpha,$$

where $\|\cdot\|_\infty$ is the matrix ∞-norm.

Condition (C1) is standard in the asymptotic theory for the Cox model (Andersen and Gill (1982)). The boundedness assumptions in Condition (C2) are convenient for technical derivations, but are not essential and can be weakened to tail bound conditions as in Lin and Lv (2013). Conditions (C3) and (C4) are two main assumptions for obtaining strong performance guarantees. The former reflects the intuition that the relevant covariates cannot be overly dependent, which is required for estimating the nonzero effects with diverging dimensionality; the latter formalizes the intuition that the set of relevant covariates and the

set of irrelevant covariates cannot be overly correlated, which is intrinsic for distinguishing between these two sets of variables and achieving model selection consistency. In the special case of ℓ_1 regularization, these conditions parallel those in Wainwright (2009) that concern linear regression models, and are also related to those in Bradic, Fan and Jiang (2011) for the Cox model.

Two new messages are conveyed by these conditions. First, since Conditions (C3) and (C4) are imposed on submatrices of the augmented matrix $\mathbf{\Sigma}^*(\boldsymbol{\beta}, \lambda_2)$, a proper choice of λ_2 and $\widetilde{\mathbf{L}}$ can substantially relax the conditions. Specifically, Weyl's inequality (Horn and Johnson (1985)) and the fact that $\widetilde{\mathbf{L}}$ is positive semidefinite entail that $\Lambda_{\min}\big(\mathbf{\Sigma}^*_{AA}(\boldsymbol{\beta}, \lambda_2)\big) \geq \Lambda_{\min}\big(\mathbf{\Sigma}_{AA}(\boldsymbol{\beta})\big)$. Hence, the Laplacian net method tends to improve on the condition number of the sparse information matrix $\mathbf{\Sigma}_{AA}(\boldsymbol{\beta}_0)$ and weaken the restriction imposed by Condition (C3); that is, it has the *conditioning effect*. On the other hand, nonzero entries in the matrix $\mathbf{\Sigma}(\boldsymbol{\beta})$ indicate that the contributions of the corresponding covariates in the partial likelihood score equation are correlated, which are shrunk toward zero by the entries of $\lambda_2\widetilde{\mathbf{L}}$ provided that the choice of $\widetilde{\mathbf{L}}$ correctly captures this relationship; that is, the Laplacian net has the *correlation shrinkage effect*, which helps to relax the restrictions in both Conditions (C3) and (C4). It is worth pointing out that the elastic net, with an identity matrix in place of $\widetilde{\mathbf{L}}$, does not have the latter effect. Note also that the (approximate) sign consistency of the initial estimator $\widetilde{\boldsymbol{\beta}}$ plays a helpful, but not essential, role in achieving these effects through the matrix $\widetilde{\mathbf{L}}$.

Second, in a different nature from the conditions in Bradic, Fan and Jiang (2011), Condition (C4) shows that restrictions on the population information matrix, rather than its empirical counterpart, are sufficient, which can then be viewed as a high-dimensional extension of the classical asymptotic regularity conditions. Such an extension is highly nontrivial and is achieved by a detailed characterization of the uniform convergence of the empirical information matrix, which is provided in the following result.

Proposition 1 (Concentration of empirical matrices)**.** *Under Conditions* (C1)-(C4)*, if* $s = O(n^{1/3})$*, then there exist constants* $D, K > 0$ *such that*

$$P\bigg(\inf_{\boldsymbol{\beta}\in\mathcal{B}_0} \Lambda_{\min}\big(\mathcal{I}^*_{AA}(\boldsymbol{\beta}, \lambda_2)\big) \leq \frac{C_{\min}}{2}\bigg) \leq s^2 D \exp\bigg(-K\frac{n}{s^2}\bigg) \tag{3.1}$$

and

$$P\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}^*_{A^cA}(\boldsymbol{\beta},\lambda_2)\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\|_\infty \geq 1-\frac{\alpha}{2}\right) \leq psD\exp\left(-K\frac{n}{s^3}\right). \quad (3.2)$$

The proof of Proposition 1, which relies on a series of novel concentration inequalities, is given in the Appendix. This result says that with high probability, the empirical matrices satisfy almost the same conditions as those imposed on their population counterparts, which are then needed in subsequent derivations.

The following theorem is our main theoretical result. It shows that, under suitable conditions, the proposed estimators correctly identify the sparse model and are also uniformly consistent in estimating the nonzero effects, i.e., they possess the weak oracle property in the sense of Lv and Fan (2009).

Theorem 1 (Weak oracle property)**.** *In addition to Conditions* (C1)-(C4)*, assume that*

$$\frac{n}{s^3(s\vee\log p)}\to\infty \quad (3.3)$$

and the regularization parameters λ_1 *and* λ_2 *are chosen to satisfy*

$$\frac{n\lambda_1^2}{\log p}\to\infty, \qquad \frac{\lambda_2}{\lambda_1}\|\widetilde{\mathbf{L}}_{\cdot,A}\boldsymbol{\beta}_{0A}\|_\infty<\frac{\alpha}{8}, \qquad \text{and} \quad d>\frac{5\sqrt{s}}{2C_{\min}}\lambda_1, \quad (3.4)$$

where $\widetilde{\mathbf{L}}_{\cdot,A}$ *is the submatrix formed by the columns of* $\widetilde{\mathbf{L}}$ *with index* $j\in A$*. Then there exist constants* $D,K>0$ *such that, with probability at least* $1-D\exp(-Kn\lambda_1^2)-D\exp(-Kn/s^3)\to 1$*, the optimization problem* (2.6) *has a unique solution that satisfies the following properties:*

(a) (Sparsity) $\widehat{\boldsymbol{\beta}}_{A^c}=\mathbf{0}$.
(b) (ℓ_∞-loss) $\|\widehat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_{0A}\|_\infty\leq 5\sqrt{s}\lambda_1/(2C_{\min})$.

The conditions and conclusions of Theorem 1 have important implications. The dimension condition (3.3) allows both s and p to grow with n, at the rates of $s=o(n^{1/4})$ and $\log p=o(n)$, respectively. This setting is especially relevant in genomic studies, where the number of features usually far exceeds the sample size and should be modeled as being exponentially growing with the latter, while the number of relevant features can also grow slightly as more features are included in the analysis. The conditions in (3.4) are the requirements on the choice of the regularization parameters. In particular, the second condition in (3.4) requires

λ_2 to be within a certain proportion of λ_1, depending on the matrix $\widetilde{\mathbf{L}}$ and the signal $\boldsymbol{\beta}_{0A}$. This is reasonable because the bias induced by the quadratic Laplacian penalty should be controlled at a certain level so as not to prevent consistent variable selection; see a related discussion in Hebiri and van de Geer (2011) for linear regression models.

In view of the last condition in (3.4), parts (a) and (b) in Theorem 1 together imply sign consistency (Zhao and Yu (2006)), which is in fact stronger than model selection consistency. The benefit of the Laplacian net method in estimation can be clearly seen from the upper bound in part (b); with appropriately chosen λ_2 and $\widetilde{\mathbf{L}}$, one obtains a larger constant $C_{\min}$ defined in Condition (C3) and hence a smaller estimation loss.

4. Simulation Studies

We conducted simulation studies to evaluate the finite-sample performance of the proposed Laplacian net (Lnet) and adaptive Laplacian net (AdaLnet) methods, and compare them with two popular variable selection procedures, Lasso and elastic net (Enet). We also made comparisons with the Cox regression method with the network-based penalty considered in Pan, Xie and Shen (2010), which is a sum of grouped penalties, each in the form of the ℓ_γ-norm of the two coefficients for a pair of neighboring nodes on a given network (GL_γ). We considered scenarios that are likely to be encountered in genomic studies, with different settings on the strengths and directions of genetic effects.

We simulated gene expression data within an assumed network. Each network consists of 100 disjoint regulatory modules, each with one transcription factor gene (TF) and ten regulated genes, resulting in a total of $p = 1100$ genes. In this setting, $d_i = 10$ for the TFs and $d_i = 1$ for the regulated genes, and $w_{ij} = 1$ between the TFs and their regulated genes and 0 otherwise. The expression value of each TF was generated from a standard normal distribution, and the expression values of the ten regulated genes were generated from a conditional normal distribution with a correlation of ρ between the expressions of these genes and that of the corresponding TF. We set $\rho = 0.7$ for five regulated genes and $\rho = -0.7$ for the other five. This mimics the fact that the TF can either activate or repress the regulated genes. We then generated failure times

from the Cox model

$$\lambda(t \mid \mathbf{X}) = \lambda_0(t) \exp\Big(\sum_{j=1}^{44} \beta_j X_j\Big),$$

which includes only the $s = 44$ relevant genes. The baseline hazard function $\lambda_0(t)$ was specified by a Weibull distribution with shape parameter 5 and scale parameter 2, and censoring times were generated from $U(2, 15)$, resulting in a censoring rate of about 30%. In each setting, the sample size was fixed at $n = 200$ and the simulations were replicated 50 times. We applied fivefold cross-validation to choose the optimal tuning parameters.

We considered six different models. In Model 1, we examined the situation where all genes within the same module have the same directions in their effects on the survival outcome. The coefficients β_j, $j = 1, \ldots, 22$, which correspond to the genes in the first two modules, were generated from the uniform distribution $U(0.1, 1)$, while β_j, $j = 23, \ldots, 44$, were generated from $U(-1, -0.1)$. In Model 2, we allowed more diversity in the directions of genetic effects by assigning a random set of three regulated genes different signs of regression coefficients from the other regulated genes within the same module, while keeping their absolute values the same as in Model 1.

We then considered models where the TFs have stronger effects than the regulated genes, as typically observed in practice. In Model 3, we set the regression coefficients of the four TFs to $(2, -2, 4, -4)$, and those of the regulated genes to $\beta_{\mathrm{TF}}/\sqrt{10}$, where β_{TF} is the coefficient of the corresponding TF. In Model 4, we changed the signs of regression coefficients of three genes in each module as in Model 2. In Model 5, we allowed the ten regulated genes within each module to have different effect sizes, with regression coefficients defined as $\beta_{\mathrm{TF}}/\sqrt{j+4}$ for $j = 1, \ldots, 10$. In Model 6, we changed the signs of regression coefficients of three genes in each module from Model 5. Finally, Models 7 and 8 have the same settings as Models 5 and 6, except that the coefficients of two randomly selected genes in each module were set to zero. Note that only Model 3 assumes that the neighboring genes have the same degree-scaled coefficients.

The variable selection performance of each method is summarized by three measures: sensitivity, specificity, and the Matthews correlation coefficient (MCC)

defined by

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

where TP, TN, FP and FN denote the numbers of true positives, true negatives, false positives and false negatives, respectively. The MCC is an overall measure of variable selection accuracy, and a larger MCC indicates a better variable selection performance.

Simulation results for Models 1 and 2 are reported in Table 1. We observed that, in general, AdaLnet and Enet gave the best overall variable selection performance, while Lasso tended to select too many variables with high false positive rates. In contrast, GL_γ tended to select the smallest number of genes and resulted in the lowest sensitivities. Since the majority of the genes were irrelevant and all methods resulted in sparse models, specificity in all cases was much higher than sensitivity and was comparable among all methods. Enet selected a slightly higher proportion of irrelevant genes and hence had slightly lower specificity compared with AdaLnet. Comparisons of the results for Models 1 and 2 suggest the additional benefit of accounting for different directions of the genetic effects from AdaLnet. In Model 1, since all genes within the same module have equal directions in their effects, Lnet and AdaLnet had similar performance, although AdaLnet showed slightly higher sensitivity because the expression levels of these relevant genes were not always positively correlated. In Model 2, where linked genes may affect the survival outcome in opposite directions, AdaLnet exhibited consistent improvement over Lnet in terms of sensitivity and MCC. All methods had similar estimation performance in terms of mean squared error (MSE). Lasso had a lightly smaller MSE than the other methods at the price of a much worse variable selection performance.

Simulation results for Models 3-8 are summarized in the rest of Table 1 and Table 2, indicating essentially the same trends as for Models 1 and 2. In these settings, where the TFs and regulated genes had different strengths of effects, the improvement of Lnet and AdaLnet over Lasso and Enet was even more dramatic, because the difference in effect sizes has been taken into account by our methods. In addition, AdaLnet always resulted in the highest MCC among the four models considered. GL_γ gave the smallest number of false positives; however, it also had

Table 1. Simulation results for Models 1-4. $(n, p, s) = (200, 1100, 44)$. Sensitivity, specificity, MCC, number of selected genes, number of false positives (FPs), and mean squared error (MSE) were averaged over 50 replicates. Lnet: Laplcian net; AdaLnet: adaptive Laplacian net; Lasso: ℓ_1-penalty; Enet: elastic net; GL_γ: group ℓ_γ-penalty. Standard errors are given in the Supplementary Material.

Method	Sensitivity	Specificity	MCC	# of genes	# of FPs	MSE
			Model 1			
Lnet	0.346	0.997	0.524	18.84	3.60	0.016
AdaLnet	0.395	0.996	0.559	21.47	4.09	0.016
Lasso	0.435	0.950	0.310	72.25	53.13	0.012
Enet	0.407	0.995	0.561	22.77	4.88	0.016
GL_γ	0.233	0.998	0.431	12.66	2.42	0.015
			Model 2			
Lnet	0.442	0.996	0.600	23.54	4.09	0.015
AdaLnet	0.557	0.996	0.682	28.79	4.23	0.015
Lasso	0.465	0.958	0.362	64.32	43.88	0.011
Enet	0.616	0.991	0.675	36.68	9.58	0.015
GL_γ	0.434	0.996	0.594	22.99	3.91	0.014
			Model 3			
Lnet	0.526	0.987	0.591	37.06	13.91	0.070
AdaLnet	0.624	0.995	0.715	33.24	5.77	0.071
Lasso	0.363	0.975	0.346	42.67	26.71	0.067
Enet	0.684	0.986	0.682	44.90	14.79	0.072
GL_γ	0.437	0.999	0.633	20.26	1.05	0.070
			Model 4			
Lnet	0.446	0.996	0.601	24.34	4.71	0.070
AdaLnet	0.633	0.995	0.728	32.62	4.76	0.070
Lasso	0.407	0.974	0.376	45.66	27.76	0.063
Enet	0.661	0.988	0.684	41.96	12.88	0.072
GL_γ	0.541	0.999	0.703	25.22	1.40	0.070

in general lower sensitivity and MCC compared to AdaLnet. Lasso and Enet resulted in large numbers of false positives. The Supplementary Material contains some additional simulation settings where the weights w_{ij} were generated by sample correlation coefficients between two gene expressions, yielding very similar results.

Our algorithm is also very fast: the average computation time for obtaining a single solution path over a grid of 50 points in our simulation setting with $(n, p) = (200, 1100)$ was about 0.7 second, only slightly above the average computation time for the Lasso from the R package `glmnet`.

Table 2. Simulation results for Models 5-8. $(n, p, s) = (200, 1100, 44)$. Sensitivity, specificity, MCC, number of selected genes, number of false positives (FPs), and mean squared error (MSE) were averaged over 50 replicates. Lnet: Laplcian net; AdaLnet: adaptive Laplacian net; Lasso: ℓ_1-penalty; Enet: elastic net; GL_γ: group ℓ_γ-penalty. Standard errors are given in the Supplementary Material.

Method	Sensitivity	Specificity	MCC	# of genes	# of FPs	MSE
			Model 5			
Lnet	0.491	0.989	0.575	10.65	12.08	0.077
AdaLnet	0.567	0.996	0.687	29.55	4.62	0.077
Lasso	0.339	0.977	0.337	39.61	24.71	0.073
Enet	0.649	0.985	0.651	44.83	16.28	0.078
GL_γ	0.377	0.999	0.586	17.46	0.88	0.076
			Model 6			
Lnet	0.439	0.996	0.600	23.52	4.22	0.076
AdaLnet	0.642	0.996	0.732	31.43	3.98	0.076
Lasso	0.404	0.973	0.369	46.74	28.98	0.069
Enet	0.650	0.988	0.675	41.61	13.00	0.078
GL_γ	0.523	0.998	0.686	24.78	1.75	0.076
			Model 7			
Lnet	0.518	0.985	0.553	35.06	16.43	0.067
AdaLnet	0.587	0.992	0.639	29.61	8.48	0.067
Lasso	0.424	0.969	0.349	47.99	32.73	0.061
Enet	0.656	0.983	0.610	42.05	18.42	0.069
GL_γ	0.507	0.994	0.606	24.87	6.62	0.066
			Model 8			
Lnet	0.483	0.993	0.582	24.83	7.45	0.067
AdaLnet	0.641	0.992	0.673	31.84	8.76	0.067
Lasso	0.458	0.969	0.373	49.22	32.74	0.059
Enet	0.676	0.984	0.632	41.10	16.78	0.069
GL_γ	0.564	0.994	0.647	26.41	6.10	0.067

5. Application to a Breast Cancer Gene Expression Study

We illustrate the proposed method by application to analyzing a gene expression data set for patients with lymph-node-negative primary breast cancer reported by Wang et al. (2005). These 286 patients were treated between 1980 and 1995 and did not receive adjuvant systemic therapy, of which 107 (37.4%) developed distant metastases in a median follow-up time of 7.2 years. Gene expression profiles were measured on these patients using Affymetrix HG-U133A arrays. To perform a network-based analysis, we focus our analysis on the genes that can be mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG)

pathways (Kanehisa and Goto (2000)). After merging the gene expression data with the KEGG pathways, we obtained a network consisting of 2563 genes and 15,028 edges. Based on this KEGG network, the edge weight $w_{ij} = 1$ if genes i and j are linked and 0 otherwise, and the node degree d_i is the number of genes that link to gene i. The focus of our analysis is to identify the genes and pathways on the KEGG network that are related to cancer survival.

5.1. Regression coefficients of linked genes on the KEGG network

We first demonstrate that the regression coefficients of linked genes on the KEGG network are closer to each other than randomly selected gene pairs. We have a total of $p = 2,554$ genes with 15,028 edges after removing all isolated genes and loops. For each of these genes, we first obtained the estimated regression coefficient from fitting the Cox model with the expression level of this gene as a covariate. Denote the estimated coefficient for gene i as $\hat{\beta}_i$. We define the difference between the absolute values of scaled coefficients of two linked genes by

$$D_{ij} = \frac{|\hat{\beta}_i|}{\sqrt{d_i}} - \frac{|\hat{\beta}_j|}{\sqrt{d_j}},$$

where d_i is the total number of genes linked to gene i. The sum of absolute differences of all linked genes is given by $D_E = \sum_{(i,j)\in E} |D_{ij}|$, where E is the edge set of all linked genes on the KEGG network.

We obtained $D_E = 2.8645$ for the 15,028 edges of the KEGG network. We then performed a randomization test to see if the regression coefficients of the linked genes are likely to be similar. Specifically, we generated an edge set consisting of randomly selected 15,028 gene pairs out of the total $p(p-1)/2 = 3,260,181$ pairs and calculated D_{E_0} using the same node degrees as in calculating D_E. With 50,000 random edge sets, we obtained the empirical distribution of D_{E_0} as shown in Figure 1. It is clear that the observed D_E is far away from the empirical distribution for randomly selected edge sets, where the range of D_{E_0} is between 10.28 and 15.05. We also observe that the coefficient difference of the linked genes on the KEGG network is much smaller than any of the randomly selected gene pairs, which indicates that the regression coefficients of two linked genes in this data set are more similar than randomly selected gene pairs. This partially supports

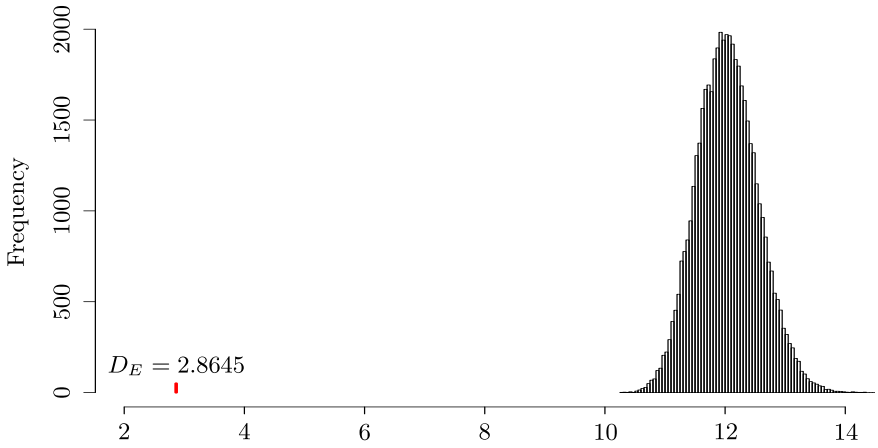

Figure 1. Analysis of breast cancer gene expression data: histogram of the sum of scaled differences between two Cox regression coefficients for 15,028 randomly selected gene pairs based on 50,000 permutations. The vertical bar represents the sum of scaled differences between two Cox regression coefficients for the 15,028 genes pairs on the KEGG network.

our biological intuition that genes connected in the KEGG network should have similar regression coefficients in the Cox model.

5.2. Genes and subnetworks selected

We applied the Lnet, AdaLnet, Lasso, and Enet methods to the data set and used tenfold cross-validation to choose the optimal tuning parameters. Lnet, AdaLnet, Lasso, and Enet selected 98, 140, 62, and 87 genes, respectively. AdaLnet identified many more genes and edges on the KEGG network than Lasso, Enet, and Lnet.

Figure 2 shows the non-isolated genes and associated subnetworks that were identified by these four methods. We observed that AdaLnet selected 47 non-isolated genes, many more than Lasso (14), Enet (19), and Lnet (27). The largest connected component on the subnetwork identified by AdaLnet includes 11 genes, most of which are involved in the mitogen-activated protein kinase (MAPK) pathway. The MAPK pathway participates in fundamental cellular processes such as proliferation, differentiation, migration, and apoptosis, and plays a key role in the development and progression of cancer (Dhillon et al. (2007)). Of particular interest is the well-known oncogene SRC; it has recently been revealed that

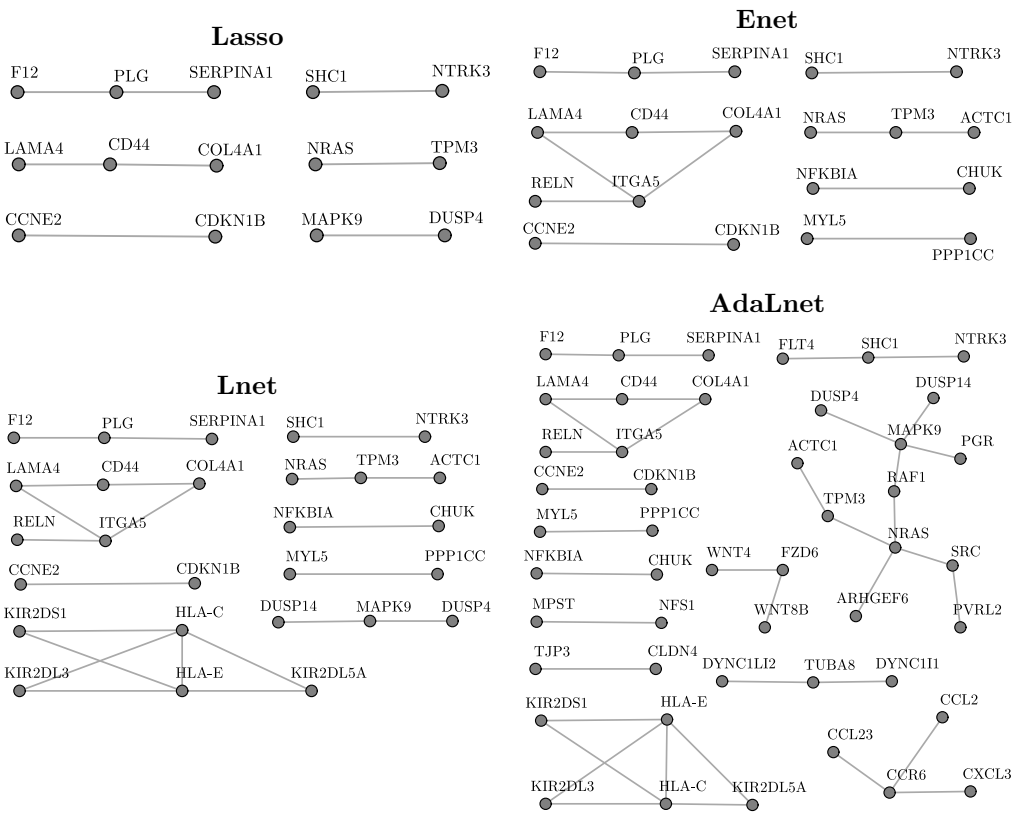

Figure 2. Subnetworks of the KEGG network identified by four different methods applied to the breast cancer gene expression data set. Only non-isolated genes are shown.

Src pathway activity is critical for the survival of disseminated breast cancer cells in the bone marrow microenvironment, leading to an extended period for latent metastasis in breast cancer (Zhang et al. (2009)). This connected subnetwork also includes DUSP4/DUSP14 genes, which negatively regulate members of the mitogen-activated protein (MAP) kinase superfamily (MAPK/ERK) and are associated with cellular proliferation and differentiation (Guan and Butch (1995)). In contrast, although Lasso and Enet identified some links in this subnetwork (e.g., NRAS-TMP3, MAPK9-DUSP4 and NRAS-TPM3-ACTC1), the results from these analyses did not provide strong evidence indicating the involvement of the MAPK pathway in distant metastases of breast cancer.

A second largest component includes two human leukocyte antigen (HLA) class I molecules and three killer immunoglobulin-like receptors (KIRs). It has

been known that altered expression of classical (e.g., HLA-C) and nonclassical (e.g., HLA-E) HLA class I molecules is among the immune escape routes most widely taken by tumor cells (Algarra et al. (2004)). The clinical impact of tumor expression of classical and nonclassical HLAs as well as their interactions have recently been confirmed in a study of 677 early breast cancer patients (de Kruijf et al. (2010)). Another second largest component includes CD44 and integrin $\alpha 5$ (ITGA5), which have been identified as target genes of microRNAs miR-373/520c and miR-31, respectively, in mediating breast cancer metastasis (Valastyan et al. (2009)). It is interesting to note that SRC was not selected by Lasso, Enet, or Lnet, HLA-C and HLA-E not selected by Lasso or Enet, and ITGA5 not selected by Lasso.

The fourth subnetwork identified by AdaLnet involves the inflammatory chemokines CCL2 and CCL23 and its receptor CCR6. A causal role was recently attributed to inflammation in many malignant diseases, including breast cancer. The different inflammatory mediators that are involved in this disease include cells, cytokines, and chemokines, and many studies have addressed the involvement and roles of the inflammatory chemokine CCL2 (MCP-1) in breast malignancy and progression (Soria and Ben-Baruch (2008)). Another subnetwork identified by AdaLnet only includes genes in the Wnt signaling pathway (WNT4, WNT8B, and FZD6), which is also implicated in breast cancer metastasis (Matsuda et al. (2009)).

5.3. Stability selection

We have observed that AdaLnet selected more genes than the other methods, and we now demonstrate that the genes selected are also quite stable. Following Meinshausen and Bühlmann (2010), let S_k be the kth random subsample of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$ without replacement, where $\lfloor x \rfloor$ is the largest integer not greater than x. To balance the censored observations, we sampled half of the censored subjects and half of the uncensored subjects. For a given pair of tuning parameters (λ, α), the selection probability of gene j is defined as

$$\Pr^{(\lambda,\alpha)}(j) = \frac{1}{K} \sum_{k=1}^{K} I\{\hat{\beta}_j^{\lambda,\alpha}(S_k) \neq 0\},$$

Table 3. Summary of stability measurements of the genes selected by four different methods. The minimum (Min), first quantile (Q1), median, mean, third quantile (Q3), and maximum (Max) are shown.

Method	# of genes	Min	Q1	Median	Mean	Q3	Max
All selected genes							
Lasso	62	0.06	0.21	0.30	0.31	0.40	0.65
Enet	87	0.35	0.65	0.79	0.76	0.87	0.99
Lnet	98	0.46	0.71	0.82	0.80	0.91	1.00
AdaLnet	140	0.34	0.60	0.75	0.73	0.87	0.99
Selected genes that are linked on the KEGG network							
Lasso	14	0.12	0.22	0.40	0.37	0.47	0.65
Enet	19	0.44	0.69	0.80	0.78	0.91	0.99
Lnet	27	0.56	0.71	0.84	0.81	0.94	1.00
AdaLnet	47	0.39	0.68	0.80	0.78	0.92	0.99

where $\hat{\beta}_j^{\lambda,\alpha}(S_k)$ is the estimate of β_j using a regularization procedure based on the subsample S_k given the tuning parameters (λ, α), and K is the number of resampling replicates. We used $K = 100$ as suggested by Meinshausen and Bühlmann (2010). A measurement of stability of gene j is then given by $\max_{\lambda,\alpha} \Pr^{(\lambda,\alpha)}(j)$. Table 3 summarizes the stability measurements of the genes selected by each of the four methods. We observed that Lnet resulted in the highest variable selection stability, followed by AdaLnet and Enet. It is also interesting to note that the selected genes that are linked on the KEGG network had in general higher stability than those isolated genes. By encouraging connectivity of the solution, genes that are highly connected in the graph tend to be more often selected, improving stability of the solution.

6. Discussion

We have proposed a network-based regularization method for high-dimensional Cox regression, as a means to incorporate prior network structural information about the covariates. We have provided theoretical results in a general high-dimensional setting that shed light on the benefits of taking into account such structural information. Simulation studies and real data analysis have confirmed the superior performance of our method in terms of variable selection accuracy and stability. In genomic studies, regularization methods that ignore current biological knowledge often result in selection of isolated genes, render-

ing interpretation of the results difficult. In contrast, network-based methods can identify many more functionally related genes and help to bridge the gap between genomic data analysis and understanding of biological mechanisms.

A practical issue in the application of the proposed methodology is to decide which existing biological network to use and how to account for its uncertainty. Choice of the network to use with measured gene expression data depends on the scientific questions asked and whether the network interactions can be reflected at the transcriptional levels. In our analysis of the breast cancer gene expression data, we chose the KEGG pathways and aimed to identify which KEGG subnetworks were associated with distant metastasis. Alternatively, we could focus on the known cancer-related pathways or the large-scale protein-protein interaction network. Instead of using the prior network information, one can build a gene co-expression network from the data and use it to determine the gene neighbors; see Section 3 of Huang et al. (2011) for a discussion of adjacency measures that can be used for the construction of such networks. Finally, if the prior network structure is inaccurate or uninformative, we expect that the tuning parameter λ_2 should be very small and therefore the Laplacian penalty will have almost no effects on variable selection and estimation. Incorporating the uncertainty of the network structure directly into our methodology and theory would be worthwhile future topics.

We have used the convex ℓ_1-penalty to induce sparsity of the regression coefficients to facilitate theoretical analysis and fast computation of a global solution. It would be interesting to explore several nonconvex extensions as in Huang et al. (2011) for linear regression models. If one would replace the ℓ_1-penalty in our method by SCAD or MCP, the main arguments used in this paper could be adapted to establish the oracle property of the modified method, under stronger conditions than those required by the weak oracle property. It is worth noting that the concentration inequalities established in this paper reflect some intrinsic properties of the Cox model in high dimensions and do not depend on any specific penalty function; hence, they will continue to play a pivotal role in the theoretical development of such nonconvex extensions.

We have demonstrated in Section 5.1 that local smoothness of regression coefficients over a gene network may be a biologically plausible assumption. One

can consider alternatively the weaker assumption that two neighboring variables are either both important or both unimportant, which would be more reasonable and likely to be satisfied in broader contexts. Network-based regularization under this assumption could be achieved by a modification of the Laplacian penalty. The discrete nature of the weaker assumption, however, makes the choice of a penalty that allows for efficient implementation much more challenging. These are interesting topics but are beyond the scope of the current paper.

Acknowledgements

This research was supported in part by NIH grants CA127334, GM097505, and GM088566. Hokeun Sun and Wei Lin contributed equally to this work. We thank Professor Wei Pan for sharing his computer code and helpful discussions. We are also grateful to the Co-Editor, Associate Editor, and two referees for constructive comments that have led to substantial improvement in the presentation of the paper.

Appendix: Proofs

We first present a few lemmas that will be essential to the proofs of our main results; their proofs can be found in the Supplementary Material. Note that constants in our proofs may vary from line to line. Lemma 1 provides optimality conditions for the optimization problem (2.6), while Lemmas 2 and 3 characterize the uniform convergence of the large vector $\mathbf{U}(\boldsymbol{\beta}_0)$ and matrix $\mathcal{I}(\cdot)$, respectively, toward their population counterparts.

Lemma 1 (Optimality conditions)**.** *A vector $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is a unique solution to the optimization problem* (2.6) *if the following conditions hold:*

$$\mathbf{U}_{\widehat{A}}(\widehat{\boldsymbol{\beta}}) - \lambda_1 \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_{\widehat{A}}) - \lambda_2 \widetilde{\mathbf{L}}_{\widehat{A},\cdot}\, \widehat{\boldsymbol{\beta}} = \mathbf{0}, \tag{A.1}$$

$$\|\mathbf{U}_{\widehat{A}^c}(\widehat{\boldsymbol{\beta}}) - \lambda_2 \widetilde{\mathbf{L}}_{\widehat{A}^c,\cdot}\, \widehat{\boldsymbol{\beta}}\|_\infty < \lambda_1, \tag{A.2}$$

and $\mathcal{I}^*_{\widehat{A}\widehat{A}}(\widehat{\boldsymbol{\beta}}, \lambda_2)$ is positive definite, where $\widetilde{\mathbf{L}}_{\widehat{A},\cdot}$ and $\widetilde{\mathbf{L}}_{\widehat{A}^c,\cdot}$ are the submatrices formed by the jth rows of $\widetilde{\mathbf{L}}$ with $j \in \widehat{A}$ and $j \in \widehat{A}^c$, respectively.

Lemma 2 (Concentration of $\mathbf{U}(\boldsymbol{\beta}_0)$)**.** *Under Conditions (C1) and (C2), there exist constants $C, D, K > 0$ such that*

$$P\big(|U_j(\boldsymbol{\beta}_0)| \geq Cn^{-1/2}(1+x)\big) \leq D\exp\big(-K(x^2 \wedge n)\big)$$

for all $x > 0$ and $j = 1, \ldots, p$, where $U_j(\boldsymbol{\beta}_0)$ is the jth component of $\mathbf{U}(\boldsymbol{\beta}_0)$.

Lemma 3 (Concentration of $\mathcal{I}(\cdot)$)**.** *Under Conditions (C1) and (C2), there exist constants $C, D, K > 0$ such that*

$$P\bigg(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0} |\mathcal{I}_{ij}(\boldsymbol{\beta}) - \sigma_{ij}(\boldsymbol{\beta})| \geq C\sqrt{s/n}(1+x)\bigg) \leq D\exp\big(-K(sx^2 \wedge n)\big)$$

for all $x > 0$ and $i, j = 1, \ldots, p$, where $\mathcal{I}_{ij}(\cdot)$ and $\sigma_{ij}(\cdot)$ are the (i,j)th entries of $\mathcal{I}(\cdot)$ and $\boldsymbol{\Sigma}(\cdot)$, respectively.

Proof of Proposition 1. By the Hoffman-Wielandt inequality (Horn and Johnson (1985)), we have

$$\begin{aligned}
&\big|\Lambda_{\min}\big(\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)\big) - \Lambda_{\min}\big(\boldsymbol{\Sigma}^*_{AA}(\boldsymbol{\beta},\lambda_2)\big)\big| \\
&\quad\leq \bigg\{\sum_{j=1}^{s}\big|\Lambda_{(j)}\big(\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)\big) - \Lambda_{(j)}\big(\boldsymbol{\Sigma}^*_{AA}(\boldsymbol{\beta},\lambda_2)\big)\big|^2\bigg\}^{1/2} \\
&\quad\leq \|\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2) - \boldsymbol{\Sigma}^*_{AA}(\boldsymbol{\beta},\lambda_2)\|_F = \|\mathcal{I}_{AA}(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_{AA}(\boldsymbol{\beta})\|_F,
\end{aligned}$$

where $\Lambda_{(j)}(\cdot)$ denotes the jth smallest eigenvalue and $\|\cdot\|_F$ is the Frobenius norm. It then follows from Lemma 3 and the union bound that

$$\begin{aligned}
&P\bigg(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\big|\Lambda_{\min}\big(\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)\big) - \Lambda_{\min}\big(\boldsymbol{\Sigma}^*_{AA}(\boldsymbol{\beta},\lambda_2)\big)\big| \geq \frac{C_{\min}}{2}\bigg) \\
&\quad\leq P\bigg(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}_{AA}(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_{AA}(\boldsymbol{\beta})\|_F \geq \frac{C_{\min}}{2}\bigg) \\
&\quad= P\bigg(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\sum_{i,\,j\in A}|\mathcal{I}_{ij}(\boldsymbol{\beta}) - \sigma_{ij}(\boldsymbol{\beta})|^2 \geq \frac{C^2_{\min}}{4}\bigg) \\
&\quad\leq \sum_{i,\,j\in A} P\bigg(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}|\mathcal{I}_{ij}(\boldsymbol{\beta}) - \sigma_{ij}(\boldsymbol{\beta})| \geq \frac{C_{\min}}{2s}\bigg) \leq s^2 D\exp\bigg(-K\frac{n}{s^2}\bigg),
\end{aligned}$$

which, together with Condition (C2), implies (3.1).

To show (3.2), we write

$$\begin{aligned}
&\mathcal{I}^*_{A^cA}(\boldsymbol{\beta},\lambda_2)\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1} - \boldsymbol{\Sigma}^*_{A^cA}(\boldsymbol{\beta},\lambda_2)\boldsymbol{\Sigma}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1} \\
&\quad= \{\mathcal{I}^*_{A^cA}(\boldsymbol{\beta},\lambda_2) - \boldsymbol{\Sigma}^*_{A^cA}(\boldsymbol{\beta},\lambda_2)\}\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1} \\
&\qquad+ \boldsymbol{\Sigma}^*_{A^cA}(\boldsymbol{\beta},\lambda_2)\{\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1} - \boldsymbol{\Sigma}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\} \\
&\quad= \{\mathcal{I}_{A^cA}(\boldsymbol{\beta},\lambda_2) - \boldsymbol{\Sigma}_{A^cA}(\boldsymbol{\beta},\lambda_2)\}\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}
\end{aligned}$$

$$
\begin{aligned}
&- \boldsymbol{\Sigma}^*_{A^cA}(\boldsymbol{\beta},\lambda_2)\boldsymbol{\Sigma}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\{\mathcal{I}_{AA}(\boldsymbol{\beta},\lambda_2)-\boldsymbol{\Sigma}_{AA}(\boldsymbol{\beta},\lambda_2)\}\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\\
&\equiv T_1 - T_2.
\end{aligned}
$$

First consider term T_1. By Lemma 3 and the union bound, we have

$$
\begin{aligned}
&P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}_{A^cA}(\boldsymbol{\beta},\lambda_2)-\boldsymbol{\Sigma}_{A^cA}(\boldsymbol{\beta},\lambda_2)\|_\infty \geq \frac{\alpha}{4}\cdot\frac{C_{\min}}{2\sqrt{s}}\Big)\\
&\quad= P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\max_{i\in A^c}\sum_{j\in A}|\mathcal{I}_{ij}(\boldsymbol{\beta},\lambda_2)-\sigma_{ij}(\boldsymbol{\beta},\lambda_2)| \geq \frac{\alpha}{4}\cdot\frac{C_{\min}}{2\sqrt{s}}\Big)\\
&\quad\leq \sum_{i\in A^c}P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\sum_{j\in A}|\mathcal{I}_{ij}(\boldsymbol{\beta},\lambda_2)-\sigma_{ij}(\boldsymbol{\beta},\lambda_2)| \geq \frac{\alpha}{4}\cdot\frac{C_{\min}}{2\sqrt{s}}\Big) \qquad \text{(A.3)}\\
&\quad\leq \sum_{i\in A^c}\sum_{j\in A}P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}|\mathcal{I}_{ij}(\boldsymbol{\beta},\lambda_2)-\sigma_{ij}(\boldsymbol{\beta},\lambda_2)| \geq \frac{\alpha}{4}\cdot\frac{C_{\min}}{2s^{3/2}}\Big)\\
&\quad\leq (p-s)sD\exp\Big(-K\frac{n}{s^3}\Big).
\end{aligned}
$$

Also, since $\|\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\|_\infty \leq \sqrt{s}\|\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\|_2 = \sqrt{s}/\Lambda_{\min}\big(\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)\big)$, (3.1) implies that

$$
\begin{aligned}
&P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\|_\infty \geq \frac{2\sqrt{s}}{C_{\min}}\Big)\\
&\quad\leq P\Big(\inf_{\boldsymbol{\beta}\in\mathcal{B}_0}\Lambda_{\min}\big(\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)\big) \leq \frac{C_{\min}}{2}\Big) \leq s^2D\exp\Big(-K\frac{n}{s^2}\Big).
\end{aligned} \qquad \text{(A.4)}
$$

Hence, we have

$$
\begin{aligned}
P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|T_1\|_\infty \geq \frac{\alpha}{4}\Big) &\leq P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}_{A^cA}(\boldsymbol{\beta},\lambda_2)-\boldsymbol{\Sigma}_{A^cA}(\boldsymbol{\beta},\lambda_2)\|_\infty \geq \frac{\alpha}{4}\cdot\frac{C_{\min}}{2\sqrt{s}}\Big)\\
&\quad + P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\|_\infty \geq \frac{2\sqrt{s}}{C_{\min}}\Big)\\
&\leq (p-s)sD\exp\Big(-K\frac{n}{s^3}\Big) + s^2D\exp\Big(-K\frac{n}{s^2}\Big).
\end{aligned}
$$

Then consider term T_2. Similar to (A.3), we have

$$
P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}_{AA}(\boldsymbol{\beta},\lambda_2)-\boldsymbol{\Sigma}_{AA}(\boldsymbol{\beta},\lambda_2)\|_\infty \geq \frac{\alpha}{4(1-\alpha)}\cdot\frac{C_{\min}}{2\sqrt{s}}\Big) \leq s^2D\exp\Big(-K\frac{n}{s^3}\Big).
$$

This, together with Condition (C3) and (A.4), leads to

$$\begin{aligned}
&P\bigg(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|T_2\|_\infty \geq \frac{\alpha}{4}\bigg)\\
&\quad\leq P\bigg(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}_{AA}(\boldsymbol{\beta},\lambda_2)-\boldsymbol{\Sigma}_{AA}(\boldsymbol{\beta},\lambda_2)\|_\infty \geq \frac{\alpha}{4(1-\alpha)}\cdot\frac{C_{\min}}{2\sqrt{s}}\bigg)\\
&\qquad+P\bigg(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\|_\infty \geq \frac{2\sqrt{s}}{C_{\min}}\bigg)\leq s^2D\exp\bigg(-K\frac{n}{s^3}\bigg).
\end{aligned}$$

Combining the bounds for T_1 and T_2 gives

$$\begin{aligned}
P\bigg(\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}^*_{A^cA}(\boldsymbol{\beta},\lambda_2)\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}-\boldsymbol{\Sigma}^*_{A^cA}(\boldsymbol{\beta},\lambda_2)\boldsymbol{\Sigma}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\|_\infty \geq \frac{\alpha}{2}\bigg)\\
\leq psD\exp\bigg(-K\frac{n}{s^3}\bigg),
\end{aligned}$$

which, along with Condition (C3), implies (3.2). This completes the proof.

Proof of Theorem 1. The idea of the proof is to first define an "ideal" event that occurs with high probability, and then analyze the behavior of the penalized estimator $\widehat{\boldsymbol{\beta}}$ conditional on that event by using deterministic arguments based on Lemma 1.

First, by Lemma 2 and the union bound, we have

$$P\bigg(\|\mathbf{U}(\boldsymbol{\beta}_0)\|_\infty \geq \frac{\alpha}{8}\lambda_1\bigg)\leq\sum_{j=1}^{p}P\bigg(|U_j(\boldsymbol{\beta}_0)|\geq\frac{\alpha}{8}\lambda_1\bigg)\leq pD\exp(-Kn\lambda_1^2).$$

This inequality, along with (3.1) and (3.2) in Proposition 1, implies that with probability at least $1-pD\exp(-Kn\lambda_1^2)-psD\exp(-Kn/s^3)$, the following inequalities hold:

$$\|\mathbf{U}(\boldsymbol{\beta}_0)\|_\infty<\frac{\alpha}{8}\lambda_1,\qquad\inf_{\boldsymbol{\beta}\in\mathcal{B}_0}\Lambda_{\min}\big(\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)\big)>\frac{C_{\min}}{2},\tag{A.5}$$

and

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|\mathcal{I}^*_{A^cA}(\boldsymbol{\beta},\lambda_2)\mathcal{I}^*_{AA}(\boldsymbol{\beta},\lambda_2)^{-1}\|_\infty<1-\frac{\alpha}{2}.\tag{A.6}$$

We now condition on the event that the above inequalities hold. It suffices to find a $\widehat{\boldsymbol{\beta}}\in\mathbb{R}^p$ that satisfies all the optimality conditions in Lemma 1 and the desired properties. Take $\widehat{\boldsymbol{\beta}}_{A^c}=\mathbf{0}$, and we will determine $\widehat{\boldsymbol{\beta}}_A$ by condition (A.1).

A Taylor expansion of $\mathbf{U}_A(\widehat{\boldsymbol{\beta}})$ gives $\mathbf{U}_A(\widehat{\boldsymbol{\beta}}) = \mathbf{U}_A(\boldsymbol{\beta}_0) - \mathcal{I}_{AA}(\bar{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A})$, where $\bar{\boldsymbol{\beta}}$ lies between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$. Also, we have $\widetilde{\mathbf{L}}_{A,\cdot}\widehat{\boldsymbol{\beta}} = \widetilde{\mathbf{L}}_{AA}\boldsymbol{\beta}_{0A} + \widetilde{\mathbf{L}}_{AA}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A})$. Substituting into the equation $\mathbf{U}_A(\widehat{\boldsymbol{\beta}}) - \lambda_1 \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A) - \lambda_2\widetilde{\mathbf{L}}_{A,\cdot}\widehat{\boldsymbol{\beta}} = \mathbf{0}$ and rearranging yield

$$\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A} = \mathcal{I}^*_{AA}(\bar{\boldsymbol{\beta}}, \lambda_2)^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0) - \lambda_1 \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A) - \lambda_2\widetilde{\mathbf{L}}_{AA}\boldsymbol{\beta}_{0A}\}. \tag{A.7}$$

Define a function $f\colon \mathbb{R}^s \to \mathbb{R}^s$ by $f(\boldsymbol{\theta}) = \boldsymbol{\beta}_{0A} + \mathcal{I}_{AA}(\bar{\boldsymbol{\theta}}, \lambda_2)^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0) - \lambda_1 \operatorname{sgn}(\boldsymbol{\theta}) - \lambda_2\widetilde{\mathbf{L}}_{AA}\boldsymbol{\beta}_{0A}\}$, where $\bar{\boldsymbol{\theta}}_{A^c} = \mathbf{0}$ and $\bar{\boldsymbol{\theta}}_A$ lies between $\boldsymbol{\beta}_{0A}$ and $\boldsymbol{\theta}$. Let $\mathcal{K}$ denote the hypercube $\{\boldsymbol{\theta} \in \mathbb{R}^s\colon \|\boldsymbol{\theta} - \boldsymbol{\beta}_{0A}\|_\infty \le 5\sqrt{s}\lambda_1/(2C_{\min})\}$. Then, by (A.4), (A.5), and the assumption $(\lambda_2/\lambda_1)\|\widetilde{\mathbf{L}}_{\cdot,A}\boldsymbol{\beta}_{0A}\|_\infty < \alpha/8$, we have, for $\boldsymbol{\theta} \in \mathcal{K}$,

$$\begin{aligned}\|f(\boldsymbol{\theta}) - \boldsymbol{\beta}_{0A}\|_\infty &\le \|\mathcal{I}^*_{AA}(\bar{\boldsymbol{\theta}}, \lambda_2)^{-1}\|_\infty\{\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty + \lambda_1 + \lambda_2\|\widetilde{\mathbf{L}}_{AA}\boldsymbol{\beta}_{0A}\|_\infty\} \\ &\le \frac{2\sqrt{s}}{C_{\min}}\left(\frac{\alpha}{8}\lambda_1 + \lambda_1 + \frac{\alpha}{8}\lambda_1\right) \le \frac{5\sqrt{s}}{2C_{\min}}\lambda_1,\end{aligned}$$

i.e., $f(\mathcal{K}) \subset \mathcal{K}$. Also, the assumption $d > 5\sqrt{s}/(2C_{\min})$ entails that $\operatorname{sgn}(\boldsymbol{\theta}) = \operatorname{sgn}(\boldsymbol{\beta}_{0A})$; hence, f is a continuous function on the convex, compact set $\mathcal{K}$. An application of Brouwer's fixed point theorem yields that equation (A.7) has a solution $\widehat{\boldsymbol{\beta}}_A$ in $\mathcal{K}$. Moreover, $\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A) = \operatorname{sgn}(\boldsymbol{\beta}_{0A})$ and hence $\widehat{A} = A$. Thus, we have found a $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ that satisfies (A.1) and the desired properties. Moreover, (A.5) implies that $\mathcal{I}^*_{AA}(\widehat{\boldsymbol{\beta}}, \lambda_2)$ is positive definite.

It remains to verify that $\widehat{\boldsymbol{\beta}}$ also satisfies (A.2). A Taylor expansion of $\mathbf{U}_{A^c}(\widehat{\boldsymbol{\beta}})$ and substituting (A.7) give

$$\begin{aligned}&\mathbf{U}_{A^c}(\widehat{\boldsymbol{\beta}}) - \lambda_2\widetilde{\mathbf{L}}_{A^c,\cdot}\widehat{\boldsymbol{\beta}} \\ &\quad= \mathbf{U}_{A^c}(\boldsymbol{\beta}_0) - \mathcal{I}_{A^cA}(\bar{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}) - \lambda_2\widetilde{\mathbf{L}}_{A^cA}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}) - \lambda_2\widetilde{\mathbf{L}}_{A^cA}\boldsymbol{\beta}_{0A} \\ &\quad= \mathbf{U}_{A^c}(\boldsymbol{\beta}_0) - \mathcal{I}^*_{A^cA}(\bar{\boldsymbol{\beta}}, \lambda_2)(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}) - \lambda_2\widetilde{\mathbf{L}}_{A^cA}\boldsymbol{\beta}_{0A} \\ &\quad= \mathbf{U}_{A^c}(\boldsymbol{\beta}_0) - \mathcal{I}^*_{A^cA}(\bar{\boldsymbol{\beta}}, \lambda_2)\mathcal{I}^*_{AA}(\bar{\boldsymbol{\beta}}, \lambda_2)^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0) - \lambda_1 \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A) - \lambda_2\widetilde{\mathbf{L}}_{AA}\boldsymbol{\beta}_{0A}\} \\ &\qquad- \lambda_2\widetilde{\mathbf{L}}_{A^cA}\boldsymbol{\beta}_{0A}.\end{aligned}$$

Then, by (A.5), (A.6), and the assumption $(\lambda_2/\lambda_1)\|\widetilde{\mathbf{L}}_{\cdot,A}\boldsymbol{\beta}_{0A}\|_\infty < \alpha/8$, we have

$$\begin{aligned}&\|\mathbf{U}_{A^c}(\widehat{\boldsymbol{\beta}}) - \lambda_2\widetilde{\mathbf{L}}_{A^c,\cdot}\widehat{\boldsymbol{\beta}}\|_\infty \\ &\quad\le \|\mathbf{U}_{A^c}(\boldsymbol{\beta}_0)\|_\infty + \|\mathcal{I}^*_{A^cA}(\bar{\boldsymbol{\beta}}, \lambda_2)\mathcal{I}^*_{AA}(\bar{\boldsymbol{\beta}}, \lambda_2)^{-1}\|_\infty \\ &\qquad\times \{\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty + \lambda_1 + \lambda_2\|\widetilde{\mathbf{L}}_{AA}\boldsymbol{\beta}_{0A}\|_\infty\} + \lambda_2\|\widetilde{\mathbf{L}}_{A^cA}\boldsymbol{\beta}_{0A}\|_\infty\end{aligned}$$

$$< \frac{\alpha}{8}\lambda_1 + \left(1 - \frac{\alpha}{2}\right)\left(\frac{\alpha}{8}\lambda_1 + \lambda_1 + \frac{\alpha}{8}\lambda_1\right) + \frac{\alpha}{8}\lambda_1$$
$$\leq \frac{\alpha}{8}\lambda_1 + \frac{\alpha}{8}\lambda_1 + \left(1 - \frac{\alpha}{2}\right)\lambda_1 + \frac{\alpha}{8}\lambda_1 + \frac{\alpha}{8}\lambda_1 = \lambda_1,$$

which verifies (A.2) and concludes the proof.

References

Algarra, I., García-Lora, A., Cabrera, T., Ruiz-Cabello, F. and Garrido, F. (2004). The selection of tumor variants with altered expression of classical and nonclassical MHC class I molecules: Implications for tumor immune escape. *Cancer Immunol. Immunother.* **53**, 904-910.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**, 1100-1120.

Antoniadis, A., Fryzlewicz, P. and Letué, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scand. J. Statist.* **37**, 531-552.

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39**, 3092-3120.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313-2404.

Chung, F. R. K. (1997). *Spectral Graph Theory.* Amer. Math. Soc., Providence, RI.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

de Kruijf, E. M., Sajet, A., van Nes, J. G. H., Natanov, R., Putter, H., Smit, V. T. H. B. M., Liefers, G. J., van den Elsen, P. J., van de Velde, C. J. H. and Kuppen, P. J. K. (2010). HLA-E and HLA-G expression in classical HLA class I-negative tumors is of prognostic value for clinical outcome of early breast cancer patients. *J. Immunol.* **185**, 7452-7459.

Dhillon, A. S., Hagan, S., Rath, O. and Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene* **26**, 3279-3290.

Engler, D. and Li, Y. (2009). Survival analysis with high-dimensional covariates: An application in microarray studies. *Stat. Appl. Genet. Mol. Biol.* **8**, Article 14.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1**, 302-332.

Guan, K.-L. and Butch, E. (1995). Isolation and characterization of a novel dual specific phosphatase, HVH2, which selectively dephosphorylates the mitogen-activated protein kinase. *J. Biol. Chem.* **270**, 7197-9203.

Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21**, 3001-3008.

Hebiri, M. and van de Geer, S. (2011). The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electron. J. Stat.* **5**, 1184-1226.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis.* Cambridge Univ. Press, New York.

Huang, J., Ma, S., Li, H. and Zhang, C.-H. (2011). The sparse Laplacian shrinkage estimator for high-dimensional regression. *Ann. Statist.* **39**, 2021-2046.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013). Oracle inequalities for the lasso in the Cox model. *Ann. Statist.* **41**, 1142-1165.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27-30.

Kong, S. and Nan, B. (2012). Non-asymptotic oracle inequalities for the high-dimensional Cox regression via lasso. *Statist. Sinica*, to appear.

Lehner, B., Crombie, C., Tischler, J., Fortunato, A. and Fraser, A. G. (2006). Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat. Genet.* **38**, 896-903.

Li, C. and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.* **4**, 1498-1516.

Lin, W. and Lv, J. (2013). High-dimensional sparse additive hazards regression. *J. Amer. Statist. Assoc.* **108**, 247-264.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498-3528.

Matsuda, Y., Schlange, T., Oakeley, E. J., Boulay, A. and Hynes, N. E. (2009). WNT signaling enhances breast cancer cell motility and blockade of the WNT pathway by sFRP1 suppresses MDA-MB-231 xenograft growth. *Breast Cancer Res.* **11**, R32.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection (with discussion). *J. Roy. Statist. Soc. Ser. B* **72**, 417-473.

Pan, W., Xie, B. and Shen, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics* **66**, 474-484.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Statist. Software* **39**, 1-13.

Soria, G. and Ben-Baruch, A. (2008). The inflammatory chemokines CCL2 and CCL5 in breast cancer. *Cancer Lett.* **267**, 271-285.

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

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385-395.

Valastyan, S., Reinhardt, F., Benaich, N., Calogrias, D., Szász, A. M., Wang, Z. C., Brock, J. E., Richardson, A. L. and Weinberg, R. A. (2009). A pleiotropically acting microRNA, miR-31, inhibits breast cancer metastasis. *Cell* **137**, 1032-1046.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55**, 2183-2202.

Wang, S., Nan, B., Zhou, N. and Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika* **96**, 307-322.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M. J. J., Atkins, D. and Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-679.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2**, 224-244.

Wu, Y. (2012). Elastic net for Cox's proportional hazards model with a solution path algorithm. *Statist. Sinica* **22**, 271-294.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94**, 691-703.

Zhang, X. H.-F., Wang, Q., Gerald, W., Hudis, C. A., Norton, L., Smid, M., Foekens, J. A. and Massagué, J. (2009). Latent bone metastasis in breast cancer tied to Src-dependent survival signals. *Cancer Cell* **16**, 67-78.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541-2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509-1566.

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, U.S.A.

E-mail: hs2674@columbia.edu

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

E-mail: weilin1@mail.med.upenn.edu

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

E-mail: ruifeng@mail.med.upenn.edu

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

E-mail: hongzhe@upenn.edu