Annals of the Institute of Statistical Mathematics (2023) 75:931–948
https://doi.org/10.1007/s10463-023-00871-9

Gene–environment interaction analysis under the Cox model

Kuangnan Fang[1] · Jingmao Li[1] · Yaqing Xu[2] · Shuangge Ma[3] · Qingzhao Zhang[1,4]

Received: 29 August 2022 / Revised: 21 February 2023 / Accepted: 24 February 2023 /
Published online: 10 April 2023
© The Institute of Statistical Mathematics, Tokyo 2023

Abstract

For the survival of cancer and many other complex diseases, gene–environment (G-E) interactions have been established as having essential importance. G-E interaction analysis can be roughly classified as marginal and joint, depending on the number of G variables analyzed at a time. In this study, we focus on joint analysis, which can better reflect disease biology and is statistically more challenging. Many approaches have been developed for joint G-E interaction analysis for survival outcomes and led to important findings. However, without rigorous statistical development, quite a few methods have a weak theoretical ground. To fill this knowledge gap, in this article, we consider joint G-E interaction analysis under the Cox model. Sparse group penalization is adopted for regularizing estimation and selecting important main effects and interactions. The "main effects, interactions" variable selection hierarchy, which has been strongly advocated in recent literature, is satisfied. Significantly advancing from some published studies, we rigorously establish the consistency properties under high dimensionality. An effective computational algorithm is developed, simulation demonstrates competitive performance of the proposed approach, and analysis of The Cancer Genome Atlas (TCGA) data on stomach adenocarcinoma (STAD) further demonstrates its practical utility.

Keywords Gene–environment interaction analysis · Cox model · Penalized estimation · Asymptotic consistency

✉ Qingzhao Zhang
qzzhang@xmu.edu.cn

Extended author information available on the last page of the article

1 Introduction

For many complex diseases such as cancer, cardiovascular diseases, diabetes, and mental disorders, gene–environment (G-E) interactions have essential importance for risk, prognosis, biomarkers, and response to treatment (Hunter 2005). Extensive biomedical research has been conducted, and accordingly, a myriad of statistical methods has been developed. Roughly, G-E interaction analysis methods can be classified as marginal and joint (Winham and Biernacka 2013; Liu et al. 2020). In marginal analysis, one G variable is analyzed at a time, and the identification of important interactions and main effects is based on p-values. Statistically, challenges mostly come from multiple comparisons adjustment. In joint analysis, all (or a large number of) G variables are analyzed in a single model. As dimensionality can be higher than sample size, regularization is usually needed to control estimation and select important variables. Here, challenges come from the need for simultaneous estimation and selection under high dimensionality. Relatively, joint analysis—which is the focus of this article—can better reflect disease biology (in that outcomes and phenotypes of complex diseases are usually attributable to the collective effects of multiple G variables and their interactions with E variables) and is statistically more challenging. For comprehensive discussions, we refer to Thomas (2010); McAllister et al. (2017); and others.

Literature review suggests that quite a few methods, although having promising numerical performance in simulation and having led to important discoveries, do not have well-established theoretical properties, rendering them a weak statistical ground. Disease outcomes are diverse and so are their statistical models. Relatively, linear models (for continuous outcomes) and generalized linear models (for categorical, count, and other outcomes) have been more extensively studied in terms of theory. Examples include Choi et al. (2010), Wu et al. (2020), Feng et al. (2021), and others, which have established estimation and selection consistency properties under high-dimensional settings comparable to those for linear and generalized linear models. In contrast, survival models have been less studied.

G-E interaction analysis under survival models faces unique challenges. With censoring, many statistical techniques for linear and generalized linear models cannot be applied. In the recent G-E interaction analysis literature, the "main effects, interactions" variable selection hierarchy has been strongly advocated. This hierarchy postulates that, if a G-E interaction is identified as important, the corresponding main G effect needs to be automatically identified. To respect this hierarchy, more complex estimation techniques are needed, bringing in additional complexity (Bien et al. 2013). In addition, certain assumptions that can be made in main-effect-only analysis cannot be made. For example, the (extreme) assumption of independence between all variables, although not sensible, is theoretically possible in main-effect-only analysis. However, it is simply impossible in interaction analysis. Our literature review leads to only a handful of theoretical investigations of G-E interaction analysis under survival models. In Wu et al. (2020), the accelerated failure time (AFT) model is studied. With the linear regression

form of this model, certain techniques developed for linear regression without censoring can be borrowed. In Xu et al. (2019), a censored quantile regression (CQR) approach is developed, which can accommodate outliers with its quantile-based estimation but may have limited applications.

In this article, we conduct joint G-E interaction analysis under the Cox model. Advancing from studies that have focused on numerical investigation, we also rigorously establish asymptotic properties, which can provide the proposed approach a uniquely strong theoretical ground. Compared to the AFT and some other models, the Cox model is more popular in practice. It is also more challenging, as the partial likelihood can be much more complicated than the likelihoods for linear and generalized linear models (Eriksson et al. 2019; Fujimori 2022). An effective computational algorithm is developed, and simulation and data analysis are conducted, complementing the theoretical investigation. Overall, this study can fill an important knowledge gap and provide a theoretically well-grounded and empirically well-behaved method for joint G-E interaction analysis for survival outcomes.

2 Methods

For n i.i.d. subjects, the observed data are $\{(\mathbf{X}_{i,\cdot}, \mathbf{E}_{i,\cdot}, Y_i, \delta_i)\}_{i=1}^n$. Here, for subject i, $\mathbf{X}_{i,\cdot} = (X_{i,1}, \ldots, X_{i,p})$ denotes the p-dimensional G variables, and $\mathbf{E}_{i,\cdot} = (E_{i,1}, \ldots, E_{i,q})$ denotes the q-dimensional E variables. Let T_i and C_i denote the event time and censoring time, respectively. Then $Y_i = \min\{T_i, C_i\}$ and $\delta_i = I(T_i \le C_i)$ denote the observed survival time and censoring indicator, respectively. In practical studies, we often also have demographic, clinical, and other variables. They can be included in the E variables—their interactions with G variables have been advocated in some recent publications. In addition, the proposed analysis can be modified to accommodate the main effects of these variables (without interactions).

Under the Cox model, the conditional hazard function is:

$$\begin{aligned} \lambda_i(t|\mathbf{X}_{i,\cdot}, \mathbf{E}_{i,\cdot}) &= \lambda_0(t)\exp\{m_i(t)\}, \\ m_i(t) &= \sum_{j=1}^{p} X_{i,j}\theta_j + \sum_{j=1}^{p}\sum_{k=1}^{q} E_{i,k}X_{i,j}\mu_{j,k} + \sum_{k=1}^{q} E_{i,k}\eta_k \\ &= \sum_{j=1}^{p} \mathbf{W}_{i,\cdot}^{(j)}\mathbf{b}_j + \mathbf{E}_{i,\cdot}\boldsymbol{\eta} \dot{=} \mathbf{A}_{i,\cdot}\boldsymbol{\phi}, \end{aligned} \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function, θ_j's ($j = 1, \ldots, p$) and η_k's ($k = 1, \ldots, q$) correspond to the main G and E effects, respectively, and $\mu_{j,k}$'s ($j = 1, \ldots, p$, $k = 1, \ldots, q$) correspond to the G-E interactions. Denote $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^\top$, $\mathbf{b}_j = (\theta_j, \mu_{j,1}, \ldots, \mu_{j,q})^\top$ ($j = 1, \ldots, p$), $\boldsymbol{\phi} = (\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_p^\top, \boldsymbol{\eta}^\top)^\top$, and the observed data matrices $\mathbf{W}_{i,\cdot}^{(j)} = (X_{i,j}, X_{i,j}E_{i,1}, \ldots, X_{i,j}E_{i,q})$, $\mathbf{A}_{i,\cdot} = (\mathbf{W}_{i,\cdot}^{(1)}, \ldots, \mathbf{W}_{i,\cdot}^{(p)}, \mathbf{E}_{i,\cdot})$.

We propose penalized estimation with objective function:

$$Q_n(\boldsymbol{\phi}) = -\mathcal{L}_n(\boldsymbol{\phi}) + \sum_{j=1}^{p} P(\mathbf{b}_j;\lambda_1,\lambda_2,\kappa), \tag{2}$$

where $\mathcal{L}_n(\boldsymbol{\phi})$ is the log partial likelihood:

$$\mathcal{L}_n(\boldsymbol{\phi}) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left[\mathbf{A}_{i,\cdot}\boldsymbol{\phi} - \log\left(\sum_{i'\in\mathcal{R}_i}\exp\{\mathbf{A}_{i',\cdot}\boldsymbol{\phi}\}\right)\right], \tag{3}$$

and the at-risk set $\mathcal{R}_i = \{i' : Y_{i'} \geq Y_i\}$.

For penalty, we consider:

$$P(\mathbf{b}_j;\lambda_1,\lambda_2,\kappa) = \rho(\|\mathbf{b}_j\|_2;\sqrt{q+1}\lambda_1,\kappa) + \sum_{k=2}^{q+1}\rho(b_{j,k};\lambda_2,\kappa), \tag{4}$$

where $\|\cdot\|_2$ denotes the L_2−norm, and $\rho(x;\lambda,\kappa)$ is the "base penalty" such as SCAD and MCP—more detailed discussions are provided below. In our numerical study, we adopt MCP, which is defined as $\rho(x;\lambda,\kappa) = \int_0^{|x|}(\lambda - t/\kappa)_+\mathrm{d}t$. λ_1, λ_2 and κ are tuning parameters.

Overall, this is a sparse group penalty (Simon et al. 2013). The regression coefficients corresponding to a G variable—including its main effect and G–E interactions—are treated as a group. The first group penalty determines whether a G variable has any impact on survival at all. If a group norm is zero, then this G variable is concluded as not important. Otherwise, its main effect is not additionally "examined," and the second penalty term determines which interactions are nonzero. By not penalizing the main effect in the second penalty, the "main effects, interactions" variable selection hierarchy is ensured. We note that sparse group penalization has been adopted for G-E interaction analysis under other models (Wu et al. 2018).

Some published studies have adopted composite penalization (Huang et al. 2009) to respect the hierarchy. The two penalizations have certain differences in computational and statistical properties. We defer the study of composite penalization to the future. In most existing G-E interaction studies, E variables have a low dimensionality. In addition, they are often manually selected based on their relevance to the disease of interest. As such, penalization is not imposed to E variables. This can be modified if E variables also have a high dimensionality and contain noises.

2.1 Computation

For optimization, we resort to the alternating direction method of multipliers (ADMM) technique (Boyd et al. 2011). With ADMM, complicated optimization problem (2) can be divided into several simpler sub-problems that can be solved iteratively. ADMM has been extensively applied to high-dimensional problems (Ma and Huang 2015; Zhang et al. 2022; Tang et al. 2021). We note that other techniques, such as iterative coordinate ascent and proximal gradient descent, may also be applicable. Developing alternative algorithms is deferred to future research. Original optimization problem (2) can be reformulated as:

$$\begin{aligned} &\min_{\boldsymbol{\phi},\mathbf{D}} \tilde{Q}(\boldsymbol{\phi},\mathbf{D}) = -\mathcal{L}_n(\boldsymbol{\phi}) + \textstyle\sum_{j=1}^{p} P(\mathbf{d}_j;\lambda_1,\lambda_2,\kappa) \\ &\quad s.t. \quad \mathbf{b}_j = \mathbf{d}_j, \quad j = 1,\ldots,p, \end{aligned} \tag{5}$$

where $\mathbf{d}_j$ ($j = 1,\ldots,p$) is a $(q+1)$-dimensional vector and $\mathbf{D} = (\mathbf{d}_1,\ldots,\mathbf{d}_p)$.

We define the augmented Lagrangian for problem (5) as:

$$\tilde{Q}_{\psi}(\boldsymbol{\phi},\mathbf{D},\boldsymbol{\Upsilon}) = \tilde{Q}(\boldsymbol{\phi},\mathbf{D}) + \frac{\psi}{2}\sum_{j=1}^{p}\left(\|\mathbf{b}_j - \mathbf{d}_j + \boldsymbol{\upsilon}_j\|_2^2 - \|\boldsymbol{\upsilon}_j\|_2^2\right), \tag{6}$$

where $\boldsymbol{\upsilon}_j$ ($j = 1,\ldots,p$) is the scaled dual variable, $\boldsymbol{\Upsilon} = (\boldsymbol{\upsilon}_1,\ldots,\boldsymbol{\upsilon}_p)$, and ψ is the augmented Lagrangian parameter. In our numerical calculation, we set $\psi = 1$. Based on (6), the ADMM algorithm iteratively updates the estimates of $\boldsymbol{\phi}$, $\mathbf{D}$, and $\boldsymbol{\Upsilon}$ until convergence. Specifically, given the estimates $\boldsymbol{\phi}^{(m)}$, $\mathbf{D}^{(m)}$, $\boldsymbol{\Upsilon}^{(m)}$ from the mth iteration, the scaled form ADMM algorithm proceeds as follows:

Step 1. $\boldsymbol{\phi}^{(m+1)} = \arg\min_{\boldsymbol{\phi}} \tilde{Q}_{\psi}(\boldsymbol{\phi},\mathbf{D}^{(m)},\boldsymbol{\Upsilon}^{(m)})$;
Step 2. $\mathbf{D}^{(m+1)} = \arg\min_{\mathbf{D}} \tilde{Q}_{\psi}(\boldsymbol{\phi}^{(m+1)},\mathbf{D},\boldsymbol{\Upsilon}^{(m)})$;
Step 3. $\boldsymbol{\upsilon}_j^{(m+1)} = \boldsymbol{\upsilon}_j^{(m)} + \mathbf{b}_j^{(m+1)} - \mathbf{d}_j^{(m+1)}$, $j = 1,\ldots,p$.

The stopping criteria are:

$$\begin{aligned} \text{Primal feasibility}: \quad & \sum_{j=1}^{p}\|\mathbf{b}_j^{(m+1)} - \mathbf{d}_j^{(m+1)}\|_2^2 < \epsilon^{\text{primal}}, \\ \text{Dual feasibility}: \quad & \sum_{j=1}^{p}\|\mathbf{d}_j^{(m+1)} - \mathbf{d}_j^{(m)}\|_2^2 < \epsilon^{\text{dual}}, \end{aligned} \tag{7}$$

where ϵ^{primal} and ϵ^{dual} are the pre-specified tolerance values.

We now examine the update steps in detail. In Step 1, the optimization problem can be rearranged as $\boldsymbol{\phi}^{(m+1)} = \arg\min_{\boldsymbol{\phi}} \tilde{Q}_{\psi 1}^{(m+1)}(\boldsymbol{\phi})$ with:

$$\tilde{Q}_{\psi 1}^{(m+1)}(\boldsymbol{\phi}) = -\mathcal{L}_n(\boldsymbol{\phi}) + \frac{1}{2}\sum_{j=1}^{p}\|\mathbf{b}_j - \mathbf{d}_j^{(m)} + \boldsymbol{\upsilon}_j^{(m)}\|_2^2. \tag{8}$$

The first-order derivative of $\tilde{Q}_{\psi 1}^{(m+1)}(\boldsymbol{\phi})$ is:

$$\frac{\partial \tilde{Q}_{\psi 1}^{(m+1)}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = -\frac{\partial \mathcal{L}_n(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} + \left[(\mathbf{b}_1 - \mathbf{d}_1^{(m)} + \boldsymbol{\upsilon}_1^{(m)})^{\top},\ldots,(\mathbf{b}_p - \mathbf{d}_p^{(m)} + \boldsymbol{\upsilon}_p^{(m)})^{\top}, \mathbf{0}^{\top}\right]^{\top}, \tag{9}$$

where

$$\frac{\partial \mathcal{L}_n(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left[\mathbf{A}_{i,\cdot}^{\top} - \left\{\sum_{i'\in\mathcal{R}_i}\exp(\mathbf{A}_{i',\cdot}\boldsymbol{\phi})\mathbf{A}_{i',\cdot}^{\top}\right\}\Big/\left\{\sum_{i'\in\mathcal{R}_i}\exp(\mathbf{A}_{i',\cdot}\boldsymbol{\phi})\right\}\right]$$

is the gradient of the log partial likelihood function. Objective function (8) is convex. With gradient (9), we adopt the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method (Nocedal and Wright 2006), which is one of the most efficient quasi-Newton methods, for optimization.

In Step 2, the optimization for $\mathbf{D}$ is separable in $\mathbf{d}_j$'s with:

$$\mathbf{d}_j^{(m+1)} = \arg\min_{\mathbf{d}_j} \frac{1}{2}\|\mathbf{b}_j^{(m+1)} - \mathbf{d}_j + \boldsymbol{\upsilon}_j^{(m)}\|_2^2 + P(\mathbf{d}_j;\lambda_1,\lambda_2,\kappa), \quad j = 1,\ldots,p. \tag{10}$$

To solve these sub-problems, we define the proximal mapping for penalty term $P(\cdot;\lambda_1,\lambda_2,\kappa)$ as follows:

$$\text{prox}(\mathbf{g};\lambda_1,\lambda_2,\kappa) = \arg\min_{\mathbf{d}} \frac{1}{2}\|\mathbf{g}-\mathbf{d}\|_2^2 + P(\mathbf{d};\lambda_1,\lambda_2,\kappa), \tag{11}$$

where $\mathbf{g}$ is a $(q+1)$-dimensional vector. In this case, the solution to (10) follows:

$$\mathbf{d}_j^{(m+1)} = \text{prox}(\mathbf{b}_j^{(m+1)} + \boldsymbol{\upsilon}_j^{(m)};\lambda_1,\lambda_2,\kappa). \tag{12}$$

An iterative method to calculate proximal mapping problem (11) is presented in Algorithm 1, and we refer to the supplementary material for additional details.

Algorithm 1: For the proximal mapping problem $\text{prox}(\mathbf{g};\lambda_1,\lambda_2,\kappa)$.

Input: $(q+1)$-vector $\mathbf{g}$; stopping criterion ϵ;
if $\|\text{ST}_{-1}(\mathbf{g})\|_2 \le \sqrt{q+1}\lambda_1$, **then**
  $\hat{\mathbf{d}} = \mathbf{0}$;
else
  Define index set $\Psi = \{k = 1,\ldots,(q+1) : k = 1 \quad \text{or} \quad |g_k| > \lambda_2\}$ and its complement Ψ^c;
  Set $\hat{\mathbf{d}}_{\Psi^c} = \mathbf{0}$;
  Initialize: $\hat{\mathbf{d}}_\Psi^{(0)} = \mathbf{g}_\Psi$.
  for *m=1,2,...* **do**

$$\hat{\mathbf{d}}_\Psi^{(m)} = \frac{\mathbf{g}_\Psi - \rho'\left(\hat{\mathbf{d}}_\Psi^{(m-1)};\lambda_2,\kappa\right)}{1+\rho'\left(\left\|\hat{\mathbf{d}}_\Psi^{(m-1)}\right\|_2;\sqrt{q+1}\lambda_1,\kappa\right)/\left\|\hat{\mathbf{d}}_\Psi^{(m-1)}\right\|_2}$$

    if $\left\|\hat{\mathbf{d}}_\Psi^{(m)} - \hat{\mathbf{d}}_\Psi^{(m-1)}\right\|_2 < \epsilon$ **then**
      terminate iteration;
    end
  end
end
Output: $(q+1)$-dimensional vector $\hat{\mathbf{d}}_\Psi^{(m)}$.
Remark:
(1) $\text{ST}_{-1}(\mathbf{x};\lambda) = (x_1, [sgn(\mathbf{x}_{-1}) \circ (|\mathbf{x}_{-1}| - \lambda)_+]^\top)^\top$;
(2) $\rho'(x;\lambda,\kappa) = \text{sign}(x) I(|x| < \kappa\lambda)(\lambda - |x|/\kappa)$.

The overall computational algorithm is summarized in Algorithm 2. Following Ma and Huang (2015), we examine convergence properties of the ADMM algorithm in the supplementary material. The proposed method involves three tuning

parameters λ_1, λ_2, and κ. In numerical study, we fix $\kappa = 3$ (Zhang 2010) and conduct a grid search using the extended Bayesian information criterion (EBIC) (Luo et al. 2015) to select λ_1 and λ_2 by minimizing:

$$\text{EBIC}(\lambda_1, \lambda_2) = -2n\mathcal{L}_n\big(\hat{\boldsymbol{\phi}}(\lambda_1, \lambda_2)\big) + \log(n)\hat{s} + 2\omega \log\{\chi(\hat{s})\},$$

where $\hat{\boldsymbol{\phi}}(\lambda_1, \lambda_2)$ is the estimated coefficient for given (λ_1, λ_2), $\hat{s}$ denotes the number of nonzero coefficients in $\hat{\boldsymbol{\phi}}(\lambda_1, \lambda_2)$, $\chi(\hat{s}) = \binom{pq+p+q}{\hat{s}}$, and $\omega \in (0, 1)$ is a parameter. In our analysis, we set $\omega = 1 - \log(n)/[2\log(pq + p + q)]$ following Chen and Chen (2008).

Algorithm 2: ADMM algorithm for the proposed method

Input: data $\{\mathbf{A}_{i\cdot}, Y_i, \sigma_i\}_{i=1}^n$; tuning parameters λ_1, λ_2, κ; stopping criteria ϵ^{primal}, ϵ^{dual}.

Initialize: $\boldsymbol{\phi}^{(0)}$, $\mathbf{D}^{(0)}$, $\boldsymbol{\Upsilon}^{(0)}$

for $m = 0, 1, 2, \ldots$ **do**
  // Step 1. Updata $\boldsymbol{\Phi}$
  update $\boldsymbol{\phi}^{(m+1)}$ based on (8) using the BFGS method;
  // Step 2. Update $\mathbf{D}$
  for $j = 1, 2, \ldots, p$ **do**
    update $\boldsymbol{d}_j^{(m+1)}$ based on (12);
  end
  // Step 3. Update $\boldsymbol{\Upsilon}$
  for $j = 1, 2, \ldots, p$ **do**
    update $\boldsymbol{\upsilon_j}^{(m+1)} = \boldsymbol{\upsilon_j}^{(m)} + \mathbf{b}_j^{(m+1)} - \mathbf{d}_j^{(m+1)}$;
  end
  // Stopping condition
  if *Stopping condition (7) holds* **then**
    terminate iteration;
  end
end

Output: $\hat{\phi} = \boldsymbol{\phi}^{(m+1)}$.

2.2 Theoretical properties

Here we establish that the proposed estimate enjoys the well-desired estimation and selection consistency properties under high-dimensional settings.

Consider the scenario where the number of G variables p increases as the sample size increases with $\log p = O(n^{\zeta_0})$, $\zeta_0 > 0$, while the number of E variables q is fixed. Denote $\boldsymbol{\phi}^* = (\mathbf{b}_1^{*\top}, \ldots, \mathbf{b}_p^{*\top}, \boldsymbol{\eta}^{*\top})^\top$ as the true value of $\boldsymbol{\phi}$. Sparsity is assumed for both the main G effects and interactions. That is, only a subset of the components of $\boldsymbol{\phi}^*$ is nonzero. Let $\Psi = \text{supp}(\boldsymbol{\phi}^*)$, and Ψ^c and s be the complement and cardinality of Ψ, respectively. Assume that s may also increase with n with $s = O(n^{\zeta})$, $\zeta \in (0, 1)$. For convenience, for any $\big[p(q+1)+q\big]$–dimensional vector $\boldsymbol{\phi}$, denote $\boldsymbol{\phi}_1$ ($\boldsymbol{\phi}_2$) as the

subvector of $\boldsymbol{\phi}$ with elements indexed by Ψ (Ψ^c). Similar notations are also adopted for any square matrix with size $p(q+1)+q$. For example, $\mathbf{G}_{12}$ denotes the submatrix of matrix $\mathbf{G}$ with its rows and columns indexed by Ψ and Ψ^c, respectively.

The true coefficient vector $\boldsymbol{\phi}^*$ is assumed to respect "main effects, interactions" hierarchy. Define $\Phi = \{j : \theta_j^* \neq 0\}$ as the index set of important main effects. For $j \in \Phi$, denote Ξ_j and Ξ_j^c as the index sets of nonzero and zero elements in $\mathbf{b}_j$, respectively. Then, coefficients $\boldsymbol{\eta}$ and $\mathbf{b}_{j,\Xi_j}$, $j \in \Phi$ constitute the nonzero coefficients $\boldsymbol{\phi}_1$, while $\mathbf{b}_{j,\Xi_j^c}$, $j \in \Phi$ and $\mathbf{b}_j$, $j \in \Phi^c$ constitute the zero coefficients $\boldsymbol{\phi}_2$.

Before proceeding further, we first make a cautious note. Although the "main effects, interactions" hierarchy has been commonly adopted and argued as statistically highly sensible, some biological studies have suggested that it may be violated. When this happens, we can revise the definition of main effects as $\Phi = \{j : \mathbf{b}_j^* \neq \mathbf{0}\}$. This means that a gene will be identified as important if its main effect and/or some of its interactions are identified. With this revision, the proposed method and theoretical development can follow through.

To facilitate theoretical development, we introduce the counting process representation of the log partial likelihood and some related notations (Bradic et al. 2011). Let $N_i(t) = I(Y_i \leq t; \delta_i = 1)$, $\bar{N}(t) = \sum_{i=1}^n N_i(t)$, and $\mathcal{Y}_i(t) = I(Y_i \geq t)$. With a slight abuse of notation, denote the intensity process of $N_i(t)$ as $\lambda_i(t) = \lambda_0(t)\mathcal{Y}(t)\exp\{\mathbf{A}_{i\cdot}\boldsymbol{\phi}^*\}$, and let $\Lambda_i(t) = \int_0^t \lambda_i(u)\mathrm{d}u$. Then, $M_i(t) = N_i(t) - \Lambda_i(t)$ is an orthogonal local square integrable martingale.

Define:

$$S_n^{(l)}(t,\boldsymbol{\phi}) = \frac{1}{n}\sum_{i=1}^n \mathcal{Y}_i(t)\left(\mathbf{A}_{i,\cdot}^\top\right)^{\otimes l}\exp(\mathbf{A}_{i,\cdot}\boldsymbol{\phi}), \quad l = 0,1,2.$$

Here $\otimes$ denotes the outer product, where for vector $\mathbf{z}$, $\mathbf{z}^{\otimes 0} = 1$, $\mathbf{z}^{\otimes 1} = \mathbf{z}$, and $\mathbf{z}^{\otimes 2} = \mathbf{z}\mathbf{z}^\top$. We can rewrite the log partial likelihood using counting process notations as:

$$\mathcal{L}_n(\boldsymbol{\phi}) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{\mathbf{A}_{i,\cdot}\boldsymbol{\phi} - \log\left[nS_n^{(0)}(t,\boldsymbol{\phi})\right]\right\}\mathrm{d}N_i(t),$$

where τ is the maximum follow-up time. Define $\big[p(q+1)+q\big]$-dimensional vector $\mathbf{H}_n(t,\boldsymbol{\phi}) = S_n^{(1)}(t,\boldsymbol{\phi})/S_n^{(0)}(t,\boldsymbol{\phi})$ and $\big[p(q+1)+q\big]$-dimensional square matrix $\mathbf{V}_n(t,\boldsymbol{\phi}) = S_n^{(2)}(t,\boldsymbol{\phi})/S_n^{(0)}(t,\boldsymbol{\phi}) - \left[S_n^{(1)}(t,\boldsymbol{\phi})/S_n^{(0)}(t,\boldsymbol{\phi})\right]^{\otimes 2}$. Furthermore, to simplify notation, for $\big[p(q+1)+q\big]$-dimensional vector $\boldsymbol{\phi}$ with $\boldsymbol{\phi}_2 = \mathbf{0}$, we denote $S^{(l)}(t,\boldsymbol{\phi}_1) = S^{(l)}(t,\boldsymbol{\phi})\,(l = 0,1,2)$. Similar notations are adopted for $\mathbf{H}_n(t,\boldsymbol{\phi})$, $\mathbf{V}_n(t,\boldsymbol{\phi})$, and some other related functions introduced latter. For two sequences a_n and b_n,

$a_n \asymp b_n$ means that a_n and b_n grow in the same rate, $a_n \gg b_n$ (or $b_n \ll a_n$) means that a_n grows in a faster rate than b_n. The following conditions are assumed.

Condition 1 There exists a compact neighborhood $\mathcal{B}$ of $\boldsymbol{\phi}^*$ such that there are functions $s^{(l)}(t, \boldsymbol{\phi})$ $(l = 0, 1, 2)$ that satisfy:

(a) Functions $s^{(l)}$ $(l = 0, 1, 2)$ are bounded, and $s^{(0)}$ is bounded away from 0 on $\mathcal{B} \times [0, \tau]$. The family of functions $s^{(l)}(t, \cdot)$, $t \in [0, \tau]$ is equicontinuous on $\boldsymbol{\phi}^*$;
(b) With probability tending to 1, as $n \to \infty$,

$$\sup_{t\in[0,\tau],\boldsymbol{\phi}_1\in\mathcal{B}_1} \|S_n^{(l)}(t, \boldsymbol{\phi}_1) - s^{(l)}(t, \boldsymbol{\phi}_1)\|_2 \to 0.$$

(c) Denote $\mathbf{h}(t, \boldsymbol{\phi}) = s^{(1)}(t, \boldsymbol{\phi}_1)/s^{(0)}(t, \boldsymbol{\phi}_1)$ and the sequences $c_n = \sup_{t\in[0,\tau)} \|\mathbf{H}_n(t, \boldsymbol{\phi}^*) - \mathbf{h}(t, \boldsymbol{\phi}^*)\|_\infty$ and $d_n = \sup_{t\in[0,\tau)} |S_n^{(0)}(t, \boldsymbol{\phi}^*) - s^{(0)}(t, \boldsymbol{\phi}^*)|$. Then c_n and d_n are bounded almost surely.

Condition 2 Denote $\epsilon_{i,j} = \int_0^\tau \left(A_{i,j} - h_j(t, \boldsymbol{\phi}^*)\right) \mathrm{d}M_i(t)$ $(i = 1, \ldots, n,\ j = 1, \ldots, p(q + 1) + q)$. Assume that the Cramer condition holds for $\epsilon_{i,j}$. That is,

$$E|\epsilon_{i,j}|^m = m! C^{m-2} \sigma_j^2/2,$$

where C is a positive constant, $m \geq 2$, and ${\sigma_j}^2 = var(\epsilon_{i,j}) < \infty$.

Condition 3 Define:

$$\mathbf{v}(t, \boldsymbol{\phi}_1) = \frac{s_{11}^{(2)}(t, \boldsymbol{\phi}_1)}{s^{(0)}(t, \boldsymbol{\phi}_1)} - \left[\frac{s_1^{(1)}(t, \boldsymbol{\phi}_1)}{s^{(0)}(t, \boldsymbol{\phi}_1)}\right]^{\otimes 2}, \quad \boldsymbol{\Sigma}(t, \boldsymbol{\phi}_1) = \int_0^t \mathbf{v}(u, \boldsymbol{\phi}_1) s^{(0)}(u, \boldsymbol{\phi}_1^*) \mathrm{d}\Lambda_0(u).$$

Let $\boldsymbol{\Sigma}(\boldsymbol{\phi}_1) = \boldsymbol{\Sigma}(\tau, \boldsymbol{\phi}_1)$. Assume that $\boldsymbol{\Sigma}(\boldsymbol{\phi}_1)$ is positive definite for all n. $\sigma_{\min}\left[\boldsymbol{\Sigma}(\boldsymbol{\phi}_1^*)\right] \asymp 1$, where $\sigma_{\min}[\cdot]$ denotes the minimal eigenvalue of a square matrix.

Condition 4

$$E\left\{\sup_{0\leq t\leq\tau} Y_i(t) \|\mathbf{A}_{i,\Psi}\|_2^2 \exp\{\mathbf{A}_{i,\Psi}\boldsymbol{\phi}_1^*\}\right\} = O(s).$$

Condition 5

$$\sup_{0\leq t\leq\tau} \sup_{\boldsymbol{\phi}_1\in\mathcal{B}(\boldsymbol{\phi}_1^*,\phi_{\min})} \|\mathbf{V}_{n21}(t, \boldsymbol{\phi}_1)\|_{2,\infty} = O_p(n^{\alpha_1}),$$

where $\mathcal{B}(\boldsymbol{\phi}_1^*, \boldsymbol{\phi}_{\min})$ is an s–dimensional ball with center $\boldsymbol{\phi}_1$ and radius $\phi_{\min} = \|\boldsymbol{\phi}_1^*\|_\infty$, $\|\mathbf{V}_{n21}(t, \boldsymbol{\phi}_1)\|_{2,\infty} = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{V}_{n21}(t, \boldsymbol{\phi}_1)\mathbf{x}\|_\infty$ denotes the $L_{2,\infty}$-norm of $\mathbf{V}_{n21}(t, \boldsymbol{\phi}_1)$, and α_1 is a positive constant.

Condition 6 The minimal signals satisfy:

$$b_{\min} = \min_{j\in\mathcal{A}} \|\mathbf{b}^*_{j,\Xi_j}\|_2 \gg \lambda_1$$
$$\mu_{\min} = \min_{j\in\mathcal{A},k\in\Xi_j\setminus\{1\}} |b^*_{j,k}| \gg \lambda_2.$$

Condition 7 $\rho(x;\lambda,\kappa)$ is a symmetric function of x. It is nondecreasing and concave on $[0,+\infty)$. $\rho(0;\lambda,\kappa)=0$. $\rho(x;\lambda,\kappa)$ is constant for $|x|\geq\kappa\lambda$. Its first-order derivative $\rho'(x;\lambda,\kappa)$ exists and is continuous for $x\in[0,\infty)$. $\rho'(0+;\lambda,\kappa)=\lambda$.

Conditions 1–5 have been commonly assumed for the Cox model under penalization. Specifically, Condition 1 contains some commonly adopted assumptions for the Cox model (Andersen and Gill 1982; Stute and Wang 1993). We can verify it following a similar path as Theorem 8.4.1 in Fleming and Harrington (2011). Condition 2 describes the tail behavior of $\epsilon_{i,j}=\int_0^\tau A_{i,j}-h_j(t,\boldsymbol{\phi}^*)\mathrm{d}M_i(t)$. If we assume that the G and E variables are bounded, Condition 2 is satisfied. Condition 3 requires that the "Fisher information matrix" for the important variables is positive definite with its eigenvalues not vanishing to 0. It is a commonly adopted condition for high-dimensional models (Bradic et al. 2011; Huang et al. 2013). Condition 5 is the "irrepresentable" condition (Zhao and Yu 2006) for censored data. It controls the uniform growth rate of the covariance matrices between the important variables and unimportant variables. With a folded concave penalty, the upper bound on the right-hand side in Condition 5 can grow to infinity at a polynomial rate. Condition 6 is the minimal signal condition, and comparable conditions have been commonly assumed in the literature. Owing to the sparse group penalty, we do not impose an explicit condition on the minimal signal for the important main effects but assume a condition for the minimal signal of the important groups and interaction effects. Condition 7 is on the folded concave penalty function $\rho(\cdot;\lambda,\kappa)$ and is commonly assumed in the literature (Fan and Li 2001). With these conditions, we can establish the following results.

Theorem 1 *Assume that Conditions 1-7 hold. If* $\max_j(\sigma_j^2)=O(n^{0.5\zeta+\alpha_1})$, $\lambda_1\gg n^{0.5\zeta+\alpha_1-0.5}$, $\lambda_2\gg n^{0.5\zeta+\alpha_1-0.5}$ *and* $\zeta_0<0.5\zeta+\alpha_1$, *then there exists a local minimizer* $\hat{\boldsymbol{\phi}}$ *of* $\mathcal{Q}_n(\boldsymbol{\phi})$ *in* (2) *such that*:

$$(a)\quad \left\|\hat{\boldsymbol{\phi}}_1-\boldsymbol{\phi}_1^*\right\|_2=O_p\left(\sqrt{s/n}\right),\quad (b)\quad \hat{\boldsymbol{\phi}}_2=\mathbf{0}.$$

Furthermore, if $s=o\left(n^{1/3}\right)$, *i.e.*, $\zeta<1/3$, *then for any* $s\times 1$ *unit vector* $\boldsymbol{\nu}_n$,

$$(c)\quad \sqrt{n}\boldsymbol{\nu}_n^\top\boldsymbol{\Sigma}(\boldsymbol{\phi}_1^*)^{1/2}\left(\hat{\boldsymbol{\phi}}_1-\boldsymbol{\phi}_1^*\right)\to_d\mathcal{N}(0,1).$$

Theorem 1 first states that, under high-dimensional settings (with an exponential rate), the proposed approach has estimation consistency. Additionally, for the zero coefficients, result (b) establishes that they will be estimated as zero. The minimal signal condition (Condition 6) and result (a) also lead to that the nonzero coefficients will be estimated as nonzero. As such, the proposed approach has variable selection consistency. Result (c) further establishes the asymptotic behavior of the estimates

of the nonzero coefficients. Overall, the results are comparable with those under simpler settings. The proof is presented in the supplementary material. Although certain components of the proof share some similar spirit with the literature, with the uniqueness of the Cox model, the developments are nontrivial.

3 Simulation

Data are generated as follows. (a) We set $n = 350$, $p = 600$, and $q = 5$. The G variables are generated from multivariate normal distributions, that is, $\mathbf{X}_{i,\cdot} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_1)$, mimicking gene expression data. For the variance matrix $\boldsymbol{\Sigma}_1$, we consider the following structures: AR(0.3), AR(0.7), Band1, Band2, CS(0.2), and CS(0.4). Under the auto-regressive correlation structure AR(ρ), $\rho \in (0, 1)$, $[\boldsymbol{\Sigma}_1]_{i,j} = \rho^{|i-j|}$. Under the first banded correlation structure Band1, $[\boldsymbol{\Sigma}_1]_{i,j} = I(i = j) + 0.4I(|i - j| = 1)$. Under the second banded correlation structure Band2, $[\boldsymbol{\Sigma}_1]_{i,j} = I(i = j) + 0.6I(|i - j| = 1) + 0.2I(|i - j| = 2)$. Under the compound symmetry structure CS(ρ), $\rho \in (0, 1)$, $[\boldsymbol{\Sigma}_1]_{i,j} = I(i = j) + \rho I(i \neq j)$. (b) For the five E variables, we simulate three continuous and two binary ones. In particular, we first generate a 5-dimensional variable $\mathbf{U}_{i,\cdot} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_2)$, where, in each simulated case, $\boldsymbol{\Sigma}_2$ has the same correlation structure as $\boldsymbol{\Sigma}_1$. Then, for the three continuous variables, $E_{i,k} = U_{i,k}$ $(k = 1, 2, 3)$, and for the two binary variables, $E_{i,k} = I(U_{i,k} > -0.7)$ $(k = 4, 5)$. (c) The nonzero regression coefficients are set to respect the "main effects, interactions" hierarchy. Their values are randomly generated from Unif(0.4, 0.8). (d) The survival times are generated with conditional hazard function $\lambda_i(t|\mathbf{X}_{i,\cdot}, \mathbf{E}_{i,\cdot}) = \exp\{1 + \mathbf{A}_{i,\cdot}\boldsymbol{\phi}\}$. We set $\lambda_0(t) = \exp(1)$ (and note that we have also examined time-dependent baseline hazard functions and made similar observations). The censoring times are generated from exponential distributions, whose parameters are adjusted to achieve a $\sim 40\%$ censoring rate.

Additionally, we consider the following four examples, which have different numbers of important main effects and interactions.

Example 1: There are 5 main E effects, 16 main G effects, and 12 G-E interactions. The "locations" of the important G effects and interactions are randomly selected (while ensuring that the variable selection hierarchy is respected).
Example 2: There are 5 main E effects, 8 main G effects, and 6 G-E interactions. The other settings are consistent with Example 1.
Example 3: There are 5 main E effects, 12 main G effects, and no G-E interactions.
Example 4: There are 5 main E effects, 4 main G effects, and 20 G-E interactions. The locations of the main G effects are randomly selected. All the G-E interactions corresponding to the important main G effects are important.

Among them, Examples 1 and 2 are "standard," and similar settings have been considered in the literature. Examples 3 and 4 are two extreme scenarios: One has no G-E interactions, and the other has full interactions.

Beyond the proposed approach, we also consider the following highly relevant alternatives: (a) MCP, which has the same goodness-of-fit function as the proposed approach and applies MCP to individual coefficients. It does not have a mechanism to ensure the "main effects, interactions" variable selection hierarchy. (b) grMCP, which has the same goodness-of-fit function as the proposed approach and applies group MCP. That is, the penalty is $\sum_{j=1}^{p} \rho(\|\mathbf{b}_j\|_2; \sqrt{q+1}\lambda, \kappa)$. Here, all the interactions corresponding to an important G variable are selected. (c) Marginal, which analyzes one G variable at a time and selects important main

Table 1 Simulation results under the AR(0.3) and AR(0.7) correlation structures. In each cell, mean (sd) based on 100 replicates

Example	Correlation	Method	Main effects (M)		Interactions (I)		SSE	C-index
			TP	FP	TP	FP		
1	AR(0.3)	Proposed	12.2(3.0)	0.1(0.4)	6.0(2.4)	1.9(1.5)	5.624(1.462)	0.810(0.034)
		MCP	8.2(2.3)	0.1(0.4)	6.0(2.4)	2.9(2.0)	7.402(1.273)	0.776(0.037)
		grMCP	7.0(4.1)	0.0(0.0)	8.1(2.8)	26.7(18.3)	7.642(1.772)	0.759(0.057)
		Marginal	0.9(1.3)	2.9(3.6)	2.3(1.6)	20.7(20.2)	-(-)	-(-)
	AR(0.7)	Proposed	14.1(1.9)	2.1(1.9)	5.8(2.2)	4.6(3.2)	5.595(1.400)	0.854(0.037)
		MCP	9.2(2.9)	1.4(1.5)	6.0(2.2)	6.2(3.3)	7.076(1.351)	0.830(0.043)
		grMCP	10.3(5.2)	0.6(0.6)	9.6(2.7)	44.7(26.1)	6.799(1.876)	0.819(0.072)
		Marginal	0.9(1.2)	5.9(9.4)	1.2(1.3)	35.0(45.1)	-(-)	-(-)
2	AR(0.3)	Proposed	7.2(1.6)	0.1(0.5)	4.8(1.3)	2.1(1.8)	2.503(0.894)	0.838(0.035)
		MCP	5.5(1.7)	0.0(0.0)	4.5(1.4)	3.3(1.9)	2.985(0.922)	0.807(0.030)
		grMCP	7.3(1.3)	0.2(0.7)	5.8(0.7)	32.1(7.6)	2.877(0.761)	0.837(0.030)
		Marginal	0.8(1.1)	4.3(6.5)	2.9(1.2)	36.0(42.5)	-(-)	-(-)
	AR(0.7)	Proposed	7.3(0.6)	0.8(1.0)	5.0(0.8)	4.8(3.4)	2.748(0.607)	0.868(0.014)
		MCP	4.5(1.6)	0.5(0.7)	4.8(0.9)	6.2(2.6)	3.353(0.673)	0.842(0.019)
		grMCP	7.2(0.7)	0.8(0.7)	6.0(0.2)	33.5(5.6)	3.302(0.524)	0.865(0.014)
		Marginal	0.9(1.2)	4.9(6.5)	1.6(1.3)	37.3(41.8)	-(-)	-(-)
3	AR(0.3)	Proposed	11.2(0.8)	0.1(0.3)	0.0(0.0)	1.9(1.7)	1.553(0.822)	0.849(0.019)
		MCP	9.9(1.7)	0.1(0.2)	0.0(0.0)	3.9(1.7)	2.501(0.955)	0.816(0.027)
		grMCP	10.1(2.7)	0.1(0.2)	0.0(0.0)	50.8(13.9)	2.966(0.928)	0.828(0.038)
		Marginal	1.2(1.4)	2.7(4.2)	0.0(0.0)	19.6(33.9)	-(-)	-(-)
	AR(0.7)	Proposed	11.3(0.8)	0.8(0.7)	0.0(0.0)	2.9(2.5)	2.217(0.819)	0.865(0.018)
		MCP	9.1(1.5)	0.5(0.8)	0.0(0.0)	5.1(2.8)	3.047(0.980)	0.837(0.024)
		grMCP	10.0(2.7)	0.4(0.5)	0.0(0.0)	52.0(15.0)	3.430(1.089)	0.842(0.039)
		Marginal	1.5(1.9)	8.9(15.1)	0.0(0.0)	41.7(72.3)	-(-)	-(-)
4	AR(0.3)	Proposed	4.0(0.0)	0.0(0.0)	14.1(3.0)	0.0(0.0)	3.848(0.728)	0.859(0.020)
		MCP	3.4(0.6)	0.0(0.0)	10.2(2.9)	0.2(0.5)	4.851(0.720)	0.846(0.022)
		grMCP	4.0(0.0)	0.0(0.0)	20.0(0.0)	0.0(0.0)	2.343(0.498)	0.877(0.016)
		Marginal	0.2(0.4)	2.4(3.9)	1.7(1.7)	25.2(30.6)	-(-)	-(-)
	AR(0.7)	Proposed	4.0(0.0)	0.4(0.7)	15.4(2.5)	0.3(0.6)	3.734(1.152)	0.882(0.020)
		MCP	1.1(1.1)	0.1(0.3)	13.6(2.1)	0.6(1.1)	4.375(0.789)	0.860(0.017)
		grMCP	4.0(0.0)	0.0(0.0)	20.0(0.0)	0.0(0.0)	1.951(0.526)	0.893(0.013)
		Marginal	0.1(0.2)	1.7(3.0)	0.7(1.0)	19.5(28.0)	-(-)	-(-)

effects and interactions based on p-values. We note that this approach has a different analysis scheme. It is considered as a benchmark, given the still high popularity of marginal analysis. We acknowledge that there are other alternatives that are applicable. However, the above may be the most relevant. For the two penalization alternatives, tuning parameters are selected in the same manner as for the proposed approach.

We evaluate the accuracy of variable selection, estimation, and prediction. For variable selection, we consider the numbers of true positive (TP) and false positive (FP) for main effects (M) and interactions (I). For estimation, we consider the sum of squared error (SSE) $\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\|_2^2$. For prediction, we generate independent testing data under the same settings and use the C-index (Uno et al. 2011) for evaluation. Here it is noted that the marginal analysis approach cannot generate a single model. As such, its estimation and prediction performance is not evaluated.

Summary results based on 100 replicates are provided in Table 1 and Tables 3–4 in the supplementary material. Across the whole spectrum of simulation, the proposed approach is observed to have competitive performance. Specifically, for Examples 1–3, the proposed approach is able to identify the majority of important main effects and interactions, while having small numbers of false positives. In addition, it also has superior estimation and prediction performance. As a representative example, in Example 1 with the AR(0.3) correlation structure, the proposed approach has (M:TP, M:FP, I:TP, I:FP)=(12.2, 0.1, 6.0, 1.9), compared to (8.2, 0.1, 6.0, 2.9) for MCP, (7.0, 0, 8.1, 26.7) for grMCP, and (0.9, 2.9, 2.3, 20.7) for marginal. Additionally, it has (SSE, C-index)=(5.624, 0.810), compared to (7.402, 0.776) for MCP and (7.642, 0.759) for grMCP. Under Example 4, the proposed

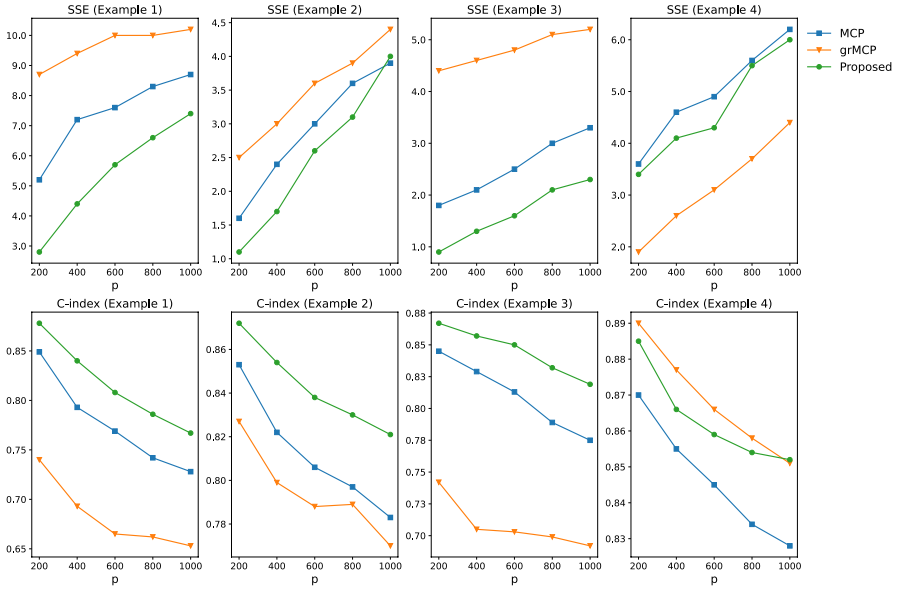

Fig. 1 SSE and C-index for different numbers of G variables under the AR(0.3) correlation structure

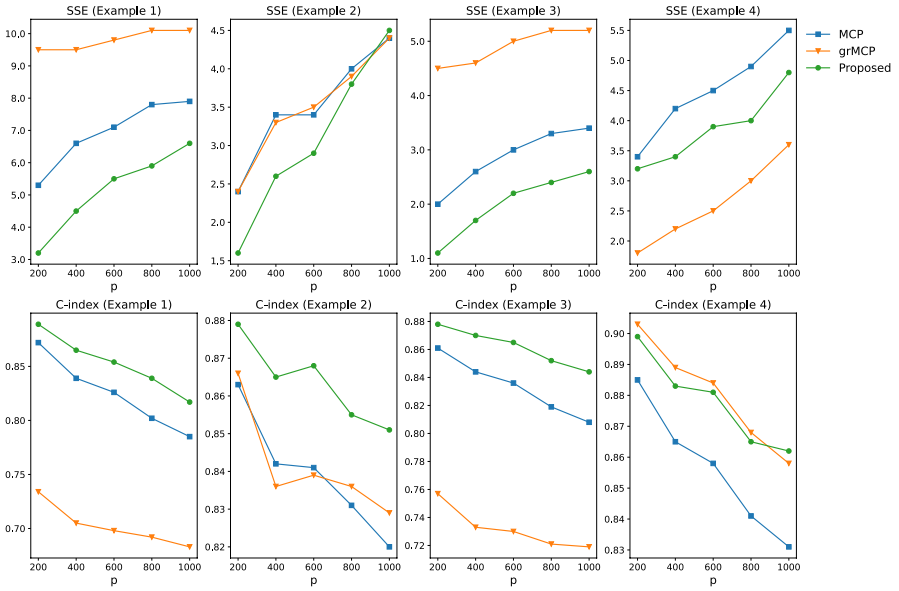

Fig. 2 SSE and C-index for different numbers of G variables under the AR(0.7) correlation structure

approach still outperforms MCP and marginal. Its performance is somewhat comparable to grMCP. This is sensible, as under this setting no within-group selection is needed, which favors grMCP. Here we note that in practical data analysis, full interactions are very rarely observed.

Additionally, we examine whether the proposed approach scales well with p. In particular, we consider $p = 200$, 400, 600, 800, and 1,000. The other settings remain unchanged. The estimation and prediction results are summarized in Figs. 1 and 2. As expected, as p increases, overall, performance of all three approaches deteriorates. However, the proposed approach clearly outperforms. We have also examined identification and made similar observations (detailed results omitted here).

4 Analysis of TCGA data on STAD

Stomach cancer is the second most common type of malignancy worldwide, and the 5-year survival rate is as low as 10%-19%. Profiling studies have been conducted on this cancer. G-E interactions have been suggested as critical for its risk, survival, and other outcomes. In this study, we analyze The Cancer Genome Atlas (TCGA) data on stomach adenocarcinoma (STAD). TCGA is a collective effort organized by the National Institutes of Health (NIH). Compared to many other data sources, TCGA is advantageous by being more recent and publicly available as well as having high data quality. TCGA data have been analyzed in many G-E interaction studies.

Table 2 Analysis of the STAD data using the proposed method: identified main effects and interactions

		PM	PN	PT	Gender	Age
Main.E		1.054	3.037	1.848	1.050	1.737
IFRD1	−0.169				−0.178	
CNTN1	0.145					
ECH1	0.322	0.284		0.414	0.321	
CHRNA1	−0.079			−0.045		
CNDP1	0.444					0.567
M1AP	−0.083					−0.204
FAM19A4	0.096	0.355				
GPRIN1	−0.172					−0.235
AC010280.2	0.425			0.284	0.273	
AC026704.1	−0.252					
AC022150.2	−0.275	−0.468				

In our analysis, the response variable of interest is overall survival, which is subject to right censoring. For E variables, we take a "looser" definition and consider age, AJCC metastasis pathologic stage (PM), AJCC nodes pathologic stage (PN), AJCC tumor pathologic stage (PT), and gender. In recent literature, interactions between demographic/clinical variables and G variables have drawn increasing interest, and many recent statistical studies have examined such interactions. Among them, age is continuous, and the rest are categorical. All of them have been suggested as relevant for stomach cancer survival. For G variables, we consider gene expressions. Level 3 processed data are downloaded from the TCGA website. Standard data processing is conducted, for which we refer to the published literature. The data available for downstream analysis include 19,037 gene expression measurements on 338 subjects. Among these subjects, 70 died during follow-up. The median observed time is 10.79 months. And the estimated median survival time is 29.3 months. The outcome information is graphically presented in Figure 4 of the supplementary material. In principle, the proposed approach can be directly applied. Considering the limited sample size, we conduct a supervised marginal screening and select the top 600 genes with the smallest marginal p-values for downstream analysis.

The proposed approach identifies 11 main G effects and 12 interactions. Detailed estimation results are presented in Table 2. The five main E effects all have positive signs, suggesting that higher PM/ PN/ PT stage and higher age are associated with a higher risk and shorter survival. In addition, males have a higher survival risk than females. All these results are consistent with the literature. By design, the identified main effects and interactions satisfy the variable selection hierarchy. A quick literature search suggests that the identified genes may have important implications. For example, upregulated CNTN1 is associated with lymph node metastasis and poor prognosis of cancer. Gene CNDP1 may constitute a marker of aggressive cancer and cancer cachexia. Low CNDP1 levels are associated with markers of poor prognosis including weight

loss, malnutrition, lipid breakdown, low circulating albumin/IGF1 levels, and poor quality of life. Gene FAM19A4 is one of immune-related genes, which are considered as important factors during cancer development. Gene GPRIN1 can bound miR-654-5p, which facilitates cancer cell proliferation, migration, and invasion.

Data are also analyzed using the three alternatives. Detailed estimation results under the alternative approaches are omitted here. Comparison in identification is summarized in Table 5 of the supplementary material, where we examine the numbers of overlapping identifications and RV-coefficients (Smilde et al. 2009). It is observed that different approaches lead to findings with small to moderate overlapping. To evaluate the proposed approach and also obtain a deeper understanding of the differences between methods, we evaluate prediction performance and stability. Specifically, data are randomly split into a training and a testing set with sizes 4:1. The training data are analyzed, and we obtain the identification results and an outcome model. With the proposed approach, MCP, and grMCP, we then make predictions for the testing data subjects. This procedure is repeated 100 times to avoid an extreme split. The mean C-index values are 0.723 (Proposed), 0.714 (MCP), and 0.708 (grMCP), suggesting a small advantage of the proposed approach in prediction. In addition, for the main effects and interactions identified using the full data, we also compute their probabilities of being identified in the 100 splittings. Such probabilities have been referred to as observed occurrence index (OOI). For all the identified effects, the mean OOI values are 0.664 (proposed), 0.486 (MCP), 0.629 (grMCP), and 0.578 (marginal). The proposed approach has better stability.

5 Discussion

In this article, we have filled an important knowledge gap by developing G-E interaction analysis for censored survival data under the Cox model. The key advancement of this study may not be its "conceptual" innovation—adopting the Cox model for survival analysis and penalization for regularized estimation and selection may seem somewhat "standard." Rather, the key contributions may come from the rigorous development of theoretical properties and an effective computational algorithm as well as careful numerical investigation via simulation and data analysis. In particular, with the theoretical guarantee of estimation and selection consistency, this approach can be used in practice with high confidence.

This study can potentially serve as the building block for others. For example, it may be of interest to examine composite and other penalizations that can also conduct selection and estimation under the variable selection hierarchy. It may also be of interest to accommodate additional structures (for example, pathway) in G variables. Finally, as there are multiple models/techniques available for G-E interaction analysis with a survival outcome, it may be of interest to systematically compare their performance in practical settings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-023-00871-9.

Acknowledgements The authors thank the Editor, Associate Editor, and two referees for their insightful comments which have led to a significant improvement of this article. This study is partly supported by National Bureau of Statistics of China (2022LZ34), National Natural Science Foundation of China (11971404, 72071169, 71988101, 82204153), National Social Science Foundation of China (21 &ZD146), and NIH (CA204120, CA121974, and CA196530).

References

Andersen, P. K., Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics, 10*(4), 1100–1120.

Bien, J., Taylor, J. E., Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics, 41*(3), 1111–1141.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends®in Machine Learning*, *3*(1), 1–122.

Bradic, J., Fan, J., Jiang, J. (2011). Regularization for cox's proportional hazards model with np-dimensionality. *Annals of Statistics, 39*(6), 3092–3120.

Chen, J., Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika, 95*(3), 759–771.

Choi, N. H., Li, W., Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association, 105*(489), 354–364.

Eriksson, F., Martinussen, T., Nielsen, S. (2019). Large sample results for frequentist multiple imputation for cox regression with missing covariate data. *Annals of the Institute of Statistical Mathematics, 72*, 969–996.

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

Feng, S., Zhang, M., Tong, T. (2021). Variable selection for functional linear models with strong heredity constraint. *Annals of the Institute of Statistical Mathematics, 74*, 321–339.

Fleming, T. R., Harrington, D. P. (2011). *Counting processes and survival analysis*. Hoboken, NJ, United States: Wiley.

Fujimori, K. (2022). The variable selection by the dantzig selector for cox's proportional hazards model. *Annals of the Institute of Statistical Mathematics, 74*(3), 515–537.

Huang, J., Ma, S., Xie, H., Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika, 96*(2), 339–355.

Huang, J., Sun, T., Ying, Z., Yu, Y., Zhang, C. (2013). Oracle inequalities for the lasso in the cox model. *Annals of Statistics, 41*(3), 1142–1165.

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nature Reviews Genetics, 6*(4), 287–298.

Liu, X., Zhong, P.-S., Cui, Y. (2020). Joint test of parametric and nonparametric effects in partial linear models for gene-environment interaction. *Statistica Sinica, 30*(1), 325–346.

Luo, S., Xu, J., Chen, Z. (2015). Extended bayesian information criterion in the cox model with a high-dimensional feature space. *Annals of the Institute of Statistical Mathematics, 67*(2), 287–311.

Ma, S., Huang, J. (2015). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association, 112*(517), 410–423.

McAllister, K. A., Mechanic, L. E., Amos, C. I., Aschard, H., Blair, I. A., Chatterjee, N., Conti, D. V., Gauderman, W. J., Hsu, L., Hutter, C., Jankowska, M. M., Kerr, J., Kraft, P., Montgomery, S. B., Mukherjee, B., Papanicolaou, G. J., Patel, C. J., Ritchie, M. D., Ritz, B. R., Witte, J. S. (2017). Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *American Journal of Epidemiology, 186*(7), 753–761.

Nocedal, J., Wright, S. (2006). *Numerical optimization*. Berlin/Heidelberg, Germany: Springer.

Simon, N., Friedman, J. H., Hastie, T., Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics, 22*, 231–245.

Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M., van Erk, M. J. (2009). Matrix correlations for high-dimensional data: The modified rv-coefficient. *Bioinformatics, 25*(3), 401–405.

Stute, W., Wang, J. (1993). The strong law under random censorship. *Annals of Statistics, 21*(3), 1591–1607.
Tang, X., Xue, F., Qu, A. (2021). Individualized multidirectional variable selection. *Journal of the American Statistical Association, 116*(535), 1280–1296.
Thomas, D. C. (2010). Gene-environment-wide association studies: Emerging approaches. *Nature Reviews Genetics, 11*(4), 259–272.
Uno, H., Cai, T., Pencina, M., D'Agostino, R., Wei, L. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine, 30*(10), 1105–1117.
Winham, S. J., Biernacka, J. M. (2013). Gene-environment interactions in genome-wide association studies: Current approaches and new directions. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 54*(10), 1120–1134.
Wu, C., Jiang, Y., Ren, J., Cui, Y., Ma, S. (2018). Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Statistics in Medicine, 37*(3), 437–456.
Wu, M., Zhang, Q., Ma, S. (2020). Structured gene-environment interaction analysis. *Biometrics, 76*(1), 23–35.
Xu, Y., Wu, M., Zhang, Q., Ma, S. (2019). Robust identification of gene-environment interactions for prognosis using a quantile partial correlation approach. *Genomics, 111*(5), 1115–1123.
Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics, 38*(3), 894–942.
Zhang, X., Liu, J., Zhu, Z. (2022) Learning coefficient heterogeneity over networks: A distributed spanning-tree-based fused-lasso regression. *Journal of the American Statistical Association*, *0*(0), 1–13.
Zhao, P., Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research, 7*, 2541–2563.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Kuangnan Fang[1] · Jingmao Li[1] · Yaqing Xu[2] · Shuangge Ma[3] · Qingzhao Zhang[1,4]

[1] Department of Statistics and Data Science, School of Economics, Xiamen University, No.422, Siming South Road, Xiamen 361005, Fujian, China

[2] School of Public Health, Shanghai Jiao Tong University School of Medicine, 227 South Chongqing Road, Shanghai 200240, China

[3] Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06520, USA

[4] The Wang Yanan Institute for Studies in Economics, Xiamen University, No.422, Siming South Road, Xiamen 361005, Fujian, China