SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
0(0) 1–14
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280218820356
journals.sagepub.com/home/smm
SAGE

Promoting sign consistency in the cure model estimation and selection

Xingjie Shi,[1] Shuangge Ma[2] and Yuan Huang[3]

Abstract

In survival analysis, when a subset of subjects has extremely long survival, the two-part cure rate model has been commonly adopted. In the two-part model, the first part is for a binary response and describes the probability of cure. The second part is for a survival response and describes the probability of survival. Despite their intuitive interconnections, most of the existing works estimate the two parts without any constraint. The existing works on proportionality promote similarity in magnitudes (i.e. quantitative similarity) and can be too restrictive. In this study, for the two-part cure rate model, we propose imposing a sign-based penalty to promote similarity in signs (i.e. qualitative similarity). The proposed strategy can be more informative than those that neglect the two-part interconnections and be less restrictive than the existing proportionality works. Penalty is also imposed to select relevant variables and accommodate high-dimensional data. Numerical studies, including simulation and two data analyses, demonstrate the advantageous performance of the proposed approach.

Keywords

Cure model, high-dimensionality, penalized estimation, sign consistency, two-part model

1 Introduction

In survival analysis, when a subset of subjects has extremely long survival, the cure rate model is commonly adopted.[1–3] Among the existing cure rate models, the two-part model, with an intuitive interpretation, has been popular.[4] In the two-part model, the first part is for a binary response, that is, whether a subject is "cured". For this part, the logistic regression model and other generalized linear models have been commonly adopted. The second part is for a (censored) survival response. For this part, the Cox and other survival models have been adopted. Extensive works have been conducted. For relevant discussions, we refer to literature.[5,6] In more recent studies, multi- and high-dimensional covariates are sometimes present. Under certain scenarios, it may be desirable to select relevant covariates and screen out the noisy ones. For this purpose, regularized especially penalized estimation has been conducted. For relevant discussions, we refer to literature.[7,8]

Although the existing works have been successful in multiple aspects, our literature review suggests that most of them have not paid sufficient attention to the interconnections between the covariate effects in the two model parts. Specifically, they usually estimate the two sets of covariate effects "freely" without imposing any constraint. Although the two model parts have different forms and are on different scales, they in fact describe two highly related processes: from infinite long survival (not susceptible) to finite survival (susceptible), and from finite longer survival to shorter survival. As such, it is reasonable to expect that the two sets of covariate effects are somewhat interconnected. In the literature, this has been considered in the proportionality works,[9,10] under which it is assumed that some covariate effects are proportional in the two model parts. Published studies[11,12] have argued convincingly the necessity of considering the interconnections in the covariate effects. Theoretical derivations[9] and numerical studies[13] suggest that, under quite general model settings, the proportionality constraint can improve estimation and variable selection accuracy.

[1]Department of Statistics, Nanjing University of Finance and Economics, Nanjing, Jiangsu, China
[2]Department of Biostatistics, Yale University, New Haven, CT, USA
[3]Department of Biostatistics, University of Iowa, Iowa City, IA, USA

Corresponding author:
Yuan Huang, Department of Biostatistics, University of Iowa, 145 N. Riverside Drive, CPHB N318, Iowa City, IA 52242, USA.
Email: yuan-huang@uiowa.edu

The proportionality works promote similarity in the magnitudes of the regression coefficients (up to a constant), that is, quantitative similarity. It needs to be recognized that, although the two model parts are related, they still describe different "regions" of survival. As such, the assumption of quantitative similarity may be too stringent. Taking into account the successes of the proportionality works as well as their limitations, in this study, we propose promoting similarity in the signs of the regression coefficients. That is, with the interconnections between the two model parts, we expect whether a covariate is positively (or negatively, or not) associated with survival is consistent to a certain extent. However, we do not further impose constraints on magnitudes. In the context of integrative analysis of multiple datasets,[14] which is dramatically different from the present settings, it has been shown that promoting qualitative similarity is needed beyond quantitative similarity under certain scenarios and demands new techniques. However, this issue has not been examined for the two-part cure models. We note that it is possible that a covariate has different effects and even opposite signs in the two model parts. For example, a treatment may have a positive short-term effect and a detrimental long-term effect. To accommodate such a scenario, in the proposed study, the sign consistency is encouraged but not required.

In this study, we consider the two-part cure rate model. Different from most of the existing works, the focus is on the structure of covariate effects. Complementary to the existing proportionality studies, our goal is to promote sign consistency, i.e. qualitative similarity in covariate effects, for the two model parts, while being flexible to accommodate possibly opposite signs. This is achieved using a novel sign-based penalization approach. In addition, both low-dimensional and high-dimensional cases are considered, advancing from many of the existing studies that are focused on one case only. A penalization strategy is adopted to accommodate high-dimensionality, select relevant variables, and conduct regularized estimation. Overall, this study can provide a practically useful alternative strategy for analyzing many practical data. It is also noted that the proposed method may be easily extended to other two-part models (for example, the logistic + linear models as in Fang et al.[11]) for heterogeneous data.

2 Methods

2.1 Data and model

For survival data where some subjects may have extremely long survival, we consider the two-part model which postulates that the population is a mixture of susceptible and "cured" subjects. Let U be the time to event and C be the time of right censoring. Denote $T = \min(U, C)$ and $\Delta = I(U \leq C)$. Note that for subjects that are not susceptible, $U = \infty$. The susceptibility status is indicated by a binary variable Y, where $Y = 1$ for those who are susceptible. Denote the length-p vector Z as the covariate of interest. One observation consists of (T, Δ, Z).

The first part of the model describes whether a subject is susceptible, for which we adopt the popular logistic regression model. That is, $Pr(Y = 1|Z) = \pi(Z) = \frac{\exp(\gamma_0 + Z'\gamma)}{1+\exp(\gamma_0 + Z'\gamma)}$, where γ_0 is the intercept and γ is the length-p vector of unknown regression coefficients. For susceptible subjects, we model their survival using the Cox model, where the conditional hazard function is $h(U|Z) = h_0(U)\exp(Z'\beta)$, $h_0(U)$ is the nonparametric baseline hazard function, and β is the length-p vector of regression coefficients. Further denote f and f_0 as the density functions and S and S_0 as the survival functions corresponding to h and h_0, respectively. The population survival function is $\pi(Z)S + 1 - \pi(Z)$.

Assume n observations. Let (t_i, δ_i, z_i) denote the ith realization of (T, Δ, Z), $i = 1, \ldots, n$. Then the full log-likelihood function is

$$\ell(\beta, \gamma, S_0) = \log \prod_i [\pi_i f(t_i, z_i)]^{\delta_i} [(1 - \pi_i) + \pi_i S(t_i, z_i)]^{1-\delta_i} \tag{1}$$

where $\pi_i = \pi(z_i)$ and $S(t_i, z_i) = S_0(t_i)^{\exp(z_i'\beta)}$

2.2 Penalized estimation

Consider the scenario with a moderate to large p, under which some covariates may not be relevant. To accommodate the potential high data dimensionality and select relevant variables, we adopt penalization. Further, to promote sign consistency, a new penalty is developed. Specifically, we propose the penalized objective function

$$-\frac{2}{n}\ell(\beta, \gamma, S_0) + \sum_{l=1}^{p} \rho(|\gamma_l|; \lambda_1, \nu) + \sum_{l=1}^{p} \rho(|\beta_l|; \lambda_2, \nu) + \frac{\lambda_3}{2}\sum_{l=1}^{p} [\text{sign}(\beta_l) - \text{sign}(\gamma_l)]^2 \tag{2}$$

where $\rho(t; \lambda_s, \nu) = \lambda_s \int_0^t (1 - \frac{x}{\nu\lambda_s})_+ \mathrm{d}x$ $(s = 1, 2)$ is the minimax concave penalty (MCP),[15] and β_l and γ_l $(l = 1, 2, \ldots, p)$ are the l-th component of β and γ, respectively. $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ are data-dependent tuning parameters, and ν is the regularization parameter.

[Remark 1] Here we consider the same set of covariates Z for both model parts for convenience. For cases where either prior knowledge or data collection lead to different sets of covariates, we can modify the sign based penalty term as $\sum_{l \in L} [\mathrm{sign}(\beta_l) - \mathrm{sign}(\gamma_l)]^2$, where L denotes the index set of covariates that are shared by the two model parts.

Overall, our proposal fits well in the penalized estimation and selection paradigm. In equation (2), the first two penalties conduct estimation and selection for γ and β, respectively. The MCP is adopted for its satisfactory properties demonstrated in the literature and can be replaced by other penalties. To reduce computational cost, it is possible to set $\lambda_1 = \lambda_2$. The newly proposed penalty $\frac{\lambda_3}{2} \sum_{l=1}^{p} [\mathrm{sign}(\beta_l) - \mathrm{sign}(\gamma_l)]^2$ directly promotes sign consistency between β and γ and has an intuitive definition. Different from those in Fan et al.[12] and some other existing works, it is built on the sign function, not magnitude, and hence promotes qualitative similarity. The tuning parameter λ_3 adjusts the degree of penalization data-dependently. When there is little support from data for sign consistency, it may take a small value, and the proposed method simplifies to the "standard" penalized estimation and selection. It is noted that if there is strong prior knowledge that some covariates have opposite signs, then they can be excluded from the sign penalty.

As a penalty involving the sign function is computationally difficult to optimize, we propose the following approximation

$$\frac{\lambda_3}{2} \sum_{l=1}^{p} [\mathrm{sign}(\beta_l) - \mathrm{sign}(\gamma_l)]^2 \approx \frac{\lambda_3}{2} \sum_{l=1}^{p} \left(\frac{\beta_l}{|\beta_l| + \xi} - \frac{\gamma_l}{|\gamma_l| + \xi} \right)^2 \tag{3}$$

where $\xi > 0$ controls the degree of approximation (more discussions below).

2.3 Computation

Computation is challenging for several reasons. First, with the cure rate model, calculation of nonparametric S_0 cannot be eliminated. Second, high-dimensionality increases computational cost. For example, the existing Lasso-penalized algorithm implemented by Liu et al.[7] cannot cope with high-dimensional data. Lastly, the sign-based penalty adds complexity to the objective function. To tackle the challenging computation problems, we propose a new efficient algorithm, the Expectation/Coordinate Descent (ECD) algorithm, which can be considered as an Expectation/Conditional Maximization (ECM) algorithm.[16] In the E step, the susceptible indicators are introduced to obtain a full likelihood function that allows the optimization respect to γ, β, and S_0 to be conducted separately, using the profile likelihood technique. In the coordinate descent (CD) steps, minimizing the penalized objective function can be performed by updating a single parameter with the remaining parameters fixed at their most recent values. Instead of iteratively updating β and γ until convergence for each CD step, we use one-step update that might lead to more iterations of the ECD algorithm, but reduce the overall computational time,[17] which makes it easier to handle high-dimensional data. The details are as follows. We have developed the code in R and made it publicly available at www.github.com/shuanggema.

2.3.1 E step

Expected complete data log-likelihood: Consider the complete data $\{(t_i, \delta_i, z_i, y_i),\ i = 1, \ldots, n\}$, which include the observed data as well as the unobserved y_i's. The complete data log-likelihood is

$$\ell_c(\beta, \gamma, S_0; y) = \sum_i [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)] + \sum_i [\delta_i \log h(t_i, z_i) + y_i \log S(t_i, z_i)] \tag{4}$$

Note that we include y in the likelihood notation to emphasize the dependence on the unobserved indicators.

The E step of the ECD algorithm calculates the expectation of equation (4) with respect to the distribution of the unobserved y_i's, given the observed data O and current estimate $\tilde{\theta}$ of $\theta = (\beta, \gamma, S_0)$. Denote $w_i = \mathrm{E}(y_i | O, \theta)$. Then the expected log-likelihood is the sum of the following functions

$$\begin{aligned} \ell_1(\gamma; w) &= \sum_i [w_i \log \pi_i + (1 - w_i) \log(1 - \pi_i)], \\ \ell_2(\beta, S_0; w) &= \sum_i [\delta_i \log h(t_i, z_i) + w_i \log S(t_i, z_i)] \end{aligned} \tag{5}$$

Note that w_i can be viewed as the posterior probability of the ith subject being susceptible and updated by

$$\tilde{w}_i = \delta_i + (1-\delta_i)\frac{\tilde{\pi}_i \tilde{S}(t_i, z_i)}{(1-\tilde{\pi}_i) + \tilde{\pi}_i \tilde{S}(t_i, z_i)} \tag{6}$$

where $\tilde{\pi}_i = \frac{\exp(\tilde{\gamma}_0 + z_i'\tilde{\gamma})}{1+\exp(\tilde{\gamma}_0 + z_i'\tilde{\gamma})}$ and $\tilde{S}(t_i, z_i) = \tilde{S}_0(t_i)^{\exp(z_i'\tilde{\beta})}$. Estimation of $S(t, z)$ in turn involves estimating the baseline hazard function h_0. For h_0, we propose using a profile likelihood construction similar to that used in the standard Cox model, motivated by Peng and Dear.[18]

Let $\tau_1 < \cdots < \tau_k$ denote the distinct uncensored event times. Denote D_j as the set of d_j tied uncensored events at τ_j. Let E_j be the set of subjects with censoring times in $[\tau_j, \tau_{j+1}), j = 0, \ldots, k$, where $\tau_0 = 0$ and $\tau_{k+1} = \infty$. Denote R_j as the at risk set at time τ_j. Following the nonparametric profile likelihood method of Breslow,[19] maximizing ℓ_2 with given parameter estimate $\tilde{\beta}$ and weights $\tilde{w}_i$'s leads to a discrete baseline hazard with $\hat{h}_0(t) = 0$ for all $t \notin \{\tau_1, \ldots, \tau_k\}$. Thus, $\left\{\hat{h}_0(\tau_1), \ldots, \hat{h}_0(\tau_k)\right\}$ that maximizes ℓ_2 also maximizes

$$\sum_{j=1}^{k} d_j \log h_0(\tau_j) - \sum_{j=1}^{k}\left[h_0(\tau_j)\sum_{i \in R_j}\tilde{w}_i \exp(z_i'\tilde{\beta})\right] \tag{7}$$

Differentiating equation (7) with respect to $h_0(\tau_j)$ gives the maximum likelihood estimate as

$$\hat{h}_0(\tau_j) = \frac{d_j}{\sum\limits_{i \in R_j}\tilde{w}_i \exp(z_i'\tilde{\beta})} \tag{8}$$

The estimated baseline survival function $S_0(t)$ is then

$$\hat{S}_0(t) = \exp\left\{-\sum_{j:\tau_j \leq t}\hat{h}_0(\tau_j)\right\} \tag{9}$$

To ensure that the susceptible subjects have zero survival at time infinity, we impose $\hat{S}_0(t) = 0$ for $t > \tau_k$.

2.3.2 CD step

Here with fixed estimates for h_0 and w, we optimize with respect to γ and β. Substitute the current estimates $\tilde{S}_0$ and $\tilde{w}$ into $\ell_2(\beta, S_0; w)$, and consider

$$\ell_2(\beta, \tilde{S}_0; \tilde{w}) = \sum_{j=1}^{k}\left[-d_j \log \sum_{i \in R_j}\tilde{w}_i \exp(z_i'\beta) + x_j'\beta\right] \tag{10}$$

where $x_j = \sum_{i \in D_j} z_i$.

With a slight abuse of notations, we abbreviate $\ell_1(\gamma; \tilde{w})$ and $\ell_2(\beta, \tilde{S}_0; \tilde{w})$ as $\ell_1(\gamma)$ and $\ell_2(\beta)$, respectively. The M-step in a standard EM algorithm consists of minimizing

$$-\frac{2}{n}[\ell_1(\gamma) + \ell_2(\beta)] + \sum_{l=1}^{p}[\rho(|\gamma_l|; \lambda_1, \nu) + \rho(|\beta_l|; \lambda_2, \nu)] + \frac{\lambda_3}{2}\sum_{l=1}^{p}[\mathrm{sign}(\beta_l) - \mathrm{sign}(\gamma_l)]^2 \tag{11}$$

(or the approximated objective function) with respect to γ and β. This can be achieved using the CD approach. Below we provide details for the CD step for γ. The procedure for β is similar.

Optimization with respect to γ: We seek to minimize equation (11) with respect to γ_l while fixing β and $\gamma_{l'}$ $(l' \neq l)$ at their current estimates. Here we adopt a local quadratic approximation approach. Let $\nabla_l \ell$ and $\nabla_l^2 \ell$ denote the first and second derivatives of ℓ with respect to the lth variable, respectively. With the quadratic approximation, we have the following objective function for γ_l

$$\begin{aligned} & -\frac{1}{n}\nabla_l^2\ell_1(\tilde{\gamma})(\gamma_l-\tilde{\gamma}_l)^2-\frac{2}{n}\nabla_l\ell_1(\tilde{\gamma})(\gamma_l-\tilde{\gamma}_l)+\rho(|\gamma_l|;\lambda_1,\nu)+\frac{\lambda_3}{2}\left(\frac{\tilde{\beta}_l}{|\tilde{\beta}_l|+\xi}-\frac{\gamma_l}{|\tilde{\gamma}_l|+\xi}\right)^2 \\ & =\frac{1}{2}a_{1l}\gamma_l^2-b_{1l}\gamma_l+c_{1l}|\gamma_l|+\text{const} \end{aligned} \tag{12}$$

where

$$\begin{aligned} a_{1l} &= -\frac{2}{n}\nabla_l^2\ell_1(\tilde{\gamma})-\frac{I(|\tilde{\gamma}_l|<\nu\lambda_1)}{\nu}+\lambda_3\frac{1}{(|\tilde{\gamma}_l|+\xi)^2}, \\ b_{1l} &= \frac{2}{n}\left[-\nabla_l^2\ell_1(\tilde{\gamma})\tilde{\gamma}_l+\nabla_l\ell_1(\tilde{\gamma})\right]+\lambda_3\frac{\tilde{\beta}_l}{(|\tilde{\beta}_l|+\xi)(|\tilde{\gamma}_l|+\xi)}, \\ c_{1l} &= \lambda_1 I(|\tilde{\gamma}_l|<\nu\lambda_1) \end{aligned} \tag{13}$$

with

$$\begin{aligned} \nabla_l\ell_1(\gamma) &= \sum_{i=1}^{n}(\tilde{w}_i-\pi_i)z_{il}, \\ \nabla_l^2\ell_1(\gamma) &= -\sum_{i=1}^{n}\pi_i(1-\pi_i)z_{il}^2 \end{aligned} \tag{14}$$

The minimizer of (12) is

$$\hat{\gamma}_l=\frac{\text{sign}(b_{1l})}{a_{1l}}(|b_{1l}|-c_{1l})_+ \tag{15}$$

For the intercept, we have

$$\hat{\gamma}_0=\tilde{\gamma}_0-\frac{\nabla_0\ell_1(\tilde{\gamma})}{\nabla_0^2\ell_1(\tilde{\gamma})} \tag{16}$$

Optimization with respect to $\boldsymbol{\beta}$: To update β_l in a similar way as equation (15), we need

$$\begin{aligned} a_{2l} &= -\frac{2}{n}\nabla_l^2\ell_2(\tilde{\beta})-\frac{I(|\tilde{\beta}_l|<\nu\lambda_2)}{\nu}+\lambda_3\frac{1}{(|\tilde{\beta}_l|+\xi)^2}, \\ b_{2l} &= \frac{2}{n}\left[-\nabla_l^2\ell_2(\tilde{\beta})\tilde{\beta}_l+\nabla_l\ell_2(\tilde{\beta})\right]+\lambda_3\frac{\tilde{\gamma}_l}{(|\tilde{\beta}_l|+\xi)(|\tilde{\gamma}_l|+\xi)}, \\ c_{2l} &= \lambda_2 I(|\tilde{\beta}_l|<\nu\lambda_2) \end{aligned} \tag{17}$$

and the derivatives of ℓ_2

$$\begin{aligned} \nabla_l\ell_2(\beta) &= \sum_{j=1}^{k}\left[x_{jl}-d_j\frac{\sum_{i\in R_j}\tilde{w}_i\exp(z_i'\beta)z_{il}}{\sum_{i\in R_j}\tilde{w}_i\exp(z_i'\beta)}\right], \\ \nabla_l^2\ell_2(\beta) &= \sum_{j=1}^{k}d_j\left[\frac{\left(\sum_{i\in R_j}\tilde{w}_i\exp(z_i'\beta)z_{il}\right)^2}{\left(\sum_{i\in R_j}\tilde{w}_i\exp(z_i'\beta)\right)^2}-\frac{\sum_{i\in R_j}\tilde{w}_i\exp(z_i'\beta)z_{il}^2}{\sum_{i\in R_j}\tilde{w}_i\exp(z_i'\beta)}\right] \end{aligned} \tag{18}$$

2.3.3 ECD algorithm

With fixed tunings, the ECD algorithm is summarized in Algorithm 1.

Algorithm 1: the ECD Algorithm

```
Initialize $m = 0$, $\hat{\theta}^m = (\hat{\gamma}^m, \hat{\beta}^m, \hat{S}_0^m)$, and $w^m$.
repeat
    • E-Step
        – Compute $\hat{S}_0^{m+1}$ from (9),
        – Update $w^{m+1}$ from (6).
    • CD-Step
        – For $l = 0, \ldots, p$, update $\gamma_l^{m+1}$ according to (16) and (15) with (13);
        – For $l = 1, \ldots, p$, update $\beta_l^{m+1}$ similarly to (15) with (17).
    $m = m + 1$;
until the change of $\theta$ is smaller than a threshold;
```

Following the same arguments as for the ECM algorithm in Meng and Rubin,[16] it is easy to see that this algorithm converges to a stationary point which is also a local optimizer.

2.3.4 Tuning parameter selection

To reduce computational cost, the value of v in MCP can be fixed, as suggested in the literature. In our numerical study, we fix $v = 6$. For approximating the sign function, a smaller value of ξ may lead to a better approximation but at the same time less stable estimation. In our numerical study, we fix $\xi = 0.1$, which leads to satisfactory results. λ is selected using V-fold cross validation. As only simple updates are involved and no iteration is needed in the CD step, the overall computational cost is affordable. For example, with fixed tunings, one replicate of Scenario (1) in Example 1 takes about 11 s on a standard PC. In practice, with the computation run over a grid of tuning parameters, the overall computational cost can increase accordingly.

3 Simulation

We set $n = 400$ and consider both the low-dimensional case with $p = 30$ and high-dimensional case with $p = 300$. The covariates are generated from a multivariate normal distribution with marginal means zero and variances one. Covariates l and l' have correlation $\Sigma_{ll'} = \rho^{|l-l'|}$ with $\rho = 0.5$. The cure indicator is generated from a logistic model. For susceptible subjects, the event times are generated from the Cox model $h(U|Z) = 2U\exp(Z'\beta)$. The censoring times are generated independently from exponential distributions. In all simulations, the overall censoring rate is about 60%, with about 37% cured. Therefore, the effective sample size (number of events) is about 160. More specifications are discussed below. We analyze data with the proposed approach as well as two alternatives. (Alt.1) The sign-based penalty is replaced by $\lambda_3 \sum_{l=1}^{p} (\gamma_l - \zeta\beta_l)^2$. This is a magnitude-based penalty and has been considered in Fan et al.[12] under a different two-part model. Here ζ also needs to be estimated data-dependently. (Alt.2) This approach does not consider any interconnection between the two model parts (i.e. $\lambda_3 = 0$ in the proposed approach). It serves as benchmark. Note that this approach is a version of Liu et al.[7] but uses MCP instead of Lasso to achieve sparsity. To compare different approaches, we consider identification accuracy measured using the true positive rate (TPR) and false positive rate (FPR). In addition, we also consider prediction performance evaluated using the relative model error (RME)[20] and estimation performance evaluated using the estimation error (ERR). For β, the four measures are defined as below

$$\mathrm{TPR} = \frac{\sum_{l=1}^{p} I(\beta_l^0 \neq 0 \cap \hat{\beta}_l \neq 0)}{\sum_{l}^{p} I(\beta_l^0 \neq 0)},$$

$$\mathrm{FPR} = \frac{\sum_{l=1}^{p} I(\beta_l^0 = 0 \cap \hat{\beta}_l \neq 0)}{\sum_{l}^{p} I(\beta_l^0 = 0)},$$

$$\mathrm{RME} = \frac{(\hat{\beta} - \beta^0)^T \Sigma (\hat{\beta} - \beta^0)}{(\hat{\beta}^* - \beta^0)^T \Sigma (\hat{\beta}^* - \beta^0)}, \qquad (19)$$
$$\mathrm{ERR} = \frac{(\hat{\beta} - \beta^0)^T (\hat{\beta} - \beta^0)}{(\hat{\beta}^* - \beta^0)^T (\hat{\beta}^* - \beta^0)}$$

where β^0, $\hat{\beta}$, and $\hat{\beta}^*$ are the true, estimated, and estimated parameters as if the nonzero ones are known. The measures can be defined similarly for γ. All the summary statistics are calculated based on 1000 replications.

Example 1. We consider cases where γ and β have the same signs and same/similar magnitudes. Let the first six components of γ and β be nonzero. Consider the following six scenarios: (1) the nonzero components of γ are generated from $Unif(0.2, 0.6)$, and $\beta = \gamma$. (2) The first three nonzero components of γ are generated from $Unif(0.2, 0.4)$, and the next three are from $Unif(0.6, 0.8)$. $\beta = \gamma$. (3) The nonzero components of γ are generated from $Unif(0.6, 0.8)$, and $\beta = \gamma$. Scenarios (1) and (2) contain some weak signals while all signals in Scenario (3) are strong. Scenarios (4) to (6) are similar to Scenarios (1) to (3), with $\beta = \gamma + N(0, 0.01)$.

Table 1 shows the summary statistics for the high-dimensional cases. Although Example 1 can favor both the proposed method and Alt.1 by design, the proposed method outperforms Alt.1, and both of them outperform Alt.2. Consider for example Scenario 5 in high-dimensional cases (Table 1). For γ, the TPR values are 0.73 (proposed), 0.58 (Alt.1), and 0.21 (Alt.2), respectively. All three approaches have close to zero FPRs. The proposed approach also excels in terms of prediction and estimation. The RME values are 0.96 (proposed), 1.32 (Alt.1), and 4.40 (Alt.2), respectively. The ERR values are 0.30 (proposed), 0.34 (Alt.1), and 1.47 (Alt.2),

Table 1. Simulation Example 1, high-dimensional data.

		γ			β		
Scenario		Proposed	Alt.1	Alt.2	Proposed	Alt.1	Alt.2
1	RME	1.24 (0.89)	2.26 (1.80)	3.91 (2.62)	3.11 (2.17)	4.23 (3.15)	5.05 (3.31)
	ERR	0.34 (0.23)	0.60 (0.30)	0.96 (0.28)	0.27 (0.17)	0.32 (0.16)	0.43 (0.21)
	TPR	0.64 (0.27)	0.41 (0.28)	0.14 (0.15)	0.73 (0.20)	0.73 (0.21)	0.62 (0.23)
	FPR	0.01 (0.01)	0.01 (0.01)	0.00 (0.00)	0.02 (0.02)	0.03 (0.02)	0.02 (0.01)
2	RME	0.94 (0.51)	1.36 (1.09)	4.99 (3.12)	2.19 (1.31)	2.70 (1.82)	3.51 (2.17)
	ERR	0.28 (0.15)	0.35 (0.17)	1.49 (0.34)	0.21 (0.11)	0.23 (0.09)	0.29 (0.12)
	TPR	0.72 (0.22)	0.55 (0.21)	0.18 (0.15)	0.78 (0.14)	0.78 (0.14)	0.73 (0.16)
	FPR	0.01 (0.01)	0.00 (0.01)	0.00 (0.00)	0.01 (0.02)	0.02 (0.02)	0.02 (0.01)
3	RME	0.74 (0.19)	0.62 (0.52)	6.76 (5.45)	0.98 (0.26)	1.31 (0.75)	2.13 (1.12)
	ERR	0.20 (0.15)	0.18 (0.17)	2.51 (0.51)	0.07 (0.05)	0.09 (0.06)	0.15 (0.10)
	TPR	0.99 (0.09)	0.95 (0.17)	0.24 (0.18)	1.00 (0.00)	1.00 (0.02)	0.99 (0.04)
	FPR	0.00 (0.01)	0.00 (0.01)	0.00 (0.00)	0.01 (0.01)	0.02 (0.02)	0.02 (0.01)
4	RME	1.28 (1.02)	2.34 (1.88)	4.09 (2.71)	3.04 (2.25)	4.01 (2.77)	5.17 (3.23)
	ERR	0.35 (0.28)	0.57 (0.31)	1.02 (0.36)	0.27 (0.18)	0.32 (0.17)	0.43 (0.20)
	TPR	0.63 (0.29)	0.44 (0.27)	0.15 (0.14)	0.74 (0.21)	0.73 (0.22)	0.63 (0.23)
	FPR	0.01 (0.01)	0.01 (0.01)	0.00 (0.00)	0.02 (0.02)	0.03 (0.02)	0.02 (0.02)
5	RME	0.96 (0.66)	1.32 (1.10)	4.40 (3.09)	2.42 (1.64)	2.90 (1.93)	3.43 (2.28)
	ERR	0.30 (0.19)	0.34 (0.20)	1.47 (0.46)	0.20 (0.10)	0.23 (0.10)	0.29 (0.12)
	TPR	0.73 (0.20)	0.58 (0.20)	0.21 (0.17)	0.78 (0.14)	0.79 (0.15)	0.73 (0.15)
	FPR	0.01 (0.01)	0.00 (0.01)	0.00 (0.00)	0.01 (0.02)	0.03 (0.02)	0.02 (0.01)
6	RME	0.73 (0.20)	0.65 (0.51)	6.72 (5.18)	1.00 (0.30)	1.28 (0.72)	2.22 (1.21)
	ERR	0.23 (0.17)	0.20 (0.17)	2.55 (0.59)	0.07 (0.05)	0.09 (0.05)	0.15 (0.10)
	TPR	0.98 (0.12)	0.95 (0.15)	0.23 (0.18)	1.00 (0.01)	1.00 (0.01)	0.99 (0.05)
	FPR	0.00 (0.01)	0.00 (0.01)	0.00 (0.00)	0.01 (0.01)	0.02 (0.02)	0.02 (0.01)

Note: In each cell, mean (sd).

respectively. Similar satisfactory performance is observed for β. Performance of Scenarios (4) to (6) is similar to that of Scenarios (1) to (3).

Signal level plays an important role in the design. As shown in Table 1, the proposed method clearly outperforms Alt.1 when there are weak signals (Scenarios (1) and (2)). For scenarios where signals are all strong, the proposed method still shows an advantage in selection, but Alt.1 may have better estimation. An inspection of the estimated coefficients reveals that Alt.1 has difficulty in identifying weak signals correctly. That is, Alt.1 may shrink both coefficients to zero when they are small, resulting in a smaller TPR.

Summary statistics for the low-dimensional cases are provided in Table A1, Supplementary material. Comparisons of the low-dimensional cases are similar to those of the high-dimensional cases. Signal level again shows a great impact over the comparison. As a result, the advantage of Alt.1 in terms of estimation shows up when signals are large by design (Scenarios (3) and (6)).

Example 2. We consider cases where γ and β have the same signs but magnitudes can get less proportional. Denote $\alpha = (0.2, 0.4, 0.6)$. The following six scenarios are considered. (1) $\gamma = \beta = (\alpha, \alpha)'$. (2) $\gamma = (2\alpha, \alpha)'$ and $\beta = (\alpha, 2\alpha)'$. (3) $\gamma = (3\alpha, \alpha)'$ and $\beta = (\alpha, 3\alpha)'$. (4) $\gamma = (4\alpha, \alpha)'$ and $\beta = (\alpha, 4\alpha)'$. (5) $\gamma = (5\alpha, \alpha)'$ and $\beta = (\alpha, 5\alpha)'$. (6) Components 1–3 of γ and 4–6 of β are generated from $Unif(2, 2.4)$. Components 4–6 of γ and 1–3 of β are generated from $Unif(0.2, 0.6)$.

Table 2 shows the summary statistics for the high-dimensional cases. Unlike Example 1 which favors both the proposed method and Alt.1, Example 2 may only favor the proposed method as the proportionality assumption is further violated. As shown in Table 2, the proposed method has superior performance over the two alternatives, and the advantage can be more evident as signal levels are large as in Scenarios (4) and (5). For γ, the TPR values for the proposed method, Alt.1, and Alt.2 for Scenario (1) are 0.63, 0.45, and 0.16, respectively, and are 0.94, 0.50, and 0.42 for Scenario (5). As signal level increases, performance of the proposed method improves, which might not hold for Alt.1. The increase of signal at the same time means more deviation from the proportionality

Table 2. Simulation Example 2, high-dimensional data.

		γ			β		
Scenario		Proposed	Alt.1	Alt.2	Proposed	Alt.1	Alt.2
1	RME	1.04 (0.60)	1.77 (1.37)	4.20 (2.89)	2.49 (1.75)	3.20 (2.27)	4.18 (2.75)
	ERR	0.27 (0.19)	0.43 (0.28)	1.11 (0.18)	0.20 (0.12)	0.24 (0.13)	0.32 (0.16)
	TPR	0.63 (0.20)	0.45 (0.22)	0.16 (0.15)	0.70 (0.14)	0.70 (0.16)	0.63 (0.18)
	FPR	0.01 (0.01)	0.00 (0.01)	0.00 (0.00)	0.01 (0.02)	0.03 (0.02)	0.02 (0.01)
2	RME	1.12 (0.76)	2.02 (1.54)	4.32 (3.17)	1.93 (1.30)	2.90 (2.03)	3.62 (2.65)
	ERR	0.35 (0.22)	0.57 (0.29)	1.41 (0.52)	0.17 (0.13)	0.24 (0.14)	0.33 (0.19)
	TPR	0.82 (0.19)	0.58 (0.20)	0.30 (0.17)	0.87 (0.13)	0.82 (0.16)	0.74 (0.18)
	FPR	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.01 (0.02)	0.02 (0.02)	0.02 (0.01)
3	RME	1.12 (0.75)	2.56 (1.98)	3.46 (2.77)	1.48 (0.89)	2.83 (2.01)	3.23 (2.34)
	ERR	0.56 (0.34)	1.08 (0.36)	1.30 (0.58)	0.13 (0.08)	0.22 (0.13)	0.30 (0.21)
	TPR	0.90 (0.15)	0.55 (0.21)	0.36 (0.14)	0.92 (0.10)	0.84 (0.14)	0.76 (0.17)
	FPR	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.02 (0.02)	0.01 (0.01)
4	RME	1.13 (0.88)	2.39 (2.20)	2.57 (2.10)	1.26 (0.75)	2.75 (2.18)	3.15 (2.36)
	ERR	0.79 (0.57)	1.47 (0.69)	1.48 (0.67)	0.12 (0.08)	0.28 (0.20)	0.36 (0.28)
	TPR	0.93 (0.11)	0.52 (0.20)	0.40 (0.14)	0.93 (0.10)	0.80 (0.16)	0.73 (0.16)
	FPR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
5	RME	1.20 (1.03)	2.71 (2.43)	2.34 (2.04)	1.21 (0.78)	3.12 (2.85)	3.12 (2.49)
	ERR	0.97 (0.76)	1.81 (1.04)	1.82 (1.08)	0.14 (0.10)	0.35 (0.28)	0.44 (0.30)
	TPR	0.94 (0.12)	0.50 (0.16)	0.42 (0.12)	0.94 (0.10)	0.74 (0.16)	0.69 (0.17)
	FPR	0.01 (0.03)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
6	RME	0.95 (0.88)	1.54 (1.40)	1.36 (0.99)	1.02 (0.65)	2.91 (2.74)	3.33 (2.81)
	ERR	0.78 (0.47)	1.20 (0.66)	1.07 (0.62)	0.13 (0.10)	0.37 (0.21)	0.43 (0.23)
	TPR	0.98 (0.10)	0.54 (0.11)	0.50 (0.06)	0.97 (0.09)	0.74 (0.17)	0.69 (0.17)
	FPR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)

Note: In each cell, mean (sd).

assumption, which negatively affects Alt.1. As an example, for γ, the TRP values for Alt.1 are 0.58, 0.55, and 0.50 for Scenarios (2), (3), and (5), respectively. This indicates the drawback of Alt.2.

Similar performance is observed in the low-dimensional cases as shown in Table A2, Supplementary material. For the proposed method, its performance constantly improves due to increased signal levels from Scenarios (1) to (5). For Alt.1, performance improves from Scenarios (1) to (2) due to increased signal levels but gets worse for Scenarios (3) to (5) due to the violation of proportionality.

Example 3. In the previous two examples, the components of γ and β have the same signs and, therefore, both favor the proposed method. Example 3 aims to study robustness of the proposed method. In this example, we consider settings where the components of γ and β can have different signs. Under Scenarios (1) to (6), the first six components of γ are generated from $Unif(0.4, 0.8)$. Scenarios (1) to (3) consider cases where β has nonzero elements different from γ. That is, we have covariates having a positive sign in γ but being zero in β. Here for β, components 3–8 (Scenario (1)), 5–10 (Scenario (2)), and 7–12 (Scenario (3)) are generated from $Unif(0.2, 0.6)$. Scenarios (4) to (6) consider the more extreme cases where β and γ have the same nonzero components but various conflicting signs. Here the first six components of β are also generated from $Unif(0.2, 0.6)$, but 2 (Scenario (4)), 4 (Scenario (5)), and 6 (Scenario (6)) of them have signs conflicting with γ.

Table 3 shows the summary statistics for the high-dimensional cases. We note that the design of coefficients' magnitudes for Example 3 is similar to that in Example 1, but signs/nonzero positions change, so that sign consistency or proportionality no longer holds. For the proposed method and Alt.1, all measures (TPR, FRP, RME, and ERR) are negatively affected as expected. However, the proposed method and also Alt.1 still have an advantage over Alt.2. On one hand, despite the violation of sign consistency/proportionality, there is still similarity shared between the two model parts due to the large number of zero coefficients. By exploiting such information, both the proposed method and Alt.1 result in better selection and estimation than Alt.2. On the other hand, the tuning parameter that controls the sign consistency or proportionality penalty is data-driven and can be adaptive

Table 3. Simulation Example 3, high-dimensional data.

		γ			β		
Scenario		Proposed	Alt.1	Alt.2	Proposed	Alt.1	Alt.2
	RME	2.73 (2.05)	2.96 (2.44)	4.09 (2.77)	3.84 (2.82)	4.05 (3.05)	5.41 (3.89)
1	ERR	1.28 (0.51)	1.33 (0.58)	1.96 (0.48)	0.39 (0.21)	0.40 (0.19)	0.57 (0.27)
	TPR	0.53 (0.24)	0.47 (0.27)	0.24 (0.17)	0.67 (0.23)	0.69 (0.25)	0.54 (0.25)
	FPR	0.01 (0.01)	0.00 (0.01)	0.00 (0.00)	0.02 (0.02)	0.02 (0.02)	0.01 (0.01)
	RME	3.86 (2.99)	3.68 (3.04)	4.16 (2.99)	4.93 (3.35)	4.79 (3.35)	5.77 (3.79)
2	ERR	1.68 (0.49)	1.52 (0.48)	1.92 (0.48)	0.51 (0.25)	0.46 (0.28)	0.62 (0.25)
	TPR	0.38 (0.18)	0.38 (0.22)	0.25 (0.16)	0.59 (0.23)	0.61 (0.25)	0.49 (0.23)
	FPR	0.01 (0.01)	0.00 (0.01)	0.00 (0.00)	0.02 (0.02)	0.02 (0.02)	0.01 (0.01)
	RME	4.11 (3.08)	4.05 (3.08)	4.05 (3.00)	6.14 (4.56)	5.55 (4.31)	6.51 (4.77)
3	ERR	1.76 (0.44)	1.60 (0.48)	1.88 (0.42)	0.60 (0.33)	0.54 (0.28)	0.68 (0.36)
	TPR	0.29 (0.17)	0.34 (0.20)	0.26 (0.15)	0.49 (0.23)	0.53 (0.23)	0.44 (0.20)
	FPR	0.01 (0.01)	0.00 (0.01)	0.00 (0.00)	0.02 (0.02)	0.02 (0.02)	0.01 (0.01)
	RME	3.81 (2.63)	3.80 (2.81)	5.30 (3.77)	4.76 (3.16)	4.98 (3.14)	5.80 (3.53)
4	ERR	1.23 (0.49)	1.22 (0.52)	1.85 (0.50)	0.42 (0.21)	0.44 (0.23)	0.52 (0.22)
	TPR	0.55 (0.22)	0.50 (0.23)	0.29 (0.19)	0.67 (0.19)	0.65 (0.22)	0.55 (0.22)
	FPR	0.01 (0.01)	0.01 (0.01)	0.00 (0.00)	0.02 (0.02)	0.03 (0.02)	0.02 (0.02)
	RME	5.03 (3.64)	5.12 (3.60)	6.01 (4.02)	6.70 (4.67)	6.37 (4.22)	7.14 (4.72)
5	ERR	1.70 (0.46)	1.55 (0.50)	1.84 (0.49)	0.64 (0.26)	0.61 (0.27)	0.70 (0.27)
	TPR	0.37 (0.16)	0.36 (0.17)	0.26 (0.16)	0.48 (0.17)	0.49 (0.20)	0.42 (0.18)
	FPR	0.00 (0.01)	0.00 (0.01)	0.00 (0.00)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)
	RME	5.86 (3.61)	6.08 (3.96)	5.89 (3.71)	8.62 (5.78)	8.56 (6.33)	8.82 (5.82)
6	ERR	1.92 (0.48)	1.84 (0.49)	1.99 (0.52)	0.79 (0.28)	0.78 (0.29)	0.80 (0.28)
	TPR	0.23 (0.14)	0.26 (0.15)	0.21 (0.15)	0.29 (0.19)	0.31 (0.20)	0.26 (0.19)
	FPR	0.00 (0.01)	0.00 (0.01)	0.00 (0.00)	0.02 (0.02)	0.02 (0.02)	0.01 (0.01)

Note: In each cell, mean (sd).

to the level of support provided by data. When data show little support to the assumption, the selected tuning may be small which can reduce the effect of the extra penalty. Therefore, this ‘‘counterintuitive’’ observation is actually reasonable.

Results for the low-dimensional cases are shown in Table A3, Supplementary material. Different from the high-dimensional cases, Alt.1 has high FPR values in all six scenarios, and the proposed method has a similar problem in Scenarios (1) to (3). Also, when the assumption is completely violated, advantages of the proposed method diminish.

To examine performance under an even smaller sample size, we conduct simulation for all three examples with a sample size of 200. Here the effective sample size is about 80. Results are shown in Tables B1–B6, Supplementary material. As expected, performance of all methods deteriorates. However, the relative performance remains similar to that observed above.

4 Data analysis

4.1 Analysis of SEER breast cancer data

Breast cancer is the most common invasive cancer among women in the USA and worldwide. It has been recognized that, with the advancement of treatment, some breast cancer patients can have extended survival, which can be viewed as cured. Here we analyze data from SEER, which is the largest cancer registry in the USA. The analyzed cohort consists of patients diagnosed between 2001 and 2005 and registered in the State of Connecticut, allowing for as long as 10 years of follow-up. The spatial variations in cancer survival have been noted in the literature. As such, it is reasonable to focus on one registry. Following the literature,[21] we consider female patients with active follow-up and breast cancer confirmed as the first primary cancer. The age at diagnosis was between 20 and 85, and the stage at diagnosis was worse than stage 0 (in situ). The analyzed covariates are provided in Table 4. It is noted that the covariates in SEER have been manually selected and are expected to have some role in breast cancer survival. After excluding records with missing covariates, 9550 cases are available for analysis. The survival plot is shown in Figure C1, Supplementary material. Although the plateau is not very long, there is very dense censoring in the tail, and the tail survival probability is greater than 0.6. To formally test whether the follow-up is sufficient, we apply the test developed by Maller and Zhou.[22] The test statistic is $(1 - N_n n)^n$, where N_n is the number of uncensored subjects in the time interval $(2T_n^* - T_n, T_n)$, where T_n is the

Table 4. Analysis of SEER breast cancer data: estimated coefficients.

	Proposed		Alt.1		Alt.2	
	γ	β	γ	β	γ	β
Race (reference=White)						
Black	0.19	0.15		0.16		0.08
Others	0.02	0.02				
Ethnicity (reference=Non-Spanish)						
Spanish	0.12	0.12		0.12		0.03
Age (reference=50–60)						
20–30	0.45	0.40				
30–40						
40–50	0.10	0.14		0.09		0.06
60–70	0.87	0.04	1.17	−0.21	1.19	−0.17
Above 70	2.47	0.12	2.65		2.70	
Marital status (reference=Married)						
Single	0.15	0.12		0.10		0.05
Separated/Divorced	0.09	0.08		0.04		
Widowed	0.09	0.10		0.06		
Receptor Status (reference=Positive)						
Negative	0.48	0.06	0.84	−0.07	0.86	−0.01
Stage (reference=Localized)						
Regional	0.74	0.06	0.93	−0.01	0.95	
Distant	4.36	1.39	4.35	1.35	4.59	1.34

last observed survival time and T_n^* is the last observed uncensored survival time. The test statistic is smaller than 10^{-10} compared to a 0.05 threshold,[22] and there is a strong evidence that the follow-up is sufficient. Also given the fact that breast cancer is medically curable, it is reasonable to apply the cure model. As there is no prior information regarding opposite signs, all the covariates are included in the sign-based penalty.

The estimated coefficients using the three approaches are shown in Table 4. It is observed that different approaches lead to different findings. The proposed approach identifies more covariate effects as being associated with cure and survival. As all of the analyzed covariates have been previously associated with breast cancer survival, this finding is reasonable. The individual findings are mostly consistent with the literature. It is observed that by introducing the sign-based penalty, the proposed approach has more consistent findings. For example for the variable receptor status, the proposed approach leads to the same sign in γ and β, whereas the two alternatives lead to conflicting signs. The sign consistent results can be easier to interpret in practice.

With real-world data, it is difficult to evaluate identification accuracy. We resort to a multi-splitting-based approach to evaluate prediction performance,[12] which may provide support to the overall validity of analysis. For the logistic and Cox parts, we use the AUC (area under the ROC curve) and C-statistic to assess prediction. They are calculated as follows: (1) split the data randomly into a training set of size $\frac{2n}{3}$ and a testing set of size $\frac{n}{3}$; (2) obtain estimates $\hat{\gamma}$ and $\hat{\beta}$ based on the training set for each method; (3) calculate risk scores $z_i^\top \hat{\gamma}$ and $z_i^\top \hat{\beta}$ for samples in the testing set; (4) compute the imputation-based AUC[23] based on $z_i^\top \hat{\gamma}$ for the logistic part and the inverse-probability-weighting-based C-statistic[24] based on $z_i^\top \hat{\beta}$ for the Cox part. With 200 splits, the median AUC and C-statistics are (0.76, 0.65), (0.78, 0.37), and (0.79, 0.36) for the proposed method, Alt.1, and Alt.2, respectively. The three approaches have similar performance for the logistic part. However, the proposed approach has significantly better prediction performance for the Cox model part.

4.2 Analysis of TCGA-KIRC data

We analyze the TCGA (The Cancer Genome Atlas) data on kidney renal clear cell carcinoma (KIRC), which is the most common subtype of kidney cancer. Data are available on 521 patients, among whom 172 died during follow-up. The survival plot is shown in Figure C2, Supplementary material. Bussy et al.[25] suggest that considering longer survivors can lead to more satisfactory results compared to the Cox model.

Applying the Maller and Zhou test gives a test statistic less than 10^{-10}, suggesting that it may be reasonable to apply the cure rate model. Similar to some published studies,[25–27] the goal is to identify genetic risk factors that are potentially associated with survival. Specifically, in this analysis we focus on gene expressions. We refer to the literature[28] for more details on gene expression profiling and data processing. In principle, the proposed approach (and alternatives) can be directly applied. To generate more reliable results, we focus on the MAPK (mitogen-activated protein kinase) signaling pathway. This pathway plays an essential role in cell proliferation and differentiation. It has been shown that its activation has an important effect in the tumorigenesis, metastasis, and angiogenesis of multiple cancers, including kidney renal cell carcinoma.[29,30] The set of analyzed genes is identified from Gene Ontology using the GSEA annotation package (http://www.broadinstitute.org/gsea). A total of 306 genes are analyzed. It is noted that the number of genes is larger than the effective sample size (number of events).

The analysis results are shown in Table 5. The estimates are sparser than in the previous example, which is reasonable considering the small set of KIRC related genes. Different approaches lead to different findings. The proposed approach has qualitatively more consistent findings for the two model parts. For example for gene SHC1, the proposed approach identifies positive effects for both model parts, whereas the two alternatives only identify its effects in γ.

Literature search suggests that the findings can be biologically sensible. For TRAF2 and WNT5A which are identified by all the three approaches, TRAF2 has been reported as a driver oncogene in many cancers,[31] and WNT5A has been reported as a tumor suppressor in KIRC.[32] There is also literature support for genes that are identified only by the proposed approach. Several of them are involved in the etiology and oncogenesis of KIRC and other cancers. For example, DAB2IP plays an important role in KIRC development as a tumor suppressor.[33] PAK1 belongs to multiple signaling pathways and contributes to the development and progression of tumor.[34]

Prediction evaluation is again conducted using the multi-splitting approach. The median AUC and C-statistic are computed as (0.78, 0.60), (0.77, 0.58), and (0.50, 0.57) for the three approaches. The proposed approach has a small advantage over Alt.1. Both approaches significantly outperform Alt.2.

Table 5. Analysis of TCGA-KIRC data: estimated coefficients.

	Proposed		Alt.1		Alt.2	
Gene	γ	β	γ	β	γ	β
BIRC7	0.02		0.11		0.06	
BMP4	−0.06	−0.07	−0.16		−0.17	
BRAF					−0.03	
DAB2IP		−0.02				
DVL3	0.14	0.27	0.21	0.37	0.20	
GHR		−0.01				
GRM1			−0.02			
IGF1R		−0.05				
IRAK1	0.02		0.12		0.09	
MAP3K13		−0.16	−0.15	−0.48		
MAPK8IP1	−0.05	−0.09	−0.03		−0.07	
MAPKAPK3		−0.01	−0.07			
NOD2		0.02				
PAK1		0.01				
PIK3R6		0.01				
RAPGEF1		−0.05				
SHC1	0.08	0.23	0.03		0.09	
TRAF2	0.14	0.23	0.07		0.14	
WNT5A	0.36	0.40	0.33		0.41	

5 Discussion

For the two-part cure rate model, in this study, we have focused on the structure of covariate effects. This may complement the existing works on estimation and selection. Different from the proportionality studies, the proposed method promotes sign consistency, that is, qualitative similarity in the two covariate effects. It adopts a novel penalization technique, has an intuitive formulation, and can be effectively realized. Extensive simulations and analysis of both low- and high-dimensional data establish its satisfactory properties. The two-part cure rate model is a special case of the mixture model. As such, the proposed method can be extended to more general mixture models. The penalization technique has been adopted for the promotion of sign consistency, variable selection, and regularized estimation. Other regularization techniques may also be applicable. In practice, there might be scientific reasoning or strong evidence on whether the long-term and short-term effects have consistent signs. If there is strong evidence on conflicting signs, the corresponding effects can be excluded from the sign penalty term. Without prior information, we can rely on the tuning parameter to decide how much weight to assign to the sign penalty. That is, penalty weight on sign difference can be data-adaptive under the penalization framework. In our data analysis, we compare different methods using AUC and C-statistic. It would be desirable to have a test that can test the significance between models with and without promoting sign consistency. Such work can be a future direction. Another future work is to establish the theoretical properties and develop methods for inference such as constructing confidence intervals using computer intensive methods.

Acknowledgements

We thank the editor and reviewers for their constructive comments which greatly improve the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was supported by the National Natural Science Foundation of China (Grant No. 71501089).

Supplemental Material

Supplemental material for this article is available online.

ORCID iD

Yuan Huang http://orcid.org/0000-0002-8011-3034

References

1. Berkson J and Gage RP. Survival curve for cancer patients following treatment. *J Am Stat Assoc* 1952; **47**: 501–515.
2. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J R Stat Soc Series B Stat Methodol* 1949; **11**: 15–53.
3. Othus M, Barlogie B, LeBlanc ML, et al. Cure models as a useful statistical tool for analyzing survival. *Clin Cancer Res* 2012; **18**: 3731–3736.
4. Farewell V. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982; **38**: 1041–1046.
5. Maller RA and Zhou X. *Survival analysis with long term survivors*. England, UK: John Wiley and sons, 1996.
6. Sy JP and Taylor JM. Estimation in a cox proportional hazards cure model. *Biometrics* 2000; **56**: 227–236.
7. Liu X, Peng Y, Tu D, et al. Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Stat Med* 2012; **31**: 2882–2891.
8. Scolas S, El Ghouch A, Legrand C, et al. Variable selection in a flexible parametric mixture cure model with interval-censored data. *Stat Med* 2016; **35**: 1210–1225.
9. Han C and Kronmal R. Two-part models for analysis of agatston scores with possible proportionality constraints. *Commun Stat Theory Meth* 2006; **35**: 99–111.
10. Liu H and Chan KS. Generalized additive models for zero-inflated data with partial constraints. *Scand J Stat* 2011; **38**: 650–665.
11. Fang K, Wang X, Shia BC, et al. Identification of proportionality structure with two-part models using penalization. *Comput Stat Data Anal* 2016; **99**: 12–24.
12. Fan X, Liu M, Fang K, et al. Promoting structural effects of covariates in the cure rate model with penalization. *Stat Methods Med Res* 2017; **26**: 2078–2092.
13. Liu A, Kronmal R, Zhou X, et al. Determination of proportionality in two-part models and analysis of multi-ethnic study of atherosclerosis (MESA). *Stat Interface* 2011; **4**: 475–487.
14. Huang Y, Zhang Q, Zhang S, et al. Promoting similarity of sparsity structures in integrative analysis with penalization. *J Am Stat Assoc* 2017; **112**: 342–350.
15. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010; **38**: 894–942.
16. Meng X and Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 1993; **80**: 267–278.
17. Rashid N, Sun W and Ibrahim JG. Some statistical strategies for dae-seq data analysis: variable selection and modeling dependencies among observations. *J Am Stat Assoc* 2014; **109**: 78–94.
18. Peng Y and Dear KB. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000; **56**: 237–243.
19. Breslow NE. Discussion on ‘‘Regression models and life tables’’ by D. R. Cox. *J R Stat Soc Series B Stat Methodol* 1972; **34**: 187–220.
20. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
21. Ries L, Young JL, Keel GE, et al. (eds) *SEER survival monograph: cancer survival among adults: U.S. SEER Program, 1988–2001, patient and tumor characteristics*. Bethesda, MD: National Cancer Institute, SEER Program, NIH Pub. No. 07—6215, 2007.
22. Maller RA and Zhou S. Testing for sufficient follow-up and outliers in survival data. *J Am Stat Assoc* 1994; **89**: 1499–1506.
23. Asano J, Hirakawa A and Hamada C. Assessing the prediction accuracy of cure in the cox proportional hazards cure model: an application to breast cancer data. *Pharm Stat* 2014; **13**: 357–363.
24. Asano J and Hirakawa A. Assessing the prediction accuracy of a cure model for censored survival data with long-term survivors: application to breast cancer data. *J Biopharm Stat* 2017; **26**: 918–932.
25. Bussy S, Guilloux A, Gaïffas S, et al. C-mix: A high-dimensional mixture model for censored durations, with applications to genetic data. *Stat Methods Med Res* 2018; 0962280218766389. DOI: 10.1177/0962280218766389
26. Zhao Q, Shi X, Xie Y, et al. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Brief Bioinform* 2014; **16**: 291–303.
27. Dimitrieva S, Schlapbach R and Rehrauer H. Prognostic value of cross-omics screening for kidney clear cell renal cancer survival. *Biol Direct* 2016; **11**: 68.
28. Yang W, Yoshigoe K, Qin X, et al. Identification of genes and pathways involved in kidney renal clear cell carcinoma. *BMC Bioinform* 2014; **15**: S2.
29. Network CGAR, et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 2013; **499**: 43–49.

30. Huang D, Ding Y, Luo WM, et al. Inhibition of MAPK kinase signaling pathways suppressed renal cell carcinoma growth and angiogenesis in vivo. *Cancer Res* 2008; **68**: 81–88.
31. Shen RR, Zhou AY, Kim E, et al. TRAF2 is an NF-κB-activating oncogene in epithelial cancers. *Oncogene* 2015; **34**: 209–216.
32. Xu Q, Krause M, Samoylenko A, et al. WNT signaling in renal cell carcinoma. *Cancers* 2016; **8**: 57.
33. Zhou J, Luo J, Wu K, et al. Loss of DAB2IP in RCC cells enhances their growth and resistance to mTOR-targeted therapies. *Oncogene* 2016; **35**: 4663–4674.
34. Radu M, Semenova G, Kosoff R, et al. Pak signaling during the development and progression of cancer. *Nat Rev Cancer* 2013; **14**: 13–25.