Omae *et al. BMC Medical Research Methodology* (2020) 20:182
https://doi.org/10.1186/s12874-020-01063-2

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Quasi-linear Cox proportional hazards model with cross-L_1 penalty

Katsuhiro Omae[1*] and Shinto Eguchi[2]

Abstract

Background: To accurately predict the response to treatment, we need a stable and effective risk score that can be calculated from patient characteristics. When we evaluate such risks from time-to-event data with right-censoring, Cox's proportional hazards model is the most popular for estimating the linear risk score. However, the intrinsic heterogeneity of patients may prevent us from obtaining a valid score. It is therefore insufficient to consider the regression problem with a single linear predictor.

Methods: we propose the model with a quasi-linear predictor that combines several linear predictors. This provides a natural extension of Cox model that leads to a mixture hazards model. We investigate the property of the maximum likelihood estimator for the proposed model. Moreover, we propose two strategies for getting the interpretable estimates. The first is to restrict the model structure in advance, based on unsupervised learning or prior information, and the second is to obtain as parsimonious an expression as possible in the parameter estimation strategy with cross-L_1 penalty. The performance of the proposed method are evaluated by simulation and application studies.

Results: We showed that the maximum likelihood estimator has consistency and asymptotic normality, and the cross-L_1-regularized estimator has root-*n* consistency. Simulation studies show these properties empirically, and application studies show that the proposed model improves predictive ability relative to Cox model.

Conclusions: It is essential to capture the intrinsic heterogeneity of patients for getting more stable and effective risk score. The proposed hazard model can capture such heterogeneity and achieve better performance than the ordinary linear Cox proportional hazards model.

Keywords: Cox's proportional hazards model, Generalized average, Heterogeneity, Mixture model, Survival analysis

Background

Medical science has made dramatic progress in recent years and reached the stage of trying to develop treatment tailored to patients' individual characteristics. In particular, it is becoming standard for doctors to prescribe selective therapeutic agents to cancer patients with specific oncogenes. Such personalized medicines are not only effective, but also economical and practical: if personalization fulfills its promise, patients no longer have to try expensive but ineffective treatments, or suffer from unnecessary side effects.

The idea of treatment individualization arose from the fact that patients often show different responses, in terms of both therapeutic and side effects, to the same specific treatments. Therefore, in order to realize individualized treatment, it is necessary to predict treatment risk accurately and carefully based on patients' characteristics. Because such a prediction should be performed in an objective manner, we need some quantified measurement of risk. This is usually achieved by a risk score, estimated by a regression model derived from various types of datasets. Survival time datasets are among the most popular source of data in medical science because

*Correspondence: katsuhiro.omae@gmail.com
[1]Department of Clinical Biostatistics, Graduate School of Medicine, Kyoto University, Yoshida Konoe-cho, Kyoto, Japan
Full list of author information is available at the end of the article

they focus on extension of time-to-event (in this case, an undesirable or bad event). Several types of events can be considered, including cancer prognosis or metastasis, myocardial infarction, and death. For time-to-event data with right-censoring, it is standard to apply the relative risk model. The main feature of the model is the assumption that hazards are proportional; i.e., the hazard ratio of two different subjects depends only on their covariates. Despite requiring this somewhat strong assumption, this type of model is used in a broad range of applications. Although the relative risk model covers a wider range, the exponential relative risk model, known as Cox's proportional hazards model, is most common for these applications. For simplicity, in the rest of this paper we assume that all covariates are time-independent and that there are no event ties, but these assumptions can be relaxed. In the Cox model, it is assumed that the log hazard is decomposed into and time-independent linear predictor of covariates vector $\boldsymbol{x}$ as $\log(h(t|\boldsymbol{x})/h_0(t)) = \boldsymbol{\beta}^\top \boldsymbol{x}$, where $h(t|\boldsymbol{x})$ is a hazard rate at time t, $h_0(t) = h(t|\mathbf{0})$ is called the baseline hazard function, and $\boldsymbol{\beta}$ is a coefficient vector.

However, the relationship between the hazard and covariates may not be common among subjects. For example, recent clinical studies focused on the fact that heterogeneity among such populations can result in different responses to the same treatment. Such complex relationships can no longer be described in a single hazard model. Therefore, we should consider a mixture hazard model to capture the different hazard patterns in the heterogeneous population. In fact, [1] and [2] introduced a general family of mixture hazard models to describe multimodal hazards, although these models were described in a limited situation under a parametric approach. Also, in the context of a cure model, a binary hazard mixture model was proposed in a semi-parametric manner [3]. In this paper, we propose a natural extension of Cox's proportional hazards model using a quasi-linear predictor that leads to a proportional model with mixture hazards.

The rest of the article is organized as follows. In "Methods" section, we derive the mixture hazard model via the quasi-linear predictor and two strategies are developed to obtain parsimonious expression: the restricted quasi-linear model and the cross-L_1-penalty estimation. In "Results" section, we investigate the estimators' asymptotic properties. Moreover, we present numerical simulations and applications to real data sets, respectively. The proofs for all propositions and theorems given as Appendix are available in Additional file 1.

Methods

Quasi-linear Cox Model

Formulations

Let t be the survival time of a subject with baseline covariate vector $\boldsymbol{x}$. Then we define the quasi-linear Cox model as

$$h(t|\boldsymbol{x},\boldsymbol{\pi},\boldsymbol{\beta}) = h_0(t)\exp\left(f_Q(\boldsymbol{x},\boldsymbol{\pi},\boldsymbol{\beta})\right). \quad (1)$$

Here, f_Q is a quasi-linear predictor function defined by the log-sum-exp averages of K linear predictors [4] as

$$f_Q(\boldsymbol{x},\boldsymbol{\pi},\boldsymbol{\beta}) = \log\left(\sum_{k=1}^{K}\pi_k\exp\left(\boldsymbol{\beta}_k^\top\boldsymbol{x}\right)\right), \quad (2)$$

where $\boldsymbol{\pi}^\top = (\pi_1,\cdots,\pi_K)$ is a vector of mixing proportion with $\sum_{k=1}^{K}\pi_k = 1$, and $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top,\cdots,\boldsymbol{\beta}_K^\top)$ is the coefficient vector. The relationship between hazard and covariates differs among subpopulations; the parameter K relies on the total number of subpopulations that satisfy this condition. We find that the quasi-linear Cox model can be understood as a mixture hazard model because from (1) and (2)

$$h(t|\boldsymbol{x},\boldsymbol{\pi},\boldsymbol{\beta}) = \sum_{k=1}^{K}\pi_k h_0(t)\exp\left(\boldsymbol{\beta}_k^\top\boldsymbol{x}\right). \quad (3)$$

The underlying hazard model in (3) is described as $h(t|\boldsymbol{x},\boldsymbol{\pi},\boldsymbol{\beta}) = \sum_{k=1}^{K}\pi_k h_k(t|\boldsymbol{x},\boldsymbol{\beta}_k)$, where $h_k(t|\boldsymbol{x},\boldsymbol{\beta}_k) = h_{0k}(t)\exp(\boldsymbol{\beta}_k^\top\boldsymbol{x})$ with the assumption that $h_{0k}(t) = h_0(t)$ for any k. Thus the proposed model can be understood as the special case of the mixture of Cox's proportional hazards models. The assumption about equality of the baseline hazard function may seem somewhat stronger, but this simplifies the model formulation and interpretability. Simulations and application studies in "Simulations" section and "Application" section show that model (3) has sufficient predictive ability. A more general model that removes the assumption of equality of the baseline hazard function is discussed in the "Discussion" section.

Partial likelihood and maximum likelihood estimator

Consider the data $(\boldsymbol{x}_i, t_i, \delta_i)$ $(i = 1,2,\cdots,n)$ from n subjects, where $\boldsymbol{x}_i$ is a p-dimensional covariates vector, t_i is observed survival or censored time, and δ_i is an event indicator which takes a value of 1 if the sample experiences the event by $t = t_i$ and 0 otherwise. We assume that t_i and δ_i are independent for all subjects. Let $\boldsymbol{\theta}^\top = (\boldsymbol{\pi}^\top, \boldsymbol{\beta}^\top)$ and $\boldsymbol{\theta}_k^\top = (\pi_k, \boldsymbol{\beta}_k^\top)$ for any k. Then, the partial log-likelihood function of the parameter $\boldsymbol{\theta}$ is written as

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\delta_i\left\{\log\left(\sum_{k=1}^{K}\eta_i(\boldsymbol{\theta}_k)\right) - \log\left(\sum_{\ell\in R(t_i)}\sum_{k=1}^{K}\eta_\ell(\boldsymbol{\theta}_k)\right)\right\}, \quad (4)$$

where $R(t_i) = \{l\in\{1,\cdots,n\}|t_i\le t_\ell\}$ and $\eta_i(\boldsymbol{\theta}_k) = \pi_k\exp(\boldsymbol{\beta}_k^\top\boldsymbol{x}_i)$. The $R(t_i)$ denotes the risk set at time t_i. The maximum partial likelihood estimator $\hat{\boldsymbol{\theta}}$ of (4) has consistency and asymptotic normality, as shown in "Asymptotic properties" section. Because we cannot get the estimates analytically, as with the Cox's proportional hazards model, we need some numerical optimization

method. As an example of such a method, the outline of the Minorization-Maximization (MM) algorithm [5] is shown here. The convergence property of the algorithm is demonstrated in Appendix A.

First, the score function of the partial likelihood (4) consists of the following elements:

$$\frac{\partial}{\partial \pi_m} l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \left(\frac{p_{mi}(\boldsymbol{\theta}) - p^*_{mi}(\boldsymbol{\theta})}{\pi_m} \right) \tag{5}$$

and

$$\frac{\partial}{\partial \boldsymbol{\beta}_m} l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \left(p_{mi}(\boldsymbol{\theta}) x_i - p^*_{mi}(\boldsymbol{\theta}) \frac{\sum_{\ell \in R(t_i)} \eta_\ell(\boldsymbol{\theta}_m) x_l}{\sum_{\ell \in R(t_i)} \eta_\ell(\boldsymbol{\theta}_m)} \right), \tag{6}$$

where

$$p_{mi}(\boldsymbol{\theta}) = \frac{\eta_i(\boldsymbol{\theta}_m)}{\sum_{k=1}^{K} \eta_i(\boldsymbol{\theta}_k)} = \frac{\pi_m \exp\left(\boldsymbol{\beta}_m^\top x_i\right)}{\sum_{k=1}^{K} \pi_k \exp\left(\boldsymbol{\beta}_k^\top x_i\right)},$$

$$\begin{aligned} p^*_{mi}(\boldsymbol{\theta}) &= \frac{\sum_{\ell \in R(t_i)} \eta_\ell(\boldsymbol{\theta}_m)}{\sum_{\ell \in R(t_i)} \sum_{k=1}^{K} \eta_\ell(\boldsymbol{\theta}_k)} \\ &= \frac{\sum_{\ell \in R(t_i)} \pi_m \exp\left(\boldsymbol{\beta}_m^\top x_i\right)}{\sum_{\ell \in R(t_i)} \sum_{k=1}^{K} \pi_k \exp\left(\boldsymbol{\beta}_k^\top x_i\right)}. \end{aligned}$$

We remark that the function $l(\boldsymbol{\theta})$ is rather complicated relative to the one of Cox's model, which typically contains summands in the logarithmic function. In fact, when $K = 1$, (4) is reduced to the standard form of the partial log-likelihood function $l(\boldsymbol{\beta}_1)$ in which the gradient vector becomes

$$\frac{\partial}{\partial \boldsymbol{\beta}_1} l(\boldsymbol{\beta}_1) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{x}_i - \frac{\sum_{\ell \in R(t_i)} \exp\left(\boldsymbol{\beta}_1^\top \boldsymbol{x}_\ell\right) \boldsymbol{x}_\ell}{\sum_{\ell \in R(t_i)} \exp\left(\boldsymbol{\beta}_1^\top \boldsymbol{x}_\ell\right)} \right). \tag{7}$$

Compared with (7), repeated calculation of (5) and (6) is computationally hard. We therefore consider a simpler function as

$$\begin{aligned} G(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = l(\boldsymbol{\theta}_0) &+ \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_i p_{ki}(\boldsymbol{\theta}_0) \log \frac{\pi_k \exp\left(\boldsymbol{\beta}_k^\top \boldsymbol{x}_i\right)}{\pi_{0k} \exp\left(\boldsymbol{\beta}_{0k}^\top \boldsymbol{x}_i\right)} \\ &- \sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{\ell \in R(t_i)} \sum_{k=1}^{K} \pi_k \exp\left(\boldsymbol{\beta}_k^\top \boldsymbol{x}_\ell\right)}{\sum_{\ell \in R(t_i)} \sum_{k=1}^{K} \pi_{0k} \exp\left(\boldsymbol{\beta}_{0k}^\top \boldsymbol{x}_\ell\right)} - 1 \right\}, \end{aligned}$$

which has only feasible terms of log hazard and log cumulative hazard functions. Thus, we observe that

$$\frac{\partial}{\partial \pi_m} G(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \sum_{i=1}^{n} \delta_i \left\{ \frac{p_{mi}(\boldsymbol{\theta}_0)}{\pi_m} - \frac{p^*_{mi}(\boldsymbol{\theta}_0)}{\pi_m} \frac{\sum_{\ell \in R(t_i)} \pi_m \exp\left(\boldsymbol{\beta}_m^\top \boldsymbol{x}_\ell\right)}{\sum_{\ell \in R(t_i)} \pi_{0m} \exp\left(\boldsymbol{\beta}_{0m}^\top \boldsymbol{x}_\ell\right)} \right\} \tag{8}$$

and

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}_m} G(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \sum_{i=1}^{n} \delta_i \Bigg\{ & p_{mi}(\boldsymbol{\theta}_0) \boldsymbol{x}_i \\ & - p^*_{mi}(\boldsymbol{\theta}_0) \frac{\sum_{\ell \in R(t_i)} \pi_m \exp\left(\boldsymbol{\beta}_m^\top \boldsymbol{x}_\ell\right) \boldsymbol{x}_\ell}{\sum_{\ell \in R(t_i)} \pi_{0m} \exp\left(\boldsymbol{\beta}_{0m}^\top \boldsymbol{x}_\ell\right)} \Bigg\}, \end{aligned} \tag{9}$$

which leads to $\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} G(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|_{\boldsymbol{\theta}_0 = \boldsymbol{\theta}}$. Exploring these properties, we propose a learning algorithm $\left\{ \boldsymbol{\theta}^{(s)} = (\boldsymbol{\pi}^{(s)}, \boldsymbol{\beta}^{(s)}) | s \in \mathcal{S} \right\}$ for the maximum partial likelihood estimator of $\boldsymbol{\theta}$ by sequential maximization of $G(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ as $\boldsymbol{\theta}^{(s+1)} = \text{argmax}_{\boldsymbol{\theta}} G\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s)}\right)$ for all s in $\mathcal{S} = \{1, 2, \cdots, S\}$, where $\boldsymbol{\theta}^{(1)}$ is an initial value and S denotes a stopping time. By definition, we obtain $G\left(\boldsymbol{\theta}^{(s+1)}, \boldsymbol{\theta}^{(s)}\right) \geq G(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^{(s)})$. It follows from (8) and (9) that the iteration step is given by $\boldsymbol{\theta}^{(s+1)} = \left(\tilde{\pi}_k\left(\boldsymbol{\theta}^{(s)}\right), \tilde{\boldsymbol{\beta}}_k\left(\boldsymbol{\theta}^{(s)}\right)\right)$, where

$$\tilde{\pi}_k(\boldsymbol{\theta}) = \frac{1}{z(\boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}_k(\boldsymbol{\theta}^{(s)}))} \frac{\sum_{i=1}^{n} \delta_i p_{ki}(\boldsymbol{\theta})}{\sum_{i=1}^{n} \delta_i p^*_{ki}(\boldsymbol{\theta}) \frac{\sum_{\ell \in R(t_i)} \exp\left(\tilde{\boldsymbol{\beta}}_k^\top \boldsymbol{x}_\ell\right)}{\sum_{\ell \in R(t_i)} \exp\left(\boldsymbol{\beta}_k^\top \boldsymbol{x}_\ell\right)}} \pi_k, \tag{10}$$

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_k(\boldsymbol{\theta}) &= \text{argsolve}_{\boldsymbol{b}_{k|\boldsymbol{\theta}}} \left\{ \sum_{i=1}^{n} \delta_i p_{ki}(\boldsymbol{\theta}) \boldsymbol{x}_i \right. \\ &\left. = \sum_{i=1}^{n} \delta_i p^*_{ki}(\boldsymbol{\theta}) \frac{\sum_{\ell \in R(t_i)} \exp\left(\boldsymbol{b}_k^\top \boldsymbol{x}_\ell\right) \boldsymbol{x}_\ell}{\sum_{\ell \in R(t_i)} \exp\left(\boldsymbol{\beta}_k^\top \boldsymbol{x}_\ell\right)} \right\}, \end{aligned} \tag{11}$$

where

$$z\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}_k\left(\boldsymbol{\theta}^{(s)}\right)\right) = \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} \delta_i p_{ki}(\boldsymbol{\theta})}{\sum_{i=1}^{n} \delta_i p^*_{ki}(\boldsymbol{\theta}) \frac{\sum_{\ell \in R(t_i)} \exp\left(\tilde{\boldsymbol{\beta}}_k^\top \boldsymbol{x}_\ell\right)}{\sum_{\ell \in R(t_i)} \exp\left(\boldsymbol{\beta}_k^\top \boldsymbol{x}_\ell\right)}} \pi_k.$$

We observe that estimating equation in (11) is a weighted variant of standard partial likelihood equation. Furthermore, we observe that the minus Hessian matrix of $G(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ with respect to $\boldsymbol{\theta}$ is positive-definite, which guarantees that $\boldsymbol{\theta}^{(s+1)} = \left(\tilde{\pi}_k\left(\boldsymbol{\theta}^{(s)}\right), \tilde{\boldsymbol{\beta}}_k\left(\boldsymbol{\theta}^{(s)}\right)\right)$, as discussed above, is the unique minimizer of $G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s)})$ in $\boldsymbol{\theta}$. We observe the basic property of the learning algorithm, $\left\{ \boldsymbol{\theta}^{(s)} = \left(\pi^{(s)}, \boldsymbol{\beta}^{(s)}\right) | s \in \mathcal{S} \right\}$, as follows.

Proposition 1 *Let* $\left\{ \boldsymbol{\theta}^{(s)} = \left(\pi^{(s)}, \boldsymbol{\beta}^{(s)}\right) | s \in \mathcal{S} \right\}$ *be the fixed-point algorithm defined by the iteration rules* (10) *and* (11). *Then, the partial log-likelihood function* $l(\boldsymbol{\theta})$ *increases on the sequence* $\left\{ \boldsymbol{\theta}^{(s)} | s \in \mathcal{S} \right\}$ *as* $l\left(\boldsymbol{\theta}^{(s+1)}\right) \geq l\left(\boldsymbol{\theta}^{(s)}\right)$ *for any* $s = 1, \cdots, S - 1$.

The proof of Proposition 1 is given in Appendix B. The convergence of the algorithm $\left\{\boldsymbol{\theta}^{(s)} : s \geq 1\right\}$ to the maximum partial likelihood estimator $\hat{\boldsymbol{\theta}}$ is not directly connected to Proposition 1. We need to make some assumption about the model in order to guarantee convergence, similar to that of expectation–maximization (EM) algorithm [6] for the analytic conditions. For example, we assume that $l(\boldsymbol{\theta})$ is unimodal, with $\boldsymbol{\theta}^*$ being the only stationary point. We note that $\partial G(\boldsymbol{\theta}, \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$ is continuous for $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Thus, the sequence $\{\boldsymbol{\theta}^{(s)}\}$ converges to the unique maximizer $\boldsymbol{\theta}^*$ [7]. In fact, the partial likelihood function $l(\boldsymbol{\theta})$ is expressed as a difference of two concave functions $\psi_1(\boldsymbol{\theta})$ and $\psi_2(\boldsymbol{\theta})$, where $\psi_1(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \delta_i \log \sum_{k=1}^{K} \pi_k \exp\left(\boldsymbol{\beta}_k^\top \boldsymbol{x}_i\right)$ and $\psi_2(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \delta_i \log \sum_{j \in R(t_i)} \sum_{k=1}^{K} \pi_k \exp\left(\boldsymbol{\beta}_k^\top \boldsymbol{x}_j\right)$. Hence, the assumption for the unimodality is necessary for convergence.

Parsimonious Model

The quasi-linear Cox model consists of a relatively large number of parameters. Moreover, each covariate has multiple roles in every linear predictor. These complexities compromise the stability of parameter estimation and the interpretability of the model overall. Accordingly, we need a more parsimonious expression as shown in Fig. 1. To obtain a more parsimonious model, we propose a variant of the proposed model and parameter estimation procedure: restricted quasi-linear Cox model and cross-L_1-penalty method. The former idea relies on restricting the model structure in advance based on prior information. If there is prior knowledge that some factors strongly depends on the hazard of subpopulation and weakly on the hazards of other subpopulations, then we use the restricted quasi-linear form to insert the knowledge into consideration. If this is not the case, then a penalty is needed to bring full model (3) closer to the parsimonious model (13). We achieve this by using cross-L_1 penalty introduced in "Quasi-linear Cox model with cross L_1 penalty" section.

Restricted quasi-linear Cox model

The first strategy is to use the idea of disjoint sets of covariates, as proposed by [4]. In the strategy, we assume that we know the disjoint decomposition of $\boldsymbol{x}_i$ as $\boldsymbol{x}_{i(1)}, \cdots, \boldsymbol{x}_{i(K)}$ with a fixed group size K, and that this is identical among individuals. We denote the size of $\boldsymbol{x}_{i(k)}$ as p_k, where $\sum_{k=1}^{K} p_k = p$. We note that such decomposition is given by prior knowledge about the disjoint structure of $\boldsymbol{x}$. The disjoint sets of covariates yield the restricted quasi-linear predictor defined by

$$f_Q^{Res}\left(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \cdots, \boldsymbol{x}_{(K)}, \boldsymbol{\pi}, \boldsymbol{\beta}\right) = \log\left(\sum_{k=1}^{K} \pi_k \exp(\boldsymbol{\beta}_k^\top \boldsymbol{x}_{(k)})\right). \quad (12)$$

The mixture hazard model (1) is modified by replacing f_Q with f_Q^{Res} as

$$h^{Res}\left(t|\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(K)}, \boldsymbol{\pi}, \boldsymbol{\beta}\right) = h_0(t) \exp\left(f_Q^{Res}\left(\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(K)}, \boldsymbol{\pi}, \boldsymbol{\beta}\right)\right). \quad (13)$$

The maximum partial likelihood estimator is calculated by the fixed-point algorithm proposed for the non-restricted version, with easy modifications.

Quasi-linear Cox model with cross L_1 *penalty*

In the second strategy, we regularize the log-likelihood function by cross-L_1 penalty defined by

$$P_c(\boldsymbol{\beta}) = n\lambda_c \sum_{\ell \neq m} \sum_{j=1}^{p} \frac{|\beta_{\ell j}\beta_{mj}|}{|\hat{\beta}_{\ell j}\hat{\beta}_{mj}|}, \quad (14)$$

where λ_c is a regularization parameter and β_{kj} and $\hat{\beta}_{kj}$ are the j-th component of k-th coefficient vector $\boldsymbol{\beta}_k$ and corresponding maximum partial likelihood estimator, respectively. When λ_c goes to infinity, the estimated parameter of the $\boldsymbol{\beta}_k$'s would be cross-sparse; if $\beta_{\ell j} \neq 0$, then $\beta_{kj} = 0$ for any $k \neq \ell$, and the estimated model belongs to the class of restricted quasi-linear models (13). We define the regularized log-likelihood function with cross-L_1 penalty as

$$l^{pen}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - P_2(\boldsymbol{\beta}) - P_c(\boldsymbol{\beta}), \quad (15)$$

where $P_2(\boldsymbol{\beta}) = n\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}$, known as the L_2 penalty. We note that an additional regularization factor such as L_1 penalty yields the elastic net-type regularization $l^{pen*}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - \nu P_1(\boldsymbol{\beta}) - (1-\nu)P_2(\boldsymbol{\beta}) - P_c(\boldsymbol{\beta})$. We refer to the maximizer $(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\beta}})$ of $l^{pen}(\boldsymbol{\theta})$ as *CLASSO* (Cross least absolute shrinkage and selection operator) estimator. The penalty (14) is a variant of adaptive L_1 penalty originally introduced by [8]. The adaptive weights $|\hat{\beta}_{\ell j}\hat{\beta}_{mj}|$ are needed to equip *CLASSO* estimator with root-n consistency (see Theorem 2 in "Asymptotic properties" section). We make the following proposition regarding the *CLASSO* estimator.

Proposition 2 *Let $\mathcal{R}_c$ be a region in $\mathbb{R}^{pK}$ defined as $\mathcal{R}_c = \left\{\boldsymbol{\beta} \in \mathbb{R}^{pK} | P_c(\boldsymbol{\beta}) \leq c, P_2(\boldsymbol{\beta}) \leq c\right\}$. Then the region $\mathcal{R}_c$ is a convex set.*

Due to the convexity of the sum of cross-L_1 and -L_2 penalties, the *CLASSO* estimator also has consistency and asymptotic normality as shown in Theorem 2. Empirically, however, we do not need to regularize the partial log-likelihood by $P_2(\boldsymbol{\beta})$ for stable estimation. Therefore, we consider only the *CLASSO* penalty in the empirical studies in the Simulations and Applications sections.

To get the *CLASSO* estimator, we use the full gradient algorithm [9] in the updating step for each $\boldsymbol{\beta}_k$ (9). For each s-th iteration, we need to update $\boldsymbol{\beta}^{(s-1)}$ to get $\boldsymbol{\beta}^{(s)}$

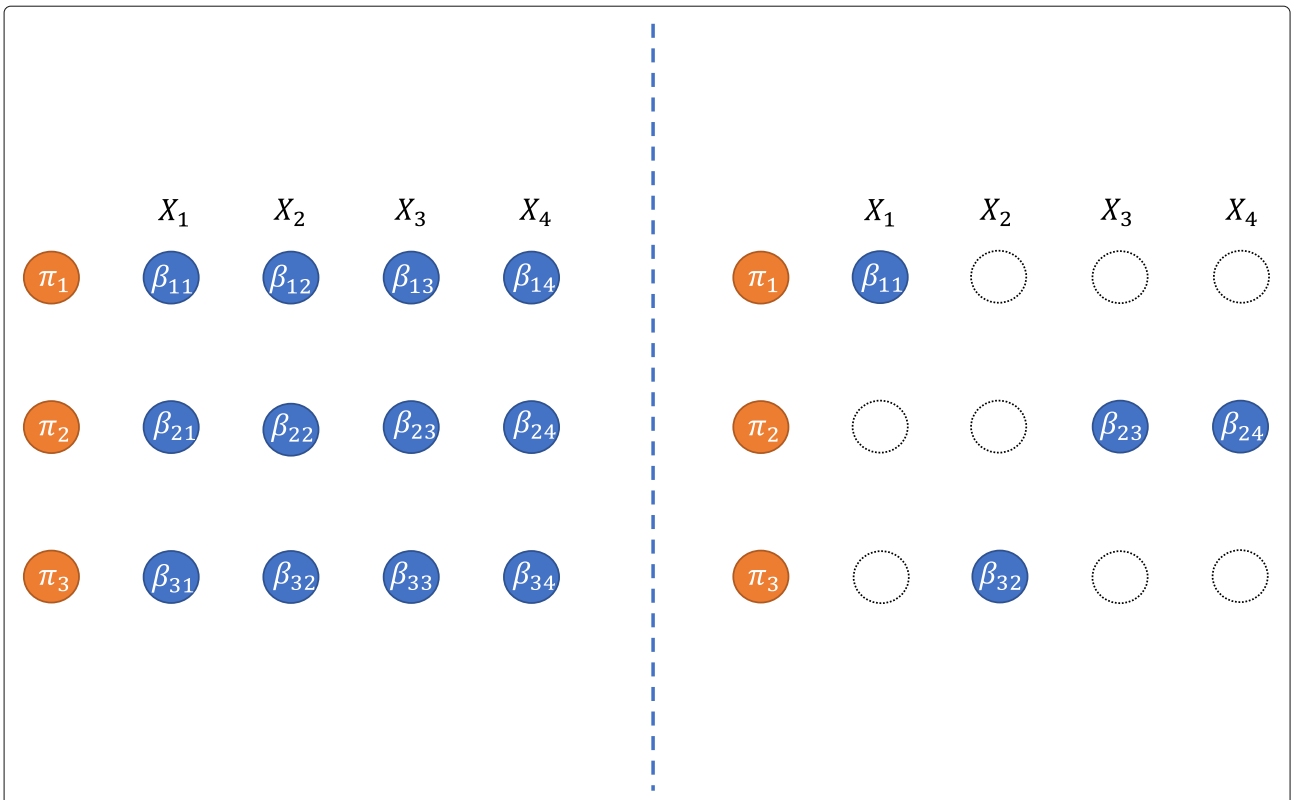

Fig. 1 The conceptual diagram of the full and parsimonious models in the setting of $p = 4$ and $K = 3$ are drawn. The left figure shows the full model by the quasi-linear predictor f_Q. The right figure shows the parsimonious model written in the same predictor but has some zero-coefficients

by gradient algorithm. Let S_s be the stopping time in the s-th iteration step. The initial value $\boldsymbol{\beta}_k^{(s+1,0)} = \boldsymbol{\beta}_k^{(s,S_s)}$ is repeatedly updated as

$$\boldsymbol{\beta}_m^{(s+1,u+1)} = \boldsymbol{\beta}_m^{(s+1,u)} + \min\left\{t_{\text{opt}}(\boldsymbol{\pi}^{(s)}, \boldsymbol{\beta}^{(s+1,u)}), t_{\text{edge}}(\boldsymbol{\pi}^{(s)}, \boldsymbol{\beta}^{(s+1,u)})\right\} \boldsymbol{d}_m\left(\boldsymbol{\pi}^{(s)}, \boldsymbol{\beta}^{(s+1,u)}\right), \tag{16}$$

where

$$\boldsymbol{d}_m(\boldsymbol{\theta}) = (d_{m1}(\boldsymbol{\theta}), d_{m2}(\boldsymbol{\theta}), \cdots, d_{mp}(\boldsymbol{\theta}))^\top,$$

$$t_{\text{edge}}(\boldsymbol{\theta}) = \min_{1\le j\le p}\left(-\frac{\beta_{mj}}{d_{mj}(\boldsymbol{\theta})} : \text{sign}(\beta_{mj}) = -\text{sign}(d_{mj}(\boldsymbol{\theta})) \neq 0\right),$$

and

$$t_{\text{opt}}(\boldsymbol{\theta}) = \frac{|d_m(\boldsymbol{\theta})|}{d_m(\boldsymbol{\theta})^\top \left\{\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right\} d_m(\boldsymbol{\theta})}.$$

Here

$$d_{mj}(\boldsymbol{\theta}) = \begin{cases} \frac{\partial G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s)})}{\partial \beta_{mj}} - \lambda_c \left(\sum_{k\neq m} |\beta_{kj}|\right) \text{sign}(\beta_{mj}) & \text{if } \beta_{mj} \neq 0 \\ \frac{\partial G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s)})}{\partial \beta_{mj}} - \lambda_c \left(\sum_{k\neq m} |\beta_{kj}|\right) \text{sign}\left(\frac{\partial G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s)})}{\partial \beta_{mj}}\right) & \text{if } \beta_{mj} = 0, \\ \left|\frac{\partial G\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s)}\right)}{\partial \beta_{mj}}\right| > \lambda_c \left(\sum_{k\neq m} |\beta_{kj}|\right) & \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

for $j = 1, \cdots, p$, where $\text{sign}(\cdot)$ is a sign function defined by setting $\text{sign}(z)$ equal to 1 for $z > 0$, 0 for $z = 0$ and -1 for $z < 0$. In each step, t_{opt} provides the optimal solution of the gradient descent algorithm, and t_{edge} controls the direction of the gradient so as not to change the signs of parameters.

For all analysis in Simulations and Applications sections, the initial values of parameters were set to the equal probability weighting parameters $\pi_k = 1/K$ for $k = 1, 2, \cdots, K$ and the coefficient vectors of Cox's proportional hazard models estimated from random K-samples sets on the parameter estimation of the quasi-linear Cox model. The tuning parameter λ_c is determined by any model selection criteria such as AIC or estimated test AUC from bootstrap estimates other than BIC. In this paper, we use Bayes Information Criteria (BIC) [10] because (i) it is one of the most popular information criteria, (ii) it is computationally easy to calculate compared with the estimator which relies on the bootstrap sampling and (iii) it has the consistency in model selection.

Results

Asymptotic properties

In this section, we provide an asymptotic property of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ and the coefficient part

of *CLASSO* estimator $\tilde{\boldsymbol{\beta}}$. In the following, our discussion is based on the stochastic process. For the i-th individual, let $N_i(t) = 1_{\{t_i \le t, \delta_i = 1\}}(t)$ be the right-continuous counting process, where each $N_i(t)$ counts the number of observed events on $(0, t]$, and let $Y_i(t) = 1_{\{t_i \ge t, c_i \ge t\}}(t)$ be the left-continuous at-risk process that shows the observation status at time t, where c_i and t_i are censoring and true survival times. Here 1_E is an indicator function defined by setting $1_E(t)$ equal to 1 for $t \in E$ and equal to 0 for $t \notin E$. Let us denote $\mathcal{F}_t = \sigma\left\{N_i(u), Y_i(u^+); i = 1, \cdots, n; 0 \le u \le t\right\}$ as the σ-algebra generated by all $N_i(u)$ and $Y_i(u)$, $0 \le u \le t$. Then, the corresponding intensity process of $N_i(t)$ is defined by $\Lambda_i(t)dt = P(dN_i(t) = 1|\mathcal{F}_{t\_})$ and the proposed model is described as

$$P(dN_i(t) = 1|\mathcal{F}_{t\_}) = Y_i(t)h_0(t)\exp(f_Q(\boldsymbol{x}, \boldsymbol{\theta}_0)), \qquad (18)$$

where $\boldsymbol{\theta}_0 = (\boldsymbol{\pi}_0^\top, \boldsymbol{\beta}_0^\top)^\top$. Then we get two theorems about the maximum partial likelihood estimator and *CLASSO* estimator.

Theorem 1 *Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}^\top, \hat{\boldsymbol{\beta}}^\top)^\top$ be the maximum partial likelihood estimator in the quasi-linear Cox model. Assume that the regularity conditions A-D (in Appendix C) hold. Then it follows that*

1. *(Consistency)* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta_0}$
2. *(Asymptotic Normality)* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0))$

Theorem 2 *Let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\pi}}^\top, \tilde{\boldsymbol{\beta}}^\top)^\top$ be the CLASSO estimator in the quasi-linear Cox model with cross* L_1 *penalty. Assume that condition A-D (in Appendix C) hold. If $\sqrt{n}\lambda_n \to \infty$, then $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_p(n^{-1/2})$.*

Theorem 1 shows that the partial maximum likelihood estimator has the consistency and asymptotic normality. Theorem 2 shows that by choosing a proper sequence λ_n, there exists a $\sqrt{n}$-consistent *CLASSO* estimator. The proofs are given in Appendix C and D.

Simulations

Settings

We conducted the simulations described in this section with two objectives in mind. The first was to ascertain whether the consistency of the maximum partial-likelihood estimator can be observed empirically. The second was to ascertain whether tuning parameter $\lambda^{(c)}$ selection can be performed efficiently using the BIC.

In all simulation studies introduced here, the inverse function method was used for data generation. First, the covariates $\boldsymbol{x}$ were generated from the multivariate normal distribution, the apparent censored time T_1 was generated from the exponential distribution with mean 1000, and the random variable U was generated from the uniform distribution on $[0, 1]$. Let the baseline survival time be followed the exponential distribution with mean 100. Then, the true survival time corresponding to the log relative risk function $f_Q^{Res}(\boldsymbol{x}; \boldsymbol{\theta})$ was given as $T_2 = -(\log(U)/100)\exp\left(f_Q^{Res}(\boldsymbol{x}; \boldsymbol{\theta})\right)$. Based on T_1 and T_2, let $T = \min(T_1, T_2)$ be the observational survival time and $\delta = I(T_1 < T_2)$ be the censored indicator before the event time.

Sample size was set to $N = 400$ in all scenarios, and it was assumed that the true number of the groups were known for all settings. The tuning parameter λ of cross-L_1 penalty was determined by BIC. We note that the maximum candidate value of the tuning parameter λ was controlled sufficiently to achieve the restricted quasi-linear form for every setting. We had the following options:

- *Independent Setting (IS) or Dependent Setting (DS)*
  Each covariate vector $\boldsymbol{x}_i$ was sampled from the standard normal distribution $\mathrm{N}(\mathbf{0}_p, \Sigma)$, where $\boldsymbol{a}_p = (a, a, \cdots, a) \in \mathbb{R}^p$. In IS, $\Sigma = 2^2 I_p$, where I_p is an identity matrix of size p. In DS, $\Sigma = (s_{ij}) \in \mathbb{R}^{p \times p}$, where $s_{ij} = 2^2 \times 0.7^{|i-j|}$.
- *Group Size and Coefficients*
  A number of combined linear predictors, namely group size, was set to $K = 2$ or $K = 3$. A number of covariates, namely dimension size, was set to $p = 2$, $p = 3$, or $p = 5$. A coefficients vector was set to cross-sparse (Scenario 1,3,5) or overlapped (Scenario 2,4,6) according to the following settings.

  1. $K = 2, p = 2, \boldsymbol{\pi}^\top = (0.3, 0.7)$ and $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top) = ((1, 0), (0, 1.5))$
  2. $K = 2, p = 2, \boldsymbol{\pi}^\top = (0.3, 0.7)$ and $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top) = ((1, 0.5), (0, 1.5))$
  3. $K = 2, p = 5, \boldsymbol{\pi}^\top = (0.3, 0.7)$ and $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top) = ((1, 1, 1, 0, 0), (0, 0, 0, 1.5, 1.5))$
  4. $K = 2, p = 5, \boldsymbol{\pi}^\top = (0.3, 0.7)$ and $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top) = ((1, 1, 1, 0, 0.5), (0, 0.25, 0.5, 1.5, 1.5))$
  5. $K = 3, p = 3, \boldsymbol{\pi}^\top = (0.2, 0.3, 0.5)$ and $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top) = ((1, 0, 0), (0, 1.5, 0), (0, 0, 1))$
  6. $K = 3, p = 3, \boldsymbol{\pi}^\top = (0.2, 0.3, 0.5)$ and $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top) = ((1, 0.5, 0), (0, 1.5, 0.5), (0.5, 0, 1))$

Simulation results

For each scenario, the mean values of the estimated coefficients and the mean-squared errors (MSEs) are shown in Table 1. Throughout the whole scenario, all coefficients were estimated with only a little bias. In particular, in the disjoint setting (Scenario 1,3,5 for IS and DS) we could almost certainly distinguish the zero coefficients from non-zero coefficients, indicating that BIC and

Table 1 Simulation results

	Setting		π_1	π_2	π_3	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{31}	β_{32}	β_{33}
	(1)	Mean	0.31	0.69	-	1.00	0.00	-	-	-	0.00	1.52	-	-	-	-	-	-
		MSE$\times 10^2$	0.34	0.34	-	0.58	0.06	-	-	-	0.07	0.84	-	-	-	-	-	-
	(2)	Mean	0.31	0.69	-	1.01	0.49	-	-	-	0.00	1.52	-	-	-	-	-	-
		MSE$\times 10^2$	1.02	1.02	-	1.10	1.04	-	-	-	0.37	0.98	-	-	-	-	-	-
IS	(3)	Mean	0.30	0.70	-	1.01	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.50	1.50	-	-	-
		MSE$\times 10^2$	0.21	0.21	-	0.02	0.01	0.01	0.67	0.53	0.41	0.51	0.42	0.00	0.00	-	-	-
	(4)	Mean	0.31	0.69	-	0.99	0.99	0.98	-0.01	0.46	-0.01	-0.23	0.48	1.48	1.48	-	-	-
		MSE$\times 10^2$	0.53	0.53	-	0.66	0.81	0.72	0.50	0.78	0.25	0.31	0.36	1.04	0.88	-	-	-
	(5)	Mean	0.21	0.30	0.50	1.00	0.00	0.00	-	-	0.00	1.52	0.00	-	-	0.00	0.00	1.01
		MSE$\times 10^2$	0.38	0.33	0.47	1.14	0.00	0.00	-	-	0.01	1.04	0.00	-	-	0.02	0.00	0.68
	(6)	Mean	0.21	0.31	0.48	1.01	0.35	0.02	-	-	0.01	1.52	0.45	-	-	0.46	0.01	1.02
		MSE$\times 10^2$	0.74	0.74	0.75	1.28	4.34	0.51	-	-	0.31	1.29	0.76	-	-	0.61	0.39	0.65
	(1)	Mean	0.31	0.69	-	1.00	0.00	-	-	-	0.00	1.52	-	-	-	-	-	-
		MSE$\times 10^2$	0.29	0.29	-	1.00	0.08	-	-	-	0.05	0.72	-	-	-	-	-	-
	(2)	Mean	0.29	0.71	-	1.11	0.38	-	-	-	-0.01	1.53	-	-	-	-	-	-
		MSE$\times 10^2$	4.33	4.33	-	9.64	9.33	-	-	-	3.64	3.90	-	-	-	-	-	-
DS	(3)	Mean	0.30	0.70	-	1.01	1.00	1.01	0.00	0.00	0.00	0.00	0.00	1.52	1.51	-	-	-
		MSE$\times 10^2$	0.22	0.22	-	0.76	1.33	1.06	0.00	0.00	0.04	0.05	0.07	0.80	0.81	-	-	-
	(4)	Mean	0.29	0.71	-	1.00	1.00	0.99	0.00	0.43	0.01	0.22	0.47	1.50	1.48	-	-	-
		MSE$\times 10^2$	0.63	0.63	-	1.12	1.83	1.82	0.77	1.82	0.34	0.71	0.82	1.09	1.10	-	-	-
	(5)	Mean	0.20	0.29	0.51	1.03	0.00	-0.01	-	-	0.00	1.54	0.00	-	-	0.00	0.00	1.01
		MSE$\times 10^2$	0.33	0.59	0.53	1.92	0.02	0.13	-	-	0.00	1.85	0.00	-	-	0.10	0.15	0.72
	(6)	Mean	0.18	0.32	0.50	1.16	0.19	0.05	-	-	0.04	1.57	0.39	-	-	0.41	0.05	1.03
		MSE$\times 10^2$	2.11	2.08	2.84	10.26	16.61	2.16	-	-	1.37	6.85	4.35	-	-	3.15	2.13	2.22

cross-L_1 penalty work well in these scenarios. Dependent situations did not have a strong effect on parameter estimation, although a slightly larger MSE was observed in comparison with the independent situations. Especially in Scenario 6 for DS, we had moderate biases in estimation of the coefficients of the first linear predictor (β_{11} and β_{12}). This is because the other two groups had enough information for fitting the model to the data. In fact, the estimated risk scores between the estimated and true parameters were almost equal. This shows that the loss of model identifiability sometimes yields bias in parameter estimation for the overlapped situation; however, the predictive performance of the estimated score is sufficient.

Application

In this section, we show the results of application studies for the breast cancer dataset in order to evaluate the performance of the quasi-linear Cox model. To evaluate the predictive ability of the learned model, we calculated the Area under the curve (AUC) of time-dependent ROC [11] using test dataset. The predictive performance was compared between Cox's proportional hazard model and the quasi-linear Cox model. A dataset from [12] was used as the training data, and a dataset from [13] as the test data. These datasets include expression levels of 70 genes and survival time with some censors. Except for samples with missing values, there were 75 samples in the training dataset and 220 samples in the test dataset. In this application, we extracted the top 10 relevant genes to evaluate the model performance. Such marker preselection has been performed in many studies [14].

Since we had no prior knowledge for these genes, we applied the proposed model with cross-L_1 penalty introduced in "Quasi-linear Cox model with cross L_1 penalty" section. The number of groups K and the regularization parameter of the cross-L_1 penalty λ_c were determined using BIC from $K \in \{2, 3, 4, 5\}$ and $\lambda_c \in \{0, 0.1, 0.2, \cdots, 5.0\}$. All gene expressions are standardized to have mean zero and variance one among the training sample to compare the estimated coefficients. The same transformation was applied for the test sample.

As a result, a group size $K = 2$ was selected. The time series of test AUCs and the bootstrap 95% intervals for each year are shown in Figs. 2 and 3. For every time point, the test AUC of the quasi-linear relative risk model was larger than that of Cox's proportional hazard model. The estimated coefficients for linear and quasi-linear Cox's proportional hazard models are shown in Fig. 4. While the overall trend was not much different between linear and quasi-linear models, the cross-L_1 penalty gave contrast to the fitted quasi-linear model. Five out of ten genes have zero coefficient in the hazard function in the

Fig. 2 The time series changes in test AUC of the breast cancer dataset. Two line graphs show the AUC values at each time (days) calculated from time dependent ROC for the linear (red) and quasi-linear (blue) predictor

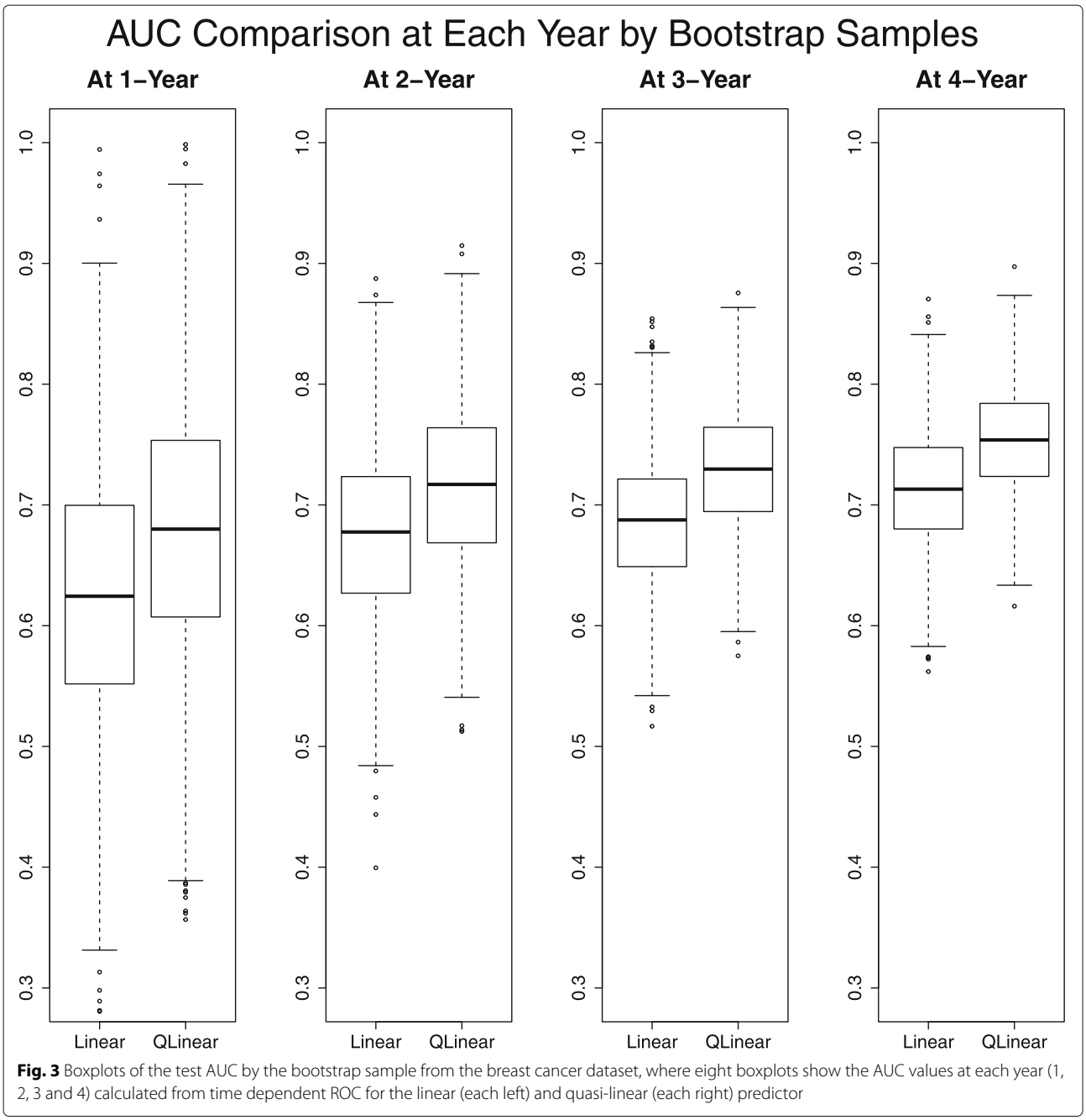

Fig. 3 Boxplots of the test AUC by the bootstrap sample from the breast cancer dataset, where eight boxplots show the AUC values at each year (1, 2, 3 and 4) calculated from time dependent ROC for the linear (each left) and quasi-linear (each right) predictor

first or second group. As a representative of them, we focus on the set of genes with the first ("gene6") and second largest ("gene4") coefficients in the first group. Those are called NUSAP1 (Nucleolar And Spindle Associated Protein 1) and TSPYL5 (Testis-Specific Y-Encoded-Like Protein 5), respectively. TSPYL5 is a well-known prognostic factor of poor outcome in breast cancer patients. Higher expression of TSPYL5 suppresses p53 protein levels causing damage to mammary cells. It may be important that those two genes have zero coefficient in the second group. Interestingly, it was reported that the TSPYL5 and NUSAP1 are biologically correlated with the same hallmarks of cancer: limitless replication potential [15]. While further investigation is needed, the proposed model thus suggests there are roughly two subpopulations which have different hazards.

Discussion

We developed a mixture hazard model, an extension of Cox's proportional hazards model, via a quasi-linear predictor. Theoretical discussion revealed that the maximum partial-likelihood estimator has properties of consistency

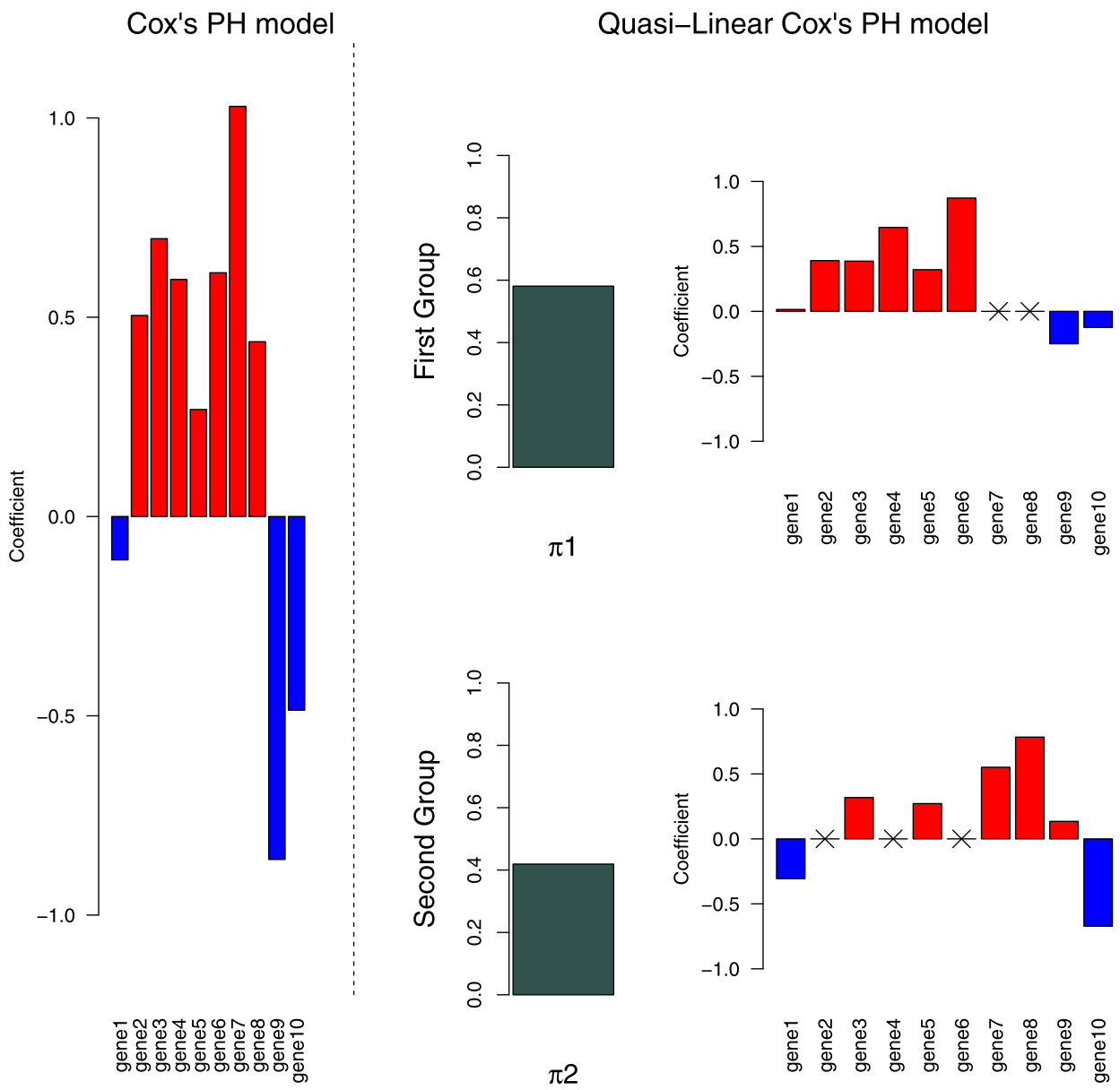

Fig. 4 The estimated coefficients for linear and quasi-linear Cox's proportional hazard models. The cross marks show the coefficients which equal exactly to zero

and asymptotic normality. Furthermore, we showed combining the cross-L_1 penalty makes the estimated model stable and interpretable. Empirical simulations and applications confirm these superior properties, and BIC have been shown to work well as a measure for selecting the number of groups and the tuning parameter of cross-L_1 penalty.

We will discuss the relationship between our study and previous work. First, the quasi-linear predictor was proposed in [16] to extend the logistic regression model for capturing heterogeneous structure in biomarkers. The quasi-linear logistic model was motivated by a Bayes risk-consistent predictor in binary classification between the mixture normal distribution and single normal distributions. The quasi-linear predictor enables us to model intrinsic heterogeneity using some linear predictors, and can be used as an extension from the standard to the heterogeneous setting of several models that rely on the linear predictor. Second, as introduced in the Introduction, several studies have proposed a mixture hazard model [1, 2], but these were limited to parametric ones. Our proposed method is one extension that is possible without

assuming a specific distribution. Also, several mixture distribution model proposals have been developed in past studies [17–19]. We note that the concepts of the *mixture hazards model* and *mixture density model* are completely distinct. In fact, while the mixture density model gives the simple weighted average of each survival function as the whole survival function $\tilde{S}(t) = \sum_{k=1}^{K} \tilde{\pi}_k \tilde{S}_k(t)$, the mixture hazard model gives the weighted average of log-survival function of each survival function as the whole log-survival function: $\log S(t) = \sum_{k=1}^{K} \pi_k \log S_k(t)$. In this context, the survival function is understood as the geometric mean of each survival function, $S(t) = \prod_{k=1}^{K} S_k^{\pi_k}$. We note that the density function is analogue to the probability while the hazards function is also, i.e. instantaneous rate of mortality. In this sense, both models can be regarded as the special case of latent variable models. It is not yet well understood what such a formal difference yields for the modeling in survival analysis, and it will be very important for our future work.

In this paper, we restricted the baseline hazard functions to be identical. There are three reasons for the restriction. First, it stabilizes estimation of model parameters. In addition to the high computational cost of the mixture model, it will be difficult to estimate the separate baseline hazard functions. Second, it improves the interpretability of the model. The assumption of different baseline hazards functions may seem a somewhat strange idea. This is because *baseline hazard function* refers to a hazard function when all observed covariate values are considered to be a reference for all patients. When a different baseline hazard function is required for each group, it means that there is heterogeneity that cannot be observed with the dataset in question. Instead, the quasi-linear Cox model enables us to model the intrinsic but observable heterogeneity. Third, regardless of such restrictions, the proposed model empirically had better predictive ability than the standard Cox's proportional hazards model. We thus achieved simultaneous modeling of group-wise proportional hazards models. On the other hand, although a stratified Cox model focuses on the heterogeneity for hazards in the population with different baseline hazards, it assumes the same relative risk function among groups. These two models thus have dualistic roles to capture hazard heterogeneity in the population. Finally, we note that in theory we should be able to loose these restriction for baseline hazard function in the proposed model, based on similar ideas for density mixture models proposed by [19].

Conclusions

In this paper, we focused on hazards mixture model. The quasi-linear Cox proportional hazards model was naturally derived by the quasi-linear predictor. It is essential to capture the intrinsic heterogeneity of patients for getting more stable and effective risk score. The proposed hazard model can capture such heterogeneity and achieve better performance than the ordinary linear Cox proportional hazards model.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12874-020-01063-2.

Additional file 1: Technical Derivations. In this file, we give proofs of Proposition 1, Proposition 2, Theorem 1 and Theorem 2.

Abbreviations

MM: Minorization-maximization; AUC: Area under curve; BID: Bayes information criteria; CLASSO: Cross least absolute shrinkage and selection operator; DS: Dependent setting; EM: Expectation–maximization; IS: Independent setting; MSE: Mean-squared error

Acknowledgements

The authors are thankful to the reviewers for their valuable suggestions to improve the article further.

Authors' contributions

KO and SE designed the methods of this article. KO carried out the simulation study and data analysis, and wrote the paper. Both authors have read and approved the final manuscript.

Funding

This work is partly supported by the Project Promoting Clinical Trials for Development of New Drugs (18lk020106Xt0003) from the Japan Agency for Medical Research and Development, AMED, and is supported by JSPS KAKENHI Grant Number 18H03211. The funding bodies had no role in the design and the conclusions of the presented study.

Availability of data and materials

All data used to perform the application described in this paper are freely available. The data of van't Veer et al. is available on the Gene Expression Omnibus data base [https://www.ncbi.nlm.nih.gov/geo/], series GSE2990. The data of Buyse et al. is available on the European Bioinformatics Institute ArrayExpress database [http://www.ebi.ac.uk/arrayexpress/], accession number E-TABM-77.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Clinical Biostatistics, Graduate School of Medicine, Kyoto University, Yoshida Konoe-cho, Kyoto, Japan. [2] The Institute of Statistical Mathematics, Midori-cho, Tachikawa, Tokyo, Japan.

Received: 2 September 2019 Revised: 22 June 2020
Published online: 06 July 2020

References

1. Louzada-Neto F, Mazucheli J, Achcar JA. Mixture hazard models for lifetime data. Biom J. 2002;44:3–14.
2. Hilton RP, Zheng Y, Serban N. Modeling heterogeneity in healthcare utilization using massive medical claims data. J Am Stat Assoc. 2018;113(521):111–21.
3. Fang HB, Li G, Sun J. Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. Scand J Stat. 2005;32(1):59–75.

4. Omae K, Komori O, Eguchi S. Quasi-linear score for capturing heterogeneous structure in biomarkers. BMC Bioinformatics. 2017;18(1):. https://doi.org/10.1186/s12859-017-1721-x.
5. Hunter DR, Lange K. A tutorial on mm algorithms. Am Stat. 2004;58(1): 30–7.
6. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc Ser B. 1977;39:1–38.
7. McLachlan GJ, Krishnan T. The EM Algorithm and Extensions, 2nd edn. In: Wiley series in probability and statistics. New Jersey: Wiley; 2008.
8. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–29.
9. Goeman JJ. $\mathbf{L_1}$ penalized estimation in the **C**ox proportional hazards model. Biom J. 2010;52:70–84.
10. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6:461–4.
11. Heagerty PJ, Lumley T, Pepe MS. Time-dependent roc curves for censored survival data and a diagnostic marker. Biometrics. 2000;56(2):337–44.
12. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6.
13. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas A, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst. 2006;98:1183–92.
14. Dettling M, Bühlman P. Boosting for tumor classification with gene expression data. Bioinformatics. 2003;19(9):1061–9. https://doi.org/10.1093/bioinformatics/btf867.
15. Tian S, P R, van't Veer LJ, Bernards R, De Snoo F, Glas AM. Biological functions of the genes in the mammaprint breast cancer profile reflect the hallmarks of cancer. Biomark Insights. 2010;5:6184.
16. Omae K, Komori O, Eguchi S. Reproducible detection of disease-associated markers from gene expression data. BMC Med Genomics. 2016;9(1):. https://doi.org/10.1186/s12920-016-0214-5.
17. Elmahdy EE, Aboutahoun AW. A new approach for parameter estimation of finite Weibull mixture distributions for reliability modeling. Appl Math Model. 2013;37:1800–10.
18. Zhang Q, Hua C, Xu G. A mixture Weibull proportional hazard model for mechanical system failure prediction utilising lifetime and monitoring data. Mech Syst Signal Process. 2014;43:103–12.
19. You N, He S, Wang X, Zhu J, Zhang H. Subtype classification and heterogeneous prognosis model construction in precision medicine. Biometrics. 2018;74(3):814–22. https://doi.org/10.1111/biom.12843.

Publisher's Note