Statistica Sinica **34** (2024), 439-458
doi:https://doi.org/10.5705/ss.202021.0050

UNBIASED BOOSTING ESTIMATION FOR CENSORED SURVIVAL DATA

Li-Pang Chen[1,2] and Grace Y. Yi*[1]

[1] *University of Western Ontario and* [2] *National Chengchi University*

Abstract: Boosting methods have been broadly discussed for various settings, and most methods handle data with complete observations. Although some methods are available for survival data with censored responses, they tend to assume a specific model for the survival process, and most provide numerical implementation procedures without rigorous theoretical justifications. In this paper, we develop an unbiased boosting estimation method for censored survival data, without assuming an explicit model, and explore three strategies for adjusting the loss functions, while accommodating censoring effects. We implement the proposed method using a functional gradient descent algorithm, and rigorously establish the theoretical results, including the consistency and optimization convergence. Our numerical studies show that the proposed method exhibits satisfactory performance in finite-sample settings.

Key words and phrases: Adjusted loss functions, boosting, consistency, empirical processes, machine learning, right-censoring, survival data.

1. Introduction

Boosting is a popular technique for deriving a strong learner from weak, yet simple learners by iteratively updating the learning results. Interest in boosting has increased since the publications of Schapire (1990) and Freund (1995), with numerous boosting algorithms developed for various settings; see Ridgeway (1999) for a summary of early boosting methods for regression and classification problems. Additional details can be found in Bühlmann and Hothorn (2007), Hastie, Tibshirani and Friedman (2008), and Schapire and Freund (2014).

Although most studies focus on settings with complete responses (e.g., Bühlmann and Yu (2003); Lugosi and Vayatis (2004); Zhang and Yu (2005)), boosting algorithms have also recently been used to analyze survival data, which typically involve censored responses, mainly under parametric or semiparametric survival models. For example, under the Cox proportional hazards model, Li and Luan (2005) present a gradient boosting procedure with cubic smoothing splines, Chen et al. (2013) derive a boosting algorithm using the concordance index, He et al. (2016) consider a component-wise gradient boosting procedure, and Bühlmann and Hothorn (2007) and Binder and Schumacher (2008) develop

*Corresponding author.

the R packages mboost and CoxBoost, respectively. Focusing on the accelerated failure time model, Schmid and Hothorn (2008) examine a boosted estimating procedure. Considering nonlinear transformation models, Lu and Li (2008) propose a gradient boosting method that uses the negative log marginal likelihood function as the loss function. Other boosting procedures related to survival data include those of Benner (2002), Mayr, Hofner and Schmid (2016), Lee, Chen and Ishwaran (2021), Bellot and van der Schaar (2018a), and Bellot and van der Schaar (2018b).

Although boosting methods under assumed survival models can be useful, they are vulnerable to model misspecification and their application scope is model-dependent, and thus restrictive. More notably, most existing boosting methods for survival data focus simply on implementation procedures whose feasibility are assessed using numerical studies. To the best of our knowledge, with the exception of the work of Lee, Chen and Ishwaran (2021), no existing studies establish theoretical results for boosting methods on survival data.

To fill this gap, we develop a boosting method under a general setup that does not impose specific models for survival data. Our research not only supplies a model-free boosting method for censored survival data, but also establishes its theoretical results. In particular, our research contributes to the literature in three ways. First, under the same framework, we examine three strategies for adjusting the usual loss functions to address the effects of right-censored responses, yielding three classes of adjusted loss functions, called *Buckley–James-type* (BJ) loss functions, *inverse-censoring probability weighted* (ICPW) loss functions, and *augmented inverse-censoring probability weighted* (AICPW) loss functions. These strategies enable us to handle survival data flexibly in order to incorporate different features. The BJ adjustment applies to settings with a good amount of information about the survival process, the ICPW scheme works when we have adequate knowledge about the censoring process, and the AICPW method allows either the survival or the censoring process to be misspecified, thus enjoying the *double robustness* property. Second, we use the functional gradient descent algorithm to devise a two-stage minimization procedure to derive the prediction function for the survival times. Finally, and importantly, we rigorously establish theoretical results for the proposed methods, including consistency and convergence.

Our work is similar to that of Hothorn et al. (2006), who consider only the ICPW scheme to adjust for censoring effects, without providing theoretical studies. In contrast, we present additional adjustment methods and provide their theoretical justifications. Wang and Wang (2010) use the BJ formulation (Buckley and James (1979)) to create a pseudo-response to account for censoring effects, using the L_2-norm loss function, but not providing any theoretical justification. Our BJ scheme adjusts for any loss functions, and we provide rigorous justifications for our results. Although Lee, Chen and Ishwaran (2021) establish theoretical results, their goal differs from ours. They estimate the hazard function nonparametrically by minimizing a scaled negative log likelihood function. In contrast, we focus on finding a prediction model for survival times by minimizing the risk function, which may assume various forms.

The remainder of the paper is organized as follows. In Section 2, we introduce our notation and framework. In Section 3, we consider censored survival data, and examine three schemes that accommodate censoring effects in loss functions. In Section 4, we implement the procedure for the proposed unbiased boosting estimation method for censored survival data. In Section 5, we rigorously establish our theoretical results to justify the validity of the proposed method. Numerical results for the real-data analysis and simulation studies are presented in Section 6 and Section S5 of the Supplementary Material, respectively. Section 7 concludes the paper; all technical details are reported in the Supplementary Material.

2. Notation and Framework

2.1. Survival data and objective

Let $T \geq 0$ represent the survival time of an individual, and let X denote the vector of associated p-dimensional covariates. To remove the positivity constraint on T, we consider a transformed outcome. In particular, let $\widetilde{T} = \log T$. We are interested in finding a function of X such that its value predicts $\widetilde{T}$ well. To this end, we consider a class of useful functions. Specifically, let $\mathcal{F}$ denote the convex set of real-valued functions from $\mathbb{R}^p$ to $\mathbb{R}$ satisfying Condition (C5) in Section S1.1 of the Supplementary Material. For $f \in \mathcal{F}$, let $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ denote the *loss function* of using $f(X)$ to predict $\widetilde{T}$, which is a convex and differentiable function in the second argument, as stated in Condition (C3) in Section S1.1 of the Supplementary Material.

Define the *risk function* as

$$R(f) = E\left[L\left\{\widetilde{T}, f(X)\right\}\right], \tag{2.1}$$

where, and hereafter, E represents the expectation with respect to the joint distribution of the random variables appearing in the loss function. By the convexity and differentiability of $L(\cdot,\cdot)$, the risk function (2.1) is convex with respect to $f \in \mathcal{F}$, as discussed by Zhang and Yu (2005), and is differentiable, provided we can interchange the expectation and the differentiation.

To find a function in $\mathcal{F}$ that predicts $\widetilde{T}$ well, we need to identify the element in $\mathcal{F}$ with

$$f_0 = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, R(f),$$

assuming the existence and uniqueness of $\min_{f\in\mathcal{F}} R(f)$, or equivalently,

$$R(f_0) = \min_{f\in\mathcal{F}} R(f). \tag{2.2}$$

Because the joint distribution of $\widetilde{T}$ and X is unknown, we use the sample information and the empirical average to replace the expectation in (2.1). That is, we estimate f_0 by finding $\widehat{f}_{\text{comp}} \in \mathcal{F}$, such that

$$\widehat{f}_{\text{comp}} = \underset{f\in\mathcal{F}}{\text{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} L\{\widetilde{T}_i, f(X_i)\} \right], \tag{2.3}$$

where we assume the availability of a random sample of independent observations $\mathcal{O}_{\text{comp}} \triangleq \{(\widetilde{T}_i, X_i) : i = 1,\ldots,n\}$, with n being the sample size and $(\widetilde{T}_i, X_i)$ denoting an independent copy of $(\widetilde{T}, X)$. For ease of exposition , throughout the article, we use uppercase letters, such as $\widetilde{T}_i$ and X_i, to represent both random variables and their realizations without differentiating them in notation.

2.2. Usual steepest descent algorithm

We can use a boosting method such as the steepest descent method (e.g., Hastie, Tibshirani and Friedman (2008, Sec. 10.10)) to solve (2.3). This method essentially uses the gradient of the loss function to enhance estimates of the function $f(\cdot)$ by iteratively using a varying learning rate that can be treated as a weak learner. Thus, we can use the method of Hastie, Tibshirani and Friedman (2008) by "parameterizing" the function $f(X)$ as $\{f(X_1),\ldots,f(X_n)\}$ for the n observations of the covariates X, and then define the partial derivative of the loss function $L\{\widetilde{T}_i, f(X_i)\}$ as

$$\partial L\{\widetilde{T}_i, f(X_i)\} \triangleq \frac{\partial L(u,v)}{\partial v}\Big|_{u=\widetilde{T}_i, v=f(X_i)}, \tag{2.4}$$

where $\partial L(u,v)/\partial v$ represents the partial derivative of the loss function $L(u,v)$ with respect to the second argument, while keeping the first argument fixed.

With the estimate of $f(\cdot)$ at iteration m, denoted as $f^{(m)}(\cdot)$, we employ the steepest descent method to enhance the estimation of $f(\cdot)$ at iteration $(m+1)$ by adding an increment term, $-\widehat{\alpha}_{m+1}\partial L\{\widetilde{T}_i, f^{(m)}(X_i)\}$, to $f^{(m)}(\cdot)$. Here, $\widehat{\alpha}_{m+1}$ is a scalar learning rate determined during the previous iteration:

$$\widehat{\alpha}_{m+1} = \underset{\alpha_{m+1}\in\mathbb{R}}{\text{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} L[\widetilde{T}_i, f^{(m)}(X_i) - \alpha_{m+1}\partial L\{\widetilde{T}_i, f^{(m)}(X_i)\}] \right). \tag{2.5}$$

That is, the estimate of $f(\cdot)$ at iteration $(m+1)$ is given by

$$f^{(m+1)}(X_i) = f^{(m)}(X_i) - \widehat{\alpha}_{m+1}\partial L\{\widetilde{T}_i, f^{(m)}(X_i)\}, \tag{2.6}$$

Table 1. Common loss functions.

Name	Loss function: $L(\widetilde{T}_i, f(X_i))$	Derivative: $\partial L(\widetilde{T}_i, f(X_i))$
L_2-norm	$\left\{\widetilde{T}_i - f(X_i)\right\}^2$	$-2\left\{\widetilde{T}_i - f(X_i)\right\}$
L_1-norm	$\left\|\widetilde{T}_i - f(X_i)\right\|$	$-\text{sign}\left\{\widetilde{T}_i - f(X_i)\right\}$ if $\left\|\widetilde{T}_i - f(X_i)\right\| \neq 0$
Huber	$\begin{cases} \frac{1}{2}\left\{\widetilde{T}_i - f(X_i)\right\}^2, & \text{if } \left\|\widetilde{T}_i - f(X_i)\right\| \leq \eta, \\ \eta\left(\left\|\widetilde{T}_i - f(X_i)\right\| - \frac{\eta}{2}\right), & \text{otherwise} \end{cases}$	$\begin{cases} -\left\{\widetilde{T}_i - f(X_i)\right\}, & \text{if } \left\|\widetilde{T}_i - f(X_i)\right\| < \eta, \\ -\eta \text{ sign}\left\{\widetilde{T}_i - f(X_i)\right\}, & \text{if } \left\|\widetilde{T}_i - f(X_i)\right\| > \eta \end{cases}$

and the final boosting estimator is $\widehat{f}_{\text{comp}}(X_i) \triangleq f^{(m+1)}(X_i)$, if the iteration stops at $(m+1)$.

2.3. Loss functions

Three loss functions are popular for continuous responses; see the middle column of Table 1 (Friedman (2001, p.1197)). Here, for the Huber loss function, η can be taken as the αth-quantile of $|\widetilde{T}_i - f(X_i)|$ for a constant α, with $0 < \alpha < 100$ and $i = 1, \ldots, n$ (Hastie, Tibshirani and Friedman (2008, p.349, p.360)).

Differentiating the loss function is crucial to implement (2.4). Although this is not a problem when using the L_2-norm, it is when we use the L_1-norm or Huber loss function, because they are not differentiable over the entire domain. In this case, we can modify $\partial L\{\widetilde{T}_i, f(X_i)\}$ using the *subdifferential*. For ease of exposition, we take $\partial L\{\widetilde{T}_i, f(X_i)\}$ as defined in (2.4). For the loss functions listed in Table 1, we present the values of $\partial L\{\widetilde{T}_i, f(X_i)\}$ in the last column of Table 1, where the constraint $|\widetilde{T}_i - f(X_i)| \neq a$ is included for the L_1-norm loss with $a = 0$ and the Huber loss with $a = \eta$. This requirement is not as restrictive as it appears, and holds in practical settings, because "$|\widetilde{T}_i - f(X_i)| = a$" occurs with zero probability.

3. Adjusting Loss Functions with Censoring Effects Accommodated

The discussion in Section 2 relies on the availability of complete observations, $\mathcal{O}_{\text{comp}}$, of a random sample. However, this is often not true for survival data, owing to the presence of censoring. For $i = 1, \ldots, n$, let C_i denote the censoring time for T_i, let $\Delta_i = \mathbb{I}(T_i \leq C_i)$ denote the censoring indicator, and write $Y_i = \min(T_i, C_i)$ and $\widetilde{Y}_i = \log Y_i$, where $\mathbb{I}(\cdot)$ is the indicator function. Let $[0, \tau]$ denote the study period with τ finite. Consistent with most discussions about survival analysis, we assume that T_i and C_i are conditionally independent, given X_i. Following Rubin and van der Laan (2007), Zhu and Kosorok (2012), and Steingrimsson et al. (2016), we further assume that C_i and X_i are independent. These assumptions are listed as Condition (C6) in Section S1.1 of the Supplementary Material.

In the presence of censoring, the survival time $\widetilde{T}_i$ is not available for every study subject, and, thus, we cannot use the estimation procedure in Section 2.2 directly. Consequently, we consider new loss functions expressed in terms of the observed censored data $\mathcal{O}_{\text{cd}} \triangleq \big\{(\widetilde{Y}_i, X_i, \Delta_i) : i = 1, \ldots, n\big\}$, while accounting for the censoring effects. The basic idea is to ensure that the expectation of a new loss function, denoted by $L^*\{\widetilde{Y}_i, f(X_i)\}$, recovers the expectation of the original loss function $L\{\widetilde{T}_i, f(X_i)\}$, expressed in terms of $\widetilde{T}_i$ and X_i; that is, $E[L^*\{\widetilde{Y}_i, f(X_i)\}] = E[L\{\widetilde{T}_i, f(X_i)\}]$. Thus, the minimizer of the expectation of the workable new loss function $E[L^*\{\widetilde{Y}_i, f(X_i)\}]$ also minimizes the risk function $R(f)$ in (2.1), as if $\widetilde{T}_i$ were always observed.

3.1. Adjustment strategies

In this subsection, we describe three strategies for constructing adjusted loss functions. Let $F_{T0}(y|X_i) = P(T_i > y|X_i)$ represent the true conditional survival function of T_i, given X_i, and let $F_T(y|X_i)$ denote a working function used to model $F_{T0}(y|X_i)$, with $f_T(t|X_i)$ denoting the corresponding conditional density of T_i. Let $G_0(c) = P(C_i > c)$ stand for the true survivor function of C_i, and let $G(c)$ denote its working function.

The first adjusted loss function is motivated by the BJ formulation (Buckley and James (1979)):

$$L_{\text{BJ}}\{\widetilde{Y}_i, f(X_i)\} = \Delta_i L\{\widetilde{Y}_i, f(X_i)\} + (1 - \Delta_i)\Psi(Y_i, X_i), \tag{3.1}$$

where $\Psi(y, X_i) = E[L\{\widetilde{T}_i, f(X_i)\}|T_i > y, X_i]$, determined by

$$\Psi(y, X_i) = \int_y^\infty \frac{L\{t, f(X_i)\} f_T(t|X_i)}{F_T(y|X_i)} dt. \tag{3.2}$$

The conditional expectation $\Psi(Y_i, X_i)$ in (3.1) facilitates the contribution from the censored subjects, but requires a working model $F_T(y|X_i)$ for the survival process.

On the other hand, we may wish to use information from the uncensored subjects only, in the hope of not needing $F_T(y|X_i)$. Hothorn et al. (2006) considered an adjusted loss function based on the ICPW scheme:

$$L_{\text{ICPW}}\{\widetilde{Y}_i, f(X_i)\} = \frac{\Delta_i L\{\widetilde{Y}_i, f(X_i)\}}{G(Y_i)}. \tag{3.3}$$

Although (3.3) frees us from using $F_T(y|X_i)$ in the formulation, as in (3.1), it calls for a working model $G(c)$ of the censoring process. With the different involvements of $F_T(y|X_i)$ and $G(c)$ in (3.1) and (3.3), what happens if we use both $F_T(y|X_i)$ and $G(c)$ to adjust the original loss function $L\{\widetilde{T}_i, f(X_i)\}$? Motivated by Rubin and van der Laan (2007), we further consider the AICPW loss function

$$L_{\text{AICPW}}\{\widetilde{Y}_i, f(X_i)\} = L_{\text{ICPW}}\{\widetilde{Y}_i, f(X_i)\} + \Gamma(Y_i, X_i, \Delta_i), \tag{3.4}$$

with

$$\Gamma(y, X_i, \Delta_i) = \frac{(1-\Delta_i)}{G(y)}\Psi(y, X_i) - \int_0^y \frac{\Psi(t, X_i)}{G^2(t)} dG(t).$$

Although (3.4) might seem more restrictive than (3.1) and (3.3) because both $F_T(y|X_i)$ and $G(c)$ are involved, it does have an advantage, as discussed in Section 3.2. Clearly, (3.3) uses information from subjects in the sample who are not censored, ignoring the partial information of subjects who are censored. Adding the term $\Gamma(Y_i, X_i, \Delta_i)$ to (3.3) enables us to use the measurements from censored subjects, possibly enhancing the efficiency. Because $F_T(y|X_i)$ and $G(c)$ appear in the denominators in the preceding formulations, they are assumed to be greater than zero (almost surely), as stated in Condition (C7) in Section S1.1 of the Supplementary Material. Note that Steingrimsson et al. (2016) use an empirical version similar to (3.4) to construct survival trees.

3.2. Properties of the proposed loss functions

The three adjusted loss functions in Section 3.1 are formulated from different perspectives to accommodate censoring effects. Their validity is justified in the following two propositions.

Proposition 1. *The proposed adjusted loss functions* (3.1) *and* (3.3) *have the same expectation as* $L\{\widetilde{T}_i, f(X_i)\}$. *That is,*

(a) $E[L_{ICPW}\{\widetilde{Y}_i, f(X_i)\}] = E[L\{\widetilde{T}_i, f(X_i)\}]$;

(b) $E[L_{BJ}\{\widetilde{Y}_i, f(X_i)\}] = E[L\{\widetilde{T}_i, f(X_i)\}]$,

where the expectations are evaluated with respect to the joint distribution of the associated random variables under the working models.

The proof of this proposition is provided in Section S4.1 of the Supplementary Material. Proposition 1 states that the expectation of the two adjusted loss functions, $L_{\text{ICPW}}(\cdot,\cdot)$ and $L_{\text{BJ}}(\cdot,\cdot)$, recovers the risk function (2.1) based on the transformed failure time $\widetilde{T}_i$. With $L(\cdot,\cdot)$ taken as the L_2-norm loss function, Bühlmann and Hothorn (2007) establish the identity in Proposition 1(a).

The formulation of $L_{\text{AICPW}}(\cdot,\cdot)$ involves the distributions of both the survival and the censoring processes, which, at first sight, appears more restrictive than either $L_{\text{BJ}}(\cdot,\cdot)$ or $L_{\text{ICPW}}(\cdot,\cdot)$. However, the following proposition ensures that $L_{\text{AICPW}}(\cdot,\cdot)$ is more flexible than $L_{\text{BJ}}(\cdot,\cdot)$ or $L_{\text{ICPW}}(\cdot,\cdot)$. As long as $F_T(y|X_i)$ or $G(c)$ is specified correctly (even if we do not know which one), $L_{\text{AICPW}}(\cdot,\cdot)$ has the expectation identical to that of the initial loss function $L(\cdot,\cdot)$.

Proposition 2 (Double Robustness). *Let $L_{AICPW,0}\{\widetilde{Y}_i, f(X_i)\}$ be determined by (3.4), with $G(c)$ and $F_T(y|X_i)$ replaced by $G_0(c)$ and $F_{T0}(y|X_i)$, respectively. Then,*

(a) the expectation of the loss function $L_{AICPW,0}\{\widetilde{Y}_i, f(X_i)\}$ is given by

$$\begin{aligned} &E\left[L_{AICPW,0}\{\widetilde{Y}_i, f(X_i)\}\right] \\ &= E\left[\frac{G(T_i)}{G_0(T_i)} L\{\widetilde{T}_i, f(X_i)\}\right] \\ &\quad - E\left\{L\{\widetilde{T}_i, f(X_i)\} \times \left(\int_0^{T_i} \frac{F_T(t|X_i)}{F_{T0}(t|X_i)} \left[\frac{d}{dt}\left\{\frac{G(t)}{G_0(t)}\right\}\right] dt\right)\right\}; \end{aligned}$$

(b) if either $F_T(y|X_i) = F_{T0}(y|X_i)$ or $G(c) = G_0(c)$, then we have

$$E\left[L_{AICPW,0}\{\widetilde{Y}_i, f(X_i)\}\right] = E\left[L\{\widetilde{T}_i, f(X_i)\}\right], \tag{3.5}$$

where the expectations are evaluated with respect to the working models for the associated random variables.

The proof of this proposition is deferred to Section S4.2 of the Supplementary Material. Proposition 2(b) resembles the property of doubly robust estimators in regression analysis (e.g., Rubin and van der Laan (2007, Thm. 1)), which requires that only one of the two models be specified correctly.

For ease of referral, we let $L^*(\cdot,\cdot)$ denote the loss function defined by (3.1), (3.3), or (3.4). By (2.1), Propositions 1 and 2 indicate $R(f) = E[L^*\{\widetilde{Y}_i, f(X_i)\}]$. Consequently, we now modify (2.3), with the complete observations $\mathcal{O}_{\text{comp}}$ of a random sample replaced by the available censored data $\mathcal{O}_{\text{cd}}$. That is, we want to find $\widehat{f}_{\text{cd}} \in \mathcal{F}$, such that

$$\widehat{f}_{\text{cd}} = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n} L^*\{\widetilde{Y}_i, f(X_i)\}\right]. \tag{3.6}$$

However, the minimization problem in (3.6) cannot proceed, because of the involvement of the adjusted loss function $L^*(\cdot,\cdot)$ with unknown (conditional) survivor functions $F_T(y|X_i)$ and/or $G(c)$. To circumvent this difficulty, we approximate $L^*(\cdot,\cdot)$, denoted by $\widehat{L}^*(\cdot,\cdot)$, by replacing $F_T(y|X_i)$ and $G(c)$ with their consistent estimators, denoted by $\widehat{F}_T(y|X_i)$ and $\widehat{G}(c)$, respectively; the construction of $\widehat{F}_T(y|X_i)$ and $\widehat{G}(c)$ is deferred to Section 4.2. Our goal is to apply the observed censored data $\mathcal{O}_{\text{cd}}$ to find the solution to the following minimization problem:

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n} \widehat{L}^*\{\widetilde{Y}_i, f(X_i)\}\right]. \tag{3.7}$$

4. Unbiased Boosting Estimation with Censored Data

4.1. Boosting estimation procedure

Propositions 1 and 2(b) state that an adjusted loss function constructed from censored data has the same expectation as that of the initial loss function $L\{\widetilde{T}_i, f(X_i)\}$. Thus, on average, the risk function induced from an adjusted loss function is identical to that derived from the original failure time $\widetilde{T}_i$, as if $\widetilde{T}_i$ were available for all $i \in \{1, \ldots, n\}$. In this sense, we regard the procedure described in this subsection as an *unbiased boosting estimation for censored (UBEC) data.*

Now, we implement the procedure for the minimization problem (3.7). With a given form of $\widehat{L}^*(\cdot, \cdot)$, it may be tempting to use (2.6) by replacing $\partial L\{\widetilde{T}_i, f(X_i)\}$ with $\partial \widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i)\} \triangleq \partial \widehat{L}^*(u, v)/\partial v|_{u=\widetilde{Y}_i, v=f^{(m)}(X_i)}$ (assuming its existence). However, as discussed by Schapire and Freund (2014), using the gradient descent update directly may lead to an entirely unconstrained new update $f^{(m+1)}(\cdot)$. A way to overcome this problem is to impose some constraints to ensure each updated estimate of $f(\cdot)$ to be contained in a class of functions.

Let $\mathcal{C}$ represent a certain class of continuous functions mapping $\mathbb{R}^p$ to $\mathbb{R}$ that are uniformly bounded over any finite domain. Instead of using the increment $-\widehat{\alpha}_{m+1}\partial L\{\widetilde{T}_i, f^{(m)}(X_i)\}$ in (2.6), with $\partial L\{\widetilde{T}_i, f^{(m)}(X_i)\}$ replaced by $\partial \widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i)\}$, to update the estimation of $f(\cdot)$, we take the increment to be $\widehat{\alpha}_{m+1}h_{m+1}(X_i)$, with $h_{m+1}(\cdot)$ taken from $\mathcal{C}$, and update the estimate of $f(\cdot)$ at iteration $(m+1)$ using a modified version of (2.6):

$$f^{(m+1)}(X_i) = f^{(m)}(X_i) + \widehat{\alpha}_{m+1}\widehat{h}_{m+1}(X_i), \tag{4.1}$$

where $\widehat{\alpha}_{m+1}$ and $\widehat{h}_{m+1}(\cdot)$ are determined by

$$\left(\widehat{\alpha}_{m+1}, \widehat{h}_{m+1}\right) = \underset{\substack{\alpha_{m+1} \in \mathbb{R} \\ h_{m+1} \in \mathcal{C}}}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n} \widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i) + \alpha_{m+1}h_{m+1}(X_i)\}\right]. \tag{4.2}$$

Although (4.2) involves determining two unknown components, α_{m+1} and $h_{m+1}(\cdot)$, rather than one unknown parameter α_{m+1}, as in (2.6), it ensures that the estimates of $f(\cdot)$ at each iteration are bounded and, thus, that the final estimate falls in the class $\mathcal{F}$.

To find the minimizer of (4.2) at iteration $(m+1)$, we use two iterative steps to find $\widehat{\alpha}_{m+1}$ and $\widehat{h}_{m+1}(\cdot)$ separately, rather than jointly. First, treating $\widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i) + \alpha_{m+1}h_{m+1}(X_i)\}$ as a function of the argument $h_{m+1}(X_i)$, with other quantities fixed, we apply the first-order Taylor series expansion around $h_{m+1}(X_i) = 0$:

Algorithm 1: Functional Gradient Descent Algorithm.

```
Let $f^{(0)} \in \mathcal{F}$ denote the initial value and set $\zeta = n^{-\varpi}$ for a given $\varpi \geq 1$ ;
for step $m$ with $m = 0, 1, 2, \ldots$ do
    (a) calculate $\partial \widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i)\}$ for $i = 1, \ldots, n$;
    (b) find $\widehat{h}_{m+1}$ by solving (4.4);
    (c) solve the minimization problem (4.5) and obtain $\widehat{\alpha}_{m+1}$ ;
    (d) update $f^{(m)}(X_i)$ using (4.1) and denote the resultant estimate as
        $f^{(m+1)}(X_i)$ ;
    if
        $\left| \frac{1}{n} \sum_{i=1}^{n} \widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i)\} - \frac{1}{n} \sum_{i=1}^{n} \widehat{L}^*\{\widetilde{Y}_i, f^{(m+1)}(X_i)\} \right| \leq \zeta$    (4.6)
    then
        Stop iteration and let
            $\widehat{f}_n(\cdot) \leftarrow f^{(\widetilde{m})}(\cdot)$    (4.7)
        be the final estimator, where $\widetilde{m}$ represents the iteration number $m$ at
        the stopping step such that (4.6) is met for $\zeta$.
    end
end
```

$$\widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i) + \alpha_{m+1} h_{m+1}(X_i)\} \approx \widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i)\} + \left[\partial \widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i)\} \times h_{m+1}(X_i)\right] \alpha_{m+1}. \tag{4.3}$$

Because the first term in (4.3) is free of $h_{m+1}(\cdot)$ and α_{m+1} is fixed, the minimizer of $h_{m+1}(\cdot)$, denoted by $\widehat{h}_{m+1}(\cdot)$, is determined by

$$\widehat{h}_{m+1} = \underset{h_{m+1} \in \mathcal{C}}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} \left[\partial \widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i)\} \times h_{m+1}(X_i) \right] \right). \tag{4.4}$$

Next, replacing $h_{m+1}(\cdot)$ in (4.2) with $\widehat{h}_{m+1}(\cdot)$ gives

$$\widehat{\alpha}_{m+1} = \underset{\alpha_{m+1} \in \mathbb{R}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \widehat{L}^*\{\widetilde{Y}_i, f^{(m)}(X_i) + \alpha_{m+1} \widehat{h}_{m+1}(X_i)\} \right]. \tag{4.5}$$

Consequently, $\widehat{\alpha}_{m+1}$ and $\widehat{h}_{m+1}(\cdot)$ can be used to update $f^{(m)}(\cdot)$ and produce $f^{(m+1)}(\cdot)$ using (4.1). This strategy is also called the *functional gradient descent algorithm* (e.g., Boyd and Vandenberghe (2004, p.475); Schapire and Freund (2014, p.190)). The pseudocode for the implementation is presented in Algorithm 1.

Algorithm 1 differs from the *greedy algorithm* of Zhang and Yu (2005) and Schapire and Freund (2014), which obtains the minimizers of α_{m+1} and h_{m+1} simultaneously. When implementing Algorithm 1, we need to set a *stopping* criterion. To highlight the feature of censored responses, we examine the difference of $\widehat{L}^*(\cdot,\cdot)$ evaluated at two successive estimates, $f^{(m+1)}(\cdot)$ and $f^{(m)}(\cdot)$, using the squared error (L_2-norm), absolute error (L_1-norm), or Huber error as a stopping criterion, and compare it with a prespecified threshold value; an example of using the L_1-norm is shown in (4.6). Alternatively, we can determine the stopping step by comparing the values of the empirical risk function against the number of iterations.

For the data $\mathcal{O}_{\mathrm{cd}}$ with size n, described in Section 3, let $\widetilde{m}$ represent the iteration number at the stopping step, such that (4.6) is met for a prespecified positive value ζ, and let $\widehat{f}_n(\cdot)$ denote the resultant estimator for the solution (3.7) determined by (4.7). The value of ζ determines when to stop the iterations. Taking $\zeta = n^{-\varpi}$, with $\varpi \geq 1$, gives us a convenient way of discussing the *asymptotic* behavior of $\widehat{f}_n(\cdot)$, as shown in Section S4.4. In applications with a finite sample, one may often set ζ as a small value, such as 10^{-6}, regardless of the value of n.

4.2. Implementation remarks

The boosting procedure described in Section 4.1 hinges on the specification of the class $\mathcal{C}$, and on the use of a consistent estimator of $F_T(y|X_i)$ or $G(c)$.

Similarly to Li and Luan (2005) and Bühlmann and Yu (2003), we employ the cubic spline method to characterize the functions in $\mathcal{C}$. Specifically, any function $h(\cdot)$ in $\mathcal{C}$ is assumed to take the additive form

$$h(X_i) = h_1(X_{i1}) + \cdots + h_p(X_{ip}),$$

with $X_i = \big(X_{i1}, \ldots, X_{ip}\big)^\top$, and each $h_j(X_{ij})$ expressed as an M-order spline with J knots, where M and J are positive integers. That is, using the truncated power basis functions $\big\{1, X_{ij}, X_{ij}^2, \ldots, X_{ij}^{M-1}, (X_{ij}-\rho_{j1})_+^{M-1}, \ldots, (X_{ij}-\rho_{jJ})_+^{M-1}\big\}$ with knots $\rho_{j1}, \ldots, \rho_{jJ}$, we write

$$h_j(X_{ij}) = \sum_{r=0}^{M-1} \beta_{jr} X_{ij}^r + \sum_{k=1}^{J} \gamma_{jk} (X_{ij} - \rho_{jk})_+^{M-1},$$

where β_{jr} for $r = 0, 1, \ldots, M-1$ and γ_{jk} for $k = 1, \ldots, J$ are unknown parameters, and $a_+ \triangleq \max(0, a)$ for a constant a. In practice, we often use a cubic spline with M set as 4, where J may be set as 2 (Hastie, Tibshirani and Friedman (2008, p.143)).

To estimate $F_T(y|X_i)$, we can use strategies based on parametric or semiparametric regression models (e.g., Lawless (2003)). Such methods are straightforward to implement, but a major drawback is the sensitivity of the results to the model assumptions. Alternatively, one may invoke the kernel conditional Kaplan–Meier estimator (e.g., Dabrowska (1989)) to consistently estimate $F_T(y|X_i)$. However, this approach requires a proper specification of the bandwidth, and may suffer from the curse of dimensionality when the dimension p of the covariates is large (e.g., Geenens (2011)).

To provide a consistent, yet robust estimation of $F_T(y|X_i)$, we can use a random survival forest (RSF) to estimate $F_T(y|X_i)$. The estimation procedure is outlined as follows. First, set a positive integer D (e.g., $D = 1000$), and draw D independent bootstrap samples from the initial sample data $\mathcal{O}_{\mathrm{cd}}$, denoted as $\mathcal{S}_1, \ldots, \mathcal{S}_D$. For each bootstrap sample $\mathcal{S}_d$, with $d = 1, \ldots, D$, build a binary survival tree using recursive random splitting rules and the procedures of, for instance, Cui et al. (2022); let $\{\mathcal{A}_{ud} : u \in \mathcal{U}_d\}$ denote the collection of the resulting terminal nodes, with $\mathcal{U}_d$ denoting a set of indices based on the dth bootstrap sample. Then, the Nelson–Aalen estimator for the cumulative baseline hazard function based on a terminal node $\mathcal{A}_{ud}$ is given by

$$\widehat{\Lambda}_{\mathcal{A}_{ud}}(t) = \sum_{u \leq t} \left\{ \frac{\sum_{i=1}^n \mathbb{I}(\Delta_i = 1)\mathbb{I}(Y_i = u)\mathbb{I}(X_i \in \mathcal{A}_{ud})}{\sum_{i=1}^n \mathbb{I}(Y_i \geq u)\mathbb{I}(X_i \in \mathcal{A}_{ud})} \right\} \quad \text{for } t > 0,$$

and the conditional cumulative baseline hazards function, given X_i, is thus given by

$$\widehat{\Lambda}_d(t|X_i) = \sum_{u \in \mathcal{U}_d} \mathbb{I}(X_i \in \mathcal{A}_{ud}) \widehat{\Lambda}_{\mathcal{A}_{ud}}(t).$$

Finally, the RSF estimate of $\Lambda(t|X_i)$ is given by $\widehat{\Lambda}(t|X_i) = (1/D)\sum_{d=1}^D \widehat{\Lambda}_d(t|X_i)$, leading to an estimate for $F_T(t|X_i)$ to be $\widehat{F}_T(t|X_i) = \exp\{-\widehat{\Lambda}(t|X_i)\}$. This estimator of $F_T(t|X_i)$ is shown to be consistent under regularity conditions (e.g., Ishwaran and Kogalur (2010); Cui et al. (2022)).

Finally, using the Kaplan–Meier estimator (e.g., Lawless (2003)), we estimate $G(c)$ by pooling the study subjects and ignoring their differences in covariates:

$$\widehat{G}(c) = \prod_{i:Y_i \leq c} \left(1 - \frac{1}{\#\{j : Y_j \geq Y_i\}}\right)^{1-\Delta_i} \quad \text{for } c > 0,$$

which is shown to be consistent (e.g., Wang (1987)), and where $\#\{j : Y_j \geq Y_i\}$ represents the count of index j satisfying $Y_j \geq Y_i$.

4.3. Finite-sample performance

To evaluate the finite-sample prediction performance of the proposed method, we use the *integrated Brier score* (IBS) (Graf et al. (1999)) as other studies have done (e.g., Benner (2002); Zhu and Kosorok (2012)). Let $F_T(t) = P(T_i \geq t)$ represent the unconditional survivor function for the survival time T_i. We want to assess the performance of applying the estimator (4.7) to predict $F_T(t)$ using the censored sample $\mathcal{O}_{\rm cd}$. To this end, we divide the original observed data $\mathcal{O}_{\rm cd}$ into training data and validation data, so that the censoring proportions in both are comparable, and let $\mathcal{T}$ and $\mathcal{V}$ denote the respective sets of subject indices.

First, we use the training data in $\mathcal{T}$ to estimate $F_T(t)$, using the procedure described in Section 4.1 to determine $\widehat{f}_n(\cdot)$ in (4.7), with $F_T(y|X_i)$ and $G(c)$ in $L^*(\cdot,\cdot)$ estimated using the RSF estimator and $\widehat{G}(c)$, respectively. Then, using $\widehat{f}_n(\cdot)$, we take

$$\widehat{F}_T(t) = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} \mathbb{I}\left\{\widehat{f}_n^*(X_i) \geq t\right\} \tag{4.8}$$

as an estimate of $F_T(t)$, where $\widehat{f}_n^*(x) = \exp\{\widehat{f}_n(x)\}$.

Next, we use the validation data in $\mathcal{V}$ by separating the measurements by censoring status, and then calculate the empirical version of the expected value for the squared difference between $\mathbb{I}(T_i > t)$ and $F_T(t)$, while accounting for the censoring effects. That is, for any $t > 0$, the Brier score (Benner (2002)) for the validation data in $\mathcal{V}$ is defined as

$$\begin{aligned} BS(t) = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \Bigg[& \left\{0 - \widehat{F}_T(t)\right\}^2 \mathbb{I}\left(Y_i \leq t, \Delta_i = 1\right) \frac{1}{\widehat{G}(Y_i)} \\ & + \left\{1 - \widehat{F}_T(t)\right\}^2 \mathbb{I}\left(Y_i > t\right) \frac{1}{\widehat{G}(t)} \Bigg], \end{aligned} \tag{4.9}$$

where $\widehat{G}(t)$ and $\widehat{F}_T(t)$ are obtained from the training data. Then, the IBS for the validation data is defined as

$$\text{IBS} = \{y_{\max}\}^{-1} \int_0^{y_{\max}} BS(t)dt, \tag{4.10}$$

where $y_{\max} = \max\{Y_i : i \in \mathcal{V}\}$.

To alleviate the effects of dividing the original data into training and validation data sets, we use the K-fold cross-validation procedure, with a positive integer K such as 5. Specifically, we split the original data $\mathcal{O}_{\rm cd}$ into K roughly equal-sized subsets, so that the censoring proportion in each subset is similar. For $k = 1, \ldots, K$, take the kth subset as validation data, and let the remaining $(K - 1)$ pooled subsets be training data; let $\mathcal{V}_k$ and $\mathcal{T}_k$ represent the class of subject indices for the kth validation and training data sets, respectively.

For $k = 1, \ldots, K$, we apply the preceding steps to the training data in $\mathcal{T}_k$ and the validation data in $\mathcal{V}_k$ to calculate the Brier score at each Y_i, with $i \in \mathcal{V}_k$, using (4.9), with $\mathcal{V}$ replaced by $\mathcal{V}_k$. Then, we approximate the IBS (4.10) for the kth validation data by

$$\mathrm{IBS}_k = \frac{1}{y_{\max,k}} \sum_{i \in \mathcal{V}_k} BS(Y_i),$$

where $y_{\max,k} = \max\{Y_i : i \in \mathcal{V}_k\}$. Consequently, the IBS for the original sample data $\mathcal{O}_{\mathrm{cd}}$ is approximated by the average of the K-fold cross-validation estimates of IBS:

$$\mathrm{IBS}_{\mathrm{CV}} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{IBS}_k. \tag{4.11}$$

5. Theoretical Results

In this section, we develop theoretical results for the proposed method, including the convergence of Algorithm 1 and the consistency of the estimator $\widehat{f}_n(\cdot)$ defined by (4.7).

5.1. Convergence of the algorithm

Theorem 1. *Assume that the regularity conditions* (C1)–(C5) *in Section* S1.1 *of the Supplementary Material hold. Suppose we are given data of a random sample* $\mathcal{O}_{cd}$ *with the given size* n. *For any initial function* $f^{(0)} \in \mathcal{F}$, *let* $f^{(m+1)}$ *denote the updated estimate of the function at step* $(m+1)$ *of Algorithm* 1. *Then,*

$$\lim_{m \to \infty} R(f^{(m+1)}) = R(f_0). \tag{5.1}$$

This theorem states that with data $\mathcal{O}_{\mathrm{cd}}$ given, iterating Algorithm 1 yields convergence as the iteration number approaches infinity.

Although Theorem 1 ensures the convergence of the iterations in Algorithm 1, in practice, we prefer a small number $\widetilde{m}$ to stop the iterations in order to avoid overfitting (e.g., Jiang (2004); Zhang and Yu (2005)). In Section S4.3 of the Supplementary Material, we show that for any given nonnegative integer m, there exist positive constants b^* and B^*, with $b^* < B^*$, such that

$$0 \le R(f^{(m+1)}) - R(f_0) \le \left(1 - \frac{b^*}{B^*}\right)^m \left\{R(f^{(0)}) - R(f_0)\right\}, \tag{5.2}$$

where f_0 is the true function satisfying (2.2).

Let $\widetilde{m}$ be a positive integer denoting the number of iterations required to stop Algorithm 1. Then, we have

$$\begin{aligned}\left|R(f^{(\widetilde{m})}) - R(f^{(\widetilde{m}+1)})\right| &\le \left|R(f^{(\widetilde{m})}) - R(f_0)\right| + \left|R(f^{(\widetilde{m}+1)}) - R(f_0)\right| \\ &\le \left(1 - \frac{b^*}{B^*}\right)^{\widetilde{m}-1} \left\{R(f^{(0)}) - R(f_0)\right\} \\ &\quad + \left(1 - \frac{b^*}{B^*}\right)^{\widetilde{m}} \left\{R(f^{(0)}) - R(f_0)\right\}\end{aligned}$$

$$= \left(1 - \frac{b^*}{B^*}\right)^{\widetilde{m}-1} \left\{R(f^{(0)}) - R(f_0)\right\} \left(2 - \frac{b^*}{B^*}\right), \quad (5.3)$$

where the second inequality comes from (5.2).

We wish to stop the iteration at $\widetilde{m}$ if the difference (in absolute value) of its following two iterations is smaller than a given threshold, say, $\zeta > 0$. By (5.3), we suggest stopping the iteration if

$$\left(1 - \frac{b^*}{B^*}\right)^{\widetilde{m}-1} \left\{R(f^{(0)}) - R(f_0)\right\} \left(2 - \frac{b^*}{B^*}\right) < \zeta.$$

That is,

$$\widetilde{m} < 1 + \frac{\log\left[\zeta/\{\left|R(f^{(0)}) - R(f_0)\right|(2 - b^*/B^*)\}\right]}{\log(1 - b^*/B^*)}, \quad (5.4)$$

suggesting that the number of iterations $\widetilde{m}$ is upper bounded by a value depending on b^*, B^*, and ζ, as well as on the initial value $f^{(0)}$.

Note that the upper bound of (5.4) involves the initial value $f^{(0)}$. Although (5.1) holds regardless of the initial values, a better choice of $f^{(0)}$ makes $\left|R(f^{(0)}) - R(f_0)\right|$ and the right-hand side of (5.4) small. In this case, a smaller number of iterations $\widetilde{m}$ can achieve the required accuracy.

Finally, another concern is whether the sample size n affects the convergence (5.1). As discussed in Section 2.2, based on n observations, the function f is "parametrized" and characterized as $\{f(X_1), \ldots, f(X_n)\}$. From Algorithm 1, the updated value at X_i in step $(m+1)$, $f^{(m+1)}(X_i)$, is determined by $\widehat{h}_{m+1}(X_i)$ and $\widehat{\alpha}_{m+1}$, the optimization of which depends on the sample size n. In addition, a larger sample size n may enable the updated value $f^{(m+1)}(X_i)$ to be more precise.

5.2. Consistency and boundness

In this subsection, we examine the asymptotic behavior of $\widehat{f}_n$. For $f \in \mathcal{F}$ and $\widehat{L}^*(\cdot, \cdot)$ defined as in (3.7), define

$$\widehat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \widehat{L}^*\{\widetilde{Y}_i, f(X_i)\}. \quad (5.5)$$

Theorem 2. *Assume that the regularity conditions* (C1)–(C5) *in Section* S1.1 *of the Supplementary Material hold. Furthermore, suppose that Algorithm* 1 *is run for a sequence of random samples $\mathcal{O}_{cd}$, with varying sizes n, and let $\widehat{f}_n$ denote the resultant estimator, defined in* (4.7) *at the stopping time $\widetilde{m}$, satisfying* (4.6). *Then, for any $\epsilon > 0$,*

$$P(\|\widehat{f}_n - f_0\|_\infty \le \epsilon) \to 1 \text{ as } n \to \infty,$$

where $\|\widehat{f}_n - f_0\|_\infty = \sup_{X_i:\ i=1,\ldots,n} \left|\widehat{f}_n(X_i) - f_0(X_i)\right|$ is the L_∞-norm of $\widehat{f}_n - f_0$

evaluated over $\{X_i : i = 1, \ldots, n\}$*.*

Theorem 2 shows the limiting behavior of a sequence of the proposed estimators $\widehat{f}_n$ obtained by applying the same method (i.e., Algorithm 1) to a sequence of data $\mathcal{O}_{\text{cd}}$ with varying sample sizes n. The result shows that the difference between the estimator $\widehat{f}_n$ and its target f_0, expressed in the L_∞-norm, converges in probability to zero, suggesting the consistency of $\widehat{f}_n$ in this sense. The proof of Theorem 2 is provided in Section S4.4 of the Supplementary Material.

Next, we examine an informative lower bound of $\widehat{f}_n - f_0$ in the infinity norm.

Theorem 3. *Under the regularity conditions and the setup in Theorem* 2*, there exist positive constants* s *and* $\alpha > 1/2$*, such that*

$$\left\|\widehat{f}_n - f_0\right\|_\infty \geq sn^{-\alpha/(2\alpha+1)},$$

for any sample size n*.*

This lower bound of $\|\widehat{f}_n - f_0\|_\infty$ is characterized in terms of the sample size n, which partially explains the finite-sample performance of $\widehat{f}_n$. This implies that with a small sample size n, the difference between $\widehat{f}_n$ and f_0 cannot be arbitrarily small, and must be lower bounded by a positive constant related to the size n.

6. Analysis of NKI Breast Cancer Data

To demonstrate the usefulness of the proposed method, we apply it to analyze the breast cancer data collected by the Netherlands Cancer Institute (NKI) (van de Vijver et al. (2002)). Tumors from 295 women with breast cancer were collected from the fresh-frozen-tissue bank of the Netherlands Cancer Institute. The tumors were primarily invasive breast cancer carcinoma about 5 cm in diameter. At diagnosis, the patients were 52 years or younger, and the diagnosis occured between 1984 and 1995. Of the 295 patients, 79 died before the study ended, yielding approximately 73.2% censoring.

Approximately 25,000 gene expressions were also collected, of which 70 genes with previously determined average profiles are useful for tumor diagnosis (van de Vijver et al. (2002, p.2002)); these gene expression values are recorded as the log intensity. We study the relationship of survival time with those 70 genes by implementing the proposed method. Here we specify $\mathcal{C}$ as in Section 4.2, and set $M = 4$ and $J = 2$, where ρ_{j1} and ρ_{j2} are the 25th and 75th percentiles, respectively, of the jth variable in X_i in the sample. We assess the prediction performance using the measure discussed in Section 4.3, with $K = 5$.

To gain a better understanding of how the IBS measure performs over a number of data sets instead of a single data set, we apply the procedure described in Section 4.3 repeatedly to N_{boot} bootstrap samples generated randomly from

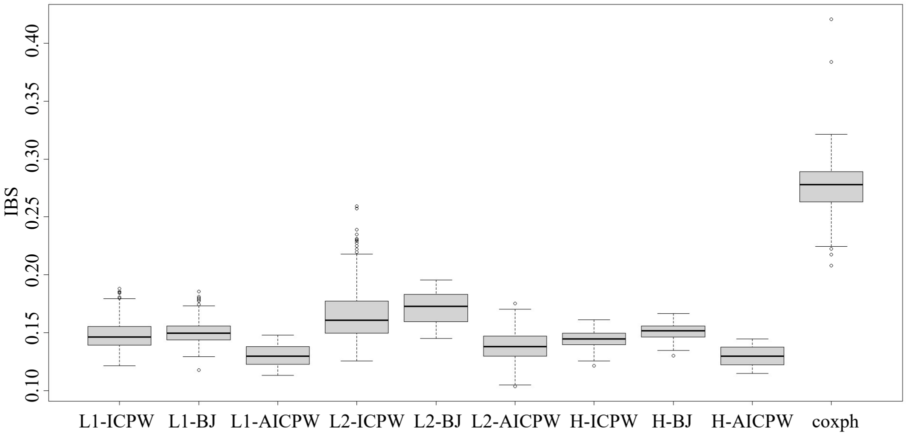

Figure 1. Box plots of the IBS with 500 repeated bootstrapping.

the initial sample $\mathcal{O}_{\text{cd}}$ with replacement, where N_{boot} is taken as 500. Let $\text{IBS}^{(d)}$ denote the value (4.11) yielded by the dth bootstrap samples, for $d = 1, \ldots, N_{\text{boot}}$, where the three adjusted loss functions (3.1), (3.3), and (3.4) are used with the three respective loss functions in Table 1. Box plots for $\{\text{IBS}^{(d)} : d = 1, \ldots, N_{\text{boot}}\}$ are displayed in Figure 1, where L1-BJ, L1-ICPW, and L1-AICPW denote the adjusted loss functions (3.1), (3.3), and (3.4), respectively, with the loss function $L(\cdot, \cdot)$ taken as the L_1-norm; L2-BJ, L2-ICPW, and L2-AICPW denote the same adjusted loss functions with the loss function $L(\cdot, \cdot)$ set as the L_2-norm; and H-BJ, H-ICPW, and H-AICPW denote the adjusted loss functions with the loss function $L(\cdot, \cdot)$ set as the Huber loss function. For comparison purposes, we also apply the "coxph" method of Chen et al. (2013) to analyze the data, and show the results in the box plot labeled "coxph".

Compared with the "coxph" method, the proposed UBEC method generally performs well with small IBS values, regardless of the choice of the loss function $L(\cdot, \cdot)$ or its adjusted version. With a given loss function form, the AICPW adjustment tends to perform best. In addition, with a given adjustment strategy, the Huber loss function outperforms the other two loss functions, and the L_2-norm loss function seems to incur the most variability.

On the other hand, the "coxph" method produces noticeably larger IBS values than our proposed estimators, suggesting unsatisfactory prediction performance. To see why the "coxph" method fails to work, we apply the Schoenfeld residuals (e.g., Lawless (2003, p.364)) and implement the R function `cox.zph` to test the proportional hazards assumption. This yields a p-value of 7.3e-06, suggesting that the proportional hazards assumption is not suitable for describing the data.

7. Discussion

To assess the finite-sample performance of our method, we conduct simulation studies under different settings. The results are provided in Section S5 of the Supplementary Material, and show that the proposed UBEC method produces satisfactory results.

As discussed in Section 3.2, the proposed method hinges on consistent estimations of $F_T(t|X_i)$ and $G(c)$, which are often calculated using existing methods, described in Section 4.2. In contrast to many available parametric or semiparametric methods that focus on inferences about the model parameters or that estimate the conditional survivor function $F_T(t|X_i)$, our goal is to find an optimal function of covariates to predict the transformed failure time T_i. Furthermore, our development enables us to estimate the unconditional survivor function of T_i without knowing the distribution of the covariates X_i or specifying any model forms. In contrast to most existing boosting methods for censored data that do not provide theoretical justifications, the validity of the proposed method is asserted discreetly.

We assume that the censoring time C_i is independent of the covariates X_i, enabling us to estimate the survivor function of the censoring process consistently using the Kaplan–Meier estimator. However, this assumption is not essential, and can be relaxed by replacing the unconditional survivor function $G(c)$ of C_i with the conditional survivor function $G(c|X_i) \triangleq P(C_i > c|X_i)$ of C_i, given X_i. Then, the development can be modified accordingly, where we require a consistent estimator of $G(c|X_i)$.

Supplementary Material

The online Supplementary Material contains a full set of regularity conditions, with discussion, detailed proofs for the results in Sections 3.2 and 5, and further details about our simulation studies.

Acknowledgments

The authors thank the editor and the review team for their comments on the initial version of the manuscript. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Yi is Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs program.

References

Bellot, A. and van der Schaar, M. (2018a). Boosted trees for risk prognosis. *Proceedings of Machine Learning Research* **85**, 1–15.

Bellot, A. and van der Schaar, M. (2018b). Multitask boosting for survival analysis with competing risks. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 1–10. Montréal, Canada.

Benner, A. (2002). Application of "aggregated classifiers" in survival time studies. In *Proceedings in Computational Statistics: COMPSTAT 2002* (Edited by W. Härdle and B. Rönz), 171–176. Physica-Verlag, Heidelberg.

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics* **9**, 1–10.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press

Buckley, J. and James, I. (1979) Linear regression with censored data. *Biometrika* **66**, 429–436.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* **22**, 477-–505.

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association* **98**, 324–339.

Chen, Y., Jia, Z., Mercola, D. and Xie, X. (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and Mathematical Methods in Medicine 2013*, 873595.

Cui, Y., Zhu, R., Zhou, M. and Kosorok, M. (2022) Consistency of survival tree and forest models: splitting bias and correction. *Statistica Sinica* **32**, 1245–1267.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *The Annals of Statistics* **17**, 1157–1167.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation* **121**, 256–285.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**, 1189–1232.

Geenens, G. (2011). Curse of dimensionality and related issues in nonparametric functional regression. *Statistics Surveys* **5**, 30–43.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529–2545.

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning.* Springer, New York.

He, K., Li, Y., Zhu, J., Liu, H., Lee, J. E., Amos, C. I. et al. (2016). Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates. *Bioinformatics* **32**, 50–57.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics* **7**, 355–373.

Ishwaran, H. and Kogalur, U. B. (2010) Consistency of random survival forests. *Statistics and Probability Letters* **80**, 1056–1064.

Jiang, W. (2004). Process consistency for AdaBoost. *The Annals of Statistics* **32**, 13–29.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data.* Wiley, New York.

Lee, D. K. K., Chen, N. and Ishwaran, H. (2021). Boosted nonparametric hazards with time-dependent covariates. *The Annals of Statistics* **49**, 2101–2128.

Li, H. and Luan, Y. (2005). Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics* **21**, 2403–2409.

Lu, W. and Li, L. (2008). Boosting method for nonlinear transformation models with censored survival data. *Biostatistics* **9**, 658–667.

Lugosi, G. and Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics* **32**, 30–55.

Mayr, A., Hofner, B. and Schmid, M. (2016). Boosting the discriminatory power of sparse survival models via optimization of the concordance index and stability selection. *BMC Bioinformatics* **17**, 1–12.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics* **31**, 172—181.

Rubin, D. and van der Laan, M. J. (2007). A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics* **3**, 1–21.

Schapire, R. E. (1990). The strength of Weak learnability. *Machine Learning* **5**, 197–227.

Schapire, R. E. and Freund, Y. (2014). *Boosting: Foundations and Algorithms.* The MIT Press, Cambridge.

Schmid, M. and Hothorn, T. (2008). Flexible boosting of accelerated failure time models. *BMC Bioinformatics* **9**, 1–13.

Steingrimsson, J. A., Diao, L., Molinaro, A. M. and Strawderman, R. L. (2016). Doubly robust survival trees. *Statistics in Medicine* **35**, 3595–3612.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W. et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine* **347**, 1999–2009.

Wang, J. G. (1987) A note on the uniform consistency of the Kaplan-Meier estimator. *The Annals of Statistics* **15**, 1313–1316.

Wang, Z. and Wang, C. Y. (2010). Buckley-James boosting for survival analysis with high-dimensional biomarker data. *Statistical Applications in Genetics and Molecular Biology* **9**, 1–31.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics* **33**, 1538–1579.

Zhu, R. and Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association* **107**, 331–340.

Li-Pang Chen

Department of Statistics, National Chengchi University, Taipei City 11605, Taiwan.

E-mail: lchen723@nccu.edu.tw

Grace Y. Yi

Department of Statistical and Actuarial Sciences, and Department of Computer Science, University of Western Ontario, London, Ontario N6A 3K7, Canada.

E-mail: gyi5@uwo.ca

(Received February 2021; accepted June 2022)