Journal of Algorithms and Computation

journal homepage: http://jac.ut.ac.ir

Survival analyses with dependent covariates: A regression tree-base approach

Mostafa Boskabadi*[1], Mahdi Doostparast†[2] and Majid Sarmad‡[3]

[1, 2, 3]Department of Statistics, Ferdowsi University of Mashhad, P.O. Box 91775-1159,
Khorasan Razavi, Iran

ABSTRACT

Cox proportional hazards models are the most common modelling framework to prediction and evaluation of covariate effects in time-to-event analyses. These models usually do not account the relationship among covariates which may have impacts on survival times. In this article, we introduce regression tree models for survival analyses by incorporating dependencies among covariates. Various properties of the proposed model are studied in details. To assess the accuracy of the proposed model, a Monte–Carlo simulation study is conducted. A real data set from assay of serum free light chain is also analysed to illustrate advantages of the proposed method in medical investigations.

Keyword: Survival tree, Cox proportional hazards model, Dependence, Copula function.

ARTICLE INFO

Article history:
Received 18,September 2019
Received in revised form 15, April 2020
Accepted 02 May 2020
Available online 01, June 2020

AMS subject Classification: Primary: 62N02 , Secondary: 62J02

*bmostafa77@yahoo.com
†Corresponding author: M. Doostparast. Email: doustparast@um.ac.ir
‡sarmad@um.ac.ir

1 Introduction

Survival analysis contains information on time-to-event data, often death or relapse after treatment for a disease. The literature contains rich sets of models and analytical methods via parametric, non–parametric and semi–parametric approaches. Unlike parametric and non–parametric methods, semi–parametric models can provide more consistent estimators under some general conditions. One of the most important semi–parametric models in the analyses of survival data sets is the Cox proportional hazards models which are used to estimate covariate effects and also to prediction of future outcomes [11].
When the response is subject to censoring, regression models are often complex. Also, there exist basic parametric assumptions imposed by these models for implementing the regression models (such as no existence linear combinations of covariates, interactions and high dimensional parameter spaces). Ignoring these assumptions is critical and may cause misleading results. Therefore, various methods for regression models have been discussed in the literature which do not need to the above mentioned conditions. For example, Morgan[22] proposed a method, known as *regression tree*, which divides the data set into homogeneous partitions. This method requires fewer assumptions than the above-mentioned regression models. The Classification And Regression Trees (CART) algorithm, introduced by Breiman[4], can be used for both quantitative and qualitative responses. Specifically, suppose that Y and $\mathbf{X} = (X_1, \cdots, X_m)$ denote the response variable and the vector of the (fixed effect) explanatory variables, respectively. In the CART algorithm, the data set is partitioned into two regions based on a rule of the form $x_j \leq s$, and then the response (quantitative) modelled using the mean of y in each region. Therefore, we seek for j and s by minimizing:

$$\sum_{i:x_j \leq s} (y_i - \bar{y})^2 + \sum_{i:x_j > s} (y_i - \bar{y})^2. \tag{1.1}$$

The algorithm is implemented by the package “`rpart`” in the statistical software `R`. Recently, it has been found that the regression trees provide more accurate estimations and predictions in data mining and computer sciences; See, e.g., Cichosz[10] and references therein. Table 1 summarizes well–known tree-based models with some applications.

Various methods for building tree models in the survival analysis have been proposed in literature. The main feature that differs between proposed tree methods is the splitting criterion. Goldman[12] emphasized importance of tree techniques in biomedical settings. Gordon[13] used distance measures between Kaplan–Meier curves and certain point masses. It is the first paper which discussed the creation of survival trees. Segal[28] extended regression trees to right-censored observations by replacing the conventional splitting criteria. Ciampi[9] employed log-rank test statistics for computing between-node heterogeneity measures and used the Akaike Information Criterion (AIC) for selecting the tree size. Butler[6] also applied the log-rank test statistic for splitting proposes. LeBlanc[19] splitted the covariate space based on a rule that maximizes the difference between the log-likelihood function of the saturated model and the maximized log-likelihood

Table 1: A summary of some famous tree-based models.

Algorithm	Reference	Type of splits	Details	Application
CART	Breiman[4]	Binary	The split criterion is the Gini index. It uses cross-validation to estimate the misclassification cost of each subtree and chooses the one with the lowest estimated cost. It also considers the sum of squared residuals as the impurity function. It is pruned with the cross-validation method.	Immunosuppression and the diagnosis of cancer
CHAID	Kass[17]	Multiple	For split selection, each variable is assessed with a Bonferroni-adjusted p-value, and the one with the smallest p-value is selected to split the node. It is pruned by using Bonferroni-adjusted significance tests.	Study of student data set from the University of Witwatersrand
ID3	Quinlan[24]	Multiple	The split criterion is the entropy index and it chooses the split that yields the highest gain ratio. It uses a conservative estimate of the error at each node to prune.	The weather classification
QUEST	Loh[21]	Multiple	It uses hypothesis tests to choose the split variables, namely, analyses of variance for non–categorical variables and chi-squared tests for categorical variables. Variable selection is unbiased.	Evaluations of teaching assistant at the University of Wisconsin Madison
GUIDE	Loh[20]	Binary	Best split on the selected variable, using an appropriate loss function. The variable selection is unbiased. It is pruned with the cross-validation method.	The mammography data analyses
BART	Chipman[7]	Binary	It provides a Bayesian approach for regression trees. It is pruned by each regression tree in a sequential Bayesian backfitting algorithm. See as Chipman[8]	Hematopoietic stem cell transplantation data
DART	Boskabadi[3]	Binary	The split criterion is based on the dependency structure and Kendall's tau correlation among covariates. It considers linear models in leaves. The tree is pruned with the generalized likelihood ratio test.	Automobile company data analyses

function. This method constructs a tree by representing the relative risk function. This algorithm is a generalization of the CART algorithm for survival data. To do this, the data set is partitioned into two regions based on a rule of the form $x_j \leq s$, and then the Kaplan–Meier estimate of the the response median, denoted by $v(.)$, is derivded in each region. Therefore, we seek for j and s by minimizing:

$$\sum_{i:x_j \leq s} |Y_i - v_L(Y)| + \sum_{i:x_j > s} |Y_i - v_R(Y)|, \tag{1.2}$$

where $v_L(Y)$ and $v_R(Y)$ are the Kaplan–Meier estimates of the response medians for regions in which $\{x_j \leq s\}$ and $\{x_j > s\}$, respectively. The node impurity measure for the j-th terminal node in a survival tree model is defined as

$$Q_j(T) = \frac{1}{n_j} \sum_{i=1}^{n_j} |Y_i - v_j(Y)|, \tag{1.3}$$

when n_j and $v_j(Y)$ are, respectively, the number of uncensored observations observations and the Kaplan–Meier estimate of the the response median in the j-th node. The cost-complexity criterion for a survival tree structure with k terminal nodes is formularized as

$$C_k(T) = \sum_k n_k Q_k(T) + \eta |T|, \tag{1.4}$$

where $|T|$ denotes the number of terminal nodes in the tree and η stands for the tuning parameter or error.
Hothorn[15] implemented an unbiased survival tree using the log-rank test as a method for splitting data sets. The algorithm is implemented in the statistical software `R` with package “`partykit`”. Bertolet[2] extended it for partitioning a data set based on time-varying Cox models. Other proposed methods include Huang[16], Xu[32] and Wallace[31]. Boskabadi[3] dealt with a new approach for regression trees which considers the dependency structures among covariates for splitting the data set. Hereafter, it is called “Dependence And Regression Trees” and abbreviated by DART. When linear models and the CART model are fitted in leaves, the DART algorithm is called LM-DART and C-DART, respectively. In this paper, the DART approach is studied in details for analysing lifetime data sets where the Cox proportional hazards model are fitted in terminal nodes (Cox-DART). Therefore, the rest of this paper is organized as follows. Section 2 defines formally the DART algorithm and a non–parametric algorithm for deriving the respective optimal splitting points. The DARTs with Cox proportional hazards models in terminal nodes (leaves) are studied in Section 3 with more details. Indeed, the assessment and accuracy criterion for the model selection and the problem of hypotheses testing are discussed. In Section 4, a simulation study is conducted to carry out the performance of the proposed DART model. A real data set on an “Assay of serum free light chain” study is also analysed using the obtained results in Section 4. Finally, Section 5 gives conclusions and further remarks.

2 DART algorithm in survival data

In this section, the DART approach is extended for survival data sets in the presence of censored observations. Suppose that T denotes the failure time of primary interest, C is the censoring time, and $\mathbf{X} = (X_1, \cdots, X_m)$ stands for a vector of covariates. Let $Y = \min(T, C)$ and $\delta = I(T \leq C)$, where $I(E)$ is the indicator function for event E, that is, $I(E) = 1$ if E occurs and $I(E) = 0$ otherwise. Suppose that the covariates X_i and X_j are dependent and the dependency among them changes in different regions of $\mathcal{S}_{X_1} \times \mathcal{S}_{X_2} \times \cdots \times \mathcal{S}_{X_k}$ (the Cartesian products of $\mathcal{S}_{X_1}$, $\mathcal{S}_{X_2} \cdots \mathcal{S}_{X_k}$). Here, $\mathcal{S}_X$ stands for the support of the random variable X; See Figure 1.

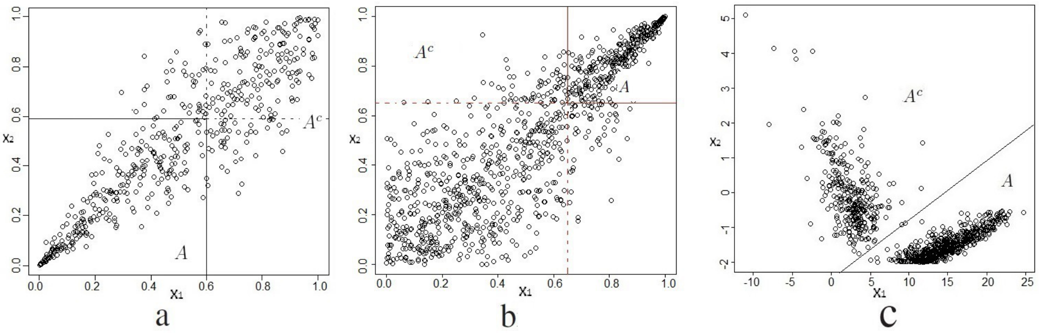


Figure 1: Some possible scatter plot between (X_i, X_j)

In practice, change in the kind of the dependency among covariates may impose some impacts on the failure time T. Then, the estimate of the survival function (SF) may be misleading if one does not consider the dependency among covariates (See Subsection 4.3). The DART model considers the dependency among covariates and produce a suitable regression tree.

For a motivation about the DART approach, let $A_k^{[strong]}$:=$\{$ Area of (X_i, X_j) with strong (either positive or negative) dependency $\}$, $A_l^{[weak]}$:=$\{$ Area of (X_i, X_j) with weak dependency or independence $\}$, for $k, l = 1, \ldots, n$. The process of fitting a DART algorithm is to find l and k that divide the area into more homogeneous areas. Assuming square error loss for prediction proposes, it minimizes

$$\sum_k E[(Y - f_k(\mathbf{X}_{A_k^{[strong]}}))^2 \mid A_k^{[strong]}]P(A_k^{[strong]}) + \sum_l E[(Y - f_l(\mathbf{X}_{\mathbf{A}_\mathbf{l}^{[\mathbf{weak}]}}))^2 \mid A_l^{[weak]}]P(A_l^{[weak]}).$$

where X_A stands for the covariates of the data set restricted to the area A. If covariates impact on the survival function vary in various regions, then we expect this occurs in regions of (X_i, X_j) that maximize $\tau_{X_i, X_j | A_k^{[strong]}}$ and minimize $\tau_{X_i, X_j | A_l^{[weak]}}$, where $\tau_{X_1, X_2 | A_k^{[strong]}}$ ($\tau_{X_1, X_2 | A_l^{[weak]}}$) stands for the Kendall's tau correlation among X_i and X_j over the set $A_k^{[strong]}$ ($A_l^{[weak]}$) provided that $|A^{[strong]}| \geq 2$ and $|A^{[weak]}| \geq 2$ where $|A|$ stands for the number of samples falling in region A; See Appendix 1.

Therefore, one should maximize $\tau_{X_i,X_j|A_k^{[strong]}}$ and minimize $\tau_{X_i,X_j|A_l^{[weak]}}$. There is a multi–object optimization problem (MOP). Various approaches may be used for solving this problem. In the DART approach, we suggest to transform the mentioned MOP into a single-object optimization problem. That is, one should maximize the object function (OF) as given by

$$OF(k,l) := \sum_k |\tau_{X_1,X_2|A_k^{[strong]}}| - \sum_l |\tau_{X_1,X_2|A_l^{[weak]}}|.$$

Remark 2.1. The Kendall's tau correlation for splitting the data set is used, since it uses the ranks of observations and hence it does not depend on the marginal distributions of covariates.

The following steps is given for implementing the DART model in survival data analyses:

1: Enter the survival time Y and affecting vector of covariates $\mathbf{X} = (X_1, \cdots, X_m)$.

2: Is there any change of dependency between covariates (This can be done with scatter plot between the two covariates. Such as Figures 1)? If no then exit from the DART algorithm and fit the other appropriate model. In this step, one can do clustering based on the same of dependency structure in various regions between covariates.

3: Select a pair (such as (X_i, X_j) for $1 \leq i, j \leq m; i \neq j$) covariates that the change of dependency is occurred with the greatest effect on the survival time. If there exist more than two pairs of covariates which exhibit dependency changes, then we choose a pair which has greater R^2 (coefficient of determination) in a simple tree model when the covariate is an indicator function which denotes that the observations belong to respect partitions. In the DART approach recommend the coefficient of determination because it stands for the percentage change in survival times expressed by the tree, but other refined criteria such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) can also be used.

4: Find thresholds for selected covariates using their dependency structure. Solve the associate MOP by maximizing the corresponding OF for the areas. Refer to Sections 2.1

5: Does the survival time on the areas differ significantly (Pre–pruning the tree using the hypotheses testing procedures by log–rank test; See Klein & Moeschberger, 1997)? If no then exclude these two covariates for re–categorization process and go to Step 2.

6: Is there any change of dependency among covariates in each partition? If yes then choose a partition and go to Step 3. If no then go to the next step. For example, assume that in the subset D_1 there is another type of dependency among covariates X_l and X_k for $1 \leq l, k \leq m; l \neq k$. Similarly, the partition D_1 is divided into two subsets D_{11} and $D_{12} = D_1 - D_{11}$.

6: Fit the appropriate models in partitions.

6: Does the fitted models on the partitions differ significantly(Post–pruning the tree using the hypotheses testing procedures)? If no then exclude these two covariates for re–categorization process and go to Step 2.

7: Construct the tree and then fit statistical models in leaves. By parametric, nonparametric and semiparametric models we can provide a hybrid approach of the existing methodologies in the leaves.

2.1 Determination of split points in DART

In DART approach once, the type of dependency among covariates was determined, it is suggested to use the following technique to derive split points. First, assume that in the scatter plot of the two covariates X_i and X_j, there exist some dependency changes.
For illustration, one can see that $A_j := \{(X_1, X_2)|X_1 \leq X_1^{[j:n]}, X_2 \leq X_2^{[j:n]}\}$, $j = 1, 2$ in Figure 1(a). Also $A_j := \{(X_1, X_2)|X_1 > X_1^{[j:n]}, X_2 > X_2^{[j:n]}\}$, $j = 1, 2$ in Figure 1(b). Here $X_i^{[1:n]} \leq X_i^{[2:n]} \leq \cdots \leq X_i^{[n:n]}$ denote the n observations of the ith covariates in magnitude order. Therefore, the corresponding MOP can be transformed into single-object optimization problem in which one maximizes

$$OF(j) := |\tau_{X_1,X_2|A_j}| - |\tau_{X_1,X_2|A_j^c}|.$$

One can easily compute $OF(j)$ for $j = 1, \ldots, n$ and then find $j^\star := \arg\max_j OF(j)$. Hence, the optimal partition is given by $A_{j^\star}$ and $A_{j^\star}^c$. In Figure 1(d), $A_{j_1,j_2} := \{(X_1, X_2)|X_1 \leq j_1 X_2 + j_2\}$ the corresponding MOP is also transformed to maximizing

$$OF(j_1, j_2) := |\tau_{X_1,X_2|A_{j_1,j_2}}| + |\tau_{X_1,X_2|A_{j_1,j_2}^c}|,$$

Then, we divide the data set (D) into partitions $D_i = \{(Y, \mathbf{X}) \mid (X_1, X_2) \in A_i\}$ for $i = 1, 2, \ldots$. Finally, one can fit various models on the subsets D_i. This method may be extended to more than two regions. For more information about the various dependency, see Boskabadi[3].

2.2 Dependency structure and thresholds

In this section, we use the dependency structure distribution function among covariates for determining split points. Sometimes, visually inspection of scatter plot is difficult to detect the dependency structures among covariates. One can use the “Copula” function among covariates and the corresponding contour diagram and the scatter plot of $(\hat{F}_{X_i}(.), \hat{F}_{X_j}(.))$, where $\hat{F}_X(.)$ is the empirical marginal distribution X. To do this, we use the concept of “Copula” function.

Definition 2.2 (Nelsen[23]). A two-dimensional function $C : [0,1]^2 \to [0,1]$ with the following properties is called *Copula*:
(i) $C(u,0) = C(0,v) = 0; \qquad \forall u, v \in [0,1]$,
(ii) $C(u,1) = u; \quad C(1,v) = v; \quad \forall u, v \in [0,1]$,
(iii) for every $B = [u_1, u_2] \times [v_1, v_2]$ in $[0,1]$ such that $v_1 < v_2$ and $u_1 < u_2$,

$$C(u_2, v_2) + C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) \geq 0.$$

In a DART model, one can use various methods for determining thresholds. For example, previous similar researches and the researcher's knowledge on the subject may be used. Scatter plots among covariates $X_1, \cdots, X_m$ provide an insight for deriving thresholds. Moreover, according to (X_i, X_j) dependency structure, one can use quantiles of covariates as follow:

1: If the dependency structures do not change in various areas, then it is not necessary to use a DART model. Example includes, linear dependencies occur known as collinearity problems. Also, some well–known copulas including Frank, FGM, Normal and Ali–Mikhail–Hag (AMH) exhibit neither upper nor lower tail dependency.

2: If the dependency is not uniform (same) in all areas, Boskabadi[3] suggested a heuristic approach for deriving split points (See Subsection 2.1). Some copulas such as Joe, Clayton and Gumbel which have a tail dependency and the split points may be derived with this approach. The tail dependency coefficients are discussed in Appendix 1.

Remark 2.3. Note that, to identify approximately regions in which the dependency structures among X_i and X_j are the same, one can use the empirical joint distribution function among X_i and X_j and the corresponding contour diagram as well the scatter plot of $\hat{F}_{X_i}(.)$ and $\hat{F}_{X_j}(.)$; See Figure 2

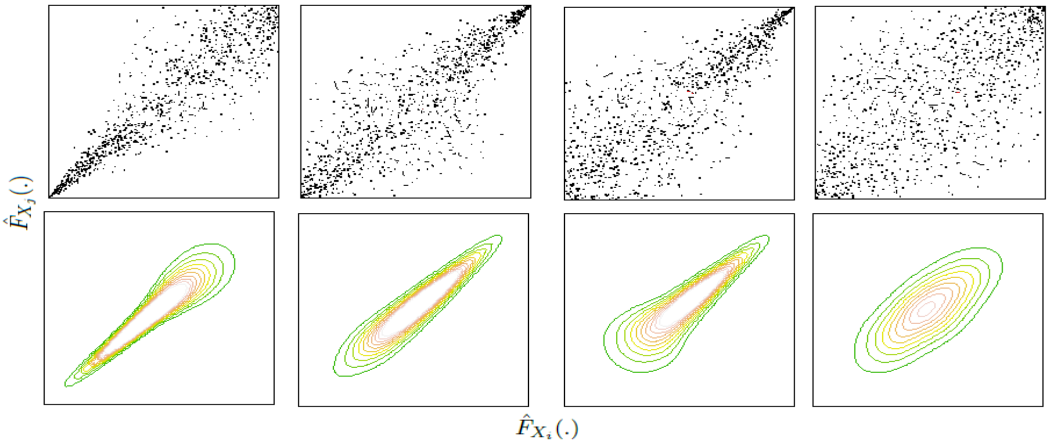


Figure 2: Some possible scatter plot and contour diagram for $\hat{F}_{X_i}(.)$ and $\hat{F}_{X_j}(.)$.

With this in mind, Boskabadi[3] derived split points for the Clayton copula

$$C(u,v) = [\max(u^{-\theta} + v^{-\theta} - 1), 0]^{-\frac{1}{\theta}}, \qquad u,v \in [0,1], \tag{2.1}$$

where $\theta \in [-1, \infty)$. Notice that in Table 2, we have the down tail dependency $\lambda_L = 2^{-1/\theta}$ and upper tail dependency $\lambda_U = 0$.

Table 2: Simulation-based average optimal split points for a DART model with Clayton copula for covariates (Parentheses denote the standard deviations of generated $j^\star$s for 1000 iterations).

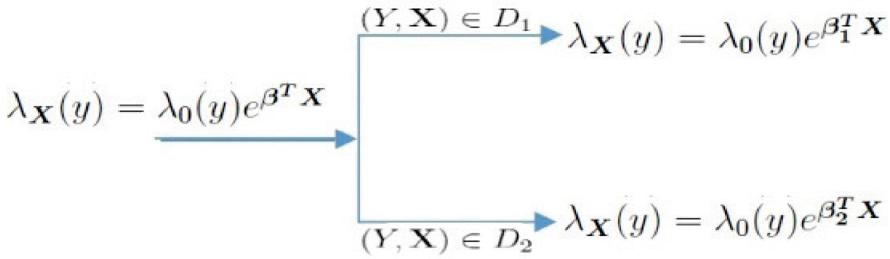
θ	3	4	5	6	7	8	9
$j^\star$	68.3 (2.5)	73.9 (2.6)	77.8 (2.9)	80.1 (2.2)	82.3 (2.1)	84.3 (1.9)	85.7 (2)

3 Cox-DART: Cox proportional hazards model in DART

In this section, the DART model is implemented in which the Cox proportional hazards models are fitted in terminal nodes. The derived model is abbreviated by Cox-DART. To assess the performance of the Cox-DART model and model selection, some criteria are also proposed. To do this, an alternative representation for the DART model is presented. Suppose that T denotes the failure time of primary interest, C is the censoring time, and $\mathbf{X} = (X_1, \ldots, X_m)$ is a vector of covariates. Let $Y = \min(T, C)$ and $\delta = I(T \leq C)$, where $I(E)$ is the indicator function for event E, that is, $I(E) = 1$ if E occurs and $I(E) = 0$ otherwise. Consider the well–known Cox model

$$\lambda_{\mathbf{X}}(y) = \lambda_0(y) \exp\{\beta^T \mathbf{X}\}, \tag{3.1}$$

where λ_0 and $\lambda_{\mathbf{X}}$ are the baseline hazard function and the with covarietes hazard function, respective. Also, β_i $(i = 1, \ldots, m)$ are the coefficients (Klein[18]). Assume that a given Cox-DART model consists two branches with two subsets D_1 and D_2 as follows:


$\lambda_{\mathbf{X}}(y) = \lambda_0(y)e^{\boldsymbol{\beta}^T\mathbf{X}}$
$(Y, \mathbf{X}) \in D_1$
$\lambda_{\mathbf{X}}(y) = \lambda_0(y)e^{\boldsymbol{\beta}_1^T\mathbf{X}}$
$(Y, \mathbf{X}) \in D_2$
$\lambda_{\mathbf{X}}(y) = \lambda_0(y)e^{\boldsymbol{\beta}_2^T\mathbf{X}}$

A unified version for the above Cox-DART model is

$$\begin{aligned} \lambda_{\boldsymbol{X}}(y) &= \lambda_0(y)e^{(\boldsymbol{\beta_1^T X})I_A+(\boldsymbol{\beta_2^T X})I_{A^c}} \\ &= \lambda_0(y)e^{(\boldsymbol{\beta_1}I_A+\boldsymbol{\beta_2}I_{A^c})^T\boldsymbol{X}} \\ &= \lambda_0(y)e^{\boldsymbol{\beta^{\star T} X}}, \end{aligned} \tag{3.2}$$

where $\boldsymbol{\beta}^\star = \boldsymbol{\beta}_1 I_A + \boldsymbol{\beta}_2 I_{A^c}$. In the jth terminal node, the probability that the event is due to failure at the fixed point time t_i is given by

$$\begin{aligned} P_j(y_i) &= \frac{\lambda_j(y_i \mid \boldsymbol{X}_i)}{\Sigma_{j\in R(y_i)}\lambda_j(y_i \mid \boldsymbol{X}_j)} \\ &= \frac{e^{\boldsymbol{\beta_j^T X_i}}}{\Sigma_{j\in R(y_i)}e^{\boldsymbol{\beta_j^T X_j}}}, \end{aligned} \tag{3.3}$$

where $j = 1, 2$ and $R(t_i)$ is the set of all individuals who are still under study at a time just prior to t_i. Notice that $R(t_i)$ is also called "the number at risk" in time t_i. From (3.3), the partial likelihood function (LF) is obtained as

$$L(\boldsymbol{\beta_j}; D_j) = \prod_{i|\delta_{ij}=1} \frac{e^{\boldsymbol{\beta_j^T X_i}}}{\Sigma_{j\in R(y_i)}e^{\boldsymbol{\beta_j^T X_j}}}, \qquad j = 1, 2. \tag{3.4}$$

The maximum (partial) likelihood estimates (MLEs) are derived by maximizing (3.4) in the jth terminal node. In general, the failure probability at the fixed time point t_i in the Cox-DART model (3.2) is

$$P_{tree}(y_i) = \frac{e^{(\boldsymbol{\beta_1^T X_i})I_A+(\boldsymbol{\beta_2^T X_i})I_{A^c}}}{\Sigma_{j\in R(y_i)}e^{(\boldsymbol{\beta_1^T X_j})I_A+\boldsymbol{\beta_2^T X_j})I_{A^c}}}.$$

Therefore, the overall partial LF in the Cox-DART model (3.2) is obtained from (3.4) as

$$\begin{aligned} L(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}; D) &= \prod_{i|\delta_{tree}=1} \frac{e^{(\boldsymbol{\beta_1^T X_i})I_A+(\boldsymbol{\beta_2^T X_i})I_{A^c}}}{\Sigma_{j\in R(y_i)}e^{(\boldsymbol{\beta_1^T X_i})I_A+(\boldsymbol{\beta_2^T X_i})I_{A^c}}} \\ &= \prod_{i|\delta_{tree}=1} \frac{e^{(\boldsymbol{\beta_1^T X_i})I_A}e^{(\boldsymbol{\beta_2^T X_i})I_{A^c}}}{\Sigma_{j\in R(y_i)_{y_i\in A}}e^{(\boldsymbol{\beta_1^T X_i})I_A}\Sigma_{j\in R(y_i)_{y_i\in A^c}}e^{(\boldsymbol{\beta_2^T X_i})I_{A^c}}} \\ &= L(\boldsymbol{\beta_1}; D_1)L(\boldsymbol{\beta_2}; D_2). \end{aligned} \tag{3.5}$$

Similarly, the partial LF in the Cox-DART model with k terminal nodes is

$$L(\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_k}; D) = \prod_{j=1}^{k} L(\boldsymbol{\beta_j}; D_j), \tag{3.6}$$

where $D = \bigcup_{j=1}^{k} D_j$ and $D_1, \ldots, D_k$ is a portion of the data set D.

3.1 Model selection and assessment

There are different criteria for selecting and comparing statistical models. In this section, the Akaike Information Criterion (AIC), the coefficient of determination based on the Schoenfeld residuals (R^2_{sch}) and the root mean squared error (RMSE) criterion in the proposed DART approach are explained.

3.1.1 AIC criterion

The Akaike's information criterion (AIC) in a given model with the LF $L(\boldsymbol{\theta}; \mathbf{X})$ is defined as (Burnham[5])

$$\mathrm{AIC} = -2\ln(L(\hat{\boldsymbol{\theta}}; \mathbf{X})) + 2m,$$

when m is the number of the model parameters and $\hat{\boldsymbol{\theta}}$ is the MLE of the parameter vector $\boldsymbol{\theta}$. As mentioned by Klein[18], one may use the partial LF (3.6) instead of the (full) LF. Suppose that in a given regression tree model, there exist k terminal nodes and each terminal node includes a sample of size n_i $(i = 1, 2, \ldots, k)$. Then

$$\begin{aligned} AIC_{tree} &= -2\ln(L(\beta_1, \cdots, \beta_k; D)) + 2\sum_{i=1}^{j} m_i \\ &= -2\ln(\prod_{j=1}^{k} L_(\beta_j; D_j)) + 2\sum_{i=1}^{j} m_i \\ &= \sum_{j=1}^{k} (-2\ln(L(\beta_j; D_j) + 2m_j) \\ &= AIC_{nod1} + AIC_{nod2} + \cdots + AIC_{nodk}. \end{aligned} \tag{3.7}$$

Remark 3.1. Note that the AIC of the DART model in (3.7) is equal to the sum of the AIC for every terminal node.

3.1.2 Coefficient of determination based on Schoenfeld residuals

The well–known regression coefficient of determination R^2 is defined by

$$R^2 = 1 - \left(\frac{L(0)}{L(\hat{\theta})}\right)^{\frac{2}{n}}, \tag{3.8}$$

where $L(0)$ and $L(\hat{\theta})$ are the null and the full LFs, respectively. A measure for goodness-of-fit assessment is defined on the basis of the squared residuals (Schoenfeld[27]). The

Schoenfeld residuals for the jth terminal node are

$$\begin{aligned} r_{sck_{ij}}(\boldsymbol{\beta}) &= X_{ij} - E[X_{ij} \mid \boldsymbol{\beta}] \\ &= X_{ij} - \sum_{j|j \in R(j)} X_{ij} P_{jk}(\boldsymbol{\beta}; t_i) \\ &= X_{ij} - \sum_{j|j \in R(j)} X_{ij} \left(\frac{e^{\boldsymbol{\beta_j^T X_i}}}{\sum_{j \in R(y_i)} e^{\boldsymbol{\beta_j^T X_j}}} \right), \end{aligned} \tag{3.9}$$

for $1 \leq i \leq n_j$ and $1 \leq j \leq k$. Therefore, residuals for the null model (without covariates, that is for $\boldsymbol{\beta}_j = \mathbf{0}$) in jth terminal node are obtained by replacing the probability $P_j(\boldsymbol{\beta}; y_i)$ with

$$P_j(\mathbf{0}; y_i) = \frac{1}{|R(i)|}. \tag{3.10}$$

So the coefficient of determination R^2 on the basis of the Schoenfeld residuals is

$$\begin{aligned} R^2_{Sch_j} &= \frac{\sum_{i=1}^{n_j} \delta_i \{\boldsymbol{\beta}_j^T r^2_{sck_{ij}}(\mathbf{0})\} - \sum_{i=1}^{n_j} \delta_i \{\boldsymbol{\beta}_j^T r^2_{sck_{ij}}(\boldsymbol{\beta})\}}{\sum_{i=1}^{n_j} \delta_i \{\boldsymbol{\beta}_j^T r^2_{sck_{ij}}(\mathbf{0})\}} \\ &= 1 - \frac{\sum_{i=1}^{n_j} \delta_i \{\boldsymbol{\beta}_j^T r^2_{sck_{ij}}(\boldsymbol{\beta})\}}{\sum_{i=1}^{n_j} \delta_i \{\boldsymbol{\beta}_j^T r^2_{sck_{ij}}(\mathbf{0})\}}. \end{aligned} \tag{3.11}$$

An overall coefficient of determination for the Cox-DART model (3.2) can be defined by

$$R^2_{Sch_{tree}} = 1 - \frac{\sum_{j=1}^{k} \sum_{i=1}^{n} \delta_i \{\boldsymbol{\beta}_j^T r^2_{sck_{ij}}(\boldsymbol{\beta})\}}{\sum_{j=1}^{k} \sum_{i=1}^{n} \delta_i \{\boldsymbol{\beta}_j^T r^2_{sck_{ij}}(\mathbf{0})\}}. \tag{3.12}$$

3.1.3 Root Mean Squared Error (RMSE) criterion

The root mean squared error (RMSE) measures the expected squared distance between the fitted survival function (SF) predicted at a specific value and the corresponding true SF. Following Ambler[1], we use the RMSE in deference models for uncensored data, say $y_1, \ldots, y_n$, that is,

$$RMSE(S, \hat{S}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\hat{S}(y_i) - S(y_i) \right)^2}, \tag{3.13}$$

where $\hat{S}(y_i)$ stands for the fitted SF at the point y_i $(1 \leq i \leq n)$ and $S(y_i)$ is the true SF. For the Cox model, we have $\hat{S}(y_i) = \exp\{-\Lambda_0(y_i)\}^{\exp\{\hat{\beta}^T \mathbf{X_i}\}}$. In this paper, it is assumed that $\Lambda_0(y_i) = y_i^{\alpha}$ (the Weibull commulative hazard function) and then $S(y_i) = \exp\{-\exp(\beta^T \mathbf{X_i})\mathbf{y_i^{\alpha}}\}$.

Remark 3.2. Note that the true SF $S(y_i)$ in (3.13) is usually unknown. So, one may use some non–parametric estimators for $S(y_i)$ such as the Kaplan–Meier $\hat{S}_{KM}(y_i)$ estimator to calculate RMSE (3.13). Therefore

$$RMSE(\hat{S}_{KM}, \hat{S}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{S}(y_i) - \hat{S}_{KM}(y_i)\Big)^2}. \tag{3.14}$$

Remark 3.3. Notice that the RMSE may be used for every parameter of interest; See Subsection 4.3.

3.2 Tree post–pruning

In the preceding sections, we consider classification of data sets for deriving better and more reasonable models by examining dependencies among predictors. Here a questions may arise in mind: When is the obtained Cox-DART model more efficient than a standard Cox proportional hazards model? In this subsection, we study the problem of hypotheses testing for the this question, which leads to the tree pruning for the obtained Cox-DART model. To do this, suppose that a given Cox-DART model consists two terminal nodes with m predictors as in Model (3.2). The answer to the above question is equivalent to the following problem of hypotheses testing

$$H_0 : \forall j : \beta_{1j} = \beta_{2j} \qquad \text{versus} \qquad H_1 : \exists j : \beta_{1j} \neq \beta_{2j} \tag{3.15}$$

Here, the data set D is partitioned into two subsets D_1 and D_2, that is $D = D_1 \cup D_2$ and $D_1 \cap D_2 = \phi$, where ϕ stands for the empty set. To answer the above mentioned question, we used deviances of models. The deviance for the root model is

$$D_{Root} = 2[L(saturated) - L_{root}(\hat{\boldsymbol{\theta}}; \mathbf{X})], \tag{3.16}$$

where $L(saturated)$ is the log-likelihood function of the saturated model and $L_{root}(\hat{\boldsymbol{\theta}}; \mathbf{X})$ is the maximized log-likelihood function for the root model. Under H_0 in (3.15), D_{Root} follows the chi-square distribution with $n - m$ degrees of freedom (df) (Hosmer[14]). Similarly, the deviance for the Cox-DART model is defined by

$$D_{Cox-DART} = 2[L(saturated) - L_{Cox-DART}(\hat{\boldsymbol{\theta}}; \mathbf{X})], \tag{3.17}$$

where $L_{Cox-DART}(\hat{\boldsymbol{\theta}}; \mathbf{X})$ is the maximized log-likelihood function for the Cox-DART model. Therefore $D_{Cox-DART}$ follows the chi-square distribution with $n-2m$ df. So, the deviance criterion for splitting is

$$\begin{aligned} D &= D_{Root} - D_{Cox-DART} \qquad (3.18) \\ &= 2[L(saturated) - L_{root}(\hat{\boldsymbol{\theta}}; \mathbf{X})] - 2[L(saturated) - L_{Cox-DART}(\hat{\boldsymbol{\theta}}; \mathbf{X})] \\ &= 2[L_{Cox-DART}(\hat{\boldsymbol{\theta}}; \mathbf{X}) - L_{root}(\hat{\boldsymbol{\theta}}; \mathbf{X})], \end{aligned}$$

where D has the chi-square distribution with $m(= 2m - m)$ df. Equivalently, the hypothesis H_0 in (3.15) is rejected if $D > \chi_{m,\alpha}$, where $\chi_{m,\alpha}$ stands for the upper αth quantile of the chi-square distribution with m df.

4 Numerical studies

For illustration purposes, we analyse a simulated sample and a real data set using the results obtained in the preceding sections. Moreover, a simulation study is conducted to carry out the performance of the proposed Cox-DART model.

4.1 A simulated illustrative example

Suppose that $D = (Y, \mathbf{X})$ be data set and $\mathbf{X} = (X_1, \cdots, X_5)$ be covariates. The Clayton copula function by parameter $\theta = 3$ is assumed for the pair (X_1, X_2) corresponding to a down tail dependency. Also, The Clayton copula function by parameter $\theta = 5$ is assumed for the pair (X_3, X_4); See Table 2. Suppose $X_5 = F^{-1}_{X_5}(U)$, where U follows the standard uniform distribution, where $X_i \sim F_{X_i}$, $1 \leq i \leq 5$. The marginal distributions of X_1, X_2 and X_5 are assumed to be the standard exponential distribution. Therefore, the change point is derived from Section 2.2 as $(q_{X_1}, q_{X_2}) = (1.139, 1.139)$ where $q_{X_1} = q_{X_2}$ is equal to 68th percentile of the standard exponential distribution and hence $A = \{X_1 < 1.139, X_2 < 1.139\}$. The marginal distributions of X_3 and X_4 are the standard uniform distribution. Therefore, the change point is similarly derived from Section 2.2 as $(q_{X_3}, q_{X_4}) = (0.78, 0.78)$ and hence $B = \{X_3 < 0.78, X_4 < 0.78\}$. Consider the hazard model

$$\lambda_{\boldsymbol{X}}(y) = \lambda_0(y) e^{\boldsymbol{\beta}^{\star T} \boldsymbol{X}}, \qquad \forall y > 0, \tag{4.1}$$

where $\lambda_0(y)$ is the baseline hazard function of the Weibull distribution with the shape parameter $\alpha = 2$ and the scale parameter $\lambda = 1$. The model parameters are given by

$$\begin{cases} \boldsymbol{\beta_1^T X} = 0.9X_1 + 0.001X_2 + 0.8X_3 + 0.5X_4 + 0.6X_5 & \text{for}(Y, \mathbf{X}) \in \{(X_1, X_2) \in A\} \\ \boldsymbol{\beta_2^T X} = 0.8X_1 + 0.9X_2 + 0.003X_3 + 0.7X_4 + 0.2X_5 & \text{for}(Y, \mathbf{X}) \in \{(X_1, X_2) \in A^c \cap B\} \\ \boldsymbol{\beta_2^T X} = 0.0012X_1 + 0.5X_2 + 0.4X_3 + 0.9X_4 + 0.001X_5 & \text{for}(Y, \mathbf{X}) \in \{(X_1, X_2) \in A^c \cap B^c\} \end{cases}$$

By the above mentioned assumptions, a sample of size $n = 1000$ is generated from the Cox-DART model and then survival times $T_1, \ldots, T_n$ are simulated using Equation (4.1). Then, the censoring times $C_1, \ldots, C_n$ are generated from the standard exponential distribution. Finally, we computed $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. Figure 3 displays a scatter plot between the empirical distributions for the observed (X_1, X_2) and (X_3, X_4).
As mentioned in Subsection 2.1, the desired quantiles of (X_1, X_2) for fitting a Cox–DART are 68th sample percentiles observed data for (X_1, X_2), i.e. $\hat{q}_{X_1} = 1.142$ and $\hat{q}_{X_2} = 1.138$. So, the whole data set D is partitioned into two subsets with $\hat{A} = \{X_1 < 1.142, X_2 < 1.138\}$. Thus, $D = D_1 \cup D_2$, where

$$D_1 = \{(Y, \mathbf{X}) \mid (X_1, X_2) \in \hat{A}\} \qquad \text{and} \qquad D_2 = \{(Y, \mathbf{X}) \mid (X_1, X_2) \in \hat{A}^c\}.$$

Similarly, the desired quantiles of (X_3, X_4) are derived as $\hat{q}_{X_3} = 0.773$ and $\hat{q}_{X_4} = 0.799$. So $\hat{B} = \{X_3 < 0.773, X_4 < 0.799\}$ and the data set D_2 is partitioned into two subsets as

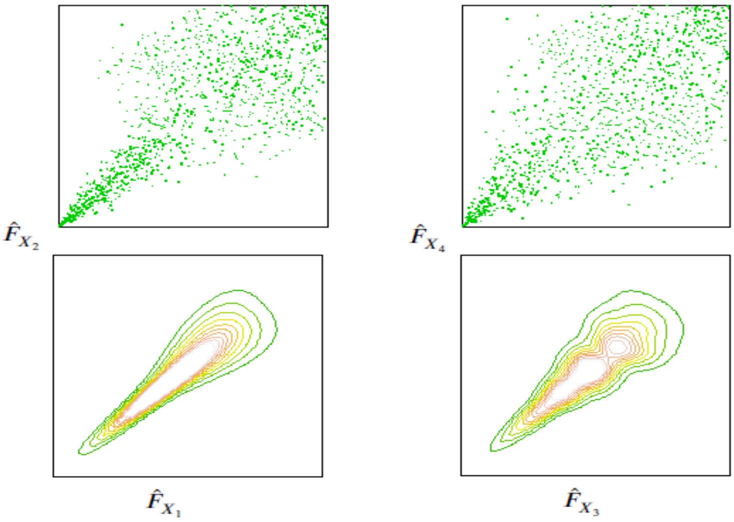


Figure 3: The scatter plot and the contour diagrams between the empirical distributions of (X_1, X_2) and (X_3, X_4)

$$D_{21} = \{(Y, \mathbf{X}) \mid (X_3, X_4) \in \hat{A}\} \qquad \text{and} \qquad D_{22} = \{(Y, \mathbf{X}) \mid (X_3, X_4) \in \hat{A}^c\}.$$

Notice that log–rank test of the split by (X_3, X_4) for survival times is not significance in the data set D_1 and therefore it is not partitioned according to the DART approach. The Cox proportional hazards models is fitted to the data set in terminal nodes by the `survival` package in the statistical software `R`. Therefore, the Cox-DART models is fitted to the data set.

In Table 3, the performance criteria AIC, R^2_{sch} and RMSE(S,$\hat{S}$) of the fitted models are presented.

Table 3: Various criteria for the illustration example.

Model	RMSE(S,$\hat{S}$)	R^2	AIC
Cox(Root)	0.29	0.19	11562
Cox-DART	0.24	0.52	9320

From Table 3, we see that the proposed Cox-DART model dominates the Cox proportional hazards model without dividing the data set (the root model). It is very important to compare the estimated model coefficients $\beta^{\star}_j$. One can see that the coefficients estimates at the terminal nodes are very close to the corresponding coefficients in the true model, while estimated model coefficient, in the root of the tree, are unrealistic and misleading.

4.2 A MCMC simulation study

One sample does not tell us so much. So, we conducted a simulation study to assess the performance of the proposed Cox-DART model with $N = 10^4$ iterations for $\mathbf{X} =$

(X_1, X_2, X_3) covariates, some selected values of model parameters and censoring rates. The Clayton copula function by parameter $\theta = 3$ is assumed for the pair (X_1, X_2). Then, one can follow the such as Subsection 4.1. In Table 4, some criteria for comparing the fitted Cox-DART model and the Cox proportional hazard model (the root model) are reported.

Table 4: Simulation-based comparing of models; Here, the true parameters in terminal nodes are different.

α	Censor rate	Model	RMSE(S,$\hat{S}$)	R^2_{sch}	AIC
0.25	0.1	Root	0.25	0.06	11084
		Cox-DART	0.13	0.39	9372
	0.2	Root	0.26	0.08	10852
		Cox-DART	0.13	0.39	9175
	0.5	Root	0.26	0.10	10484
		Cox-DART	0.13	0.40	8864
0.5	0.1	Root	0.26	0.08	10654
		Cox-DART	0.05	0.51	8797
	0.2	Root	0.26	0.11	10097
		Cox-DART	0.05	0.52	8324
	0.5	Root	0.26	0.16	9175
		Cox-DART	0.05	0.52	7544
1	0.1	Root	0.26	0.11	9679
		Cox-DART	0.04	0.55	7945
	0.2	Root	0.26	0.17	8458
		Cox-DART	0.04	0.56	6905
	0.5	Root	0.27	0.27	6506
		Cox-DART	0.05	0.58	5255
2	0.1	Root	0.27	0.14	8076
		Cox-DART	0.04	0.56	6571
	0.2	Root	0.27	0.24	5979
		Cox-DART	0.05	0.59	4798
	0.5	Root	0.28	0.40	3143
		Cox-DART	0.06	0.64	2436
3	0.1	Root	0.26	0.15	7052
		Cox-DART	0.05	0.57	5705
	0.2	Root	0.27	0.26	4534
		Cox-DART	0.05	0.60	3589
	0.5	Root	0.28	0.46	1652
		Cox-DART	0.07	0.68	1229

From Tables 4, one can see the following:

1: The Cox-DART model dominates the classic Cox model according to RMSE, R^2_{sch} and AIC criteria.

2: RMSE and R^2_{sch} are increasing in the *censor rate* while AIC is decreasing in the *censor rate*.

3: As we expected, the AIC decreases as the censor rate increases, since the corresponding (partial) likelihood function decreases.

It is mentioned that the Weibull distribution with $\alpha = 1$ simplifies to the exponential distribution with the constant hazard rate function. For $\alpha > 1$ ($\alpha < 1$), the Weibull hazard function increases (decreases).

4.3 Assay of serum free light chain data set analyses

In this subsection, a Cox-DART model is fitted to an assay of serum free light chain (`flchain`) data set. It contains 7874 subjects and is available in the statistical software `R` by package "`survival`" (Therneau[29]). The objective of the study is to determine whether the free light chain (FLC) assay provides prognostic information relevant to the general population.

The explanatory variables are sex, age (in years), kappa (serum free light chain, kappa portion), lambda (serum free light chain, lambda portion) and creatinine (serum creatinine). The response "futime" is days from enrolment until death. The binary variable death is 1 if we observe the death occurs and 0 otherwise (that is if censoring occurs). A Cox model is fitted to this data set and the explanatory variables age, lambda and kappa are significant at level 0.05.

Figure 4 displays the scatter plot and the contour diagram of the joint empirical distribution for kappa and lambda. One can see that the association among kappa and lambda varies in different regions. The tail dependence estimators among kappa and lambda based on Equations (A.2) and (A.3) are obtained as $\lambda_L = 0.27$ and $\lambda_U = 0.57$. As mentioned in Subsection 2.2, the desired split points are equal to 20th quantiles of $(kappa, lambda)$, that is $\hat{q}_{kappa} = 0.89$ and $\hat{q}_{lambda} = 1.14$. Thus, the whole data set is partitioned into two subsets as

$$D_1 = \{D \mid (kappa, lambda) \in \hat{A}\}, \qquad D_2 = \{D \mid (kappa, lambda) \in \hat{A}^c\},$$

where $\hat{A} = \{kappa > 0.89, lambda > 1.14\}$. Using the Step 5 algorithm mentioned in Section 2, we conducted a log–rank test on the survival time based on two subsets A and A^c and obtained p-value≈ 0, which means that the splitting based on the regions A and A^c is significant and must be considered in the root of the decision tree.

The Kendall's tau correlation coefficients among kappa and lambda in the two regions $\hat{A}$ and $\hat{A}^c$ are obtained as $\hat{\tau}_{(kappa,lambda)|\hat{A}} = 0.81$ and $\hat{\tau}_{(kappa,lambda)|\hat{A}^c} = 0.07$, respectively. As

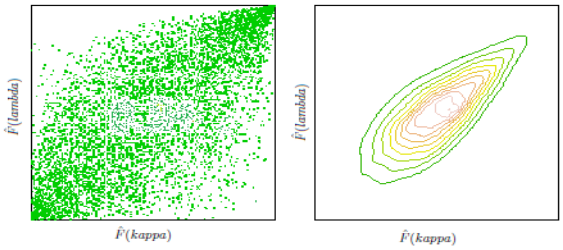
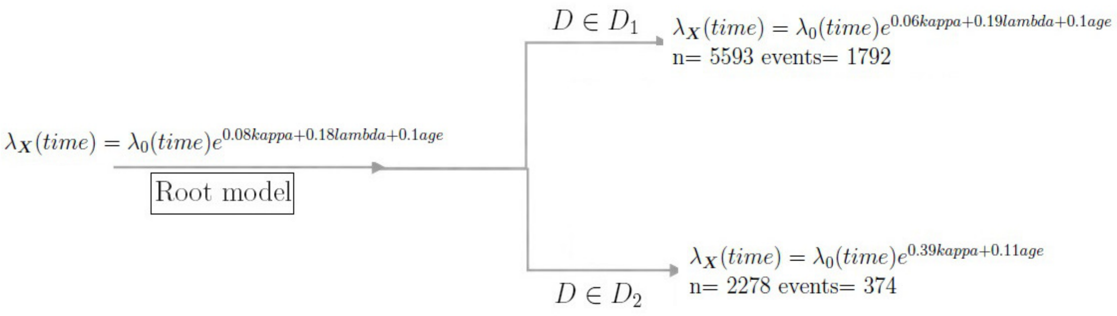


Figure 4: The scatter plot and contour diagram between empirical distribution *kappa* and *lambda*.

one can see, there is a strong dependency in $\hat{A}$ while a weak dependence in $\hat{A}^c$. The result of plot curves in the test the proportional hazards assumption for a Cox regression of the whole data set showed that the proportional hazards assumption model is appropriate. Therefore, a Cox–DART model is fitted to the data set and reported as follows:


$D \in D_1$
$\lambda_X(time) = \lambda_0(time)e^{0.06kappa+0.19lambda+0.1age}$
n= 5593 events= 1792
$\lambda_X(time) = \lambda_0(time)e^{0.08kappa+0.18lambda+0.1age}$
Root model
$\lambda_X(time) = \lambda_0(time)e^{0.39kappa+0.11age}$
n= 2278 events= 374
$D \in D_2$

(4.2)

In Table 5, some criteria for comparing the fitted Cox-DART model and the Cox proportional hazards model (the root model) (4.2) are given.

Table 5: Comparing Models for the assay of serum free light chain data set

Models	Sample size	Number of event	R^2_{sch}	AIC	$RMSE(\hat{S}_{KM}, \hat{S})$
Cox(Root)	7871	2166	0.30	34905	0.058
Cox-DART	$n_1 = 5593, n_2 = 2278$	$1792, 374$	0.80	32927	0.057

We see from Table 5, the Cox-DART model dominants the root regression model.

Remark 4.1. Note that one may consider various models in terminal nodes. For example, Boskabadi[3] proposed a non–parametric method, called *C-DART* model, where CART models are fitted in terminal nodes on the basis of samples without censoring observations. They approach may be used for samples including censored lifetime observations. To do this, the data sets is partioned properly according to Subsection 2.1 and then the extended CART defined by Equation (1.2) is implimented. In sequel, the CART and C-DART models were fitted to this data set with the turning parameter $\eta = 0.001$ in Equation (1.4). The results are obtained $RMSE_{CART}(futime, fut\hat{i}me) = 0.73$ and $RMSE_{C-DART}(futime, fut\hat{i}me) = 0.72$, also $R^2_{CART} = 0.46$ and $R^2_{C-DART} = 0.78$. So the DART approach provides more accurate results.

5 Conclusions and future research

In this article, we extended the DART approach in survival analyses. In addition, the Cox proportional hazards models are fitted in leaves of a regression tree with the DART approach and studied in details. By a simulation study and analysing a real data set, we have shown, that the DART approach can be successfully implemented in the survival setting. A diagram for construction of a DART model is given in Figure 5. This is helpful particularly for big data sets, where the relationship between the covariates may have some impacts on the survival times. This paper may be extended in various directions. Study on multivariate regression trees is another interesting topic.

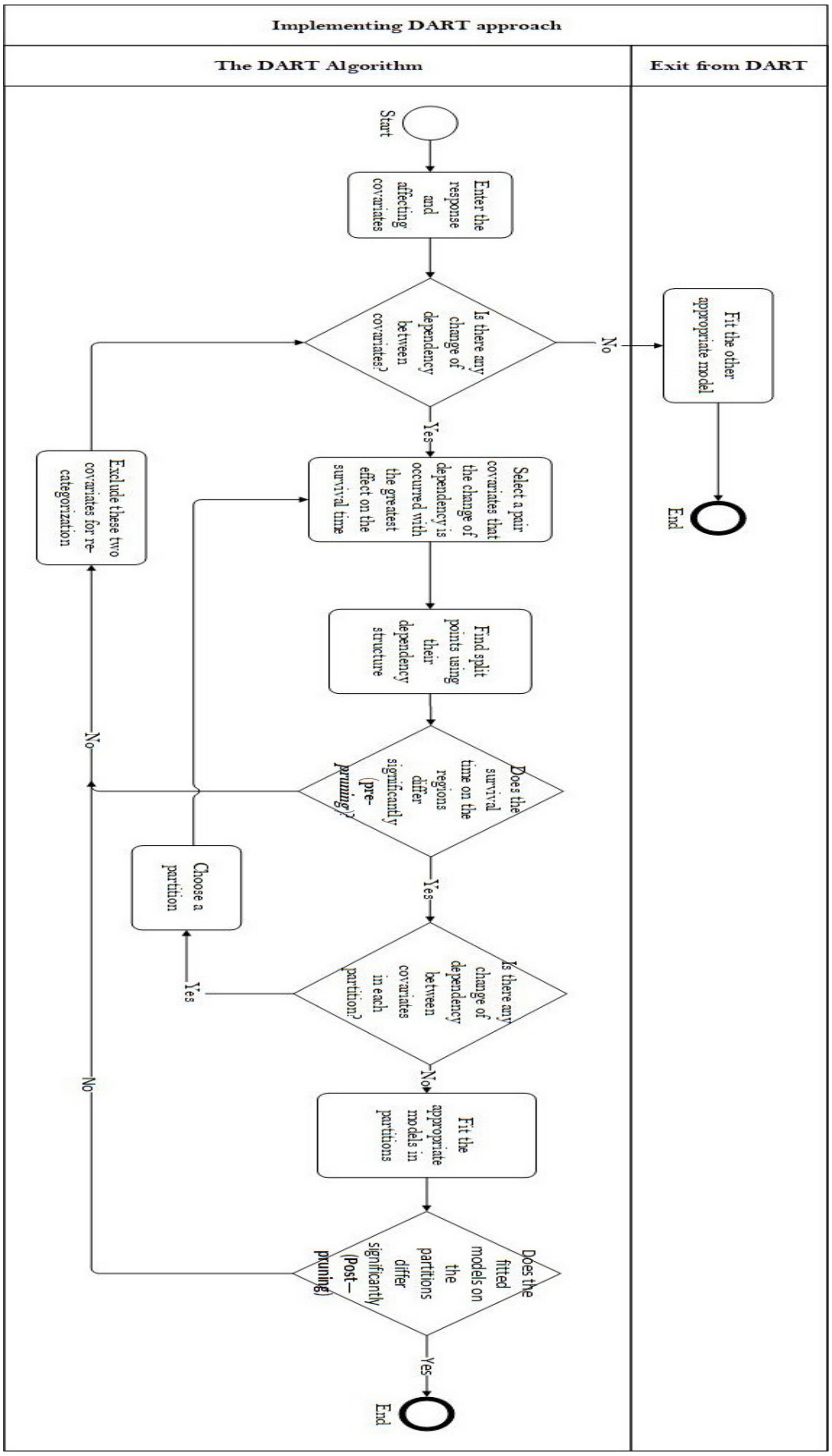

Implementing DART approach
The DART Algorithm
Exit from DART
Start
Enter the response and affecting covariates
Is there any change of dependency between covariates?
No
Fit the other appropriate model
End
Yes
Select a pair covariates that the change of dependency is occurred with the greatest effect on the survival time
Exclude these two covariates for re-categorization
Find split points using their dependency structure
Does the survival time on the regions differ significantly (pre-pruning)?
No
Yes
Choose a partition
Is there any change of dependency between covariates in each partition?
Yes
No
Fit the appropriate models in partitions
Does the fitted models on the partitions differ significantly (Post—pruning)
No
Yes
End

Figure 5: Process of implementing the DART approach.

A Copula function and dependence coefficients

Definition A.1. (Nelsen[23]) Let (X_1, Y_1) and (X_2, Y_2) be two independent and identically distributed random vectors each with distribution function H. The population version of " Kendall's tau correlation" is defined as the probability of concordance minus the probability of discordance as

$$\begin{aligned} \tau_{X,Y} &= P((X_1 - X_2)(Y_1 - Y_1) > 0) - P((X_1 - X_2)(Y_1 - Y_1) < 0) \\ &= E\{sgn\,((X_1 - X_2)(Y_1 - Y_2))\}, \end{aligned} \tag{A.1}$$

where "$sgn(u)$" is the sign of u, -1 for $u < 0$, 0 for $u = 0$, and 1 for $u > 0$. Tsai[30] introduced the conditional Kendall's τ as

$$\begin{aligned} \tau_{X,Y|A} &= E\{sgn\,((X_1 - X_2)(Y_1 - Y_2)) \mid A\} \\ &= E\{sgn\,((X_1 - X_2)(Y_1 - Y_2))\, I_A\}/P(A), \end{aligned}$$

where A denotes a given event. Following Randles[25], an estimator for the conditional Kendall's τ over the set A is

$$\hat{\tau}_{X,Y|A} = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \{sgn((X_i - X_j)(Y_i - Y_j))\} I_A,$$

where $M = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I_A$.

Definition A.2. Let X_1 and X_2 be two random variables with marginal functions $F_1(x) = P(X_1 \leq x)$ and $F_2(x) = P(X_2 \leq x)$, respectively, with the joint distribution functions $F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$. The lower and upper tail dependency coefficients are defined by

$$\lambda_L = \lim_{t \to 0^+} P(X_1 \leq F_1^{-1}(t) \mid X_2 \leq F_2^{-1}(t)) \tag{A.2}$$

and

$$\lambda_U = \lim_{t \to 1^-} P(X_1 > F_1^{-1}(t) \mid X_2 > F_2^{-1}(t)), \tag{A.3}$$

respectively, where $F_i^{-1}(x) = \inf\{t : F(t) \geq x\}$ $(i = 1, 2)$ is the inverse function of the marginal distribution $F_i(x)$.

Here, we review some properties of copula. For more information, see Nelsen (2006).

Theorem A.3 (Sklar, 1959)**.** *Let H be a joint distribution function with margins F and G. Then, there exists a copula C such that*

$$H(x, y) = C(F(x), G(y)) \quad \forall x, y \in R. \tag{A.4}$$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on $Ran(F) \times Ran(G)$, where $Ran(F)$ and $Ran(G)$ are ranges of distributions F and G, respectively. Conversely, if C is a copula and F and G are distribution functions, then the function H defined by (A.4) is a joint distribution function with margins F and G.

Remark A.4. From (A.4), we have

$$C(u,v) = H(F^{-1}(u), G^{-1}(v)), \qquad \forall u, v \in [0,1].$$

Theorem A.5. *Let $C(u,v)$ be the copula of X_1 and X_2. If the limit exists, then*

$$\lambda_L = \lim_{t \to 0^+} \frac{C(t,t)}{t}, \tag{A.5}$$

and

$$\lambda_U = 2 - \lim_{t \to 1^-} \frac{1 - C(t,t)}{1-t}. \tag{A.6}$$

Remark A.6. If the copula function is unknown, one may use the non–parametric inversion tail dependence (Schmidt[26]). Let $R_i^{x_1}$ and $R_i^{x_2}$ denote the rank of X_{1i} and X_{2i}, $i = 1, ..., m$, respectively. Then C_m is an experimental copula function. The first set of estimators are based on Equations (A.5) and (A.6), as

$$\hat{\lambda}_L := \frac{m}{k} C_m \left(\frac{kx}{m}, \frac{ky}{m} \right) \approx \frac{1}{k} \sum_{i=1}^{n} I_{\{R_i^{x1} \le kx, R_i^{y1} \le ky\}} \tag{A.7}$$

and

$$\hat{\lambda}_U := \frac{m}{k} C_m \left((1 - \frac{kx}{m}, 1], (1 - \frac{ky}{m}, 1] \right) \approx \frac{1}{k} \sum_{i=1}^{n} I_{\{R_i^{x1} > n-kx, R_i^{y1} > n-ky\}}, \tag{A.8}$$

with a parameter $k \in \{1, 2, \ldots, n\}$, chosen by the researcher; For a greater detail, see Schmidt[26].

References

[1] Ambler, G., Seaman, S., Omar, R. Z. An evaluation of penalised survival methods for developing prognostic models with rare events. *Statistics in Medicine*, **31** (2011) 1150–1161.

[2] Bertolet, M., Brooks, M. M., and Bittner, V. Tree-based identification of subgroups for time-varying covariate survival data. *Statistical Methods in Medical Research,* (2012) doi:10.1177/0962280212460442.

[3] Boskabadi, M., Doostparast, M. Regression trees with splitting based on changes of dependencies among covariates. *Intelligent Data Analysis, Accepted,* (2020).

[4] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. *Classification and Regression Trees*, CRC Press; New York, (1984).

[5] Burnham, K. P., Anderson, D. R. *Model selection and multimodel inference: A practical information-theoretic approach.* New York: Springer, (2002).

[6] Butler, J. H., Gilpin, E., Gordon, L., Olshen, R. A. Tree-structured survival analyses, ii. *Technical Report* 133, Division of Biostatistics, Stanford University, Stanford University, (1989).

[7] Chipman, H. A., George, E. I., McCulloch, R. E. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, **4** (2010) 266–98.

[8] Chipman, H. A., McCulloch, R. E., Dorie, V. dbarts: discrete Bayesian additive regression trees sampler, (Available from: http:// lib.stat.cmu.edu /R/ CRAN/ web/ packages/dbarts/index.html) [Accessed on 28 January 2016].

[9] Ciampi, A., Chang, C. H., Hogg, S., McKinney, S. Recursive partition: A versatile method for exploratory data analysis in biostatistics, *Biostatistics*, **16** (1987) 23–50.

[10] Cichosz, P. Data Mining Algorithms: Explained Using R, John Wiley and Sons, Ltd., West Sussex, United Kingdom, (2015).

[11] Cox, D. R. Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, **34** (1972) 187–220.

[12] Goldman, L., Weinberg, M., Weisberg, M., Olshen, R., et al. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. *New England Journal of Medicine*, **307** (1982) 588–596.

[13] Gordon, L., Olshen, R. A. Tree-structured survival a analysis. *Cancer Treatment Reports*, **69** (1985) 1065–1069.

[14] Hosmer, D. W., and Lemeshow, S. *Applied Survival Analysis*, New York, John Wiley and Sons, (1999).

[15] Hothorn, T., Hornik, K., and Zeileis, A. Unbiased recursive partitioning: A conditional inference framework.*Journal of Computational and Graphical Statistics*, **15** (2006) 651–674.

[16] Huang, X., Chen, S., Soong, S. j. Piecewise exponential survival trees with time-dependent covariates. *Biometrics*, **54** (1998) 1420–1433.

[17] Kass, G. V. An exploratory technique for investigating large quantities of categorical data. *Annals of Applied Statistics*, **29** (1980) 119–127.

[18] Klein, J. P. and Moeschberger, M. L. *Survival Analysis Techniques for Censored and truncated data*,Springer. Lagakos et al. Biometrika **68** (1981) (1997) 515-523.

[19] LeBlanc, M. and Crowley, J. Relative risk trees for censored survival data, *Biometrics*, **29** (1992) 411–425.

[20] Loh, W. Y. Regression trees with unbiased variable selection and interaction detection.*Statistica Sinica*, **12** (2002) 361–386.

[21] Loh, W. Y., Shih, Y. S. Split selection methods for classification trees. *Statistica Sinica*, **7** (1997) 815–840.

[22] Morgan, J. N., Sonquist, J. A. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association,* **58** (1963) 415–434.

[23] Nelsen R. B. *An Introduction to Copulas.* J. Springer Science, (2006).

[24] Quinlan, J. R. Induction of decision trees. *Machine Learning*, **1** (1986) 81–106.

[25] Randles, R. H., Wolfe, D. A. *Introduction to the theory of non–parametric statistics,* Malabar, FL: Krieger, (1991).

[26] Schmidt, R., Stadtm. *Nonparametric estimation of tail dependence.J. The Scandinavian Journal of Statistics,* **33** (2003) 307–335.

[27] Schoenfeld, D. Partial Residuals for The Proportional Hazards Regression Model.*Biometrika,* **69** (1982) 239–241.

[28] Segal, M. R. Regression trees for censored data. *Biometrics,* **33** (1988) 35–47.

[29] Therneau, T. A Package for Survival Analysis in S. version 2.38, URL: https://CRAN.R-project.org/package=survival, (2015).

[30] Tsai, W. Y. Testing the assumption of independence of truncation time and failure time, *Biometrika,* **77** (1990) 169–177.

[31] Wallace, M. Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. , *Statistics in Medicine,* **33** (2014) 4790–4804.

[32] Xu, R., Adak, S. Survival analysis with time-varying regression effects using a tree-based approach, *Biometrika,* **58** (2002) 305–315.