

Machine Learning for Survival Analysis: A Survey

PING WANG, Virginia Tech

YAN LI, University of Michigan, Ann Arbor

CHANDAN K. REDDY, Virginia Tech

Accurately predicting the time of occurrence of an event of interest is a critical problem in longitudinal data analysis. One of the main challenges in this context is the presence of instances whose event outcomes become unobservable after a certain time point or when some instances do not experience any event during the monitoring period. Such a phenomenon is called censoring which can be effectively handled using survival analysis techniques. Traditionally, statistical approaches have been widely developed in the literature to overcome this censoring issue. In addition, many machine learning algorithms are adapted to effectively handle survival data and tackle other challenging problems that arise in real-world data. In this survey, we provide a comprehensive and structured review of the representative statistical methods along with the machine learning techniques used in survival analysis and provide a detailed taxonomy of the existing methods. We also discuss several topics that are closely related to survival analysis and illustrate several successful applications in various real-world application domains. We hope that this paper will provide a more thorough understanding of the recent advances in survival analysis and offer some guidelines on applying these approaches to solve new problems that arise in applications with censored data.

CCS Concepts: • **Mathematics of computing** → **Survival analysis**; • **Computing methodologies** → **Machine learning**; • **Information systems** → **Data mining**;

Additional Key Words and Phrases: Longitudinal data; censoring; survival analysis; regression; hazard rate; Cox model; Concordance index.

ACM Reference Format:

Ping Wang, Yan Li, Chandan K. Reddy, 2017. Machine Learning for Survival Analysis: A Survey. *ACM Comput. Surv.* 1, 1, Article 1 (March 2017), 38 pages.

DOI: 0000001.0000001

1. INTRODUCTION

Due to the development of various data acquisition and big data technologies, the ability to collect a wide variety of data and monitor the observation over long-term periods have been attained in different disciplines. For most of the real-world applications, the primary objective of monitoring these observations is to obtain a better estimate of the time of occurrence of a particular event of interest. One of the main challenges for such time-to-event data is that usually there exist censored instances, i.e., the event of interests is not observed for these instances due to either the time limitation of the study period or losing track during the longitudinal observation period. More precisely, certain instances have experienced event (or labeled as event) and the information about

This material is based upon work supported by, or in part by, the U.S. National Science Foundation grants IIS-1707498, IIS-1619028, IIS-1646881.

Author's addresses: P. Wang is with the Department of Computer Science, Virginia Tech, 900 N. Glebe Road, Arlington, VA, 22203. E-mail: ping@vt.edu; Y. Li is with the Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI, 48109. E-mail: yanli@umich.edu; C. K. Reddy is with the Department of Computer Science, Virginia Tech, 900 N. Glebe Road, Arlington, VA, 22203. E-mail: reddy@cs.vt.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2017 ACM. 0360-0300/2017/03-ART1 \$15.00

DOI: 0000001.0000001

the outcome variable for the remaining instances is only available until a specific time point in the study. Therefore, it is not suitable to directly apply predictive algorithms using the standard statistical and machine learning approaches to analyze the longitudinal data. Survival analysis, which is an important subfield of statistics, provides various mechanisms to handle such censored data problems that arise in modeling such longitudinal data (also referred to as time-to-event data when modeling a particular event of interest is the main objective of the problem) which occurs ubiquitously in various real-world application domains.

In addition to the difficulty in handling the censored data, there are also several unique challenges to perform the predictive modeling with such survival data and hence several researchers have, more recently, developed new computational algorithms for effectively handling such complex challenges. To tackle such practical concerns, some related works have adapted several machine learning methods to solve the survival analysis problems and machine learning researchers have developed more sophisticated and effective algorithms which either complement or compete with the traditional statistical methods. In spite of the importance of these problems and relevance to various real-world applications, this research topic is scattered across different disciplines. Moreover, there are only few surveys that are available in the literature on this topic and, to the best of our knowledge, there is no comprehensive review paper about survival analysis and its recent developments from a machine learning perspective. Almost all of these existing survey articles describe solely statistical methods and either completely ignore or barely mention the machine learning advancements in this research field. One of the earliest surveys may be found in [Chung et al. 1991], which gives an overview of the statistical survival analysis methods and describes its applications in criminology by predicting the time until recidivism. Most of the existing books about survival analysis [Kleinbaum and Klein 2006; Lee and Wang 2003; Allison 2010] focus on introducing this topic from the traditional statistical perspective instead of explaining from the machine learning standpoint. Recently, the authors in [Cruz and Wishart 2006] and [Kourou et al. 2015] discussed the applications in cancer prediction and provided a comparison of several machine learning techniques.

The primary purpose of this survey article is to provide a comprehensive and structured overview of various machine learning methods for survival analysis along with the traditional statistical methods. We demonstrate the commonly used evaluation metrics and advanced related formulations that are commonly investigated in this research topic. We will discuss a detailed taxonomy of all the survival analysis methods that were developed in the traditional statistics as well as more recently in the machine learning community. We will also provide links to various implementations and sources codes which will enable the readers to further dwell into the methods discussed in this article. Finally, we will discuss various applications of survival analysis.

The rest of this paper is organized as follows. We will give a brief review of the basic concepts, notations and definitions that are necessary to comprehend the survival analysis algorithms and provide the formal problem statement for survival analysis problem in Section 2. A taxonomy of the existing survival analysis methods, including both statistical and machine learning methods will also be provided to elucidate the holistic view of the existing works in the area of survival analysis. We will then review the well-studied representative conventional statistical methods including non-parametric, semi-parametric, and parametric models in Section 3. Section 4 describes several basic machine learning approaches, including survival trees, Bayesian methods, support vector machines and neural networks developed for survival analysis. Different kinds of advanced machine learning algorithms such as ensemble learning, transfer learning, multi-task learning and active learning for handling survival data will also be discussed. Section 5 demonstrates the evaluation metrics for survival mod-

els. In addition to the survival analysis algorithms, some interesting topics related to this topic have received considerable attention in various fields. In Section 6, several related concepts such as early prediction and complex events will be discussed. Various data transformation techniques such as uncensoring and calibration which are typically used in conjunction with existing predictive methods will also be mentioned briefly. A discussion about topics in complex event analysis such as competing risks and recurrent events will also be provided. In Section 7, various real-world applications of survival analysis methods will be briefly explained and more insights into these application domains will be provided. In Section 8, the details about the implementations and software packages of the survival analysis methods are discussed. Finally, Section 9 concludes our discussion.

2. DEFINITION OF SURVIVAL ANALYSIS

In this section, we will first provide the basic notations and terminologies used in this paper. We will then give an illustrative example which explains the structure of the survival data and give a more formal problem statement for survival analysis. At last, we also give a complete taxonomy of the existing survival analysis methods that are available in the literature, including both the conventional statistical methods and the machine learning approaches. It provides a holistic view of the field of survival analysis and will aid the readers to gain the basic knowledge about the methods used in this field before getting into the detailed algorithms.

Table I: Notations used in this paper.

Notations	Descriptions
P	The number of features
N	The number of instances
X	$\mathbb{R}^{N \times P}$ feature vector
X_i	$\mathbb{R}^{1 \times P}$ covariate vector of instance i
T	$\mathbb{R}^{N \times 1}$ vector of event times
C	$\mathbb{R}^{N \times 1}$ vector of last follow up times
y	$\mathbb{R}^{N \times 1}$ vector of observed time which is equal to $\min(T, C)$
δ	$N \times 1$ binary vector for event status
β	$\mathbb{R}^{P \times 1}$ coefficient vector
$f(t)$	Death density function
$F(t)$	Cumulative event probability function
$S(t)$	Survival probability function
$h(t)$	Hazard function
$h_0(t)$	Baseline hazard function
$H(t)$	Cumulative hazard function

2.1. Survival Data and Censoring

During the study of a survival analysis problem, it is possible that the events of interest are not observed for some instances; this scenario occurs because of the limited observation time window or missing traces caused by other uninterested events. This concept is known as censoring [Klein and Moeschberger 2005].

For a survival problem, the time to the event of interest (T) is known precisely only for those instances who have the event occurred during the study period. For the remaining instances, since we may lose track of them during the observation time or

their time to event is greater than the observation time, we can only have the censored time (C) which may be the time of withdrawn, lost or the end of the observation. They are considered to be censored instances in the context of survival analysis. In other words, here, we can only observe either survival time (T_i) or censored time (C_i) but not both, for any given instance i . If and only if $y_i = \min(T_i, C_i)$ can be observed during the study, the dataset is said to be right-censored, which is a common scenario that arises in many practical problems [Marubini and Valsecchi 2004].

In Figure 1, an illustrative example is given for a better understanding of the definition of censoring and the structure of survival data. Six instances are observed in this longitudinal study for 12 months and the event occurrence information during this time period is recorded. From Figure 1, we can find that only subjects S4 and S6 have experienced the event (marked by 'X') during the follow-up time and the observed time for them is the event time. While the event did not occur within the 12 months period for subjects S1, S2, S3 and S5, which are considered to be censored and marked by red dots in the figure. More specifically, subjects S2 and S5 are censored since there was no event occurred during the study period, while subjects S1 and S3 are censored due to the withdrawal or being lost to follow-up within the study time period.

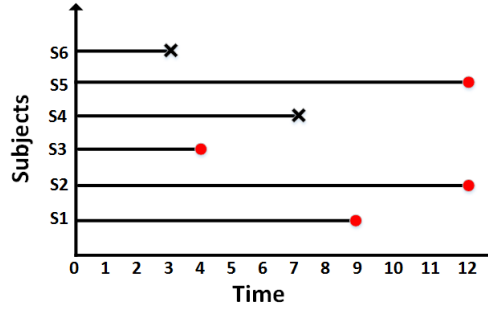


Fig. 1: An illustration to demonstrate the survival analysis problem.

Problem Statement: For a given instance i , represented by a triplet (X_i, y_i, δ_i) , where $X_i \in \mathbb{R}^{1 \times P}$ is the feature vector; δ_i is the binary event indicator, i.e., $\delta_i = 1$ for an uncensored instance and $\delta_i = 0$ for a censored instance; and y_i denotes the observed time and is equal to the survival time T_i for an uncensored instance and C_i otherwise, i.e.,

$$y_i = \begin{cases} T_i & \text{if } \delta_i = 1 \\ C_i & \text{if } \delta_i = 0 \end{cases} \quad (1)$$

It should be noted that T_i is a latent value for censored instances since these instances did not experience any event during the observation time period.

The goal of survival analysis is to estimate the time to the event of interest T_j for a new instance j with feature predictors denoted by X_j . It should be noted that, in survival analysis problem, the value of T_j will be both non-negative and continuous.

2.2. Survival and Hazard Function

The **survival function**, which is used to represent the probability that the time to the event of interest is not earlier than a specified time t [Lee and Wang 2003; Klein and Moeschberger 2005], is one of the primary goals in survival analysis. Conventionally,

survival function is represented by S , which is given as follows:

$$S(t) = \Pr(T \geq t). \quad (2)$$

The survival function monotonically decreases with t , and the initial value is 1 when $t = 0$, which represents the fact that, in the beginning of the observation, 100% of the observed subjects survive; in other words, none of the events of interest have occurred.

On the contrary, the *cumulative death distribution function* $F(t)$, which represents the probability that the event of interest occurs earlier than t , is defined as $F(t) = 1 - S(t)$, and **death density function** can be obtained as $f(t) = \frac{d}{dt}F(t)$ for continuous cases, and $f(t) = [F(t + \Delta t) - F(t)]/\Delta t$, where Δt denotes a small time interval, for discrete cases. Figure 2 shows the relationship among these functions.

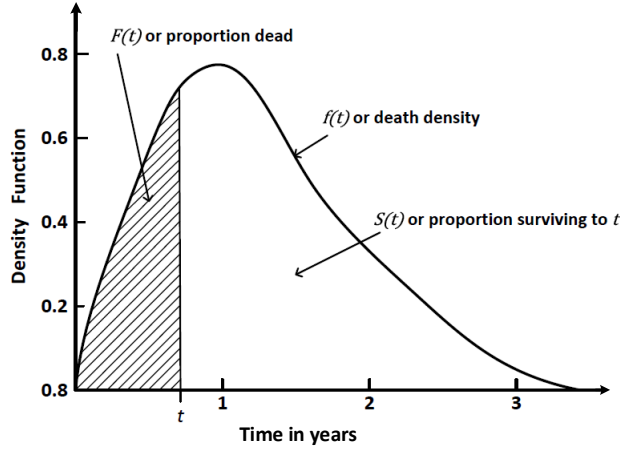


Fig. 2: Relationship among different entities $f(t)$, $F(t)$ and $S(t)$.

In survival analysis, another commonly used function is the **hazard function** ($h(t)$), which is also called the *force of mortality*, the *instantaneous death rate* or the *conditional failure rate* [Dunn and Clark 2009]. The hazard function does not indicate the chance or probability of the event of interest, but instead it is the rate of event at time t given that no event occurred before time t . Mathematically, the hazard function is defined as:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)} \quad (3)$$

Similar to $S(t)$, $h(t)$ is also a non-negative function. While all the survival functions, $S(t)$, decrease over time, the hazard function can have a variety of shapes. Consider the definition of $f(t)$, which can also be expressed as $f(t) = -\frac{d}{dt}S(t)$, so the hazard function can be represented as:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}S(t) \cdot \frac{1}{S(t)} = -\frac{d}{dt}[\ln S(t)]. \quad (4)$$

Thus, the survival function defined in Eq. (2) can be rewritten as

$$S(t) = \exp(-H(t)) \quad (5)$$

where $H(t) = \int_0^t h(u)du$ represents the *cumulative hazard function* (CHF) [Lee and Wang 2003].

2.3. Taxonomy of Survival Analysis methods

Broadly speaking, the survival analysis methods can be classified into two main categories: statistical methods and machine learning based methods. Statistical methods share the common goal with machine learning methods to make predictions of the event time. However, they focus more on both the distributions of the event times and the statistical properties of the parameter estimation. Machine learning methods are usually applied to the high-dimensional problems, while statistical methods are generally developed for the low-dimensional data. In addition, machine learning methods for survival analysis offer more effective algorithms by incorporating survival problems with both statistical methods and machine learning methods and taking advantages of the recent developments in machine learning and optimization to learn the dependencies between covariates and survival times in different ways.

Based on the assumptions and the usage of the parameters used in the model, the traditional statistical methods can be subdivided into three categories: (i) non-parametric models, (ii) semi-parametric models and (iii) parametric models. Machine learning algorithms, such as survival trees, Bayesian methods, neural networks and support vector machines, which have become more popular in the recent years are included under a separate branch. Several advanced machine learning methods, including ensemble learning, active learning, transfer learning and multi-task learning methods, are also included. The overall taxonomy also includes some of the research topics that are related to survival analysis such as complex events, data transformation and early prediction. A complete taxonomy of these survival analysis methods is shown in Figure 3.

3. TRADITIONAL STATISTICAL METHODS

In this section, we will introduce three different types of statistical methods to estimate the survival/hazard functions: non-parametric, semi-parametric and parametric methods. Table II shows both the advantages and disadvantages of each type of methods based on theoretical and experimental analysis and lists the specific methods in each type.

In non-parametric methods, an empirical estimate of the survival function is obtained using Kaplan-Meier (KM) method, Nelson-Aalen estimator (NA) or Life-Table (LT) method. More generally, any KM estimator for the survival probability at the

Table II: Summary of different types of statistical methods for survival analysis.

Type	Advantages	Disadvantages	Specific methods
Non-parametric	More efficient when no suitable theoretical distributions known.	Difficult to interpret; yields inaccurate estimates.	Kaplan-Meier Nelson-Aalen Life-Table
Semi-parametric	The knowledge of the underlying distribution of survival times is not required.	The distribution of the outcome is unknown; not easy to interpret.	Cox model Regularized Cox CoxBoost Time-Dependent Cox
Parametric	Easy to interpret, more efficient and accurate when the survival times follow a particular distribution.	When the distribution assumption is violated, it may be inconsistent and can give sub-optimal results.	Tobit Buckley-James Penalized regression Accelerated Failure Time

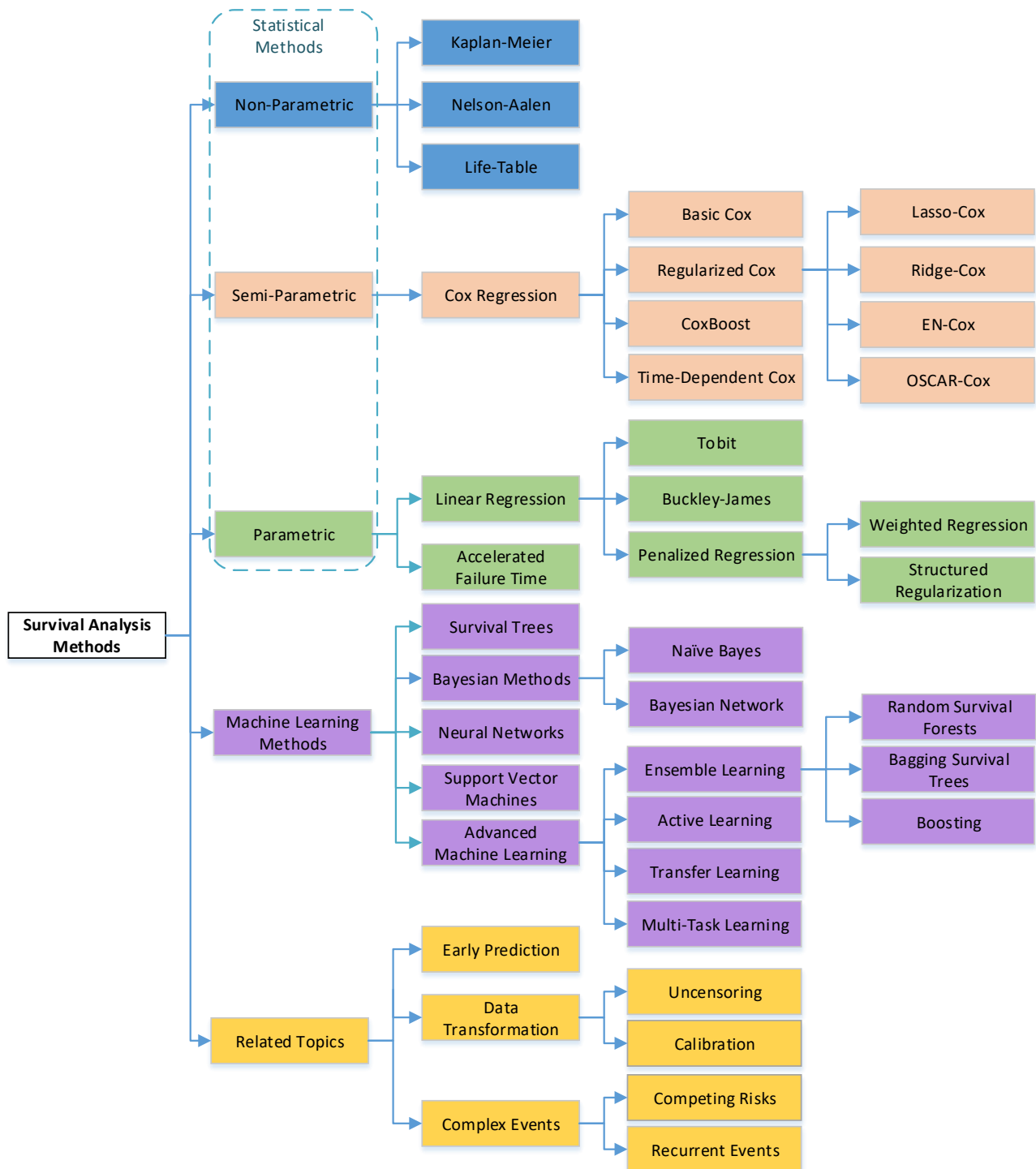


Fig. 3: Taxonomy of the methods developed for survival analysis.

specified survival time is a product of the same estimate up to the previous time and the observed survival rate for that given time. Thus, KM method is also referred to as a *product-limit method* [Kaplan and Meier 1958; Lee and Wang 2003]. NA method is an estimator based on modern counting process techniques [Andersen et al. 2012]. LT [Cutler and Ederer 1958] is the application of the KM method to the interval grouped survival data.

Under the semi-parametric category, Cox model is the most commonly used regression analysis approach for survival data and it is built on the proportional hazards assumption and employs partial likelihood for parameter estimation. Cox regression method is described as semi-parametric method since the distribution of the outcome remains unknown even if it is based on a parametric regression model. In addition, several useful variants of the basic Cox model, such as penalized Cox models, Cox-Boost algorithm and Time-Dependent Cox model (TD-Cox), are also proposed in the literature.

Parametric methods are more efficient and accurate for estimation when the time to the event of interest follows a particular distribution specified in terms of certain parameters. It is relatively easy to estimate the times to the event of interest with parametric models, but it becomes awkward or even impossible to do so with the Cox model [Allison 2010]. Linear regression method is one of the main parametric survival methods, while the Tobit model, Buckley-James regression model and the penalized regression are the most commonly used linear models for survival analysis. In addition, other parametric models, such as accelerated failure time (AFT) which models the survival time as a function of covariates [Kleinbaum and Klein 2006], are also widely used. We will now describe these three types of statistical survival methods in this section.

3.1. Non-parametric Models

Among all functions, the survival function or its graphical presentation is the most widely used one. In 1958, Kaplan and Meier [Kaplan and Meier 1958] developed the Kaplan-Meier (KM) Curve or the product-limit (PL) estimator to estimate the survival function using the actual length of the observed time. This method is the most widely used one for estimating survival function. Let $T_1 < T_2 < \dots < T_K$ be a set of distinct ordered event times observed for N ($K \leq N$) instances. In addition to these event times, there are also censoring times for instances whose event times are not observed. For a specific event time T_j ($j = 1, 2, \dots, K$), the number of observed events is $d_j \geq 1$, and r_j instances will be considered to be “at risk” since their event time or censored time is greater than or equal to T_j . It should be noted that we cannot simply consider r_j as the difference between r_{j-1} and d_{j-1} due to the censoring. The correct way to obtain r_j is $r_j = r_{j-1} - d_{j-1} - c_{j-1}$, where c_{j-1} is the number of censored instances during the time period between T_{j-1} and T_j . Then the conditional probability of surviving beyond time T_j can be defined as:

$$p(T_j) = \frac{r_j - d_j}{r_j} \quad (6)$$

Based on this conditional probability, the product-limit estimate of survival function $S(t) = P(T \geq t)$ is given as follows:

$$\hat{S}(t) = \prod_{j: T_j < t} p(T_j) = \prod_{j: T_j < t} \left(1 - \frac{d_j}{r_j}\right) \quad (7)$$

However, if the subjects in the data are grouped into some interval periods according to the time, or if the number of subjects is very large, or when the study is for

a large population, the Life Table (LT) analysis [Cutler and Ederer 1958] will be a more convenient method. Different from KM and LT method, Nelson-Aalen estimator [Nelson 1972; Aalen 1978] is a method to estimate the cumulative hazard function for censored data based on counting process approach. It should be noted that when the time to event of interest follows a specific distribution, nonparametric methods are less efficient compared to the parametric methods.

3.2. Semi-Parametric Models

As a hybrid of the parametric and non-parametric approaches, semi-parametric models can obtain a more consistent estimator under a broader range of conditions compared to parametric models, and a more precise estimator than the non-parametric methods [Powell 1994]. Cox model [David 1972] is the most commonly used survival analysis method in this category. Unlike parametric methods, the knowledge of the underlying distribution of time to event of interest is not required, but the attributes are assumed to have an exponential influence on the outcome. We will now discuss the details of Cox model more elaborately and then describe different variants and extensions of the basic Cox model such as regularized Cox, CoxBoost and Time-Dependent Cox.

3.2.1. The Basic Cox Model. For a given instance i , represented by a triplet (X_i, y_i, δ_i) , the hazard function $h(t, X_i)$ in the Cox model follows the proportional hazards assumption given by

$$h(t, X_i) = h_0(t) \exp(X_i \beta), \quad (8)$$

for $i = 1, 2, \dots, N$, where the *baseline hazard function*, $h_0(t)$, can be an arbitrary non-negative function of time, $X_i = (x_{i1}, x_{i2}, \dots, x_{iP})$ is the corresponding covariate vector for instance i , and $\beta^T = (\beta_1, \beta_2, \dots, \beta_P)$ is the coefficient vector. The Cox model is a semi-parametric algorithm since the baseline hazard function, $h_0(t)$, is unspecified. For any two instances X_1 and X_2 , the hazard ratio is given by

$$\frac{h(t, X_1)}{h(t, X_2)} = \frac{h_0(t) \exp(X_1 \beta)}{h_0(t) \exp(X_2 \beta)} = \exp[(X_1 - X_2) \beta]. \quad (9)$$

which means that the hazard ratio is independent of the baseline hazard function. Cox model is a proportional hazards model since the hazard ratio is a constant and all the subjects share the same baseline hazard function. Based on this assumption, the survival function can be computed as follows:

$$S(t) = \exp(-H_0(t) \exp(X \beta)) = S_0(t) \exp(X \beta) \quad (10)$$

where $H_0(t)$ is the *cumulative baseline hazard function*, and $S_0(t) = \exp(-H_0(t))$ represents the baseline survival function. The Breslow's estimator [Breslow 1972] is the most widely used method to estimate $H_0(t)$, which is given by

$$\hat{H}_0(t) = \sum_{t_i \leq t} \hat{h}_0(t_i) \quad (11)$$

where $\hat{h}_0(t_i) = 1 / \sum_{j \in R_i} e^{X_j \beta}$ if t_i is an event time, otherwise $\hat{h}_0(t_i) = 0$. Here, R_i represents the set of subjects who are at risk at time t_i .

Because the baseline hazard function $h_0(t)$ in Cox model is not specified, it is not possible to fit the model using the standard likelihood function. In other words, the hazard function $h_0(t)$ is a nuisance function, while the coefficients β are the parameters of interest in the model. To estimate the coefficients, Cox proposed a partial likelihood [David 1972; 1975] which depends only on the parameter of interest β and is

free of the nuisance parameters. The hazard function refers to the probability that an instance with covariate X fails at time t on the condition that it survives until time t can be expressed by $h(t, X)dt$ with $dt \rightarrow 0$. Let J ($J \leq N$) be the total number of events of interest that occurred during the observation period for N instances, and $T_1 < T_2 < \dots < T_J$ is the distinct ordered time to event of interest. Without considering the ties, let X_j be the corresponding covariate vector for the subject who fails at T_j , and R_j be the set of risk subjects at T_j . Thus, conditional on the fact that the event occurs at T_j , the individual probability corresponding to covariate X_j can be formulated as

$$\frac{h(T_j, X_j)dt}{\sum_{i \in R_j} h(T_j, X_i)dt} \quad (12)$$

and the partial likelihood is the product of the probability of each subject; referring to the Cox assumption and the presence of the censoring, the partial likelihood is defined as

$$L(\beta) = \prod_{j=1}^N \left[\frac{\exp(X_j \beta)}{\sum_{i \in R_j} \exp(X_i \beta)} \right]^{\delta_j} \quad (13)$$

It should be noted that here $j = 1, 2, \dots, N$; if $\delta_j = 1$, the j^{th} term in the product is the conditional probability; otherwise, when $\delta_j = 0$, the corresponding term is 1, which means that the term will not have any effect on the final product. The coefficient vector $\hat{\beta}$ is estimated by maximizing this partial likelihood, or equivalently, minimizing the negative *log-partial likelihood* for improving efficiency.

$$LL(\beta) = - \sum_{j=1}^N \delta_j \{X_j \beta - \log[\sum_{i \in R_j} \exp(X_i \beta)]\}. \quad (14)$$

The maximum partial likelihood estimator (MPLE) [David 1972; Lee and Wang 2003] can be used along with the numerical Newton-Raphson method [Kelley 1999] to iteratively find an estimator $\hat{\beta}$ which minimizes $LL(\beta)$.

3.2.2. Regularized Cox models. With the development of data collection and detection techniques, most real-world domains tend to encounter high-dimensional data with lots of features. In some cases, the number of variables (P) in the given data is almost equal to or even exceeds the number of instances (N). It is challenging to build the prediction model with all the features and the model might provide inaccurate results because of the overfitting problem [van Houwelingen and Putter 2011]. This motivates using sparsity norms to select vital features in high-dimension under the assumption that most of the features are not significant [Friedman et al. 2001]. For the purpose of identifying the most relevant features to the outcome variable among tens of thousands of features, different penalty functions, including lasso, group lasso, fused lasso and graph lasso, are also used to develop the prediction models using the sparse learning methods. The family of ℓ -norm penalty functions $\ell_\gamma : \mathbb{R}^\gamma \rightarrow \mathbb{R}$, with the form of $\ell_\gamma(\beta) = \|\beta\|_\gamma = (\sum_{i=1}^P \|\beta_i\|^\gamma)^{\frac{1}{\gamma}}$, $\gamma > 0$ are the commonly used penalty functions. The smaller the value of γ , the sparser the solution, but when $0 \leq \gamma < 1$, the penalty is non-convex, which makes the optimization problem more challenging to solve. Here, we will introduce the commonly used regularized Cox models, whose regularizers are summarized in Table III.

Lasso-Cox: Lasso [Tibshirani 1996] is a ℓ_1 -norm regularizer which is good at performing feature selection and estimating the regression coefficients simultaneously.

Table III: Different regularizers used in the variants of Cox model.

Regularized Cox models	Regularizers
Lasso-Cox	$\lambda \sum_{p=1}^P \beta_p $
Ridge-Cox	$\frac{\lambda}{2} \sum_{p=1}^P \beta_p^2$
EN-Cox	$\lambda[\alpha \sum_{p=1}^P \beta_p + \frac{1}{2}(1 - \alpha) \sum_{p=1}^P \beta_p^2]$
OSCAR-Cox	$\lambda_1 \ \beta\ _1 + \lambda_2 \ T\beta\ _1$

In [Tibshirani 1997], the ℓ_1 -norm penalty was incorporated into the log-partial likelihood shown in Eq. (14) to obtain the Lasso-Cox algorithm, which inherits the properties of ℓ_1 -norm in feature selection.

There are also some extensions of Lasso-Cox method. Adaptive Lasso-Cox [Zhang and Lu 2007] is based on a penalized partial likelihood with adaptively weighted ℓ_1 penalties $\lambda \sum_{j=1}^P \tau_j |\beta_j|$ on regression coefficients, with small weights τ_j for large coefficients and large weights for small coefficients. In fused Lasso-Cox [Tibshirani et al. 2005], the coefficients and their successive differences are penalized using the ℓ_1 -norm. In graphical Lasso-Cox [Friedman et al. 2008], the sparse graphs are estimated using coordinate descent method by applying a ℓ_1 -penalty to the inverse covariance matrix. These extensions solve the survival problems in a similar way as the regular Lasso-Cox model by incorporating different ℓ_1 penalties.

Ridge-Cox: Ridge regression was originally proposed by Hoerl and Kennard [Hoerl and Kennard 1970] and was successfully used in the context of Cox regression by Verweij et al. [Verweij and Van Houwelingen 1994]. It incorporates a ℓ_2 -norm regularizer to select the correlated features and shrink their values towards each other.

Feature-based regularized Cox method (FEAR-Cox) [Vinzamuri and Reddy 2013] uses feature-based non-negative valued regularizer $R(\beta) = |\beta|^T M |\beta|$ for the modified least squares formulation of Cox regression and the cyclic coordinate descent method is used to solve this optimization problem, where $M \in \mathbb{R}^{P \times P}$ (P is the number of features) is a positive semi-definite matrix. Ridge-Cox is a special case of FEAR-Cox when M is the identity matrix.

EN-Cox: Elastic net (EN), which combines the ℓ_1 and squared ℓ_2 penalties, has the potential to perform the feature selection and deal with the correlation between the features simultaneously [Zou and Hastie 2005]. The EN-Cox method was proposed by Noah Simon et al. [Simon et al. 2011] where the Elastic Net penalty term shown in Table III with $0 \leq \alpha \leq 1$ and introduced into the log-partial likelihood function in Eq. (14). Different from Lasso-Cox, EN-Cox can select more than N features if $N \leq P$.

Kernel Elastic Net (KEN) algorithm [Vinzamuri and Reddy 2013], which uses the concept of kernels, compensates for the drawbacks of the EN-Cox which is partially effective at dealing with the correlated features in survival data. In KEN-Cox, it builds a kernel similarity matrix for the feature space in order to incorporate the pairwise feature similarity into the Cox model. The regularizer used in KEN-Cox is defined as $\lambda \alpha \|\beta\|_1 + \lambda(1 - \alpha) \beta^T K \beta$, where K is a symmetric kernel matrix with $K_{ij} = \exp(-\|x_i - x_j\|_2^2 / 2\sigma^2)$ ($i, j = 1, \dots, P$) as its entries. We can see that the equation for KEN-Cox method includes both smooth and non-smooth ℓ_1 terms.

OSCAR-Cox: The modified graph Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) [Yang et al. 2012; Ye and Liu 2012] regularizer is incorporated in the basic Cox model as the OSCAR-Cox algorithm [Vinzamuri and Reddy 2013], which can perform the variable selection for highly correlated features in regression

problem. The main advantage of OSCAR regularizer is that it tends to have equal coefficients for the features which relate to the outcome in similar ways. In addition, it can simultaneously obtain the advantages of the individual sparsity because of the ℓ_1 norm and the group sparsity due to the ℓ_∞ norm. The regularizer used in the formulation of the OSCAR-Cox is given in Table III, where T is the sparse symmetric edge set matrix generated by building a graph structure which considers each feature as an individual node. By using this way, a pairwise feature regularizer can be incorporated into the basic Cox regression framework.

Among the regularizers shown in Table III, the parameters $\lambda \geq 0$ can be tuned to adjust the influence introduced by the regularizer term. The performance of these penalized estimators significantly depend on λ , and the optimal λ_{opt} can be chosen via cross-validation.

3.2.3. CoxBoost. While there exists several algorithms (such as the penalized parameter estimation) which can be applied to fit the sparse survival models on the high-dimensional data, none of them are applicable in the situation that some mandatory covariates should be taken into consideration explicitly in the models. CoxBoost [Binder and Schumacher 2008] approach is proposed to incorporate the mandatory covariates into the final model. The CoxBoost method also aims at estimating the coefficients β in Eq. (8) as in the Cox model. It considers a flexible set of candidate variables for updating in each boosting step by employing the offset-based gradient boosting approach. This is the key difference from the regular gradient boosting approach, which either updates only one component of β in component-wise boosting or fits the gradient by using all covariates in each step.

3.2.4. Time-dependent (TD) Cox Model. Cox regression model is also effectively adapted to handle time-dependent covariates, which refer to the variables whose values may change with time t for a given instance. Typically, the time-dependent variable can be classified into three categories [Kleinbaum and Klein 2006]: internal time-dependent variable, ancillary time-dependent variable and defined time-dependent variable. The reason for a change in the internal time-dependent variable depends on the internal characteristics or behavior that is specific to the individual. In contrast, a variable is called an ancillary time-dependent variable if its value changes primarily due to the environment that may affect several individuals simultaneously. Defined variable, with the form of the product of a time-independent variable multiplied by a function of time, is used to analyze a time independent predictor not satisfying the PH assumption in the Cox model. The commonly used layout of the dataset in time-dependent Cox model is in the form of counting process (CP) [Kleinbaum and Klein 2006].

Given a survival analysis problem which involves both time-dependent and time-independent features, we can denote the variables at time t as $X(t) = (X_{.1}(t), X_{.2}(t), \dots, X_{.P_1}(t), X_{.1}, X_{.2}, \dots, X_{.P_2})$, where P_1 and P_2 represent the number of time-dependent and time-independent variables, respectively. And $X_{.j}(t)$ and $X_{.i}$ represent the j^{th} time-dependent feature and the i^{th} time-independent feature, respectively. Then, by involving the time-dependent features into the basic Cox model given in Eq. (8), the time-dependent Cox model can be formulated as:

$$h(t, X(t)) = h_0(t) \exp \left[\sum_{j=1}^{P_1} \delta_j X_{.j}(t) + \sum_{i=1}^{P_2} \beta_i X_{.i} \right] \quad (15)$$

where δ_j and β_i represent the coefficients corresponding to the j^{th} time-dependent variable and the i^{th} time-independent variable, respectively. For the two sets of predictors at time t : $X(t) = (X_{.1}(t), X_{.2}(t), \dots, X_{.P_1}(t), X_{.1}, X_{.2}, \dots, X_{.P_2})$ and $X^*(t) = (X_{.1}^*(t), X_{.2}^*(t), \dots, X_{.P_1}^*(t), X_{.1}^*, X_{.2}^*, \dots, X_{.P_2}^*)$, the hazard ratio for the time-dependent Cox

model can be computed as follows:

$$\hat{H}R(t) = \frac{\hat{h}(t, X^*(t))}{\hat{h}(t, X(t))} = \exp \left[\sum_{j=1}^{P_1} \delta_j [X_{.j}^*(t) - X_{.j}(t)] + \sum_{i=1}^{P_2} \beta_i [X_{.i}^* - X_{.i}] \right] \quad (16)$$

Since the first component in the exponent of Eq. (16) is time-dependent, we can consider the hazard ratio in the TD-Cox model as a function of time t . This means that it does not satisfy the PH assumption mentioned in the standard Cox model. It should be noted that the coefficient δ_j is in itself not time-dependent and it represents the overall effect of the j^{th} time-dependent variable at various survival time points. The likelihood function of time-dependent Cox model can be constructed in the same manner as done in the Cox model.

3.3. Parametric Models

The parametric censored regression models assume that the survival times or the logarithm of the survival times of all instances in the data follow a particular theoretical distribution [Lee and Wang 2003]. These models are important alternatives to the Cox-based semi-parametric models and are also widely used in many application domains. It is simple, efficient and effective in predicting the time to event of interest using parametric methods. The parametric survival models tend to obtain the survival estimates that are consistent with a theoretical survival distribution. The commonly used distributions in parametric censored regression models are: normal, exponential, weibull, logistic, log-logistic and log-normal. If the survival times of all instances in the data follow these distributions, the model is referred as linear regression model. If the logarithm of the survival times of all instances follow these distributions, the problem can be analyzed using the accelerated failure time model, in which we assume that the variable can affect the time to the event of interest of an instance by some constant factor [Lee and Wang 2003]. It should be noted that if no suitable theoretical distribution is known, nonparametric methods are more efficient.

The maximum-likelihood estimation (MLE) method [Lee and Wang 2003] can be used to estimate the parameters for these models. Let us assume that the number of instances is N with c censored observations and $(N - c)$ uncensored observations, and use $\beta = (\beta_1, \beta_2, \dots, \beta_P)^T$ as a general notation to denote the set of all parameters [Li et al. 2016e]. Then the death density function $f(t)$ and the survival function $S(t)$ of the survival time can be represented as $f(t, \beta)$ and $S(t, \beta)$, respectively. For a given instance i , if it is censored, the actual survival time will not be available. However, we can conclude that the instance i did not experience the event of interest before the censoring time C_i , so the value of the survival function $S(C_i, \beta)$ will be a probability closed to 1. In contrast, if the event occurs for instance i at T_i , then the death density function $f(T_i, \beta)$ will have a high probability value. Thus, we can denote $\prod_{\delta_i=1} f(T_i, \beta)$

as the joint probability of all the uncensored observations and $\prod_{\delta_i=0} S(T_i, \beta)$ to represent the joint probability of the c censored observations [Li et al. 2016e]. Therefore, we can estimate the parameters β by optimizing the likelihood function of all N instances in the form of

$$L(\beta) = \prod_{\delta_i=1} f(T_i, \beta) \prod_{\delta_i=0} S(T_i, \beta) \quad (17)$$

Table IV shows the death density function $f(t)$ and its corresponding survival function $S(t)$ and hazard function $h(t)$ for these commonly used distributions. Now we will discuss more details about these distributions.

Table IV: Density, Survival and Hazard functions for the distributions commonly used in the parametric methods in survival analysis.

Distribution	PDF $f(t)$	Survival $S(t)$	Hazard $h(t)$
Exponential	$\lambda \exp(-\lambda t)$	$\exp(-\lambda t)$	λ
Weibull	$\lambda k t^{k-1} \exp(-\lambda t^k)$	$\exp(-\lambda t^k)$	$\lambda k t^{k-1}$
Logistic	$\frac{e^{-(t-\mu)/\sigma}}{\sigma(1+e^{-(t-\mu)/\sigma})^2}$	$\frac{e^{-(t-\mu)/\sigma}}{1+e^{-(t-\mu)/\sigma}}$	$\frac{1}{\sigma(1+e^{-(t-\mu)/\sigma})}$
Log-logistic	$\frac{\lambda k t^{k-1}}{(1+\lambda t^k)^2}$	$\frac{1}{1+\lambda t^k}$	$\frac{\lambda k t^{k-1}}{1+\lambda t^k}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(t-\mu)^2}{2\sigma^2})$	$1 - \Phi(\frac{t-\mu}{\sigma})$	$\frac{1}{\sqrt{2\pi}\sigma(1-\Phi((t-\mu)/\sigma))} \exp(-\frac{(t-\mu)^2}{2\sigma^2})$
Log-normal	$\frac{1}{\sqrt{2\pi}\sigma t} \exp(-\frac{(\log(t)-\mu)^2}{2\sigma^2})$	$1 - \Phi(\frac{\log(t)-\mu}{\sigma})$	$\frac{\frac{1}{\sqrt{2\pi}\sigma t} \exp(-\frac{(\log(t)-\mu)^2}{2\sigma^2})}{1 - \Phi(\frac{\log(t)-\mu}{\sigma})}$

Exponential Distribution: Among the parametric models in survival analysis, exponential model is the simplest and prominent one since it is characterized by a constant hazard rate, λ , which is the only parameter. In this case, the failure or the death is assumed to be a random event independent of time. A larger value of λ indicates a higher risk and a shorter survival time period. Based on the survival function shown in Table IV, we can have $\log S(t) = -\lambda t$, in which the relationship between the logarithm of survival function and time t is linear with λ as the slope. Thus, it is easy to determine whether the time follows an exponential distribution by plotting $\log \hat{S}(t)$ against time t [Lee and Wang 2003].

Weibull Distribution: The Weibull model, which is characterized by two parameters $\lambda > 0$ and $k > 0$, is the most widely used parametric distribution for survival problem. The shape of the hazard function is determined using the shape parameter k , which provides more flexibility compared to the exponential model. If $k = 1$, the hazard will be a constant, and in this case, the Weibull model will become an exponential model. If $k < 1$, the hazard function will be decreasing over time. The scaling of the hazard function is determined by the scaling parameter λ .

Logistic and Log-logistic Distribution: In contrast to Weibull model, the hazard functions of both logistic and log-logistic models allow for non-monotonic behavior in the hazard function, which is shown in Table IV. The survival time T and the logarithm of survival time $\log(T)$ will follow the logistic distribution in logistic and log-logistic models, respectively. For logistic model, μ is the parameter to determine the location of the function, while σ is the scale parameter. For log-logistic model, the parameter $k > 0$ is the shape parameter. If $k \leq 1$, the hazard function is decreasing over time. However, if $k > 1$, the hazard function will increase over time to the maximum value first and then decrease, which means that the hazard function is unimodal if $k > 1$. Thus, the log-logistic distribution may be used to describe a monotonically decreasing hazard or a first increasing and then decreasing hazard [Lee and Wang 2003].

Normal and Log-normal Distribution: If the survival time T satisfies the condition that T or $\log(T)$ is normally distributed with mean μ and variance σ^2 , then T is normally or log-normally distributed. This is suitable for the survival patterns with an initially increasing and then decreasing hazard rate.

Based on the framework given in Eq. (17), we will discuss these commonly used parametric methods.

3.3.1. Linear regression models. In data analysis, the linear regression model, together with the least squares estimation method, is one of the most commonly used approach. We cannot apply it directly to solve survival analysis problems since the actual event times are missing for censored instances. Some linear models [Miller and Halpern 1982; Koul et al. 1981; Buckley and James 1979; Wang et al. 2008; Li et al. 2016e] including Tobit regression and Buckley-James (BJ) regression were proposed to handle censored instances in survival analysis. Strictly speaking, linear regression is a specific parametric censored regression, however, this method is fundamental in data analysis, and hence we discuss the linear regression methods for censored data separately here.

Tobit Regression: The Tobit model [Tobin 1958] is one of the earliest attempts to extend linear regression with the Gaussian distribution for data analysis with censored observations. In this model, a latent variable y^* is introduced and the assumption made here is that it linearly depends on X via the parameter β as $y^* = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, where ϵ is a normally distributed error term. Then, for the i^{th} instance, the observable variable y_i will be y_i^* if $y_i^* > 0$, otherwise it will be 0. This means that if the latent variable is above zero, the observed variable equals to the latent variable and zero otherwise. Based on the latent variable, the parameters in the model can be estimated with maximum likelihood estimation (MLE) method.

Buckley-James Regression: The Buckley-James (BJ) regression [Buckley and James 1979] estimates the survival time of the censored instances as the response value based on Kaplan-Meier (KM) estimation method, and then fits a linear (AFT) model by considering the survival times of uncensored instances and the approximated survival times of the censored instances at the same time. To handle high-dimensional survival data, Wang et al. [Wang et al. 2008] applied the elastic net regularizer in the BJ regression (EN-BJ).

Penalized Regression: Penalized regression methods [Kyung et al. 2010] are well-known for their nice properties of simultaneous variable selection and coefficient estimation. The penalized regression method can provide better prediction results in the presence of either multi-collinearity of the covariates or high-dimensionality. Recently, these methods have received a great attention in survival analysis. The weighted linear regression model with different regularizers for high-dimensional censored data is an efficient method to handle the censored data by giving different weights to different instances [Li et al. 2016e]. In addition, the structured regularization based linear regression algorithm [Bach et al. 2012; Vinzamuri et al. 2017] for right censored data has a good ability to infer the underlying structure of the survival data.

- **Weighted Regression:** Weighted regression method [Li et al. 2016b] can be used when the constant variance assumption about the error in the ordinary least squares methods is violated (which is called heteroscedasticity), which is different from the constant variance in the errors (which is called homoscedasticity) in ordinary least squares regression methods. Instead of minimizing the residual sum of squares, the weighted regression method minimizes the weighted sum of squares $\sum_{i=1}^n w_i (y_i - X_i\beta)^2$. The ordinary least squares is a special case of this where all the weights $w_i = 1$. Weighted regression method can be solved in the same manner as the ordinary linear least squares problem. In addition, using the weighted regression method, we can assign higher weights to the instances that we want to emphasize or ones where mistakes are especially costly. If we give the samples high weights, the model will be pulled towards matching the data. This will be very helpful for survival analysis to put more emphasis on the instances whose information may contribute more to the model.

- **Structured Regularization:** The ability to effectively infer latent knowledge through tree-based hierarchies and graph-based relationships is extremely crucial in survival analysis. This is also supported by the effectiveness of structured sparsity based regularization methods in regression [Bach et al. 2012]. Structured regularization based Linear REgression algorithm for right Censored data (SLIREC) in [Vinzamuri et al. 2017] infers the underlying structure of the survival data directly using sparse inverse covariance estimation (SICE) method and uses the structural knowledge to guide the base linear regression model. The structured approach is more robust compared to the standard statistical and Cox based methods since it can automatically adapt to different distributions of events and censored instances.

3.3.2. Accelerated Failure Time (AFT) Model. In the parametric censored regression methods discussed previously, we assume that the survival time of all instances in the given data follows a specific distribution and that the relationship between either the survival time or the logarithm of the survival time and the features is linear. Specially, if the relationship between the logarithm of survival time T and the covariates is linear in nature, it is also termed as Accelerated failure time (AFT) model [Kalbfleisch and Prentice 2011]. Thus, we consider these regression methods as the generalized linear models.

In the AFT model, it assumes that the relationship of the logarithm of survival time T and the covariates is linear and can be written in the following form.

$$\ln(T) = X\beta + \sigma\epsilon \quad (18)$$

where X is the covariate matrix, β represents the coefficient vector, $\sigma(\sigma > 0)$ denotes an unknown scale parameter, and ϵ is an error variable which follows a similar distribution to $\ln(T)$. Typically, we make a parametric assumption on ϵ which can follow any of the distributions given in Table IV. In this case, the survival is dependent on both the covariate and the underlying distribution. Then, the only distinction of an AFT model compared to regular linear methods would be the inclusion of censored information in the survival analysis problem. The AFT model is additive with respect to $\ln(T)$, while multiplicative with respect to T , and is written in the form of $T = e^{X\beta}e^{\sigma\epsilon}$.

In order to demonstrate the idea of the assumption in AFT model, let us compare the survival function among smokers $S_1(t)$ and non-smokers $S_2(t)$. Then, the assumption of AFT model is in the form of

$$S_2(t) = S_1(\gamma t) \quad (19)$$

where $t \geq 0$ and γ represents a constant which is named as an acceleration factor for comparison of the smokers and non-smokers. For the linear regression method, we can parameterize γ as $\exp(\alpha)$, where α can be estimated using the given data. Then, the assumption in AFT method will be updated to $S_2(t) = S_1(\exp(\alpha)t)$. The acceleration factor which is the key measure of the relationships in the AFT method is used to evaluate the effect of features on the survival time.

4. MACHINE LEARNING METHODS

In the past several years, due to the advantages of machine learning techniques, such as its ability to model the non-linear relationships and the quality of their overall predictions made, they have achieved significant success in various practical domains. In survival analysis, the main challenge of machine learning methods is the difficulty to appropriately deal with censored information and the time estimation of the model. Machine learning is effective when there are a large number of instances in a reasonable dimensional feature space, but this is not the case for certain problems in survival analysis [Zupan et al. 2000]. In this section, we will do a comprehensive review of commonly used machine learning methods in survival analysis.

4.1. Survival Trees

Survival trees are one form of classification and regression trees which are tailored to handle censored data. The basic intuition behind the tree models is to recursively partition the data based on a particular splitting criterion, and the objects that are similar to each other based on the event of interest will be placed in the same node. The earliest attempt at using a tree structure for survival data was made in [Ciampi et al. 1981]. However, [Gordon and Olshen 1985] is the first paper which discussed the creation of survival trees.

The primary difference between a survival tree and the standard decision tree is in the choice of splitting criterion. The decision tree method performs recursive partitioning on the data by setting a threshold for each feature, however, it can neither consider the interactions between the features nor the censored information in the model [Safavian and Landgrebe 1991]. The splitting criteria used for survival trees can be grouped into two categories: (i) maximizing between-node heterogeneity and (ii) minimizing within-node homogeneity. The first class of approaches minimizes the loss function using the within-node homogeneity criterion. The authors in [Gordon and Olshen 1985] measured the homogeneity and Hellinger distances between the estimated distribution functions using the Wasserstein metric. An exponential log-likelihood function was employed in [Davis and Anderson 1989] for recursive partitioning based on the sum of residuals from the Cox model. Leblanc and Crowley [LeBlanc and Crowley 1992] measured the node deviance based on the first step of a full likelihood estimation procedure. In the second class of splitting criteria, Ciampi et al. [Ciampi et al. 1986] employed log-rank test statistics for between-node heterogeneity measures. Later, Ciampi et al. [Ciampi et al. 1987] proposed a likelihood ratio statistic to measure the dissimilarity between two nodes. Based on the Tarone-Ware class of two-sample statistics, Segal [Segal 1988] introduced a procedure to measure the between-node dissimilarity. The main improvement of a survival tree over the standard decision tree is its ability to handle the censored data using the tree structure.

Another important aspect of building a survival tree is the selection of the final tree. Procedures such as backward selection or forward selection can be followed for choosing the optimal tree [Bou-Hamad et al. 2011]. However, an ensemble of trees (described in Section 4.5) can avoid the problem of final tree selection with better performance compared to a single tree.

4.2. Bayesian Methods

Bayes theorem is one of the most fundamental principles in probability theory and mathematical statistics; it provides a link between the *posterior probability* and the *prior probability*, so that one can see the changes in probability values before and after accounting for a certain event. Using the Bayes theorem, there are two models, namely, Naïve Bayes (NB) and Bayesian network (BN) [Friedman et al. 1997]. Both of these approaches, which provide the probability of the event of interests as their outputs, are commonly studied in the context of clinical prediction [Kononenko 1993; Pepe 2003; Zupan et al. 2000]. The experimental results of using Bayesian methods on survival data show that Bayesian methods have good properties of both interpretability and uncertainty reasoning [Raftery et al. 1995].

Naïve Bayes, a well-known probabilistic method in machine learning, is one of the simplest yet effective prediction algorithms. In [Bellazzi and Zupan 2008], the authors build a naïve Bayesian classifier to make predictions in clinical medicine by estimating various probabilities from the data. Recently, the authors in [Fard et al. 2016] effectively integrate Bayesian methods with an AFT model by extrapolating the prior event probability to implement early stage prediction on survival data for the future time points. One drawback of Naïve Bayes method is that it makes the independence

assumption between all the features, which may not be true for many problems in survival analysis.

A Bayesian network, in which the features can be related to each other at various levels, can graphically represent a theoretical distribution over a set of variables. Bayesian networks can visually represent all the relationships between the variables which makes it interpretable for the end user. It can acquire knowledge information by using procedures of estimating the network structures and parameters from a given dataset. In [Lisboa et al. 2003], the authors proposed a Bayesian neural network framework to perform model selection for longitudinal data using automatic relevance determination [MacKay 1995]. In [Raftery 1995], a Bayesian model averaging for Cox proportional hazards models is proposed and also used to evaluate the Bayes factors in the problem. More recently, in [Fard et al. 2016], the authors proposed a novel framework which combines the power of Bayesian network representation with the AFT model by extrapolating the prior probabilities to future time points.

4.3. Artificial Neural Networks

Inspired by biological neural systems, in 1958, Frank Rosenblatt published the first paper [Rosenblatt 1958] about artificial neural network (ANN). In this approach, the simple artificial nodes denoted by “neurons” are connected based on a weighted link to form a network which simulates a biological neural network. A neuron in this context is a computing element which consists of sets of adaptive weights and generates the output based on a certain kind of *activation function*. Artificial neural network (ANN) has been widely used in survival analysis. Three kinds of methods are proposed in the literature which employ the neural network method to solve the survival analysis problems.

- (1) The neural network survival analysis has been employed to predict the survival time of a subject directly from the given inputs.
- (2) The authors in [Faraggi and Simon 1995] extended the Cox PH model to the non-linear ANN predictor and suggested to fit the neural network which has a linear output layer and a single logistic hidden layer. The authors in [Mariani et al. 1997] used both the standard Cox model and the neural network method proposed in [Faraggi and Simon 1995] to assess the prognostic factors for the recurrence of breast cancer. Although these extensions for Cox model allowed for preserving most of the advantages of a typical PH model, they were still not the optimal way to model the baseline variation [Baesens et al. 2005].
- (3) Many approaches [Liestbl et al. 1994; Biganzoli et al. 1998; Brown et al. 1997; Ravdin and Clark 1992; Lisboa et al. 2003] take the survival status of a subject, which can be represented by the survival or hazard probability, as the output of the neural network. The authors in [Biganzoli et al. 1998] apply the partial logistic artificial neural network (PLANN) method to analyze the relationship between the features and the survival times in order to obtain a better predictability of the model. Recently, feed-forward neural networks are used to obtain a more flexible non-linear model by considering the censored information in the data using a generalization of both continuous and discrete time models [Biganzoli et al. 1998]. In [Lisboa et al. 2003], the PLANN was extended to a Bayesian neural framework with covariate-specific regularization to carry model selection using automatic relevance determination [MacKay 1995].

4.4. Support Vector Machine

Support Vector Machine (SVM), a very successful supervised learning approach, is used mostly for classification and can also be modified for regression problems [Smola

and Schölkopf 2004]. It has also been successfully adapted to in survival analysis problems.

A naive way is to consider only those instances which have events in support vector regression (SVR), in which the ϵ -insensitive loss function, $f(X_i) = \max(0, |f(X_i) - y_i| - \epsilon)$, is minimized with a regularizer [Smola and Schölkopf 1998]. However, the main disadvantage of this approach is that the order information included in the censored instances will be completely ignored [Shivaswamy et al. 2007]. Another possible approach to handle the censored data is to use support vector classification using the constraint classification approach [Har-Peled et al. 2002] which imposes constraints in the SVM formulation for two comparable instances in order to maintain the required order. However, the computational complexity for this algorithm is quadratic with respect to the number of instances. In addition, it only focuses on the ordering among the instances, and ignores the actual values of the output.

The authors in [Khan and Zubek 2008] proposed support vector regression for censored data (SVRc), which takes advantage of the standard SVR and also adapts it for censored cases by using an updated asymmetric loss function. In this case, it considers both the uncensored and censored instances in the model. The work in [Van et al. 2007] studies a learning machine designed for predictive modeling of independently right censored survival data by introducing a health index which serves as a proxy between the instance's covariates and the outcome. The authors in [Van et al. 2011] introduces a SVR based approach which combines the ranking and regression methods in the context of survival analysis.

Relevance Vector Machine (RVM) [Widodo and Yang 2011; Kiaee et al. 2016], which obtains the parsimonious estimations for regression and probabilistic problems using Bayesian inference, has the same formulation as SVM but provides probabilistic classification. RVM adopts a Bayesian approach by considering the prior over the weights controlled by some parameters. Each of these parameters corresponds to a weight, the most probable value of which can be estimated iteratively using the data. The Bayesian representation of the RVM can avoid these parameters in SVM (the optimization methods based on cross-validation are usually used.). However, it is possible that RVMs converge to the local minimum since EM algorithm is used to learn the parameters. This is different from the regular sequential minimal optimization (SMO) algorithm used in SVM, which can guarantee the convergence to a global minimum.

4.5. Advanced Machine Learning Approaches

Over the past few years, more advanced machine learning methods have been developed to deal with and predict from censored data. These methods have various unique advantages on survival data compared to the other methods described so far.

4.5.1. Ensemble Learning. Ensemble learning methods [Dietterich 2000] generate a committee of classifiers and then predict the class labels for the new coming data points by taking a weighted vote among the prediction results from all these classifiers. It is often possible to construct good ensembles and obtain a better approximation of the unknown function by varying the initial points, especially in the presence of insufficient data. To overcome the instability of a single method, bagging [Breiman 1996] and random forests [Breiman 2001], proposed by Breiman, are commonly used to perform the ensemble based model building. Such ensemble models have been successfully adapted to survival analysis.

Bagging Survival Trees: Bagging is one of the oldest and most commonly used ensemble method which typically reduces the variance of the base models that are used. In bagging survival trees, the aggregated survival function can be calculated by averaging the predictions made by a single survival tree instead of taking a majority

vote [Hothorn et al. 2004]. There are mainly three steps in this method: (i) Draw B bootstrap samples from the given data. (ii) For each bootstrap sample, build a survival tree and ensure that, for all the terminal nodes, the number of events is greater than or equal to the threshold d . (iii) By averaging the leaf nodes' predictions, calculate the bootstrap aggregated survival function. For each leaf node the survival function is estimated using the KM estimator, and all the individuals within the same node are assumed to have the same survival function.

Random Survival Forests: Random forest is an ensemble method specifically proposed to make predictions using the tree structured models [Breiman 2001]. It is based on a framework similar to Bagging; the main difference between random forest and bagging is that, at a certain node, rather than using all the attributes, random forest only uses a random subset of the residual attributes to select the attributes based on the splitting criterion. It is shown that randomization can reduce the correlation among the trees and thus improve the prediction performance.

Random survival forest (RSF) [Ishwaran et al. 2008] extended Breiman's random forest method by using a forest of survival trees for prediction. There are mainly four steps in RSF: (i) Draw B bootstrap samples randomly from the given dataset. This is also called out-of-bag (OOB) data because around 37% of the data is excluded in each sample. (ii) For each sample, build a survival tree by randomly selecting features and split the node using the candidate feature which can maximize the survival difference between the child nodes. (iii) Build the tree to the full size with a constraint that the terminal node has greater than or equal to a specific unique deaths. (iv) Using the non-parametric Nelson-Aalen estimator, calculate the ensemble cumulative hazard function (CHF) of OOB data by taking the average of the CHF of each tree. In addition, the authors in [Ishwaran et al. 2011] provide an effective way to apply RSF for high-dimensional survival analysis problems by regularizing forests.

Boosting: Boosting algorithm is one of the widely used ensemble methods designed to combine base learners into a weighted sum that represents the final output of the strong learner. It iteratively fits the appropriately defined residuals based on the gradient descent algorithm [Hothorn et al. 2006; Bühlmann and Hothorn 2007]. The authors in [Hothorn et al. 2006] extend the gradient boosting algorithm to minimize the weighted risk function $\hat{\beta}_{\tilde{U},X} = \arg \min_{\beta} \sum_{i=1}^N w_i (\tilde{U}_i - h(X_i|\beta))$, where \tilde{U} is a pseudo-response variable with $\tilde{U}_i = -\frac{\partial L(y_i, \phi)}{\partial \phi} \Big|_{\phi=f_m(X_i)}$; β is a vector of parameters; $h(\cdot|\beta_{U,X})$ is the prediction made by regressing U using a base learner. Then the steps to optimize this problem are as follows: (i) Initialize $\tilde{U}_i = y_i$ ($i = 1, \dots, N$), $m = 0$ and $\hat{f}_0(\cdot|\hat{\beta}_{\tilde{U},X})$; fix the number of iterations M ($M > 1$). (ii) Fit $h(\cdot|\hat{\beta}_{U,X})$ after updating residuals \tilde{U}_i ($i = 1, \dots, N$). (iii) Iteratively update $\hat{f}_{m+1}(\cdot) = \hat{f}_m(\cdot) + v h(\cdot|\hat{\beta}_{U,X})$, where $0 < v \leq 1$ represents the step size. (iv) Repeat the procedures in steps (ii) and (iii) until $m = M$.

4.5.2. Active Learning. Active learning based on the data with censored observations can be very helpful for survival analysis since the opinions of an expert in the domain can be incorporated into the models. Active learning mechanism allows the survival model to select a subset of subjects by learning from a limited set of labeled subjects first and then query the expert to get the label of survival status before considering it in the training set. The feedback from the expert is particularly useful for improving the model in many real-world application domains [Vinzamuri et al. 2014]. The goal of active learning for survival analysis problems is to build a survival regression model by utilizing the censored instances completely without deleting or modifying the

instance. In [Vinzamuri et al. 2014], the active regularized Cox regression (ARC) algorithm based on a discriminative gradient sampling strategy is proposed by integrating the active learning method with the Cox model. The ARC framework is an iteration based algorithm with three main steps: (i) Build a regularized Cox regression using the training data, (ii) Apply the model obtained in (i) to all the instances in the unlabeled pool, (iii) Update the training data and the unlabeled pool, select the instance whose influence on the model is the highest and label it before running the next iteration. One of the main advantages of the ARC framework is that it can identify the instances and get the feedback about event labeling from the domain expert.

4.5.3. Transfer Learning. Collecting labeled information in survival problems is very time consuming, i.e., one has to wait for the event occurrence from a sufficient number of training instances to build robust models. A naive solution for this insufficient data problem is to merely integrate the data from related tasks into a consolidated form and build prediction models on such integrated data. However, such approaches often do not perform well because the target task (for which the predictions are to be made) will be overwhelmed by auxiliary data with different distributions. In such scenarios, knowledge transfer between related tasks will usually produce much better results compared to a data integration approach. Transfer learning method has been extensively studied to solve standard regression and classification problems [Pan and Yang 2010]. Recently, in [Li et al. 2016c], a regularized Cox PH model named Transfer-Cox, is proposed to improve the prediction performance of the Cox model in the target domain through knowledge transfer from the source domain in the context of survival models built on multiple high-dimensional datasets. The Transfer-Cox model employs $\ell_{2,1}$ -norm to penalize the sum of the loss functions (negative partial log-likelihood) for both source and target domains. Thus, the model will not only select important features but will also learn a shared representation across source and target domains to improve the model performance on the target task.

4.5.4. Multi-task Learning. In [Li et al. 2016d], the survival time prediction problem is reformulated as a multi-task learning problem. In survival data, the outcome labeling matrix is incomplete since the event label of each censored instance is unavailable after its corresponding censoring time; therefore, it is not suitable to handle the censored information using the standard multi-task learning methods. To solve this problem, the multi-task learning model for survival analysis (MTLSA) translates the original event labels into a $N \times K$ indicator matrix I , where $K = \max(y_i) (\forall i = 1, \dots, N)$ is the maximum follow-up time of all the instances in the dataset. The element I_{ij} ($i = 1, \dots, N; j = 1, \dots, K$) of the indicator matrix will be 1 if the event occurred before time y_j for instance i , otherwise it will be 0. One of the primary advantages of the MTLSA approach is that it can capture the dependency between the outcomes at various time points by using a shared representation across the related tasks in the transformation, which will reduce the prediction error on each task. In addition, the model can simultaneously learn from both uncensored and censored instances based on the indicator matrix. One important characteristic of non-recurring events, i.e., once the event occurs it will not occur again, is encoded via the *non-negative non-increasing list* structure constraint. In the MTLSA algorithm, the $\ell_{2,1}$ -norm penalty is employed to learn a shared representation across related tasks and hence compute the relatedness between the individual models built for various unique event time points.

5. PERFORMANCE EVALUATION METRICS

Due to the presence of the censoring in survival data, the standard evaluation metrics for regression such as root of mean squared error and R^2 are not suitable for measuring the performance in survival analysis [Heagerty and Zheng 2005]. Instead, the predic-

tion performance in survival analysis needs to be measured using more specialized evaluation metrics.

5.1. C-index

In survival analysis, a common way to evaluate a model is to consider the relative risk of an event for different instance instead of the absolute survival times for each instance. This can be done by computing the concordance probability or the concordance index (C-index) [Harrell et al. 1984; Harrell et al. 1982; Pencina and D'Agostino 2004]. The survival times of two instances can be ordered for two scenarios: (1) both of them are uncensored; (2) the observed event time of the uncensored instance is smaller than the censoring time of the censored instance [Steck et al. 2008]. This can be visualized by the ordered graph given in Figure 4. Figure 4(a) and Figure 4(b) are used to illus-

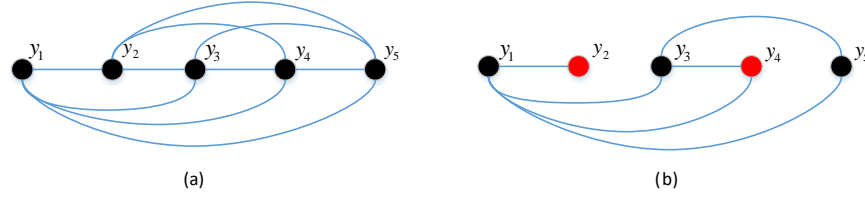


Fig. 4: Illustration of the ranking constraints in survival data for C-index calculations ($y_1 < y_2 < y_3 < y_4 < y_5$). Here, black circles indicate the observed events and red circles indicate the censored observations. (a) No censored data and (b) with censored data.

trate the possible ranking comparisons (denoted by edges between instances) for the survival data without and with censored instances, respectively. There are $\binom{5}{2} = 10$ possible pairwise comparisons for the five instances in the survival data without censored cases shown in Figure 4(a). Due to the presence of censored instances (represented by red circles) in Figure 4(b), only 6 out of the 10 comparisons are feasible. It should be noted that, for a censored instance, only an earlier uncensored instance (for example y_2 & y_1) can be compared with. However, any censored instance cannot be compared with both censored and uncensored instances after its censored time (for example, y_2 & y_3 and y_2 & y_4) since its actual event time is unknown.

Consider both the observations and prediction values of two instances, (y_1, \hat{y}_1) and (y_2, \hat{y}_2) , where y_i and \hat{y}_i represent the actual observation time and the predicted value, respectively. The concordance probability between them can be computed as

$$c = Pr(\hat{y}_1 > \hat{y}_2 | y_1 \geq y_2) \quad (20)$$

By this definition, for the binary prediction problem, C-index will have a similar meaning to the regular area under the ROC curve (AUC), and if y_i is binary, then the C-index is the AUC [Li et al. 2016d]. As the definition above is not straightforward, in practice, there are multiple ways of calculating the C-index.

- (1) When the output of the model is a hazard ratio (such as the outcome obtained by Cox based models), C-index can be computed using

$$\hat{c} = \frac{1}{num} \sum_{i: \delta_i=1} \sum_{j: y_i < y_j} I[X_i \hat{\beta} > X_j \hat{\beta}] \quad (21)$$

where $i, j \in \{1, \dots, N\}$, num denotes the number of all comparable pairs, $I[\cdot]$ is the indicator function and $\hat{\beta}$ is the estimated parameters from the Cox based models.

- (2) For the survival methods which aim at directly learning the survival time, the C-index should be calculated as:

$$\hat{c} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i < y_j} I[S(\hat{y}_j|X_j) > S(\hat{y}_i|X_i)] \quad (22)$$

where $S(\cdot)$ is the estimated survival probabilities.

In order to evaluate the performance during a follow-up period, Heagerty and Zheng defined the C-index for a fixed follow-up time period $(0, t^*)$ as the weighted average of AUC values at all possible observation time points [Heagerty and Zheng 2005]. The time-dependent AUC for any specific survival time t can be calculated as

$$AUC(t) = P(\hat{y}_i < \hat{y}_j | y_i < t, y_j > t) = \frac{1}{num(t)} \sum_{i:y_i < t} \sum_{j:y_j > t} I(\hat{y}_i < \hat{y}_j) \quad (23)$$

where $t \in T_s$ which is the set of all possible survival times and $num(t)$ represents the number of comparable pairs for the time point t . Then the C-index during the time period $(0, t^*)$, which is the weighted average of the time-dependent AUC obtained by Eq. (23), is computed as

$$c_{t^*} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i < y_j} I(\hat{y}_i < \hat{y}_j) = \sum_{t \in T_s} AUC(t) \cdot \frac{num(t)}{num} \quad (24)$$

Thus c_{t^*} is the probability that the predictions are concordant with their outcomes for a given data during the time period $(0, t^*)$.

5.2. Brier Score

Named after the inventor Glenn W. Brier, the Brier score (BS) [Brier 1950] is developed to predict the inaccuracy of probabilistic weather forecasts. It can only evaluate the prediction models which have probabilistic outcomes; that is, the outcome must remain within the range $[0,1]$, and the sum of all the possible outcomes for a certain individual should be 1. When we consider the binary outcome prediction with a sample of N instances and for each X_i ($i = 1, 2, \dots, N$), the predicted outcome at t is $\hat{y}_i(t)$, and the actual outcome is $y_i(t)$; then, the empirical definition of the Brier score at the specific time t can be given by

$$BS(t) = \frac{1}{N} \sum_{i=1}^N [\hat{y}_i(t) - y_i(t)]^2 \quad (25)$$

where the actual outcome $y_i(t)$ for each instance can only be 1 or 0.

Brier score was extended in [Graf et al. 1999] to be a performance measure for survival problems with censored information to evaluate the prediction models where the outcome to be predicted is either binary or categorical in nature. When incorporating the censoring information in the dataset, the individual contributions to the empirical Brier score are reweighted according to the censored information. Then, the Brier score can be updated as follows:

$$BS(t) = \frac{1}{N} \sum_{i=1}^N w_i(t) [\hat{y}_i(t) - y_i(t)]^2 \quad (26)$$

In Eq.(26), $w_i(t)$, given in Eq. (27), denotes the weight for the i^{th} instance and it is estimated by incorporating the Kaplan-Meier estimator of the censoring distribution

G obtained on the given dataset $(X_i, y_i, 1 - \delta_i), i = 1, \dots, N$.

$$w_i(t) = \begin{cases} \delta_i/G(y_i) & \text{if } y_i \leq t \\ 1/G(y_i) & \text{if } y_i > t \end{cases} \quad (27)$$

With this weight distribution, the weights for the instances that are censored before t will be 0. However, they contribute indirectly to the calculation of the Brier score since they are used for calculating G . The weights for the instances that are uncensored at t are greater than 1, so that they contribute their estimated survival probability to the calculation of the Brier score.

5.3. Mean Absolute Error

For survival analysis problems, the mean absolute error (MAE) can be defined as an average of the differences between the predicted time values and the actual observation time values. It is calculated as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^N (\delta_i |y_i - \hat{y}_i|) \quad (28)$$

where y_i ($i = 1, \dots, N$) represents the actual observation times, and \hat{y}_i ($i = 1, \dots, N$) denotes the predicted times. It should be noted that only the samples for which the event occurs are being considered in this metric since if $\delta_i = 0$, the corresponding term will become zero. MAE can only be used for the evaluation of survival models which can provide the event time as the predicted target value such as AFT models.

6. RELATED TOPICS

Besides the machine learning methods introduced in Section 4 and the traditional statistical survival methods discussed in Section 3, there are few other topics that are closely related to survival analysis and we will summarize them now.

6.1. Early Prediction

One of the primary challenges in the context of survival analysis, and in general longitudinal studies, is that a sufficient number of events in the training data can be collected only by waiting for a long period. This is the most significant difference from the regular supervised learning problems, in which the labels for each instance can be given by a domain expert in a reasonable time period. Therefore, a good survival model should have the ability to forecast the event occurrence at future time by using only a limited event occurrence information at the early stage of a longitudinal problem.

There are many real-world applications which motivate the need for new prediction models which can work using only the early stage data in longitudinal studies. For example, in the healthcare domain, it is critical to study the effect of a new treatment in order to understand the treatment or drug efficacy, which should be estimated as early as possible. In this case, the patients will be monitored over a certain time period and the event of interest will be the patient admission to the hospital due to the treatment failure. This scenario clearly indicates the need for algorithms which can predict the event occurrence effectively using only a few events.

To solve this problem, an Early Stage Prediction (ESP) approach trained at early stages of longitudinal studies to predict the time-to-event is proposed in [Fard et al. 2016]. Two algorithms based on Naïve Bayes and Bayesian Networks are developed by estimating the posterior probability of event occurrence based on different extrapolation techniques using Weibull, Log-logistic and Log-normal distributions discussed in Section 3. The ESP framework is a two-stage algorithm: (1) Estimate the conditional probability distribution based on the training data collected until the early stage time

point (t_c) of the study; (2) Extrapolating the prior probability of the event for the future time (t_f) using AFT model with different distributions. According to the experimental results in these works, the ESP framework can provide more accurate predictions when the prior probability at the future time is appropriately estimated using the current information of event occurrence.

6.2. Data Transformation

In this section, we will discuss two data transformation techniques that will be useful for data pre-processing in survival analysis. Both of these approaches transform the data to a more conducive form so that other survival-based (or sometimes even the standard algorithms) can be applied effectively.

6.2.1. Uncensoring approach. In survival data, the incompleteness in the event (outcome) information makes it difficult for standard machine learning methods to learn from such data. The censored observations in survival data might look similar to unlabeled samples in classification or unknown response in regression problem in the sense that status or time-to-event is not known for some of the observations. However, different from unlabeled samples where the labeling information is completely missing, the censored instances actually have partial informative labeling information which provides the possible range of the corresponding true response (survival time). Such censored data have to be handled with special care within any machine learning method in order to make good predictions. Also, in survival analysis problems, only the information before a certain time point (before censoring occurs) is available for the censored instances and this information should be integrated into the prediction algorithm to obtain the most optimal result.

Typically, there are two naive ways of handling such censored data. One is to delete the censored instances, and it performs well if the number of the samples are large enough and the censoring instances are not censored randomly. However, it will provide a sub-optimal model because of neglecting the available information in those censored instances [Delen et al. 2005; Burke et al. 1997]. Treating censoring as event-free is another naive and simple choice. This method performs well for data with only a few censored instances, but it underestimates the true performance of the model. Although these methods are simple for handling the censored data, they lose useful information available in the data. Here, we list two other approaches proposed in the literature to handle censored data.

- (1) Group the instances in the given data into three categories [Zupan et al. 2000]: (i) instances which experience the event of interest during the observation will be labeled as event; (ii) instances whose censored time is later than a predefined time point are labeled as event-free; (iii) for instances whose censored time is earlier than a predefined time point, a copy of these instances will be labeled as event and another copy of the same instances will be labeled as event-free, respectively, and all these instances will be weighted by a marginal probability of event occurrence estimated by the Kaplan-Meier method.
- (2) For each censored instance, estimate the probability of event and probability of being censored (considering censoring as a new event) using Kaplan-Meier estimator and give a new class label based on these probability values [Fard et al. 2016]. For each instance in the data, when the probability of event exceeds the probability of being censored, then it is labeled as event; otherwise, it will be labeled as event-free which indicates that even if there is complete follow-up information for that instance, there is extremely low chance of event occurrence by the end of the observation time period.

6.2.2. Calibration. Censoring causes missing time-to-event labels, and this effect is compounded when dealing with datasets which have high amounts of censored instances. Instead of using the uncensoring approach, calibration methods for survival analysis can also be used to solve this problem by learning more optimal time-to-event labels for the censored instances. Generally, there are mainly two reasons which motivate calibration. First, the survival analysis model is built using the given dataset where the missing time-to-events for the censored instances are assigned to a value such as the duration of the study or last known follow up time. However, this approach is not suitable for handling data with many censored instances. In other words, for such data, these inappropriately labeled censored instances cannot provide much information to the survival algorithm. Calibration method can be used to overcome this missing time-to-events problem in survival analysis. Secondly, dependent censoring in the data, where censoring is dependent on the covariates, may lead to some bias in standard survival estimators, such as KM method. This motivates an imputed censoring approach which calibrates the time-to-event attribute to decrease the bias of the survival estimators.

In [Vinzamuri et al. 2017], a calibration survival analysis method which uses a regularized inverse covariance based imputation is proposed to overcome the problems mentioned above. It has the ability to capture correlations between censored instances and correlations between similar features. In calibrated survival analysis, through imputing an appropriate label value for each censored instance, a new representation of the original survival data can be learned effectively. This approach fills the gap in the current literature by estimating the calibrated time-to-event values for these censored instances by exploiting row-wise and column-wise correlations among censored instances in order to effectively impute them.

6.3. Complex Events

Until now, the discussion in this paper has been primarily focused on survival problems in which each instance can experience only a single event of interest. However, in many real-world domains, each instance may experience different types of events and each event may occur more than once during the observation time period. For example, in the healthcare domain, one patient may be hospitalized multiple times due to different medical conditions. Since this scenario is more complex than the survival problems we discussed before, we consider them to be complex events. In this section, we will discuss two techniques, namely, competing risks and recurrent events, to tackle such complex events.

6.3.1. Competing Risks. In the survival problem, if several different types of events are considered, but only one of them can occur for each instance over the follow-up period, then the competing risks will be defined as the probabilities of different events. In other words, the competing risks will only exist in survival problems with more than one possible event of interest, but only one event will occur at any given time. For example, in healthcare domain, a patient may have both heart attack and lung cancer before his death, but the reason of his death can be either lung cancer or heart attack, but not both. In this case, competing risks are the events that prevent an event of interest from occurring which is different from censoring. It should be noted that in the case of censoring, the event of interest still occurs at a later time, while the event of interest is impeded.

To solve this problem, the standard way is to analyze each of these events separately using the survival analysis approach by considering other competing events as censored [Kleinbaum and Klein 2006]. However, there are two primary drawbacks with such an approach. One problem is that this method assumes that the competing

risks are independent of each other. In addition, it would be difficult to interpret the survival probability estimated for each event separately by performing the survival analysis for each event of interest in the competing risks.

To overcome these drawbacks, two methods are developed in the survival analysis literature: Cumulative Incidence Curve (CIC) Approach and Lunn-McNeil (LM) Approach.

Cumulative Incidence Curve (CIC) Approach: To avoid the questionable interpretation problem, the cumulative incidence curve [Putter et al. 2007] is one of the main approaches for competing risks which estimates the marginal probability of each event q . The CIC is defined as

$$CIC_q(t) = \sum_{j:t_j \leq t} \hat{S}(t_{j-1}) \hat{h}_q(t_j) = \sum_{j:t_j \leq t} \hat{S}(t_{j-1}) \frac{n_{qj}}{n_j} \quad (29)$$

where $\hat{h}_q(t_j)$ represents the estimated hazard at time t_j for event q ($q = 1, \dots, Q$), n_{qj} is the number of events for the event q at t_j , n_j denotes the number of instances who are at the risk of experiencing events at t_j , and $\hat{S}(t_{j-1})$ denotes the survival probability at last time point t_{j-1} .

Lunn-McNeil (LM) Approach [Lunn and McNeil 1995]: It is an alternative approach to analyze the competing risks in the survival problems and it also allows the flexibility to conduct statistical inference from the features in the competing risk models. It fits a single Cox PH model which considers all the events in competing risks rather than separate models for each event [Kleinbaum and Klein 2006]. It should be noted that the LM approach is implemented using an augmented data, in which a dummy variable is created for each event to distinguish different competing risks.

6.3.2. Recurrent Events. In many application domains, the event of interest in survival problems may occur several times during the observation time period. This is significantly different from the death of the patients in healthcare domain. In such cases, the outcome event can occur for each instance more than once during the observation time period. In survival analysis, we refer to such events which occur more than once as *recurrent events*, which contrasts with the competing risks discussed above. Typically, if all the recurring events for each instance are of the same type, the counting process (CP) algorithm [Andersen et al. 2012] can be used to tackle this problem. If there are different types of events or the order of the events is the main goal, other methods using stratified Cox (SC) approaches can be used [Ata and Sözer 2007]. These methods include stratified CP, Gap Time and Marginal approach. These approaches differ not only in the way they determine the risk set but also in the data format.

Counting Process: In Counting Process method, the data processing procedure is as follows: (i) For each instance, identify the time interval for each recurrent event and add one record to the data. It should be noted that an additional record for the event-free time interval should also be included for each instance. (ii) For each instance, each record of data should be labeled by a starting time and ending time of the corresponding time interval. These properties of the data format distinguish the counting process method from other methods. They are significantly different from the regular survival data format for non-recurrent event problems, which provides only the ending time and contain only one record for each instance in the dataset.

The key idea to analyze the survival data with recurrent events is to treat the different time intervals for each instance as independent records from different instances. The basic Cox model is used to perform the counting process approach. Each instance will not be removed from the risk set until the last time interval during the observation period. In other words, for the survival problem with recurrent events, the partial

likelihood function formula is different from that in the non-recurrent event survival problems [Kleinbaum and Klein 2006].

Stratified Cox: Stratified CP [Prentice et al. 1981], Marginal [Wei et al. 1989] and Gap Time [Prentice et al. 1981] are three approaches using stratified Cox method to differentiate the event occurrence order. (1) In Stratified CP approach, the data format is exactly the same as that used in the CP approach, and the risk set for the future events is affected by the time of the first event. (2) In Marginal approach, it uses the same data format as the non-recurrent event survival data. This method considers the length of the survival time from the starting time of the follow-up until the time of a specific event occurrence and it assumes that each event is independent of other events. For the k^{th} event ($k = 1, 2, \dots$) in this method, the risk set contains those instances which are at the risk of experience the corresponding event after their entry into the observation. (3) In the Gap Time approach, the data format (start, stop) is used, but the starting time for each data record is 0 and the ending time is the length of the interval from the previous experienced event. In this method, the risk set for the future events will not be affected by the time of the first event.

7. APPLICATION DOMAINS

In this section, we will demonstrate the applications of survival analysis in various real-world domains. Table V summarizes the events of interest, expected goal and the features that are typically used in each specific application described in this section.

7.1. Healthcare

In the healthcare domain, the starting point of the observation is usually a particular medical intervention such as a hospitalization admission, the beginning of taking a certain medication or a diagnosis of a given disease [Klein and Moeschberger 2005; Miller Jr 2011]. The event of interest might be death, hospital readmission, discharge from the hospitalization or any other interesting incident that can happen during the observation period. The missing trace of the observation is also an important characteristic of the data collected in this domain. For example, during a given hospitalization, some patients may be moved to another hospital and in such cases, that patient will become unobserved from the study with respect to the first hospital after that time point. In healthcare applications, survival prediction models primarily aim at estimating the failure time distribution and the prognostic evaluation of different features, including histological, biochemical and clinical characteristics [Marubini and Valsecchi 2004].

7.2. Reliability

In the field of reliability, it is a common practice to collect data over a period of time and record the interesting events that occur within this period. Reliability prediction focuses on developing methods which are good at accurately estimating the reliability of the new products [Modarres et al. 2009; Lyu 1996]. The event of interest here corresponds to the time taken for a device to fail. In such applications, it is desirable to be able to estimate which devices will fail and if they do, when they will fail. Survival analysis methods can help in building such prediction models using the available information about these devices. These models can provide early warnings about potential failures, which is significantly important to either prevent or reduce the likelihood of failures and to identify and correct the causes of device failures.

7.3. Crowdfunding

In recent years, the topic of crowdfunding has gained a lot of attention. Although the crowdfunding platforms have been successful, the percentage of the projects which

Table V: Summary of various real-world application domains where survival analysis was successfully used.

Application	Event of interest	Estimation	Features
Healthcare [Miller Jr 2011] [Reddy and Li 2015]	Rehospitalization Disease recurrence Cancer survival	Likelihood of hospitalization within t days of discharge.	Demographics: age, gender, race. Measurements: height, weight, disease history, disease type, treatment, comorbidities, laboratory, procedures, medications.
Reliability [Lyu 1996] [Modarres et al. 2009]	Device failure	Likelihood of a device being failed within t days.	Product: model, years after production, product performance history. Manufactory: location, no. of products, average failure rate of all the products, annual sale of the product, total sale of the product. User: user reviews of the product.
Crowdfunding [Rakesh et al. 2016] [Li et al. 2016a]	Project success	Likelihood of a project being successful within t days.	Projects: duration, goal amount, category. Creators: past success, location, no. of projects. Twitter: no. of promotions, backings, communities. Temporal: no. of backers, funding, no. of retweets.
Bioinformatics [Li et al. 2016d] [Beer et al. 2002]	Cancer survival	Likelihood of cancer within time t .	Clinical: demographics, labs, procedures, medications. Genomics: gene expression measurements.
Student Retention [Murtaugh et al. 1999] [Ameri et al. 2016]	Student dropout	Likelihood of a student being dropout within t days.	Demographics: age, gender, race. Financial: cash amount, income, scholarships. Pre-enrollment: high-school GPA, ACT scores, graduation age. Enrollment: transfer credits, college, major. Semester performance: semester GPA, % passed credits, % dropped credits.
Customer Lifetime Value [Zeithaml et al. 2001] [Berger and Nasr 1998]	Purchase behavior	Likelihood of a customer purchasing from a given service supplier within t days.	Customer: age, gender, occupation, income, education, interests, purchase history. Store/Online store: location, customer review, customer service, price, quality, shipping fees and time, discount.
Click Through Rate [Yin et al. 2013] [Barbieri et al. 2016]	User clicking	Likelihood of a user clicking the advertisement within time t .	User: gender, age, occupation, interests, users click history. Advertisement (ad): time of the ad, location of the ad on the website, topics of the ad, ad format, total click times of the ad. Website: no. of users of the website, page view each day of the website, no. of websites linking to the website.
Unemployment Duration in Economics [Kiefer 1988]	Getting a job	Likelihood of a person finding a new job within t days.	People: age, gender, major, education, occupation, work experience, city, expected salary. Economics: job openings, unemployment rates every year.

achieved their desired goal amount is less than 50% [Rakesh et al. 2015]. Moreover, many of the prominent crowdfunding platforms follow the “all-or-nothing” policy. In other words, if the goal is achieved before the pre-determined time period, the pledged

funding can be collected. Therefore, in the crowdfunding domain, one of the most important challenges is to estimate the success probability of each project. The need to estimate the project success probability motivates the development of new prediction approaches which can integrate the advantages of both regression (for estimating the time for success) and classification (for considering both successful and failed projects simultaneously in the model) [Li et al. 2016a]. For the successful projects, the time to the success can be collected easily. However, for the projects that failed, it is not possible to collect the information about the length of the time for project success. The only information that can be collected is the funding amount that they raised until the pre-determined project end date. The authors in [Li et al. 2016a] consider both the failed and successful projects simultaneously by using censored regression methods. It fits the probability of project success with log-logistic and logistic distributions and predicts the time taken for a project to become potentially successful.

7.4. Bioinformatics

One of the most popular applications of survival analysis in the domain of bioinformatics is gene expression. Gene expression is the process of synthesizing a functional gene product from the gene information and can be quantified by measuring either message RNA (mRNA) or proteins. Gene expression profiling is developed as a powerful technique to study the cell transcriptome. In recent years, multiple studies [Li et al. 2016d; Beer et al. 2002] have correlated gene expression with survival outcomes in cancer applications in a genome-wide scale. Survival analysis methods are helpful in assessing the effect of single gene on survival prognosis and then identifying the most relevant genes as biomarkers for patients. In this scenario, the event of interest is the specific type of cancer (or any disease), and the goal is to estimate the likelihood of cancer using the gene expression measurements values. Generally, the survival prediction based on gene expression data is a high-dimensional problem since each cell contains tens of thousands of mRNA molecules. The authors in [Antonov et al. 2014] developed a statistical tool for biomedical researchers to define the clinical relevance of genes under investigation via their effect on the patient survival outcome. Survival analysis methods have shown to be effective in predicting the gene expression for different cancer data with the censored information.

7.5. Student Retention

In higher education, student retention rate can be evaluated by the percentage of students who return to the same university for the following semester after completing a semester of study. In the U.S. and around the world, one of the long-term goals of a university is to improve the student retention. Higher the student retention rate, more probable for the university to be positioned higher, secure more government funds, and have an easier path to program accreditations. In view of these reasons, directors and administrators in higher education constantly try to implement new strategies to increase student retention. Survival analysis has success in student retention problem [Murtaugh et al. 1999; Ameri et al. 2016]. The goal of survival analysis is to estimate the time of event occurrence, which is critical in student retention problems because both correctly identifying whether a student will dropout and estimating when the dropout will happen are important. In such cases, it will be helpful if one can reliably estimate the dropout risk at the early stage of student education using both pre-enrollment and post-enrollment information.

7.6. Customer Lifetime Value

Customer lifetime value (LTV) [Berger and Nasr 1998; Zeithaml et al. 2001] of a customer refers to the profit that the customer brings to the store based on the purchase

history. In the marketing domain, the customer LTV is used to evaluate the relationships between the customers and the store. It is important for a store to improve the LTV in order to maintain or increase its profits in the long term since it is often quite expensive to acquire new customers. In this case, the main goal of this problem is to identify purchase patterns of the customers who have a high LTV and provide recommendations for a relatively new user who has similar interest. Identifying loyal customers using LTV estimation has been studied by various researchers [Rosset et al. 2003; Mani et al. 1999] using survival analysis methods and data mining approaches which are helpful in identifying the purchase patterns. Then, LTV will be defined using a survival function, which can be used to estimate the time of purchase for every customer from a given store, using the store information and also the available customer demographic information in the store database, such as the gender, income and age.

7.7. Click-Through Rate

Nowadays, many free web services, including online news portals, search engines and social networks present users with advertisements [Barbieri et al. 2016]. Both the topics and the display orders of the ads will affect the user clicking probability [Richardson et al. 2007]. Studies have mostly focused on predicting the click-through rate (CTR) which indicates the percentage of the users who click on a given ad. It can be calculated as the ratio of the clicking times and the corresponding presentation times (no. of ad impressions). The CTR value indicates the attraction effect of the ad to the users [Barbieri et al. 2016]. The goal is to predict how likely the user will click the ads based on the available information about the website, users and ads. The time taken to click the ad is considered to be the event time. Those users who did not click on the ads are considered to be censored observations.

7.8. Duration Modeling in Economics

Traditionally, duration data which measures how long individuals remain in a certain state is analyzed in biometrics and medical statistics using survival analysis methods. Actually, duration data also appears in a wide variety of situations in economics such as unemployment duration, marital instability and time-to-transaction in the stock market. Among them, the unemployment duration problem, which is the most widely studied one, analyzes the time people spend without a job [Gamerman and West 1987]. Generally, in the domain of economics, the time of being unemployed is extremely important since the length of unemployment of people plays a critical role in economics theories of job search [Kiefer 1988]. For this problem, the data contains information on the time duration of unemployment for each individual in the sample. The event of interest here is getting a new job for each person and the objective is to predict the likelihood of getting a new job within a specific time period. It is desirable to understand how the re-employment probability changes over the period of the spell and to know more about the effect of the unemployment benefits on these probabilities.

8. RESOURCES

This section provides a list of software implementations developed in various statistical methods and machine-learning algorithms for survival analysis. Table VI summarizes the basic information of the software packages for each survival method. We can find that most of the existing survival analysis methods can be implemented in R.

- (1) *Non-parametric methods*: All the three non-parametric survival analysis methods can be implemented by employing the function *coxph* and *survfit* in the *survival* package in R.

Table VI: Summary of software packages for various survival analysis methods.

Algorithm	Software	Language	Link
Kaplan-Meier	survival	R	https://cran.r-project.org/web/packages/survival/index.html
Nelson-Aalen			
Life-Table			
Basic Cox	survival	R	https://cran.r-project.org/web/packages/survival/index.html
TD-Cox			
Lasso-Cox	fastcox	R	https://cran.r-project.org/web/packages/fastcox/index.html
Ridge-Cox			
EN-Cox			
Oscar-Cox	RegCox	R	https://github.com/MLSurvival/RegCox
CoxBoost	CoxBoost	R	https://cran.r-project.org/web/packages/CoxBoost/
Tobit	survival	R	https://cran.r-project.org/web/packages/survival/index.html
BJ	bujar	R	https://cran.r-project.org/web/packages/bujar/index.html
AFT	survival	R	https://cran.r-project.org/web/packages/survival/index.html
Baysian Methods	BMA	R	https://cran.r-project.org/web/packages/BMA/index.html
RSF	randomForestSRC	R	https://cran.r-project.org/web/packages/randomForestSRC/
BST	ipred	R	https://cran.r-project.org/web/packages/ipred/index.html
Boosting	mboost	R	https://cran.r-project.org/web/packages/mboost/
Active Learning	RegCox	R	https://github.com/MLSurvival/RegCox
Transfer Learning	TransferCox	C++	https://github.com/MLSurvival/TransferCox
Multi-Task Learning	MTLSA	Matlab	https://github.com/MLSurvival/MTLSA
Early Prediction	ESP	R	https://github.com/MLSurvival/ESP
Uncensoring			
Calibration	survutils	R	https://github.com/MLSurvival/survutils
Competing Risks	survival	R	https://cran.r-project.org/web/packages/survival/index.html
Recurrent Events	survrec	R	https://cran.r-project.org/web/packages/survrec/

- (2) *Semi-parametric methods*: Both the basic Cox model and the time-dependent Cox model can be trained using *coxph* function in the *survival* package in R. The Lasso-Cox, Ridge-Cox and EN-Cox in the regularized Cox methods can be trained using the *cocktail* function in the *fastcox* package. The *RegCox* package can be used to implement OSCAR-Cox method. The *CoxBoost* function in the *CoxBoost* package can fit a Cox model by likelihood based boosting algorithm.
- (3) *Parametric methods*: The Tobit regression can be trained using the *survreg* function in the *survival* package. Buckley-James Regression can be fitted using the *bujar*

- package. The parametric AFT models can be trained using the *survreg* function with various distributions.
- (4) *Machine learning methods*: The *BMA* package can be used to train a Bayesian model by averaging for Cox models. Bagging survival tree methods can be implemented using the *bagging* function in the R package *ipred*. Random survival forest is implemented in the *rfsrc* function in the package *randomForestSRC*. The *mboost* function in the package *mboost* can be used to implement the boosting algorithm. The *arc* function in *RegCox* package can be used to train the active learning survival model. Transfer-Cox model is written in C++ language. The multi-task learning survival method is implemented using *MTLSA* package in MATLAB.
 - (5) *Related topics*: The *ESP* package which performs the early stage prediction for survival analysis problem also incorporates the uncensoring functions in the data pre-processing part. The *survutils* package in R can be used to implement the calibration for the survival datasets. The *survfit* function in *survival* package can also be used to train the model for competing risks. In addition, the function *Surv* in package *survrec* can be used to train the survival analysis model for recurrent event data.

9. CONCLUSION

The primary goal of survival analysis is to predict the occurrence of specific events of interest at future time points. Due to the widespread availability of longitudinal data from various real-world domains combined with the recent developments in various machine learning methods, there is an increasing demand for understanding and improving methods for effectively handling survival data. In this survey article, we provided a comprehensive review of the conventional survival analysis methods and various machine learning methods for survival analysis, and described other related topics along with the evaluation metrics. We first introduced the basic notations and concepts in survival analysis, including the structure of survival data and the common functions used in survival analysis. Then, we introduced the well-studied statistical survival methods and the representative machine learning based survival methods. Furthermore, the related topics in survival analysis, including data transformation, early prediction and complex events, were also discussed. We also provided the implementation details of these survival methods and described the commonly used performance evaluation metrics for these models. Besides the traditional applications in healthcare and biomedicine, survival analysis was also successfully applied in various real-world problems, such as reliability, student retention and user behavior modeling.

REFERENCES

- Odd Aalen. 1978. Nonparametric inference for a family of counting processes. *The Annals of Statistics* 6, 4 (1978), 701–726.
- Paul D Allison. 2010. *Survival analysis using SAS: a practical guide*. Sas Institute.
- Sattar Ameri, Mahtab J Fard, Ratna B Chinnam, and Chandan K Reddy. 2016. Survival analysis based framework for early prediction of student dropouts. In *Proceedings of ACM International Conference on Conference on Information and Knowledge Management*. ACM, 903–912.
- Per Kragh Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. 2012. *Statistical models based on counting processes*. Springer Science & Business Media.
- A V Antonov, M Krestyaninova, R A Knight, I Rodchenkov, G Melino, and NA Barlev. 2014. PPISURV: a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome. *Oncogene* 33, 13 (2014), 1621–1628.
- Nihal Ata and M Tekin Sözer. 2007. Cox regression models with nonproportional hazards applied to lung cancer survival data. *Hacettepe Journal of Mathematics and Statistics* 36, 2 (2007), 157–167.
- Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. 2012. Structured sparsity through convex optimization. *Statist. Sci.* 27, 4 (2012), 450–468.

- Bart Baesens, Tony Van Gestel, Maria Stepanova, Dirk Van den Poel, and Jan Vanthienen. 2005. Neural network survival analysis for personal loan data. *Journal of the Operational Research Society* 56, 9 (2005), 1089–1098.
- Nicola Barbieri, Fabrizio Silvestri, and Mounia Lalmas. 2016. Improving post-click user engagement on native ads via survival analysis. In *Proceedings of the 25th International Conference on World Wide Web*. 761–770.
- David G Beer, Sharon LR Kardia, Chiang-Ching Huang, Thomas J Giordano, Albert M Levin, David E Misek, Lin Lin, Guoan Chen, Tarek G Gharib, Dafydd G Thomas, and others. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine* 8, 8 (2002), 816–824.
- Riccardo Bellazzi and Blaz Zupan. 2008. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics* 77, 2 (2008), 81–97.
- Paul D Berger and Nada I Nasr. 1998. Customer lifetime value: Marketing models and applications. *Journal of interactive marketing* 12, 1 (1998), 17–30.
- Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. 1998. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine* 17, 10 (1998), 1169–1186.
- Harald Binder and Martin Schumacher. 2008. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC bioinformatics* 9, 1 (2008), 1–10.
- Imad Bou-Hamad, Denis Larocque, Hatem Ben-Ameur, and others. 2011. A review of survival trees. *Statistics Surveys* 5 (2011), 44–71.
- Leo Breiman. 1996. Bagging predictors. *Machine learning* 24, 2 (1996), 123–140.
- Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- Norman E Breslow. 1972. Discussion of the paper by D.R. Cox. *J R Statist Soc B* 34 (1972), 216–217.
- Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.
- Stephen F Brown, Alan J Branford, and William Moran. 1997. On the use of artificial neural networks for the analysis of survival data. *Neural Networks, IEEE Transactions on* 8, 5 (1997), 1071–1077.
- Jonathan Buckley and Ian James. 1979. Linear regression with censored data. *Biometrika* 66, 3 (1979), 429–436.
- Peter Bühlmann and Torsten Hothorn. 2007. Boosting algorithms: Regularization, prediction and model fitting. *Statist. Sci.* 22, 4 (2007), 477–505.
- Harry B Burke, Philip H Goodman, David B Rosen, Donald E Henson, John N Weinstein, Frank E Harrell, Jeffrey R Marks, David P Winchester, and David G Bostwick. 1997. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 79, 4 (1997), 857–862.
- Ching-Fan Chung, Peter Schmidt, and Ana D Witte. 1991. Survival analysis: A survey. *Journal of Quantitative Criminology* 7, 1 (1991), 59–98.
- A. Ciampi, R.S. Bush, M. Gospodarowicz, and J.E. Till. 1981. An approach to classifying prognostic factors related to survival experience for non-Hodgkin's lymphoma patients: Based on a series of 982 patients: 1967–1975. *Cancer* 47, 3 (1981), 621–627.
- A. Ciampi, C-H Chang, S. Hogg, and S. McKinney. 1987. Recursive partition: a versatile method for exploratory-data analysis in biostatistics. In *Biostatistics*. Springer, 23–50.
- Antonio Ciampi, Johanne Thiffault, Jean-Pierre Nakache, and Bernard Asselain. 1986. Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational statistics & data analysis* 4, 3 (1986), 185–204.
- Joseph A Cruz and David S Wishart. 2006. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics* 2 (2006).
- Sidney J Cutler and Fred Ederer. 1958. Maximum utilization of the life table method in analyzing survival. *Journal of chronic diseases* 8, 6 (1958), 699–712.
- Cox R David. 1972. Regression models and life tables. *Journal of the Royal Statistical Society* 34, 2 (1972), 187–220.
- Cox R David. 1975. Partial likelihood. *Biometrika* 62, 2 (1975), 269–276.
- Roger B Davis and James R Anderson. 1989. Exponential survival trees. *Statistics in Medicine* 8, 8 (1989), 947–961.
- Dursun Delen, Glenn Walker, and Amit Kadam. 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine* 34, 2 (2005), 113–127.

- Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.
- Olive J Dunn and Virginia A Clark. 2009. *Basic statistics: a primer for the biomedical sciences*. John Wiley & Sons.
- David Faraggi and Richard Simon. 1995. A neural network model for survival data. *Statistics in medicine* 14, 1 (1995), 73–82.
- Mahtab J Fard, Ping Wang, Sanjay Chawla, and Chandan K Reddy. 2016. A bayesian perspective on early stage event prediction in longitudinal data. *IEEE Transactions on Knowledge and Data Engineering* 28, 12 (2016), 3126–3139.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 3 (2008), 432–441.
- Nir Friedman, Dan Geiger, and Moises Goldszmidt. 1997. Bayesian network classifiers. *Machine learning* 29, 2 (1997), 131–163.
- Dani Gamerman and Mike West. 1987. An application of dynamic survival models in unemployment studies. *The Statistician* (1987), 269–274.
- Louis Gordon and Richard A Olshen. 1985. Tree-structured survival analysis. *Cancer treatment reports* 69, 10 (1985), 1065–1069.
- Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine* 18, 17-18 (1999), 2529–2545.
- Sariel Har-Peled, Dan Roth, and Dav Zimak. 2002. Constraint classification: A new approach to multiclass classification. In *Algorithmic Learning Theory*. Springer, 365–379.
- Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. 1982. Evaluating the yield of medical tests. *Journal of the American Medical Association* 247, 18 (1982), 2543–2546.
- Frank E Harrell, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. 1984. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine* 3, 2 (1984), 143–152.
- Patrick J Heagerty and Yingye Zheng. 2005. Survival model predictive accuracy and ROC curves. *Biometrics* 61, 1 (2005), 92–105.
- Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.
- Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. 2006. Survival ensembles. *Biostatistics* 7, 3 (2006), 355–373.
- Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. 2004. Bagging survival trees. *Statistics in medicine* 23, 1 (2004), 77–91.
- Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. 2008. Random survival forests. *The annals of applied statistics* 2, 3 (2008), 841–860.
- Hemant Ishwaran, Udaya B Kogalur, Xi Chen, and Andy J Minn. 2011. Random survival forests for high-dimensional data. *Statistical analysis and data mining* 4, 1 (2011), 115–132.
- John D Kalbfleisch and Ross L Prentice. 2011. *The statistical analysis of failure time data*. Vol. 360. John Wiley & Sons.
- Edward L Kaplan and Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* 53, 282 (1958), 457–481.
- Carl T Kelley. 1999. *Iterative methods for optimization*. Vol. 18. SIAM.
- Faisal M Khan and Valentina B Zubek. 2008. Support vector regression for censored data (SVRc): a novel tool for survival analysis. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, 863–868.
- Farkhondeh Kiaee, Hamid Sheikhzadeh, and Samaneh Eftekhari Mahabadi. 2016. Relevance vector machine for survival analysis. *IEEE transactions on neural networks and learning systems* 27, 3 (2016), 648–660.
- Nicholas M Kiefer. 1988. Economic duration data and hazard functions. *Journal of economic literature* 26, 2 (1988), 646–679.
- John P Klein and Melvin L Moeschberger. 2005. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.
- David G Kleinbaum and Mitchel Klein. 2006. *Survival analysis: a self-learning text*. Springer Science & Business Media.

- Igor Kononenko. 1993. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal* 7, 4 (1993), 317–337.
- H Koul, V Susarla, and J Van Ryzin. 1981. Regression analysis with randomly right-censored data. *The Annals of Statistics* 9, 6 (1981), 1276–1288.
- Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* 13 (2015), 8–17.
- Minjung Kyung, Jeff Gill, Malay Ghosh, George Casella, and others. 2010. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* 5, 2 (2010), 369–411.
- Michael LeBlanc and John Crowley. 1992. Relative risk trees for censored survival data. *Biometrics* 48, 2 (1992), 411–425.
- Elisa T Lee and John Wang. 2003. *Statistical methods for survival data analysis*. Vol. 476. John Wiley & Sons.
- Yan Li, Vineeth Rakesh, and Chandan K Reddy. 2016a. Project success prediction in crowdfunding environments. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 247–256.
- Yan Li, Bhanukiran Vinzamuri, and Chandan K Reddy. 2016b. Regularized weighted linear regression for high-dimensional censored data. In *Proceedings of SIAM International Conference on Data Mining*. SIAM, 45–53.
- Yan Li, Jie Wang, Jieping Ye, and Chandan K Reddy. 2016d. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1715–1724.
- Yan Li, Lu Wang, Jie Wang, Jieping Ye, and Chandan K Reddy. 2016c. Transfer learning for survival analysis via efficient L2,1-norm regularized Cox regression. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, 231–240.
- Yan Li, Kevin S Xu, and Chandan K Reddy. 2016e. Regularized parametric regression for high-dimensional survival analysis. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 765–773.
- Knut Liestbl, Per Kragh Andersen, and Ulrich Andersen. 1994. Survival analysis and neural nets. *Statistics in medicine* 13, 12 (1994), 1189–1200.
- Paulo JG Lisboa, H Wong, P Harris, and Ric Swindell. 2003. A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial intelligence in medicine* 28, 1 (2003), 1–25.
- Mary Lunn and Don McNeil. 1995. Applying Cox regression to competing risks. *Biometrics* (1995), 524–532.
- Michael R Lyu. 1996. *Handbook of software reliability engineering*. Vol. 222. IEEE computer society press CA.
- David JC MacKay. 1995. Probable networks and plausible predictions—a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* 6, 3 (1995), 469–505.
- DR Mani, James Drew, Andrew Betz, and Piew Datta. 1999. Statistics and data mining techniques for lifetime value modeling. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 94–103.
- L Mariani, D Coradini, E Biganzoli, P Boracchi, E Marubini, S Pilotti, B Salvadori, R Silvestrini, U Veronesi, R Zucali, and others. 1997. Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. *Breast cancer research and treatment* 44, 2 (1997), 167–178.
- Ettore Marubini and Maria Grazia Valsecchi. 2004. *Analysing survival data from clinical trials and observational studies*. Vol. 15. John Wiley & Sons.
- Rupert Miller and Jerry Halpern. 1982. Regression with censored data. *Biometrika* 69, 3 (1982), 521–531.
- Rupert G Miller Jr. 2011. *Survival analysis*. Vol. 66. John Wiley & Sons.
- Mohammad Modarres, Mark P Kaminskiy, and Vasilii Krivtsov. 2009. *Reliability engineering and risk analysis: a practical guide*. CRC press.
- Paul A Murtaugh, Leslie D Burns, and Jill Schuster. 1999. Predicting the retention of university students. *Research in higher education* 40, 3 (1999), 355–371.
- Wayne Nelson. 1972. Theory and applications of hazard plotting for censored failure data. *Technometrics* 14, 4 (1972), 945–966.
- Sinno J Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2010), 1345–1359.

- Michael J Pencina and Ralph B D'Agostino. 2004. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine* 23, 13 (2004), 2109–2123.
- Margaret S Pepe. 2003. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.
- James L Powell. 1994. Estimation of semiparametric models. *Handbook of econometrics* 4 (1994), 2443–2521.
- Ross L Prentice, Benjamin J Williams, and Arthur V Peterson. 1981. On the regression analysis of multivariate failure time data. *Biometrika* 68, 2 (1981), 373–379.
- Hein Putter, M Fiocco, and R. B. Geskus. 2007. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine* 26, 11 (2007), 2389–2430.
- Adrian Raftery, David Madigan, and Chris T. Volinsky. 1995. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics* 5 (1995), 323–349.
- Adrian E Raftery. 1995. Bayesian model selection in social research. *Sociological methodology* 25 (1995), 111–163.
- Vineeth Rakesh, Jaegul Choo, and Chandan K Reddy. 2015. Project recommendation using heterogeneous traits in crowdfunding. In *Ninth International AAAI Conference on Web and Social Media*. 337–346.
- Vineeth Rakesh, Wang-Chien Lee, and Chandan K Reddy. 2016. Probabilistic group recommendation model for crowdfunding domains. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 257–266.
- Peter M Ravdin and Gary M Clark. 1992. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment* 22, 3 (1992), 285–293.
- Chandan K Reddy and Yan Li. 2015. A review of clinical prediction models. *Healthcare Data Analytics* 36 (2015), 343–378.
- Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 521–530.
- Frank Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* 65, 6 (1958), 386–408.
- Saharon Rosset, Einat Neumann, Uri Eick, and Nurit Vatnik. 2003. Customer lifetime value models for decision support. *Data mining and knowledge discovery* 7, 3 (2003), 321–339.
- S Rasoul Safavian and David Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics* 21, 3 (1991), 660–674.
- Mark R Segal. 1988. Regression trees for censored data. *Biometrics* 44, 1 (1988), 35–47.
- Pannagadatta K Shivaswamy, Wei Chu, and Martin Jansche. 2007. A support vector approach to censored targets. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, 655–660.
- Noah Simon, Jerome Friedman, Trevor Hastie, Rob Tibshirani, and others. 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of statistical software* 39, 5 (2011), 1–13.
- Alex J Smola and Bernhard Schölkopf. 1998. *Learning with kernels*. <https://mitpress.mit.edu/books/learning-kernels>.
- Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14, 3 (2004), 199–222.
- Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, and Vikas C Raykar. 2008. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*. 1209–1216.
- Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58, 1 (1996), 267–288.
- Robert Tibshirani. 1997. The lasso method for variable selection in the Cox model. *Statistics in medicine* 16, 4 (1997), 385–395.
- Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 1 (2005), 91–108.
- James Tobin. 1958. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society* 26, 1 (1958), 24–36.

- Belle Van, Kristiaan Pelckmans, Huffel S Van, and Johan AK Suykens. 2011. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial intelligence in medicine* 53, 2 (2011), 107–118.
- Belle V Van, Kristiaan Pelckmans, Johan AK Suykens, and S Van Huffel. 2007. Support vector machines for survival analysis. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*. 1–8.
- Hans C. van Houwelingen and Hein Putter. 2011. *Dynamic prediction in clinical survival analysis*. CRC Press.
- Pierre JM Verweij and Hans C Van Houwelingen. 1994. Penalized likelihood in Cox regression. *Statistics in medicine* 13, 23-24 (1994), 2427–2436.
- Bhanukiran Vinzamuri, Yan Li, and Chandan K Reddy. 2014. Active learning based survival regression for censored data. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. ACM, 241–250.
- Bhanukiran Vinzamuri, Yan Li, and Chandan K Reddy. 2017. Calibrated survival analysis using regularized inverse covariance estimation for right censored data. *Submitted* (2017).
- Bhanukiran Vinzamuri and Chandan K Reddy. 2013. Cox regression with correlation based regularization for electronic health records. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, 757–766.
- Sijian Wang, Bin Nan, Ji Zhu, and David G Beer. 2008. Doubly penalized Buckley–James method for survival data with high-dimensional covariates. *Biometrics* 64, 1 (2008), 132–140.
- Lee-Jen Wei, Danyu Y Lin, and Lisa Weissfeld. 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association* 84, 408 (1989), 1065–1073.
- Achmad Widodo and Bo-Suk Yang. 2011. Application of relevance vector machine and survival probability to machine degradation assessment. *Expert Systems with Applications* 38, 3 (2011), 2592–2599.
- Sen Yang, Lei Yuan, Ying-Cheng Lai, Xiaotong Shen, Peter Wonka, and Jieping Ye. 2012. Feature grouping and selection over an undirected graph. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 922–930.
- Jieping Ye and Jun Liu. 2012. Sparse methods for biomedical data. *ACM SIGKDD Explorations Newsletter* 14, 1 (2012), 4–15.
- Peifeng Yin, Ping Luo, Wang-Chien Lee, and Min Wang. 2013. Silence is also evidence: interpreting dwell time for recommendation from psychological perspective. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 989–997.
- Valarie A Zeithaml, Katherine N Lemon, and Roland T Rust. 2001. *Driving customer equity: How customer lifetime value is reshaping corporate strategy*. Simon and Schuster.
- Hao H Zhang and Wenbin Lu. 2007. Adaptive Lasso for Cox’s proportional hazards model. *Biometrika* 94, 3 (2007), 691–703.
- Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 2 (2005), 301–320.
- Blaž Zupan, Janez Demšar, Michael W Kattan, Robert J Beck, and Ivan Bratko. 2000. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine* 20, 1 (2000), 59–75.