Machine Learning (2021) 110:349–391
https://doi.org/10.1007/s10994-020-05930-3

Kernel machines for current status data

Yael Travis-Lumer[1] · Yair Goldberg[1]

Received: 17 July 2019 / Revised: 22 September 2020 / Accepted: 4 November 2020 /
Published online: 30 November 2020

Abstract
In survival analysis, estimating the failure time distribution is an important and difficult task, since usually the data is subject to censoring. Specifically, in this paper we consider current status data, a type of data where the failure time cannot be directly observed. The format of the data is such that the failure time is restricted to knowledge of whether or not the failure time exceeds a random monitoring time. We propose a flexible kernel machine approach for estimation of the failure time expectation as a function of the covariates, with current status data. In order to obtain the kernel machine decision function, we minimize a regularized version of the empirical risk with respect to a new loss function. Using finite sample bounds and novel oracle inequalities, we prove that the obtained estimator converges to the true conditional expectation for a large family of probability measures. Finally, we present a simulation study and an analysis of real-world data that compares the performance of the proposed approach to existing methods. We show empirically that our approach is comparable to current state of the art, and in some cases is even better.

Keywords Kernel machines · Oracle inequalities · Support vector regression · Survival analysis · Universal consistency

1 Introduction

In this paper we aim to develop a general model free method for analyzing current status data using machine learning techniques. In particular, we propose a kernel machine learning method for estimation of the failure time expectation with current status data. Kernel machines, also known as support vector machines, were originally introduced by Vapnik in the 1990's and are firmly related to statistical learning theory (Vapnik 1999). Kernel machines are learning algorithms that utilize positive definite kernels (Hofmann et al. 2008). The choice of kernel machines for current status data is motivated by the fact that kernel machines can be implemented easily, have fast training speed, produce decision functions that have a strong generalization ability, and can guarantee

Editor: Jean-Philippe Vert.

✉ Yael Travis-Lumer
travis-lumer@campus.technion.ac.il

[1] The Faculty of Industrial Engineering and Management, Technion, 3200003 Haifa, Israel

convergence to the optimal solution, under some weak assumptions (Shivaswamy et al. 2007).

The format of current status data is such that the failure time T is restricted to knowledge of whether or not T exceeds a random monitoring time C. Current status data is also known in the literature as type I interval censored data (Huang and Wellner 1997). This data format is quite common and includes examples from various fields. Jewell and van der Laan (2004) mention a few examples including: studying the distribution of the age of a child at weaning given observation points; when conducting a partner study of HIV infection over a number of clinic visits; and when a tumor under investigation is occult and an animal is sacrificed at a certain time point in order to determine presence or absence of the tumor. For instance, in the last example, when performing carcinogenicity testing, T is the time from exposure to a carcinogen and until the presence of a tumor, and C is the time point at which the animal is sacrificed in order to determine presence or absence of the tumor. Clearly, it is difficult to estimate the failure time distribution since we cannot observe the failure time T. These examples illustrate the importance of this topic and the need to find advanced tools for analyzing such data.

There are several approaches for analyzing current status data. Traditional methods include parametric models where the underlying distribution of the survival time is assumed to be known, such as Weibull, Gamma, and other distributions with non-negative support. Other approaches include semiparametric models, such as the Cox proportional hazards model, and the accelerated failure time (AFT) model (see, for example, Klein and Moeschberger 2005). Several works including Diamond et al. (1986), Shiboski and Jewell (1992), Jewell and van der Laan (2004) and others, have suggested the Cox proportional hazard model for current status data, where the Cox model can be represented as a generalized linear model with a log-log link function. Other works, including Tian and Cai (2006), discussed the use of the AFT model for current status data and suggested different algorithms for estimating the model parameters. Additional semiparametric regression models for current status data include proportional odds (Rossini and Tsiatis 1996), additive hazards (Lin et al. 1998), additive transformations (Cheng and Wang 2011), linear transformations (Sun and Sun 2005), and linear regression (Shen 2000). Needless to say that both parametric and semiparametric models demand stringent assumptions on the distribution of interest which can be restrictive. For this reason, additional estimation methods are needed.

Nonparametric methods for analyzing current status data were also investigated in the literature. Nonparametric maximum likelihood estimation (NPMLE) of the failure time distribution function is commonly used with this type of data, and relies on the PAV algorithm of Ayer et al. (1955). Burr and Gomatam (2002) studied nonparametric estimation of the conditional distribution function of the failure time given the covariates, based on a locally smoothed modification of the NPMLE. Peto (1973), followed by Turnbull (1976), both suggested a non-parametric estimator of the survival function, with interval censored data, that can also be applied to current status data. Dehghan and Duchesne (2011) generalized Turnbull's estimator to include also covariates, by adding weights that depend on a univariate kernel function. Wang et al. (2012) studied nonparametric estimation of the marginal distribution function of the failure time using the copula model approach. Honda (2004) constructed an estimator for the regression function utilizing a modification of maximum rank correlation, and estimated the difference between the regression function at some value, to the regression function at a standard fixed point. Note that these works are not specifically intended for estimation of the conditional expectation and thus might not yield accurate estimates.

Over the past two decades, some learning algorithms for censored data have been proposed. However, most of these algorithms cannot be applied to current status data but only to other, more common, censored data formats. A few recent exceptions include Fu and Simonoff (2017), which studied survival trees for interval-censored data, and the subsequent works of Yao et al. (2019) and Cho et al. (2020), which proposed random forests for interval censored data. Over the last decade, several authors suggested the use of kernel machines, or similarly support vector machines, for survival data, including Van Belle et al. (2007), Khan and Zubek (2008), Eleuteri and Taktak (2011), Shiao and Cherkassky (2013), Wang et al. (2016), Pölsterl et al. (2016), and Goldberg and Kosorok (2017). These examples illustrate that initial steps in this direction have already been taken. However, as far as we know, the only work based on kernel machines that can also be applied to current status data is by Shivaswamy et al. (2007) which has a more computational and less theoretic nature. The authors studied the use of kernel machines for regression problems with interval censoring and, using simulations, showed that the method is comparable to other missing data tools.

We present a kernel machine framework for current status data. We propose a learning method, denoted by KM-CSD, for estimation of the failure time conditional expectation. We investigate the theoretical properties of the KM-CSD, and in particular, prove consistency for a large family of probability measures. In order to estimate the conditional expectation we use a modified version of the quadratic loss, using the methodology of van der Laan and Robins (1998, 2003). Since the failure time T is not observed, our new modified loss function is based on the censoring time C and on the current status indicator. Finally, in order to obtain the KM-CSD estimator, we minimize a regularized version of the empirical risk with respect to our new proposed loss. Note that the terminology decision function is used in the kernel machine context to describe the obtained estimator.

The kernel machine we present in this work may be referred to as an inverse probability weighted complete-case estimator (van der Laan and Robins 2003; Tsiatis 2006, Chapter 6). It is tempting to use the tools described in these books to derive doubly-robust kernel machine estimators. In the context of estimating equations with missing data, doubly-robust estimators are typically constructed by adding an augmentation term. This term is constructed by projecting the estimating equation onto the augmentation space (see Tsiatis 2006, Section 7.4, and Theorem 10.1). However, in our kernel machine setting, the estimator is obtained as the minimizer of a weighted loss function over a reproducing kernel Hilbert space (RKHS) and thus it is not clear how meaningful it is to project the loss function on the augmentation space. It is also not trivial to add a term to the proposed regularized empirical risk minimization problem in a way that yields a convex optimization problem over an RKHS, which is essential for deriving the results presented in this paper. While doubly-robust estimators for current status data were derived in the semiparametric literature (Andrews et al. 2005), we do not consider such estimators in this work. To the best of our knowledge, the only work that studied doubly-robust estimators in the context of kernel machines was done by Liu and Goldberg (2018), however this was done in the context of missing responses, and cannot be applied to our case.

The contribution of this work includes the development of a nonparametric estimator of the conditional expectation, the development of a kernel machine framework for current status data, the development of new oracle inequalities for censored data, and the study of the theoretical properties and the consistency of the KM-CSD.

The paper is organized as follows. In Sect. 2 we describe the formal setting of current status data and discuss the choice of the quadratic loss for estimating the conditional expectation. In Sect. 3 we present the proposed KM-CSD and its corresponding loss function. Section 4 contains the main theoretical results, including finite sample bounds and consistency. Section 5 contains the simulations and Sect. 6 contains an analysis of real world data. Concluding remarks are presented in Sect. 7. The proofs appear in "Appendix C". The R code for both the algorithm and for the simulations, as well as the artificially censored data from Sect. 6.2, can be found in the Supplementary Materials.

2 Preliminaries

In this section we present the notation used throughout the paper. First we describe the data setting and then we discuss briefly loss functions and risks.

Assume that the data consists of n independent and identically distributed random triplets $D = \{(Z_1, C_1, \Delta_1), \dots, (Z_n, C_n, \Delta_n)\}$. The random vector Z is a vector of covariates that takes its values in a compact set $\mathcal{Z} \subset \mathbb{R}^d$. The failure-time T is non-negative, the random variable C is the non-negative censoring time, where both C and T are contained in the interval $[0, \tau] \equiv \mathcal{Y}$, for some constant $\tau > 0$. The indicator $\Delta = \mathbf{1}\{T \leq C\}$ is the current status indicator at time C, obtaining the value 1 when $T \leq C$, and 0 otherwise. For example, in carcinogenicity testing, an animal is sacrificed at a certain time point in order to determine presence or absence of the tumor. In this example, T is the time from exposure to a carcinogen and until the presence of a tumor, Z can be any explanatory information collected such as the weight of the animal, C is the time point at which the animal is sacrificed, and Δ is the current status indicator at time C (indicating whether the tumor has developed before the censoring time, or not).

We now move to discuss a few definitions of loss functions and risks, following Steinwart and Christmann (2008). Let $(\mathcal{Z}, \mathcal{A})$ be a measurable space and $\mathcal{Y} \subset \mathbb{R}$ be a closed subset. Then a loss function is any measurable function L from $\mathcal{Y} \times \mathbb{R}$ to $[0, \infty)$.

Let $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a probability measure on $\mathcal{Z} \times \mathcal{Y}$. For a measurable function $f : \mathcal{Z} \mapsto \mathbb{R}$, the L-risk of f is defined by $R_{L,P}(f) \equiv E_P\left[L(Y, f(Z))\right] = \int_{Z \times Y} L(y, f(z)) dP(z, y)$. A function f that achieves the minimum L-risk is called a *Bayes decision function* and is denoted by f^*, and the minimal L-risk is called the *Bayes risk* and is denoted by $R^*_{L,P}$. Finally, the empirical L-risk is defined by $R_{L,D}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(z_i))$. It is well known (see, for example, Hastie et al. 2013) that the conditional expectation is the Bayes decision function with respect to the quadratic loss. That is, $E[Y|Z] = f^* = \operatorname{argmin}_f R_{L,P}(f)$, where L is the quadratic loss defined by $L(Y, f(Z)) = (Y - f(Z))^2$.

Recall that our goal is to estimate the conditional expectation of the failure-time T given the covariates Z. However, in the setting of current status data, the response variable (failure-time) is not observed, making the estimation procedure more complex. It is not even clear if and how loss functions can be defined with current status data. In the following section we construct a new modification of the quadratic loss that is based on the censoring time and on the current status indicator, and use it to estimate the conditional expectation of the unobservable failure-time.

3 Kernel machines for current status data

This section is divided into three subsections. We start by describing general kernel machines for uncensored data. Then we define a new loss function for current status data, utilizing an equality between risks, and incorporate it into the kernel machine framework. Finally we define the proposed estimator of the conditional expectation of the failure-time, with current status data, and discuss some assumptions regarding the censoring mechanism.

3.1 Kernel machines for uncensored data

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) of functions from $\mathcal{Z}$ to $\mathbb{R}$, where an RKHS is a function space that can be characterized by some kernel function $k : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$. For more information on reproducing kernel Hilbert spaces, we refer the reader to (Steinwart and Christmann 2008, Chapter 4). A continuous kernel k for which the corresponding RKHS $\mathcal{H}$ is dense in the space of continuous functions on $\mathcal{Z}$, $C(\mathcal{Z})$, is called a universal kernel (see, for example, Steinwart and Christmann 2008, Definition 4.52). Fix such an RKHS $\mathcal{H}$ and denote its norm by $\|\cdot\|_{\mathcal{H}}$. Let $\{\lambda_n\} > 0$ be some sequence of regularization constants. A kernel machine decision function for uncensored data is defined by:

$$f_{D,\lambda_n} = \arg\min_{f\in\mathcal{H}} \lambda_n \|f\|^2_{\mathcal{H}} + \frac{1}{n}\sum_{i=1}^{n} L(T_i, f(Z_i)) .$$

3.2 Equality between risks

In this subsection we show that the risk can be represented as the sum of two terms

$$E\left[\frac{(1-\Delta)\ell(C, f(Z))}{g(C|Z)}\right] + E[L(0, f(Z))].$$

We recall that current status data consists of n independent and identically-distributed random triplets $D = \{(Z_1, C_1, \Delta_1), \dots, (Z_n, C_n, \Delta_n)\}$.

Let $F(\cdot|Z = z)$ and $G(\cdot|Z = z)$ be the cumulative distribution functions of the failure time and censoring, respectively, given the covariates $Z = z$. Let $g(\cdot|Z = z)$ be the density of $G(\cdot|Z = z)$. Throughout this work we will assume the following:

(A1) The censoring time C is independent of the failure time T given the covariates Z.
(A2) C and T take values in the interval $[0, \tau] \equiv \mathcal{Y}$ and $\inf_{z\in\mathcal{Z}, c\in\mathcal{Y}} g(c|z) \geq 2\kappa > 0$, for some $\kappa > 0$.

The conditional independence assumption (A1) is a standard identifiability assumption in survival analysis (see, for example, Klein and Goel 1992; Klein and Moeschberger 2005). Assumption (A2) is needed in order to guarantee that integration with respect to T and C can be exchanged, and in order to allow for division by the censoring density. Similar assumptions were made by van der Laan and Robins (1998).

Let $L : \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty)$ be a loss function differentiable in the first variable. Let $\ell : \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}$ be the derivative of L with respect to the first variable.

For current status data, we introduce the following two sets of identities. In (1) we use integration by parts to show that the risk can be represented as the sum of two terms: $a = E_Z\big[\int_0^\tau \ell(t,f(Z))(1 - F(t|Z))dt\big]$ and $b = E[L(0,f(Z))]$. In (2) we show that the term a is equivalent to $E\Big[\frac{(1-\Delta)\ell(C,f(Z))}{g(C|Z)}\Big]$, which can be seen as a generalization of van der Laan and Robins (1998, 2003) that also includes loss functions and covariates.

We would like to find the minimizer of $R_{L,P}(f)$ over a set $\mathcal{H}$ of functions f. Note that

$$
\begin{aligned}
R_{L,P}(f) &\equiv E_Z E_{T|Z} L(T,f(Z)) = E_Z\left[\int_0^\tau L(t,f(Z))dF(t|Z)\right] \\
&= E_Z\left[\int_0^\tau \ell(t,f(Z))(1 - F(t|Z))dt - L(t,f(Z))(1 - F(t|Z))|_0^\tau\right] \\
&= E_Z\left[\int_0^\tau \ell(t,f(Z))(1 - F(t|Z))dt\right] + E[L(0,f(Z))],
\end{aligned} \tag{1}
$$

and that $(1 - \Delta) = \mathbf{1}\{T > C\}$ and thus

$$
\begin{aligned}
E\left[\frac{(1-\Delta)\ell(C,f(Z))}{g(C|Z)}\right] &= E_{Z,T}\left[E_C\left[\frac{\mathbf{1}\{T > C\}\ell(C,f(Z))}{g(C|Z)}\middle| Z, T\right]\right] \\
&= E_{Z,T}\left[\int_0^\tau \frac{\mathbf{1}\{T > c\}\ell(c,f(Z))g(c|Z)}{g(c|Z)}dc\right] \\
&= E_{Z,T}\left[\int_0^\tau \mathbf{1}\{T > c\}\ell(c,f(Z))dc\right] \\
&= E_Z\left[\int_0^\tau \ell(c,f(Z))\int_0^\tau \mathbf{1}\{t > c\}dF(t|Z)dc\right] \\
&= E_Z\left[\int_0^\tau \ell(c,f(Z))(1 - F(c|Z))dc\right].
\end{aligned} \tag{2}
$$

In summary, we show that the risk can be represented as

$$
R_{L,P}(f) = E\left[\frac{(1-\Delta)\ell(C,f(Z))}{g(C|Z)}\right] + E[L(0,f(Z))]. \tag{3}
$$

The motivation for these equations arises from the fact that we are interested in minimizing the empirical risk, but unfortunately have no observed failure times T. However, we do observe the censoring times C and the current status indicator Δ, and would like to use the observed data for estimation. To that end, we represent the empirical risk as the empirical version of the two terms on the RHS of (3). Such an estimator is known as an inverse probability of censoring weighted (IPCW) average, as it involves re-weighting by the inverse of the censoring density.

3.3 Kernel machines for current status data

Hence, in order to estimate the minimizer of $R_{L,P}(f)$, one can minimize a regularized version of the empirical risk with respect to a new loss function defined by

$$L^n(D,(Z,C,\Delta,s)) = \frac{(1-\Delta)\ell(C,s)}{g(C|Z)} + L(0,s)\,.$$

Note that this function need not be convex nor a loss function. Recall that we are interested in estimating the conditional expectation. This means that we would like to minimize the risk with respect to the quadratic loss. For the quadratic loss, our new loss function becomes

$$L^n(D,(Z,C,\Delta,s)) = \frac{(1-\Delta)2(C-s)}{g(C|Z)} + s^2\,.$$

Note that this function is convex but not necessarily a loss function since it can obtain negative values. However, one can always add a constant to ensure positivity. Since this constant does not effect optimization it will be neglected hereafter. For a detailed explanation, see "Appendix B".

In order to implement this result into the kernel machine framework, we propose to define the KM-CSD decision function for current status data by

$$f_{D,\lambda} = \arg\min_{f\in\mathcal{H}} \lambda\|f\|^2_{\mathcal{H}} + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-\Delta_i)2(C_i - f(Z_i))}{g(C_i|Z_i)} + (f(Z_i))^2\right]. \tag{4}$$

This is a quadratic programming problem that has a closed form solution. The solution requires inverting an $(n+1)\times(n+1)$ symmetric PSD matrix and can be found in "Appendix A".

Note that if the censoring mechanism is unknown, we can replace the density g in (4) with its estimate $\hat{g}$, as long as $\hat{g}$ is strictly positive on $[0,\tau]\equiv\mathcal{Y}$; in this case the kernel machine decision function is

$$f_{D,\lambda} = \arg\min_{f\in\mathcal{H}} \lambda\|f\|^2_{\mathcal{H}} + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-\Delta_i)2(C_i - f(Z_i))}{\hat{g}(C_i|Z_i)} + (f(Z_i))^2\right]$$

(note the use of $\hat{g}$ instead of g in the denominator).

We note that for current status data, the assumption of some knowledge of the censoring distribution is reasonable, for example, when it is chosen by the researcher (Jewell and van der Laan 2004). In other cases, the density can be estimated using either parametric or nonparametric density estimation techniques such as kernel estimates. It should be noted that the censoring variable itself is fully observed (not censored) and thus simple density estimation techniques can be used in order to estimate the density g.

4 Theoretical results

The main goal of our work is to find a 'good' estimator of the failure time conditional expectation. A good estimator should first and foremost be consistent, that is, its risk should converge in probability to the Bayes risk. Additionally, we would like such an estimator to be consistent for a large family of probability measures. The consistency proof is based on novel oracle inequalities that are presented below.

We start by proving risk consistency of the KM-CSD learning method for a large family of probability measures. We first assume that the censoring mechanism is known, which means that the true density of the censoring variable g is known. Using this assumption, and some additional conditions, we bound the difference between the risk of the KM-CSD decision function and the Bayes risk in order to form finite sample bounds. We use this result to show that the KM-CSD converges in probability to the Bayes risk. That is, we demonstrate that for a large family of probability measures, the KM-CSD learning method is consistent. We then consider the case in which the censoring mechanism is unknown, and thus the density g needs to be estimated. We estimate the density g using nonparametric kernel density estimation, and develop a novel finite sample bound. We use this bound to prove that the KM-CSD is consistent even when the censoring distribution is unknown.

For simplicity, we use the normalized version of the quadratic loss.

Definition 1 Let $L(y,s) = \frac{(y-s)^2}{\tau^2}$ be the normalized quadratic loss, let $l(y,s) = 2(y-s)\tau^{-2}$ be its derivative with respect to the first variable, and let $L^n(D,(Z,C,\Delta,s)) = \frac{1}{\tau^2}\left(\frac{(1-\Delta)2(C-s)}{g(C|Z)} + s^2\right)$ be the proposed modified version of this loss.

Since both L and l are convex functions with respect to s, then for any compact set $\mathcal{S} = [-S,S] \subset \mathbb{R}$, Both L and l are bounded and Lipschitz continuous with constants c_L and c_l that depend on $\mathcal{S}$.

Remark 1 $L(y,0) \le 1$ for all $y \in \mathcal{Y}$ and $\ell(y,s) \le B_1$ for all $(y,s) \in \mathcal{Y} \times \mathcal{S}$ and for some constant $B_1 > 0$.

We need the following additional assumptions:

(A3) $\mathcal{Z} \subset \mathbb{R}^d$ is compact,
(A4) $\mathcal{H}$ is an RKHS of a continuous kernel k with $\|k\|_\infty \le 1$.

Assumptions (A3-A4) are standard technical assumptions in the kernel machines literature.

Define the approximation error by $A_2(\lambda) = \inf_{f\in\mathcal{H}} \lambda\|f\|^2_{\mathcal{H}} + R_{L,P}(f) - R^*_{L,P}$. Define $B_2 = c_L\lambda^{-1/2} + 1$ and $B = \frac{B_1}{2\kappa} + B_2$, where B_1 is defined in Remark 1, κ is defined in Assumption (A2), c_L is the Lipschitz constant of the normalized quadratic loss L, and λ is the regularization parameter.

4.1 Case I: the censoring density g is known

In this section we develop finite sample bounds assuming that the censoring density g is known.

Theorem 1 *Assume that* (A1)–(A4) *hold. Then for fixed* $\lambda > 0$, $n \ge 1$, $\varepsilon > 0$, *and* $\theta > 0$, *with probability not less than* $1 - e^{-\theta}$

$$\lambda\|f_{D,\lambda}\|^2_{\mathcal{H}} + R_{L,P}(f_{D,\lambda}) - R^*_{L,P} - A_2(\lambda)$$
$$\le B\sqrt{\frac{2\log\left(2N(\sqrt{\frac{1}{\lambda}}B_H, \|\cdot\|_\infty, \epsilon)\right) + 2\theta}{n}} + \frac{2c_l\varepsilon}{\kappa} + 4c_L\varepsilon$$

where $N(\lambda^{-\frac{1}{2}}B_H, \|\cdot\|_\infty, \epsilon)$ *is the covering number of the* ϵ *− net of* $\sqrt{\frac{1}{\lambda}\overline{B_H}}$ *with respect to the supremum norm and where* $\overline{B_H}$ *is the closure of the unit ball of* $\mathcal{H}$ *(for further details see* Steinwart and Christmann 2008) .

The proof of this theorem appears in "Appendix C.1".

We now move to discuss consistency of the KM-CSD learning method. By definition, P-universal consistency means that for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(D \in (\mathcal{Z}\times\mathcal{Y})^n \ : \ R_{L,P}(f_{D,\lambda_n}) \le R^*_{L,P} + \epsilon) = 1, \tag{5}$$

where $R^*_{L,P}$ is the Bayes risk. Universal consistency means that (5) holds for all probability measures P on $\mathcal{Z}\times\mathcal{Y}$. However, in survival analysis we have the problem of identifiability and thus we will limit our discussion to probability measures that satisfy some identification conditions. Let $\mathcal{P}$ be the set of all probability measures that satisfy Assumptions (A1)–(A2). We say that a learning method is $\mathcal{P}$-universal consistent when (5) holds for all probability measures $P \in \mathcal{P}$.

In order to show $\mathcal{P}$-universal consistency, we utilize the finite sample bounds of Theorem 1. The following assumption is also needed for proving $\mathcal{P}$-universal consistency:

(A5) k is a universal kernel.

Universal kernels are a wide family of kernel functions that include Gaussian and Taylor kernels. A kernel k is called universal if the RKHS $\mathcal{H}$ of k is dense in the space of continuous functions on $\mathcal{Z}$, $C(\mathcal{Z})$, with respect to the sup norm. Assumption (A5) means that $\inf_{f\in\mathcal{H}} R_{L,P}(f) = R^*_{L,P}$, for all probability measures P on $\mathcal{Z}\times\mathcal{Y}$.

Corollary 1 *Assume the setting of Theorem* 1 *and that Assumption* (A5) *holds. Assume that there exist constants* $a \ge 1$ and $p > 0$ such that $\log\left(N(B_H, \|\cdot\|_\infty, \epsilon)\right) \le a\epsilon^{-2p}$. *Let* λ_n *be a sequence such that* $\lambda_n \underset{n\to\infty}{\to} 0$ *and* $\lambda_n^{1+p} n \underset{n\to\infty}{\to} \infty$. *Then the KM-CSD learning method is* $\mathcal{P}$ *-universal consistent.*

The proof of this theorem appears in "Appendix C.2".

Note that the bound on the covering number $N(B_H, \|\cdot\|_\infty, \epsilon)$ in Corollary 1 is satisfied for smooth kernels, such as polynomial and Gaussian kernels, for arbitrarily small $p > 0$ (see Steinwart and Christmann 2008, Section 6.4).

4.2 Case II: the censoring density *g* is unknown

Here we consider the case in which the censoring mechanism is unknown, and thus the density g needs to be estimated. We estimate the density g using nonparametric kernel density estimation, and develop a novel finite sample bound. We use this bound to prove that the KM-CSD is consistent even when the censoring distribution is unknown. Note that asymptotic results for kernel density estimators are well known in the literature (see, for example, Silverman 1978). However, to the best of our knowledge, finite sample bounds for this case do not exist and hence are developed here.

For simplicity, we assume here that the censoring time C is independent of the covariates Z. One can generalize the estimation procedure to include dependence of the censoring time C on the covariates Z; for example, the conditional density estimate can be computed by the ratio of the joint density estimate to the marginal density estimate. In Lemma 1 we construct finite sample bounds on the difference between the estimated density $\hat{g}$ and the true density g. In Theorem 2 we utilize this bound to form finite sample bounds for the KM-CSD learning method.

Definition 2 We say that $K_m : \mathbb{R} \mapsto \mathbb{R}$ (not to be confused with the kernel function k of the RKHS $\mathcal{H}$) is a kernel of order m, if the functions $u \mapsto u^j K_m(u)$ $, j = 0, 1, \dots, m$ are integrable and satisfy $\int_{-\infty}^{\infty} K_m(u)du = 1$ and $\int_{-\infty}^{\infty} u^j K_m(u)du = 0,\ j = 1, \dots, m$.

Definition 3 The Hölder class $\sum(\beta, \mathcal{L})$ of functions $f : \mathbb{R} \mapsto R$ is the set of $m = \lfloor \beta \rfloor$ times differentiable functions whose derivative $f^{(m)}$ satisfies

$$\left| f^{(m)}(x) - f^{(m)}(x') \right| \le \mathcal{L} \left| x - x' \right|^{\beta - m}$$

for any $x, x' \in \mathbb{R}$ and for some constant $\mathcal{L} > 0$.

Lemma 1 *Let $K_m : \mathbb{R} \mapsto \mathbb{R}$ be a kernel function of order $m = \lfloor \beta \rfloor$ satisfying $\int_{-\infty}^{\infty} K_m^2(u)du < \infty$ and define $\hat{g}(x) = (hn)^{-1} \sum_{i=1}^{n} K_m\big((C_i - x)/h\big)$ where h is the bandwidth. Suppose that the true density g* and *its estimate $\hat{g}$ both satisfy $g(c), \hat{g}(c) \le g_{max} < \infty$. Let us also assume that $g(c)$ belongs to the* Höl*der class $\sum(\beta, \mathcal{L})$. Finally, assume that $\int_{-\infty}^{\infty} |u|^{\beta} \big|K_m(u)\big| du < \infty$. Then for any $\theta > 0$,*

$$Pr\left(\frac{1}{n} \sum_{i=1}^{n} \left| \hat{g}(C_i) - g(C_i) \right| > \sqrt{\frac{2D_1\theta}{n^2 h}} + \frac{2g_{max}\theta}{3n} + D_2 \cdot h^{\beta} \right) \le \exp(-\theta)$$

where $D_1 = g_{max} \int_{-\infty}^{\infty} K_m^2(v)dv$ and $D_2 = \mathcal{L}|\pi|^{\beta - m}/m! \int_{-\infty}^{\infty} \big|K_m(v)\big| |v|^{\beta} dv$ are constants, and for some $\pi \in [0, 1]$.

The proof of the lemma is based on Tsybakov (2008, Propositions 1.1 and 1.2) together with basic concentration inequalities; the proof can be found in "Appendix C.3".

We now move to construct finite sample bounds for the KM-CSD learning method when g is unknown using the above lemma. We assume that $\hat{g}$ is the kernel density estimate of g, such that the conditions of Lemma 1 hold.

Theorem 2 *Assume that* (A1)–(A4) *hold. Assume the setting of Lemma* 1 *and that* $\inf_{c \in C} \hat{g}(c) \ge \kappa > 0$, *for some $\kappa > 0$. Then for fixed $\lambda > 0$, $\theta > 0$, $n \ge 1$, $\epsilon > 0$,* we *have with probability not less than* $1 - 2e^{-\theta}$ *that*

$$\begin{aligned} &\lambda \|f_{D,\lambda}\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* - A_2(\lambda) \\ &\le B \sqrt{\frac{2log\left(2N(\sqrt{\frac{1}{\lambda}} B_H, \|\cdot\|_{\infty}, \epsilon) \right) + 2\theta}{n}} + \frac{3c_l\epsilon}{\kappa} + 4c_L\epsilon + 2\eta \end{aligned}$$

where $\eta \equiv \frac{B_1}{2\kappa^2} \left(\sqrt{\frac{2D_1\theta}{n^2 h}} + \frac{2g_{max}\theta}{3n} + D_2 \cdot h^{\beta} \right)$.

The proof of the theorem appears in "Appendix C.4".

Using the above theorem we show that under some mild conditions, the KM-CSD decision function converges in probability to the conditional expectation.

Corollary 2 *Assume the setting of Theorem 2 and assume that Assumption* (A5) *holds. Assume that there exist constants* $a \geq 1$ and $p > 0$ *such that* $\log\left(N(B_H, \|\cdot\|_\infty, \epsilon)\right) \leq a\epsilon^{-2p}$. Let λ_n *be a sequence such that* $\lambda_n \underset{n\to\infty}{\to} 0$ *and that* $\lambda_n^{1+p} n \underset{n\to\infty}{\to} \infty$. *Then the KM-CSD learning method is* $\mathcal{P}$*-universal consistent.*

The proof of this theorem appears in "Appendix C.5".

We refer the readers to "Appendix D" for a straightforward derivation of learning rates that are based on the same oracle inequalities of Theorem 1 and 2.

5 Simulation study

We test the KM-CSD learning method on simulated data and compare its performance to current state of the art. We construct six different data-generating mechanisms, including one-dimensional and multi-dimensional settings. For each data type, we compute the squared difference between the KM-CSD decision function and the true failure time. We compare this result to results obtained by the Cox model, the AFT model, the proportional odds (PO) model, and the survival forests method (ICcforest). As a reference, we compare all these methods to the Bayes risk, which we calculated using the Monte Carlo method.

For each data setting, we considered three cases: (1) the censoring density g is known, (2) the censoring density is unknown, and (3) the censoring density is misspecified. For simplicity, we assumed that the censoring time C is independent of the covariates Z. For the first setting, we used a uniform distribution on $[0, \tau]$ with density $g(C) = \frac{1}{\tau}$. For the second setting, the distribution of the censoring variable was estimated using univariate nonparametric kernel density estimation with a Gaussian kernel. For the third setting, we misspecified the censoring distribution using a beta distribution $Beta(0.9, 0.9)$, rescaled to the interval $[0, \tau]$, with density $g(C) = \frac{1}{\tau}\frac{C^{-0.1}(1-C)^{-0.1}}{\mathbf{B}(0.9,0.9)}$, where $\mathbf{B}(\alpha, \beta)$ is the beta function. To keep the manuscript short, case (3) appears in "Appendix E".

Our code is written as a package in R (R Core Team 2020), and uses the R packages 'kernlab' (Karatzoglou et al. 2019) for kernel function computations, 'ks' (Duong et al. 2020) for kernel density estimates, and 'mlr3' (Lang et al. 2020) and 'mlr3tuning' (Becker et al. 2020) for hyper-parameter tuning. Figures were produced using the package 'ggplot2' (Wickham 2016). In order to fit the Cox model to current status data, we used the state of the art 'ICsurv' R package (McMahan and Wang 2014). In this package, monotone splines are used to estimate the cumulative baseline hazard function, and the model parameters are then chosen via the EM algorithm. We chose the most commonly used cubic splines. To choose the number and locations of the knots, we followed Ramsay (1988) and McMahan et al. (2013) who both suggested using a fixed small number of knots and thus we placed the knots evenly at the quartiles. For the AFT model, we used the 'survreg' function in the 'Survival' R package (Therneau and Lumley 2016), and for the PO model, we used the 'ic_sp' function in the 'icenReg' R package (Anderson-Bergman 2020). For the random forest implementation we used the 'ICcforest' function in the 'ICcforest' R package (Yao et al. 2020).

For the kernel of the RKHS $\mathcal{H}$, we used both a linear kernel and a Gaussian RBF kernel $k(x_i, x_j) = \exp\left(-\left\|x_i - x_j\right\|_2^2/2\sigma^2\right)$, where σ and λ were chosen using fivefold cross-validation. Cross validation is commonly used for kernel machine parameter selection (see, for example, Steinwart and Christmann 2008). Oracle inequalities for penalized risk minimization with multi-fold cross validation were developed by van der Vaart et al. (2006). This result can be applied to kernel machines and justifies the use of cross validation for parameter selection. Since in our case the failure time T is not observed, using cross-validation with current status data is not trivial. Hence we used fivefold cross-validation with respect to the empirical risk obtained by our proposed loss.

We consider the following six failure time distributions, corresponding to the six different data-generating mechanisms: (1) Exponential, (2) Weibull, (3) Multi-Weibull, (4) Multi-Log-Normal, (5) an example where the failure time expectation is triangle shaped, and (6) an example where the failure time expectation is U-shaped. We present below the KM-CSD risks for each case and compare them to risks obtained by other methods. The risks are based on 100 iterations per sample size. The Bayes risk is also plotted as a reference. The Bayes risk was calculated based on the Monte Carlo method where a large number of observations were drawn from the true failure time distribution; the empirical risk was then calculated.

In Setting 1 (Exponential failure-time), the covariates Z are generated uniformly on [0, 1], the censoring variables C is generated uniformly on $[0, \tau]$, and the failure time T is generated from an Exponential distribution with parameter $\lambda = \exp(-0.5Z)$. The failure time was then truncated at $\tau = 3$. Figure 1 compares the results obtained by the KM-CSD to results obtained by the other methods, for the exponential distribution, and for different sample sizes.

In Setting 2 (Weibull failure-time), the covariates Z are generated uniformly on [0, 1], the censoring variables C is generated uniformly on $[0, \tau]$, and the failure time T is generated from a Weibull distribution with parameters $scale = \exp(-0.5Z)$, and $shape = 2$. The failure time was then truncated at $\tau = 1$. The results appear in Fig. 2.

Figures 1 and 2 shows that when g is known, the KM-CSD with a linear kernel produces risks that are comparable to those of the Cox model and the AFT model, and are better than those of the other methods. However, when g is not known, the Cox model produces the smallest risks, but its superiority reduces as the sample size grows. Note that the exponential distribution in Setting 1 and the Weibull distribution in Setting 2 both satisfy the Cox proportional hazards (PH) assumption, and the AFT assumption. In particular, when the PH assumption holds, estimation based on the Cox model is consistent and efficient; hence, when the PH assumption holds, we will use the Cox model as a benchmark.

In Setting 3 (Multi-Weibull failure-time), the covariates Z are generated uniformly on $[0, 1]^{10}$, and the censoring variable C is generated uniformly on $[0, \tau]$, as in Setting 1. The failure time T is generated from a Weibull distribution with parameters $scale = -0.5Z_1 + 2Z_2 - Z_3$ and $shape = 2$. The failure time was then truncated at $\tau = 2$. Note that this model depends only on the first three variables. In Fig. 3, boxplots of risks are presented, for the Weibull distribution, with multivariate covariates. Figure 3 shows that the ICcforest method produces the lowest risks for smaller sample sizes, but the KM-CSD with a linear kernel produces the lowest risks for larger sample sizes. This trend is observed for both cases (i) and (ii); however, the convergence rate of the KM-CSD is

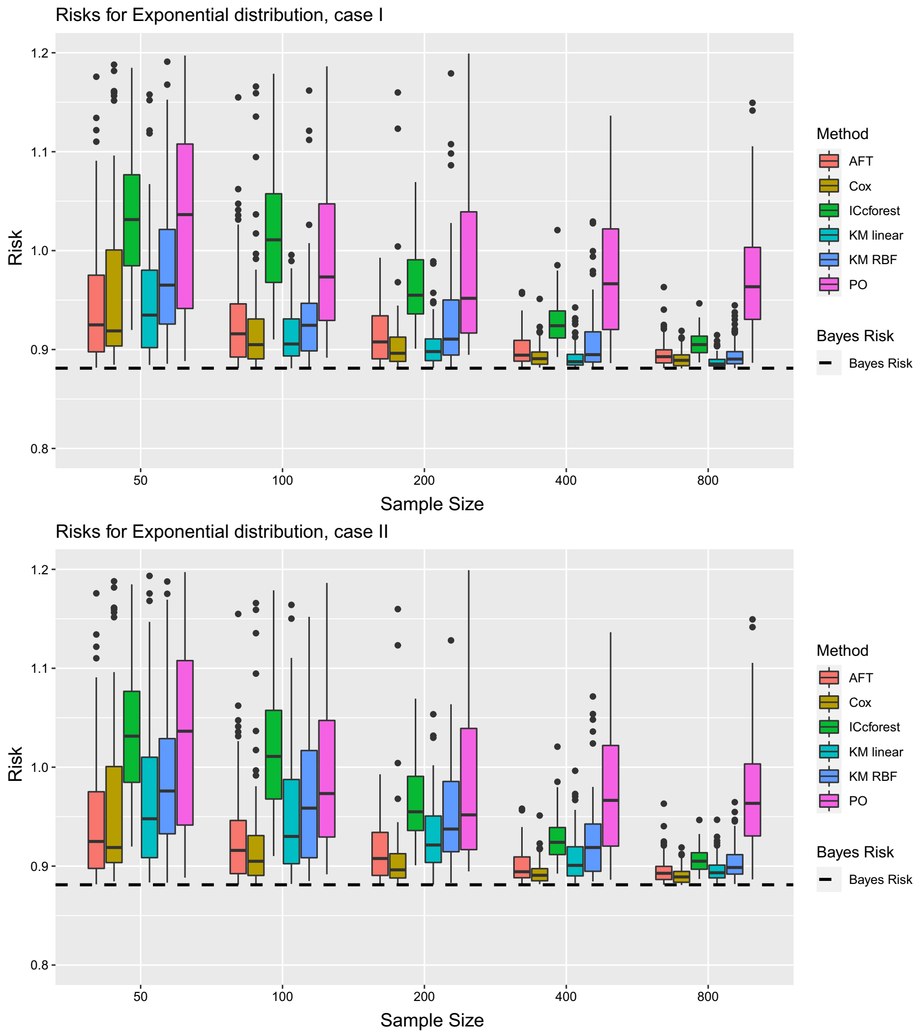


Fig. 1 Exponential failure time distribution.The Bayes risk is the dashed line and the boxplots of the following risks are compared: the KM-CSD with an RBF kernel, the KM-CSD with a linear kernel, AFT, Cox, ICcforest, and PO, for sample sizes $n = 50, 100, 200, 400, 800$

slower for the case where g is unknown, which corresponds to the rates derived in Theorem 3 of "Appendix D".

In Setting 4 (Multi-Log-Normal), the covariates Z are generated uniformly on $[0, 1]^{10}$, C was generated as before and the failure time T was generated from a Log-Normal distribution with parameters $\mu = \frac{1}{2}(0.3Z_1 + 0.5Z_2 + 0.2Z_3)$ and $\sigma = 1$. The failure time was then truncated at $\tau = 7$. Figure 4 presents the risks of the compared methods. Figure 4 shows that the nonparametric methods produce the lowest risks, with a slight preference

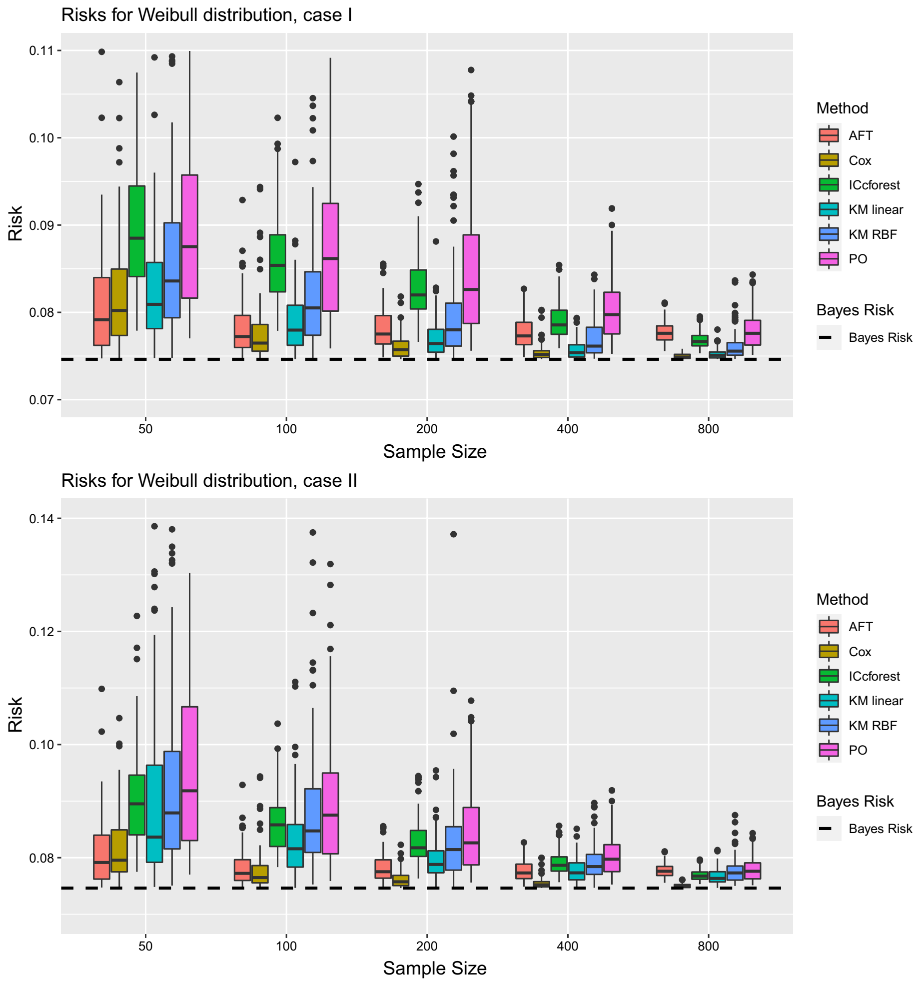


Fig. 2 Weibull failure time distribution. The Bayes risk is the dashed line and the boxplots of the following risks are compared: the KM-CSD with an RBF kernel, the KM-CSD with a linear kernel, AFT, Cox, ICcforest, and PO, for sample sizes $n = 50, 100, 200, 400, 800$

for the KM-CSD with an RBF kernel in case (i), and a slight preference for the ICcforest method in case (ii).

In Setting 5, we considered a non-smooth conditional expectation function in the shape of a triangle. The covariates Z are generated uniformly on $[0, 1]$, C is generated uniformly on $[0, \tau]$, and T is generated according to the following

$$T = \begin{cases} 4 + 6 \cdot Z + \epsilon, & Z \leq 0.5 \\ 10 - 6 \cdot Z + \epsilon, & Z > 0.5 \end{cases}, \quad \text{where } \epsilon \sim N(0, 1).$$

The failure time was then truncated at $\tau = 8$. In Fig. 5, the boxplots of risks are presented. As expected, Fig. 5 shows that for a large enough sample size, the non-parametric methods ICcforest and KM-CSD with an RBF kernel both manage to correctly estimate the

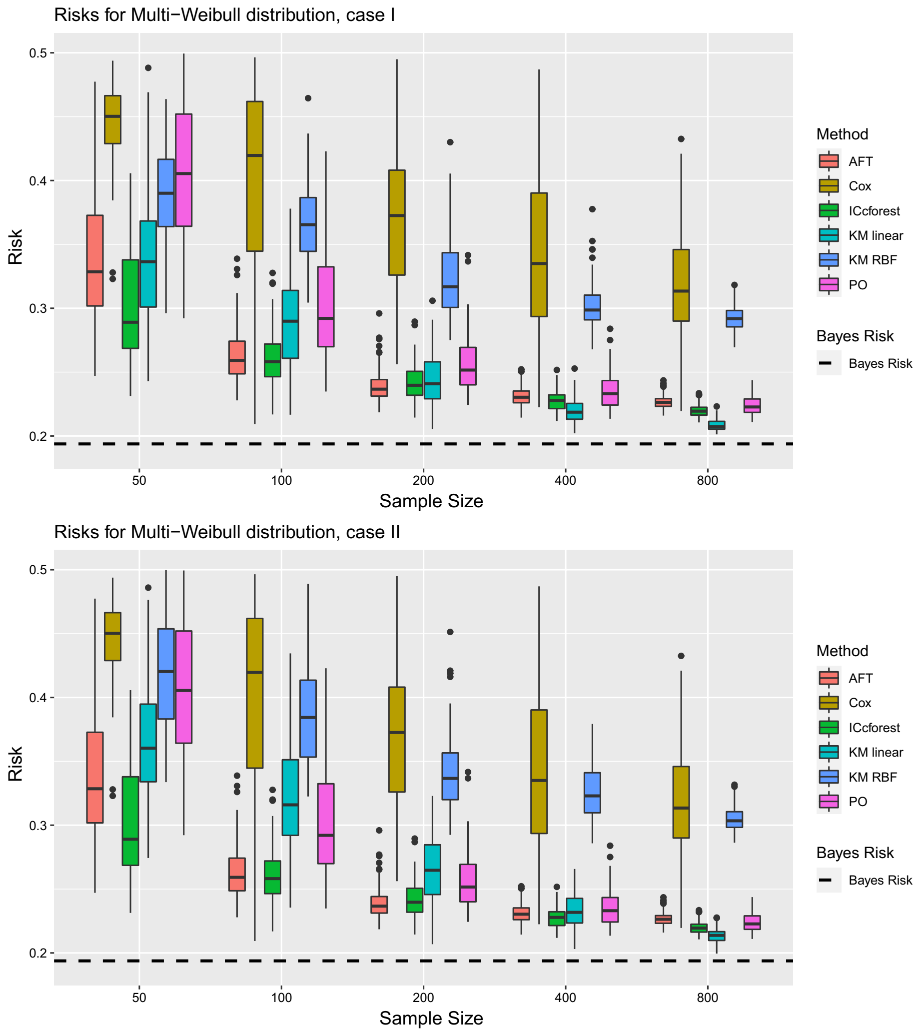
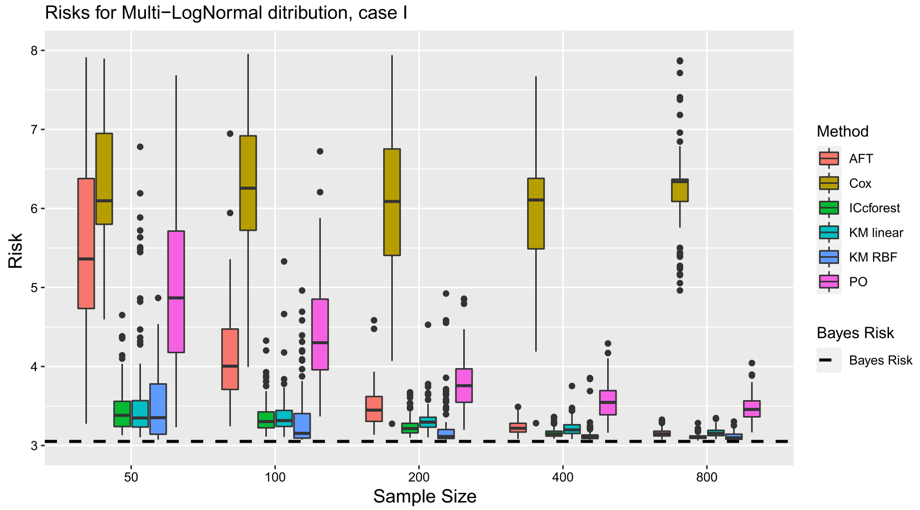


Fig. 3 Multi-Weibull failure time distribution. The Bayes risk is the dashed line and the boxplots of the following risks are compared: the KM-CSD with an RBF kernel, the KM-CSD with a linear kernel, AFT, Cox, ICcforest, and PO, for sample sizes $n = 50, 100, 200, 400, 800$

non-linear conditional expectation function, with some preference for the ICcforest method over the KM-CSD with an RBF kernel. This preference may be explained by the fact that an RBF kernel produces smooth functions and cannot capture non-differentiable points, such as the triangle's vertex. To test this hypothesis, we derived an additional simulation setting where the conditional expectation is quadratic and U-shaped.

In Setting 6, we considered a non-linear conditional expectation function that is U-shaped. The two-dimensional covariates Z are generated uniformly on $[0, 1]^2$, C is generated uniformly on $[0, \tau]$, and T is generated according to the following

$$T = 1 + 8(\beta^T Z - 1)^2 + \epsilon, \quad \text{where } \epsilon \sim N(0, 1) \text{ and } \beta = (1, 1)^T.$$



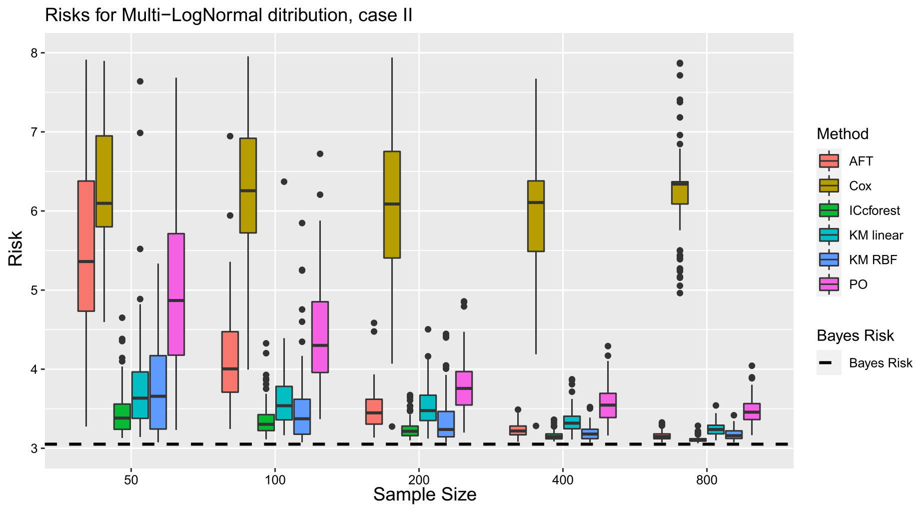
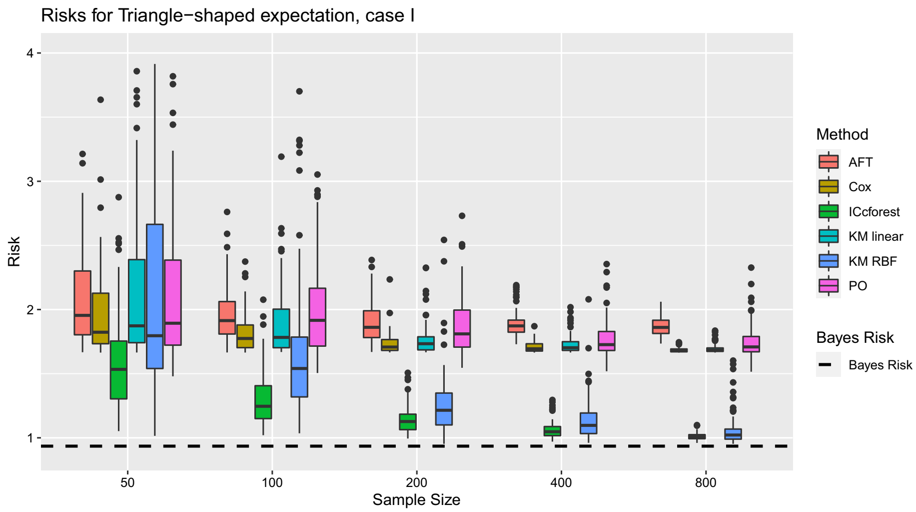


Fig. 4 Multi-LogNormal failure time distribution. The Bayes risk is the dashed line and the boxplots of the following risks are compared: the KM-CSD with an RBF kernel, the KM-CSD with a linear kernel, AFT, Cox, ICcforest, and PO, for sample sizes $n = 50, 100, 200, 400, 800$

The failure time was then truncated at $\tau = 9$. In Fig. 6, the boxplots of risks are presented. Figure 6 shows that the KM-CSD with an RBF kernel produces the lowest risks, for both cases g known and unknown, and is the only method that converges to the Bayes risk.

Finally, Table 1 presents a summary of training times (in milliseconds) and memory usage (in bytes) for all the methods considered in the experimental evaluations. The summary is based on a sample of size $n = 200$ drawn from the univariate Weibull simulation setting, and 100 iterations. For benchmarking, we used the 'mark' function from the R package 'bench' (Hester 2020). Table 1 shows that the memory allocated to the KM-CSD with either kernel is larger than that of AFT and PO, but smaller than that of Cox and ICcforest. Additionally, the KM-CSD with either kernel is slower only than



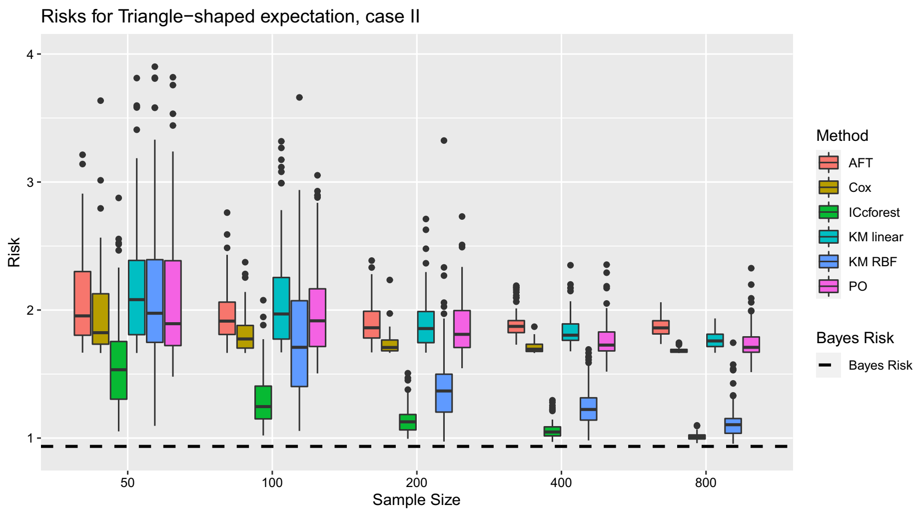


Fig. 5 Triangle shaped failure time expectation. The Bayes risk is the dashed line and the boxplots of the following risks are compared: the KM-CSD with an RBF kernel, the KM-CSD with a linear kernel, AFT, Cox, ICcforest, and PO, for sample sizes $n = 50, 100, 200, 400, 800$

the parametric AFT model, but is faster than all other semiparametric methods (PO and Cox), and is substantially faster than the nonparametric method (ICcforest).

To summarize, Figs. 1, 2, 3, 4, 5 and 6 showed that the KM-CSD is comparable to other known methods for estimating the failure time distribution with current status data, and in certain cases is even better. Specifically, we found that the KM-CSD with an appropriate kernel was superior in most settings, especially when the sample size n was large enough, and when true density g was known. It should be noted that even when the assumptions of the other (semi-) parametric models were true, the KM-CSD estimates were comparable. Additionally, when these assumptions fail to hold, the KM-CSD estimates were generally better. Furthermore, it seems that the KM-CSD can perform well in higher dimensions. Finally, when compared to the non-parametric ICcforest method, the performance of the

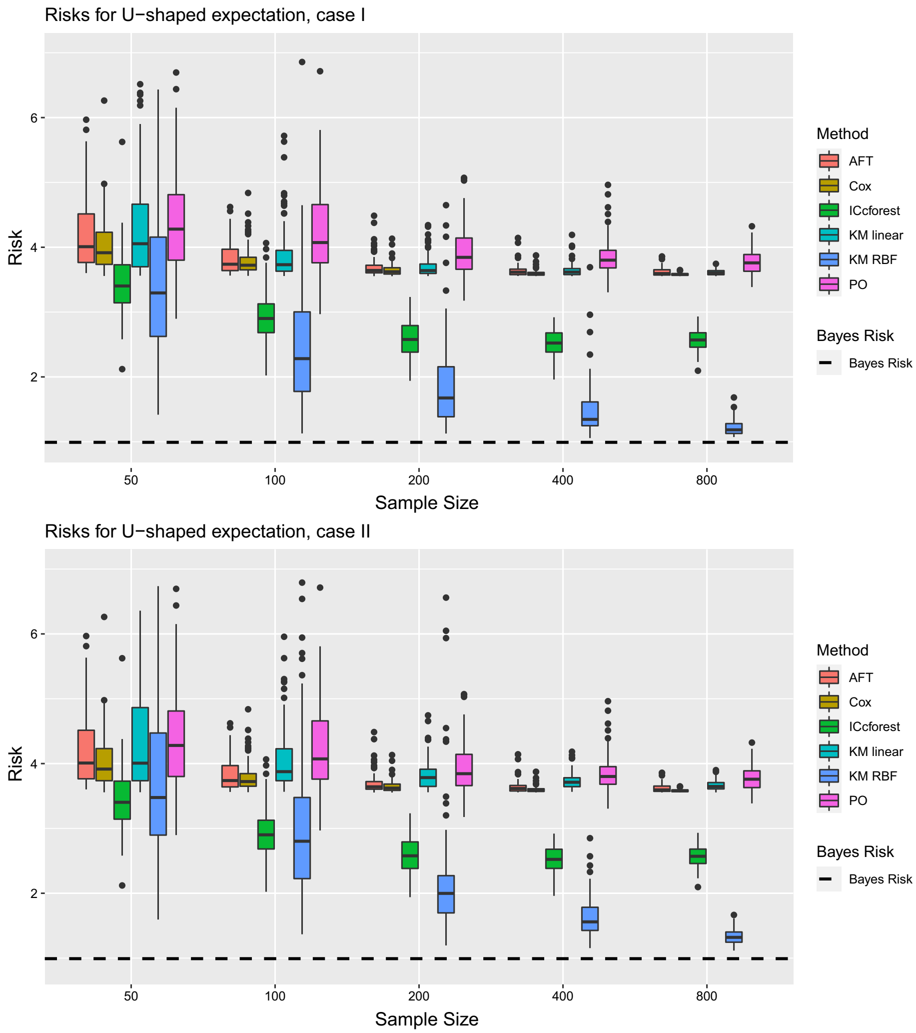


Fig. 6 U-shaped failure time expectation. The Bayes risk is the dashed line and the boxplots of the following risks are compared: the KM-CSD with an RBF kernel, the KM-CSD with a linear kernel, AFT, Cox, ICcforest, and PO, for sample sizes $n = 50, 100, 200, 400, 800$

Table 1 Comparison of training run times (in milliseconds) and memory usage (in bytes) for the different methods, for failure time data generated according to the univariate Weibull distribution, and for a sample of size $n = 200$

	n_itr	Min	Median	Total_time	Mem_alloc
AFT	100	4.06	5.19	523.30	642,456.00
KM linear	100	7.21	8.79	959.19	2,210,784.00
KM RBF	100	11.21	13.49	2189.14	3,216,008.00
PO	100	38.79	41.24	4200.90	415,048.00
Cox	100	108.59	151.48	19,975.32	45,004,712.00
ICcforest	100	5192.21	6830.46	702,177.32	122,276,448.00

KM-CSD was similar, and for larger sample sizes was even better. In summary, when the sample size is large enough ($n \gtrapprox 500$), the KM-CSD is comparable and/or better than existing methods, for all the different simulation settings, including both linear and non-linear conditional expectation functions. Additionally, Table 1 showed that the implementation of the KM-CSD is computationally cheap, relatively to the non-parametric ICcforest method, and the semi-parametric Cox model with splines used to estimate the baseline hazard. As nowadays training sets are usually quite large, the requirement on the size of the training set does not seem to be restrictive, and is complemented nicely by the relatively low computational cost.

6 Real world data analysis

In this section we test our approach on two real-world data sets, and compare its performance to current state of the art. The first data set is current status data from immunological studies, and the second is real world data concerning news popularity, with artificial censoring. Note that the second data set was artificially censored by us, allowing us to train our method on current status data, and to test it on the true uncensored data. We used the mean squared error (MSE) in order to determine the best fit.

6.1 Current status data from immunological studies

We present an analysis of real world serological data[1] on PVB19 and VZV infections. Both PVB19 and VZV cause a variety of diseases that mainly occur in childhood. The data was collected in Belgium between 2001 and 2003, as described in Hens et al. (2012). Blood samples were tested for presence of infection-specific IgG antibodies, reflecting infection experience. In addition, age at the time of data collection was registered. These blood samples are classified as either being seropositive or seronegative, based on some cut-off level, thus yielding current status data, with patient age being the monitoring time. The statistical analysis included in this paper is based on serological data on 2382 subjects with known immunological status for both PVB19 and VZV.

For our analysis, we use the patient's age at the time of data collection as the monitoring time (C). We consider the continuous IgG antibody level of B19 as a covariate (Z) explaining the presence of the current status indicator VZV (Δ). Note that we are treating the IgG antibody level of B19 as a baseline covariate, since we only have a single measurement of this antibody level. Also note that Hens et al. (2008) and Abrams and Hens (2015) have investigated the association between VZV and B19, and have shown that they share the same transmission route. Hence, there is a scientific justification for using the continuous IgG antibody level of B19 as a covariate explaining the presence of VZV.

We test our proposed KM-CSD on this data and compare it to estimates obtained by the Cox model, the AFT model, the PO model, and the survival forest ICcforest. For the kernel of the RKHS $\mathcal{H}$, we used both a linear kernel and a Gaussian RBF kernel, where the kernel width σ and the regularization parameter λ were chosen using fivefold cross-validation. It should be noted that we first standardized the covariates Z (PVB19 antibody level) in order to suggest a reasonable selection of kernel widths. As before, the density of

[1] This dataset can be found at https://www.dropbox.com/s/h120ml7pc68u63d/RCodeBook.zip?dl=0.

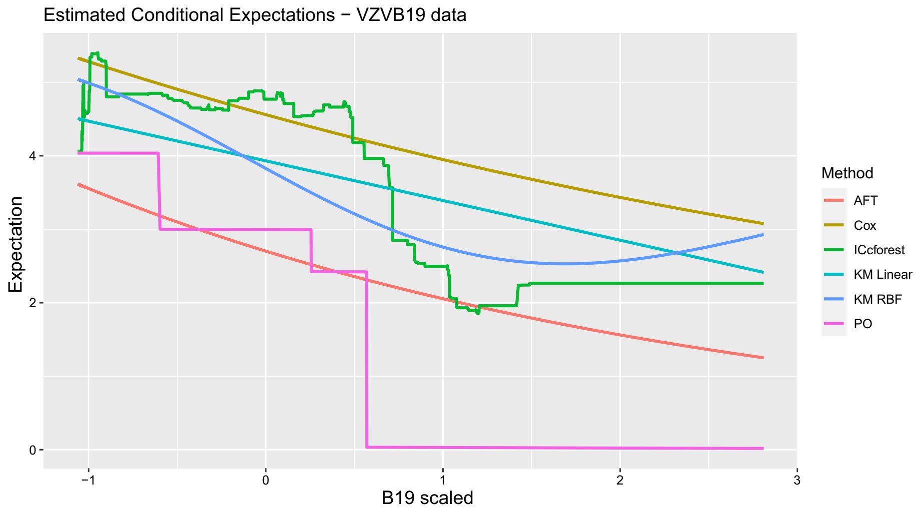


Fig. 7 Conditional expectation of time to infection of VZV as a function of the standardized antibody level of B19. The following estimates are compared: the KM-CSD with an RBF kernel, the KM-CSD with a linear kernel, AFT, Cox, ICcforest and PO

the censoring variable was estimated using nonparametric kernel density estimation with a Gaussian kernel. In Fig. 7, we present the results of the estimated expectation of time-to-infection of VZV, as a function of the covariates, for all six methods: KM with an RBF kernel, KM with a linear kernel, AFT, Cox, ICcforest and PO. It should be noted that since we do not know the true time-to-infection, we cannot argue that any model is superior. All six methods agree that there is a decreasing connection between time to infection of VZV, and B19 antibody level. In other words, the higher the level of PVB19, the lower the age of infection with VZV. This outcome supports previous research on joint transmission routes of VZV and B19. Further serological research can be done in order to better understand this relationship.

6.2 Artificially censored real-world data

For our second analysis, we used real-world data on news popularity,[2] with artificial censoring. The original data summarizes a set of features regarding articles published by Mashable, in a period of two years, as described in Fernandes et al. (2015). The goal is to predict the number of shares of an article in social networks, referred to as ‘popularity’. Since the number of shares is non-negative, we consider it as our failure-time T. The original dataset contains 58 predictive attributes. As before, we first standardized the covariates Z. In order to reduce the dimensionality of the data, we used the LASSO method for subset selection (Tibshirani 1996). For the sake of our analysis, we used the six most important explanatory variables. In order to obtain current status data, we generated the monitoring times $C_1, \dots, C_n$ as random exponential

[2] This dataset can be found at https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity.

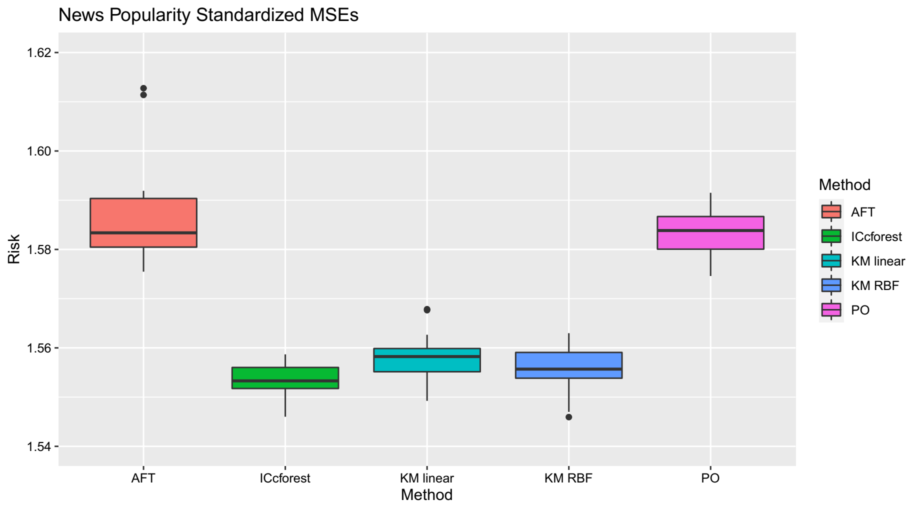


Fig. 8 SMSEs of predicted number of shares, based on 35 training sets. The following estimates are compared: the KM-CSD with an RBF kernel, the KM-CSD with a linear kernel, AFT, ICcforest, and PO

variables with mean equal to the mean number of shares. We then calculated the current status indicator by $\Delta = \mathbf{1}_{\{T<C\}}$. In summary, the artificially censored data consists of six covariates, the current status indicator, and the monitoring time generated from an exponential distribution. The uncensored data after standardization and dimensionality reduction, and its artificially censored version, can be found in the Supplementary Materials.

Since the original dataset contains 39,644 entries, we divided it randomly into 35 training sets of 1000 observations, and one testing set of 4644 observations. The training sets consisted of the artificially censored data, whereas the testing data contained the original uncensored scaled number of shares. We trained the KM-CSD, with both a linear and a Gaussian RBF kernel, as well as Cox, AFT, PO, and ICcforest, on each training set. As before, the kernel width σ and the regularization parameter λ were chosen using fivefold cross-validation. For a fair comparison, we estimated the density of the censoring variable using nonparametric kernel density estimation with a Gaussian kernel, and did not use our knowledge regarding the censoring mechanism.

For each training set, we computed the model predictions on the testing set and calculated the corresponding MSE. Since the MSE is sensitive to the overall scale of the response variable, we divided the MSE by the empirical variance of the number of shares in order to achieve standardized MSE. We did not include the Cox model in the graphical representation, since the Cox model produced very high risks. Figure 8 presents the boxplot of the standardized MSEs (SMSEs), for all remaining five methods: KM-CSD with an RBF kernel, KM-CSD with a linear kernel, AFT, ICcforest, and PO. It should also be noted that for some training sets, the AFT SMSEs were so high that we had to omit them from the graphical representation. Figure 8 shows that the nonparametric methods are comparable, and produce risk values that are much lower than those of the AFT and the PO model.

7 Concluding remarks

We proposed a kernel-machine approach for estimation of the failure time expectation, studied its theoretical properties, presented a simulation study, and tested our approach on two real-world data sets. Specifically, we proved that our method is consistent, and showed by simulations and analysis of real-world data that our approach is just as good as current state of the art, and sometimes even better. We believe this work demonstrates an important approach in applying machine learning techniques to current status data. However, many open questions remain and many possible generalizations exist. First, note that we only studied the problem of estimating the failure time expectation and not other distribution related quantities. Further work needs to be done in order to extend the kernel machines approach to other estimation problems with current status data, and is beyond the scope of this paper. Note that the theory developed here might not hold in such generalizations, as the corresponding modified loss function will no longer be a convex function. Second, we assumed that the censoring is independent of the failure time given the covariates and that the censoring density is positive given the covariates over the entire observed time range. It would be worthwhile to study the consequences of violation of some of these assumptions. Third, it could be interesting to extend this work to other censored data formats such as interval censoring. We believe that further development and generalization of kernel machine learning methods for different types of censored data is of great interest. Some additional generalization of this work can include derivation of doubly-robust estimators and inclusion of time-dependent covariates. For the case of time-dependent covariates, one first needs to define an RKHS over the covariate process space and then to define the appropriate empirical risk minimization. Since this space is rich, the covering number results discussed in Sect. 4 may not hold for this space.

8 Supplementary materials

The appendices referenced in Sects. 3–5, the R code for the algorithm and simulations, and the artificially censored data used in Sect. 6.2, are available with this article.

8.1 R code and artificially censored data set

An R package 'KMforCSD' containing the R code for both the algorithm and for the simulations is available at https://github.com/Yael-Travis-Lumer/KMforCSD. The 'KMforCSD' package also contains the data used in Sect. 6.2, including the uncensored data after standardization and dimensionality reduction, and its artificially censored version. An additional R package 'mlr3learners.KMforCSD', that wraps the 'KMforCSD' package as an MLR3 learner, is available at https://github.com/Yael-Travis-Lumer/mlr3learners.KMforCSD. The 'mlr3learners.KMforCSD' package enables cross validation.

Acknowledgements The authors would like to thank the editor and the reviewers for their valuable comments and suggestions. This work was supported in part by the National Science Foundation [Grant Number DMS-1407732], and by the Israel Science Foundation [Grant Number 849/17]. The authors would also like to thank Niel Hens for sharing the serological dataset on VZV and PVB19, and for helpful discussions.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A

Computation of the decision function

Equation (4) is a quadratic optimization problem. Such problems are vastly studied in the literature (see, for example, Suykens and Vandewalle 1999) and known solutions exist. Specifically, using the representer theorem (Steinwart and Christmann 2008, Theorem 5.5), the solution of Eq. (4) is given by

$$f_{D,\lambda}(Z) = \sum_{i=1}^{n} \alpha_i k(Z, Z_i) + b.$$

Using the Lagrange method, the quadratic optimization problem in (4) can be simplified to a set of linear equations (see, for example, Fletcher 1987). Hence, it can be shown that the coefficients $\alpha_1, \dots, \alpha_n$ and b in the representation of $f_{D,\lambda}$ above can be obtained by

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_n \\ b \end{pmatrix} = \begin{pmatrix} K_{11} + n\lambda & K_{12} & \dots & K_{1n} & 1 \\ K_{21} & K_{22} + n\lambda & \dots & K_{2n} & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ K_{n1} & K_{n2} & \dots & K_{nn} + n\lambda & 1 \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \\ 0 \end{pmatrix},$$

where $\mathbf{K}_{n\times n}$ is the kernel matrix with entries $\mathbf{K}_{ij} = k(Z_i, Z_j)$, and where $v_i = (1 - \Delta_i)/g(C_i|Z_i)$, for $1 \le i,j \le n$. That is, the KM-CSD decision function has a closed form.

Appendix B

Non-negative new modified loss function

Recall that our proposed loss function is

$$L^n(D, (Z, C, \Delta, s)) = \frac{(1 - \Delta)2(C - s)}{g(C|Z)} + s^2 .$$

Note that this function is convex but not necessarily a loss function since it can obtain negative values. By completing the square,

$$
\begin{aligned}
L^n(D,(Z,C,\Delta,s)) &= \frac{(1-\Delta)2(C-s)}{g(C|Z)} + s^2 \\
&= \left(s - \frac{1-\Delta}{g(C|Z)}\right)^2 + \frac{2(1-\Delta)C}{g(C|Z)} - \left(\frac{1-\Delta}{g(C|Z)}\right)^2
\end{aligned}
$$

we observe that the only possibly negative component is $-\left(\frac{1-\Delta}{g(C|Z)}\right)^2 = -\frac{1-\Delta}{g(C|Z)^2}$. In order to ensure positivity we add a constant term that does not depend on f, and so our loss becomes

$$
\widetilde{L^n}(D,(Z,C,\Delta,f(Z))) = \frac{(1-\Delta)2(C-f(Z))}{g(C|Z)} + (f(Z))^2 + a,
$$

where for a fixed dataset of length n, the constant a is $a = \max_{1\le i\le n}\left\{(1-\Delta_i)/\left(g(C_i|Z_i)\right)^2\right\}$. Note that this additional term will not effect the optimization (since $\widetilde{L^n}$ is just a shift by a constant of L^n) and thus will be neglected hereafter.

Appendix C

Proof of Theorem 1

Proof Since $L^n(D,(Z,C,\Delta,s)) = \tau^{-2}\left((1-\Delta)2(C-s)/g(C|Z) + s^2\right)$ is convex, it implies that there exists a unique decision function (see Steinwart and Christmann 2008, Section 5.1). For all distributions Q on $\mathcal{Z}\times\mathcal{Y}$, we define the kernel machine decision function by $f_{Q,\lambda} = \inf_{f\in\mathcal{H}} \lambda\|f\|^2_{\mathcal{H}} + R_{L,Q}(f)$. We note that for an RKHS $\mathcal{H}$ of a continuous kernel k with $\|k\|_\infty \le 1$,

$$
\left\|f_{Q,\lambda}\right\|_\infty \le \|k\|_\infty \left\|f_{Q,\lambda}\right\|_{\mathcal{H}} \le \left\|f_{Q,\lambda}\right\|_{\mathcal{H}}.
$$

Hence,

$$
\begin{aligned}
\lambda\left\|f_{Q,\lambda}\right\|^2_{\mathcal{H}} &\le \lambda\left\|f_{Q,\lambda}\right\|^2_{\mathcal{H}} + R_{L,Q}(f_{Q,\lambda}) = \inf_{f\in\mathcal{H}} \lambda\|f\|^2_{\mathcal{H}} + R_{L,Q}(f) \\
&\le \lambda\|0\|^2_{\mathcal{H}} + R_{L,Q}(0) = R_{L,Q}(0),
\end{aligned}
$$

Hence $\left\|f_{Q,\lambda}\right\|_\infty \le \left\|f_{Q,\lambda}\right\|_{\mathcal{H}} \le \lambda^{-1/2}\sqrt{R_{L,Q}(0)}$ for all $f\in\mathcal{H}$. By Remark 1, $L(y,0)\le 1$ for all $y\in\mathcal{Y}$ and so we conclude that $R_{L,Q}(0)\le 1$ and thus $\left\|f_{Q,\lambda}\right\|_\infty \le \left\|f_{Q,\lambda}\right\|_{\mathcal{H}} \le \lambda^{-1/2}$ for all distributions Q on $\mathcal{Z}\times\mathcal{Y}$.

Recall that the unit ball of $\mathcal{H}$ is denoted by B_H and its closure by $\overline{B_H}$; since $\left\|f_{P,\lambda}\right\|_{\mathcal{H}} \le \lambda^{-1/2}$ we can write $f\in\lambda^{-1/2}\overline{B_H}$. Since $\mathcal{Z}\subset\mathbb{R}^d$ is compact, it implies that the $\|\cdot\|_\infty$ − closure $\overline{B_H}$ of the unit ball B_H is compact in $\ell_\infty(\mathcal{Z})$ (see Steinwart and Christmann 2008, Corollary 4.31).

Denote by $R_{L^n,D}(f)$ the empirical risk with respect to the data-dependent loss L^n. Since $f_{D,\lambda}$ minimizes $\lambda\|f\|^2_{\mathcal{H}} + R_{L^n,D}(f)$,

$$
\lambda\left\|f_{D,\lambda}\right\|^2_{\mathcal{H}} + R_{L^n,D}(f_{D,\lambda}) \le \lambda\left\|f_{P,\lambda}\right\|^2_{\mathcal{H}} + R_{L^n,D}(f_{P,\lambda}).
$$

Recall that the approximation error is defined by $A_2(\lambda) = \inf_{f\in\mathcal{H}} \lambda\|f\|^2_{\mathcal{H}} + R_{L,P}(f) - R^*_{L,P}$, and thus, as in Steinwart and Christmann (2008, Eq. 6.18),

$$
\begin{aligned}
&\lambda\|f_{D,\lambda}\|^2_{\mathcal{H}} + R_{L,P}(f_{D,\lambda}) - R^*_{L,P} - A_2(\lambda) \\
&\quad= \lambda\|f_{D,\lambda}\|^2_{\mathcal{H}} + R_{L,P}(f_{D,\lambda}) - \lambda\|f_{P,\lambda}\|^2_{\mathcal{H}} - R_{L,P}(f_{P,\lambda}) \\
&\quad= \lambda\|f_{D,\lambda}\|^2_{\mathcal{H}} + R_{L^n,D}(f_{D,\lambda}) - R_{L^n,D}(f_{D,\lambda}) + R_{L,P}(f_{D,\lambda}) - \lambda\|f_{P,\lambda}\|^2_{\mathcal{H}} - R_{L,P}(f_{P,\lambda}) \\
&\quad\le \lambda\|f_{P,\lambda}\|^2_{\mathcal{H}} + R_{L^n,D}(f_{P,\lambda}) - R_{L^n,D}(f_{D,\lambda}) + R_{L,P}(f_{D,\lambda}) - \lambda\|f_{P,\lambda}\|^2_{\mathcal{H}} - R_{L,P}(f_{P,\lambda}) \\
&\quad= R_{L^n,D}(f_{P,\lambda}) - R_{L^n,D}(f_{D,\lambda}) + R_{L,P}(f_{D,\lambda}) - R_{L,P}(f_{P,\lambda}) \\
&\quad\le 2 \sup_{\|f\|_{\mathcal{H}}\le\lambda^{-1/2}} |R_{L,P}(f) - R_{L^n,D}(f)|.
\end{aligned}
$$

That is,

$$
\lambda\|f_{D,\lambda}\|^2_{\mathcal{H}} + R_{L,P}(f_{D,\lambda}) - R^*_{L,P} - A_2(\lambda) \le 2 \sup_{\|f\|_{\mathcal{H}}\le\lambda^{-1/2}} |R_{L,P}(f) - R_{L^n,D}(f)| \tag{6}
$$

Note that since L is Lipschitz continuous, $|L(y,s) - L(y,s')| \le c_L|s - s'|$ for all $s, s' \in \underline{S}.$

From the discussion above, we are only interested in bounded functions $f \in \lambda^{-1/2}\overline{B_H}$.

Then for all $f \in \lambda^{-1/2}\overline{B_H}$ we have

$$
|L(y,f(z))| \le |L(y,f(z)) - L(y,0)| + L(y,0) \le c_L|f(z)| + 1 \le c_L\lambda^{-1/2} + 1 \equiv B_2
$$

thus we obtain that for functions $f \in \lambda^{-1/2}\overline{B_H}$, the loss $L(y, f(z))$ is bounded.

For any $\epsilon > 0$, let $\mathcal{F}_\epsilon$ be an $\varepsilon - net$ of $\lambda^{-1/2}\overline{B_H}$. Since $\overline{B_H}$ is compact, then the cardinality of the $\varepsilon - net$ is

$$
|\mathcal{F}\varepsilon| = N\left(\lambda^{-1/2}B_H, \|\cdot\|_\infty, \epsilon\right) = N(B_H, \|\cdot\|_\infty, \sqrt{\lambda}\epsilon) < \infty.
$$

Thus for every $f \in \lambda^{-1/2}\overline{B_H}$, there exists a function $h \in \mathcal{F}_\varepsilon$ with $\|f - h\| \le \varepsilon$, and thus

$$
\begin{aligned}
&|R_{L,P}(f) - R_{L^n,D}(f)| \\
&\quad\le |R_{L,P}(f) - R_{L,P}(h)| + |R_{L,P}(h) - R_{L^n,D}(h)| + |R_{L^n,D}(h) - R_{L^n,D}(f)| \\
&\quad\equiv A_n + B_n + C_n
\end{aligned} \tag{7}
$$

First we will bound C_n;

$$
\begin{aligned}
C_n &\equiv \left|R_{L^n,D}(h) - R_{L^n,D}(f)\right| \\
&\le \left|\frac{1}{n}\sum_{i=1}^n\left[\frac{(1-\Delta_i)\ell(C_i, h(Z_i))}{g(C_i|Z_i)}\right] - \frac{1}{n}\sum_{i=1}^n\left[\frac{(1-\Delta_i)\ell(C_i, f(Z_i))}{g(C_i|Z_i)}\right]\right| \\
&\quad+ \left|\frac{1}{n}\sum_{i=1}^n[L(0, h(Z_i))] - \frac{1}{n}\sum_{i=1}^n[L(0, f(Z_i))]\right| \\
&\equiv C_{n,1} + C_{n,2},
\end{aligned}
$$

where

$$
\begin{aligned}
C_{n,1} &\equiv \left| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{(1-\Delta_i)\ell(C_i, h(Z_i))}{g(C_i|Z_i)} - \frac{(1-\Delta_i)\ell(C_i, f(Z_i))}{g(C_i|Z_i)} \right] \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{(1-\Delta_i)}{g(C_i|Z_i)} \big(\ell(C_i, h(Z_i)) - \ell(C_i, f(Z_i)) \big) \right] \right| \\
&\leq \left| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{g(C_i|Z_i)} \big(\ell(C_i, h(Z_i)) - \ell(C_i, f(Z_i)) \big) \right] \right| \\
&\leq \frac{1}{2\kappa} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\ell(C_i, h(Z_i)) - \ell(C_i, f(Z_i)) \right] \right| \leq \frac{1}{2n\kappa} \sum_{i=1}^{n} |\ell(C_i, h(Z_i)) - \ell(C_i, f(Z_i))| \\
&\leq \frac{1}{2n\kappa} \sum_{i=1}^{n} c_l |h(Z_i) - f(Z_i)| \leq \frac{1}{2n\kappa} \sum_{i=1}^{n} c_l \varepsilon = \frac{c_l \varepsilon}{2\kappa},
\end{aligned}
$$

and where

$$
\begin{aligned}
C_{n,2} &\equiv \left| \frac{1}{n} \sum_{i=1}^{n} [L(0, h(Z_i)) - L(0, f(Z_i))] \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \big| L(0, h(Z_i)) - L(0, f(Z_i)) \big| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} c_L |h(Z_i) - f(Z_i)| \leq \frac{1}{n} \sum_{i=1}^{n} \left[c_L \varepsilon \right] = c_L \varepsilon
\end{aligned}
$$

So we were able to bound C_n by $c_l\varepsilon/2\kappa + c_L\varepsilon$.

Similarly, using to the property that $E[\alpha] = \alpha$ for any constant α, it can be shown that $A_n \leq c_l\varepsilon/2\kappa + c_L\varepsilon$.

As an interim summary, we showed that

$$
\sup_{f \in \lambda^{-1/2} B_H} |R_{L,P}(f) - R_{L^n,D}(f)| \leq \underbrace{\sup_{h \in \mathcal{F}_\varepsilon} |R_{L,P}(h) - R_{L^n,D}(h)|}_{=B_n} + \frac{1}{\kappa} c_l \varepsilon + 2c_L \varepsilon. \tag{8}
$$

Recall that the loss $L(y, f(z))$ is bounded by B_2 and that by Remark 1, $\ell(y, s) \leq B_1$.

We note that

$$
\frac{(1-\Delta)\ell(C, h(Z))}{g(C|Z)} + L(0, h(Z)) \leq \frac{\ell(C, h(Z))}{g(C|Z)} + L(0, h(Z)) \leq \frac{B_1}{2\kappa} + B_2 \equiv B
$$

Combining this with Eq. (6), we obtain that

$$
\begin{aligned}
&Pr\left(\lambda\|f_{D,\lambda}\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* - A_2(\lambda) \geq B\sqrt{\frac{2\eta}{n}} + \frac{2c_l\varepsilon}{\kappa} + 4c_L\varepsilon \right) \\
&\quad \leq Pr\left(2 \sup_{\|f\|_{\mathcal{H}} \leq \lambda^{-1/2}} |R_{L,P}(f) - R_{L^n,D}(f)| \geq B\sqrt{\frac{2\eta}{n}} + \frac{2c_l\varepsilon}{\kappa} + 4c_L\varepsilon \right) \qquad \text{(by eq. 6)} \\
&\quad \leq Pr\left(2\left(\sup_{h\in\mathcal{F}_\varepsilon} |R_{L,P}(h) - R_{L^n,D}(h)| + \frac{1}{\kappa}c_l\varepsilon + 2c_L\varepsilon \right) \geq B\sqrt{\frac{2\eta}{n}} + \frac{2c_l\varepsilon}{\kappa} + 4c_L\varepsilon \right) \qquad \text{(by eq. 8)} \\
&\quad = Pr\left(2\left(\sup_{h\in\mathcal{F}_\varepsilon} B_n + \frac{1}{\kappa}c_l\varepsilon + 2c_L\varepsilon \right) \geq B\sqrt{\frac{2\eta}{n}} + \frac{2c_l\varepsilon}{\kappa} + 4c_L\varepsilon \right) \\
&\quad = Pr\left(\sup_{h\in\mathcal{F}_\varepsilon} B_n \geq B\sqrt{\frac{\eta}{2n}} \right) = Pr\left(\sup_{h\in\mathcal{F}_\varepsilon} |R_{L,P}(h) - R_{L^n,D}(h)| \geq B\sqrt{\frac{\eta}{2n}} \right).
\end{aligned}
$$

By the union bound, the last expression is bounded by

$$
\sum_{h\in\mathcal{F}_\varepsilon} Pr\left(|R_{L,P}(h) - R_{L^n,D}(h)| \geq B\sqrt{\frac{\eta}{2n}} \right),
$$

which can then be bounded again by $2|\mathcal{F}_\varepsilon|\exp(-\eta)$, using Hoeffding's inequality (Steinwart and Christmann 2008, Theorem 6.10); where $\mathcal{F}_\epsilon$ is an ε-net of $\lambda^{-1/2}B_H$ with cardinality

$$
|\mathcal{F}\varepsilon| = N\left(\lambda^{-1/2}B_H, \|\cdot\|_\infty, \epsilon\right) < \infty.
$$

Define $\eta = \log(2|\mathcal{F}_\varepsilon|) + \theta$, then

$$
\begin{aligned}
&Pr\left(\lambda\|f_{D,\lambda}\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* - A_2(\lambda) \geq B\sqrt{\frac{2(\log(2|\mathcal{F}_\varepsilon|) + \theta)}{n}} + \frac{2c_l\varepsilon}{\kappa} + 4c_L\varepsilon \right) \\
&\quad \leq \exp(-\theta),
\end{aligned}
$$

which concludes the proof. □

Proof of Corollary 1

Proof In Theorem 1 we showed that

$$
\begin{aligned}
&\lambda\|f_{D,\lambda}\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* - A_2(\lambda) \\
&\quad \leq B\sqrt{\frac{2\log\left(2N(\sqrt{\frac{1}{\lambda}}B_H, \|\cdot\|_\infty, \epsilon)\right) + 2\theta}{n}} + \frac{2c_l\varepsilon}{\kappa} + 4c_L\varepsilon,
\end{aligned}
$$

with probability not less than $1 - e^{-\theta}$.

Choose $\lambda = \lambda_n$; from Assumption (A5) together with Lemma 5.15 of Steinwart and Christmann (Steinwart and Christmann 2008, 5.15), $A_2(\lambda_n)$ converges to zero as n goes to infinity. By the assumption $\log\left(N(B_H, \|\cdot\|_\infty, \epsilon)\right) \leq a\epsilon^{-2p}$, we have that

$$
\begin{aligned}
&\log\left(2N(B_H, \|\cdot\|_\infty, \sqrt{\lambda}\epsilon)\right)\\
&\quad = \log(2) + \log\left(N(B_H, \|\cdot\|_\infty, \sqrt{\lambda}\epsilon)\right)\\
&\quad \le \log(2) + a\left(\sqrt{\lambda}\epsilon\right)^{-2p}\\
&\quad \le 1 + a\left(\sqrt{\lambda}\epsilon\right)^{-2p}.
\end{aligned}
$$

Choose $\epsilon = \left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\frac{1}{\sqrt{\lambda}}$ and recall that $a \ge 1$. Then for $n \ge \frac{p^2 a}{2}$ we have

$$
\begin{aligned}
&\log\left(2N(B_H, \|\cdot\|_\infty, \sqrt{\lambda}\epsilon)\right)\\
&\quad \le 1 + a\left(\sqrt{\lambda}\epsilon\right)^{-2p}\\
&\quad = 1 + a\left(\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right)^{-2p}\\
&\quad \le 2a\left(\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right)^{-2p}
\end{aligned}
$$

.

Recall that B is defined by $B = \frac{B_1}{2\kappa} + c_L\lambda^{-\frac{1}{2}} + 1$. Hence, from the assumption on the covering number we have that

$$
\begin{aligned}
&B\sqrt{\frac{2\log\left(2N(\sqrt{\frac{1}{\lambda}}B_H, \|\cdot\|_\infty, \epsilon)\right) + 2\theta}{n}}\\
&\quad \le \left(\frac{B_1}{2\kappa} + c_L\lambda^{-\frac{1}{2}} + 1\right)\sqrt{\frac{4a\left(\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right)^{-2p} + 2\theta}{n}}
\end{aligned}
$$

and since $\lambda_n^{1+p}n \underset{n\to\infty}{\to} \infty$, the right hand side of this converges to 0 as $n \to \infty$. Finally, from the choice of ϵ, it follows that both $\frac{2c_l\varepsilon}{\kappa}$ and $4c_L\varepsilon$ converge to 0 as $n \to \infty$. Hence for every fixed θ,

$$
\begin{aligned}
&\lambda_n\left\|f_{D,\lambda_n}\right\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda_n}) - R^*_{L,P}\\
&\quad \le A_2(\lambda_n) + B\sqrt{\frac{2\log\left(2N(\sqrt{\frac{1}{\lambda_n}}B_H, \|\cdot\|_\infty, \epsilon)\right) + 2\theta}{n}} + \frac{2c_l\varepsilon}{\kappa} + 4c_L\varepsilon
\end{aligned}
$$

with probability not less than 1-$e^{-\theta}$. The right hand side of this converges to 0 as $n \to \infty$, which implies consistency (Steinwart and Christmann 2008, Lemma 6.5). Since this holds for all probability measures $P \in \mathcal{P}$, we obtain $\mathcal{P}$-universal consistency. □

Proof of Lemma 1

Proof For the sake of completeness, we develop here a finite sample bound on the difference between the kernel density estimator $\hat{g}$ and the true density g. While asymptotic results for kernel density estimators are well known in the literature (see, for example, Silverman 1978), finite sample bounds were not previously studied. In order to develop our bound, we incorporate Bernstein's inequality in our analysis as described below.

Note that

$$\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}(C_i)-g(C_i)\right|$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}(C_i)-E\left[\hat{g}(C_i)\right]\right| + \frac{1}{n}\sum_{i=1}^{n}\left|E\left[\hat{g}(C_i)\right]-g(C_i)\right| \equiv A+B$$

As in Tsybakov (Tsybakov 2008, Proposition 1.1), for any $c_0 \in \mathcal{Y}$, define

$$\eta_i(c_0) = K_m\left(\frac{C_i-c_0}{h}\right) - E_g\left[K_m\left(\frac{C_i-c_0}{h}\right)\right].$$

Then $\eta_i(c_0)$, for $i = 1, \dots, n$ are i.i.d. random variables with zero mean and with variance:

$$\begin{aligned}\mathrm{Var}\left[\eta_i(c_0)\right] &= E_g\left[\left(\eta_i(c_0)\right)^2\right] = E_g\left[\left(K_m\left(\frac{C_i-c_0}{h}\right) - E_g\left[K_m\left(\frac{C_i-c_0}{h}\right)\right]\right)^2\right]\\ &\leq E_g\left[K_m^2\left(\frac{C_i-c_0}{h}\right)\right] = \int_u K_m^2\left(\frac{u-c_0}{h}\right)g(u)du \leq g_{max}\int_u K_m^2\left(\frac{u-c_0}{h}\right)du\\ &= g_{max}h\int_v K_m^2(v)dv = D_1 h\end{aligned}$$

where the equality before last follows from change of variables and where $D_1 = g_{max}\int_v K_m^2(v)dv$. Thus $\mathrm{Var}(\hat{g}(c_0)) = E_g\left[\left(\frac{1}{nh}\sum_{i=1}^{n}\eta_i(c_0)\right)^2\right] = \frac{1}{nh^2}E_g\left[\eta_1^2(c_0)\right] \leq \frac{D_1 h}{nh^2} = \frac{D_1}{nh}$.

Note that $\mathrm{Var}(\left|\hat{g}(c_0)-E\left[\hat{g}(c_0)\right]\right|) = E\left[\left(\left|\hat{g}(c_0)-E\left[\hat{g}(c_0)\right]\right|\right)^2\right] = \mathrm{Var}(\hat{g}(c_0)-E\left[\hat{g}(c_0)\right]) = \mathrm{Var}(\hat{g}(c_0))$. Hence $\mathrm{Var}(\left|\hat{g}(c_0)-E\left[\hat{g}(c_0)\right]\right|) = \mathrm{Var}(\hat{g}(c_0)) \leq \frac{D_1}{nh}$. Using Bernstein's inequality, for any $\theta > 0$ we have

$$\begin{aligned}Pr\left(A > \sqrt{\frac{2D_1\theta}{n^2h}} + \frac{2g_{max}\theta}{3n}\right) &\equiv Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}(C_i)-E\left[\hat{g}(C_i)\right]\right| > \sqrt{\frac{2D_1\theta}{n^2h}} + \frac{2g_{max}\theta}{3n}\right)\\ &\leq \exp(-\theta)\end{aligned}$$

For the second term, as in Tsybakov (Tsybakov 2008, Proposition 1.2), we have that

$$B \equiv \frac{1}{n}\sum_{i=1}^{n}\left|E\left[\hat{g}(C_i)\right]-g(C_i)\right| \leq D_2 h^{\beta}$$

where $D_2 = \mathcal{L}|\pi|^{\beta-m}/m!\int_{-\infty}^{\infty}\left|K_m(v)\right||v|^{\beta}dv < \infty$, and for some $\pi \in [0,1]$.

In conclusion, we showed that

$$
\begin{aligned}
&Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}(C_i)-g(C_i)\right| > \sqrt{\frac{2D_1\theta}{n^2h}}+\frac{2g_{max}\theta}{3n}+D_2\cdot h^{\beta}\right)\\
&\leq Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}(C_i)-E\left[\hat{g}(C_i)\right]\right|+\frac{1}{n}\sum_{i=1}^{n}\left|E\left[\hat{g}(C_i)\right]-g(C_i)\right|\right.\\
&\left.> \sqrt{\frac{2D_1\theta}{n^2h}}+\frac{2g_{max}\theta}{3n}+D_2\cdot h^{\beta}\right)\\
&\leq Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}(C_i)-E\left[\hat{g}(C_i)\right]\right|+D_2\cdot h^{\beta}\right.\\
&\left.> \sqrt{\frac{2D_1\theta}{n^2h}}+\frac{2g_{max}\theta}{3n}+D_2\cdot h^{\beta}\right)\\
&= Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}(C_i)-E\left[\hat{g}(C_i)\right]\right| > \sqrt{\frac{2D_1\theta}{n^2h}}+\frac{2g_{max}\theta}{3n}\right)\leq \exp(-\theta)
\end{aligned}
$$

where h is the bandwidth. □

Proof of Theorem 2

Proof Note that the proof of this theorem is similar to the proof of of Theorem 1 and thus we will only discuss the parts of the proof where they differ. As in Theorem 1, Eq. 7,

$$
\begin{aligned}
&\lambda\|f_{D,\lambda}\|_{\mathcal{H}}^2+R_{L,P}(f_{D,\lambda})-R_{L,P}^{*}-A_2(\lambda)\\
&\leq 2\left(A_n+B_n+C_n\right)
\end{aligned}
$$

where

$$A_n\equiv|R_{L,P}(f)-R_{L,P}(v)|,\ B_n\equiv|R_{L,P}(v)-R_{L^n,D}(v)|,\ \text{and where}\ C_n\equiv|R_{L^n,D}(v)-R_{L^n,D}(f)|,$$

Since A_n does not depend on the data-set D, the same bound holds as in the proof of Theorem 1, that is, $A_n\leq c_l\epsilon/2\kappa+c_L\epsilon$.

We bound C_n as follows:

$$
\begin{aligned}
C_n&\equiv\left|R_{L^n,D}(v)-R_{L^n,D}(f)\right|\\
&\leq\left|\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-\Delta_i)\ell(C_i,v(Z_i))}{\hat{g}(C_i)}\right]-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-\Delta_i)\ell(C_i,f(Z_i))}{\hat{g}(C_i)}\right]\right|\\
&\quad+\left|\frac{1}{n}\sum_{i=1}^{n}[L(0,v(Z_i))]-\frac{1}{n}\sum_{i=1}^{n}[L(0,f(Z_i))]\right|\\
&\equiv C_{n,1}+C_{n,2}
\end{aligned}
$$

Using the same arguments as in Theorem 1, we can bound C_n by $c_l\epsilon/\kappa+c_L\epsilon$. Note that the only difference is in the denominator of $C_{n,1}$ since $g^{-1}\leq(2\kappa)^{-1}$ and $\hat{g}^{-1}\leq\kappa^{-1}$.

Recall that the loss $L(y, f(z))$ is bounded by B_2. Define $R_{L^n,D,g}(v)$ by

$$R_{L^n,D,g}(v) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-\Delta_i)\ell(C_i, v(Z_i))}{g(C_i)}\right] + \frac{1}{n}\sum_{i=1}^{n}[L(0, v(Z_i))].$$

In other words, $R_{L^n,D,g}(v)$ is the empirical risk with the true censoring density function g.

We bound B_n as follows

$$\begin{aligned} B_n &= |R_{L,P}(v) - R_{L^n,D}(v)| \\ &\leq \left|R_{L,P}(v) - R_{L^n,D,g}(v)\right| + \left|R_{L^n,D,g}(v) - R_{L^n,D}(v)\right| \equiv B_{n,1} + B_{n,2} \end{aligned}$$

where

$$\frac{(1-\Delta)\ell(C, v(Z))}{g(C)} + L(0, v(Z)) \leq \frac{\ell(C, v(Z))}{g(C)} + L(0, v(Z)) \leq \frac{B_1}{2\kappa} + B_2 = B$$

and where

$$\begin{aligned} B_{n,2} &= \left|R_{L^n,D,g}(v) - R_{L^n,D}(v)\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-\Delta_i)\ell(C_i, v(Z_i))}{g(C_i)}\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-\Delta_i)\ell(C_i, v(Z_i))}{\hat{g}(C_i)}\right]\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^{n}\left[(1-\Delta_i)\ell(C_i, v(Z_i))\left(\frac{1}{g(C_i)} - \frac{1}{\hat{g}(C_i)}\right)\right]\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left[\left|\ell(C_i, v(Z_i))\left(\frac{1}{g(C_i)} - \frac{1}{\hat{g}(C_i)}\right)\right|\right] \\ &= \frac{B_1}{n}\sum_{i=1}^{n}\left[\left|\frac{\hat{g}(C_i) - g(C_i)}{g(C_i)\hat{g}(C_i)}\right|\right] \leq \frac{B_1}{2\kappa^2 n}\sum_{i=1}^{n}\left[\left|\hat{g}(C_i) - g(C_i)\right|\right]. \end{aligned}$$

Note that these inequalities hold for all functions $v \in \mathcal{F}_\varepsilon \subseteq \lambda^{-1/2}B_H$. We would like to bound the last expression using Lemma 1. Let

$$\eta = \frac{B_1}{2\kappa^2}\left(\sqrt{\frac{2D_1\theta}{n^2 h}} + \frac{2g_{max}\theta}{3n} + D_2 \cdot h^\beta\right),$$

then by Lemma 1

$$\begin{aligned} Pr(B_{n,2} > \eta) &\leq Pr\left(\frac{B_1}{2\kappa^2 n}\sum_{i=1}^{n}\left[\left|\hat{g}(C_i) - g(C_i)\right|\right] > \eta\right) \\ &= Pr\left(\frac{B_1}{2\kappa^2 n}\sum_{i=1}^{n}\left[\left|\hat{g}(C_i) - g(C_i)\right|\right] > \frac{B_1}{2\kappa^2}\left(\sqrt{\frac{2D_1\theta}{n^2 h}} + \frac{2g_{max}\theta}{3n} + D_2 \cdot h^\beta\right)\right) \\ &= Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left[\left|\hat{g}(C_i) - g(C_i)\right|\right] > \left(\sqrt{\frac{2D_1\theta}{n^2 h}} + \frac{2g_{max}\theta}{3n} + D_2 \cdot h^\beta\right)\right) \\ &\leq \exp(-\theta). \end{aligned}$$

We need to bound the term $B_{n,1}(v) \equiv \left|R_{L,P}(v) - R_{L^n,D,g}(v)\right|$. By the union bound, for all $\mu > 0$

$$Pr\left(\sup_{v\in\mathcal{F}_\varepsilon} B_{n,1}(v) \geq B\sqrt{\frac{\mu}{2n}}\right) = Pr\left(\sup_{v\in\mathcal{F}_\varepsilon}\left|R_{L,P}(v) - R_{L^n,D,g}(v)\right| \geq B\sqrt{\frac{\mu}{2n}}\right)$$
$$\leq \sum_{v\in\mathcal{F}_\varepsilon} Pr\left(|R_{L,P}(v) - R_{L^n,D,g}(v)| \geq B\sqrt{\frac{\mu}{2n}}\right).$$

We showed that $(1-\Delta)\ell(C, v(Z))/g(C) + L(0, v(Z)) \leq B$. Note also that $R_{L,P}(v) = R_{L^n,P}(v) = R_{L^n,P,g}(v)$; That is, $R_{L,P}(v)$ is the expectation of $R_{L^n,D,g}(v)$. Hence by Hoeffding's inequality, the last term can then be bounded again by $2|\mathcal{F}_\varepsilon|\exp(-\mu)$, where $\mathcal{F}_\epsilon$ is an ε-net of $\lambda^{-1/2}\overline{B_H}$ with cardinality

$$|\mathcal{F}\varepsilon| = N\left(\lambda^{-1/2}B_H, \|\cdot\|_\infty, \epsilon\right) < \infty.$$

Define $\mu = \log(2|\mathcal{F}_\varepsilon|) + \theta$, then

$$Pr\left(\sup_{v\in\mathcal{F}_\varepsilon} B_{n,1}(v) \geq B\sqrt{\frac{ln(2|\mathcal{F}_\varepsilon|) + \theta}{2n}}\right) \leq \exp(-\theta)$$

In conclusion we have that

$$\begin{aligned}
&Pr\left(\lambda\|f_{D,\lambda}\|^2_{\mathcal{H}} + R_{L,P}(f_{D,\lambda}) - R^*_{L,P} - A_2(\lambda) \geq B\sqrt{\frac{2\mu}{n}} + \frac{3c_l\varepsilon}{\kappa} + 4c_L\varepsilon + 2\eta\right)\\
&\leq Pr\left(2\sup_{\|f\|_{\mathcal{H}}\leq\lambda^{-1/2}} |R_{L,P}(f) - R_{L^n,D}(f)| \geq B\sqrt{\frac{2\mu}{n}} + \frac{3c_l\varepsilon}{\kappa} + 4c_L\varepsilon + 2\eta\right)\\
&\leq Pr\left(2\left(\sup_{v\in\mathcal{F}_\varepsilon} |R_{L,P}(v) - R_{L^n,D}(v)| + \frac{3}{2\kappa}c_l\varepsilon + 2c_L\varepsilon\right) \geq B\sqrt{\frac{2\mu}{n}} + \frac{3c_l\varepsilon}{\kappa} + 4c_L\varepsilon + 2\eta\right)\\
&\leq Pr\left(2\left(\sup_{v\in\mathcal{F}_\varepsilon} B_{n,1}(v) + B_{n,2}(v)\right) \geq B\sqrt{\frac{2\mu}{n}} + 2\eta\right)\\
&\leq Pr\left(\sup_{v\in\mathcal{F}_\varepsilon} B_{n,1}(v) + B_{n,2}(v) \geq B\sqrt{\frac{\mu}{2n}} + \eta\right)\\
&\leq Pr\left(\sup_{v\in\mathcal{F}_\varepsilon} B_{n,1} \geq B\sqrt{\frac{ln(2|\mathcal{F}_\varepsilon|) + \theta}{2n}}\right) + Pr\left(\sup_{v\in\mathcal{F}_\varepsilon} B_{n,2}(v) \geq \eta\right)\\
&\leq \exp(-\theta) + \exp(-\theta) = 2\exp(-\theta)
\end{aligned}$$

and the result follows. □

Proof of Corollary 2

Proof Note that the only difference between Corollary 2 and Corollary 1 is in the term 2η. Recall that η is defined by $\eta \equiv \frac{B_1}{2\kappa^2}\left(\sqrt{\frac{2D_1\theta}{n^2h}} + \frac{2g_{max}\theta}{3n} + D_2 \cdot h^{\beta}\right)$. Choose h such that $h \underset{n\to\infty}{\to} 0$ and that $h^{0.5}n \underset{n\to\infty}{\to} \infty$. Then $\eta \underset{n\to\infty}{\to} 0$. Choose $\lambda = \lambda_n$ and $\epsilon = \left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\frac{1}{\sqrt{\lambda}}$. Then as in Corollary 1, all other terms converge to zero as $n \to \infty$ which implies consistency (Steinwart and Christmann 2008, Lemma 6.5). Since this holds for all probability measures $P \in \mathcal{P}$, we obtain $\mathcal{P}$-universal consistency. □

Appendix D

In this section we derive learning rates for cases I and II.

Definition 4 A learning method is said to learn with rate $\epsilon_n \subset (0,1]$ that converges to zero if for all $n \geq 1$ and all $\tau \in (0,1]$, $Pr\left(R_{L,P}(f_D) - R^*_{L,P} \leq c_P c_\tau \epsilon_n\right) \geq 1 - \tau$, where c_τ and c_P are constants such that $c_\tau \in [1,\infty)$ and $c_P > 0$.

We demonstrate how to derive learning rates from the same oracle inequalities used for the consistency proofs. While faster learning rates can be achieved under further assumptions in a similar manner, they further complicate the calculations and are beyond the scope of this paper.

Theorem 3 *Assume that* (A1)–(A4) *hold. Choose* $0 < \lambda_n < 1$ *and assume that there exist constants* $a \geq 1$, $p > 0$ *such that* $\log\left(N(B_H, \|\cdot\|_\infty, \epsilon)\right) \leq a\epsilon^{-2p}$. *Additionally, assume that there exist constants* $c > 0$, $\gamma \in (0,1]$ *such that* $A_2(\lambda) \leq c\lambda^\gamma$. *Then*

(i) *If g is known, the learning rate is given by* $n^{-\frac{\gamma}{(1+p)(2\gamma+1)}}$.
(ii) *If g is not known and the setup of Theorem 2 holds, then the leraning rate is given by* $n^{-\min\left(\frac{\gamma}{(1+p)(2\gamma+1)}, \frac{2\beta}{2\beta+1}\right)}$.

Proof of Theorem 3

Proof *Case I*

By Theorem 1,

$$\lambda\|f_{D,\lambda}\|^2_{\mathcal{H}} + R_{L,P}(f_{D,\lambda}) - R^*_{L,P} - A_2(\lambda)$$
$$\leq B\sqrt{\frac{2\log\left(2N(\lambda^{-1/2}B_H, \|\cdot\|_\infty, \epsilon)\right) + 2\theta}{n}} + \frac{2c_l\epsilon}{\kappa} + 4c_L\epsilon$$

with probability not less than $1 - \exp(-\theta)$. For any compact set $\mathcal{S} = [-S, S] \subset \mathbb{R}$, Both L and l are bounded and Lipschitz continuous with Lipschitz constants $c_L \leq 2\tau^{-2}(S+\tau)$ and $c_l = 2\tau^{-2}$. Hence,

$$
\begin{aligned}
&\lambda\|f_{D,\lambda}\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* - A_2(\lambda) \\
&\le B\sqrt{\frac{2\log\left(2N(B_H,\|\cdot\|_\infty,\sqrt{\lambda}\epsilon)\right)+2\theta}{n}} + \frac{2c_l\varepsilon}{\kappa} + 4c_L\varepsilon \\
&\le B\sqrt{\frac{2\log\left(2N(B_H,\|\cdot\|_\infty,\sqrt{\lambda}\epsilon)\right)+2\theta}{n}} + \frac{4\varepsilon}{\kappa\tau^2} + \frac{8(S+\tau)}{\tau^2}\varepsilon \\
&= B\sqrt{\frac{2\log\left(2N(B_H,\|\cdot\|_\infty,\sqrt{\lambda}\epsilon)\right)+2\theta}{n}} + M\cdot\epsilon
\end{aligned}
\tag{9}
$$

where $M = 4\tau^{-2}\left(\kappa^{-1} + 2(S+\tau)\right)$.

By the assumption $\log\left(N(B_H,\|\cdot\|_\infty,\epsilon)\right) \le a\epsilon^{-2p}$, we have that:

$$
\begin{aligned}
&\log\left(2N(B_H,\|\cdot\|_\infty,\sqrt{\lambda}\epsilon)\right) = \log(2) + \log\left(N(B_H,\|\cdot\|_\infty,\sqrt{\lambda}\epsilon)\right) \\
&\le \log(2) + a\left(\sqrt{\lambda}\epsilon\right)^{-2p} \le 2a\left(\sqrt{\lambda}\epsilon\right)^{-2p}.
\end{aligned}
$$

Choose $\epsilon = \left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\frac{1}{\sqrt{\lambda}}$. Then

$$
a\left(\sqrt{\lambda}\epsilon\right)^{-2p} = a\left(\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right)^{-2p}. \tag{10}
$$

By (9) and (10),

$$
\begin{aligned}
&\lambda\|f_{D,\lambda}\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* - A_2(\lambda) \\
&\le B\sqrt{\frac{4a\left(\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right)^{-2p}+2\theta}{n}} + M\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\frac{1}{\sqrt{\lambda}} \\
&\le B\left(\sqrt{\frac{4a\left(\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right)^{-2p}}{n}} + \sqrt{\frac{2\theta}{n}}\right) + \frac{M}{\sqrt{\lambda}}\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} \\
&= B\left(\frac{\sqrt{4a}\left(\left(\frac{p}{2}\right)^{\frac{-p}{1+p}}\left(\frac{2a}{n}\right)^{\frac{-p}{2+2p}}\right)}{\sqrt{n}}\right) + \frac{M}{\sqrt{\lambda}}\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + B\sqrt{\frac{2\theta}{n}} \\
&= \left(\frac{p}{2}\right)^{\frac{-p}{1+p}}\left[B\sqrt{2}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \frac{M}{\sqrt{\lambda}}\frac{p}{2}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \right] + B\sqrt{\frac{2\theta}{n}}
\end{aligned}
\tag{11}
$$

Recall that $B_2 = c_L\lambda^{-1/2} + 1$ and $B = B_1/2\kappa + B_2$, where B_1 is some bound on the derivative of the loss. Since $0 < \lambda < 1$, then $1 < \lambda^{-1/2}$, and therefor

$$B_2 \leq c_L \lambda^{-1/2} + \lambda^{-1/2} = \lambda^{-1/2}(c_L + 1) \leq \lambda^{-1/2}\left(\frac{2(S+\tau)}{\tau^2} + 1\right).$$

Earlier we defined M such that $\kappa = 4/M\tau^2 - 8(S+\tau)$. Thus,

$$\begin{aligned} B &\leq \frac{B_1}{2\kappa} + \frac{1}{\sqrt{\lambda}}\left(\frac{2(S+\tau)+\tau^2}{\tau^2}\right) = \frac{B_1(M\tau^2 - 8(S+\tau))}{8} + \frac{1}{\sqrt{\lambda}}\left(\frac{2(S+\tau)+\tau^2}{\tau^2}\right) \\ &= \frac{\sqrt{\lambda}B_1(M\tau^2 - 8(S+\tau)) + 8\left(\frac{2(S+\tau)+\tau^2}{\tau^2}\right)}{8\sqrt{\lambda}} \leq \frac{B_1(M\tau^2) + 8 + 16\left(\frac{S+\tau}{\tau^2}\right)}{8\sqrt{\lambda}} \equiv \frac{N}{\sqrt{\lambda}}, \end{aligned}$$

where we define $N \equiv 8^{-1}\left(B_1 M\tau^2 + 8 + 16\tau^{-2}(S+\tau)\right)$.

Hence we can bound (11) by

$$\begin{aligned} &\left(\frac{p}{2}\right)^{\frac{-p}{1+p}}\left[\frac{\sqrt{2}N}{\sqrt{\lambda}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \frac{M}{\sqrt{\lambda}}\frac{p}{2}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right] + \frac{N}{\sqrt{\lambda}}\sqrt{\frac{2\theta}{n}} \\ &\leq \left(\frac{p}{2}\right)^{\frac{-p}{1+p}}\frac{N}{\sqrt{\lambda}}\left[\sqrt{2}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \frac{Mp}{2N}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right] + \frac{N}{\sqrt{\lambda}}\sqrt{\frac{2\theta}{n}} \\ &\leq \left(\frac{p}{2}\right)^{\frac{-p}{1+p}}\frac{N}{\sqrt{\lambda}}\left[2\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \frac{Mp}{N}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right] + \frac{N}{\sqrt{\lambda}}\sqrt{\frac{2\theta}{n}} \end{aligned}$$

Choose

$$B_1 \geq \frac{4}{\tau^2} - \left(2 + 4\left(\frac{S+\tau}{\tau^2}\right)\right)\left(\frac{1}{\kappa} + 2S + 2\tau\right)^{-1}.$$

Note that

$$M = \frac{4}{\tau^2}\left(\frac{1}{\kappa} + 2(S+\tau)\right) \leq \frac{B_1(M\tau^2)}{4} + 2 + 4\left(\frac{S+\tau}{\tau^2}\right) = 2N.$$

Consequently, for our choice of B_1, we have that $M \leq 2N$ or $M/2N \leq 1$. Note also that $(p+1)(2/p)^{p/1+p} \leq 3$, hence:

$$\begin{aligned} &\left(\frac{p}{2}\right)^{\frac{-p}{1+p}}\frac{N}{\sqrt{\lambda}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\left(2 + \frac{M}{N}p\right) + \frac{N}{\sqrt{\lambda}}\sqrt{\frac{2\theta}{n}} \leq \left(\frac{p}{2}\right)^{\frac{-p}{1+p}}(p+1)2\frac{N}{\sqrt{\lambda}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} \\ &+ \frac{N}{\sqrt{\lambda}}\sqrt{\frac{2\theta}{n}} \leq \frac{N}{\sqrt{\lambda}}\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right]. \end{aligned}$$

Since $A_2(\lambda) \leq c\lambda^\gamma$ for constants $c > 0$, and $\gamma \in (0, 1]$,

$$\lambda\|f_{D,\lambda}\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* \leq c\lambda^\gamma + \frac{N}{\sqrt{\lambda}}\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right] \tag{12}$$

We would like to choose a sequence λ_n that will minimize the bound in (12).

Define

$$W(\lambda) = c\lambda^{\gamma} + \frac{N}{\sqrt{\lambda}}\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right].$$

Differentiating W with respect to λ and setting to zero yields:

$$\begin{aligned}
\frac{dW(\lambda)}{d\lambda} &= c\gamma\lambda^{\gamma-1} - \frac{1}{2}N\lambda^{-\frac{3}{2}}\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right] = 0 \\
&\Leftrightarrow \\
c\gamma\lambda^{\gamma-1} &= \frac{1}{2}N\lambda^{-\frac{3}{2}}\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right] \\
&\Leftrightarrow \lambda = \left(\frac{1}{2c\gamma}N\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right]\right)^{\frac{1}{\gamma+\frac{1}{2}}} \propto \left(\frac{1}{n}^{\frac{1}{2+2p}} + \left(\frac{1}{n}\right)^{\frac{1}{2}}\right)^{\frac{2}{2\gamma+1}} \\
&\Rightarrow \lambda \propto n^{-\frac{1}{(1+p)(2\gamma+1)}}
\end{aligned}$$

Since the second derivative of W (with respect to λ) is positive, λ is the minimizer. by (12),

$$Pr\left(R_{L,P}(f_{D,\lambda}) - R^*_{L,P} \leq c\lambda^{\gamma} + \frac{N}{\sqrt{\lambda}}\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right]\right) \geq 1 - \exp(-\theta). \tag{13}$$

By the choice of λ_n, the bound in Eq. (13) can be written as

$$\begin{aligned}
&cn^{-\frac{\gamma}{(1+p)(2\gamma+1)}} + Nn^{\frac{1}{2(1+p)(2\gamma+1)}}\left[6(2a)^{\frac{1}{2+2p}}n^{-\frac{1}{2+2p}} + (2\theta)^{\frac{1}{2}}n^{-\frac{1}{2}}\right] \\
&= cn^{-\frac{\gamma}{(1+p)(2\gamma+1)}} + N\cdot 6(2a)^{\frac{1}{2+2p}}n^{-\frac{\gamma}{(1+p)(2\gamma+1)}} + N(2\theta)^{\frac{1}{2}}n^{-\frac{2\gamma(1+p)+p}{2(1+p)(2\gamma+1)}} \\
&\leq cn^{-\frac{\gamma}{(1+p)(2\gamma+1)}} + N\cdot 6(2a)^{\frac{1}{2+2p}}n^{-\frac{\gamma}{(1+p)(2\gamma+1)}} + N(2\theta)^{\frac{1}{2}}n^{-\frac{\gamma}{(1+p)(2\gamma+1)}} \\
&= n^{-\frac{\gamma}{(1+p)(2\gamma+1)}}\left(c + N\cdot 6(2a)^{\frac{1}{2+2p}} + N(2\theta)^{\frac{1}{2}}\right) \\
&\leq Q(1+\sqrt{\theta})n^{-\frac{\gamma}{(1+p)(2\gamma+1)}}
\end{aligned}$$

where Q is a constant that does not depend on n or on θ.

In conclusion, by choosing a sequence λ_n that behaves like $n^{-1/(1+p)(2\gamma+1)}$, we have that the resulting learning rate is given by

$$Pr\left(R_{L,P}(f_{D,\lambda}) - R^*_{L,P} \leq Q(1+\sqrt{\theta})n^{-\frac{\gamma}{(1+p)(2\gamma+1)}}\right) \geq 1 - \exp(-\theta).$$

Case II

By Theorem 2,

$$\begin{aligned}
&\lambda\|f_{D,\lambda}\|^2_{\mathcal{H}} + R_{L,P}(f_{D,\lambda}) - R^*_{L,P} - A_2(\lambda) \\
&\geq B\sqrt{\frac{2\log\left(2N(\lambda^{-1/2}B_H, \|\cdot\|_\infty, \epsilon)\right) + 2\theta}{n}} + \frac{3c_l\varepsilon}{\kappa} + 4c_L\varepsilon + 2\eta
\end{aligned}$$

with probability not greater than $2\exp(-\theta)$ and where

$$\eta \equiv \frac{B_1}{2\kappa^2}\left(\sqrt{\frac{2D_1\theta}{n^2h}} + \frac{2g_{max}\theta}{3n} + D_2 \cdot h^\beta\right).$$

Choose

$$\begin{aligned}
\epsilon &= \left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\frac{1}{\sqrt{\lambda}},\\
M &= \frac{2}{\tau^2}\left(\frac{3}{\kappa} + 4(S+\tau)\right),\\
B_1 &\geq \frac{6}{\tau^2} - \left(6 + 12\left(\frac{S+\tau}{\tau^2}\right)\right)\left(\frac{3}{\kappa} + 4S + 4\tau\right)^{-1},
\end{aligned}$$

and define $N = 12^{-1}\left(B_1 M\tau^2 + 12 + 24\tau^{-2}(S+\tau)\right)$, then as in (12), a very similar calculation shows that

$$\lambda\|f_{D,\lambda}\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* \leq c\lambda^\gamma + \frac{N}{\sqrt{\lambda}}\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right] + 2\eta.$$

We would like to choose the bandwidth h that minimizes η. The minimum is achieved at h^* where

$$h^* = \left(\frac{D_1\theta}{2D_2^2\beta^2n^2}\right)^{1/2\beta+1}.$$

Substituting this result into η yields

$$\begin{aligned}
\eta = \frac{B_1}{2\kappa^2}\Bigg(&\sqrt{2D_1}\left(\frac{D_1}{2{D_2}^2\beta^2}\right)^{\frac{-1}{2(2\beta+1)}} n^{-\frac{2\beta}{2\beta+1}}\theta^{\frac{\beta}{2\beta+1}} + \frac{2g_{max}\theta}{3n}\\
&+D_2\left(\frac{D_1}{2D_2^2\beta^2}\right)^{\frac{\beta}{2\beta+1}}\theta^{\frac{\beta}{2\beta+1}}n^{-\frac{2\beta}{2\beta+1}}\Bigg)
\end{aligned}$$

or

$$\eta = \tilde{D}n^{-\min\left(1,\frac{2\beta}{2\beta+1}\right)}\max\left(\theta,\theta^{\frac{\beta}{2\beta+1}}\right) = \tilde{D}n^{-\frac{2\beta}{2\beta+1}}\max\left(\theta,\theta^{\frac{\beta}{2\beta+1}}\right)$$

Where $\tilde{D}$ is a constant that does not depend on θ or on n.

Hence,

$$\begin{aligned}
\lambda\|f_{D,\lambda}\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* &\leq c\lambda^\gamma + \frac{N}{\sqrt{\lambda}}\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right] + 2\eta\\
&\leq c\lambda^\gamma + \frac{N}{\sqrt{\lambda}}\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right] + \tilde{D}n^{-\frac{2\beta}{2\beta+1}}\max\left(\theta,\theta^{\frac{\beta}{2\beta+1}}\right)
\end{aligned}$$

Fig. 9 Misspecification for Settings 1–3. The Bayes risk is the dashed line and the boxplots of the following ▶ risks are compared: the KM-CSD with an RBF kernel, the KM-CSD with a linear kernel, AFT, Cox, ICc-forest, and PO, for sample sizes $n = 50, 100, 200, 400, 800$

Similarly to Case I, choosing $\lambda_n \propto n^{-\frac{1}{(1+p)(2\gamma+1)}}$ minimizes the last bound (note that the choice of λ_n does not depend on η). Hence the resulting learning rate is given by

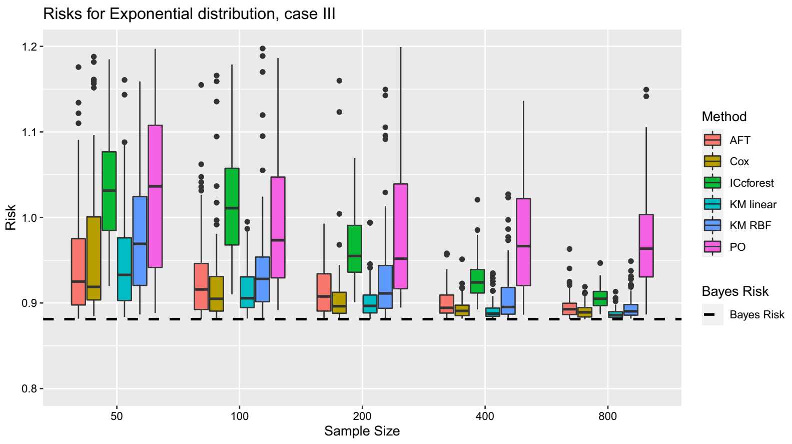
$$Pr(\mathcal{R}_{L,P}(f_{D,\lambda_n}) - \mathcal{R}^*_{L,P} \leq Q \max\left(\theta, 1+\sqrt{\theta}\right) n^{-min\left(\frac{\gamma}{(1+p)(2\gamma+1)}, \frac{2\beta}{2\beta+1}\right)}) \geq 1 - \exp(-\theta)$$

where Q is a constant that does not depend on n or on θ. □

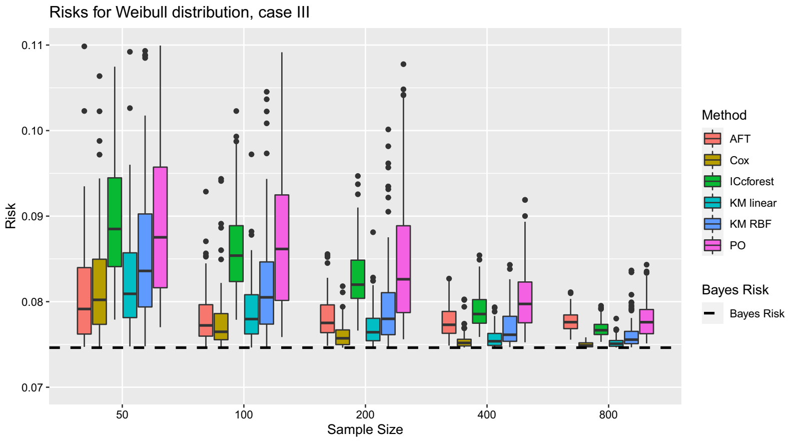
Appendix E

Simulations with misspecification of the censoring density

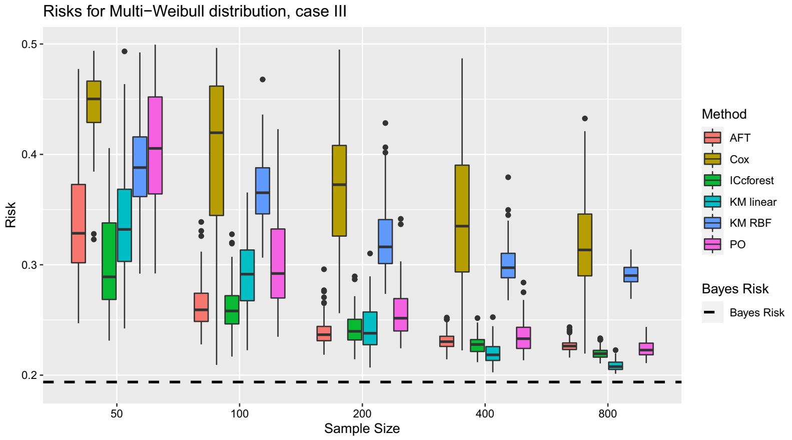
In this subsection we examine the effect of misspecification on our kernel machine estimator, for all 6 data-generating mechanisms from Sect. 5. Figures 9 and 10 present the boxplots of risks for Settings 1–6, where the censoring density is misspecified. We misspecified the censoring distribution using a beta distribution $Beta(0.9, 0.9)$, rescaled to the interval $[0, \tau]$, with density $g(C) = \frac{1}{\tau}\frac{C^{-0.1}(1-C)^{-0.1}}{\mathbf{B}(0.9,0.9)}$, where $\mathbf{B}(\alpha, \beta)$ is the beta function. Figures 9 and 10 shows that when the difference between the true censoring density and the misspecified density estimate is relatively small (as in our case), misspecification has a negligible effect on our estimator.



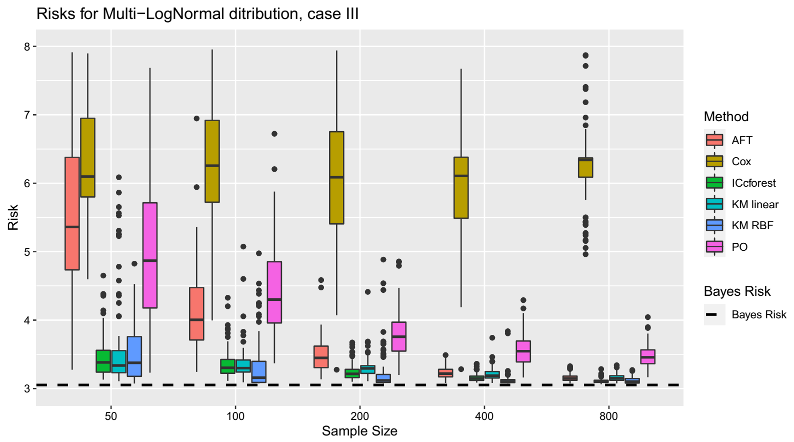
(a) Setting 1: Exponential distribution



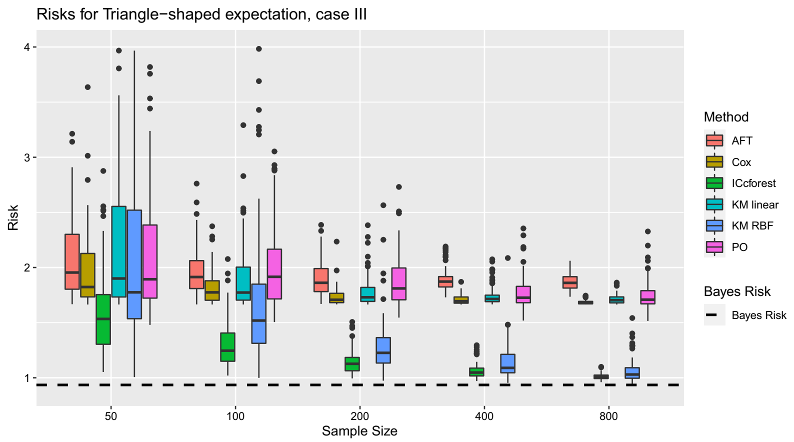
(b) Setting 2: Weibull distribution



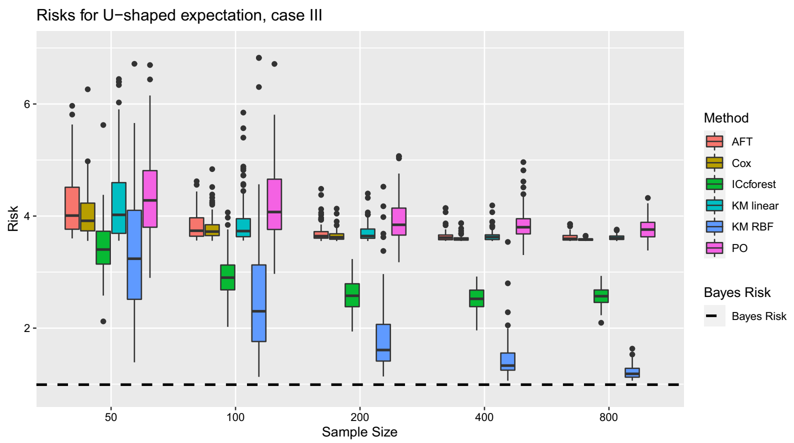
(c) Setting 3: Multi-Weibull distribution



(a) Setting 4: Multi-LogNormal distribution



(b) Setting 5: Triangle-shaped expectation



(c) Setting 6: U-shaped expectation

◀ **Fig. 10** Misspecification for Settings 4–6. The Bayes risk is the dashed line and the boxplots of the following risks are compared: the KM-CSD with an RBF kernel, the KM-CSD with a linear kernel, AFT, Cox, ICcforest, and PO, for sample sizes $n = 50, 100, 200, 400, 800$

References

Abrams, S., & Hens, N. (2015). Modeling individual heterogeneity in the acquisition of recurrent infections: An application to parvovirus B19. *Biostatistics*, *16*(1), 129–142. https://doi.org/10.1093/biostatistics/kxu031.

Anderson-Bergman, C. (2020). icenReg: Regression models for interval censored data. https://CRAN.R-project.org/package=icenReg.

Andrews, C., van der Laan, M., & Robins, J. (2005). Locally efficient estimation of regression parameters using current status data. *Journal of Multivariate Analysis*, *96*(2), 332–351.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, *26*(4), 641–647.

Becker, M., Lang, M., Richter, J., Bischl, B., & Schalk, D. (2020). mlr3tuning: Tuning for 'mlr3'. https://CRAN.R-project.org/package=mlr3tuning.

Burr, D., & Gomatam, S. (2002). On nonparametric regression for current status data. Technical report 2002–2014. Department of Statistics, Stanford University.https://statistics.stanford.edu/research/nonparametric-regression-current-status-data.

Cheng, G., & Wang, X. (2011). Semiparametric additive transformation model under current status data. *Electronic Journal of Statistics*, *5*, 1735–1764.

Cho, H., Jewell, N. P., & Kosorok, M. R. (2020). Interval censored recursive forests. arXiv:1912.09983 [stat].

Dehghan, M. H., & Duchesne, T. (2011). A generalization of Turnbull's estimator for nonparametric estimation of the conditional survival function with interval-censored data. *Lifetime Data Analysis*, *17*(2), 234–255.

Diamond, I. D., McDonald, J. W., & Shah, I. H. (1986). Proportional hazards models for current status data: Application to the study of differentials in age at weaning in Pakistan. *Demography*, *23*(4), 607–620. https://doi.org/10.2307/2061354.

Duong, T., Wand, M., Chacon, J., & Gramacki, A. (2020). ks: Kernel smoothing. https://CRAN.R-project.org/package=ks.

Eleuteri, A., & Taktak, A. F. G. (2011). Support vector machines for survival regression. In E. Biganzoli, A. Vellido, F. Ambrogi, & R. Tagliaferri (Eds.), *Computational intelligence methods for bioinformatics and biostatistics* (Vol. 7548, pp. 176–189). Berlin: Springer.

Fernandes, K., Vinagre, P., & Cortez, P. (2015). A proactive intelligent decision support system for predicting the popularity of online news. In *Portuguese conference on artificial intelligence* (pp. 535–546). Springer.

Fletcher, R. (1987). *Practical methods of optimization* (2nd ed.). New York: Wiley.

Fu, W., & Simonoff, J. S. (2017). Survival trees for interval-censored survival data. *Statistics in Medicine*, *36*(30), 4831–4842.

Goldberg, Y., & Kosorok, M. R. (2017). Support vector regression for right censored data. *Electronic Journal of Statistics*, *11*(1), 532–569.

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hens, N., Aerts, M., Shkedy, Z., Theeten, H., Van Damme, P., & Beutels, P. (2008). Modelling multisera data: The estimation of new joint and conditional epidemiological parameters. *Statistics in Medicine*, *27*(14), 2651–2664. https://doi.org/10.1002/sim.3089.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P., & Beutels, P. (2012). *Modeling infectious disease parameters based on serological and social contact data: A modern statistical perspective*. Berlin: Springer.

Hester, J. (2020). bench: High precision timing of R expressions. https://CRAN.R-project.org/package=bench. R package version 1.1.1.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, *36*(3), 1171–1220. https://doi.org/10.1214/009053607000000677.

Honda, T. (2004). Nonparametric regression with current status data. *Annals of the Institute of Statistical Mathematics*, *56*(1), 49–72.

Huang, J., & Wellner, J. A. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the first seattle symposium in biostatistics* (pp. 123–169). Springer.

Jewell, N. P., & van der Laan, M. (2004). Current status data: Review, recent developments and open problems. In N. Balakrishnan & C. R. Rao (Eds.), *Handbook of statistics, advances in survival analysis* (Vol. 23, pp. 625–642). Amsterdam: Elsevier.

Karatzoglou, A., Smola, A., Hornik, K., National ICT Australia (NICTA), Maniscalco, M. A., & Teo, C. H. (2019). kernlab: Kernel-based machine learning lab. https://CRAN.R-project.org/package=kernlab.

Khan, F. M., & Zubek, V. B. (2008). Support vector regression for censored data (SVRc): A novel tool for survival analysis. In *Eighth IEEE international conference on data mining, 2008. ICDM '08* (pp. 863–868).

Klein, J. P., & Goel, P. K. (Eds.). (1992). *Survival analysis: State of the art*. Dordrecht: Springer. https://doi.org/10.1007/978-94-015-7983-4.

Klein, J. P., & Moeschberger, M. L. (2005). *Survival analysis: Techniques for censored and truncated data* (2nd ed.). New York, NY: Springer.

Lang, M., Bischl, B., Richter, J., Schratz, P., Casalicchio, G., Coors, S., Au, Q., & Binder, M. (2020). mlr3: Machine learning in R—Next generation. https://CRAN.R-project.org/package=mlr3.

Lin, D. Y., Oakes, D., & Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika*, *85*(2), 289–298.

Liu, T., & Goldberg, Y. (2018). Kernel machines with missing responses. arXiv:1806.02865.

McMahan, C. S., & Wang, L. (2014). ICsurv: A package for semiparametric regression analysis of interval-censored data. https://CRAN.R-project.org/package=ICsurv. R package version 1.0. https://CRAN.R-project.org/package=ICsurv.

McMahan, C. S., Wang, L., & Tebbs, J. M. (2013). Regression analysis for current status data using the EM algorithm. *Statistics in Medicine*, *32*(25), 4452–4466. https://doi.org/10.1002/sim.5863.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *22*(1), 86–91. https://doi.org/10.2307/2346307.

Pölsterl, S., Navab, N., & Katouzian, A. (2016). An efficient training algorithm for kernel survival support vector machines. arXiv:1611.07054.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, *3*(4), 425–441. https://doi.org/10.1214/ss/1177012761.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rossini, A. J., & Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, *91*(434), 713. https://doi.org/10.2307/2291666.

Shen, X. (2000). Linear regression with current status data. *Journal of the American Statistical Association*, *95*(451), 842–852.

Shiao, H.-T., & Cherkassky, V. (2013). SVM-based approaches for predictive modeling of survival data. In *Proceedings of the international conference on data mining (DMIN)*.

Shiboski, S. C., & Jewell, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Association*, *87*(418), 360. https://doi.org/10.2307/2290266.

Shivaswamy, P. K., Chu, W., & Jansche, M. (2007). A support vector approach to censored targets. In *Seventh IEEE international conference on data mining (ICDM 2007)* (pp. 655–660). IEEE.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, *6*(1), 177–184.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. Berlin: Springer.

Sun, J., & Sun, L. (2005). Semiparametric linear transformation models for current status data. *Canadian Journal of Statistics*, *33*(1), 85–96.

Suykens, J. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, *9*(3), 293–300. https://doi.org/10.1023/A:1018628609742.

Therneau, T. M., & Lumley, T. (2016). survival: Survival analysis. https://CRAN.R-project.org/package=survival. R package version 2.40-1. https://CRAN.R-project.org/package=survival.

Tian, L., & Cai, T. (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika*, *93*(2), 329–342.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tsiatis, A. (2006). *Semiparametric theory and missing data*. Berlin: Springer.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Berlin: Springer.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *38*(3), 290–295.
Van Belle, V., Pelckmans, K., Suykens, J. A. K., & Van Huffel, S. (2007). Support vector machines for survival analysis. In *Proceedings of the third international conference on computational intelligence in medicine and healthcare (CIMED2007)* (pp. 1–8).
van der Laan, M. J., & Robins, J. M. (1998). Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association*, *93*(442), 693–701. https://doi.org/10.1080/01621459.1998.10473721.
van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Berlin: Springer.
van der Vaart, A. W., Dudoit, S., & van der Laan, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, *24*(3), 351–371.
Vapnik, V. (1999). *The nature of statistical learning theory* (2nd ed.). New York: Springer.
Wang, C., Sun, J., Sun, L., Zhou, J., & Wang, D. (2012). Nonparametric estimation of current status data with dependent censoring. *Lifetime data analysis*, *18*(4), 434–445. ISSN 1380-7870.
Wang, Y., Chen, T., & Zeng, D. (2016). Support vector hazards machine: A counting process framework for learning risk scores for censored outcomes. *Journal of Machine Learning Research*, *17*(167), 1–37.
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer.
Yao, W., Frydman, H., & Simonoff, J. S. (2019). An ensemble method for interval-censored time-to-event data. *Biostatistics*, 1–16. arXiv:1901.04599.
Yao, W., Frydman, H., & Simonoff, J. S. (2020). ICcforest: An ensemble method for interval-censored survival data. https://CRAN.R-project.org/package=ICcforest.