Support Vector Machines for Current Status Data

Yael Travis-Lumer

Thesis Submitted in Partial Fulfilment of the Requirements

for the Master's Degree

University of Haifa

Faculty of Social Sciences

Department of Statistics

May, 2015

Support Vector Machines for Current Status Data

By: **Yael Travis-Lumer**

Supervised by: **Dr. Yair Goldberg**

Thesis Submitted in Partial Fulfilment of the Requirements

for the Master's Degree

University of Haifa

Faculty of Social Sciences

Department of Statistics

May, 2015

Approved by:_____ Date:_____

(Supervisor)

Approved by:_____ Date:_____

(Chairperson of Master's studies Committee)

Contents

Support Vector Machines for Current Status Data

Yael Travis-Lumer

ABSTRACT

Current status data is a data format where the time to event is restricted to knowledge of whether or not the failure time exceeds a random monitoring time. We develop a support vector machine learning method for current status data that estimates the failure time expectation as a function of the covariates. In order to obtain the support vector machine decision function, we minimize a regularized version of the empirical risk with respect to a data-dependent loss. We show that the decision function has a closed form. Using finite sample bounds and novel oracle inequalities, we prove that the obtained decision function converges to the true conditional expectation for a large family of probability measures and study the associated learning rates. Finally we present a simulation study that compares the performance of the proposed approach to current state of the art.

List of Figures

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 support vector machine (SVM) learning method for estimation of the failure time expectation for current status data. SVM was originally introduced by Vapnik in the 1990's and is firmly related to statistical learning theory (Vapnik, 1999). The choice of SVMs for current status data is motivated by the fact that SVMs can be implemented easily, have fast training speed, produce decision functions that have a strong generalization ability and can guarantee convergence to the optimal solution, under some weak assumptions (Shivaswamy et al., 2007).

Current status data is a data format where the failure time T is restricted to knowledge of whether or not T exceeds a random monitoring time C. 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 of 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.

We present a support vector machine framework for current status data. We propose a learning method, denoted by CSD-SVM, for estimation of the failure time expectation. We investigate the theoretical properties of the CSD-SVM, 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), we construct a data dependent version of the quadratic loss. Since the failure time T is not observed, our data dependent loss function is based on the censoring time C and on the current status indicator. Finally, in order to obtain

a CSD-SVM decision function for current status data, we minimize a regularized version of the empirical risk with respect to this data-dependent loss.

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 hazard model, and the accelerated failure time (AFT) model (see, for example, Klein and Moeschberger, 2013). In the Cox model, the hazard function is assumed to be proportional to the exponent of a linear combination of the covariates. In the AFT model, the log of the failure time is assumed to be a linear function of the covariates. Several works including Diamond et al. (1986), Jewell and van der Laan (2004) and others have suggested the Cox proportinal 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. 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.

Over the past two decades, some learning algorithms for censored data have been proposed (such as neural networks and splitting trees), but mostly with no theoretical justification. Additionally, most of these algorithms cannot be applied to current status data but only to other, more common, censored data formats. Recently, several works suggested the use of SVMs for survival data. Van Belle et al. (2007) suggested the use of SVMs for survival analysis, and formulated the task as a ranking problem. Shortly after, Khan and Zubek (2008) suggested the use of SVMs for regression problems with censored data; this was done by asymmetrically modifying the ε-insensitive loss function. Both examples were empirically tested but lacked theoretical justification. **?** proposed an empirical quantile risk estimator, which can also be applied to right censoring, and studied the estimator's performance. Goldberg and Kosorok (2012) studied an SVM framework for right censored data and proved that the algorithm converges to the optimal solution. Shiao and Cherkassky (2013) suggested two SVM-based formulations for classification problems with survival data. These examples illustrate that initial steps in this direction

have already been taken. However, as far as we know, the only SVM-based work 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 SVM for regression problems with interval censoring and, using simulations, showed that the method is comparable to other missing data tools and performs significantly well when the majority of the training data is censored.

The contribution of this work includes the development of an SVM framework for current status data, the study of the theoretical properties of the CSD-SVM, and the development of new oracle inequalities for censored data. These inequalities, together with finite sample bounds, allow us to prove consistency and to compute learning rates.

The paper is organized as follows. In section 2 we describe the formal setting of current status data and discuss the choice of the quadratic loss for estimating the conditional expectation. In section 3 we present the proposed CSD-SVM and its corresponding data-dependent loss function. Section 4 contains the main theoretical results, including finite sample bounds, consistency proofs and learning rates. In section 5 we illustrate the estimation procedure and show that the solution has a closed form. Section 6 contains the simulations. Concluding remarks are presented in section 7. The lengthier proofs appear in Appendix A. The Matlab code for both the algorithm and for the simulations can be found in the 7.

2 Preliminaries

In this section we present the notations 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 i.i.d. random triplets $D = \{(Z_1, C_1, \Delta_1), \ldots, (Z_n, C_n, \Delta_n)\}$. 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 censoring time, the indicator $\Delta = \mathbf{1}\{T \leq C\}$ is the current status indicator at time C, and is contained in the interval $[0, \tau] \equiv \mathcal{Y}$ for some constant $\tau > 0$. 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, 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 $\mathcal{Z} \times \mathcal{Y} \times \mathbb{R}$ to $[0, \infty)$.

Let $L : \mathcal{Z} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ 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 $R_{L,P}(f) \equiv E_P[L(Z, Y, f(Z))] = \int_{Z \times Y} L(z, y, f(z))\, dP(z, y)$. 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 $R_{L,D}(f) = \frac{1}{n}\sum_{i=1}^{n} L(z_i, y_i, f(z_i))$.

For example, it is well known (see, for example, Hastie et al., 2009) that the conditional expectation is the Bayes decision function with respect to the quadratic loss.

3 Support Vector Machines for Current Status 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 $k : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$. By definition, if k is a universal kernel, then $\mathcal{H}$ is dense in the space of continuous functions on $\mathcal{Z}$, $C(\mathcal{Z})$ (see, for example, Steinwart and Christmann 2008, Definition 4.52). Let us fix such an RKHS $\mathcal{H}$ and denote its norm by $\|\cdot\|_{\mathcal{H}}$ and let $\{\lambda_n\} > 0$ be some sequence of regularization constants. An SVM decision function for uncensored data is defined by:

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

We recall that current status data consists of n independent and identically-distributed random triplets $D = \{(Z_1, C_1, \Delta_1), \ldots, (Z_n, C_n, \Delta_n)\}$. 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 $G(\cdot|Z = z)$. For current status data, we introduce the following identity between risks, following van der Laan and Robins (1998). We extend this notion and incorporate loss functions and covariates in the following identity. 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.

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

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

where the equality before last follows from integration by parts. Note also that $(1 - \Delta) =$

$\mathbf{1}\{T > C\}$ and thus

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

This shows that the risk can be represented as the sum of two terms

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

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 the data-dependent loss function

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

Note that this function need not be convex nor a loss function. For the quadratic loss function, our data-dependent 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. In order to ensure positivity we add a constant term that does not depend on f, and so our loss becomes $\widetilde{L^n}(D, (Z, C, \Delta, f(Z))) = \frac{(1-\Delta)2(C-f(Z))}{\hat{g}(C|Z)} + (f(Z))^2 + a$, where for a fixed dataset of length n, $a = \max_{1\le i\le n}\left\{\frac{(1-\Delta_i)}{(\hat{g}(C_i|Z_i))^2}\right\}$. 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 here after.

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

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

Note that if the censoring mechanism is not known, we can replace the density g with its estimate $\hat{g}$; in this case our loss function becomes $L^n(D,(Z,C,\Delta,s)) = \frac{(1-\Delta)2(C-s)}{\hat{g}(C|Z)} + (s)^2$ and the SVM decision function is

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

(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 not censored and thus simple density estimation techniques can be used in order to estimate the density g.

4 Theoretical Results

In this section we prove consistency of the CSD-SVM learning method for a large family of probability measures and construct learning rates. We first assume that the censoring mechanism is known, and thus 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 CSD-SVM decision function and the Bayes risk in order to form finite sample bounds. We use this result, together with oracle inequalities, to show that the CSD-SVM converges in probability to the Bayes risk. That is, we demonstrate that for a very large family of probability measures, the CSD-SVM learning method is consistent. We then consider the case in which the censoring mechanism is not known 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 CSD-SVM is consistent even when the censoring distribution is not known. Finally we construct learning rates for the CSD-SVM learning method for both g known and unknown.

Definition 1. *Let* $L(y,s) = \frac{(y-s)^2}{\tau^2}$ *be the normalized quadratic loss, let* $l(y,s) = \frac{2(y-s)}{\tau^2}$ *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 data-dependent version of this loss.*

For simplicity, we use the normalized version of the quadratic 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. $L(y,0) \leq 1$ *for all* $y \in \mathcal{Y}$ *and* $\ell(y,s) \leq B_1$ *for all* $(y,s) \in \mathcal{Y} \times \mathcal{S}$ *and for some constant* $B_1 > 0$.

We need the following assumptions:

(A1) The censoring time C is independent of the failure time T given Z.

(A2) C takes its values in the interval $[0,\tau]$ and $\inf_{z\in\mathcal{Z},c\in C} g(c|z) \geq 2K > 0$, for some $K > 0$.

(A3) $\mathcal{Z} \subset \mathbb{R}^d$ is compact .

(A4) $\mathcal{H}$ is an RKHS of a continuous kernel k with $\|k\|_\infty \leq 1$.

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 $B = \frac{B_1}{2K} + B_2$, where B_1 is defined in Remark 1.

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 $\lambda > 0$, $n \geq 1$, $\varepsilon > 0$, and $\theta > 0$, with probability not less than $1 - e^{-\theta}$*

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

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

The proof of this theorem appears in Appendix A.1.

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

$$\lim_{n\to\infty} P(D \in (\mathcal{Z}\times\mathcal{Y})^n \,:\, \mathcal{R}_{L,P}(f_{D,\lambda_n}) \leq \mathcal{R}^*_{L,P} + \epsilon) = 1 \tag{2}$$

where $\mathcal{R}^*_{L,P}$ is the Bayes risk. Universal consistency means that (2) 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 probablity measures that satisfy assumptions (A1)-(A2). We say that a learning method is $\mathcal{P}$-universal consistent when (2) 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) $\inf_{f\in\mathcal{H}} \mathcal{R}_{L,P}(f) = \mathcal{R}^*_{L,P}$, for all probability measures P on $\mathcal{Z}\times\mathcal{Y}$

Assumptio (A5) means that our RKHS $\mathcal{H}$ is rich enough to include the Bayes decision function.

Corollary 1. *Assume the setting of Theorem 1 and that Assumptio (A5) holds. Let* λ_n *be a sequence such that* $\lambda_n \underset{n\to\infty}{\to} 0$ *and* $\lambda_n n \underset{n\to\infty}{\to} \infty$. *Choose* $\epsilon = n^{-\rho}$, *for some* $\rho > 0$. *Then the CSD-SVM learning method is* $\mathcal{P}$*-universal consistent.*

Proof. In Theorem 1 we showed that

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

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

Choose $\lambda = \lambda_n$; from Assumption (A5) together with Lemma 5.15 of Steinwart and Christmann (2008, 5.15), $A_2(\lambda_n)$ converges to zero as n converges to infinity. Clearly

$$B\sqrt{\frac{2log(2N(\sqrt{\frac{1}{\lambda}}B_H, \left\|\cdot\right\|_\infty, \epsilon)) + 2\tau}{n}} \underset{n\to\infty}{\longrightarrow} 0.$$

Finally, from the choice of ϵ, it follows that both $\frac{2c_l\varepsilon}{K}$ and $4c_L\varepsilon$ converge to 0 as $n \to \infty$. Hence for every fixed θ,

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

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. □

4.2 Case II - The Censoring Density g is Unknown

In this section we form finite sample bounds for the case in which the censoring density is not known and needs to be estimated. We assume that the density of the censoring variable is estimated using nonparametric kernel density estimation. In Lemma 1 we construct finite sample bounds on the differnce 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 CSD-SVM learning method.

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

Definition 3. *The Hölder class $\sum(\beta, \mathcal{L})$ of functions $f : \mathbb{R} \mapsto \mathbb{R}$ is the set of $m = \lfloor \beta \rfloor$ times differentiable functions whose derivative $f^{(m)}$ satisfies $\left| f^{(m)}(x) - f^{(m)}(x') \right| \leq \mathcal{L} \left| x - x' \right|^{\beta - m}$ for some constant $\mathcal{L} > 0$.*

Lemma 1. *Let $K : \mathbb{R} \mapsto \mathbb{R}$ be a kernel function of order m satisfying $\int_{-\infty}^{\infty} K^2(u)du < \infty$ and define $\hat{g}(x) = \frac{1}{hn} \sum_{i=1}^{n} K\left(\frac{C_i - x}{h}\right)$ where h is the bandwidth. Suppose that the true density g satisfies $g(c) \leq g_{max} < \infty$. Let us also assume that $g(c)$ belongs to the* Hölder *class $\sum(\beta, \mathcal{L})$. Finally, assume that $\int_{-\infty}^{\infty} |u|^{\beta} |K(u)| \, du < \infty$. Then for any $\epsilon > 0$,*

$$Pr\left(\frac{1}{n} \sum_{i=1}^{n} |\hat{g}(c_i) - g(c_i)| > \epsilon + C_2 \cdot h^{\beta} \right) \leq \sqrt{\frac{C_1}{nh\epsilon^2}},$$

where $C_1 = g_{max} \int_{-\infty}^{\infty} K^2(v) \, dv$ and $C_2 = \frac{\mathcal{L}|\pi|^{\beta - m}}{m!} \int_{-\infty}^{\infty} |K(v)| \, |v|^{\beta} \, dv$ 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 A.2.

We would like to choose h that minimizes the sum of $C_2 \cdot h^{\beta}$ and $\sqrt{\frac{C_1}{nh\epsilon^2}}$. Define $U(h) = C_2 \cdot h^{\beta} + \sqrt{\frac{C_1}{nh\epsilon^2}}$. Taking the deivative of U with respect to h and setting to zero yields:

$$\frac{dU(h)}{dh} = \beta C_2 h^{\beta - 1} - \frac{1}{2} \sqrt{\frac{C_1}{n\epsilon^2}} h^{-\frac{3}{2}} = 0$$

$$\Leftrightarrow h = \left(\frac{\sqrt{C_1}}{2\beta C_2 \epsilon \sqrt{n}} \right)^{\frac{2}{2\beta+1}} = \kappa \left(n^{-\frac{1}{2}} \right)^{\frac{2}{2\beta+1}} \propto n^{-\frac{1}{2\beta+1}} \tag{3}$$

where $\kappa = \frac{(C_1)^{\frac{1}{2\beta+1}}}{(2\beta C_2 \epsilon)^{\frac{2}{2\beta+1}}}$. It can be shown that the second derivative of U is positive which guarentees that the zero of the derivative above is the minimizer. After substituting $h = \kappa n^{-\frac{1}{2\beta+1}}$ in U, we obtain that $U(\kappa n^{-\frac{1}{2\beta+1}}) \propto n^{-\frac{\beta}{2\beta+1}}$.

Choosing $\epsilon > 0$ such that $ln(\epsilon) = \frac{2\beta+1}{2\beta}\theta + \frac{1}{2}ln(C_1) - \frac{1}{2}ln(n) + \frac{1}{2\beta}ln(2\beta C_2)$ and substi-

tuting $h = \kappa n^{-\frac{1}{2\beta+1}}$, we obtain by Lemma 1 that

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n}|\hat{g}(c_i) - g(c_i)| > \epsilon + C_2\kappa^{\beta}n^{-\frac{\beta}{2\beta+1}}\right) \leq \sqrt{\frac{C_1 n^{\frac{1}{2\beta+1}}}{\kappa n\epsilon^2}} = e^{-\theta}.$$

We now move to construct finite sample bounds for the CSD-SVM 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* $\inf_{z\in\mathcal{Z},c\in C}\hat{g}\left(c|z\right) \geq K > 0$, *for some* $K > 0$. *Choose* α *such that*

$$0 < \left(C_1\right)^{\frac{1}{2}}\left(2\beta C_2\right)^{\frac{1}{2\beta}}n^{-\frac{1}{2}} < \alpha < 2\left(C_1\right)^{\frac{1}{2}}\left(2\beta C_2\right)^{\frac{1}{2\beta}}n^{-\frac{1}{2}}$$

and

$$ln(\alpha) = \frac{2\beta+1}{2\beta}\theta + \frac{1}{2}ln(C_1) - \frac{1}{2}ln(n) + \frac{1}{2\beta}ln(2\beta C_2).$$

Then for fixed $\lambda > 0$, $\theta > 0$, $n \geq 1$, $\varepsilon > 0$, *we have with probability not less than* $1 - 2e^{-\theta}$ *that*

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

where $\eta \equiv \frac{B_1\left(\alpha + C_2\cdot h^{\beta}\right)}{2K^2}$.

The proof of the theorem appears in Appendix A.3.

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

Corollary 2. *Let* λ_n *be a sequence such that* $\lambda_n \underset{n\to\infty}{\to} 0$ *and that* $\lambda_n n \underset{n\to\infty}{\to} \infty$. *Choose* $\epsilon = n^{-\rho}$, *for some* $\rho > 0$. *Assume the setting of Theorem 3, then the CSD-SVM learning method is consistent.*

The proof of the corollary is derived similarly to the proof of Corollary 1 (consistency - case I).

4.3 Learning rates

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* $\tau \in (0,1]$, $Pr\left(\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \leq c_P c_\tau \epsilon_n\right) \geq 1-\tau$, *where* c_τ *and* c_P *are constants such that* $c_\tau \in [1,\infty)$ *and* $c_P > 0$.

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

The proof of the theorem appears in Appendix A.4.

5 Estimation of the Failure Time Expectation

In this section we demonstrate how to compute the CSD-SVM decision function with respect to the quadratic loss. In addition we show that the solution has a closed form. Since $L^n(D,(Z,C,\Delta,s)) = \frac{(1-\Delta)2(C-s)}{g(C|Z)} + s^2$ is convex, then for any RKHS $\mathcal{H}$ over $\mathcal{Z}$ and for all $\lambda > 0$, it follows that there exists a unique SVM solution $f_{D,\lambda}$. In addition, by the Representer Theorem (Steinwart and Christmann, 2008, 5.5), there exists constants $\alpha = (\alpha_1, ..., \alpha_n)^T \in \mathbb{R}^n$ such that $f_{D,\lambda}(z) = \sum_{i=1}^n \alpha_i k(z, z_i),\ z \in \mathcal{Z}$. Hence the optimization problem reduces to estimation of the vector α. A more general approach will also include an intercept term b such that $f_{D,\lambda}(z) = \sum_{i=1}^n \alpha_i k(z, z_i) + b$.

Let $\Phi : \mathcal{Z} \to \mathcal{H}$ be the feature map that maps the input data into an RKHS $\mathcal{H}$ such that $\langle \Phi(z_j), \Phi(z) \rangle = k(z_j, z)$. Our goal is to find a function $f^c_{D,\lambda}$ that is the solution of (1). From the Representer Theorem, there exists a unique SVM decision function of the form $f_{D,\lambda} = \sum_{j=1}^n \bar{\alpha}_j \Phi(z_j) + b$.

Define for each $\alpha \in \mathbb{R}^n$ the function $w(\alpha)$ by $w(\alpha) = \sum_{j=1}^n \alpha_j \Phi(z_j)$.

Then for $C_\lambda = \frac{1}{n\lambda}$, the optimization problem reduces to:

$$\min_{w,r \in \mathbb{R}^n} \frac{C_\lambda}{2} \sum_{i=1}^n \left[\frac{(1-\Delta_i)2r_i}{\hat{g}(C_i|Z_i)} + (t_i - r_i)^2 \right] + \frac{1}{2}\|w\|^2$$

$$such\ that\ r_i = c_i - f(z_i)$$

$$where\ f(z_i) \equiv < w, \Phi(z_i) > + b.$$

This is an optimization problem under equality constraints and hence we will use the method of Lagrange multipliers. The Lagrangian is given by

$$\text{Lagrange}_\text{P} = \frac{C_\lambda}{2} \sum_{i=1}^n \left[\frac{(1-\Delta_i)2r_i}{\hat{g}(C_i|Z_i)} + (c_i - r_i)^2 \right] + \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i \left(c_i - < w, \Phi(z_i) > - b - r_i \right)$$

Minimizing the original problem Lagrange_P yields the following conditions for optimality:

$$w = \sum_{i=1}^n \alpha_i \Phi(z_i)$$

$$ri = \frac{\alpha_i}{C_\lambda} + c_i - \frac{(1-\Delta_i)}{\hat{g}(C_i|Z_i)}$$

$$\sum_{i=1}^{n} \alpha_i = 0.$$

Since these are equality constraints in the dual formulation, we can substitute them into Lagrange_P to obtain the dual problem Lagrange_D. By the strong duality theorem (Bazaraa et al., 2006, Theorem 6.2.4), the solution of the dual problem is equivalent to the solution of the primal problem.

$$\begin{aligned}
\text{Lagrange}_\text{D} =& \frac{C_\lambda}{2} \sum_{i=1}^{n} \left[\frac{(1-\Delta_i) 2 \left(\frac{\alpha_i}{C_\lambda} + c_i - \frac{(1-\Delta_i)}{2\hat{g}(C_i|Z_i)} \right)}{\hat{g}(C_i|Z_i)} + \left(\frac{(1-\Delta_i)}{2\hat{g}(C_i|Z_i)} - \frac{\alpha_i}{C_\lambda} \right)^2 \right] \\
&+ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(z_i, z_j) \\
&+ \sum_{i=1}^{n} \alpha_i \left(c_i - \sum_{j=1}^{n} \alpha_j k(z_i, z_j) - b - \left(\frac{\alpha_i}{C_\lambda} + c_i - \frac{(1-\Delta_i)}{2\hat{g}(C_i|Z_i)} \right) \right).
\end{aligned}$$

Some calculations yield:

$$\begin{aligned}
\text{Lagrange}_\text{D} =& \sum_{i=1}^{n} \frac{(1-\Delta_i)}{\hat{g}(C_i|Z_i)} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(z_i, z_j) - \frac{1}{2} \sum_{i=1}^{n} \frac{\alpha_i^2}{C_\lambda} \\
=& v^T \alpha - \frac{1}{2} \alpha^T \left(K + \frac{1}{C_\lambda} I \right) \alpha
\end{aligned}$$

subject to the constraint $\sum_{i=1}^{n} \alpha_i = 0$, and where $v^T = \left(\frac{(1-\Delta_1)}{\hat{g}(C_1|Z_1)}, ..., \frac{(1-\Delta_n)}{\hat{g}(C_n|Z_n)} \right)$.

This is a quadratic programming problem subject to equality constraints. Its solution satisfies:

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

Note that if we do not require an intercept term, the solution is $\alpha = \left(K + \frac{1}{C_\lambda} I\right)^{-1} v$. It is interesting to note that this solution is equivalent to the solution attained by the Representer Theorem for differentiable loss functions: $\alpha_i = \frac{-1}{2\lambda n} L^{'}\left(x_i, y_i, f_{D,\lambda}(x_i)\right)$ (Steinwart and Christmann, 2008, Section 5.2). In our case, $L_n(C_i, f(Z_i)) = \frac{(1-\Delta_i)2(C_i - f(Z_i))}{\hat{g}(C_i|Z_i)} + (f(Z_i))^2$; hence $\alpha_i = \frac{-1}{2\lambda n} L_n^{'}\left(C_i, f(Z_i)\right) = \frac{-1}{2\lambda n}\left(\frac{(1-\Delta_i)(-2)}{\hat{g}(C_i|Z_i)} + 2f(Z_i)\right)$ and since $f(Z_i) = \sum_{j=1}^{n} \alpha_j k(z_i, z_j)$, we see that $\alpha = \frac{1}{\lambda n} v - \frac{1}{\lambda n} K\alpha$, i.e., $\alpha = \left(K + \frac{1}{C_\lambda} I\right)^{-1} v$.

6 Simulation Study

In this section we test the CSD-SVM learning method on simulated data and compare its performance to current state of the art. We construct four different data-generating mechanisms, including one-dimensional and multi-dimentional settings. For each data type, we compute the difference between the CSD-SVM decision function and the true expectation. We compare this result to results obtained by the Cox model and by the AFT model. As a reference, we compare all these methods to the Bayes risk.

For each data setting, we considered two cases;: (i) the censoring density g is known; and (ii) the censoring density is not known. For the second setting, the distribution of the censoring variable was estimated using nonparametric kernel density estimation with a normal kernel. The code was written in Matlab, using the Spider library[1]. In order to fit the Cox model to current status data, we downloaded the 'ICsurv' R package (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 quantiles. For the AFT model, we used the 'surv reg' function in the 'Survival' R package (Therneau and Lumley, 2014). In order to call R through Matlab, we installed the R package rscproxy (Baier, 2012), installed the statconnDCOM server[2], and download the Matlab R-Link toolbox (Henson, 2004). For the kernel of the RKHS $\mathcal{H}$, we used both a linear kernel and a Gaussian RBF kernel $k(x_i, x_j) = exp\left(\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)$, where σ and C_λ were chosen using 5-fold cross-validation. The code for the algorithm and for the simulations is available for download; a link to the code can be found in the 7.

We consider the following four failure time distributions, corresponding to the four different data-generating mechanisms: (1) Weibull, (2) Multi-Weibull, (3) Multi-Log-Normal, and (4) an additional example where the failure time expectation is triangle shaped. We present below the CSD-SVM risks for each case and compare them to risks obtained by other methods. The risks are based on 100 iterations per sample size. The

[1]The Spider library for Matlab can be downloaded from http://www.kyb.tuebingen.mpg.de/bs/people/spider/

[2]Baier Thomas, & Neuwirth Erich (2007). Excel :: COM :: R. Computational Statistics, Volume 22, Number 1/April 2007. Physica Verlag.

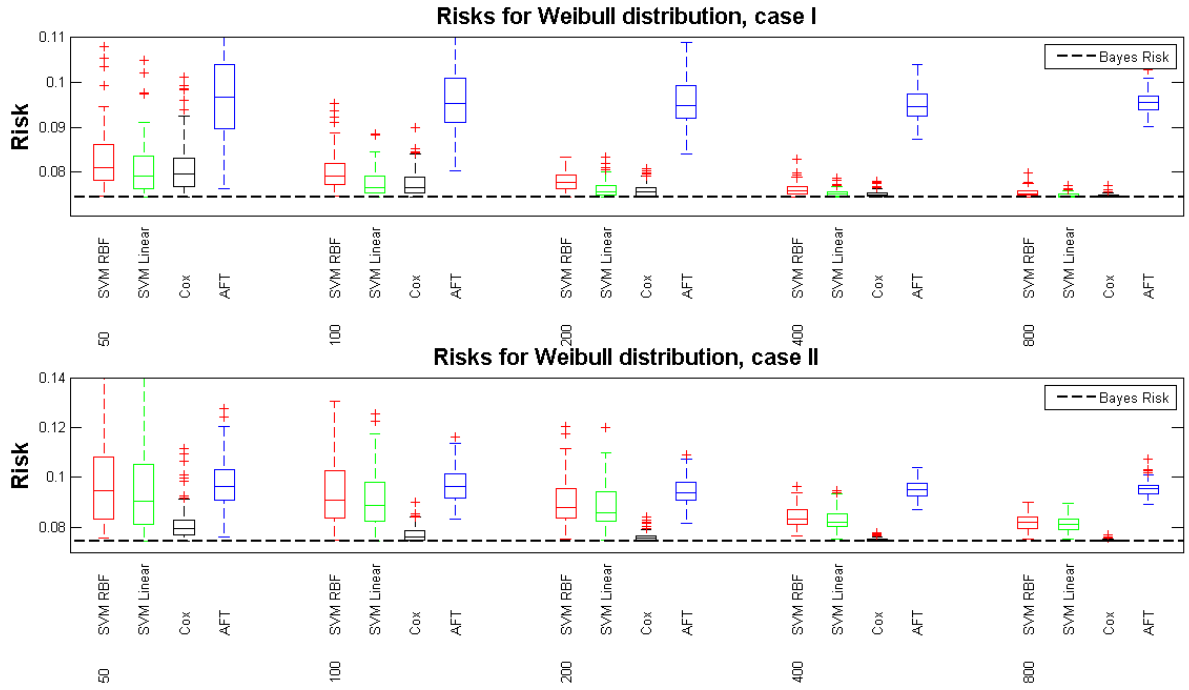

Figure 1: Weibull failure time distribution. The Bayes risk is the dashed black line and the boxlpots of the following risks are compared: CSD-SVM with an RBF kernel, CSD-SVM with a linear kernel, Cox and AFT, for sample sizes $n = 50, 100, 200, 400, 800$.

Bayes risk is also plotted as a reference.

In Setting 1 (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 $scale = e^{-\frac{1}{2}Z}$, $shape = 2$. The failure time was then truncated at $\tau = 1$.

Figure 1 compares the results obtained by the CSD-SVM to results achieved by the Cox model and by the AFT model, for different sample sizes. It should be noted that both the PH and the AFT assumption hold for the Weibull failure time distribution. In particular, when the PH assumption holds, estimation based on the Cox regression is consistent and efficient; hence, when the PH assumption holds, we will use the Cox regression as a benchmark. Figure 1 shows that when g is known, even though the CSD-SVM does not use the PH assumption or the AFT assumption, the results are comparable to those of the Cox regression, and are better than the AFT estimates, especially for larger sample sizes. However, when g is not known, the Cox model produces the smallest risks, but its superiority reduces as the sample size grows. This coincides with the fact that when g is not known, the learning rate of the CSD-SVM is slower.

In Setting 2 (Multi-Weibull failure-time), the covariates Z are generated uniformly

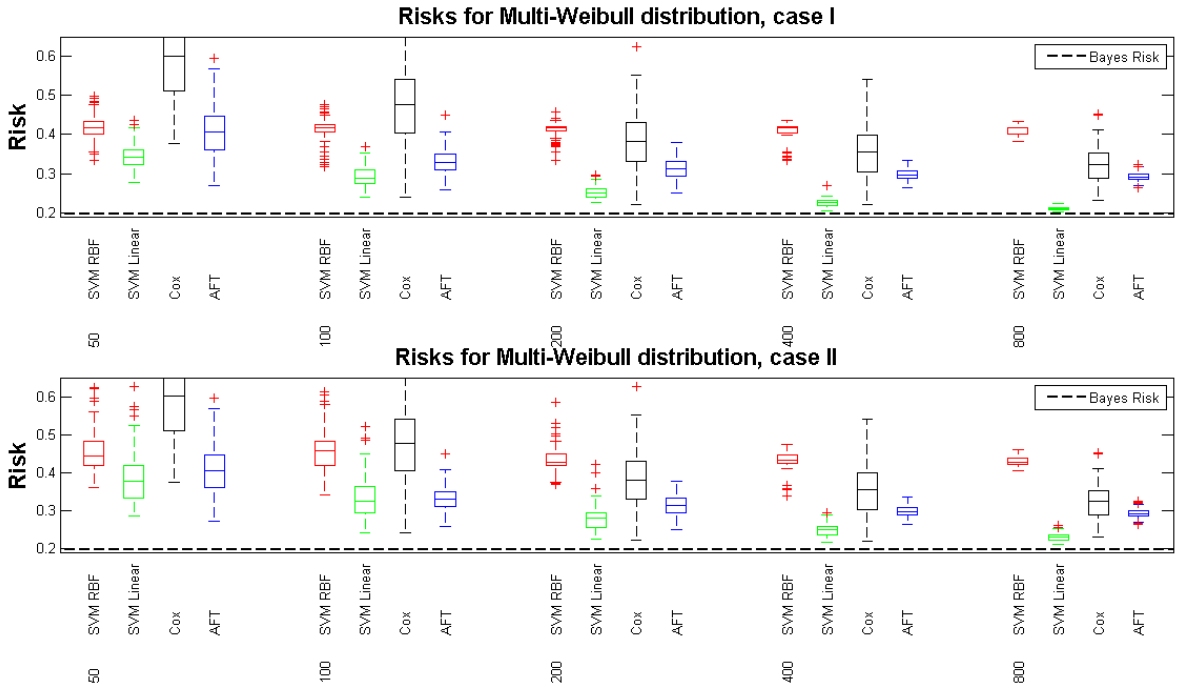

Figure 2: Multi-Weibull failure time distribution. The Bayes risk is the dashed black line and the boxlpots of the following risks are compared: the CSD-SVM with an RBF kernel, the CSD-SVM with a linear kernel, Cox and AFT for sample sizes $n = 50, 100, 200, 400, 800$.

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 $scale = -0.5Z_1 + 2Z_2 - Z_3$, $shape = 2$. The failure time was then truncated at $\tau = 2$. Note that this model depends only on the first three variables. In Figure 2, boxplots of risks are presented. Figure 2 illustrates that the CSD-SVM with a linear kernel is superior to the other methods, for all sample sizes and for both the cases g known and uknown. However, since the data may be sparse in the feature space, the CSD-SVM with an RBF kernel might require a larger sample size to converge.

In Setting 3 (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 $\mu = \frac{1}{2}(0.3Z_1 + 0.5Z_2 + 0.2Z_3)$, $\sigma = 1$. The failure time was then truncated at $\tau = 7$. Figure 3 presents the risks of the compared methods. This example illustrates that for small sample sizes, the CSD-SVM risks are significantly superior and converge quickly to the Bayes risk. As the sample size grows, the AFT also converges to the Bayes risk whereas the Cox estimates does not, as can be seen by the very high risks they produce. Note that for the Log-Normal distribution, even though

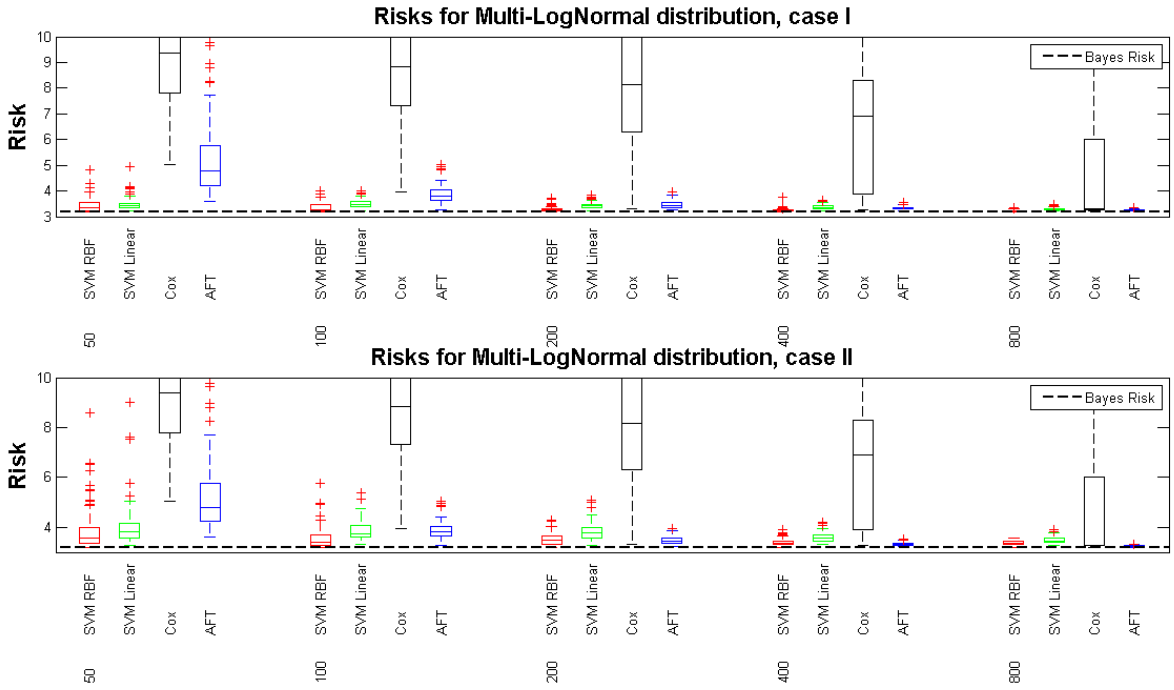

Figure 3: Multi-LogNormal failure time distribution. The Bayes risk is the dashed black line and the boxlpots of the following risks are compared: the CSD-SVM with an RBF kernel, the CSD-SVM with a linear kernel, Cox and AFT for sample sizes $n = 50, 100, 200, 400, 800$.

the AFT assumption is correct, the CSD-SVM manages to produce better or equivalent results.

In Setting 4, 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 was generated according to the following

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

The failure time was then truncated at at $\tau = 8$.

In Figure 4, the boxplots of risks are presented. As can be seen, the CSD-SVM with an RBF kernel is superior in both cases, for sufficently large sample sizes.

To illustrate the flexibility of the CSD-SVM, we also present a graphical representation of the true conditional expectation and its estimates, as a function of the covariates. Figure 5 compares the true expectation to the computed estimates for the case that g is known; these estimates are based on the first iteration. As can be seen, the CSD-SVM with an RBF kernel produces the most superior results.

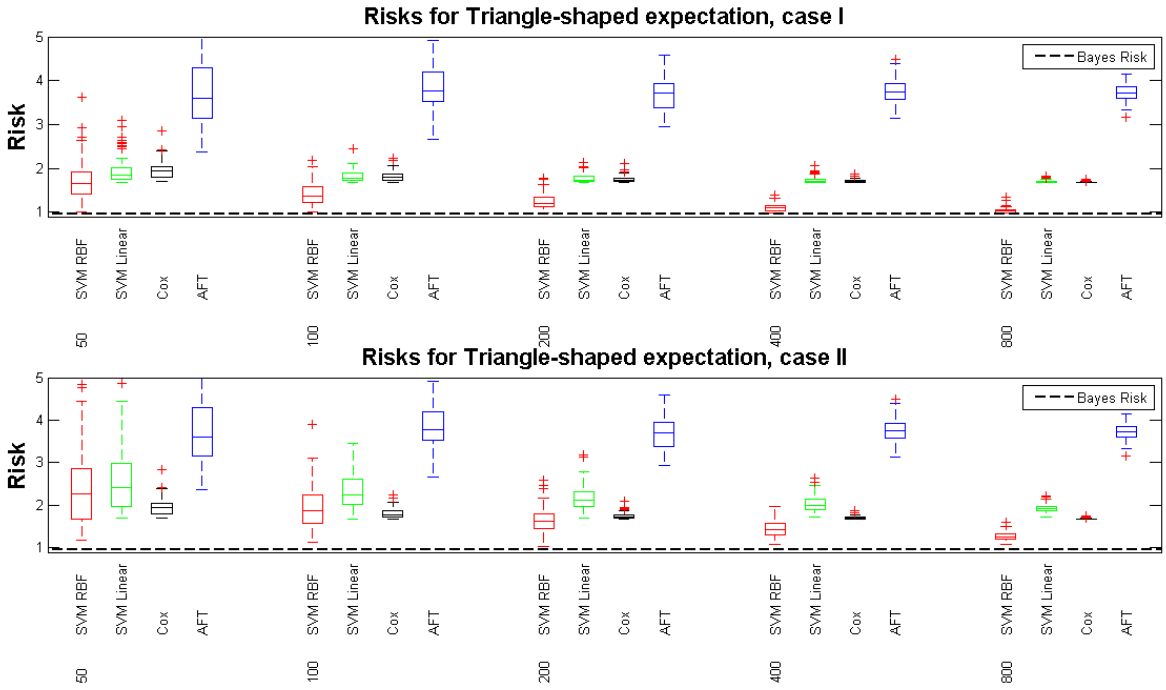

Figure 4: Triangle shaped failure time expectation. The Bayes risk is the dashed black line and the boxlpots of the following risks are compared: the CSD-SVM with an RBF kernel, the CSD-SVM with a linear kernel, Cox and AFT for sample sizes $n = 50, 100, 200, 400, 800$.

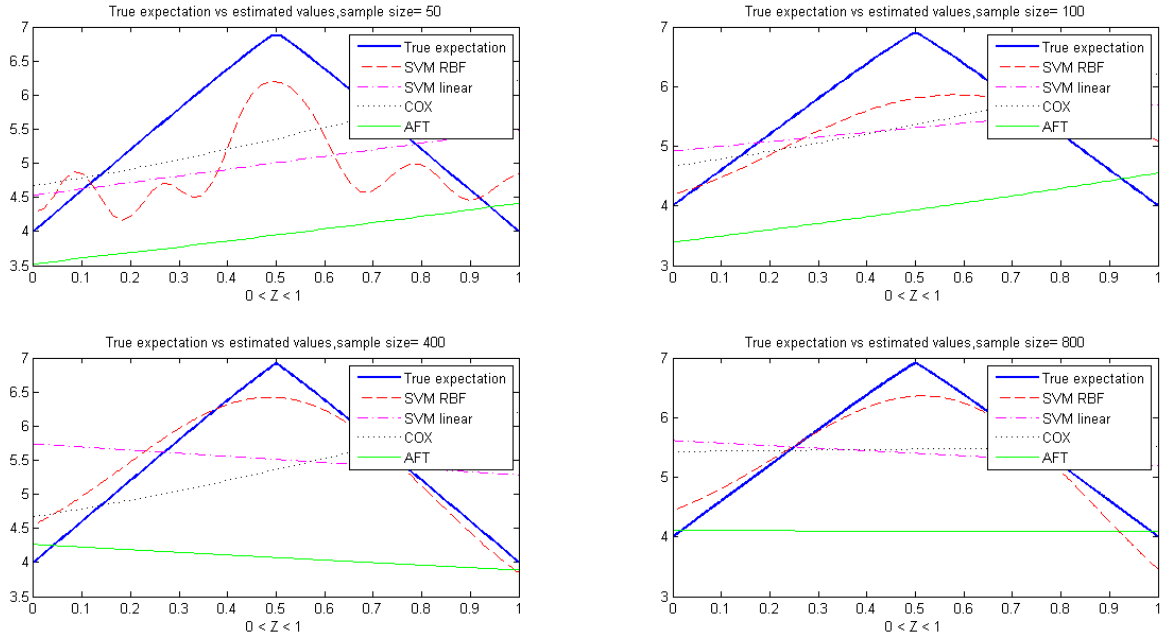

Figure 5: Triangle shaped failure time expectation, case I (g is known). The true expectation is the blue line. The following estimates are compared: the CSD-SVM with an RBF kernel, the CSD-SVM with a linear kernel, Cox and AFT for sample sizes $n = 50, 100, 400, 800$.

To summarize, Figures 1-5 showed that the CSD-SVM 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 CSD-SVM with an appropriate kernel was superior in three out of the four examples, especially when the true density g is known. It should be noted that even when the assumptions of the other models were true the CSD-SVM estimates were comparable. Additionally, when these assumptions fail to hold, the CSD-SVM estimates were generally better. The main advantage of the proposed SVM approach is that it does not assume any parametric form and thus may be superior, especially when the assumptions of other models fail to hold. Additionally, it seems that the CSD-SVM can perform well in higher dimensions.

7 Concluding Remarks

We proposed an SVM approach for estimation of the failure time expectation, studied its theoretical properties and presented a simulation study. 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 SVM approach to other estimation problems with current status data. 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 SVM learning methods for different types of censored data is of great interest.

Supplementary Material

The Matlab code is available for download and can be found at `http://stat.haifa.ac.il/~ygoldberg/research.html`. Please read the README.pdf for details on the files in this folder.

A Proofs

A.1 Proof of Theorem 1

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

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

Hence,

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

Hence $\|f_{Q,\lambda}\|_\infty \le \|f_{Q,\lambda}\|_{\mathcal{H}} \le \sqrt{\frac{R_{L,Q}(0)}{\lambda}}$ 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 $\|f_{Q,\lambda}\|_\infty \le \|f_{Q,\lambda}\|_{\mathcal{H}} \le \sqrt{\frac{1}{\lambda}}$ 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$\|f_{P,\lambda}\|_{\mathcal{H}} \le \sqrt{\frac{1}{\lambda}}$ we can write $f \in \sqrt{\frac{1}{\lambda}}\overline{B_H}$. 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).

Since $f_{D,\lambda}$ minimizes $\lambda\|f\|_{\mathcal{H}}^2 + R_{L,D}(f)$,

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

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

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

That is,

$$
\lambda \|f_{D,\lambda}\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* - A_2(\lambda) \leq 2 \sup_{\|f\|_{\mathcal{H}} \leq \sqrt{\frac{1}{\lambda}}} |R_{L,P}(f) - R_{L,D}(f)| \tag{4}
$$

Note that since L is Lipschitz continuous, $|L(y,s) - L(y,s')| \leq c_L|s - s'|$ for all $s, s' \in S$. From the discussion above, we are only interested in bounded functions $f \in \sqrt{\frac{1}{\lambda}}\overline{B_H}$. Then for all $f \in \sqrt{\frac{1}{\lambda}}\overline{B_H}$ we have

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

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

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

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

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

$$
|R_{L,P}(f) - R_{L,D}(f)| \leq |R_{L,P}(f) - R_{L,P}(h)| + |R_{L,P}(h) - R_{L,D}(h)| + |R_{L,D}(h) - R_{L,D}(f)| \equiv A_n + B_n + C_n \tag{5}
$$

First we will bound C_n;

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

where

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

and where

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

So we were able to bound C_n by $\frac{c_l \varepsilon}{2K} + c_L \varepsilon$.

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

As an interim summary, we showed that

$$\sup_{f\in\sqrt{\frac{1}{\lambda}}B_H}|R_{L,P}(f)-R_{L,D}(f)|\leq\sup_{h\in\mathcal{F}_\varepsilon}\underbrace{|R_{L,P}(h)-R_{L,D}(h)|}_{=B_n}+\frac{1}{K}c_l\varepsilon+2c_L\varepsilon. \tag{6}$$

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

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

Combining this with equation (4), we obtain that

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

By the union bound, the last expression is bounded by

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

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

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

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

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

which concludes the proof. □

A.2 Proof of Lemma 1

Proof. Note that

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

As in Tsybakov (2008, Proposition 1.1), define $\eta_i(c) = K\left(\frac{C_i - c}{h}\right) - E_g\left[K\left(\frac{C_i - c}{h}\right)\right]$. Then $\eta_i(c)$, for $i = 1, ..., n$ are i.i.d. random variables with zero mean and with variance:

$$\begin{aligned}
\mathrm{Var}\left[\eta_i(c)\right] = E_g\left[(\eta_i(c))^2\right] &= E_g\left[\left(K\left(\frac{C_i - c}{h}\right) - E_g\left[K\left(\frac{C_i - c}{h}\right)\right]\right)^2\right] \leq E_g\left[K^2\left(\frac{C_i - c}{h}\right)\right] \\
&= \int_u K^2\left(\frac{u - c}{h}\right) g(u)du \leq g_{max}\int_u K^2\left(\frac{u - c}{h}\right)du = g_{max}\int_v K^2(v)\,dv = C_1 h
\end{aligned}$$

where the equality before last follows from change of variables and where $C_1 = g_{max}\int_v K^2(v)\,dv$. Thus $\mathrm{Var}(\hat{g}(c)) = E_g\left[\left(\frac{1}{nh}\sum_{i=1}^{n}\eta_i(c)\right)^2\right] = \frac{1}{nh^2}E_g\left[\eta_1^2(c)\right] \leq \frac{C_1 h}{nh^2} = \frac{C_1}{nh}$.

By the Cauchy–Schwarz inequality we have that

$$E\left[|\hat{g}(c) - E\left[\hat{g}(c)\right]|\right] \leq \sqrt{E\left[|\hat{g}(c) - E\left[\hat{g}(c)\right]|^2\right]} = \sqrt{V(\hat{g}(c))}.$$

Hence $E\left[|\hat{g}(c) - E\left[\hat{g}(c)\right]|\right] \leq \sqrt{\frac{C_1}{nh}}$.Therfore, by Markov's inequality,

$$Pr(A > \epsilon) = Pr\left(\frac{1}{n}\sum_{i=1}^{n}|\hat{g}(c_i) - E\left[\hat{g}(c_i)\right]| > \epsilon\right) \leq \frac{E\left[|\hat{g}(c) - E\left[\hat{g}(c)\right]|\right]}{\epsilon} \leq \sqrt{\frac{C_1}{nh\epsilon^2}}.$$

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

$$B \equiv \frac{1}{n}\sum_{i=1}^{n} \left| E\left[\hat{g}(c_i)\right] - g(c_i) \right| \leq C_2 h^{\beta}$$

where $C_2 = \frac{\mathcal{L}|\pi|^{\beta-m}}{m!} \int_{-\infty}^{\infty} |K(v)| \, |v|^{\beta} \, dv < \infty$, and for some $\pi \in [0,1]$.

In conclusion, we showed that

$$\begin{aligned}
&Pr\left(\frac{1}{n}\sum_{i=1}^{n} |\hat{g}(c_i) - g(c_i)| > \epsilon + C_2 \cdot h^{\beta}\right) \\
\leq &Pr\left(\frac{1}{n}\sum_{i=1}^{n} |\hat{g}(c_i) - E\left[\hat{g}(c_i)\right]| + \frac{1}{n}\sum_{i=1}^{n} \left|E\left[\hat{g}(c_i)\right] - g(c_i)\right| > \epsilon + C_2 \cdot h^{\beta}\right) \\
\leq &Pr\left(\frac{1}{n}\sum_{i=1}^{n} |\hat{g}(c_i) - E\left[\hat{g}(c_i)\right]| + C_2 \cdot h^{\beta} > \epsilon + C_2 \cdot h^{\beta}\right) \\
= &Pr\left(\frac{1}{n}\sum_{i=1}^{n} |\hat{g}(c_i) - E\left[\hat{g}(c_i)\right]| > \epsilon\right) \leq \sqrt{\frac{C_1}{nh\epsilon^2}}
\end{aligned}$$

where h is the bandwidth. □

A.3 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, equation 5,

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

where

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

Since $A_n$does not depend on the data-set D, the same bound holds as in the proof of Theorem 1, that is, $A_n \leq \frac{c_l \varepsilon}{2K} + c_L \varepsilon$.

We bound C_n as follows:

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

Using the same arguments as in Theorem 1, we can bound C_n by $\frac{c_l \varepsilon}{K} + c_L \varepsilon$. Note that the only difference is in the denominator of $C_{n,1}$ since $\frac{1}{g} \leq \frac{1}{2K}$ and $\frac{1}{\hat{g}} \leq \frac{1}{K}$.

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

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

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

We bound B_n as follows

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

where

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

and where

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

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. By equation 3, let $h = \kappa n^{-\frac{1}{2\beta+1}}$, choose α such that

$$
0 < (C_1)^{\frac{1}{2}} (2\beta C_2)^{\frac{1}{2\beta}} n^{-\frac{1}{2}} < \alpha < 2 (C_1)^{\frac{1}{2}} (2\beta C_2)^{\frac{1}{2\beta}} n^{-\frac{1}{2}}
$$

and

$$
ln(\alpha) = \frac{2\beta+1}{2\beta}\theta + \frac{1}{2} ln(C_1) - \frac{1}{2} ln(n) + \frac{1}{2\beta} ln(2\beta C_2),
$$

and let $\eta = \frac{B_1\left(\alpha + C_2 \cdot h^\beta\right)}{2K^2}$, then by Lemma 1

$$
\begin{aligned}
& Pr(B_{n,2} > \eta) \leq Pr\left(\frac{B_1}{2K^2 n} \sum_{i=1}^{n} \left[|\hat{g}(C_i|Z_i) - g(C_i|Z_i)| \right] > \eta \right) \\
=& Pr\left(\frac{B_1}{2K^2 n} \sum_{i=1}^{n} \left[|\hat{g}(C_i|Z_i) - g(C_i|Z_i)| \right] > \frac{B_1\left(\alpha + C_2 \cdot h^\beta\right)}{2K^2} \right) \\
=& Pr\left(\frac{1}{n} \sum_{i=1}^{n} \left[|\hat{g}(C_i|Z_i) - g(C_i|Z_i)| \right] > \alpha + C_2 \cdot h^\beta \right) \\
\leq& \sqrt{\frac{C_1}{nh\alpha^2}} = e^{-\theta}.
\end{aligned}
$$

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

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

We showed that $\frac{(1-\Delta)\ell(C,v(Z))}{g(C|Z)} + L(0,v(Z)) \leq B$. Hence by Hoeffdings inequality, the last term can then be bounded again by $2|\mathcal{F}_\varepsilon|e^{-\mu}$, where $\mathcal{F}_\epsilon$ is an ε-net of $\sqrt{\frac{1}{\lambda}\overline{B_H}}$ with cardinality

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

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

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

In conclusion we have that

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

and the result follows. □

A.4 Proof of Theorem 3

Proof. **Case I**

By Theorem 1,

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

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

$$\begin{aligned}
&\lambda \left\| f_{D,\lambda} \right\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* - A_2(\lambda) \\
\leq & B\sqrt{\frac{2log(2N(B_H, \left\|\cdot\right\|_\infty, \sqrt{\lambda}\epsilon)) + 2\theta}{n}} + \frac{2c_l\varepsilon}{K} + 4c_L\varepsilon \\
\leq & B\sqrt{\frac{2log(2N(B_H, \left\|\cdot\right\|_\infty, \sqrt{\lambda}\epsilon)) + 2\theta}{n}} + \frac{4\varepsilon}{K\tau^2} + \frac{8(S+\tau)}{\tau^2}\varepsilon \\
= & B\sqrt{\frac{2log(2N(B_H, \left\|\cdot\right\|_\infty, \sqrt{\lambda}\epsilon)) + 2\theta}{n}} + M \cdot \epsilon
\end{aligned} \tag{7}$$

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

By the assumption $log(N(B_H, \left\|\cdot\right\|_\infty, \epsilon)) \leq a\epsilon^{-2p}$. Hence:

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

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

$$\begin{aligned}
& a\left(\sqrt{\lambda}\epsilon\right)^{-2p} \\
= & a\left(\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right)^{-2p}.
\end{aligned} \tag{8}$$

By (7) and (8),

$$
\begin{aligned}
&\lambda \left\| f_{D,\lambda} \right\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* - A_2(\lambda) \\
\leq & B\sqrt{\frac{4a\left(\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right)^{-2p} + 2\theta}{n}} + M\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\frac{1}{\sqrt{\lambda}} \\
\leq & B\left(\sqrt{\frac{4a\left(\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\right)^{-2p}}{n}} + \sqrt{\frac{2\theta}{n}}\right) + \frac{M}{\sqrt{\lambda}}\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} \\
= & B\left(\frac{\sqrt{4a}\left(\left(\frac{p}{2}\right)^{\frac{-p}{1+p}}\left(\frac{2a}{n}\right)^{\frac{-p}{2+2p}}\right)}{\sqrt{n}}\right) + \frac{M}{\sqrt{\lambda}}\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + B\sqrt{\frac{2\theta}{n}} \\
= & \left(\frac{p}{2}\right)^{\frac{-p}{1+p}}\left[B\sqrt{2}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \frac{M}{\sqrt{\lambda}}\frac{p}{2}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \right] + B\sqrt{\frac{2\theta}{n}}
\end{aligned}
\tag{9}
$$

Recall that $B_2 = c_L\lambda^{-1/2} + 1$ and $B = \frac{B_1}{2K} + B_2$, where B_1 is some bound on the derivative of the loss. Since $0 < \lambda < 1$, then $1 < \frac{1}{\sqrt{\lambda}}$, and therefor $B_2 \leq c_L\lambda^{-1/2} + \lambda^{-1/2} = \lambda^{-1/2}(c_L + 1) \leq \lambda^{-1/2}(\frac{2(S+\tau)}{\tau^2} + 1)$. Earlier we defined M such that $K = \frac{4}{M\tau^2 - 8(S+\tau)}$. Thus,

$$
B \leq \frac{B_1}{2K} + \frac{1}{\sqrt{\lambda}}\left(\frac{2(S+\tau)+\tau^2}{\tau^2}\right) = \frac{B_1(M\tau^2 - 8(S+\tau))}{8} + \frac{1}{\sqrt{\lambda}}\left(\frac{2(S+\tau)+\tau^2}{\tau^2}\right) =
$$

$$
= \frac{\sqrt{\lambda}B_1(M\tau^2 - 8(S+\tau)) + 8\left(\frac{2(S+\tau)+\tau^2}{\tau^2}\right)}{8\sqrt{\lambda}} \leq \frac{B_1(M\tau^2) + 8 + 16\left(\frac{S+\tau}{\tau^2}\right)}{8\sqrt{\lambda}} = \frac{N}{\sqrt{\lambda}}
$$

where we define $N \equiv {}^{B_1(M\tau^2)}/_8 + 1 + 2\left(\frac{S+\tau}{\tau^2}\right)$.

Hence we can bound (9) 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 $B_1 \geq \frac{4}{\tau^2} - \left(2 + 4\left(\frac{S+\tau}{\tau^2}\right)\right)\left(\frac{1}{K} + 2S + 2\tau\right)^{-1}$. Note that $M = \frac{4}{\tau^2}\left(\frac{1}{K} + 2(S+\tau)\right) \leq \frac{B_1(M\tau^2)}{4} + 2 + 4\left(\frac{S+\tau}{\tau^2}\right) = 2N$. Consequently, for $B_1 \geq \frac{4}{\tau^2} - \left(2 + 4\left(\frac{S+\tau}{\tau^2}\right)\right)\left(\frac{1}{K} + 2S + 2\tau\right)^{-1}$, we have that $M \leq 2N$ or $\frac{M}{2N} \leq 1$. Note also that $\left(\frac{2}{p}\right)^{\frac{p}{1+p}}(1+p) \leq 3$, hence:

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

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

$$\lambda \left\| f_{D,\lambda} \right\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* \leq c\lambda^\gamma + \frac{N}{\sqrt{\lambda}} \left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right] \tag{10}$$

We would like to choose a sequence λ_n that will minimize the bound in (10). 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 (10),

$$Pr\left(R_{L,P}(f_{D,\lambda}) - R_{L,P}^* \leq c\lambda^\gamma + \frac{N}{\sqrt{\lambda}}\left[6\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}}\right]\right) \geq 1 - e^{-\theta}. \tag{11}$$

By the choice of λ_n, the bound in equation (11) can be written as

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

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 $n^{-\frac{1}{(1+p)(2\gamma+1)}}$, we have that the resulting learning rate is given by

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

Case II

By Theorem 2,

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

where $\eta=\frac{2K^2\left(\alpha+C_2\cdot h^\beta\right)}{B_1}$ and with probability not greater than $2e^{-\theta}$. Choose $\epsilon=\left(\frac{p}{2}\right)^{\frac{1}{1+p}}\left(\frac{2a}{n}\right)^{\frac{1}{2+2p}}\frac{1}{\sqrt{\lambda}}$, $M=\frac{2}{\tau^2}\left(\frac{3}{K}+4(S+\tau)\right)$, $B_1\geq\frac{6}{\tau^2}-\left(6+12\left(\frac{S+\tau}{\tau^2}\right)\right)\left(\frac{3}{K}+4S+4\tau\right)^{-1}$, and define $N=\frac{B_1(M\tau^2)}{12}+1+2\left(\frac{S+\tau}{\tau^2}\right)$, then as in (10), a very similar calculation shows that

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

Choose $h=\kappa n^{-\frac{1}{2\beta+1}}$ as in (3) and choose α such that $ln(\alpha)=\frac{2\beta+1}{2\beta}\theta+\frac{1}{2}ln(C_1)-\frac{1}{2}ln(n)+\frac{1}{2\beta}ln(2\beta C_2)$ as in Theorem 2. Then by the definition of η,

$$\begin{aligned}\eta =& \frac{2K^2\left(\alpha + C_2 \cdot h^{\beta}\right)}{B_1} \\ =& \frac{2K^2\left(\alpha + C_2 \cdot \kappa^{\beta} n^{-\frac{\beta}{2\beta+1}}\right)}{B_1} \\ =& \frac{2K^2 e^{\frac{2\beta+1}{2\beta}} \left(C_1\right)^{\frac{1}{2}} \left(2\beta C_2\right)^{\frac{1}{2\beta}}}{B_1 n^{\frac{1}{2}}} + \frac{2K^2 C_2 \kappa^{\beta}}{B_1 n^{\frac{\beta}{2\beta+1}}} \\ \leq& \frac{2K^2\left(e^{\frac{2\beta+1}{2\beta}} \left(C_1\right)^{\frac{1}{2}} \left(2\beta C_2\right)^{\frac{1}{2\beta}} + C_2 \kappa^{\beta}\right)}{B_1 n^{\frac{\beta}{2\beta+1}}}.\end{aligned}$$

Hence,

$$\begin{aligned}&\lambda \left\| f_{D,\lambda} \right\|_{\mathcal{H}}^2 + R_{L,P}(f_{D,\lambda}) - R_{L,P}^* \leq c\lambda^{\gamma} + \frac{N}{\sqrt{\lambda}} \left[6 \left(\frac{2a}{n} \right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}} \right] + 2\eta \\ \leq& c\lambda^{\gamma} + \frac{N}{\sqrt{\lambda}} \left[6 \left(\frac{2a}{n} \right)^{\frac{1}{2+2p}} + \sqrt{\frac{2\theta}{n}} \right] + \frac{4K^2\left(e^{\frac{2\beta+1}{2\beta}} \left(C_1\right)^{\frac{1}{2}} \left(2\beta C_2\right)^{\frac{1}{2\beta}} + C_2 \kappa^{\beta}\right)}{B_1 n^{\frac{\beta}{2\beta+1}}}\end{aligned}$$

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 that the resulting learning rate is given by

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

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

B Bibliography

T Baier. rscproxy: statconn: provides portable c-style interface to r (StatConnector), 2012.

Mokhtar S. Bazaraa, Hanif D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. Wiley, 2006.

I. D. Diamond, J. W. McDonald, and I. H. Shah. Proportional hazards models for current status data: Application to the study of differentials in age at weaning in pakistan. *Demography*, 23(4):607–620, 1986.

A. Eleuteri and A.F.G. Taktak. Support vector machines for survival regression. In E. Biganzoli, A. Vellido, F. Ambrogi, and R. Tagliaferri, editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, volume 7548, pages 176–189. Springer Berlin Heidelberg, 2012a.

Y. Goldberg and M. R. Kosorok. Support vector regression for right censored data. Unpublished manuscript, 2012.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009.

V. Henschel, U. Mansmann, and C. Heiss. intcox: Iterated convex minorant algorithm for interval censored event data, 2013.

R. Henson. MATLAB R-link - File Exchange - MATLAB Central, 2004.

N. Jewell and M. van der Laan. Current status data: Review, recent developments and open problems. In N. Balakrishnan and C.R. Rao, editors, *Handbook of Statistics, Advances in Survival Analysis*, number 23, pages 625–642. Elsevier, 2004.

F.M. Khan and V.B. Zubek. Support Vector Regression for Censored Data (SVRc): A Novel Tool for Survival Analysis. In *Eighth IEEE International Conference on Data Mining, 2008. ICDM '08*, pages 863–868, December 2008.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, NY u.a., October 2013. 00000.

C. S. McMahan and L. Wang. *ICsurv: A package for semiparametric regression analysis of interval-censored data*, 2014.

C. S. McMahan, L. Wang, and J. M. Tebbs. Regression analysis for current status data using the EM algorithm. *Statistics in Medicine*, 32(25):4452–4466, 2013.

J. O. Ramsay. Monotone regression splines in action. *Statistical Science*, 3(4):425–441, 1988.

H.-T. Shiao and V. Cherkassky. SVM-based approaches for predictive modeling of survival data. *The 2013 International Conference on Data Mining*, 2013.

P.K. Shivaswamy, W. Chu, and M. Jansche. A support vector approach to censored targets. In *Seventh IEEE International Conference on Data Mining, 2007. ICDM 2007*, pages 655–660, 2007.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

T. M. Therneau and T. (original S.->R port and maintainer until Lumley. survival: Survival analysis, 2014.

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

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

V. Van Belle, K. Pelckmans, J.A.K. Suykens, and S Van Huffel. Support vector machines for survival analysi. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, Plymouth (UK), 2007.

M. J. van der Laan and J. M. Robins. Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association*, 93(442):693–701, 1998.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 2nd edition edition, 1999.

C. S. McMahan and L. Wang. ICsurv: A package for semiparametric regression analysis of interval-censored data, 2014.

הצנזור. דוגמא זו ממחישה את חשיבות הנושא ואת הצורך במציאת כלים חדשניים לניתוח נתונים מסוג זה. בעבודה זו אנו מתמקדים באמידת תוחלת זמן המאורע.

המטרה שלנו היא לאמוד את התוחלת המותנה של T באמצעות שימוש בשיטת ה-SVM, עבור נתונים מסוג Current Status Data. חלק ניכר מהעבודה עוסק בחקירת התכונות התיאורטיות של הגישה המוצעת, ובפרט עוסק בעקביות. SVMs ינסו ללמוד את פונק' הקשר האופטימאלית ביחס למזעור הסיכון האמפירי, ובתוספת קנס. כלומר בשיטות SVM עלינו להגדיר ראשית פונקצית הפסד מתאימה. מאחר ו- T אינו נצפה, יש קושי בהגדרת פונקצית הפסד זו. כדי להתגבר על בעיה זו, אנו מגדירים פונקצית הפסד התלויה במשתנה הצנזור C ובאינדיקטור הסטטוס הנוכחי $\Delta = \mathbf{1}\{T \le C\}$, אך לא בזמן המאורע. התכונה המעניינת של הצגה זו היא שהסיכון (תוחלת ההפסד) ביחס לפונקצית ההפסד המקורית, שווה לסיכון ביחס לפונקצית ההפסד החדשה.

מכיוון שאנו מעוניינים באמידת התוחלת המותנה, אנו משתמשים בפונקצית הפסד ריבועית, שכן התוחלת המותנה היא פונקצית החלטה בייס ביחס להפסד ריבועי. על מנת למצוא את פונקצית ההחלטה ה-SVM-ית עבור נתוני CSD, אנו ממזערים את הסיכון האמפירי ביחס לפונקצית הפסד התלויה בנתונים, ובתוספת קנס. תוצאה מעניינת היא שניתן להציג בעיית מזעור זו כבעיית תכנות ריבועי תחת אילוצי שוויון; לבעיה זו מצאנו פתרון סגור.

סוגיה נוספת בה נתקלנו היא האם ניתן להניח כי פונקצית הצפיפות של משתנה הצנזור הינה ידועה. פעמים רבות אכן ניתן להניח כי התפלגות זמן הצנזור הינה ידועה, לדוגמא, כאשר זו נקבעת ע"י החוקר. עם זאת, הנחה זו אינה תקפה תמיד. על מנת לפתור את הסוגיה, החלטנו לפצל את תהליך הניתוח לשני מקרים: (1) כאשר התפלגות זמן הצנזור ידועה ו-(2) כאשר התפלגות זמן הצנזור אינה ידועה ועלינו לאמוד את הצפיפות. במקרה שבו התפלגות זמן הצנזור אינה ידועה, אמדנו את הצפיפות בעזרת שיטות לא פרמטריות מבוססות גרעין. יש לציין שמשתנה הצנזור בעצמו אינו מצונזר ולכן ניתן להשתמש בשיטות רגילות לאמידת פונקצית צפיפות. פיתחנו תיאוריה מתאימה לכל אחד משני המקרים ובפרט הוכחנו עקביות ע"י כך שהראנו שהסיכון ביחס לפונקצית ההפסד החדשה מתכנס לסיכון בייס, וכן חישבנו קצבי לימוד לכל אחד מן המקרים.

לסיום השווינו את יעילות השיטה המוצעת לגישות קיימות אחרות בעזרת סימולציות. הראינו שהגישה שלנו ברת-השוואה לשיטות קיימות, ומציגה ביצועים טובים במיוחד כאשר התפלגות זמן המאורע אינה מגיעה ממשפחה פרמטרית או כאשר המשתנים המסבירים הם רב-מימדיים.

תקציר

נתונים מצונזרים מסוג Current Status Data הינם נתונים מתחום השרידות, עבורם המידע היחידי הזמין לגבי זמן המאורע T הוא האם T גדול או קטן מזמן הצנזור C. אנו פיתחנו גישת Support Vector Machines לנתונים מסוג זה אשר אומדת את התוחלת המותנה, מבלי להניח מודל פרמטרי. פונקצית הקשר מתקבלת ע״י מזעור הסיכון האמפירי ביחס לפונקצית הפסד התלויה בנתונים, ובתוספת קנס. הראינו שלפונקצית הקשר יש פתרון סגור והוכחנו, בעזרת אי-שוויונות אורקל חדשים, שפונקצית קשר זו מתכנסת לתוחלת המותנה האמיתית, עבור משפחה גדולה של מידות הסתברות. כמו כן, חישבנו את קצב הלימוד של השיטה. לסיום, הצגנו מספר סימולציות ובדקנו את הביצועים של גישה זו ביחס לגישות מתחרות. הראינו שהגישה שלנו ברת-השוואה לשיטות קיימות, ומציגה ביצועים טובים במיוחד כאשר התפלגות זמן המאורע אינה מגיעה ממשפחה פרמטרית או כאשר המשתנים המסבירים הם רב-מימדיים.

תקציר מורחב

בניתוח שרידות אנו מתעניינים בניתוח משך הזמן עד להופעת אירוע מסוים, כגון נסיגה של גידול בהקשר הרפואי, קלקול של מכונה בהקשר המכני, ומוות בהקשר הביולוגי. אמידת התפלגות זמן המאורע הינה בעלת חשיבות מכרעת במגוון תחומים, ובהקשר הרפואי בפרט. בהרבה מקרים, נתונים מתחום השרידות הינם מצונזרים, כלומר לא ניתן לצפות לחלוטין בזמני המאורע מכיוון שהמידע חסר. ספציפית, בעבודה זו אנו דנים בסוג מסוים של נתונים מצונזרים שנקראים Current Status Data, עבורם לכל דגימה (או פציינט), באיזושהי נקודת זמן, ידוע רק אם המאורע כבר התרחש, או לא. המטרה שלנו היא לפתח גישה כללית, נטולת מודל, לניתוח נתונים מסוג Current Status Data בעזרת שיטות מתחום הלמידה הסטטיסטית. בפרט, אנו מציעים גישה של Support Vector Machines (SVMs) לניתוח נתונים מסוג Current Status Data. הבחירה ב-SVMs לנתונים מצונזרים נובעת מכך שניתן ליישם את השיטה יחסית בקלות, קצב הלמידה של השיטה מהיר, יכולת ההכללה של השיטה טובה, ומובטחת התכנסות לפתרון האופטימאלי.

נתונים מסוג Current Status Data הינם נתונים מצונזרים, עבורם המידע היחידי הזמין לגבי זמן המאורע T הוא האם T גדול או קטן מזמן הצנזור C. באופן פורמאלי, נניח והנתונים מורכבים מ-n שלשות מהצורה $D=\{(Z_1,C_1,\Delta_1),\ldots,(Z_n,C_n,\Delta_n)\}$; כשאר $Z\in\Re^d$ וקטור של משתנים מסבירים, זמן המאורע T אינו שלילי, משתנה הצנזור C מקבל ערכים בקטע $[0,\tau]\equiv\mathcal{Y}$ עבור איזשהו $\tau>0$, ואינדיקטור הסטטוס הנוכחי מוגדר ע״י $\Delta=\mathbf{1}\{T\le C\}$. סוג זה של נתונים יחסית נפוץ וכולל דוגמאות ממגוון תחומים כגון חקר של דמוגרפיה, מחלות מדבקות, וסרטן. לדוגמא, בחקר הסרטן, T הוא הזמן מעת החשיפה לגורם מסרטן ועד להופעת גידול סרטני, ו- C הוא זמן אקראי בו מנתחים חיה על מנת לבדוק נוכחות או אי נוכחות של הגידול. מכאן שקשה לאמוד את התפלגות זמן המאורע T מכיוון שאנו לא צופים בזמן המאורע עצמו אלא רק בזמן

Support Vector Machines for Current Status Data

מאת: יעל טרוויס-לומר

בהנחיית: ד״ר יאיר גולדברג

עבודת גמר מחקרית (תזה) המוגשת כמילוי חלק מהדרישות

לקבלת התואר ״מוסמך האוניברסיטה"

אוניברסיטת חיפה

הפקולטה למדעי החברה

החוג לסטטיסטיקה

מאי, 2015

Support Vector Machines for Current Status Data

יעל טרוויס-לומר

עבודת גמר מחקרית (תזה) המוגשת כמילוי חלק מהדרישות

לקבלת התואר "מוסמך האוניברסיטה"

אוניברסיטת חיפה

הפקולטה למדעי החברה

החוג לסטטיסטיקה

מאי, 2015