Nearest Neighbor and Kernel Survival Analysis: Nonasymptotic Error Bounds and Strong Consistency Rates

George H. Chen [1]

Abstract

We establish the first nonasymptotic error bounds for Kaplan-Meier-based nearest neighbor and kernel survival probability estimators where feature vectors reside in metric spaces. Our bounds imply rates of strong consistency for these nonparametric estimators and, up to a log factor, match an existing lower bound for conditional CDF estimation. Our proof strategy also yields nonasymptotic guarantees for nearest neighbor and kernel variants of the Nelson-Aalen cumulative hazards estimator. We experimentally compare these methods on four datasets. We find that for the kernel survival estimator, a good choice of kernel is one learned using random survival forests.

1. Introduction

Survival analysis arises in numerous applications where we want to reason about the amount of time until some critical event happens. For example, in health care, we may be interested in using electronic health records to predict how long a patient with a particular disease will live (e.g., Botsis et al. 2010; Ganssauge et al. 2016), or how much time a patient has before a disease relapses (e.g., Zupan et al. 2000). In criminology, we may be interested in predicting the time until a convicted criminal reoffends (Chung et al., 1991).

A fundamental task in survival analysis is estimating the survival probability over time for a specific subject (for ease of exposition, we stick to using standard survival analysis terminology in which the critical event of interest is death). Formally, suppose a subject has feature vector X (a random variable that takes on values in a feature space $\mathcal{X}$) and survival time T (a nonnegative real-valued random variable). For a given feature vector $x \in \mathcal{X}$, our goal is to estimate the conditional survival function $S(t|x) := \mathbb{P}(T > t|X = x)$ for time $t \geq 0$.

To estimate S, we assume that we have access to n training subjects. For the i-th subject, we have the subject's feature vector $X_i \in \mathcal{X}$ as well as two observations: $\delta_i \in \{0, 1\}$ indicates whether we observe the survival time for the i-th subject, and $Y_i \in \mathbb{R}_+$ is the survival time for the i-th subject if $\delta_i = 1$ or the "censoring time" if $\delta_i = 0$. The censoring time gives a lower bound for the i-th subject's survival time (e.g., when we stop collecting training data, the i-th subject might still be alive, in which case that is when the subject's true survival time is "censored" and we only know that the subject survives beyond the time of censoring).

Many approaches have been devised for estimating the conditional survival function S. Most standard approaches impose strong structural assumptions on S via constraining the hazard function $h(t|x) := -\frac{\partial}{\partial t} \log S(t|x)$. For example, the Cox proportional hazards model decouples the effects of time $t \geq 0$ and of feature vector $x \in \mathbb{R}^d$ by assuming the factorization $h(t|x) = h_0(t) \exp(\beta^\top x)$, where positive-valued function h_0 and vector $\beta \in \mathbb{R}^d$ are parameters (Cox, 1972). After estimating h_0 and β from training data, then for any feature vector x, we can estimate the hazard function $h(t|x)$ by plugging in estimates for h_0 and β. Integrating the estimate for $h(t|x)$ thus yields an estimate for $S(t|x) = \exp(-\int_0^t h(s|x)ds)$. Other standard approaches such as the Aalen additive model (Aalen, 1989) and accelerated failure time models (Kalbfleisch & Prentice, 2002, Chapter 7) also impose structure on hazard function $h(t|x)$ and are typically used with parametric assumptions. More recent approaches include, for instance, modifying the Cox proportional hazards model by replacing the inner product $\beta^\top x$ with a nonlinear function of x that is encoded as a deep net (Katzman et al., 2018), or completely specifying S via a hierarchical generative model (Ranganath et al., 2016).

Rather than making structural assumptions on S, Beran (1981) takes a nonparametric approach using nearest neighbors and kernels. The idea is simple: there already is a nonparametric estimator for the marginal survival function $S_{\text{marg}}(t) := \mathbb{P}(T > t)$ known as the Kaplan-Meier estimator (Kaplan & Meier, 1958). This estimator does not use feature vectors. We can incorporate feature vectors in a straightforward manner. For a test subject with feature vector x, we first find training subjects whose feature vectors

[1] Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: George H. Chen <georgechen@cmu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

are sufficiently close to x (e.g., pick the k closest). We apply the Kaplan-Meier estimator to just these nearby subjects to estimate the conditional survival probability function $S(t|x)$ (the kernel variant can weight training subjects differently). Beran (1981) provided consistency results for these k-NN and kernel estimates for S, while Dabrowska (1989), Van Keilegom & Veraverbeke (1996), and Van Keilegom (1998) established nonasymptotic error bounds for the kernel variant when feature vectors are Euclidean.

In this paper, we present the first nonasymptotic error bounds for nearest neighbor and kernel estimators for S where feature vectors reside in the general setting of separable metric spaces (Euclidean space is a special case). Our error bounds lead to rates of strong consistency for both estimators across a wide range of distributions. Furthermore, our bounds are essentially optimal with respect to the number of training data n. In particular, note that $1 - S(\cdot|x)$ is a conditional CDF. If there is no right-censoring, the problem reduces to conditional CDF estimation. Up to a log factor, our error rates match an existing conditional CDF estimation error lower bound by Chagny & Roche (2014).

Our proof strategy also yields nonasymptotic error bounds for Nelson-Aalen-based nearest neighbor and kernel estimates of the conditional cumulative hazard function $-\log S(t|x)$. These bounds turn out to be crucial in how we derive generalization guarantees for automatic parameter selection (choosing the number of nearest neighbors or the kernel bandwidth) via a validation set.

Despite our theory handling a wide range of distances and kernels, both of these still have to be pre-specified by the user and, in practice, can lead to large prediction accuracy differences. As a simple heuristic, we propose using random survival forests (Ishwaran et al., 2008) to learn a kernel for the kernel survival estimator. We experimentally show that the resulting *adaptive* kernel estimator has prediction accuracy on par with regular random survival forests and is, in particular, typically as good as or better than other methods tested.

2. Model and Nonparametric Estimators

Model. The training data $(X_1, Y_1, \delta_1), \dots, (X_n, Y_n, \delta_n) \in \mathcal{X} \times \mathbb{R}_+ \times \{0, 1\}$ are assumed to be generated i.i.d. by the following process, stated for a generic data point (X, Y, δ):

1. Sample feature vector $X \sim \mathbb{P}_X$.
2. Sample nonnegative survival time $T \sim \mathbb{P}_{T|X}$.
3. Sample nonnegative censoring time $C \sim \mathbb{P}_{C|X}$. (Note that T and C are independent given X.)
4. Set $Y = \min\{T, C\}$, and $\delta = \mathbb{1}\{T \leq C\}$.

We refer to Y as the *observed time*, and δ as the *censoring indicator* (0 means censoring happened). For test feature vector $x \in \mathcal{X}$, we aim to estimate the conditional survival function $S(t|x) = \mathbb{P}(T > t | X = x)$ using the training data.

Nonparametric survival function estimators. All nonparametric estimators for S in this paper are based on the Kaplan-Meier estimator (Kaplan & Meier, 1958), restricted to a subset of the n training subjects. This estimator works as follows. Let $[n] := \{1, 2, \dots, n\}$ denote the set of all training subjects. For any subset of training subjects $\mathcal{I} \subseteq [n]$, the Kaplan-Meier estimator first identifies the unique times when death occurred, given by the set $\mathcal{Y}_\mathcal{I} := \{Y_j : j \in \mathcal{I} \text{ s.t. } \delta_j = 1\}$ (repeated observed times get counted once). Next, we keep track of how many deaths and how many subjects are at risk at any given time $t \geq 0$:

$$d_\mathcal{I}(t) := \sum_{j \in \mathcal{I}} \delta_j \mathbb{1}\{Y_j = t\}, \quad n_\mathcal{I}(t) := \sum_{j \in \mathcal{I}} \mathbb{1}\{Y_j \geq t\}.$$

Then the Kaplan-Meier estimator restricted to training subjects $\mathcal{I}$ is given by

$$\widehat{S}^{\text{KM}}(t|\mathcal{I}) := \prod_{t' \in \mathcal{Y}_\mathcal{I}} \Big(1 - \frac{d_\mathcal{I}(t')}{n_\mathcal{I}(t')}\Big)^{\mathbb{1}\{t' \leq t\}}.$$

This equation has a simple interpretation: if we sort the unique death times $\mathcal{Y}_\mathcal{I}$ as $t_1 < t_2 < \cdots < t_{|\mathcal{Y}_\mathcal{I}|}$, then the terms being multiplied above are estimated probabilities of a subject surviving from time 0 to t_1, from t_1 to t_2, and so forth until reaching time t. The standard Kaplan-Meier estimator has $\mathcal{I} = [n]$.

We now state four nonparametric estimators for the conditional survival function S. The first two are by Beran (1981) and are the estimators that we provide theoretical analysis for in the next section. Distances between feature vectors are measured via a user-specified metric $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$.

k-NN survival estimator. For a test feature vector $x \in \mathcal{X}$, we first find the k training subjects with feature vectors closest to x according to metric ρ, breaking ties uniformly at random. Let $\mathcal{N}_{k\text{-NN}}(x) \subseteq [n]$ denote these k subjects' indices. Then the k-NN estimate for S is $\widehat{S}^{k\text{-NN}}(t|x) := \widehat{S}^{\text{KM}}(t|\mathcal{N}_{k\text{-NN}}(x))$.

Kernel survival estimator. For a user-specified kernel function $K : \mathbb{R}_+ \to \mathbb{R}_+$ and bandwidth $h > 0$, we can measure how similar training subject $j \in [n]$ is to x by the weight $K(\frac{\rho(x, X_j)}{h})$. We generalize the unique death times, death counts, and survivor counts as follows:

$$\mathcal{Y}_K(x; h) := \Big\{Y_j \text{ for } j \in [n] \text{ s.t. } \delta_j K\Big(\frac{\rho(x, X_j)}{h}\Big) > 0\Big\},$$

$$d_K(t|x; h) := \sum_{j=1}^{n} K\Big(\frac{\rho(x, X_j)}{h}\Big)\delta_j \mathbb{1}\{Y_j = t\},$$

$$n_K(t|x; h) := \sum_{j=1}^{n} K\Big(\frac{\rho(x, X_j)}{h}\Big)\mathbb{1}\{Y_j \geq t\}.$$

Then the kernel estimate for S is given by

$$\widehat{S}^K(t|x; h) := \prod_{t' \in \mathcal{Y}_K(x;h)} \Big(1 - \frac{d_K(t'|x; h)}{n_K(t'|x; h)}\Big)^{\mathbb{1}\{t' \leq t\}}. \quad (1)$$

In our numerical experiments later, we benchmark the above methods against the random survival forests method by Ishwaran et al. (2008) along with our proposed variant of it that combines it with the kernel survival estimator.

Random survival forests. Random survival forests are much like standard random forests. During training, each tree is grown using a survival-analysis-based splitting rule. Each leaf is associated with some subset of the training data for which a Kaplan-Meier survival estimate is produced. In other words, for each tree, each leaf is associated with a particular survival function estimate. Then, for a test point x, we find the tree leaves that x belongs to. We average these leaves' survival function estimates to produce the final random survival forest estimate for $S(\cdot|x)$.

Adaptive kernel survival estimator. We propose an alternative approach to making predictions using random survival forests without changing their training procedure. For a test point x, to make a final prediction, we instead use the kernel survival estimator given by equation (1), where we replace the expression $K(\frac{\rho(x,X_j)}{h})$ by $\widehat{K}(x, X_j)$, defined as the fraction of trees for which x and training point X_j show up in the same leaf node in the learned forest. Note that interpreting standard random forests as learning kernels was originally done by Lin & Jeon (2006).

Relating to the Nelson-Aalen estimator. The Nelson-Aalen estimator estimates the marginal cumulative hazard function $H_{\text{marg}}(t) := -\log S_{\text{marg}}(t) = -\log \mathbb{P}(T > t)$ (Nelson, 1969; Aalen, 1978). The Nelson-Aalen estimator restricted to training subjects $\mathcal{I}$ is given by

$$\widehat{H}^{\text{NA}}(t|\mathcal{I}) := \sum_{t' \in \mathcal{Y}_{\mathcal{I}}} \frac{d_{\mathcal{I}}(t')}{n_{\mathcal{I}}(t')} \mathbb{1}\{t' \le t\},$$

using the same variables introduced for the Kaplan-Meier estimator. We can relate the Nelson-Aalen estimator to the Kaplan-Meier one: the first-order Taylor approximation of $-\log \widehat{S}^{\text{KM}}(t|\mathcal{I})$ is $\widehat{H}^{\text{NA}}(t|\mathcal{I})$. Because our theoretical analysis of k-NN and kernel variants of the Kaplan-Meier survival estimator is in terms of Taylor series expansions of $\log S$, our proofs extend (with small changes) to k-NN and kernel variants of the Nelson-Aalen estimator.

For exposition clarity, the rest of the paper uses k-NN and kernel estimators to refer to the Kaplan-Meier versions rather than the Nelson-Aalen ones unless stated otherwise.

3. Theoretical Guarantees

We first introduce some notation used throughout the paper. We denote closed and open balls centered at $x \in \mathcal{X}$ with radius $r > 0$ as

$$\mathcal{B}_{x,r} := \{x' \in \mathcal{X} : \rho(x, x') \le r\},$$
$$\mathcal{B}^o_{x,r} := \{x' \in \mathcal{X} : \rho(x, x') < r\}.$$

We define the "support" of feature distribution $\mathbb{P}_X$ as

$$\text{supp}(\mathbb{P}_X) := \{x \in \mathcal{X} : \mathbb{P}_X(\mathcal{B}_{x,r}) > 0 \text{ for all } r > 0\},$$

where $\mathbb{P}_X(\mathcal{B}_{x,r})$ is the probability that a feature vector sampled from distribution $\mathbb{P}_X$ lands in $\mathcal{B}_{x,r}$.

We denote tail probability functions using "S" with and without subscripts. S without a subscript always refers to the tail of the conditional survival time T distribution $S(t|x) = \mathbb{P}(T > t|X = x)$. The tails of the conditional censoring time C and observed time Y distributions are $S_{\text{C}}(t|x) := \mathbb{P}(C > t|X = x)$ and $S_{\text{Y}}(t|x) := \mathbb{P}(Y > t|X = x)$. PDF's of distributions $\mathbb{P}_{T|X=x}$ and $\mathbb{P}_{C|X=x}$ are denoted by $f_{\text{T}}(t|x)$ and $f_{\text{C}}(t|x)$. Note that $S_{\text{Y}}(t|x) = S(t|x)S_{\text{C}}(t|x)$, $S(t|x) = 1 - \int_0^t f_{\text{T}}(s|x)ds$, and $S_{\text{C}}(t|x) = 1 - \int_0^t f_{\text{C}}(s|x)ds$.

Our guarantees depend on the following four assumptions:

A1. *Feature space $\mathcal{X}$ and distance ρ form a separable metric space, and feature distribution $\mathbb{P}_X$ is a Borel probability measure.* This assumption is technical and ensures that the probability of a feature vector landing in a ball (whether open or closed) is well-defined, and that we only need to care about feature vectors that land in $\text{supp}(\mathbb{P}_X)$ (the probability of a feature vector landing outside of this support is 0). This assumption is also used in establishing consistency of nearest neighbor classification in metric spaces (Cérou & Guyader, 2006; Chaudhuri & Dasgupta, 2014).

A2. *For all $x \in \text{supp}(\mathbb{P}_X)$, distributions $\mathbb{P}_{T|X=x}$ and $\mathbb{P}_{C|X=x}$ exist and correspond to continuous random variables.* This assumption ensures that functions S, S_{C}, S_{Y}, f, and g described above are well-defined. Moreover, continuity here makes ties in observed times Y_i's happen with probability 0.

A3. *There exists $\theta \in (0, \frac{1}{2}]$ and $\tau \in (0, \infty)$ such that*

$$S_{\text{Y}}(\tau|x) \ge \theta \quad \textit{for all } x \in \textit{supp}(\mathbb{P}_X).$$

In practice, we cannot estimate conditional survival function $S(t|x)$ accurately for time t that is arbitrarily large (e.g., $t > \max_{i=1,\dots,n} Y_i$). We shall only guarantee accurate estimation of $S(t|x)$ for $t \in [0, \tau]$.

A4. *For any time $t \in [0, \tau]$, density function $f_{\text{T}}(t|x)$ and $f_{\text{C}}(t|x)$ are Hölder continuous in x with the same exponent $\alpha > 0$ but with potentially different constants $\lambda_{\text{T}} > 0$ and $\lambda_{\text{C}} > 0$, i.e., for all $x, x' \in \textit{supp}(\mathbb{P}_X)$,*

$$|f_{\text{T}}(t|x) - f_{\text{T}}(t|x')| \le \lambda_{\text{T}} \rho(x, x')^\alpha,$$
$$|f_{\text{C}}(t|x) - f_{\text{C}}(t|x')| \le \lambda_{\text{C}} \rho(x, x')^\alpha.$$

This assumption ensures that nearby feature vectors have similar conditional survival and censoring distributions. Thus, feature vectors near x can help us estimate $S(\cdot|x)$.

A wide range of distributions $\mathbb{P}_X$, f_{T}, and f_{C} satisfy the above assumptions. We provide a few examples at the end of this section.

Since $f_{\text{T}}(t|\cdot)$ and $f_{\text{C}}(t|\cdot)$ are Hölder continuous with common exponent α, then so are $S_{\text{Y}}(t|\cdot)$ and $S_{\text{C}}(t|\cdot)f_{\text{T}}(t|\cdot)$, which appear in our analysis. With a bit of algebra, one can show that $S_{\mathrm{Y}}(t|\cdot)$ is Hölder continuous with parameters $(\lambda_{\mathrm{T}} + \lambda_{\mathrm{C}})t$ and α. Meanwhile, $S_{\mathrm{C}}(t|\cdot) f_{\mathrm{T}}(t|\cdot)$ is Hölder continuous with parameters $(\lambda_{\mathrm{T}} + f_{\mathrm{T}}^* \lambda_{\mathrm{C}} t)$ and α, where

$$f_{\mathrm{T}}^* := \sup_{t \in [0,\tau], x \in \mathrm{supp}(\mathbb{P}_X)} f_{\mathrm{T}}(t|x).$$

Our k-NN result depends on the constant

$$\Lambda := \max\Big\{ \frac{2\tau}{\theta}(\lambda_{\mathrm{T}} + \lambda_{\mathrm{C}}),\ \lambda_{\mathrm{T}}\tau + \frac{f_{\mathrm{T}}^* \lambda_{\mathrm{C}} \tau^2}{2} \Big\}.$$

As we explain shortly, the k-NN survival estimator is closely related to two subproblems: k-NN CDF estimation and a special case of k-NN regression. In the definition of Λ above, the two parts of the maximization correspond precisely to the CDF estimation and regression components.

We state each of our main theoretical guarantees as a pointwise result, i.e., for any point $x \in \mathrm{supp}(\mathbb{P}_X)$ and error tolerance $\varepsilon \in (0,1)$, how to guarantee $\sup_{t\in[0,\tau]} |\widehat{S}(t|x) - S(t|x)| \le \varepsilon$ with high probability using estimator $\widehat{S}$. Translating pointwise guarantees to account for randomness in sampling $X = x$ from $\mathbb{P}_X$ can easily be done using standard proof techniques, as we discuss momentarily.

k-NN estimator results

We begin with the nonasymptotic k-NN estimator guarantee. Proofs are deferred to the appendix. As a disclaimer, no serious attempt has been made to optimize constants.

Theorem 3.1 (k-NN pointwise bound)**.** *Under Assumptions A1–A4, let $\varepsilon \in (0,1)$ be a user-specified error tolerance and define critical distance $h^* := (\frac{\varepsilon\theta}{18\Lambda})^{1/\alpha}$. For any feature vector $x \in$ supp$(\mathbb{P}_X)$ and any choice of number of nearest neighbors $k \in [\frac{72}{\varepsilon\theta^2}, \frac{n\mathbb{P}_X(\mathcal{B}_{x,h^*})}{2}]$, we have, over randomness in training data,*

$$\begin{aligned}\mathbb{P}\Big(&\sup_{t\in[0,\tau]} |\widehat{S}^{k\text{-NN}}(t|x) - S(t|x)| > \varepsilon \Big) \\ &\le \exp\Big(-\frac{k\theta}{8}\Big) + \exp\Big(-\frac{n\mathbb{P}_X(\mathcal{B}_{x,h^*})}{8}\Big) \\ &\quad + 2\exp\Big(-\frac{k\varepsilon^2\theta^4}{648}\Big) + \frac{8}{\varepsilon}\exp\Big(-\frac{k\varepsilon^2\theta^2}{162}\Big). \qquad (2)\end{aligned}$$

The four terms in the above bound correspond to penalties for the following bad events:

1. Too few of the k nearest neighbors survive beyond time τ (in the worst case, none do, so from the data alone, we would suspect Assumption A3 to not hold)
2. The k nearest neighbors are not all within critical distance h^* of x (by Assumption A4, the nearest neighbors should be close to x to guarantee that they provide accurate information about $S(\cdot|x)$)
3. The number of nearest neighbors k is too small such that when we form an empirical distribution using their Y_i values, this empirical distribution has not converged to its expectation, which is a CDF (note that when the previous bad event does not happen, then this CDF is approximately $1 - S_{\mathrm{Y}}(\cdot|x)$)
4. The k-NN survival estimator can be viewed as solving a specific k-NN regression problem, which averages over the k nearest neighbors' "labels" (if X_i is one of the k nearest neighbors of x, then its label is taken to be $-\frac{\delta_i \mathbb{1}\{Y_i \le t\}}{S_{\mathrm{Y}}(Y_i|x)}$, i.e., this label depends on an accurate estimate for $S_{\mathrm{Y}}(\cdot|x)$, which the previous bad event is about). This last bad event is that the average of these k labels is not close to its expectation due to k being too small.

In our analysis, preventing bad event #1 is pivotal to upper-bounding the k-NN survival estimator's error by those of the k-NN CDF estimation and k-NN regression problems. Subsequently, bad event #2 is about controlling the bias of these k-NN CDF and k-NN regression estimators, i.e., making sure their expectations are close to desired target values. Bad events #3 and #4 relate to controlling the variances of these k-NN CDF and k-NN regression estimates.

The observation that CDF estimation and regression subproblems arise is based on nonasymptotic analysis of the standard Kaplan-Meier estimator by Földes & Rejtö (1981). For controlling the bias and variance of k-NN CDF and k-NN regression estimators, we use proof techniques by Chaudhuri & Dasgupta (2014).

To understand the consequences of Theorem 3.1, especially how it relates to the rate of convergence for the k-NN survival estimator, we examine sufficient conditions for which the RHS of bound (2) is at most a user-specified error probability $\gamma \in (0,1)$. To achieve this, we can ask that each of the four terms be bounded above by $\gamma/4$. In doing so, a simple calculation reveals that the theorem's conditions on k and n are met if

$$k \ge \frac{648}{\varepsilon^2\theta^4}\log\frac{32}{\varepsilon\gamma}, \quad \text{and} \quad n \ge \frac{2k}{\mathbb{P}_X(\mathcal{B}_{x,h^*})}. \qquad (3)$$

This pointwise guarantee highlights a key feature of nearest neighbor methods in that they depend on the *intrinsic* dimension of the data (Kpotufe, 2011; Kpotufe & Garg, 2013). For example, consider when the feature space is $\mathcal{X} = \mathbb{R}^d$. Even though the data have *extrinsic* dimension d, it could be that $\mathbb{P}_X(\mathcal{B}_{x,h^*})$ scales as $(h^*)^{d'}$ for some $d' < d$. This could happen if the data reside in a low dimensional portion of the higher dimensional space (e.g., supp$(\mathbb{P}_X)$ is a convex polytope of $d' < d$ dimensions within $\mathbb{R}^d$). Thus, examining the second inequality of (3), the number of training data n sufficient for guaranteeing a low error in estimating $S(\cdot|x)$ scales exponentially in the intrinsic dimension at x (roughly, the smallest $d' > 0$ for which $\mathbb{P}_X(\mathcal{B}_{x,r}) \sim r^{d'}$ for all small enough r).

Sufficient conditions (3) also tell us when we can consistently estimate $S(\cdot|x)$ for a fixed x. Specifically for any error tolerance $\varepsilon > 0$, to have the error probability γ go to 0, the condition on k suggests that we take $k \to \infty$, which also means that $n \to \infty$. At the same time, the condition relating n and k says that we should have $k/n \le \mathbb{P}_X(\mathcal{B}_{x,h^*})/2$. Recall that $h^* = (\frac{\varepsilon\theta}{18\Lambda})^{1/\alpha}$, so if we pick ε to be arbitrarily small, then $\mathbb{P}_X(\mathcal{B}_{x,h^*}) \to 0$, so we want $k/n \to 0$. We remark that choosing k as a function of n to satisfy $k \to \infty$ and $k/n \to 0$ are the usual conditions on k for k-NN classification and regression to be weakly consistent (Cover & Hart, 1967; Stone, 1977).

As for how k should scale with n, this depends on $\mathbb{P}_X(\mathcal{B}_{x,h^*})$. For example, if $\mathbb{P}_X(\mathcal{B}_{x,h^*}) \sim (h^*)^d$, then the second inequality of sufficient conditions (3) says that k should scale at most as $(h^*)^d n \sim \varepsilon^{d/\alpha} n$. In this case, our next result shows that the k-NN estimator is strongly consistent. Since h^* is a function of ε, which we now take to go to 0, formally we shall assume that $\mathbb{P}_X(\mathcal{B}_{x,r}) \ge p_{\min} r^d$ for all $r \in (0, r^*]$ for some positive constants $p_{\min}$, d, and r^*. Thus, as we shrink ε toward 0, once ε becomes small enough (namely $\varepsilon \le \frac{18\Lambda(r^*)^\alpha}{\theta}$), then $h^* = (\frac{\varepsilon\theta}{18\Lambda})^{1/\alpha} \in (0, r^*]$ and so $\mathbb{P}_X(\mathcal{B}_{x,h^*}) \ge p_{\min}(h^*)^d$.

Corollary 3.1 (k-NN strong consistency rate)**.** *Under Assumptions A1–A4, let $x \in$ supp$(\mathbb{P}_X)$, and suppose that there exist constants $p_{\min} > 0$, $d > 0$, and $r^* > 0$ such that $\mathbb{P}_X(\mathcal{B}_{x,r}) \ge p_{\min} r^d$ for all $r \in (0, r^*]$. Then there are positive numbers $c_1 = \Theta\big(\frac{1}{(\theta\Lambda)^{2d/(2\alpha+d)}}\big)$, $c_2 = \Theta\big(\frac{\theta^{(4\alpha+d)/(5\alpha+2d)}}{\Lambda^{d/(5\alpha+2d)}}\big)$, and $c_3 = \Theta\big(\frac{\Lambda^{d/(2\alpha+d)}}{\theta^{(4\alpha+d)/(2\alpha+d)}}\big)$ such that by choosing the number of nearest neighbors to be $k_n := \lfloor c_1 n^{2\alpha/(2\alpha+d)}\big(\log(c_2 n)\big)^{d/(2\alpha+d)}\rfloor$, with probability 1,*

$$\limsup_{n\to\infty}\left\{\frac{\sup_{t\in[0,\tau]}|\widehat{S}^{k_n\text{-NN}}(t|x) - S(t|x)|}{c_3\big(\frac{\log(c_2 n)}{n}\big)^{\alpha/(2\alpha+d)}}\right\} < 1.$$

The above corollary follows from setting error probability $\gamma = 1/n^2$ in sufficient conditions (3), solving the inequalities in the sufficient conditions for ε, n, and k (and thus finding coefficients c_1, c_2, and c_3 above), and finally applying the Borel-Cantelli lemma. Closed-form equations for c_1, c_2, and c_3 are in Appendix D.

Near-optimality. Our nonasymptotic bound (2) turns out to essentially be optimal. Consider when the censoring times always occur after the survival times, i.e., nothing is censored. Then the problem reduces to conditional CDF estimation ($1 - S(\cdot|x)$ is a conditional CDF), for which the minimax lower bound for *expected squared error* under slightly more assumptions than we impose is $n^{-2\alpha/(2\alpha+d)}$ (Chagny & Roche, 2014, Theorem 3). Our result implies an upper bound on the expected squared error. First, note that

$$\mathbb{E}\Big[\int_0^\tau (\widehat{S}^{k_n\text{-NN}}(t|x) - S(t|x))^2 dt\Big] \le \tau\mathbb{E}\Big[\sup_{t\in[0,\tau]}|\widehat{S}^{k_n\text{-NN}}(t|x) - S(t|x)|^2\Big]. \qquad (4)$$

Next, sufficient conditions (3) say that with probability at least $1-\gamma$, none of the bad events happen so $\sup_{t\in[0,\tau]}|\widehat{S}^{k_n\text{-NN}}(t|x) - S(t|x)| \le \varepsilon$ (for which we can square both sides and bring the square into the supremum); otherwise the supremum norm error is at worst 1. Hence,

$$\mathbb{E}\Big[\sup_{t\in[0,\tau]}|\widehat{S}^{k_n\text{-NN}}(t|x) - S(t|x)|^2\Big] \le \varepsilon^2\cdot 1 + 1\cdot\gamma, \quad (5)$$

where on the RHS, the first term is the worst-case squared supremum norm error ε^2 when none of the bad events happen (this happens with probability at least $1-\gamma \le 1$), and the second term is the worst-case squared supremum norm error of 1 (this happens with probability at most γ).

It suffices to set $\gamma = \varepsilon^2$ and find precise conditions on k, n, and ε so that sufficient conditions (3) hold (the calculation is similar to the one for deriving Corollary 3.1). By doing this calculation and combining inequalities (4) and (5), we get that the k-NN survival estimator has expected squared error $\widetilde{\mathcal{O}}(n^{-2\alpha/(2\alpha+d)})$, even if there is right-censoring.

Results for random test feature vectors. As there are a number of standard approaches for translating pointwise guarantees to ones accounting for randomness in sampling $X = x \sim \mathbb{P}_X$, we only focus on one such technique and briefly mention some others. Specifically, we consider a simple approach in which we partition the feature space $\mathcal{X}$ into a "good" region $\mathcal{X}_{\text{good}}$ with sizable probability mass (where many training data are likely to be), and a bad region $\mathcal{X}_{\text{bad}}$ where we tolerate error (where there are likely to be too few training data). Using the same idea as described in Section 3.3.1 of Chen & Shah (2018), we define the *sufficient mass region* as

$$\mathcal{X}_{\text{good}}(\mathbb{P}_X; p_{\min}, d, r^*) := \{x \in \text{supp}(\mathbb{P}_X) : \mathbb{P}_X(\mathcal{B}_{x,r}) \ge p_{\min} r^d \;\; \forall r \in (0, r^*]\},$$

and $\mathcal{X}_{\text{bad}}(\mathbb{P}_X; p_{\min}, d, r^*) = \mathcal{X} \setminus \mathcal{X}_{\text{good}}(\mathbb{P}_X; p_{\min}, d, r^*)$. The sufficient mass region for feature distribution $\mathbb{P}_X$ corresponds to portions of supp$(\mathbb{P}_X)$ that behave like they have dimension d. Returning to the previous example, if $\mathcal{X} = \mathbb{R}^d$ and supp$(\mathbb{P}_X)$ is a full-dimensional convex polytope, then there exists a $p_{\min} > 0$ and $r^* > 0$ such that $\mathcal{X}_{\text{good}}(\mathbb{P}_X; p_{\min}, d, r^*) = \text{supp}(\mathbb{P}_X)$.

In general, when feature vector $X \sim \mathbb{P}_X$ lands in $\mathcal{X}_{\text{good}}(\mathbb{P}_X; p_{\min}, d, h^*)$, then the conditions of Theorem 3.1 are satisfied and, moreover, $\mathbb{P}_X(\mathcal{B}_{x,h^*}) \ge p_{\min}(h^*)^d$. We readily obtain the following corollary.

Corollary 3.2 (k-NN bound for random test point)**.** *Under the same conditions as Theorem 3.1 except now sampling test point $X \sim \mathbb{P}_X$, then over randomness in the training data and X,*

$$\begin{aligned}\mathbb{P}\Big(\sup_{t\in[0,\tau]}|\widehat{S}^{k\text{-NN}}(t|X) - S(t|X)| > \varepsilon\Big) &\le \exp\Big(-\frac{k\theta}{8}\Big) + \exp\Big(-\frac{np_{\min}(h^*)^d}{8}\Big)\\ &\quad + 2\exp\Big(-\frac{k\varepsilon^2\theta^4}{648}\Big) + \frac{8}{\varepsilon}\exp\Big(-\frac{k\varepsilon^2\theta^2}{162}\Big)\\ &\quad + \mathbb{P}_X\big(\mathcal{X}_{bad}(\mathbb{P}_X; p_{\min}, d, h^*)\big).\end{aligned}$$

Thus, if there exists $p_{\min} > 0$, $d > 0$, and $r^* > 0$ such that $\mathcal{X}_{\text{good}}(\mathbb{P}_X; p_{\min}, d, r^*) = \text{supp}(\mathbb{P}_X)$, then strong consistency of $\widehat{S}^{k\text{-NN}}(\cdot|X)$ at the rate of Corollary 3.1 holds over randomness in training data and $X \sim \mathbb{P}_X$.

Other approaches are possible to obtain guarantees over randomness in both training data and X from guarantees for fixed $X = x$. For example, there are notions similar to the sufficient mass region specific to Euclidean space such as the *strong minimal mass assumption* of Gadat et al. (2016) and the *strong density assumption* of Audibert & Tsybakov (2007). An alternative strategy that stays in separable metric spaces is to use covering numbers from metric entropy. For details, see Section 3.3.3 of Chen & Shah (2018).

Kernel estimator results

Our kernel result uses an additional decay assumption:

A5. *The kernel function K monotonically decreases, and there exists a standardized distance $\phi > 0$ such that $K(s) > 0$ for all $s \in [0, \phi]$ and $K(s) = 0$ for $s > \phi$.* This assumption ensures that training data sufficiently far from x have no impact on our estimation of $S(\cdot|x)$. (Small proof changes can be made to allow $K(\phi) = 0$, e.g., to handle triangle and Epanechnikov kernels.)

Our kernel result depends on the kernel function's maximal and minimal positive values, namely $K(0)$ and $K(\phi)$. We let $\kappa := K(\phi)/K(0)$, and define

$$\Lambda_K := \max\Big\{\frac{2\tau}{\theta\kappa}(\lambda_{\text{T}} + \lambda_{\text{C}}),\ \lambda_{\text{T}}\tau + \frac{f_{\text{T}}^* \lambda_{\text{C}} \tau^2}{2}\Big\}.$$

The first term in the maximization (related to CDF estimation) has an extra $1/\kappa$ factor compared to Λ.

As our kernel survival estimator guarantee is similar to that of the k-NN estimator, we only present its pointwise version. Deriving a corresponding strong consistency rate, accounting for randomness in sampling $X \sim \mathbb{P}_X$, and showing near-optimality can be done as before. In particular, the two methods have similar asymptotic behavior.

Theorem 3.2 (Kernel pointwise guarantee)**.** *Under Assumptions A1–A5, let $\varepsilon \in (0, 1)$ be a user-specified error tolerance. Suppose that the threshold distance satisfies $h \in (0, \frac{1}{\phi}(\frac{\varepsilon\theta}{18\Lambda_K})^{1/\alpha}]$, and the number of training data satisfies $n \geq \frac{144}{\varepsilon\theta^2\mathbb{P}_X(\mathcal{B}_{x,\phi h})\kappa}$. For any $x \in$ supp($\mathbb{P}_X$),*

$$\begin{aligned}\mathbb{P}\Big(&\sup_{t\in[0,\tau]} |\widehat{S}^K(t|x; h) - S(t|x)| > \varepsilon\Big) \\ &\leq \exp\Big(-\frac{n\mathbb{P}_X(\mathcal{B}_{x,\phi h})\theta}{16}\Big) + \exp\Big(-\frac{n\mathbb{P}_X(\mathcal{B}_{x,\phi h})}{8}\Big) \\ &\quad + \frac{216}{\varepsilon\theta^2\kappa}\exp\Big(-\frac{n\mathbb{P}_X(\mathcal{B}_{x,\phi h})\varepsilon^2\theta^4\kappa^4}{11664}\Big) \\ &\quad + \frac{8}{\varepsilon}\exp\Big(-\frac{n\mathbb{P}_X(\mathcal{B}_{x,\phi h})\varepsilon^2\theta^2\kappa^2}{324}\Big). \qquad (6)\end{aligned}$$

As with the k-NN analysis, the kernel estimator analysis involves two subproblems, a kernel CDF estimation (i.e., using weighted samples to construct an empirical distribution function) and a kernel regression. We remark that k-NN CDF estimation is straightforward to analyze because the different data points have equal weight, so we can apply the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality. To handle weighted empirical distributions, we establish the following nonasymptotic bound.

Proposition 3.1 (Weighted empirical distribution inequality)**.** *Let real-valued random variables $Z_1, \dots, Z_\ell$ be i.i.d. samples drawn from a continuous CDF F. Let $w_1, \dots, w_\ell$ be any sequence of nonnegative constants such that $\sum_{i=1}^{\ell} w_i > 0$. Consider the following weighted empirical distribution function:*

$$\widehat{F}(t) := \sum_{i=1}^{\ell} \frac{w_i}{\sum_{j=1}^{\ell} w_j} \mathbb{1}\{Z_i \leq t\} \quad \textit{for } t \in \mathbb{R}.$$

For every $\varepsilon \in (0, 1]$,

$$\mathbb{P}\Big(\sup_{t\in\mathbb{R}} |\widehat{F}(t) - F(t)| > \varepsilon\Big) \leq \frac{6}{\varepsilon}\exp\Big(-\frac{2\varepsilon^2(\sum_{j=1}^{\ell} w_j)^2}{9\sum_{i=1}^{\ell} w_i^2}\Big).$$

Box kernel, weighted k-NN. If instead the kernel survival estimator is used with a box kernel (uniform weights), then we can use the DKW inequality instead of Proposition 3.1, leading to a slightly stronger pointwise guarantee (Theorem A.1 in the appendix). We remark that proof ideas for our k-NN and kernel survival estimators can be combined to derive results for weighted k-NN survival estimators.

Choosing k and h via a validation set. Our main results choose k and h in a way that depends on unknown model parameters. In practice, validation data could be used to select k and h via minimizing the IPEC score (Gerds & Schumacher, 2006; Lowsky et al., 2013). We obtain a nonasymptotic guarantee for a slight variant of the validation strategy by Lowsky et al. (2013) in Appendix H. The high-level proof idea is simple. For example, for the k-NN estimator $\widehat{S}^{k\text{-NN}}$, suppose we have an independent validation set of size n. Provided that the choices of k that the user sweeps over for validation include one good choice according to Theorem 3.1, then for large enough n, estimator $\widehat{S}^{k\text{-NN}}$ has a validation error that approaches that of S. Our proof is a bit nuanced and requires controlling both additive and multiplicative error in tail probability estimates, using our analysis for Nelson-Aalen-based nearest neighbor and kernel estimators (given in Appendix J).

Distributions satisfying Assumptions A1–A4

We now provide example models that satisfy Assumptions A1–A4. In these examples, the feature space $\mathcal{X}$ and distance ρ are Euclidean, and the Hölder exponent is $\alpha = 1$ (so λ_{T} and λ_{C} are Lipschitz constants).

Example 3.1 (Exponential regression)**.** *Let $\mathcal{X} = \mathbb{R}^d$, and $\mathbb{P}_X$ be any Borel probability measure with compact, con-*

vex support (so Assumption A1 is met). We define conditional survival function $S(t|x)$ *using the hazard function* $h_{\text{T}}(t|x) = -\frac{\partial}{\partial t}\log S(t|x) = h_{\text{T},0}\exp(x^\top\beta_{\text{T}})$ *with parameters* $h_{\text{T},0} > 0$ *and* $\beta_{\text{T}} \in \mathbb{R}^d$*. Then*

$$S(t|x) = \exp\Big(-\int_0^t h_{\text{T},0}\exp(x^\top\beta_{\text{T}})ds\Big)$$
$$= \exp(-h_{\text{T},0}e^{x^\top\beta_{\text{T}}}t),$$

which implies that the distribution $\mathbb{P}_{T|X=x}$ *(which has CDF* $1 - S(\cdot|x)$*) is exponentially distributed with parameter* $h_{\text{T},0}e^{x^\top\beta_{\text{T}}}$*. We could similarly define the censoring time conditional distribution through the hazard function* $h_{\text{C}}(t|x) = h_{\text{C}}\exp(x^\top\beta_{\text{C}})$*, with* $h_{\text{C},0} > 0$ *and* $\beta_{\text{C}} \in \mathbb{R}^d$*. In this case, distribution* $\mathbb{P}_{C|X=x}$ *is exponentially distributed with parameter* $h_{\text{C},0}e^{x^\top\beta_{\text{C}}}$*. At this point, Assumption A2 is also met since for any* $x \in$ *supp*$(\mathbb{P}_X)$*, distributions* $\mathbb{P}_{T|X=x}$ *and* $\mathbb{P}_{C|X=x}$ *correspond to continuous random variables.*

We now present valid choices for θ *and* τ *for Assumption A3. Recall that the observed time is* $Y = \min\{T, C\}$*. Conditioned on* $X = x$*, the minimum of independent exponential random variables is exponential. In particular, distribution* $\mathbb{P}_{Y|X=x}$ *is exponentially distributed with parameter* $\omega(x) := h_{\text{T},0}e^{x^\top\beta_{\text{T}}} + h_{\text{C},0}e^{x^\top\beta_{\text{C}}}$*. Thus, if we pick* $\theta = 1/2$*, then a valid choice for* τ *would be the smallest possible median of distribution* $\mathbb{P}_{Y|X=x}$ *across all* $x \in$ *supp*$(\mathbb{P}_X)$*. Note that the median of* $\mathbb{P}_{Y|X=x}$ *is* $(\log 2)/\omega(x)$*. Thus, we can pick* $\tau = \min_{x\in \textit{supp}(\mathbb{P}_X)}\{(\log 2)/\omega(x)\}$*.*

Lastly, for Assumption A4, due to supp$(\mathbb{P}_X)$ *being compact and convex, the conditional survival time density* $f_{\text{T}}(t|\cdot)$ *has finite Lipschitz constant*

$$\lambda_{\text{T}} = \sup_{x\in \textit{supp}(\mathbb{P}_X), t\in[0,\tau]} \Big\|\frac{\partial f_{\text{T}}(t|x)}{\partial x}\Big\|_2,$$

where $\|\cdot\|_2$ *is Euclidean norm, and* $\frac{\partial f_{\text{T}}(t|x)}{\partial x} = f_{\text{T}}(t|x)(1 - h_{\text{T},0}e^{x^\top\beta_{\text{T}}}t)\beta_{\text{T}}$*. We could similarly choose Lipschitz constant* λ_{C} *for the conditional censoring time density* $f_{\text{C}}(t|\cdot)$*.*

This exponential regression example can easily be generalized to Weibull regression, which is another proportional hazards model (see Appendix I).

Example 3.2 (Weibull mixture)**.** *To give an example that is not a proportional hazard model that satisfies Assumptions A1–A4, consider an integer-valued one-dimensional feature vector* $X \sim$ *Uniform*$\{1, 2, \ldots, 100\}$*. For a threshold* $\nu \in (1, 100)$*, if* $X \le \nu$*, then we sample survival time* T *from a Weibull distribution with shape parameter* $q > 0$ *and scale parameter* $\psi_{\text{T},1} > 0$*. Otherwise if* $X > \nu$*, then we sample* T *from a Weibull distribution still with shape parameter* q *but a different scale parameter* $\psi_{\text{T},2} > 0$*. Thus, the marginal distribution of* T *is a mixture of two Weibull distributions. We similarly define the censoring time* C *to be a mixture of two Weibull distributions with common shape parameter* q *and different scale parameters* $\psi_{\text{C},1} > 0$ *and* $\psi_{\text{C},2} > 0$*; we sample* C *from the first component using the same threshold* ν *as before, i.e., when* $X \le \nu$*.*

Dataset	Description	# subjects	# dim.
PBC	primary biliary cirrhosis	276	17
GBSG2	breast cancer	686	8
RECID	recidivism	1445	14
KIDNEY	dialysis	1044	53

Table 1. Characteristics of the survival datasets used.

Conditioned on X*, the distribution of observed time* $Y = \min\{T, C\}$ *is now one of two possible Weibull distributions (the minimum of independent Weibull distributions with shape parameter* q *is still Weibull with shape* q*): if* $X \le \nu$*, then* Y *is Weibull with shape* q *and scale* $(\psi_{\text{T},1}^{-q} + \psi_{\text{C},1}^{-q})^{-1/q}$*. Otherwise* Y *is Weibull with shape* q *and scale* $(\psi_{\text{T},2}^{-q} + \psi_{\text{C},2}^{-q})^{-1/q}$*. For Assumption A3, we can choose* $\theta = 1/2$ *and* τ *to be the smaller median of the two possible Weibull distributions for* Y*, i.e.,* $\tau = \big[\min\big\{\frac{1}{\psi_{\text{T},1}^{-q}+\psi_{\text{C},1}^{-q}}, \frac{1}{\psi_{\text{T},2}^{-q}+\psi_{\text{C},2}^{-q}}\big\}\log 2\big]^{1/q}$*. Lastly, for Assumption A4, since* $|\textit{supp}(\mathbb{P}_X)|$ *is finite, we can set the Lipschitz constant* λ_{T} *to be*

$$\lambda_{\text{T}} = \sup_{x,x'\in\{1,2,\ldots,100\} \textit{ s.t. } x\neq x', t\in[0,\tau]} \frac{|f_{\text{T}}(t|x) - f_{\text{T}}(t|x')|}{|x - x'|}.$$

Lipschitz constant λ_{C} *can be chosen similarly.*

4. Experimental Results

We benchmark the four nonparametric estimators stated in Section 2 against two baselines: the Cox proportional hazards model (Cox, 1972), and a second baseline that explicitly solves the k-NN CDF estimation and k-NN regression subproblems (in succession) that arise in the theoretical analysis for the k-NN survival estimator (we refer to this method as CDF-REG; for simplicity we only consider the k-NN variant and not the kernel variant). According to our theory, the k-NN survival estimator's error should be upper-bounded by that of CDF-REG. For the k-NN, CDF-REG, and kernel methods, we standardize features and use ℓ_2 and ℓ_1 distances. For the k-NN and CDF-REG methods, we also consider their weighted versions using a triangle kernel.[*] For the kernel method, we use box and triangle kernels. We also have results for more kernel choices in Appendix K (the Epanechnikov kernel performs as well as the triangle kernel, and truncated Gaussian kernels tend to perform poorly).

We run the above methods on four datasets. Three are publicly available: the Mayo Clinic primary biliary cirrhosis dataset (abbreviated PBC) (Fleming & Harrington, 1991), the German Breast Cancer Study Group 2 dataset (GBSG2) (Schumacher et al., 1994), and the recidivism

[*] Let $X_{(i)}$ denote the i-th nearest neighbor of test point x. Then weighted k-NN assigns $X_{(i)}$ to have weight $K\big(\frac{\rho(x,X_{(i)})}{\rho(x,X_{(k)})}\big)$.

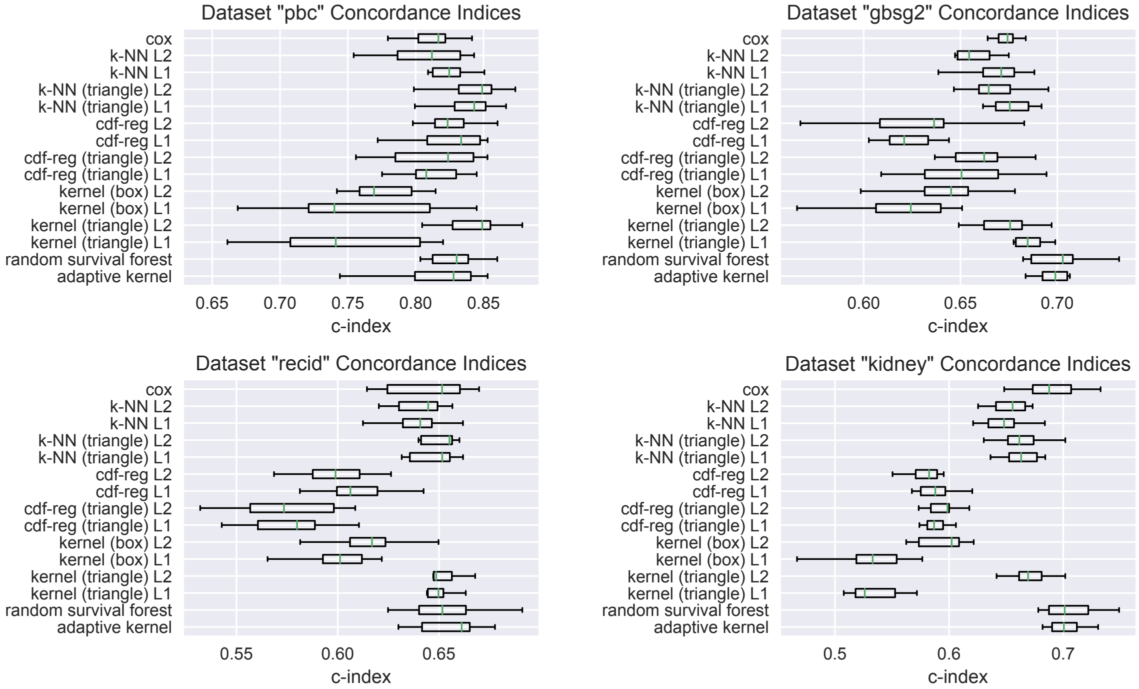

Figure 1. Survival analysis prediction results on four datasets using the concordance index (c-index; higher means more accurate prediction). Each dataset is randomly split into 10 train/test splits, resulting in the different c-index scores per method.

dataset (RECID) from Chung et al. (1991). The fourth dataset we use is from a study on dialysis patients (KIDNEY) by Ganssauge et al. (2016). For PBC, GBSG2, and KIDNEY, the survival time refers to time until death whereas for RECID, the “survival time” refers to time until a convicted criminal reoffends. The dataset sizes and number of features are reported in Table 1. In all cases, subjects with any missing features are removed. For the KIDNEY dataset, features with too many missing entries are also removed.

For each dataset, the basic experiment we run is as follows. We randomly divide the dataset into a 70%/30% train/test split. Using the training portion, for all methods except Cox proportional hazards, we run 5-fold cross-validation to select algorithm parameters before training on the full training set and predicting on the test set; prediction error is measured using the standard survival analysis accuracy metric of concordance index (c-index) (Harrell Jr et al., 1982) (details on c-index calculation and the parameter grids used are in Appendix K). This basic experiment is repeated 10 times, so that every dataset gets randomly divided into train/test sets 10 different ways. Results are shown in Figure 1.

We find that random survival forests and the adaptive kernel method (with a kernel learned using random survival forests) tend to have similar performance per dataset. These two methods have the best performance in the GBSG2, RECID, and KIDNEY datasets. However, on the smallest dataset considered (PBC with 276 subjects), while random survival forests and the adaptive kernel method outperform nearly all the other methods, their concordance indices are noticeably lower than those of the weighted k-NN and kernel survival estimators (both using triangle kernels). Separately, we find that the k-NN survival estimator always outperforms its corresponding CDF-REG variant. This agrees with our theory in which the k-NN estimator's error is upper-bounded by that of CDF-REG.

5. Conclusions

By combining contemporary metric-space-based nearest neighbor theory by Chaudhuri & Dasgupta (2014) with the classic Kaplan-Meier analysis of Földes & Rejtö (1981), we have established new guarantees for nearest neighbor and kernel variants of Kaplan-Meier and Nelson-Aalen estimators. We suspect that other recent theoretical developments in nearest neighbor and kernel methods also carry over to the survival analysis setting, such as adaptive methods for choosing the number of nearest neighbors k or kernel bandwidth h (Goldenshluger & Lepski, 2011; Kpotufe, 2011; Goldenshluger & Lepski, 2013; Kpotufe & Garg, 2013; Anava & Levy, 2016), and error bounds that are uniform over test feature vectors rather than only over a randomly chosen test vector (Kpotufe, 2011; Kpotufe & Garg, 2013). However, these developments do not explain the success of random survival forests and the proposed adaptive kernel variant. When and why do these nonparametric survival estimators work well, and how does their theory differ from that of standard random forests for regression and classification? Are there better ways of learning a kernel for use with kernel survival estimation? These questions outline promising directions for future exploration.

Acknowledgments

The author thanks Donald K.K. Lee for extremely helpful discussions.

References

Aalen, O. O. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pp. 701–726, 1978.

Aalen, O. O. A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8):907–925, 1989.

Anava, O. and Levy, K. Y. k^*-nearest neighbors: from global to local. In *Advances in Neural Information Processing Systems*, pp. 4916–4924, 2016.

Audibert, J.-Y. and Tsybakov, A. B. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2): 608–633, 2007.

Beran, R. Nonparametric regression with randomly censored survival data. *Technical report, University of California, Berkeley*, 1981.

Botsis, T., Hartvigsen, G., Chen, F., and Weng, C. Secondary use of EHR: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010.

Cérou, F. and Guyader, A. Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, 10:340–355, 2006.

Chagny, G. and Roche, A. Adaptive and minimax estimation of the cumulative distribution function given a functional covariate. *Electronic Journal of Statistics*, 8(2):2352–2404, 2014.

Chatzigeorgiou, I. Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17(8):1505–1508, 2013.

Chaudhuri, K. and Dasgupta, S. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pp. 3437–3445, 2014. We use the numbering of lemmas from arXiv: 1407.0067v2 [cs.LG].

Chen, G. H. and Shah, D. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends® in Machine Learning*, 10(5-6):337–588, 2018.

Chung, C.-F., Schmidt, P., and Witte, A. D. Survival analysis: A survey. *Journal of Quantitative Criminology*, 7(1): 59–98, 1991.

Cover, T. M. and Hart, P. E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13 (1):21–27, 1967.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):87–22, 1972.

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

Fleming, T. R. and Harrington, D. P. *Counting Processes and Survival Analysis*. John Wiley & Sons, 1991.

Földes, A. and Rejtö, L. Strong uniform consistency for nonparametric survival curve estimators from randomly censored data. *The Annals of Statistics*, pp. 122–129, 1981.

Gadat, S., Klein, T., and Marteau, C. Classification in general finite dimensional spaces with the k-nearest neighbor uule. *The Annals of Statistics*, 44(3):982–1009, 2016.

Ganssauge, M., Padman, R., Teredesai, P., and Karambelkar, A. Exploring dynamic risk prediction for dialysis patients. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2016.

Gerds, T. A. and Schumacher, M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6): 1029–1040, 2006.

Goldenshluger, A. and Lepski, O. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.

Goldenshluger, A. and Lepski, O. General selection rule from a family of linear estimators. *Theory of Probability & Its Applications*, 57(2):209–226, 2013.

Harrell Jr, F. E., Califf, R. M., Pryor, D. B., et al. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546, 1982.

Hoorfar, A. and Hassani, M. Inequalities on the Lambert W function and hyperpower function. *Journal of Inequalities in Pure and Applied Mathematics*, 9(2):5–9, 2008.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data (2nd ed.)*. John Wiley & Sons, 2002.

Kaplan, E. L. and Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

Kpotufe, S. k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pp. 729–737, 2011.

Kpotufe, S. and Garg, V. K. Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems*, pp. 3075–3083, 2013.

Lin, Y. and Jeon, Y. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

Lowsky, D. J., Ding, Y., Lee, D. K., McCulloch, C. E., Ross, L. F., Thistlethwaite, J. R., and Zenios, S. A. A K-nearest neighbors survival probability prediction method. *Statistics in Medicine*, 32(12):2062–2069, 2013.

Massart, P. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pp. 1269–1283, 1990.

Nelson, W. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52, 1969.

Ranganath, R., Perotte, A., Elhadad, N., and Blei, D. Deep survival analysis. In *Machine Learning for Healthcare*, 2016.

Schumacher, M., Bastert, G., Bojar, H., Huebner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R., and Rauschecker, H. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994.

Stone, C. J. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.

Van Keilegom, I. *Nonparametric estimation of the conditional distribution in regression with censored data*. PhD thesis, UHasselt Diepenbeek, 1998.

Van Keilegom, I. and Veraverbeke, N. Uniform strong convergence results for the conditional Kaplan-Meier estimator and its quantiles. *Communications in Statistics–Theory and Methods*, 25(10):2251–2265, 1996.

Zupan, B., Demšar, J., Kattan, M. W., Beck, J. R., and Bratko, I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine*, 20(1):59–75, 2000.