A kernel log-rank test of independence for right-censored data

Abstract

We introduce a general non-parametric independence test between right-censored survival times and covariates, which may be multivariate. Our test statistic has a dual interpretation, first in terms of the supremum of a potentially infinite collection of weight-indexed log-rank tests, with weight functions belonging to a reproducing kernel Hilbert space (RKHS) of functions; and second, as the norm of the difference of embeddings of certain finite measures into the RKHS, similar to the Hilbert-Schmidt Independence Criterion (HSIC) test-statistic. We study the asymptotic properties of the test, finding sufficient conditions to ensure our test correctly rejects the null hypothesis under any alternative. The test statistic can be computed straightforwardly, and the rejection threshold is obtained via an asymptotically consistent Wild Bootstrap procedure. Extensive investigations on both simulated and real data suggest that our testing procedure generally performs better than competing approaches in detecting complex non-linear dependence.

Keywords: right-censoring, independence testing, log-rank test, reproducing kernel Hilbert space.

1 Introduction

Right-censored data appear in survival analysis and reliability theory, where the time-to-event variable one is interested in modelling may not be observed fully, but only in terms of a lower bound. This is a common occurrence in clinical trials as, usually, the follow-up is restricted to the duration of the study, or patients may decide to withdraw from the study.

An important task when dealing with such data is to test independence between the survival times and the covariates. For instance, in a clinical trial setting, we may wish to test if the survival times differ across treatments, e.g., chemotherapy vs radiation; or if there is dependence on other covariates, such as ages of the patients, gender, or any other measured variables. The main challenge of testing independence in this setting is that we need to deal with censored observations, where the censoring mechanism may be dependent on the covariates while the time of interest may not. E.g., patients' withdrawal times from a study can be associated to their gender even if gender is independent of the survival time.

The problem of testing independence has been widely studied by the statistical community. In the context of right-censored data, this problem has often been addressed through the mechanism of two-sample tests, in which the covariate takes one out of two possible values. For the two-sample problem, the main tool is the log-rank test [22, 26] and its generalizations, namely weighted log-rank tests [17, 3, 10]. The more general case, in which the covariates belong to $\mathbb{R}^d$, is much more challenging, and most of the current approaches are ad-hoc for specific semi-parametric models, e.g. [6, 13, 34]. In particular, the most popular of these approaches is the Cox proportional hazards model [6], which assumes a linear effect of the covariates on the log-hazard function. Non-parametric approaches are more scarce, however

[20, 23, 27]. In [20], a nonparametric test for independence is obtained by measuring monotonic relationships between a censored survival time and an ordinal covariate; and in [23], the authors propose an omnibus test that can detect any type of association between a censored survival time and a 1-dimensional covariate. The recent non-parametric test in [27] was introduced to deal with general covariates on $\mathbb{R}^d$. This approach deals with censored data by transforming it into uncensored samples, and then applies a well-known kernel independence test based on the Hilbert-Schmidt Independence Criterion (HSIC) [15, 5].

In this paper we propose a non-parametric test which can potentially detect any type of dependence between right-censored survival times and general covariates. Our testing procedure is based on a dependence measure between survival times and covariates which is constructed using weighted log-rank tests and the theory of reproducing kernel Hilbert spaces (RKHSs). We provide asymptotic results for our test-statistic and propose an approximation of the rejection region of our test by using a Wild Bootstrap procedure. Under mild regularity conditions, we prove that both the oracle and the testing procedure based on the Wild Bootstrap approximations are asymptotically consistent, meaning that our test can detect any type of dependence.

The closest prior works to our approach are [10] and [27], which are both based on kernel methods. In the former, the authors specifically address the two-sample problem for right-censored data. While the present test may be seen as related, the two-sample analysis is quite different, and heavily relies on the binary nature of the covariates: the main results apply ad-hoc theory developed for log-rank tests, which is not available in our setting. As noted above, [27] bypass the problem of right-censored data by transforming it into uncensored samples, however this comes at the cost of losing considerable information in the data. By contrast, our approach deals directly with the censored observations without loss of information, resulting in a

major performance advantage in practice, as we demonstrate in our experiments.

The paper is structured as follows. In Section 2 we introduce relevant notation. In Section 3 we define the kernel log-rank test and show that it can be interpreted as i) the supremum of a collection of score tests associated to a particular family of cumulative hazard functions, and, ii) as an RKHS distance, revealing a similarity with the HSIC [14]. In Section 4 we study the asymptotic behavior of our statistic under both the null and alternative hypothesis, and we establish connections with known approaches such as the two-sample test proposed in [10], and the Cox score test. Section 5 shows how to effectively approximate the null distribution by using Wild Bootstrap [7]. Sections 6 and 7 contain extensive experiments investigating the performance of the kernel log-rank test on a range of synthetic and real datasets.

2 Notation

Survival analysis notation Let $((Z_i, C_i, X_i))_{i\in[n]}$ be a collection of random variables taking values on $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^d$, where Z_i denotes a survival time of interest, C_i is a censoring time, and X_i is a vector of covariates taking values on $\mathbb{R}^d$, and $d \geq 1$. In practice, we do not observe (Z_i, C_i, X_i) directly, but instead observe triples (T_i, Δ_i, X_i), where $T_i = \min\{Z_i, C_i\}$ and $\Delta_i = \mathbb{1}_{\{T_i=Z_i\}}$. This type of data is known as right-censored data. Additionally, we assume $Z \perp C|X$, which is known as independent right-censoring.

We denote by F_T, F_Z, F_C and F_X, the marginal distribution functions associated to T, Z, C and X, respectively. We use standard notation to denote joint and conditional distributions, e.g. $F_{ZC|X=x}$ denotes the joint distribution of Z and C conditional on $X = x$. We denote by $S_T(t) = 1 - F_T(t)$, $S_Z(t) = 1 - F_Z(t)$ and $S_C(t) = 1 - F_C(t)$ the marginal survival functions associated to T, Z and

C, respectively, and by $S_{T|X=x}(t) = 1 - F_{T|X=x}(t)$, $S_{Z|X=x}(t) = 1 - F_{Z|X=x}(t)$ and $S_{C|X=x}(t) = 1 - F_{C|X=x}(t)$, the respective survival functions conditioned on $X = x$. In this work, we assume that $Z|X = x$ and $C|X = x$ are continuous random variables for almost all $x \in \mathbb{R}^d$, with densities denoted by $dF_{Z|X=x}(t)$ and $dF_{C|X=x}(t)$ respectively. We further assume that Z and C are proper random variables, meaning that $\mathbb{P}(Z < \infty|X = x) = 1$ and $\mathbb{P}(C < \infty|X = x) = 1$ for almost all $x \in \mathbb{R}^d$. The marginal cumulative hazard function of Z is defined as $\Lambda_Z(t) = \int_0^t S_Z(s)^{-1} dF_Z(s)$. Similarly, the conditional cumulative hazard of Z given $X = x$ is $\Lambda_{Z|X=x}(t) = \int_0^t S_{Z|X=x}(s)^{-1} dF_{Z|X=x}(s)$. We define $\tau_n = \max\{T_1, \ldots, T_n\}$, $\tau_x = \sup\{t : S_{T|X=x}(t) > 0\}$ and $\tau = \sup\{t : S_T(t) > 0\}$; note that $\tau_n \overset{a.s.}{\to} \tau$.

Counting processes notation We use standard survival analysis/counting processes notation. For $i \in [n]$, we define the individual and pooled counting processes by $N_i(t) = \Delta_i \mathbb{1}_{\{T_i \leq t\}}$ and $N(t) = \sum_{i=1}^n N_i(t)$, respectively. Similarly, we define the individual and pooled risk functions by $Y_i(t) = \mathbb{1}_{\{T_i \geq t\}}$ and $Y(t) = \sum_{i=1}^n Y_i(t)$.

We assume that all our random variables take values on a common filtrated probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, where the sigma-algebra $\mathcal{F}_t$ is generated by $\left\{\mathbb{1}_{\{T_i \leq s, \Delta_i = 0\}}, \mathbb{1}_{\{T_i \leq s, \Delta_i = 1\}}, X_i : s \leq t, i \in [n]\right\}$, and the $\mathbb{P}$-null sets of $\mathcal{F}$. Under the null hypothesis, for $i \in [n]$, we define the individual and pooled $(\mathcal{F}_t)$-martingales, $M_i(t) = N_i(t) - \int_{(0,t]} Y_i(s) d\Lambda_Z(s)$ and $M(t) = N(t) - \int_{(0,t]} Y(s) d\Lambda_Z(s)$, respectively. Finally, we denote by $d\widehat{\Lambda}(t) = dN(t)/Y(t)$ the Nelson Aalen estimator of $d\Lambda_Z(t)$ under the null hypothesis. For more information about counting processes martingales, we refer the reader to Fleming and Harrington [11, Chapters 1 and 2].

In this work $\int_a^b$ means integration over $(a, b]$ unless $b = \tau$, in which case we integrate over (a, τ). Due to the simple nature of the martingales that appear in this work (which arise from counting processes), properties such as (squared-

)integrability of these processes are trivial, and thus we state them without proof. Also, note that for any $t > \tau_n$, it holds that $N(t) = N(\tau_n)$ and $M(t) = M(\tau_n)$. Hence $\int_{\mathbb{R}_+} g(t)dN(t) = \int_0^{\tau} g(t)dN(t) = \int_0^{\tau_n} g(t)dN(t)$; the same holds for the martingale M.

For simplicity of exposition and notation we assume $X \in \mathbb{R}^d$, however our results also apply straightforwardly to general covariate spaces, as our statistic is based on kernel functions that may be defined on more general domains: see next section.

3 Construction of the test

We are interested in testing if the failure times Z are independent of the covariates X. Specifically, we would like to test the null hypothesis,

$$H_0 : F_{ZX} = F_Z F_X, \quad \text{against} \quad H_1 : F_{ZX} \neq F_Z F_X.$$

One of the most popular approaches to solve this problem is the log-rank test for proportional hazard functions. This test can be obtained as a score test from a partial likelihood function for the Cox's proportional hazards model given by $\Lambda_{Z|X=x}(t) = e^{\beta^\intercal x}\Lambda_Z(t)$. This approach fails in many scenarios, however, since it only considers a linear effect of the covariates on the log hazard, which is given by the term $\beta^\intercal x$.

Our method generalizes the previous method by defining a general collection of log-rank tests in which the association between time and covariates is modeled through general functions $\omega(t, x)$, instead of the simple expression $\beta^\intercal x$.

General score test We obtain a general log-rank test, for a fixed function $\omega : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$, by computing the score test associated to the model defined in terms of the conditional cumulative hazard function,

$$\Lambda_{Z|X=x}(t; \theta, \omega) = \int_0^t e^{\theta\omega(s,x)} d\Lambda_Z(s) \quad \theta \in \Theta, \tag{1}$$

where $\omega : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ is some non-zero fixed function, Θ is an open subset of $\mathbb{R}$ containing $\theta = 0$, and $\Lambda_Z(s)$ is the marginal (baseline) cumulative hazard function associated to the failure time Z. Under the assumption that our data is generated by this model for some fixed function $\omega(t, x)$, testing the null hypothesis $H_0 : Z \perp X$ is equivalent to testing $H_0 : \theta = 0$, which can be done using a score test.

A score test is a hypothesis test used to check whether a restriction imposed on a model estimated by maximum likelihood is violated by the data. The score test assesses the gradient of the log-likelihood function, known as score function, evaluated at some parameter θ under the null hypothesis. Intuitively, if the maximizer of the log-likelihood function is close to 0, the score, evaluated at $\theta = 0$, should not differ from zero by more than sampling error.

The likelihood function associated to the right-censored data $(T_i, \Delta_i, X_i)_{i=1}^n$ can be computed as follows. Given X_i, the contribution to the likelihood of an uncensored observation (T_i, Δ_i) (that is, for $\Delta_i = 1$) is $dF_{Z|X_i}(T_i) = d\Lambda_{Z|X_i}(T_i)S_{Z|X_i}(T_i)$. When (T_i, Δ_i) is censored, $\Delta_i = 0$, the contribution corresponds to $S_{Z|X_i}(T_i)$. The latter follows from the fact that when T_i is censored, we only know that $Z_i > T_i$.

Thus, given the covariates $(X_i)_{i=1}^n$, the likelihood function for the data $(T_i, \Delta_i)_{i=1}^n$ under the model in Equation (1) corresponds to

$$\begin{aligned} L_n(\theta; \omega) &= \prod_{i=1}^n d\Lambda_{Z|X_i}(T_i)^{\Delta_i} S_{Z|X_i}(T_i) \\ &= \prod_{i=1}^n e^{\theta \Delta_i \omega(T_i, X_i)} d\Lambda_Z(T_i)^{\Delta_i} \exp\left\{ - \int_0^{T_i} e^{\theta \omega(s, X_i)} d\Lambda_Z(s) \right\}, \end{aligned}$$

where the second equality follows since $S_{Z|X}(t) = \exp\left\{ - \int_0^t d\Lambda_{Z|X}(s) \right\}$.

The score function is then defined as

$$U_n(\theta; \omega) = \frac{d}{d\theta} \log L_n(\theta; \omega) = \sum_{i=1}^n \left(\Delta_i \omega(T_i, X_i) - \int_0^{T_i} \omega(t, X_i) e^{\theta \omega(t, X_i)} d\Lambda_Z(t) \right),$$

and $U_n(0,\omega)$ is the score statistic associated to the null hypothesis, $H_0 : \theta = 0$. A normalized version of $U_n(0,\omega)$ can be obtained using the variance/covariance matrix of $U_n(\theta;\omega)$, written as $\Sigma(\theta;\omega) = \mathbb{E}(-\frac{\partial^2}{\partial\theta^2}\log L_n(\theta;\omega))$, and then writing $S_n(0;\omega) = U_n(0;\omega)^\intercal \Sigma(0;\omega)^{-1} U_n(0;\omega)$. By the Neyman-Pearson Lemma [25], it follows that the test based on $S_n(0;\omega)$ is the most powerful test for small deviations from the null under the model defined in Equation (1).

In general the marginal hazard function $d\Lambda_Z(s)$ is unknown, and thus $U_n(0;\omega)$ cannot be evaluated in practice. However, under the null, $d\Lambda_Z(s)$ can be estimated from the data using the Nelson-Aalen estimator [1] $d\widehat{\Lambda}_Z(t) = dN(t)/Y(t)$, yielding

$$\widehat{U}_n(0;\omega) = \sum_{i=1}^{n} \int_{\mathbb{R}_+} (\omega(t, X_i) - \bar{\omega}_n(t)) dN_i(t), \tag{2}$$

where $\bar{\omega}_n(t) = \sum_{j=1}^{n} \omega(t, X_j) Y_j(t)/Y(t)$.

Log-rank formulation The expression for the un-normalized score statistic given in Equation (2) can be written as a discrepancy between two empirical measures with respect to the weight function $\omega(t,x)$. In survival analysis terminology, this is known as weighted log-rank test, and, in our scenario, it takes the form

$$\mathrm{LR}_n(\omega) = \frac{1}{n}\widehat{U}_n(0;\omega) = \int_{\mathbb{R}_+}\int_{x\in\mathbb{R}^d} \omega(t,x)\left(d\nu_1^n(t,x) - d\nu_0^n(t,x)\right), \tag{3}$$

where ν_1^n and ν_0^n are empirical measures defined as

$$d\nu_1^n(t,x) = \frac{1}{n}\sum_{i=1}^{n} dN_i(t)\delta_{X_i}(x) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \delta_{T_i,X_i}(t,x) \tag{4}$$

and

$$d\nu_0^n(t,x) = \frac{dN(t)}{n}\sum_{i=1}^{n}\frac{Y_i(t)}{Y(t)}\delta_{X_i}(x) = \frac{1}{n}\sum_{j=1}^{n}\Delta_j\delta_{T_j}(t)\sum_{i=1}^{n}\frac{Y_i(t)}{Y(t)}\delta_{X_i}(x). \tag{5}$$

The next theorem gives a consistency limit result for $\mathrm{LR}(\omega)$.

Theorem 3.1. *Let $\omega : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ be a bounded measurable function. Then*

$$LR_n(\omega) \xrightarrow{\mathbb{P}} \int_{\mathbb{R}^d} \int_0^\tau \omega(s,x)(d\nu_1(s,x) - d\nu_0(s,x)),$$

where $d\nu_1(t,x) = S_{C|X=x}(t)dF_{ZX}(t,x)$, $d\nu_0(t,x) = S_{T|X=x}(t)d\alpha(t)dF_X(x)$ and $\alpha(I) = \int_I \int_{\mathbb{R}^d} S_{C|X=x}(t)/S_T(t)dF_{ZX}(t,x)$ for any measurable $I \subseteq (0,\tau)$.

Simple algebra shows that, under the null hypothesis (i.e., $H_0 : Z \perp X$), $\nu_1 = \nu_0$, and consequently $\mathrm{LR}_n(\omega) \xrightarrow{\mathbb{P}} 0$ for any weight function $\omega(t,x)$. Under some regularity conditions which we state in Assumption 3.2, we prove that $\nu_1 = \nu_0$ implies $Z \perp X$.

Assumption 3.2. *For almost all $x \in \mathbb{R}^d$, $S_{C|X=x}(t) = 0$ implies $S_{Z|X=x}(t) = 0$.*

Proposition 3.3. *Under Assumption 3.2, it holds that $\nu_1 = \nu_0$ if and only if $Z \perp X$.*

Note that $\mathrm{LR}_n(\omega) \xrightarrow{\mathbb{P}} 0$ does not necessarily imply $\nu_0 = \nu_1$, since, if we choose ω equal to the zero function, then $\mathrm{LR}_n(\omega) = 0$ trivially. Thus, when using log-rank tests, it is very important to use a relevant weight function for the problem at hand. Instead of choosing a single weight function, we propose to optimize over a large collection of candidate functions.

RKHS approach While normalized log-rank tests exhibit good statistical properties for small deviations from alternatives belonging to the model in Equation (1), this good behavior is only guaranteed for a single weight function ω at a time. In practice, it is very unlikely that the dependence structure of Z and X is known beforehand, and thus choosing the correct weight $\omega(t,x)$ (if it exists) seems hard.

In order to avoid choosing a particular weight $\omega(t,x)$ in advance, we consider a family of weighted log-rank statistics, and compute

$$\Psi_n^2 = \left(\sup_{\omega \in \mathcal{H} : \|\omega\|_{\mathcal{H}}^2 \leq 1} \mathrm{LR}_n(\omega) \right)^2, \tag{6}$$

where the function $\omega(t, x)$ is allowed to take values in a potentially infinite-dimensional space of functions $\mathcal{H}$. We refer to Ψ_n^2 as the *kernel log-rank* statistic.

In particular, we choose $\mathcal{H}$ as a reproducing kernel Hilbert space (RKHS) of functions. One of the main advantages of choosing this particular space, is that it gives a simple closed-form solution for the optimization problem of Equation (6). For general spaces of functions, finding the function $\widehat{\omega}$ that maximizes the likelihood function, or solving the optimization problem of Equation (6), might be much harder problem, as it is likely that $\widehat{\omega}$ does not have a closed-form solution. We will prove that, under some mild regularity assumptions, a sufficiently rich choice of RKHS will be able to detect any type of dependencies. Comparing with works that consider a maximum among normalised log-rank statistics (i.e., divided by the standard deviation) [18, 31, 12], our test will use the un-normalised statistic $\mathrm{LR}_n(\omega)$. This is fundamental to our result, as the linearity in ω of $\mathrm{LR}_n(\omega)$, combined with the properties of the RKHS, leads to a simple closed formula to evaluate Ψ_n^2. This being said, note that we are indirectly normalizing by choosing ω in the unit ball of $\mathcal{H}$.

Reproducing Kernel Hilbert Spaces An RKHS $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ is a Hilbert space of functions in which the evaluation operator is continuous. By the Riesz representation theorem, for all $(t, x) \in \mathbb{R}_+ \times \mathbb{R}^d$ there exists a unique element $\mathfrak{K}_{(t,x)} \in \mathcal{H}$ such that, for all $\omega \in \mathcal{H}$, it holds $\omega(t, x) = \langle \omega, \mathfrak{K}_{(t,x)} \rangle_{\mathcal{H}}$; this property is known as the *reproducing property*. We define the so-called reproducing kernel $\mathfrak{K} : (\mathbb{R}_+ \times \mathbb{R}^d)^2 \to \mathbb{R}$ as $\mathfrak{K}((t, x), (t', x')) = \langle \mathfrak{K}_{(t',x')}, \mathfrak{K}_{(t,x)} \rangle_{\mathcal{H}}$ for any $(t, x), (t', x') \in \mathbb{R}_+ \times \mathbb{R}^d$. By the Moore-Aronszajn theorem, for any symmetric positive-definite kernel $\mathfrak{K}$, there exists a unique RKHS for which $\mathfrak{K}$ is its reproducing kernel. Finally, for any finite signed Radon measure (not necessarily a probability measure), we define its embedding into

$\mathcal{H}$ as

$$\phi_\nu(\cdot) = \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \mathfrak{K}((t,x),\cdot) d\nu(t,x) \in \mathcal{H},$$

which existence is guaranteed by $\int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \sqrt{\mathfrak{K}((t,x),(t,x))} d\nu(t,x) < \infty$, see [4].

RKHS distance We define the embeddings of the empirical measures ν_1^n and ν_0^n (introduced in Equations (4) and (5)) into $\mathcal{H}$ with reproducing kernel $\mathfrak{K}$ as

$$\phi_1^n(\cdot) = \int_0^\tau \int_{\mathbb{R}^d} \mathfrak{K}((t,x),\cdot) d\nu_1^n(t,x) \quad \text{and} \quad \phi_0^n(\cdot) = \int_0^\tau \int_{\mathbb{R}^d} \mathfrak{K}((t,x),\cdot) d\nu_0^n(t,x), \quad (7)$$

respectively. Notice both ϕ_1^n and ϕ_0^n are well-defined elements of $\mathcal{H}$, as they are finite sums of elements of $\mathcal{H}$.

The next Theorem gives a closed-form expression for the kernel log-rank statistic in terms of the distance (induced by the norm) of the embeddings ϕ_0^n and ϕ_1^n.

Theorem 3.4.

$$\Psi_n^2 = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i \Delta_j \bar{\mathfrak{K}}_n((T_i,X_i),(T_j,X_j)), \quad (8)$$

where

$$\bar{\mathfrak{K}}_n((t,x),(t',x')) = \mathfrak{K}((t,x),(t',x')) - \sum_{k=1}^n \mathfrak{K}((t,x),(t',X_k)) \frac{Y_k(t')}{Y(t')}$$
$$- \sum_{j=1}^n \mathfrak{K}((t,X_j),(t',x')) \frac{Y_j(t)}{Y(t)} + \sum_{j,k=1}^n \mathfrak{K}((t,X_j),(t',X_k)) \frac{Y_j(t)}{Y(t)} \frac{Y_k(t')}{Y(t')}.$$

Moreover, if $\mathfrak{K}((t,x),(t',x')) = L(t,t')K(x,x')$*, then*

$$\Psi_n^2 = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 = \frac{1}{n^2} \operatorname{trace}(\boldsymbol{L}^{\boldsymbol{\Delta}} (\boldsymbol{I} - \boldsymbol{A}) \boldsymbol{K} (\boldsymbol{I} - \boldsymbol{A})^\intercal) \quad (9)$$

where $\boldsymbol{K}, \boldsymbol{L}^{\boldsymbol{\Delta}}$ *and* $\boldsymbol{A}$ *are* $(n \times n)$*-dimensional matrices whose entries* (i,j) *are defined as* $(\boldsymbol{K})_{i,j} = K(X_i, X_j)$*,* $(\boldsymbol{L}^{\boldsymbol{\Delta}})_{i,j} = \Delta_i \Delta_j L(T_i, T_j)$ *and* $(\boldsymbol{A})_{i,j} = A_{ij} = \frac{Y_j(T_i)}{Y(T_i)}$*, and* $\boldsymbol{I}$ *denotes the identity matrix.*

4 Asymptotic Analysis

Asymptotic null distribution We study the asymptotic null distribution of $n\Psi_n^2$ which is fundamental to construct a testing procedure. The key step is to show that we can rewrite Ψ_n^2 as a V-statistic plus an asymptotically negligible term. The asymptotic null distribution of $n\Psi_n^2$ then follows from the standard theory of V-statistics. We refer to [28, Section 5.5.2] for a discussion of V-statistics.

Proposition 4.1. *Under the null hypothesis $H_0 : Z \perp X$, the kernel log-rank statistic can be written as*

$$\Psi_n^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} J_n((T_i, \Delta_i, X_i), (T_j, \Delta_j, X_j)), \tag{10}$$

where $J_n : (\mathbb{R}_+ \times \{0,1\} \times \mathbb{R}^d)^2 \to \mathbb{R}$ is a symmetric random function defined as

$$J_n((s,c,x),(s',c',x')) = \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \bar{\mathfrak{K}}_n((t,x),(t',x'))dm_{s,c}(t)dm_{s',c'}(t),$$

and $dm_{s,c}(t) = c\delta_s(t) - \mathbb{1}_{\{s\geq t\}}d\Lambda_Z(t)$.

The expression in Equation (10) suggests a V-statistic representation for the kernel log-rank statistic. In the next result, we prove that $n\Psi_n^2$ can, indeed, be approximated by a V-statistic, by showing that J_n can be replaced by its population version J, which follows from replacing the random kernel $\bar{\mathfrak{K}}_n$ by its corresponding population version $\bar{\mathfrak{K}}$, given by

$$\begin{aligned}
&\bar{\mathfrak{K}}((t,x),(t',x')) \qquad\qquad (11)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left(\mathfrak{K}((t,x),(t',x')) - \mathfrak{K}((t,x),(t',y'))\frac{S_{C|X=y'}(t')}{S_C(t')}\right.\\
&\left. - \mathfrak{K}((t,y),(t',x'))\frac{S_{C|X=y}(t)}{S_C(t)} + \mathfrak{K}((t,y),(t',y'))\frac{S_{C|X=y}(t)S_{C|X=y'}(t')}{S_C(t)S_C(t')}\right)dF_X(y)dF_X(y'),
\end{aligned}$$

which is valid under the null as $S_T(t) = S_Z(t)S_C(t)$.

Assumption 4.2. $\mathfrak{K}((t,x),(t',x')) = L(t,t')K(x,x')$, *and both* K *and* L *are bounded.*

Lemma 4.3. *Under Assumption 4.2 and the null hypothesis, it holds*

$$\Psi_n^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} J((T_i,\Delta_i,X_i),(T_j,\Delta_j,X_j)) + o_p(n^{-1}),$$

where

$$J((s,c,x),(s',c',x')) = \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \bar{\mathfrak{K}}((t,x),(t',x'))dm_{s,c}(t)dm_{s',c'}(t). \tag{12}$$

It can be easily checked that $\mathbb{E}(J((t,c,x),(T_1,\Delta_1,X_1))) = 0$ for any $(t,c,x) \in \mathbb{R}_+ \times \{0,1\} \times \mathbb{R}^d$ under the null (since $dm_{T_i,\Delta_i}(t) = dM_i(t)$). The statistic Ψ_n^2 is then approximately a degenerate V-statistic, and thus we deduce its limit distribution from the classical theory of degenerate V-statistics [28, Section 5.5.2].

Theorem 4.4. *Under Assumption 4.2 and the null hypothesis, it holds that*

$$n\Psi_n^2 \xrightarrow{\mathcal{D}} \int_{x\in\mathbb{R}^d}\int_0^\tau \bar{\mathfrak{K}}((t,x),(t,x))S_{C|X=x}(t)dF_Z(t)dF_X(x) + \mathcal{Y}$$

where $\mathcal{Y} = \sum_{i=1}^{\infty}\lambda_i(\xi_i^2 - 1)$, $\xi_1, \xi_2, \ldots$ *are i.i.d. standard normal random variables, and* $\lambda_1, \lambda_2, \ldots$ *are non-negative constants which depend on the distribution of the random variables* (Z,C,X) *and the kernel* $\mathfrak{K}$.

The next result states that if we directly replace $\bar{\mathfrak{K}}_n$ by its limit $\bar{\mathfrak{K}}$ in Equation (8), the resulting test-statistic has the same asymptotic null distribution as $n\Psi_n^2$.

Theorem 4.5. $n\Psi_n^2$ *and* $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i\Delta_j\bar{\mathfrak{K}}((T_i,X_i),(T_j,X_j))$ *have the same asymptotic distribution under the null hypothesis.*

Power under alternatives We next analyze the asymptotic behavior of Ψ_n^2 under the alternative hypothesis, i.e., $H_1 : F_{ZX} \neq F_Z F_X$. To this end, we first establish a consistency result for Ψ_n^2.

Lemma 4.6. *Under Assumption 4.2, it holds* $\Psi_n^2 \xrightarrow{\mathbb{P}} \|\phi_0 - \phi_1\|_{\mathcal{H}}^2$, *where*

$$\phi_1(\cdot) = \int_0^\tau \int_{\mathbb{R}^d} K(\cdot, x)L(\cdot, t)d\nu_1(t, x) \quad \textit{and} \quad \phi_0(\cdot) = \int_0^\tau \int_{\mathbb{R}^d} K(\cdot, x)L(\cdot, t)d\nu_0(t, x),$$

and ν_1 *and* ν_0 *are the population measures defined in Theorem 3.1.*

The next step is to ensure that $\|\phi_0 - \phi_1\|_{\mathcal{H}}^2$ is zero if and only if the null hypothesis holds. This result will follow from assuming conditions on the kernel $\mathfrak{K}$ that ensure the embeddings of the measures ν_0 and ν_1 onto $\mathcal{H}$ are injective, and from Proposition 3.3, which proves that $\nu_0 = \nu_1$ if and only if the null hypothesis holds.

Theorem 4.7. *Let* $\gamma > 0$ *be any constant. Suppose that both* L *and* K, *are bounded, continuous, characteristic [29], translation invariant, and* c_0*-kernels. Then, under the alternative hypothesis, and under Assumptions 3.2 and 4.2,* $n\Psi_n^2 \to \infty$ *as* n *grows to infinity, and thus*

$$\limsup_{n\to\infty} \mathbb{P}\left(n\Psi_n^2 > \gamma\right) = 1.$$

In the previous result we say K is a c_0-kernel if $K(x, \cdot) \in \mathcal{C}_0(\mathbb{R}^d)$, where $\mathcal{C}_0(\mathbb{R}^d)$ denotes the class of continuous functions in $\mathbb{R}^d$ that vanish at infinity. An example of a kernel that satisfies the conditions stated in the previous Theorem is the exponentiated quadratic kernel, given by $K(x, y) = \exp\{-(x - y)^\intercal \Sigma^{-1}(x - y)\}$.

Under the assumptions of Theorem 4.7, our testing procedure has asymptotic power tending to one for any alternative, and thus *it is able to detect any type of dependency* between survival times and covariates, given enough observations. Even if the kernel does not satisfy the properties stated in Theorem 4.7, however, we can guarantee the power of the test for alternatives following the model of Equation (1), that is, for alternatives of the form $\Lambda_{Z|X=x}(t; \theta) = \int_0^t e^{\theta\omega^\star(s,x)} d\Lambda_Z(s)$ for some $\omega^\star \in \mathcal{H}$

on the unit ball and $\theta \neq 0$, as the log-rank statistic $\mathrm{LR}_n(\omega^\star)^2 \to c > 0$, with c a positive constant. Then,

$$\Psi_n^2 = \left(\sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}^2 \leq 1} \mathrm{LR}_n(\omega) \right)^2 \geq \mathrm{LR}_n(\omega^\star)^2 \to c,$$

and thus when re-scaling by n, it holds that $n\Psi_n^2 \to \infty$.

Recovering existing tests We show our approach can also recover certain known tests for specific choices of the kernel function.

Example 4.8 (Two-sample weighted log-rank test)**.** Consider $X \in \{0, 1\}$, i.e., the two-sample problem. We can recover the standard weighted log-rank test with arbitrary weight function $\tilde{\omega} : \mathbb{R}_+ \to \mathbb{R}$, by choosing $\omega(t, 1) = -\omega(t, 0)$ and $\omega(t, 0) = \tilde{\omega}(t)/2$. Then, by replacing ω into Equation (3), we obtain

$$\mathrm{LR}_n(\omega) = \frac{1}{n} \int_0^\tau \tilde{\omega}(t) L(t) (d\widehat{\Lambda}^0(t) - d\widehat{\Lambda}^1(t)), \tag{13}$$

where $d\widehat{\Lambda}^j$ denotes the Nelson-Aalen estimator for each group $j \in \{0, 1\}$. Furthermore, $\tilde{\Psi}_n = \sup_{\omega: \|\omega\|_{\mathcal{H}}=1} \mathrm{LR}_n(\omega)$ recovers the general test proposed in [10].

Example 4.9 (Cox proportional hazards model)**.** Consider the Hilbert space of functions $\omega(t, x) = V^{1/2}\beta^\intercal x$, where $\beta \in \mathbb{R}^d$ and V is a positive-definite matrix of length-scales. By using this space of functions, our kernel log-rank statistic becomes

$$\Psi_n = \sup_{\beta \in \mathbb{R}^d: \|V^{1/2}\beta\|^2 \leq 1} \mathrm{LR}_n(\beta), \tag{14}$$

and it can be computed using Equation (9) with a linear kernel on the covariates, $K(x, x') = (V^{1/2}x)^\intercal (V^{1/2}x')$, and a constant kernel on times, $L(t, t') = 1$. Then

$$n\Psi_n^2 = \frac{1}{n} \sum_{i=1}^n \sum_{l=1}^n \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} \bar{\mathfrak{K}}_n((t, X_i), (t', X_l)) dM_i(t) dM_l(t') = U_{\mathrm{Cox}}(0)^\intercal V U_{\mathrm{Cox}}(0),$$

where $U_{\text{Cox}}(0) = \frac{d}{d\beta} l_{\text{Cox}}(\beta)\Big|_{\beta=0}$ is the score function associated to the so-called *Cox partial likelihood* $l_{\text{Cox}}(\beta)$. By choosing V equal to the inverse of the Fisher information matrix, the Cox score test and our Ψ_n are asymptotically equivalent.

5 Wild Bootstrap

In practice, the asymptotic null distribution is unknown, and thus we propose to use a Wild Bootstrap approximation to it. The Wild Bootstrap test-statistic $(\Psi_n^W)^2$ is given by

$$(\Psi_n^W)^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_i W_j \Delta_i \Delta_j \bar{\mathfrak{K}}_n((T_i, X_i), (T_j, X_j)),$$

where $W = (W_1 \ldots, W_n)$ are a collection of i.i.d. Rademacher random variables, which are independent of the data $\mathcal{D} = \{(T_i, \Delta_i, X_i)\}_{i=1}^n$.

In this section we prove two main results. The first result establishes that, under the null hypothesis, the asymptotic distribution of $n(\Psi_n^W)^2$ coincides with the asymptotic distribution of the kernel log-rank test-statistic $n\Psi_n^2$. The second result is analogous to Theorem 4.7, but replacing γ by the $1-\alpha$ quantile obtained by the Wild Bootstrap procedure.

Lemma 5.1. *Under Assumption 4.2, it holds that*

$$(\Psi_n^W)^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_i W_j \Delta_i \Delta_j \bar{\mathfrak{K}}((T_i, X_i), (T_j, X_j)) + o_p(n^{-1}). \tag{15}$$

Our first main result is given in the following Theorem.

Theorem 5.2. *Suppose that the null hypothesis and Assumption 4.2 hold true, and let $\mathcal{L}$ denote the asymptotic distribution of $n\Psi_n^2$. Then, for almost all sequences $(t_i, \delta_i, x_i)_{i\geq 1}$ sampled from $(T_i, \Delta_i, X_i)_{i\geq 1}$, $n(\Psi_n^W)^2 \xrightarrow{\mathcal{D}} \mathcal{L}$, as $n \to \infty$.*

The previous result guarantees $\lim_{n\to\infty} \mathbb{P}(n(\Psi^W)_n^2 > Q_{1-\alpha}) = \alpha$, where $Q_{1-\alpha}$ is the $1-\alpha$ quantile of $\mathcal{L}$. Note that $\mathcal{L}$ depends on the distribution θ of the triple (X, Z, C) that defines the null. Changing this distribution to a distribution θ' (that still satisfies the null) will lead to a potentially different asymptotic distribution $\mathcal{L}'$ of the test-statistic. Therefore, the speed of convergence of the above result may depend on θ. It is thus important to emphasize that our result ensures a pointwise asymptotic level, but not uniformly asymptotic level.

Our second main result is given in the following theorem.

Theorem 5.3. *Consider Assumptions 3.2 and 4.2, and assume that both L and K, are bounded, continuous, characteristic, translation invariant, and c_0-kernels. Let $\alpha \in (0,1)$, and let $Q_{n,M}^W$ denote the $1-\alpha$ quantile obtained from a sample of fixed size M of the Wild Bootstrap test-statistic, $n(\Psi_n^W)_1^2, \ldots, n(\Psi_n^W)_M^2$. Then, under the alternative hypothesis*

$$\mathbb{P}\left(n\Psi_n^2 > Q_{n,M}^W\right) \to 1 \quad \text{as } n \to \infty.$$

From the previous result, we deduce that, under the Assumptions of Theorem 5.3, the test based on the Wild Bootstrap approximation of the null distribution is able to *detect any type of dependency between survival times and covariates* asymptotically, as long as censoring does not hide the regions in which the dependence occurs.

Implementation Under Assumption 4.2, the Wild Bootstrap test-statistic can be easily evaluated as follows,

$$n(\Psi_n^2)^W = \frac{1}{n^2}\operatorname{trace}(\boldsymbol{L}^{\boldsymbol{\Delta},\boldsymbol{W}}(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{A})^\intercal), \tag{16}$$

where $\boldsymbol{L}^{\boldsymbol{\Delta},\boldsymbol{W}}$ is a $(n \times n)$-matrix defined as $(\boldsymbol{L}^{\boldsymbol{\Delta},\boldsymbol{W}})_{i,j} = (\Delta_i W_i)(\Delta_j W_j)L(T_i, T_j)$, and $\boldsymbol{K}, \boldsymbol{A}$ and $\boldsymbol{I}$ are defined in Theorem 3.4. Algorithm 1 below describes the implementation of our testing procedure.

Computational time By the following Proposition, our algorithm has the same computational complexity as the HSIC based permutation test of [14].

Proposition 5.4. *$n\Psi_n^2$ and $n(\Psi_n^W)^2$ can be computed in $\mathcal{O}(n^2)$ time.*

Using a simple Python implementation that does not use a GPU, running on a CPU with 4 cores at 1.6GHz, computation of the kernel log-rank statistic takes about 10 seconds for a sample of size 10000, and about 0.1 second for a sample of size 1000. If faster computation is required, we may adopt the large-scale approximations proposed in [33]. Moreover, Wild Bootstrap statistics can be computed in parallel, and matrix computations can be done on a GPU.

Algorithm 1: Wild Bootstrap.

Input: data $\{T_i, \Delta_i, X_i\}_{i=1}^n$, α and M

for *k in 1* $\to M$ **do**

  Sample $W = (W_1, \ldots, W_n) \overset{i.i.d.}{\sim}$ Rademacher

  Compute $(\Psi_n^W)_k^2$ as in equation (16)

Denote by $Q_{n,M}^W$ the $1-\alpha$ quantile of the sample $n(\Psi_n^W)_1^2, \ldots, n(\Psi_n^W)_M^2$

Compute $n\Psi_n^2$ as in Equation (9)

Reject if $n\Psi_n^2 > Q_{n,M}^W$

6 Experiments

We study the performance of the proposed kernel log-rank test for various choices of kernels. We choose the kernels to be products of a kernel on the covariates, K, and a kernel on the times, L. We denote the product kernel by (K, L). We study the following four cases: 1. $(K = \text{Lin}, L = 1)$, 2. $(\text{Gau}, 1)$, 3. $(\text{Fis}, 1)$ and 4. (Gau, Gau), where "Lin" denotes the linear kernel, "Gau" the exponentiated quadratic (Gaussian) kernel, "Fis" the linear kernel scaled by the Fisher information (see Example 4.9) and

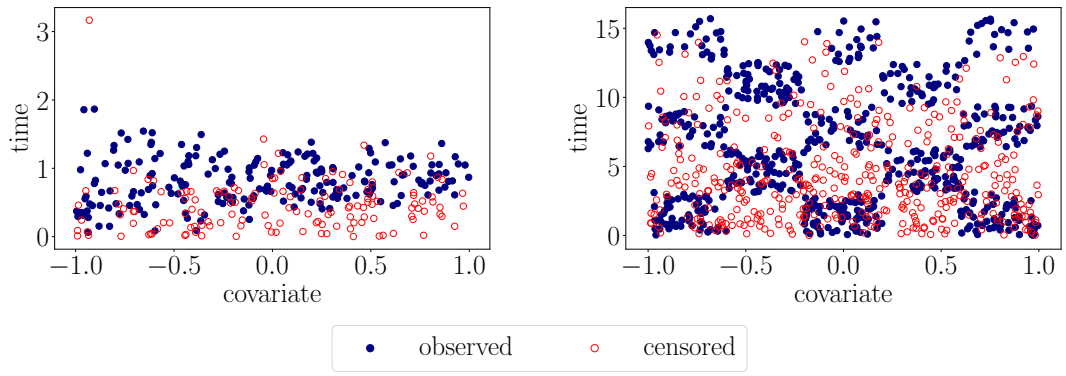

Figure 1: Scatterplots of samples from D.3 (left) and D.4 (right).

"1" denotes the constant kernel, i.e. $L = 1$ implies $L(t, s) = 1$ for all $t, s \in \mathbb{R}_+$. In all experiments we use the *median heuristic* to select the bandwidth of the Gaussian kernel: we choose $\sigma^2 = \text{median}\{||x_i - x_j||^2 \ : i \neq j\}/2$. We discuss the sensitivity of the test to different choices of bandwidth later this section. We set the level of the test to $\alpha = 0.05$ and use Algorithm 1 of Section 5 to perform the test with $M = 2000$ (or $M = 5000$ to estimate type 1 error) Wild Bootstrap samples to estimate the rejection region. We compare the kernel log-rank test with the traditional Cox likelihood ratio (Cox LR) test [6], denoted by Cph in the legends, and the optHSIC test [27], denoted by Opt in the legends. As in [27], we use the Brownian covariance kernel in optHSIC. The assessment of the Type I error can be found in Appendix A.1. Code to implement the kernel log-rank test and reproduce the experiments below can be found at `https://github.com/davidrindt/kernel_logrank_python_code`.

Power for 1-dimensional covariates In this section we investigate the power of the different tests using data simulated from distributions in which $X \not\perp Z$. In each case, we use an exponential censoring distribution in which $C \perp X$ and the mean of

C is chosen such that 60% of the events are observed. Throughout this subsection we let $X \sim \text{Unif}[-1, 1]$. Consider the following four distributions.

D.1 *A CPH distribution:* $Z|X = x \sim \text{Exp}(\text{mean} = \exp\{x/3\})$ and $C|X = x \sim \text{Exp}(\text{mean} = 1.5)$.

D.2 *A non-linear log-hazard:* $Z|X = x \sim \text{Exp}(\text{mean} = \exp\{x^2\})$ and $C|X = x \sim \text{Exp}(\text{mean} = 2.25)$.

D.3 *A family of Weibull distributions:* $Z|X = x \sim \text{Weib}(\text{shape} = 3.35 + 1.75 \cdot x, \text{scale} = 1)$ and $C|X = x \sim \text{Exp}(\text{mean} = 1.75)$.

D.4 *A checkerboard pattern:* See Figure 1. Because the pattern is more complicated, we let the sample size range from 500 to 2000 in steps of 500.

The top row of Figure 2 displays the rejection rates of the tests for samples from D.1 and D.2. Note that in D.1 the kernel log-rank test with linear kernel performs roughly equivalently to the LR test, which is ideally suited for this distribution as the CPH assumption holds. While the kernel log-rank tests with the rich kernels $(\text{Gau}, 1)$ and (Gau, Gau) do not lose much power in detecting this CPH dependency, we do observe a small tradeoff between richness of the kernels and power: the (Gau, Gau) kernel has slightly less power than the $(\text{Gau}, 1)$ kernel, which in turn has slightly less power than the $(\text{Lin}, 1)$ kernel and CPH LR test in D.1. In D.2 the quadratic term in the log-hazard violates the CPH assumption. The top right panel of Figure 2 shows that the linear kernel and the CPH LR test are unable to detect the dependency. Again, we find that the least rich kernel that is still able to model the dependency has the highest power, which in this case is the $(\text{Gau}, 1)$ kernel. In D.3 and D.4 the hazard function does not factorize into a function of covariates and a function of time, and a kernel is needed on time to model the dependency. Rejection rates are displayed in the bottom of Figure 2, confirming that indeed the kernel log-rank test with kernel (Gau, Gau) is the only test able to correctly reject the null hypothesis.

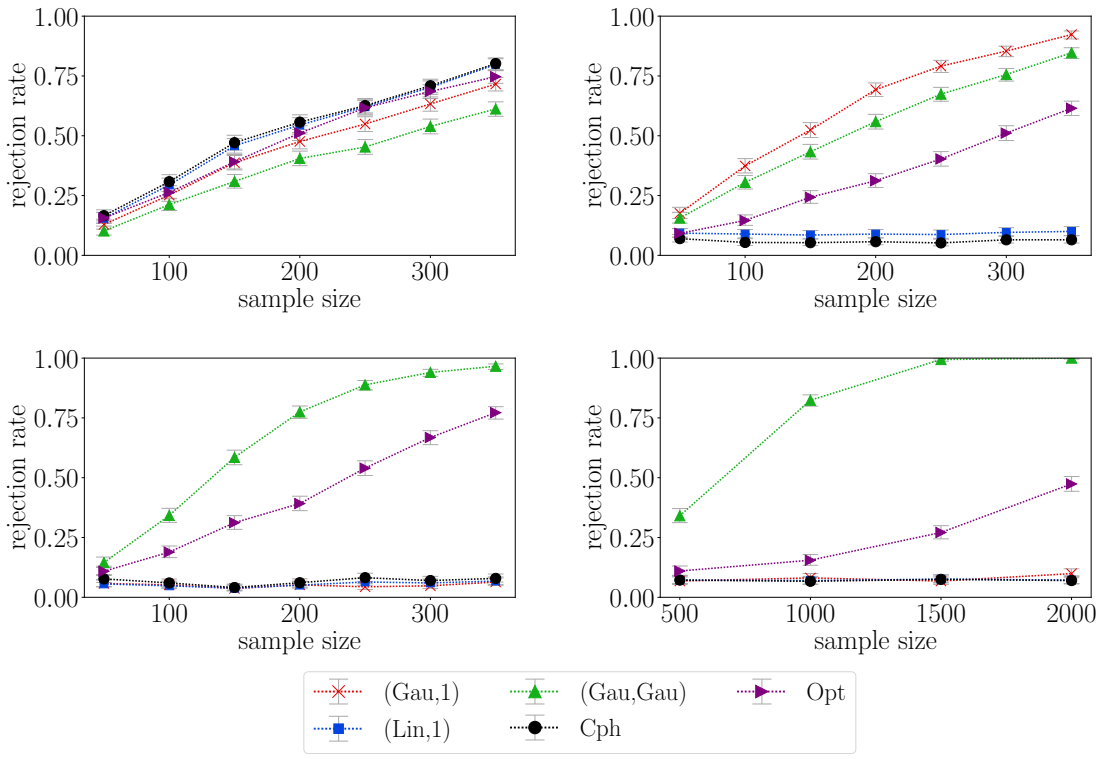

Figure 2: Rejection rates of the various test for D.1 (top left), D.2 (top right), D.3 (bottom left) and D.4 (bottom right).

Power for multidimensional covariates Let $X \sim \text{Normal}(\text{mean} = 0_d \text{ and cov} = \Sigma_d)$ where $0_d = (0, \ldots, 0) \in \mathbb{R}_d$, $\Sigma_d = MM^T$ and M is a $d \times d$ matrix of independent Normal$(0, 1)$ entries. Consider the following four distributions:

D.5: A CPH dependence on all covariates: $Z|X = x \sim \text{Exp}(\text{mean} = \exp\{1_d^T x/20\})$ and $C|X = x \sim \text{Exp}(\text{mean} = 1.5)$ where $1_d = (1, \ldots, 1) \in \mathbb{R}^d$.

D.6: A CPH dependence on single covariate: $Z|X = x \sim \text{Exp}(\text{mean} = \exp\{x_1/60\})$ and $C|X = x \sim \text{Exp}(\text{mean} = 1.5)$.

D.7: A non-CPH dependence on single covariate: $Z|X = x \sim \text{Exp}(\text{mean} = \exp\{x_1^2/60\})$ and $C|X = x \sim \text{Exp}(\text{mean} = 1.5)$.

D.8: A mixed dependence on 2-covariates: $Z|X = x \sim \text{Exp}(\text{mean} = \exp\{(x_1^2 + 3x_2)/60\})$ and $C|X = x \sim \text{Exp}(\text{mean} = 2)$.

Figure 3 displays rejection rates for both varying dimension and sample sizes. As our first finding, we observe that, while the CPH assumption holds for D.5 and D.6, the power of the kernel log-rank test is similar to that of the CPH LR test, which is an encouraging result. In D.7 and D.8 we again observe that in the presence of a non-CPH dependence, the kernel log-rank test has good power.

Varying censoring rates and distributions We study the rejection rates in cases where censoring depends on the covariate, and where the percentage of observed ($\Delta = 1$) events is $15, 30, 45, 60, 75, 90$ or 100%. The experiments are described in Appendix A.3. Under these varying censoring percentages, the main findings from the previous section remain true: the kernel log-rank test shows competitive power when the CPH assumption holds, is able to detect non-CPH dependencies, and achieves correct type 1 error. A final important observation is that optHSIC loses power compared to the kernel log-rank test for higher censoring rates.

Sensitivity to choice of bandwidth In the experiments presented thus far we set the bandwidth of the Gaussian kernel to be $\sigma^2 = \text{median}\{||x_i - x_j||^2 \ : i \neq j\}/2$. We now study the effect of varying the bandwidth of the Gaussian kernel on the rejection rate. Details of the experiments are in Appendix A.4. We find that, while for most scenarios a bandwidth can be selected that slightly outperforms the median heuristic, the median heuristic has a consistently good performance across all scenarios.

Choice of kernel When using the kernel log-rank test it is important to choose an appropriate kernel. While in principle we can choose $\mathfrak{K}$ to be any kernel, choos-

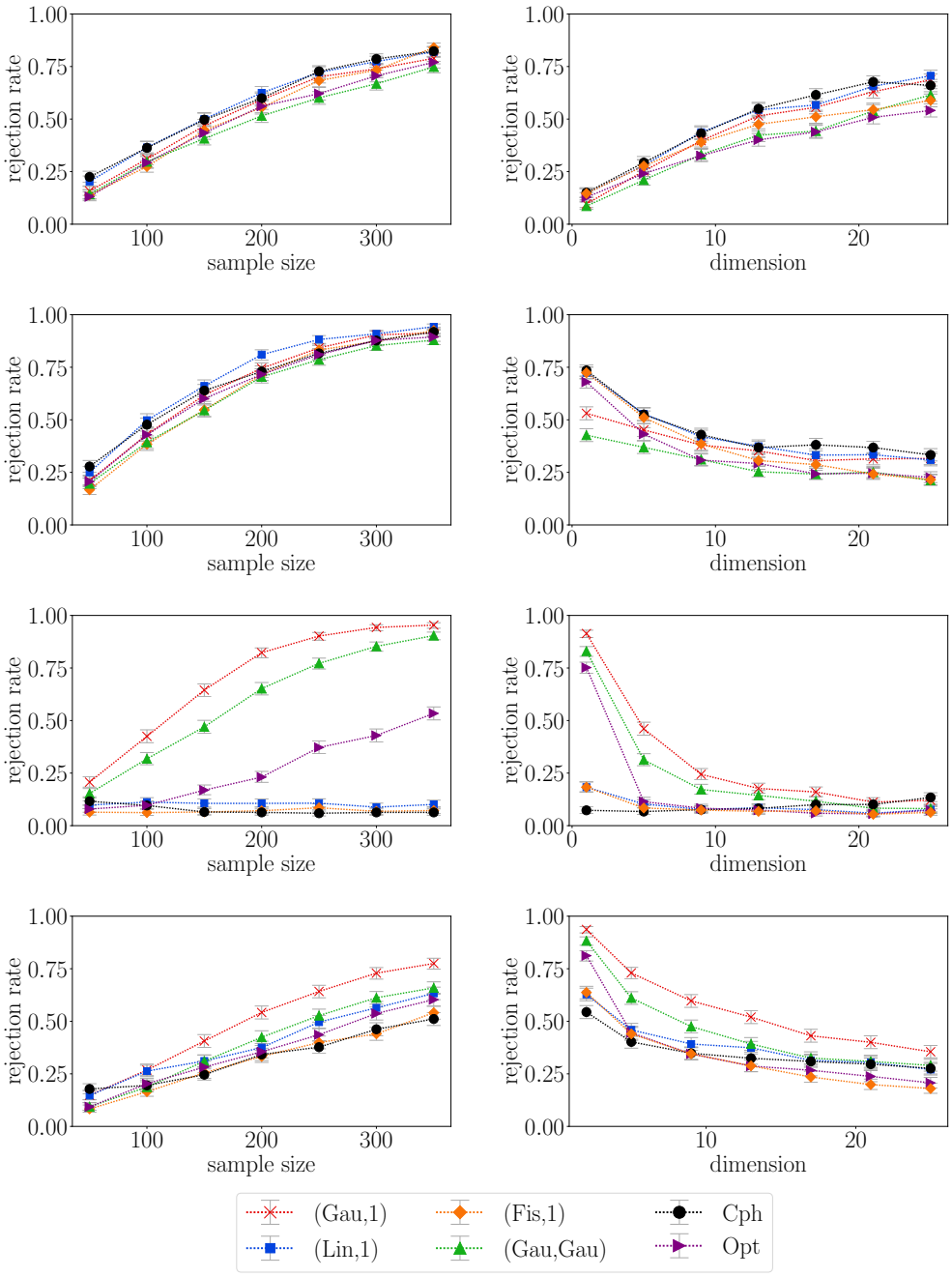

Figure 3: Rejection rates for D.5 (top) - D.8 (bottom). Left: rejection rates as the sample size increases. Right: rejection rates as the dimension increases.

ing a kernel that factorizes into a kernel on time and a kernel on the covariates has the important advantage that it gives the simple closed-form expression for our test-statistic in Theorem 3.4. In Theorems 4.7 and 5.3, we prove that our testing procedure is consistent for sufficiently expressive kernels K and L.

Our experiments show that in some scenarios there is a trade off between richness of the RKHS and the power of the test. For example, when the data is sampled from a distribution satisfying the CPH assumption, the kernel $(\mathrm{Lin}, 1)$ is generally more powerful than the richer $(\mathrm{Gau}, 1)$ or $(\mathrm{Gau}, \mathrm{Gau})$ kernels. Hence, if we have prior knowledge about the relationship, we may use this knowledge to choose the appropriate kernel. On the whole, however, we found the richer kernels to have competitive power, even in cases where less rich kernels were optimal.

We may also consider improvements on the median heuristic when selecting kernel bandwidth. In [16, 30, 21], which deal with the case of two-sample testing in the uncensored setting, part of the data is used to select parameters that result in the lowest approximate asymptotic p-value, and the test is then performed with the selected parameters on the remaining data. In [2], which addresses independence testing in the uncensored setting, an aggregation procedure is proposed over kernel bandwidths, which does not require data splitting, and is adaptive in a minimax sense over Sobolev classes of alternatives. It will be an interesting topic for future research to extend these kernel selection strategies to the censored setting.

7 Application to real-world datasets

We next apply our proposed method to two real-world datasets. We compare the p-values obtained by the kernel log-rank test with kernels $(\mathrm{Gau}, 1)$ and $(\mathrm{Gau}, \mathrm{Gau})$ to the p-values obtained by the CPH likelihood ratio test. We use 10000 Wild Bootstrap

samples in these examples.

		p-value		
Data	Covariates	Cph	$(\text{Gau}, 1)$	(Gau, Gau)
Biofeedback	(Trt, Recov)	0.496	0.014	0.023
	Trt	0.462	0.458	0.050
	Recov	0.301	0.007	0.029
	(Trt, log(Recov))	0.027	0.013	0.025
Colon	Age	0.627	0.080	0.097
	(Age, Perfor, Adhere)	0.102	0.017	0.018

Table 1: p-values obtained by the various tests for the Biofeedback and Colon data.

Biofeedback data The biofeedback treatment data studies the time until patients suffering from aspiration after head and neck surgery achieve full swallowing rehabilitation. The study is presented in [8] and the data was made available in the R package Coxphw [9]. In our presentation we name the event-time variable Rehab. Covariates in the dataset are a binary variable indicating biofeedback treatment, denoted Trt, and the time after the surgery until treatment could be started, denoted Recov. This dataset contains 33 individuals, of whom 70% were observed to fully rehabilitate (coded as $\delta = 1$).

In the first row of Table 1, we observe that the kernel log-rank test results in the p-values 0.014 for the kernel $(\text{Gau}, 1)$ and 0.023 for the kernel (Gau, Gau) when including both covariates. In contrast, the CPH likelihood ratio test results in the higher p-value of 0.496. This suggests the possibility of a non-linear relationship between the event-time of interest, Rehab, and the covariates Trt and Recov. This interpretation is strengthened by the following two observations. First, when we apply a logarithm transformation to the covariate Recov, as suggested in [8], the

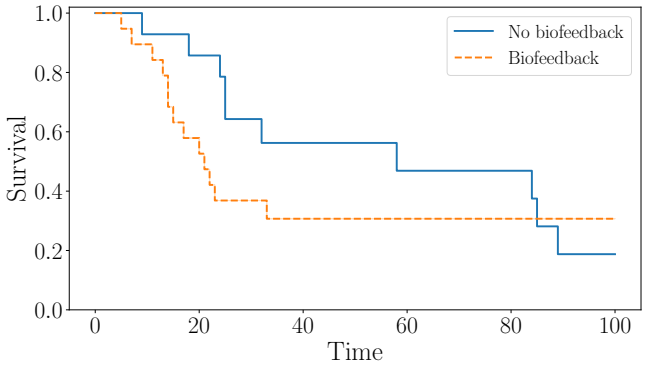

Figure 4: Kaplan-Meier estimates of the survival curves associated to the group that receives the Biofeedback treatment and one that does not receive the treatment. We observe that both curves cross which violets the CPH assumption.

CPH likelihood ratio test (shown in the fourth row of Table 1) results in a p-value of 0.027. The main advantage of the kernel log-rank test is that it does not require to manually transform the data to detect potential non-linear dependencies. Second, we observe that the kernel log-rank test using the $(\mathrm{Gau}, \mathrm{Gau})$ kernel is the only test that rejects the null hypothesis of independence between Rehab and Trt (see the second row of Table 1) at a significance level of 0.05. This result may be attributed to the fact that the survival curves associated to the two treatment groups cross (see Figure 4).

Colon data The Colon dataset studies the recurrence of tumors and survival in patients undergoing treatment for stage B/C colon cancer. The study is presented in [19] and [24]. Data from the 929 patients in the study is available in the R package Survival [32]. Each individual has 11 covariates. We consider the time of death in our analysis as the event-time, which was observed, i.e. $\delta = 1$, for 49% of the individuals.

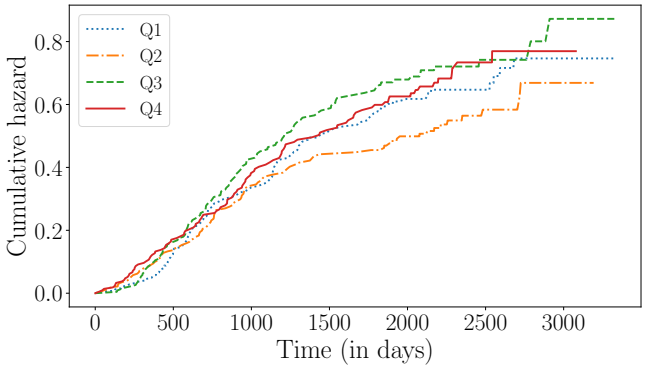

Figure 5: Nelson-Aalen estimates of the cumulative hazard for the four quartiles of age. Q1 denotes the youngest quartile and Q4 the oldest. Note that the curves do not satisfy the CPH assumption.

We focus our analysis on the variables Age, Perfor (a binary variable indicating the perforation of the colon) and Adhere (a binary variable indicating adherence of the tumor to nearby organs).

The p-values of different tests of independence are given in Table 1. Testing for independence between the age covariate and the event-time, the kernel log-rank test results in p-values of 0.080 and 0.097 for the $(\text{Gau}, 1)$ and (Gau, Gau) kernels respectively. In contrast, the CPH likelihood ratio test produces a much higher p-value of 0.627. This suggests the possibility of a non-linear dependence between the age covariate and the event-time of interest. This interpretation is strengthened by Figure 5, which displays cumulative hazard functions for different age quartiles. Indeed, these curves could suggest a non-linear effect of age, as the curves are not ordered by age. In particular we observe that first age quartile has a higher cumulative hazard function than the second quartile, but lower than the third. Finally, in the last row of Table 1, we observe differences in the conclusions of the kernel log-rank test (for

both kernels) and the CPH likelihood ratio test at a significance of $\alpha = 0.05$ when we include the binary variables Perfor and Adhere to the model. This difference in p-values suggests there may be a non-linear relationship between the event-time of interest and these covariates.

8 Conclusions

We have introduced a novel non-parametric independence test between right-censored survival times Z and covariates X. Our approach uses an infinite-dimensional exponential family of cumulative hazard functions, which are parameterized by functions in a reproducing kernel Hilbert space. By choosing an expressive Hilbert space of functions, we show that our testing procedure is able to detect any type of dependence, while for simpler Hilbert spaces, we recover ubiquitous approaches such as the Cox score test. The test statistic furthermore has an easily computed closed form. We provide a simple testing procedure based on the Wild Bootstrap, and demonstrate strong performance on a range of synthetic and real datasets.

SUPPLEMENTARY MATERIAL

A. Additional experiments In Section A.1 we study the Type 1 error, in Section A.2 we show additional experiments regarding the power of tests, in Section A.3 we show experiments for varying censoring percentages, and in Section A.4 we show experiments for varying bandwidths of the Gaussian kernel.

B. Preliminary results: In this section, and in order for this paper to be self-contained, we review some preliminary results that will be used in our proofs.

C. Auxiliary results: In this section we state some auxiliary results used for the proofs of the main results of the paper.

D. Main results: In this section we prove the main results of the paper.

E. Proofs of auxiliary results In this section we prove the auxiliary results introduced in Section C.

References

[1] Odd Aalen. Nonparametric estimation of partial transition probabilities in multiple decrement models. *Ann. Statist.*, 6(3):534–545, 1978.

[2] Mélisande Albert, Béatrice Laurent, Amandine Marrel, and Anouar Meynaoui. Adaptive test of independence based on HSIC measures. *arXiv preprint arXiv:1902.06441*, 2021.

[3] Vilijandas Bagdonavičius, Julius Kruopis, and Mikhail S. Nikulin. *Nonparametric tests for censored data*. ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, 2011.

[4] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004. With a preface by Persi Diaconis.

[5] Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. In *ICML'14: Proceedings of the 31st International Conference on International Conference on Machine Learning*, page II–1422–II–1430. Proceedings of Machine Learning Research, 2014.

[6] David R. Cox. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220, 1972.

[7] Herold Dehling and Thomas Mikosch. Random quadratic forms and the bootstrap for U-statistics. *J. Multivariate Anal.*, 51(2):392–413, 1994.

[8] Doris-Maria Denk and Alexandra Kaider. Videoendoscopic biofeedback: a simple method to improve the efficacy of swallowing rehabilitation of patients after head and neck surgery. *ORL*, 59(2):100–105, 1997.

[9] Daniela Dunkler, Meinhard Ploner, Michael Schemper, and Georg Heinze. Weighted cox regression using the R package coxphw. *Journal of Statistical Software, Articles*, 84(2):1–26, 2018.

[10] Tamara Fernández and Nicolás Rivera. A reproducing kernel hilbert space log-rank test for the two-sample problem. *Scandinavian Journal of Statistics*, 2021.

[11] Thomas R. Fleming and David P. Harrington. *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1991.

[12] Valérie Garès, Sandrine Andrieu, Jean-François Dupuy, and Nicolas Savy. An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials. *Stat. Med.*, 34(4):541–557, 2015.

[13] Robert J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951, 1992.

[14] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[15] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J. Smola. A kernel statistical test of independence. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 585–592, Red Hook, NY, USA, 2007. Curran Associates Inc.

[16] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1205–1213. Curran Associates, Inc., 2012.

[17] David P. Harrington and Thomas R. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566, 1982.

[18] Wolfgang Kössler. Max-type rank tests, U-tests, and adaptive tests for the two-sample location problem—an asymptotic power study. *Comput. Statist. Data Anal.*, 54(9):2053–2065, 2010.

[19] John A Laurie, Charles G Moertel, Thomas R Fleming, Harry S Wieand, John E Leigh, Jebal Rubin, Greg W McCormack, James B Gerstner, James E Krook, and James Malliard. Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. the north central cancer treatment group and the mayo clinic. *Journal of Clinical Oncology*, 7(10):1447–1456, 1989.

[20] Chap T. Le, Patricia M. Grambsch, and Thomas A. Louis. Association between survival time and ordinal covariates. *Biometrics*, 50(1):213–219, 1994.

[21] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *ICML '20: Proceedings of the 37th International Conference on Machine Learning*, pages 6316–6326, 2020.

[22] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports*, 50(3):163, 1966.

[23] Ian W. McKeague, A. M. Nikabadze, and Yan Qing Sun. An omnibus test for independence of a survival time from a covariate. *Ann. Statist.*, 23(2):450–475, 1995.

[24] Charles G Moertel, Thomas R Fleming, John S Macdonald, Daniel G Haller, John A Laurie, Catherine M Tangen, James S Ungerleider, William A Emerson, Douglass C Tormey, John H Glick, et al. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage iii colon carcinoma: a final report. *Annals of internal medicine*, 122(5):321–326, 1995.

[25] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[26] Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135(2):185–207, 1972.

[27] David Rindt, Dino Sejdinovic, and David Steinsaltz. A kernel-and optimal transport-based test of independence between covariates and right-censored lifetimes. *The International Journal of Biostatistics*, 1(ahead-of-print), 2020.

[28] Robert J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, Inc., New York, 1980. Wiley Series in Probability and Mathematical Statistics.

[29] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, 12:2389–2410, 2011.

[30] Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[31] Robert E. Tarone. On the distribution of the maximum of the logrank statistic and the modified wilcoxon statistic. *Biometrics*, 37(1):79–85, 1981.

[32] Terry M Therneau and Thomas Lumley. Package 'survival'. *R Top Doc*, 128(10):28–33, 2015.

[33] Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Stat. Comput.*, 28(1):113–130, 2018.

[34] David M. Zucker and Alan F. Karr. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Statist.*, 18(1):329–353, 1990.