A maximum-mean-discrepancy goodness-of-fit test for censored data

Tamara Fernández
Gatsby Computational Neuroscience Unit
University College London
t.a.fernandez@ucl.ac.uk
UCL

Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London
gretton@gatsby.ucl.ac.uk
UCL

Abstract

We introduce a kernel-based goodness-of-fit test for censored data, where observations may be missing in random time intervals: a common occurrence in clinical trials and industrial life-testing. The test statistic is straightforward to compute, as is the test threshold, and we establish consistency under the null. Unlike earlier approaches such as the Log-rank test, we make no assumptions as to how the data distribution might differ from the null, and our test has power against a very rich class of alternatives. In experiments, our test outperforms competing approaches for periodic and Weibull hazard functions (where risks are time dependent), and does not show the failure modes of tests that rely on user-defined features. Moreover, in cases where classical tests are provably most powerful, our test performs almost as well, while being more general.

1 Introduction

Survival analysis is a branch of statistics focused on the study of time-to-event data, usually called survival times. This type of data usually appears in applications such as industrial life-testing, death times of patients in clinical trials or duration of unemployment in a population. An important characteristic of this type of data is that survival times may be censored, meaning that we do not observe the actual value of a survival time but a bound for it. For instance, it is not uncommon that in a clinical trial, the actual death time of a patient is only known to be within an interval of time.

Arguably, the most common type of censoring is independent right-censoring which occurs when, instead of observing the actual survival time, say X, we observe a lower bound T for it, i.e., we observe T and we know that $X > T$. Other less common types of censoring mechanisms are independent left and interval censoring. Respectively, left and interval censoring arise when we observe either an upper bound T instead of the failure time X or an interval $(T_l, T_u) \subseteq \mathbb{R}_+$ in which the failure time X falls. A reasonable assumption we make is that the censoring mechanisms are non-informative about the distribution of the survival times X. We will provide a more extensive description of our setting and terminology in Section 2.

While in most statistical/machine-learning applications the cornerstone for analysing data is the distribution function, in survival analysis the main objects of study are the hazard function and the survival function. For a survival time X with density f and distribution F, its survival function S is defined by $1 - F$, and its hazard function λ by f/S. While the survival $S(t)$ function gives us the probability a patient survives up to time t, the hazard $\lambda(t)$ function is the instant risk of death given that she has survived until time t. Additionally, we define the cumulative hazard function by $\Lambda(t) = \int_0^t \lambda(x)dx$. It can be checked that $S(t) = e^{-\Lambda(t)}$ and thus S and λ are in a 1-1 correspondence.

The hazard function is extremely important in applications, and different families of hazard functions give rise to different problems in the area. Examples of important families are proportional hazards, which in a clinical trial may represent treatments with constant benefit/dis-benefit over time (when compared with the baseline), and crossing hazards, representing treatments that may have a negative impact at the beginning but long-term benefits, e.g. chemotherapy, among other behaviours. Distinguishing between dif-

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

ferent hazard functions is a fundamental problem in survival analysis.

In this paper, we study goodness of fit, i.e. the problem of testing the null hypothesis $H_0 : \lambda = \lambda_0$, or alternatively, $H_0 : S = S_0$, where λ_0 and S_0 denote some specified hazard and survival functions in the setting of independent right censoring.

A few methodologies have been proposed to attack this problem. The Log-rank test is the most popular among practitioners. This test is based on the simple idea of comparing the cumulative hazard function under the null against the empirical cumulative hazard function. Among the good properties of this test, we have that the Log-rank test is the most powerful test for proportional hazard alternatives. Unfortunately, when the true relationship of the hazards is time-dependent it may lead to wrong decisions, i.e. low power [1]. An option to increase the power of Log-rank tests against time-dependent alternatives is to consider weighted Log-rank tests. These tests have been extensively studied [9, 18, 16]; see [1, 22] for details. By choosing an appropriate weight function, the weighted Log-rank test can be tailored to be optimal under specific alternatives, at the expense of reduced performance against other alternatives. Modern approaches attempt to increase overall test power by considering the combination of several weighted Log-rank tests into a single test-statistic, e.g. [6, 8, 12]. Nonetheless, weight-based approaches require us to hand-design the weight functions in advance, in anticipation of a particular set of alternatives. Moreover, the amount of data required by the test grows with the number of weights chosen.

As an alternative to log-rank tests, there exist a number of Chi-squared tests under censoring [25], [2] and [17], where the space is first partitioned, and the empirical probability of uncensored events is then compared in chi squared distance with its expectation (the latter requiring an estimate of the censoring distribution). See [1] and [13] for more detail.

Yet another approach, described in [3] is based on defining a *kernel density estimate* for the survival function, i.e. for $S = 1-F$ (obtained using a slightly modified Kaplan Meier procedure), with the test statistic then defined as the squared difference between this density estimate and the model density. Since this procedure relies on density estimation as an intermediate step, it has been found to be relatively data-inefficient, compared with more direct tests (see e.g. the recent discussion in [12]). We have independently confirmed this issue in our own experiments.

In the present work, we propose a new goodness of fit test for censored data, based on distances between probability distribution embeddings in a reproducing kernel Hilbert space (RKHS) [7, Chapter 4], [29]. A particular challenge arises due to the unknown censoring distribution: correcting for this directly using e.g. a Kaplan-Meier estimate of the survival time [21] leads to a more complex estimator when compared to the uncensored case, making standard tasks as bootstrapping or computing limiting distributions very difficult. Indeed, in this setting, naively applying the Kaplan-Meier estimator together with standard kernel test [5, 32] would lead to a potential incorrectly calibrated test.

Instead, we construct a sample mapping that requires no such correction, which we describe in Section 3. We emphasise that our approach does not require evaluation nor integration of the hazard function under the null, which can be challenging. In Section 4 we define a test statistic for these transformed data based on the maximum mean discrepancy [14], which is the RKHS distance between two distribution embeddings. The resulting test statistic is a simple V-statistic, and hence the test threshold can readily be obtained using a wild bootstrap procedure.

In Section 6, we illustrate the performance of our model for a number of use-cases, both for proportional hazards (where the classical Log-rank test is provably most powerful) and for time-varying hazards, including periodic and Weibull hazard functions. In the case of proportional hazards, we perform almost as well as the most powerful model, despite our test being more general; in the case of periodic hazards, we greatly outperform alternative approaches and in the case of Weibull hazards we also outperform alternative approaches.

2 Problem setting

We briefly explain the setting of censored data, and associated challenges. Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution of interest F on $\mathbb{R}_+$ and, independent from this sample, consider $C_1, \ldots, C_n \overset{i.i.d.}{\sim} G$ from a nuisance distribution G. The data we observe correspond to the pairs $(T_1, \Delta_1), \ldots (T_n, \Delta_n)$, where $T = \min\{X, C\}$ and $\Delta = \mathbb{1}\{T = X\}$. This data-framework corresponds to *independent right censoring.*

Given this data, the task of estimating the cumulative distribution function $F = 1 - S$ of the failure times can not be solved by using the empirical distribution. Indeed, the empirical estimator given by $H_n(x) = \sum_{T_i \leq x} \frac{1}{n}$ approximates the distribution of the minimum of X and C, i.e., $H = 1-(1-F)(1-G)$. Alternatively, if we drop all the observations that we know are censored, that is, where $\Delta_i = 0$ and thus $T_i = \min\{X_i, C_i\} = C_i$, the empirical distribution $H_n^1(x) = \sum_{T_i \leq x, \Delta_i = 1} \frac{1}{n}$ approximates the function $H^1(x) = \int_0^x (1 - G(t))dF(t)$. The lack of natural empirical-type estimators for F takes out of competition all the testing approaches which heavily rely on them.

The non-parametric maximum likelihood estimator of F under independent right censoring is the Kaplan-Meier estimator [21], defined as $\bar{F}_n(x) = \sum_{T_{[i:n]} \leq x}^n W_i$, where $W_i = \frac{\Delta_{[i:n]}}{n} \prod_{j=1}^{i-1} \left(1 + \frac{1-\Delta_{[j:n]}}{n-j}\right)$, $T_{[i:n]}$ denotes the i-th order statistic of the sample $\{T_i\}_{i=1}^n$, and $\Delta_{[i:n]}$ is its corresponding censoring indicator. In the particular case in which all the observations are uncensored, the Kaplan-Meier estimator simplifies to the empirical estimator $\bar{F}_n(x) = \hat{F}_n(x) = \sum_{X_i \leq x} \frac{1}{n}$. An immediate drawback of the Kaplan-Meier estimator is that it can not be written as the sum of independent random variables (note that W_i depends explicitly on the first i data points), and all statistical approaches based on this estimator must address this.

3 Construction of a null distribution

Consider the independent right-censoring scheme. Under the null hypothesis $H_0 : F = F_0$, it holds $F_0(X_i) \sim \mathcal{U}(0, 1)$, then testing the null hypothesis is equivalent to test for $H_0 : FF_0^{-1} = F_{\text{unif}}$, where F_{unif} denotes the uniform distribution function on $(0, 1)$. Notice that since we have right censored data we do not observe the failure time X_i but instead we observe $T_i = \min(X_i, C_i)$. Nevertheless, since F_0 is increasing (it is a distribution function), it holds $F_0(T_i) = F_0(\min\{X_i, C_i\}) = \min\{F_0(X_i), F_0(C_i)\}$, and thus the indicator function Δ_i is consistent with the order of $F_0(X_i)$ and $F_0(C_i)$. Then, we transform our initial problem into testing whether $\{F_0(X_i)\}_{i=1}^n$ follows a uniform distribution based on the right censored data $\{U_i, \Delta_i\}_{i=1}^n$, with $U_i = F_0(T_i)$. For left and interval censoring the same argument applies.

Up to this point, we have just transformed our problem to test for uniformity, but we still need to deal with the censored data. We overcome this problem by introducing an estimator of the distribution function FF_0^{-1}, based on the censored data $\{U_i, \Delta_i\}_{i=1}^n$ which can be written as the sum of independent random variables, and which is unbiased under the null hypothesis. The form of this estimator allows us to use the classical theory of U-statistics to derive the asymptotic distributions related to our test approach.

For right censored data, whenever $\Delta_i = 1$, we know that $U_i = F_0(X_i)$ which is exactly the random variable we are interested in. Then, by following the approach of the empirical distribution, we put a point mass of size $1/n$ on U_i. Otherwise, if $\Delta_i = 0$ then $U_i = F_0(C_i)$. From this event, the only information we can deduce is $F_0(X_i) > F_0(C_i)$. To reflect our lack of information, we distribute the weight $1/n$ associated to U_i uniformly on $(U_i, 1)$. The estimate we just described corresponds to

$$\tilde{F}(x) = \frac{1}{n} \sum_{U_i \leq x} \Delta_i + (1 - \Delta_i) \frac{x - U_i}{1 - U_i}. \tag{1}$$

It is clear that this estimator can be generalized to deal with left and interval censoring by distributing the weights associated to these censored random variables uniformly over their corresponding intervals: in the case of left censoring, uniformly on the interval $(0, U_i)$; and for interval censoring, uniformly on $(U_{i,l}, U_{i,u})$, where $U_{i,l}$ and $U_{i,u}$ denote a lower an upper limit for the true value $F_0(X_i)$. The next proposition, whose proof is deferred to the supplementary material, establishes that our estimator is unbiased.

Proposition 3.1. *Under the null hypothesis, the estimator $\tilde{F}$, based on the data $\{U_i, \Delta_i\}_{i=1}^n$, is an unbiased estimator of the uniform distribution function F_{unif}.*

4 A kernel-based test

In defining a statistic for testing goodness of fit for censored data, we make use of kernel distribution embeddings: that is, embeddings of probability measures to a reproducing kernel Hilbert space (RKHS) [7, Chapter 4], [29]. The distance between distribution embeddings is denoted the maximum mean discepancy (MMD) [14].

Goodness of fit can be tested using the MMD between the model and sample: this is the approach in [5, 32], bearing in mind the equivalence of distance and kernel-based measures of divergence [27].[1] This approach is inappropriate for our setting, given the censoring. Instead, we will use as our statistic the MMD between a uniform distribution and the sample-based distribution introduced in eq. (1).

We begin with the reproducing kernel Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ of functions from $[0, 1] \to \mathbb{R}$, with kernel $K : [0, 1] \times [0, 1] \to \mathbb{R}$ such that $K(\cdot, x) \in \mathcal{H}$ for all $x \in [0, 1]$, and $f(x) = \langle f(\cdot), K(\cdot, x) \rangle_{\mathcal{H}}$ for each $f \in \mathcal{H}$ and $x \in [0, 1]$. The mean embedding of a probability measure P on the RKHS $\mathcal{H}$ is $\mu_P(\cdot) = \mathbb{E}_P(K(\cdot, X))$, where the expectation is to be understood in terms of

[1] We note that an alternative approach is to modify the RKHS using a Stein operator, yielding a class of functions with zero expectation under the model: [10, 24]. Yet another approach would be to define a Fisher statistic between the model and sample in an RKHS [15, 4].

the Bochner integral. In particular, μ_P is well-defined whenever $\mathbb{E}_P(\sqrt{K(X,X)}) < \infty$, which is guaranteed under the following assumption:

Assumption 4.1. *There exists a constant $M \geq 1$ such that $|K(x,y)| \leq M$ for all $x, y \in [0,1]$.*

For a sufficiently rich RKHS, mean embeddings are injective, and uniquely represent their respective probability measures [30]: such RKHS are called *characterisitc*. The exponentiated quadratic kernel used in the present work satisfies this property.

We base our test on the maximum mean discrepancy between the uniform distribution F_{unif} and FF_0^{-1}, i.e., $MMD(F_{\text{unif}}, FF_0^{-1}) := \|\mu_{FF_0^{-1}} - \mu_{\text{unif}}\|_{\mathcal{H}}$, where we use the estimator $\tilde{F}$ of equation (1) for FF_0^{-1}.

We now proceed to study the asymptotics of $MMD(F_{\text{unif}}, \tilde{F}(x))$. As we will see, the main advantage of our goodness-of-fit estimator is that it can be expressed as the sum of independent random variables, and thus it allows us to use standard machinery from the theory of U-statistics [28] in deriving the asymptotic properties of our statistic.

From the definition of the kernel mean embedding and by the reproducing kernel property, it holds that

$$\begin{aligned} &MMD(F_{\text{unif}}, \tilde{F})^2 = \|\mu_{\text{unif}} - \mu_{\tilde{F}}\|^2_{\mathcal{H}} \\ = &\int_0^1 \int_0^1 K(x,y)(dx - d\tilde{F}(x))(dy - d\tilde{F}(y)). \quad (2) \end{aligned}$$

Recall the definition of $\tilde{F}(x) = \sum_{i=1}^n h_{U_i,\Delta_i}(x)$, where $h_{U_i,\Delta_i}(x) = \frac{\mathbb{1}_{\{U_i \leq x\}}}{n}\left(\Delta_i + (1-\Delta_i)\frac{x-U_i}{1-U_i}\right)$, then $d\tilde{F}(x)$ is given by

$$d\tilde{F}(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{U_i < x\}} \frac{1-\Delta_i}{1-U_i} dx + \Delta_i \delta_{U_i}(x), \quad (3)$$

where $\delta_U(x)$ denotes a delta measure on U. Based on the reproducing kernel K, we define the U-statistic kernel $J : ([0,1] \times \{0,1\})^2 \to \mathbb{R}$ as

$$\begin{aligned} J((u,\delta),(u',\delta')) &= \textstyle\int_0^1 \int_0^1 K(x,y)(dx - dh_{u,\delta}(x)) \\ &\qquad (dy - dh_{u',\delta'}(y)). \quad (4) \end{aligned}$$

Note that the U-statistic kernel $J : \mathbb{R} \times \{0,1\} \to \mathbb{R}$ depends neither on F_0 nor on the censoring distribution G; i.e., its implementation is distribution free, therefore it need only be computed once. By contrast, as we will see in Section 5, Log-rank and Pearson tests require us to evaluate and integrate the hazard function $\lambda_0 = f_0/(1-F_0)$, which may be not trivial as λ_0 may not be easily computable. Moreover, since the test statistic depends explicitly on the function λ_0, it needs to be recomputed for each different hypothesis.

Proposition 4.2. *Denote by Q_0 and Q_a the distribution of the pair (U, Δ) under the null and alternative hypotheses, respectively. Then, under assumption 4.1 and $\epsilon > 0$, it holds*

$$\mathbb{P}_{Q_0}\left(\left|MMD^2 - \mathbb{E}_{Q_0}(MMD^2)\right| \geq \epsilon\right) \leq \exp\left\{-\frac{\epsilon^2 n}{32M}\right\}$$

where $MMD^2 = MMD^2(\tilde{F}, F_{\text{unif}})$ and $\mathbb{E}_{Q_0}(MMD^2) \leq 4M/n$. The same result holds when replacing Q_0 by Q_a, and F_{unif} by the mean measure $\mathbb{E}(\tilde{F})$, respectively.

Proposition 4.3. *The $MMD^2(F_{\text{unif}}, \tilde{F})$ can be written as a V-statistic of degree 2 in the product space of survival-times and censoring-indicators. In particular*

$$MMD^2(F_{\text{unif}}, \tilde{F}) = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n J((U_i, \Delta_i), (U_j, \Delta_j)), \quad (5)$$

where J is the U-statistic kernel defined in equation (4). Moreover, under the null hypothesis, an unbiased estimate of $MMD^2(F_{\text{unif}}, FF_0^{-1})$ corresponds to the U-statistic

$$MMD_U^2(F_{\text{unif}}, \tilde{F}) = \binom{n}{2}^{-1} \sum_{i<j} J((U_i, \Delta_i), (U_j, \Delta_j)). \quad (6)$$

Lemma 4.4. *Under the null hypothesis and under assumption 4.1, it holds*

$$nMMD^2(F_{\text{unif}}, \tilde{F}) \xrightarrow{\mathcal{D}} Y + \mathbb{E}_{Q_0}(J(U,\Delta),(U,\Delta)), \quad (7)$$

where $Y = \sum_{j=1}^\infty \lambda_j(\xi_j^2 - 1)$ and ξ_j are independent normal random variables with zero mean and unit variance. Moreover, λ_j are the eigenvalues of the linear transformation $T : ([0,1] \times \{0,1\}, Q_0) \to L_2([0,1] \times \{0,1\}, Q_0)$ given by

$$(Tg)(u,\delta) = \mathbb{E}_{Q_0}(J((U,\Delta),(u,\delta))g(U,\Delta)),$$

where Q_0 denotes the joint distribution of the pair (U,Δ) under the null hypothesis.

Corollary 4.5. *Under the null hypothesis $nMMD_U^2(F_{\text{unif}}, \tilde{F}) \xrightarrow{\mathcal{D}} Y$, with Y defined as in Lemma 4.4. Under the alternative, when $E_{Q_a}(\tilde{F}) \neq F_{\text{unif}}$, $\sqrt{n}MMD_U^2(F_{\text{unif}}, \tilde{F})$ is asymptotically normal.*

Notice that a possible failure regime for our test occurs when $\tilde{F} \to F_{\text{unif}}$ for $F \neq F_0$ and $G(x) \neq \delta_0(x)$ (extreme censoring). Nevertheless by analysing the expected value of $\tilde{F}(u)$, see $E(Z^u)$ in the supplementary material, this situation seems very unlikely. Indeed, we conjecture that under the alternative, the map from censoring distributions to our estimator, i.e., $G \to \tilde{F}$ is injective, thus only $G(.) = \delta_0(.)$ (all censored) makes $\tilde{F}(x) = x$. We support our conjecture from our extensive experiments.

Although we have an expression for the asymptotic distribution of our test statistic, the parameters of this distribution are hard to compute. Instead, we propose to use the wild bootstrap to approximate the rejection regions of our test statistic under the null hypothesis. Specifically, we can re-sample from the distribution of our statistic $MMD(F_{\text{unif}}, \tilde{F}_n)$ by repeatedly sampling

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{W}_i\mathcal{W}_j J((U_i,\Delta_i),(U_j,\Delta_j)), \tag{8}$$

where $\{\mathcal{W}_i\}_{i=1}^n$ is a sequence of independent random variables with zero mean (to preserve the degeneracy property) and variance one. It was proved in [11] that (8) has the same limit distribution as our test-statistic $MMD(F_{\text{unif}}, \tilde{F}_n)$.

Algorithm 1: Wild bootstrap.

```
Input: data $\{U_i, \Delta_i\}_{i=1}^n$
Consider $\mathcal{W}_1, \ldots, \mathcal{W}_n$ a random sample with
    $\mathbb{E}(W_i) = 0$ and $\mathbb{V}ar(\mathcal{W}_i^2) = 1$
Return $MMD_k^B =$
    $\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathcal{W}_i\mathcal{W}_j J((U_i,\Delta_i),(U_j,\Delta_j))$
```

Given a nominal α value, the rejection region can be approximated by using a bootstrap sample $\{MMD_k^B\}_{k=1}^N$, as shown in Algorithm 1.

5 Competing approaches

5.1 Pearson-type goodness-of-fit: This approach, due to Akritas [2], considers the partition of the sample space of the random variables (T_i, Δ_i), for $\Delta \in \{0,1\}$, into k cells, and studies the distribution of the k-dimensional vector of observed minus expected frequencies. Let $A_j = [a_{j-1}, a_j)$ with $j = 1, \ldots, k$ and $0 = a_0 < a_1 < \ldots < a_{k-1} < a_k = \infty$. We define the observed frequencies as $N_{1,j} = \sum_{i=1}^n \mathbb{1}\{T_i \in A_j, \Delta_i = 1\}$ and the expected frequencies as $p_{1,j} = \int_{A_j}(1 - G)dF_0$. Observe that the expected frequencies depend on the unknown censoring distribution G, thus they are estimated by replacing the distribution G by the estimate $\hat{G} = 1-(1-\hat{H})/(1-F_0)$, where $\hat{H}$ is the empirical distribution of the minimum $T = \min\{X, C\}$. Estimators for the expected frequencies are then given by $\hat{p}_{1,j} = \int_{A_j}(1-\hat{G})dF_0 = \int_{A_j}(1-\hat{H})d\Lambda_0$. Under the null hypothesis $\mathbb{H}_0 : F = F_0$, where F_0 is a specified continuous function, the Pearson-type statistic corresponds to $\sum_{j=1}^k (N_{1j} - n\hat{p}_{1j})^2/(n\hat{p}_{1j})$, whose asymptotic distribution is χ_k^2.

Implementation: For the Pearson test, "Pearson" in the tables, we consider $k = 4$ cells, each of them accumulating 0.25 probability under the null hypothesis. The reason for using $k = 4$ cells is due to the trade-off between the number of data points required and the number of cells used: as the number of cells increases more data is needed to achieve the correct level of the test (i.e. Type-I error). We found that for $k = 4$, the test is competitive even though the Type-I error is a little bit overestimated for small sample sizes $(30, 50$ data points).

5.2 Log-rank test and Weighted Log-rank tests: Arguably, the Log-rank test is the most commonly-used statistical test for comparing survival curves. The test is performed by comparing differences of area as $Z = \int_0^\infty Y(t)d(\Lambda(t) - \Lambda_0(t))$, where $Y(t)$ denotes the so-called risk function, given by $Y(t) = \sum_{i=1}^n \mathbb{1}\{T_i \geq t\}$. The true cumulative hazard function, Λ, is estimated by the non-parametric Nelson-Aalen estimator. From the theory of counting processes [13], Z is asymptotically normal under appropriate scaling.

Weighted Log-rank tests generalize the classical Log-rank test, by considering an extra weight function W when comparing areas, giving the more general statistic $Z(W) = \int_0^\infty W(t)Y(t)d(\Lambda(t) - \Lambda_0(t))$.

Implementation: We consider weighted Log-rank tests, "LR1" and "LR2" in the tables, with weight functions $W_1(t) = 1$ and $W_2(t) = Y(t)$, respectively ($Y(t)$ risk function). Observe that $Z(W_1)$ is the classical Log-rank test. The second statistic $Z(W_2)$ corresponds to the so-called Gehan-Breslow test [1].

5.3 Combined Log-rank tests: These tests automate the procedure of choosing weight functions to create a more flexible test with broader power. We describe here the approach of [12]. A vector of weighted Log-rank tests $\boldsymbol{Z} = (Z(\omega_1 \circ \hat{F}_n), \ldots, Z(\omega_k \circ \hat{F}_n))$ is defined, with weight functions $\omega_1, \ldots, \omega_k$. The ω_i's are continuous and of bounded variation, and $\hat{F}_n$ is the Kaplan-Meier estimator. A combined-Log-rank test-statistic is computed as $S_n = \boldsymbol{Z}^\intercal \hat{\Sigma}^+ \boldsymbol{Z}$, where $\hat{\Sigma}$ is the empirical covariance matrix of $\boldsymbol{Z}$, and Σ^+ represents the pseudo-inverse of Σ. Under some regularity conditions it is shown that $S_n \to \chi_k^2$ as n grows to infinity.

Implementation: We consider four different functions for the combined Log-rank test, denoted as "WLR" in the tables. These functions correspond to i) the constant weight function, $\omega_1(t) = 1$, which weights all time points equally, ii) an early weight function, $\omega_2(t) = t(1-t)^3$, which gives more weights to departures of the null at early times, iii) a central weight function, $\omega_3(t) = (1-t)t$, giving more weight at central times, iv) and a crossing weight function $\omega_4(t) = 1-2t$, which has a sign switch at $t = 1/2$.

5.4 Kernel test: In the approach of [3], a modified Kaplan-Meier estimate of the density was used for goodness of fit testing. As noted already in [12], however, the procedure suffers from low test power at reasonable sample sizes, which we have also confirmed independently (see Appendix, Section 2). For this reason, our experiments in the next section will focus on the chi-squared and log-rank approaches.

Implementation Code by the authors.

6 Experiments

For our MMD-based test, we choose the kernel to be $K_l(x,y) = \exp\{-(x-y)^2/l^2\}$. For the length-scale parameter, we either use $l = 1$ (for simple settings), or (for more complex settings) $\hat{l}_n$ computed by taking the median of the pair-wise differences of survival times, without making a distinction between censored or uncensored time points. For the estimation of the rejection regions, we implement wild bootstrap as described in Algorithm 1. In particular, we consider three different sets of random variables $\{\mathcal{W}_i\}_{i=1}^n$: i)$\mathcal{W}_i \overset{i.i.d.}{\sim} N(0,1)$, ii) $\mathcal{W}_i \sim Multinomial(n, 1/n, \ldots, 1/n) - 1$ and iii) $\mathcal{W}_i \overset{i.i.d.}{\sim} Rademacher$. Each test is denoted by "MW1", "MW2" and "MW3" in the tables.

In all our experiments, we consider the null hypothesis to be $H_0 : \lambda(t) = 1$, or alternatively $H_0 : S(t) = e^{-t}$. Then, as the data for different experiments is very similar under the null (we just vary the censoring distribution), we only show results in the main body for the estimated Type-I error for our first experiment: parallel hazards. These results are presented in Tables 1. (See sections 3,4 and 5 of the supplementary material for the remaining tables)

Type-I error; Censoring 30%; Fixed length-scale 1							
$\boldsymbol{\alpha}$	MW1	MW2	MW3	Pearson	LR1	LR2	WLR
Sample size $n = 30$							
10 %	10.10	11.70	10.20	14.50	11.15	11.10	19.65
5 %	5.10	6.55	4.85	8.70	5.90	6.60	13.65
1 %	1.15	2.05	1.10	3.05	1.60	1.30	6.30
Sample size $n = 200$							
10 %	9.35	9.55	9.45	9.80	10.05	9.80	11.40
5 %	4.90	4.85	4.75	5.65	5.10	5.10	6.85
1 %	1.15	1.35	1.20	1.30	1.25	1.30	1.80

Table 1: Estimated Type-I error, $\alpha \in \{10\%, 5\%, 1\%\}$. Sample size $n = 30, 200$. Censoring percentage 30%. Fixed-length-scale 1

From table 1, we can observe that all the tests achieve the correct level for larger sample sizes, i.e. $n = 200$. For small sample sizes, i.e. $n = 30$, the Pearson test and the combined Log-rank test have a wrong level (in red in the tables); thus, the measured performance of these tests under the alternative is not meaningful. This may be due to incomplete convergence of the test statistics, since the associated thresholds are based on asymptotic results. In the case of the Pearson test, the test-statistic uses a plug-in estimator $\hat{G}$ of the censoring distribution G, making the testing procedure more complex. Similarly, for the combined Log-rank tests (WLR) the more weight functions we consider, the more complex is our testing procedure and thus, the more data we need.

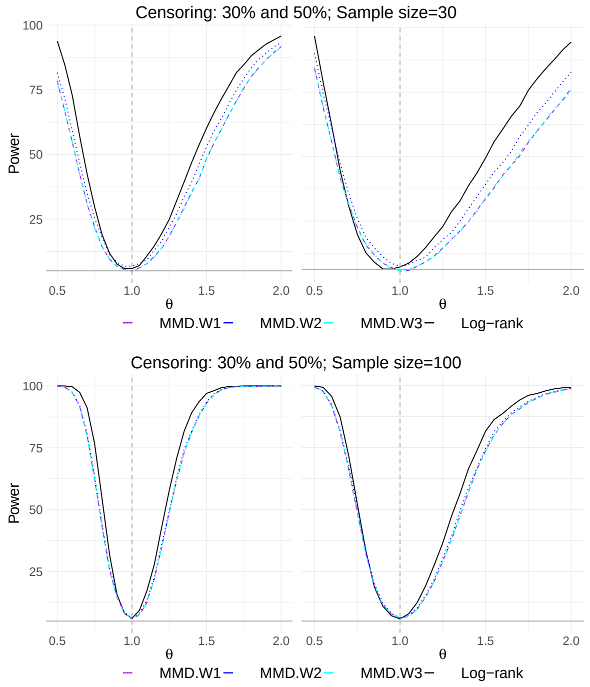

Figure 1: Power related to the family of parallel hazard alternatives $\lambda_\theta(t) = \theta$. The null is recovered for $\theta = 1$.

6.1 Proportional hazard functions

In this experiment, we consider testing the family of alternatives $\lambda_\theta(t) = \theta$ where $\theta \in \{0.5 + 0.05k; k = 1, \ldots, 30\}$ against the null hypothesis $H_0 : \lambda(t) = 1$. We consider the censoring distribution $G(t) = 1 - e^{-\gamma t}$, where γ is such that it generates a censoring percentage of 30% and 50% for each combination of alternative. The level of the test is fixed at $\alpha = 0.05$, and we consider a sample size $n \in \{30, 50, 100, 200\}$. Each experiment considers $N = 2000$ independent repetitions.

It is a well-known that the log-rank test is the most powerful against proportional hazards alternatives. The aim of our first experiment is thus to compare our test performance to a test that we know is optimal. Results are shown in Figure 1.

Overall, our test performs strongly (across all wild bootstrap distributions), despite being more general: for the case of 30 data point our test is quite competitive, and when we observe 100 data points, the

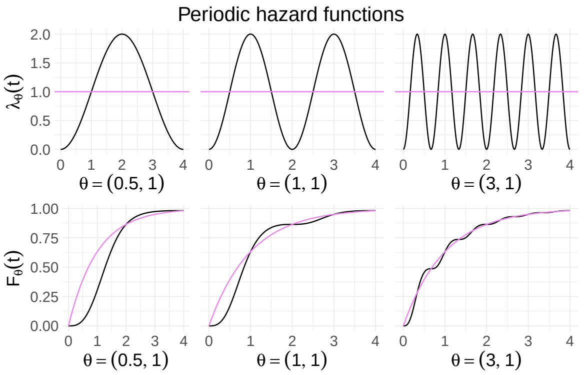

Figure 2: Top row: hazard function $\lambda_\theta(t) = 1 - \cos(\theta_1 \pi t)$ for $\theta_1 \in \{0.5, 1, 3\}$. Bottom row: corresponding cumulative distribution. In pink: exponential distribution (null).

tests behave almost equally. Recall that $\theta = 1$ represents the null hypothesis, so the closer to 1 we are, the harder it is for the test to discriminate from the null. See Section 3 of the supplementary material for a table with the numerical values, more sample sizes and combination of parameters, and other competitors.

6.2 Time-dependent hazard functions

In this section, we consider time-dependent hazard alternatives, which describe a risk that varies over time. Examples include clinical treatment that becomes less effective over time, or seasonal trends in selling patterns. We consider two particular instances of time-dependent hazard functions: i) periodic hazard functions and ii) Weibull hazard functions.

Periodic hazards: Periodic hazard functions arise in scenarios exhibiting periodic patterns, including failure of machines in industrial life-testing, consumer behaviour, and labour market participation among other scenarios: see [26] for discussion.

We consider the family of hazards functions given by $\lambda_\theta(t) = 1 - \theta_2 \cos(\theta_1 \pi t)$, with $\theta = (\theta_1, \theta_2)$ such that $\theta_2 < 1$ and $\theta_1 \in \mathbb{R}$. The null hypothesis is given by $H_0 : \lambda(t) = 1$ (equivalently, $\theta_2 = 0$), and we consider the alternatives $\lambda_\theta(t)$ with $\theta_2 = 1$ and $\theta_1 \in \{0.5, 0.7, 0.9, 1, 1.5, 3\}$. In Figure 2, we show examples of hazards and their corresponding c.d.f. for different parameters of θ. Notice that as the frequency parameter θ_1 increases, the alternative distribution functions F_θ appears closer to the null distribution function $F_0(t) = 1 - \exp\{-t\}$, shown in pink.

In tables 2 and 3 we show the results of our experiments for sample size $n = 30$ and $n = 200$, respectively. Our test strongly outperforms the other tests, and is able to discriminate the alternative from the null in the first two cases, while in the third case ($\theta = (3, 1)$) it has the better overall result for both sample sizes. These results also confirm that periodic hazards with high frequencies present a more challenging task. We note moreover from Table 2 that Pearson and WLR tests (in red) do not have the correct level for 30 samples (see Section 4 of the supplementary material), and thus their reported power might be over-optimistic. In Section 4 of the supplementary material we include more sample sizes and more parameters, which achieve similar results.

Periodic hazard function

Sample size n=30; Cens. param. $\gamma = 1/2$; Adaptive length-scale

$\boldsymbol{\alpha}$	MW1	MW2	MW3	Pearson	LR1	LR2	WLR
$\theta = (0.5, 1)$							
10 %	99.95	99.95	99.95	99.40	89.65	99.95	62.15
5 %	99.95	99.95	99.95	96.95	74.50	99.80	44.60
1 %	99.75	99.80	99.75	76.80	31.25	94.20	17.30
$\theta = (1, 1)$							
10 %	98.60	99.05	98.60	87.35	16.20	62.80	31.35
5 %	95.50	96.25	95.30	73.60	8.05	39.55	21.55
1 %	82.00	83.75	82.40	42.05	0.80	6.55	10.30
$\theta = (3, 1)$							
10 %	23.40	25.80	23.40	20.00	8.95	5.35	19.35
5 %	13.40	15.65	13.60	12.70	4.55	2.20	13.60
1 %	4.00	5.60	4.10	4.90	0.95	0.25	6.70

Table 2: Power (from 0% to 100%) under different alternatives θ for the Periodic hazards experiment. Sample size 30. Censoring Parameter 1/2. Adaptive length-scale.

Periodic hazard functions

Sample size n=200; Cens. param. $\gamma = 1/2$; Adaptive length-scale

	MW1	MW2	MW3	Pearson	LR1	LR2	WLR
$\theta = (0.5, 1)$							
10 %	100.00	100.00	100.00	100.00	100.00	100.00	100.00
5 %	100.00	100.00	100.00	100.00	100.00	100.00	100.00
1 %	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$\theta = (1, 1)$							
10 %	100.00	100.00	100.00	100.00	83.00	100.00	100.00
5 %	100.00	100.00	100.00	100.00	71.00	100.00	99.95
1 %	100.00	100.00	100.00	100.00	39.95	100.00	99.30
$\theta = (3, 1)$							
10 %	98.20	98.35	98.30	44.70	10.70	21.85	44.35
5 %	89.75	89.95	89.85	31.70	4.95	11.40	31.45
1 %	48.90	48.00	47.85	14.20	1.00	2.35	12.85

Table 3: Power (from 0% to 100%) under different alternatives θ for the Periodic hazards experiment. Sample size 200. Censoring Parameter 1/2. Adaptive length-scale.

Weibull hazards: Weibull models are popular in survival analysis and reliability problems where there is either an increasing or decreasing hazard rate. This is due to their flexibility, despite only being parametrised by two values. Weibull hazards have been used to model failure of composite materials, and fracture strength of glass among other examples, see [23].

The Weibull hazard function is given by $\lambda_{\theta_w}(t) = \theta_{w1}/\theta_{w2}(t/\theta_{w2})^{\theta_{w1}-1}$, where $\theta_w = (\theta_{w1}, \theta_{w2}) \in \mathbb{R}_+^2$. In particular, θ_{w2} denotes the scale parameter which

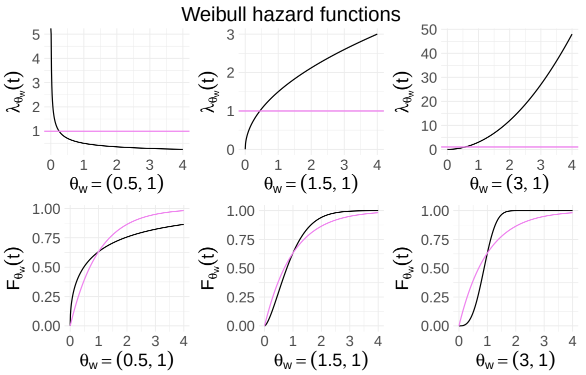

Figure 3: Top row: hazard $\lambda_{\theta_w}(t) = \theta_{w1}(t)^{\theta_{w1}-1}$, for $\theta_{w1} \in \{0.5, 1.5, 3\}$. Bottom row: The corresponding cdf.

has an proportional effect on the hazard function, and θ_{w1} is the shape parameter.

For $\theta_{w1} < 1$ we have a decreasing hazard, for $\theta_{w1} > 1$ we have an increasing hazard, and for $\theta_{w_1} = 1$ we recover a constant hazard representing the exponential distribution. The null hypothesis occurs when $\theta_w = (1, 1)$. In figure 3 we show three different Weibull hazard functions (and their corresponding survival functions) with three types of behaviour: decreasing hazard, increasing concave, and increasing convex.

In tables 4 and 5 we show the result of our experiment for sample sizes $n = 30$ and 200. Our test again yields the best performance. It also noticeable that Log-rank performs extremely poorly, which is a well-known failure mode for such hazards.

Weibull hazard functions							
Sample size $n = 30$; Censoring 30%; Adaptive length-scale							
	MW1	MW2	MW3	Pearson	LR1	LR2	WLR
$\theta_w = (0.5, 1)$							
10 %	67.20	70.70	67.30	70.25	28.55	71.30	70.55
5 %	52.30	56.10	52.15	60.15	20.45	65.15	63.70
1%	21.95	29.40	21.95	42.05	9.60	51.80	52.55
$\theta_w = (1.5, 1)$							
10 %	45.35	48.20	45.50	52.20	3.65	8.45	33.40
5 %	32.50	35.80	32.05	40.80	1.55	3.45	24.10
1%	12.70	15.05	12.35	21.10	0.05	0.25	11.55
$\theta_w = (3, 1)$							
10 %	100.00	100.00	100.00	100.00	0.05	43.40	99.55
5 %	100.00	100.00	100.00	100.00	0.00	16.30	99.00
1%	99.80	99.95	99.85	99.90	0.00	0.45	96.20

Table 4: Power (from 0% to 100%) under different alternatives θ_{w1} for the Weibull hazards experiment. Sample size 30. Censoring percentage 30%. Adaptive length-scale.

We remark that we should again treat the Pearson and WLR results with caution for sample size 30 as they given an incorrect Type-1 error. See tables in Section 5 of the supplementary material to see the later and more experiments with different parameters.

Weibull hazard functions							
Sample size $n = 200$; Censoring 30%; Adaptive length-scale							
	MW1	MW2	MW3	Pearson	LR1	LR2	WLR
$\theta_w = (0.5, 1)$							
10 %	100.00	100.00	100.00	100.00	46.05	99.95	99.95
5 %	100.00	100.00	100.00	100.00	36.80	99.90	99.95
1%	100.00	100.00	100.00	100.00	20.90	99.30	99.60
$\theta_w = (1.5, 1)$							
10 %	99.95	100.00	99.95	100.00	6.10	72.65	98.85
5 %	99.90	99.85	99.90	100.00	2.75	56.50	96.85
1%	98.70	98.95	98.65	99.80	0.45	22.40	88.35
$\theta_w = (3, 1)$							
10 %	100.00	100.00	100.00	100.00	2.15	100.00	100.00
5 %	100.00	100.00	100.00	100.00	0.20	100.00	100.00
1%	100.00	100.00	100.00	100.00	0.00	100.00	100.00

Table 5: Power (from 0% to 100%) under different alternatives θ_{w1} for the Weibull hazards experiment. Sample size 200. Censoring percentage 30%. Adaptive length-scale.

7 Discussion

We have presented a novel testing procedure for goodness-of-fit for right-censored data, based on the MMD distance between a transformation of the observed variables and the uniform distribution. Being based on kernels, it is not necessary to specify features in advance (as for the weighted log-rank test): rather, we take advantage of the infinite dictionary of features provided implicitly by the kernel. Our approach has several advantages: First, it is simple to implement, since we only need to be able to evaluate the distribution F_0 in the survival times T_i to generate the data $\{F_0(T_i), \Delta_i\}$, and we do not need to know/evaluate F_0^{-1}. Second, the U-statistic kernel $J : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$ of equation (4) is distribution free, and need be computed/tabulated only once. Third, being a U-statistic, the asymptotic analysis is straightforward, as is the bootstrap approach for the test threshold.

We emphasize that extensions to other type of censoring (left and interval) are straightforward, as our test depends on censoring only through the estimate $\tilde{F}$, in which the mass of a censored interval is distributed uniformly over such an interval.

Further improvements in the performance of our test might be achieved by a better choice of kernel function for the problem at hand. In the case of the maximum mean discrepancy on uncensored data, test power is improved by choosing a kernel to optimise the ratio of the statistic to its variance [31]. Adaptive linear-time test statistics may also be constructed for two-sample [19] and Stein goodness-of-fit [20] tests, where the features are again chosen to maximise test power. It would be of interest to extend these ideas to the present setting.

Acknowledgements: Tamara Fernández was supported by Biometrika Trust.

References

[1] Odd O. Aalen, Ø rnulf Borgan, and Håkon K. Gjessing. *Survival and event history analysis*. Statistics for Biology and Health. Springer, New York, 2008. A process point of view.

[2] Michael G. Akritas. Pearson-type goodness-of-fit tests: the univariate case. *J. Amer. Statist. Assoc.*, 83(401):222–230, 1988.

[3] D. Bagkavos, D. Ioannides, and A. Kalamatianou. A goodness of fit test for the survival function under random right censoring. *Electronic Journal of Statistics*, 7:2550–2576, 2013.

[4] K. Balasubramanian, T. Li, and M. Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. Technical Report 1709.08148v1, arxiv, 2017.

[5] L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35:339–348, 1988.

[6] Arne Bathke, Mi-Ok Kim, and Mai Zhou. Combined multiple testing by censored empirical likelihood. *Journal of Statistical Planning and Inference*, 139(3):814–827, 2009.

[7] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[8] Michael Brendel, Arnold Janssen, Claus-Dieter Mayer, and Markus Pauly. Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics*, 41(3):742–761.

[9] Norman E Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57, 1975.

[10] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *ICML*, pages 2606–2615, 2016.

[11] Herold Dehling and Thomas Mikosch. Random quadratic forms and the bootstrap for u-statistics. *Journal of Multivariate Analysis*, 51(2):392–413, 1994.

[12] Marc Ditzhaus and Sarah Friedrich. More powerful logrank permutation tests for two-sample survival data. *arXiv preprint arXiv:1807.05504*, 2018.

[13] 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.

[14] A. Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13:723–773, 2012.

[15] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. pages 609–616. MIT Press, Cambridge, MA, 2008.

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

[17] Myles Hollander and Edsel A Pena. A chi-squared goodness-of-fit test for randomly censored data. *Journal of the American Statistical Association*, 87(418):458–463, 1992.

[18] Myles Hollander and Frank Proschan. Testing to determine the underlying distribution using randomly censored data. *Biometrics*, pages 393–401, 1979.

[19] W. Jitkrittum, Z. Szabo, K. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In *NIPS*, 2016.

[20] W. Jitkrittum, W. Xu, Z. Szabo, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. Technical report, NIPS, 2017.

[21] E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481, 1958.

[22] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

[23] Chin-Diew Lai, D.N. Murthy, and Min Xie. *Weibull Distributions and Their Applications*, pages 63–78. Springer London, London, 2006.

[24] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML*, pages 276–284, 2016.

[25] Daniel P Mihalko and David S Moore. Chi-square tests of fit for type ii censored data. *The Annals of Statistics*, pages 625–644, 1980.

[26] Ulrich Pötter, Kai Kopperschmidt, and AC Nielsen. Covariate effects in periodic hazard rate models, 2001.

[27] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *Annals of Statistics*, 41(5):2263–2702, 2013.

[28] R. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

[29] A. Smola, A. Gretton, L. Song, and B. Schoelkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory: 18th International Conference*, volume LNAI4754, pages 13–31, Berlin/Heidelberg, 2007. Springer-Verlag.

[30] B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schoelkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

[31] D. Sutherland, H-Y Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR*, 2017.

[32] G. Székely and M. Rizzo. A new test for multivariate normality. *J. Multivariate Anal.*, 93:58–80, 2005.