Kernelized Stein Discrepancy Tests of Goodness-of-fit for Time-to-Event Data

Tamara Fernández [1] **Arthur Gretton** [1] **Nicolás Rivera** [2] **Wenkai Xu** [1]

Abstract

Survival Analysis and Reliability Theory are concerned with the analysis of time-to-event data, in which observations correspond to waiting times until an event of interest, such as death from a particular disease or failure of a component in a mechanical system. This type of data is unique due to the presence of censoring, a type of missing data that occurs when we do not observe the actual time of the event of interest but instead we have access to an approximation for it given by random interval in which the observation is known to belong. Most traditional methods are not designed to deal with censoring, and thus we need to adapt them to censored time-to-event data. In this paper, we focus on non-parametric goodness-of-fit testing procedures based on combining the Stein's method and kernelized discrepancies. While for uncensored data, there is a natural way of implementing a kernelized Stein discrepancy test, for censored data there are several options, each of them with different advantages and disadvantages. In this paper, we propose a collection of kernelized Stein discrepancy tests for time-to-event data, and we study each of them theoretically and empirically; our experimental results show that our proposed methods perform better than existing tests, including previous tests based on a kernelized maximum mean discrepancy.

1. Introduction

An important topic of study in statistics is the distribution of times to a critical event, otherwise known as survival times: examples include the infection time from a disease (Andersen et al., 2012; Mirabello et al., 2009); the death time of a patient in a clinical trial (Collett, 2015; Biswas et al., 2007); or the possible re-offending times for released criminals (Chung et al., 1991). Survival data are frequently subject to censoring: the time of interest is not observed, but rather a bound on it. The most common scenario studied is *right censoring*, where a lower bound on the survival time is observed, for instance, a patient might leave a clinical trial before it is completed, meaning that we only obtain a lower bound on the time of death (the definitions and terminologies for the survival analysis setting will be provided in Section 2).

We address the setting where a model of survival times is proposed, and it is desired to test this model against observed data in the presence of censoring: this is known as *goodness-of-fit* testing. When departures from the model follow a known parametric family, a number of classical tests are available, being the most popular in practice the Log-rank test (Hollander & Proschan, 1979), and its generalization, the weighted Log-rank test (Brendel et al., 2014). For an overview of these and other methods we refer the reader to (Klein & Moeschberger, 2006)

In the event of more general departures from the null, kernel methods may be used to construct a powerful class of non-parametric tests to detect a greater range of alternative scenarios. For the uncensored case, a popular class of kernel goodness-of-fit tests utilize Stein's method (Barbour & Chen, 2005; Chen et al., 2010; Ley et al., 2017; Gorham & Mackey, 2015) to develop a test statistic (Liu et al., 2016; Chwialkowski et al., 2016; Gorham & Mackey, 2017; Jitkrittum et al., 2017), which can be computed even when the model is known only up to normalization. In this paper we consider the particular case of kernel Stein discrepancies (KSDs) which are described in Section 2. While an alternative strategy would be simply to run a two-sample test using samples from the model, using for instance the maximum mean discrepancy (MMD) (Gretton et al., 2012), Stein tests are more computationally efficient (no additional sampling is needed), and can take advantage of model structure to achieve better test power. KSD tests have been extended to various settings such as discrete variable models (Yang et al., 2018), point process (Yang et al., 2019), latent variable models (Kanagawa et al., 2019), and directional data (Xu & Matsuda, 2020).

[1] Gatsby Computational Neuroscience Unit, University College London, United Kingdom [2] Department of Computer Science and Technology, University of Cambridge, United Kingdom. Correspondence to: Tamara Fernández <t.a.fernandez@ucl.ac.uk >.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

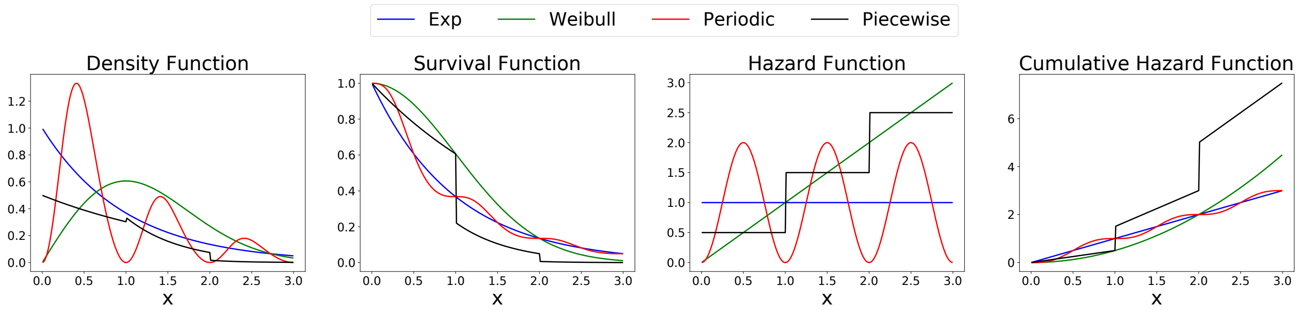

Figure 1. Example functions related to survival analysis.

In the present work, we propose to generalize Stein goodness-of-fit tests to the setting of survival analysis with right-censored data. In Section 3, we introduce three separate approaches to constructing a Stein operator in the presence of censoring: the first, the *Survival Stein Operator*, is the most direct generalization of the Stein operator used in the uncensored KSD test. The second, the *Martingale Stein Operator*, uses a different construction, based on a classical martingale studied in the survival analysis literature. The third, the *Proportional Stein Operator*, is designed for composite null hypotheses: in this case, the *hazard function* (that is, the instantaneous probability of an event at a given time, conditioned on survival to that time) is known only up to a constant of proportionality. For instance, we may wish to use a constant hazard as the null hypothesis, without specifying in advance the value of the constant.

The rest of the paper is structured as follows: in Section 4, we construct kernel statistics of goodness-of-fit, based on each of the operators previously introduced. We characterize the asymptotics of each statistic in Section 5. We find that in order to guarantee convergence in distribution under the null, the kernel statistic based on the Survival Stein Operator requires more restrictive conditions than the statistic built on the Martingale Stein Operator. In other words, the straightforward extension of the uncensored test is in fact the more restrictive approach of the two. Stronger assumptions again are required in obtaining convergence in distribution for the Proportional Stein Operator statistic, which should come as no surprise, given that the null is now an entire model class. For each statistic, we propose a wild bootstrap approach to obtain the test threshold. Empirical studies and results are presented in Section 6, where we compare with a recent state-of-the-art non-parametric test for censored data (Fernandez & Gretton, 2019) based on the MMD, which has been shown to outperform classical tests. For challenging cases, our Stein tests surpass the MMD test.

2. Background

Kernel Stein Discrepancy We briefly review the kernel Stein discrepancy (KSD) in the absence of censoring (Chwialkowski et al., 2016; Liu et al., 2016), which is inspired from (Gorham & Mackey, 2015; Ley et al., 2017). Let f_0 be a smooth probability density on $\mathbb{R}$. For a bounded smooth function $\omega : \mathbb{R} \to \mathbb{R}$, the Stein operator $\mathcal{T}_0$ is

$$\mathcal{T}_0\,\omega(x) = \omega(x)(\log f_0(x))' + \omega'(x), \tag{1}$$

where $'$ denotes derivative w.r.t x. Since f_0 vanishes at the boundary and ω is bounded, integration by parts on $\mathbb{R}$ results in Stein's Lemma,

$$\mathbb{E}_0[\mathcal{T}_0\omega] = \int (\mathcal{T}_0\omega)(x) f_0(x) = 0,$$

under some regularity conditions. Since the Stein operator $\mathcal{T}_0$ depends on the density f_0 only through the derivative of $\log f_0$, it does not involve the normalization constant of f_0, which is a useful property for dealing with unnormalized models (Hyvärinen, 2005).

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) on $\mathbb{R}$ with associated kernel K. By using the Stein operator above, the kernel Stein discrepancy (KSD) (Chwialkowski et al., 2016; Liu et al., 2016) between two densities f_X and f_0 is defined as

$$\mathrm{KSD}(f_X \| f_0) = \sup_{\omega \in B_1(\mathcal{H})} \mathbb{E}_X[\mathcal{T}_0\omega], \tag{2}$$

where $B_1(\mathcal{H})$ denotes the unit ball of $\mathcal{H}$, and $\mathbb{E}_X$ denotes the expectation w.r.t. the density f_X. It is easy to see that $\mathrm{KSD}(f_X, f_0) \geq 0$ and that $\mathrm{KSD}(f_X, f_0) = 0$ for $f_X = f_0$. Moreover, under some regularity conditions, we have that $\mathrm{KSD}(f_X, f_0) = 0$ if and only if $f_X = f_0$ (Chwialkowski et al., 2016).

By using standard properties of RKHSs, we can conveniently write $\mathrm{KSD}(f_X, f_0)$ as

$$\mathrm{KSD}^2(f_X \| f_0) = \mathbb{E}_{x,y \sim f_X}[h_0(x,y)], \tag{3}$$

where $h_0(x,y) =$

$$\langle \log f_0(x)' K(x,\cdot) + K'(x,\cdot), \log f_0(y)' K(y,\cdot) + K'(y,\cdot)\rangle,$$

with $\langle \cdot, \cdot \rangle$ denoting the inner product of $\mathcal{H}$.

Censored Data Let $(X_1,\dots,X_n) \overset{\text{i.i.d.}}{\sim} F_X$ be the survival times, which are non-negative real-valued random variables of interest, and let $(C_1,\dots,C_n) \overset{\text{i.i.d.}}{\sim} F_C$ be another collection of non-negative random variables called censoring times. In this work, we assume the non-informative censoring setting, where the censoring times are independent of the survival times. The data we observe correspond to (T_i,Δ_i) where $T_i = \min\{X_i, C_i\}$ and $\Delta_i = \mathbb{1}_{\{X_i \le C_i\}}$. We can imagine that X_i is the time of interest (death of a patient) and C_i is the time a patient leaves the study for some other reason, thus, for some patients we observe their actual death time, whereas for others we just observe a lower bound (the time they left the study). Δ_i indicates if we are observing X_i or C_i.

We denote by f_T, f_X and f_C, the respective density functions associated with the random variables T, X and C. Similarly, we denote by F_T, F_X and F_C, the respective cumulative distribution functions; and by $S_T = 1 - F_T$, $S_X = 1 - F_X$ and $S_C = 1 - F_C$, the survival functions. An important element in survival analysis is the hazard function which represents the instantaneous risk of dying at a given time (as X usually refers to a death time). Given a distribution with density f_X and survival function S_X, the hazard function $\lambda_X(x)$ is given by $f_X(x)/S_X(x)$, which can be seen as the density at x of a random variable X conditioned on the event $\{X \ge x\}$. The corresponding cumulative hazard function is defined as $\Lambda_X(x) = \int_0^x \lambda_X(t)dt$. A useful feature of the hazard function is that there is a one-to-one relation between hazard and density functions through the relation $S_X(x) = e^{-\Lambda_X(x)}$. For the random variables T and C, we denote by λ_T and λ_C their respective hazard functions, and by Λ_T and Λ_C, their cumulative hazards functions. As a remark, every continuous non-negative function $\lambda : \mathbb{R}_+ \to \mathbb{R}$ can be a hazard function, as long as $\int_{\mathbb{R}_+} \lambda(t)dt = \infty$, thus, describing hazards is much easier than describing densities, as we do not need to worry about normalization constants. Examples of corresponding functions for different models are displayed in Figure 1.

As observations come as pairs (T_i, Δ_i), it is convenient to consider the joint measure μ on $\mathbb{R}_+ \times \{0,1\}$ induced by the pair (T,Δ). We write μ_X to denote the measure μ when the survival times of interest X_i are generated according to f_X, and μ_0 if they are generated under f_0 (i.e., under the null). Note that μ_X and μ_0 also depend on f_C, however we don't make this dependence explicit, since for goodness-of-fit we only care about f_0 and f_X.

Finally, for any function ϕ, the following identities hold, which the reader should keep in mind for later use:

$$\mathbb{E}_X[\Delta\phi(T)] = \int_0^\infty \phi(s) f_X(s) S_C(s) ds, \tag{4}$$

$$\mathbb{E}_X[(1-\Delta)\phi(T)] = \int_0^\infty \phi(s) f_C(s) S_X(s) ds. \tag{5}$$

Here $\mathbb{E}_X = \mathbb{E}_{\mu_X}$ means that we are taking expectation w.r.t. $(T,\Delta) \sim \mu_X$. Similarly, we write $\mathbb{E}_0$ to indicate $(T,\Delta) \sim \mu_0$ (under the null hypothesis).

3. Stein Operator for Censored Data

In this section, we describe a set of Stein operators for censored data. We denote by Ω the set of functions $\mathbb{R}_+ \times \{0,1\} \to \mathbb{R}$, and recall that μ_0 is the measure induced by data (T,Δ) under the null hypothesis.

Definition 1. *Let $\mathcal{H} \subseteq L^2(f_0)$. We call $\mathcal{T}_0 : L^2(f_0) \to \Omega$ a Stein operator for $\mathcal{H}$ if for each $\omega \in \mathcal{H}$*

$$\mathbb{E}_0\left[(\mathcal{T}_0\omega)(T,\Delta)\right] = 0. \tag{6}$$

An interesting technical point is that our operator takes functions $\omega : \mathbb{R}_+ \to \mathbb{R}$ and maps them to Ω. The idea behind having these two spaces is that while our data of interest is a time (hence the space $\mathcal{H}$ of functions $\mathbb{R}_+ \to \mathbb{R}$), we actually observe pairs (T_i, Δ_i), hence we need functions in Ω.

We choose the general class $\mathcal{H}$ to be an RKHS. We assume that $\mathcal{H}$ contains only differentiable and bounded functions, and that if $\omega \in \mathcal{H}$ then $\omega' \in \mathcal{H}$. These requirements are not restrictive and most of the standard kernels in the literature generate RKHSs with these properties, including the Gaussian kernel (furthermore, we can avoid this restriction, but we keep it as it is convenient for the exposition of the paper). Further properties of $\mathcal{H}$ will be imposed if needed in particular cases.

3.1. Survival Stein Operator

Observe that $T_i = X_i$ if and only if $\Delta_i = 1$. One might be tempted to use only the uncensored observations to approximate $\int_0^\infty (\mathcal{T}_0\omega)(x) f_0(x)dx$ (where $\mathcal{T}_0$ is the standard Stein operator in (1)) by computing

$$\frac{1}{n}\sum_{i=1}^n \Delta_i (\mathcal{T}_0\omega)(T_i) = \frac{1}{n}\sum_{i=1}^n \Delta_i (\mathcal{T}_0\omega)(X_i),$$

however, this sum does not converge to $\int_0^\infty (\mathcal{T}_0\omega)(x) f_0(x)dx$ as the term Δ_i introduces bias due to censoring. Indeed, such an empirical average converges to $\int_0^\infty (\mathcal{T}_0\omega)(x) S_C(x) f_X(x)$. To account for this bias we redefine $\mathcal{T}_0 : \mathcal{H}^{(s)} \to \Omega$ as

$$(\mathcal{T}_0\omega)(x,\delta) = \delta \frac{(\omega(x) S_C(x) f_0(x))'}{S_C(x) f_0(x)} + \omega(0) f_0(0). \tag{7}$$

Here we write $\mathcal{H}^{(s)}$ instead of $\mathcal{H}$ whenever we assume that the additional condition is satisfied,

$$\int_{\mathbb{R}+} |\left(\omega(x)S_C(x)f_0(x)\right)'|dx < \infty, \quad \forall \omega \in \mathcal{H}, \qquad (8)$$

which guarantees that the operator is well-defined. Notice that $\omega(0)f_0(0)$ in equation (7) appears since we do not necessarily assume a vanishing boundary at 0.

Under the null hypothesis, $(T_i, \Delta_i) \sim \mu_0$, it holds that

$$\frac{1}{n}\sum_{i=1}^{n}(\mathcal{T}_0\omega)(T_i,\Delta_i) \to \mathbb{E}_0[(\mathcal{T}_0\omega)(T,\Delta)] \qquad (9)$$

as the number of data points tends to infinity, and $\mathbb{E}_0[(\mathcal{T}_0\omega)(T,\Delta)] = 0$ due to Equation (4) and the fact that

$$\int_{\mathbb{R}+}(\omega(x)S_C(x)f_0(x))'dx + \omega(0)f_0(0) = 0, \qquad (10)$$

which is proved using integration by parts. Notice that in this argument we use that $\mathcal{H}^{(s)}$ only contains bounded functions, allowing us to get rid of the boundary at infinity.

The operator $\mathcal{T}_0$ can be seen as a natural extension of the Stein operator (Gorham & Mackey, 2015) to censored data. Observe that in the uncensored case, $S_C(x) \equiv 1$ recovers the standard Stein operator.

Unfortunately, in the goodness-of-fit setting, we only have access to the null distribution $f_0(x)$ but not to the censoring distribution $f_C(x)$, thus $S_C(x)$ needs to be estimated. The standard estimator for S_C is the Kaplan-Meier estimator (Kaplan & Meier, 1958) which is very data inefficient, leading to an unsatisfactory testing procedure.

To bypass the approximation of S_C we define the survival Stein operator $\mathcal{T}_0^{(s)} : \mathcal{H}^{(s)} \to \Omega$ as

$$\begin{aligned}(\mathcal{T}_0^{(s)}\omega)(x,\delta) = \delta\omega'(x) + \frac{\lambda_0'(x)}{\lambda_0(x)}\delta\omega(x) \\ - \lambda_0(x)\omega(x) + \lambda_0(0)\omega(0) \qquad (11)\end{aligned}$$

Proposition 2. *Consider $\mathcal{T}_0$ and $\mathcal{T}_0^{(s)}$ defined in equations* (7) *and* (11)*, respectively. Let $(T,\Delta) \sim \mu_0$. Then*

$$\mathbb{E}_0[(\mathcal{T}_0^{(s)}\omega)(T,\Delta)] = \mathbb{E}_0[(\mathcal{T}_0\omega)(T,\Delta)] = 0, \quad \forall\omega \in \mathcal{H}^{(s)}.$$

The previous proposition says that if the data we observed was generated from μ_0 then the expectation of the operators $\mathcal{T}_0$ and $\mathcal{T}_0^{(s)}$ are equal for each function in $\mathcal{H}^{(s)}$. However, the relation between $\mathcal{T}_0$ and $\mathcal{T}_0^{(s)}$ is stronger than merely equality in expectation, indeed, under a slightly stronger condition on the form of the distribution f_0 and f_C we get the following result, which is proven in Appendix A.

Proposition 3. *Assume that*

$$\int_0^\infty(\lambda_C(x)+\lambda_0(x))f_C(x)f_0(x) < \infty, \qquad (12)$$

then, under the null hypothesis, i.e. $(T_i,\Delta_i) \sim \mu_0$, we have that, as the number of data points tends to infinity,

$$\sup_{\omega\in B_1(\mathcal{H})}\frac{1}{n}\sum_{i=1}^{n}(\mathcal{T}_0^{(s)}\omega)(T_i,\Delta_i) - (\mathcal{T}_0\omega)(T_i,\Delta_i) \xrightarrow{\mathbb{P}} 0.$$

To better understand the survival Stein operator, we interpret the proposed Stein operator by making connections to the Stein operator used in the uncensored case.

A careful computation gives the following equivalent expression for the expectation of $(\mathcal{T}_0^{(s)}\omega)(T,\Delta)$ for $(T,\Delta) \sim \mu_X$:

$$\mathbb{E}_X[(\mathcal{T}_0^{(s)}\omega)(T,\Delta)] = \mathbb{E}_X\left[\omega(T)\Delta\left(\log\frac{f_0(T)}{f_X(T)}\right)'\right]$$
$$-\,\mathbb{E}_X\left[\omega(T)(1-\Delta)(\lambda_0-\lambda_X)(T)\right] + \omega(0)(\lambda_0-\lambda_X)(0).$$

Here, we can relate the first expectation to uncensored observations: $\Delta = 1$; the second expectation to censored observations: $\Delta = 0$; and the third term describes a shift due to boundary conditions.

The expectation of the uncensored part is equal to

$$\int_0^\infty \omega(x)\left(\log\frac{f_0(x)}{f_X(x)}\right)' S_C(x)f_X(x)dx,$$

which is analogous to what we obtain in the uncensored case, with an additional S_C weighting. If we have no censoring, then $S_C \equiv 1$, recovering the expression found in (Chwialkowski et al., 2016). On the other hand, the expectation of the censored part is equal to

$$\int_0^\infty \omega(x)\left(\frac{S_X(x)}{S_0(x)}f_0(x) - f_X(x)\right)f_C(x)dx,$$

which measures the discrepancy between f_0 and f_X through survival weights, under the measure of censoring f_C. In the absence of censoring, $f_C = 0$ a.e., so this term appears due to the censoring variable. Notice that if differences between f_0 and f_X occur at times t where $S_C(t) = 0$, then no method will detect these differences (the observations at this time are entirely censored).

3.2. Martingale Stein Operator

While the previous approach mimics the classic Stein operator, it has similar drawbacks. Similarly to what we observe in (Chwialkowski et al., 2016) and (Liu et al., 2016), our Stein operator $\mathcal{T}_0^{(s)}$ requires very strong integrability conditions on the involved distribution functions. In our setting, we find, for example condition c.1 in Section 5.1, which involves integrals with respect to hazard functions which are known to satisfy $\int \lambda_0(x)dx = \infty$, leading to a testing procedure with weak theoretical guarantees. While these conditions may hold for some models, it is not hard to find simple examples where they do not hold.

In order to get a more robust test, we exploit a well-known identity in survival analysis, allowing us to deduce a more natural Stein operator. Such an identity is given by

$$\mathbb{E}_0\left[\Delta\phi(T) - \int_0^T \phi(t)\lambda_0(x)dx\right] = 0, \qquad (13)$$

which holds for any function ϕ such that $\mathbb{E}_0(|\phi(T)|) < \infty$ under μ_0 (Aalen et al., 2008). This equality is derived by using a martingale identity that appears in the derivation of classical estimators in survival analysis (see Appendix B).

Assuming $\lambda_0(t) > 0$, we replace $\phi = \omega'/\lambda_0$ in (31) to get

$$\mathbb{E}_0\left[\Delta\frac{\omega'(T)}{\lambda_0(T)} - (\omega(T) - \omega(0))\right] = 0.$$

Define the martingale Stein Operator $\mathcal{T}_0^{(m)} : \mathcal{H}^{(m)} \to \Omega$ as

$$(\mathcal{T}_0^{(m)}\omega)(x,\delta) = \delta\frac{\omega'(x)}{\lambda_0(x)} - (\omega(x) - \omega(0)) \qquad (14)$$

where we write $\mathcal{H}^{(m)}$ instead of $\mathcal{H}$ whenever $\mathcal{H}$ satisfies

$$\int_{\mathbb{R}+}\left|\frac{\omega'(x)}{\lambda_0(x)}\right| S_C(x)f_0(x)dx < \infty, \quad \forall\omega \in \mathcal{H}. \qquad (15)$$

From its definition, it is clear that $\mathbb{E}_0[(\mathcal{T}_0^{(m)}\omega)(T,\Delta)] = 0$. Note that, by the definition of the hazard functions, condition (15) is equivalent to

$$\int_{\mathbb{R}+}|\omega'(x)|\, S_C(x)S_0(x)dx < \infty, \quad \forall\omega \in \mathcal{H}, \qquad (16)$$

which holds true if the kernel is bounded (recall we assume that $\omega' \in \mathcal{H}$), therefore, compared to $\mathcal{T}_0^{(s)}$, the testing procedure associated to $\mathcal{T}_0^{(m)}$ has very strong theoretical guarantees. Indeed, we observe that condition c.2 in Section 5.1 is much simpler to satisfy because, this time, we consider integrals with respect to the inverse of the hazard function.

Model-Free Implementation: Inspired by the test of uniformity via a F_0 transformation (Fernandez et al., 2019), we transform our data $U_i = F_0(T_i)$ to generate pairs (U_i, Δ_i). Notice that since F_0 is monotone $U_i = F_0(T_i) = \min\{F_0(X_i), F_0(C_i)\}$, thus Δ_i remains consistent. Under this transformation, testing the null hypothesis is equivalent to test whether $F_0(X_i)$ is distributed as a uniform random variable, thus, in this setting, $\lambda_0 = \lambda_{\mathcal{U}} = \frac{1}{1-x}$ and

$$(\mathcal{T}_0^{(m)}\omega)(u,\delta) = \delta\omega'(u)(1-u) - \omega(u) + \omega(0)$$

for $u = F_0(x)$ (notice that $F_0(0) = 0$). It will be shown in the experiments that this transformation is beneficial in terms of power performance. Similarly, we can exploit that $\Lambda_0(X) \sim \text{Exp}(1)$ under the null when the model is described via the cumulative hazard function.

3.3. Proportional Stein Operator

In some scenarios, we are interested in the shape of the hazard function up to a multiplicative constant, i.e. $\lambda_0(t) = \gamma\lambda(t)$ where we know $\lambda(t)$ but not the constant γ. The family indexed by γ is called a proportional hazards family and it is one of the key objects of study in Survival Analysis. This object is fundamental because sometimes it is more important to test for qualitative results as "the hazard rate is growing at a constant speed", rather than obtaining precise values of the hazard function. If we only know $\lambda_X(t)$ up to constant and we can ensure that $\omega(0)\lambda(0) = 0$, then we can define a Stein operator based on unnormalized hazard.

In order to define our operator, we assume that

$$\int_{\mathbb{R}_+}|(\omega(x)\lambda_0(x))'|dx < \infty, \quad \text{and}$$
$$\omega(0)\lambda_0(0) = \lim_{x\to\infty}\omega(x)\lambda_0(x) = 0, \quad \forall\omega \in \mathcal{H}. \qquad (17)$$

As usual, we write $\mathcal{H}^{(p)}$ to indicate that $\mathcal{H}$ satisfies property (17). Note that for any function $\omega \in \mathcal{H}^{(p)}$ it holds that

$$\int_0^\infty \frac{(\omega(x)\lambda_0(x))'}{\lambda_0(x)}\lambda_0(x)dx = 0.$$

The integral above can be estimated using the Nelson-Aalen estimator (Nelson, 1972), leading to the statistic

$$\frac{1}{n}\sum_{i=1}^n \frac{(\omega(T_i)\lambda_0(T))'}{\lambda_0(T_i)}\frac{\Delta_i}{Y(T_i)/n},$$

where $Y(t) = \sum_{k=1}^n \mathbb{1}_{\{T_k \geq t\}}$ is the so-called *risk function*, which counts the number of individuals at risk at time t. This suggests the following operator

$$(\widehat{\mathcal{T}_0}^{(p)}\omega)(x,\delta) = \left(\omega'(x) + \frac{\omega(x)\lambda_0'(x)}{\lambda_0(x)}\right)\frac{\delta}{Y(x)/n}. \qquad (18)$$

In the definition above we use the notation $\widehat{\mathcal{T}_0}^{(p)}$ to indicate that, the function $Y(t)$ depends on all data points, hence $\widehat{\mathcal{T}_0}^{(p)}$ can be seen as an empirical estimator of a deterministic operator. Indeed, if $(T_i, \Delta_i) \sim \mu_0$, then $\frac{Y(x)}{n} \to S_C(x)S_0(x)$, which indicates that under the null hypothesis, the operator $\widehat{\mathcal{T}_0}^{(p)}$ is similar to $\mathcal{T}_0^{(p)}$, given by

$$(\mathcal{T}_0^{(p)}\omega)(x,\delta) = \left(\omega'(x) + \frac{\omega(x)\lambda_0'(x)}{\lambda_0(x)}\right)\frac{\delta}{S_C(x)S_X(x)}.$$

This operator cannot be directly evaluated since we do not have access to S_C. The following proposition establishes the formal relation between $\widehat{\mathcal{T}_0}^{(p)}$ and $\mathcal{T}_0^{(p)}$.

Proposition 4. *Let $(T_i, \Delta_i) \sim \mu_0$, then for every $\omega \in \mathcal{H}^{(p)}$.*

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathcal{T}}_0^{(p)}\omega)(T_i, \Delta_i) \xrightarrow{\mathbb{P}} \mathbb{E}_0\left[(\mathcal{T}_0^{(p)}\omega)(T_1, \Delta_1)\right] = 0. \tag{19}$$

4. Censored-Data Kernel Stein Discrepancy

In this section, we derive censored-data Kernel Stein Discrepancies (c-KSD) using each of our three Stein operators defined in the previous section. The idea is to compare the largest discrepancy between two distributions f_X and f_0 over a class of test functions in the RKHS $\mathcal{H}$. Since we have access to censored data, we compare f_X and f_0 through the measures μ_X and μ_0, defined in Section 2.

We proceed to defined three censored-data kernel Stein discrepancies: the Survival Kernel Stein Discrepancy (s-KSD), the Martingale Kernel Stein Discrepancy (m-KSD), and the Proportional Kernel Stein Discrepancy (p-KSD) based on the respective Stein operators $\mathcal{T}_0^{(s)}$, $\mathcal{T}_0^{(m)}$ and $\widehat{\mathcal{T}_0}^{(p)}$. In general, for any given Stein operator $\mathcal{T}_0^{(c)} : \mathcal{H}^{(c)} \to \Omega$ we define the c-KSD as

$$\text{c-KSD}(f_X \| f_0) = \sup_{\omega \in B_1(\mathcal{H}^{(c)})} \mathbb{E}_X[({\mathcal{T}_0}^{(c)}\omega)(T, \Delta)].$$

Denote by $K^{(c)}$ the reproducing kernel of $\mathcal{H}^{(c)}$. By using this kernel we can get a close-form expression for c-KSD: For any of the operators $\mathcal{T}_0^{(c)}$, we define the application of $\mathcal{T}_0^{(c)}$ on $K^{(c)}(x, \cdot)$ as a function $\mathbb{R}_+ \to \mathbb{R}$ which is defined as $(\mathcal{T}_0^{(c)}\omega)(x, \delta)$ but replacing $\omega(x)$ by $K^{(c)}(x, \cdot)$ and $\omega'(x)$ by $\frac{\partial}{\partial x}K^{(c)}(x, \cdot)$. For example, for $c = m$, we get that $\left[(\mathcal{T}_0^{(m)}K^{(m)})(x, \delta)\right](\cdot)$ equals

$$\frac{\delta}{\lambda_0(x)}\left(\frac{\partial}{\partial x}K^{(m)}(x, \cdot)\right) - (K^{(m)}(x, \cdot) - K^{(m)}(0, \cdot)),$$

which the reader should compare with equation (14).

Recall that for $c \in \{s, m, p\}$, we assumed that if $\omega \in \mathcal{H}^{(c)}$ then $\omega' \in \mathcal{H}^{(c)}$, and thus $\xi^{(c)}(x, \delta)(\cdot) = \left[(\mathcal{T}^{(c)}K^{(c)})(x, \delta)\right](\cdot) \in \mathcal{H}^{(c)}$ since all operators involve ω or ω'. Define the Stein kernel $h^{(c)} : (\mathbb{R}_+ \times \{0, 1\})^2 \to \mathbb{R}$ by

$$h^{(c)}((x, \delta), (x', \delta')) = \langle \xi^{(c)}(x, \delta), \xi^{(c)}(x', \delta') \rangle_{\mathcal{H}^{(c)}}$$

The following proposition gives a closed form for the kernel Stein discrepancies $\text{c-KSD}(f_X \| f_0)$.

Proposition 5. *For $c \in \{s, m, p\}$, and let (T, Δ) and (T', Δ') be independent samples from μ_X, and suppose that*

$$\mathbb{E}_X\left[\sqrt{h^{(c)}((T, \Delta), (T, \Delta))}\right] < \infty, \tag{20}$$

then

$$(\text{c-KSD}(f_X \| f_0))^2 = \mathbb{E}_X\left[h^{(c)}((T, \Delta), (T', \Delta'))\right].$$

Detailed forms and the derivation for Stein kernels $h^{(c)}((x, \delta), (x', \delta'))$ can be found in Appendix A.3.2.

5. Goodness-of-fit Test via c-KSD

In this section, we study goodness-of-fit testing procedures based on c-KSD. We begin by estimating c-KSD^2 using

$$\widehat{\text{c-KSD}}^2(f_X \| f_0) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h^{(c)}((T_i, \Delta_i), (T_j, \Delta_j))$$

where (T_i, Δ_i) are independent samples from μ_X. By construction, under the null hypothesis, the estimator above should be close to zero, while under the alternative we expect it to be separated from zero.

5.1. Theoretical Analysis

We state some technical conditions that feature our analysis in order to establish the asymptotic behavior of $\widehat{\text{c-KSD}}^2$.

TECHNICAL CONDITIONS

a) ***Reproducing kernel conditions:*** We assume that K has continuous second-order derivatives, and that $K(x, y)$ and $\frac{\partial^2}{\partial x \partial y}K(x, y)$ are bounded and c_0-universal kernels.

b) ***Boundary condition:*** $\lim_{x \to 0+} \sqrt{K(x, x)}\lambda_0(x) < \infty$.

c) ***Null integrability conditions:*** Let $(T, \Delta), (T', \Delta') \overset{i.i.d.}{\sim} \mu_0$, and recall that $\mathbb{E}_0 = \mathbb{E}_{\mu_0}$. Depending on $c \in \{s, m, p\}$, we assume:

1) s-KSD:

   i) $\mathbb{E}_0[\phi(T, \Delta)^2 |K(T, T)|] < \infty$, and

   ii) $\mathbb{E}_0[\phi(T, \Delta)^2 \phi(T', \Delta')^2 K(T, T')^2] < \infty$,

   where $\phi(x, \delta) = \delta\frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)$.

2) m-KSD:

   i) $\mathbb{E}_0\left[\frac{|K^\star(T,T)|\Delta}{\lambda_0(T)^2}\right] < \infty$, and

   ii) $\mathbb{E}_0\left[\frac{K^\star(T,T')^2\Delta\Delta'}{\lambda_0(T)^2\lambda_0(T')^2}\right] < \infty$, where $K^\star(x,y) = \frac{\partial^2}{\partial x \partial y} K(x,y)$.

3) p-KSD:

i) $\mathbb{E}_0\left[\frac{|K^\star(T,T)|\Delta}{(f_0(T)S_C(T))^2}\right] < \infty$, and

ii) $\mathbb{E}_0\left[\frac{K^\star(T,T')^2\Delta\Delta'}{(f_0(T)f_0(T')S_C(T)S_C(T'))^2}\right] < \infty$,

where $K^\star(x,y) = \left(\frac{\partial^2}{\partial x \partial y} K(x,y)\lambda_0(x)\lambda_0(y)\right)$.

d) *Alternative integrability conditions:* Consider samples $(T,\Delta),(T',\Delta') \overset{i.i.d.}{\sim} \mu_X$. Then, for each $c \in \{s,m,p\}$ we assume:

1) s-KSD:

i) $\mathbb{E}_X[\phi(T,\Delta)^2|K(T,T)|] < \infty$,

where $\phi(x,\delta) = \delta\frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)$.

2) m-KSD:

i) $\mathbb{E}_X\left[\frac{|K^\star(T,T)|\Delta}{\lambda_0(T)^2}\right] < \infty$,

where $K^\star(x,y) = \frac{\partial^2}{\partial x \partial y} K(x,y)$.

3) p-KSD:

i) $\mathbb{E}_X\left[\frac{|K^\star(T,T)|\Delta}{S_T(T)^2\lambda_0(T)^2}\right] < \infty$,

where $K^\star(x,y) = \left(\frac{\partial^2}{\partial x \partial y} K(x,y)\lambda_0(x)\lambda_0(y)\right)$.

The following theorem establishes consistency of our empirical kernel Stein discrepancies to their population versions.

Theorem 6. *[Asymptotics under the alternative H_1] Let $c \in \{s,m,p\}$, and suppose that f_X satisfies conditions a), b), and the corresponding condition d). Then it holds*

$$\left(\widehat{\text{c-KSD}}(f_X\|f_0)\right)^2 \overset{\mathbb{P}}{\to} \left(\text{c-KSD}(f_X\|f_0)\right)^2.$$

The previous theorem is not enough to ensure good behavior under the alternative as we need to be sure that the discrepancy of two different distribution functions f_X and f_0 is different from 0 (regardless of censoring). We can prove this for c-KSD for $c \in \{s,m\}$. This does not hold true for p-KSD since it is designed to test if the hazard function λ_X is proportional to λ_0, and not for goodness-of-fit testing purposes. Indeed, whenever the hazards are in a proportional relation, p-KSD is 0.

Theorem 7. *Let $c \in \{s,m\}$. Assume $S_C(x) = 0$ implies $S_X(x) = 0$ and that K is c_0-universal. Then, under Conditions a), b) and d), $f_0 \neq f_X$ implies* $\text{c-KSD}(f_0\|f_X) > 0$.

Under the null distribution, $f_X = f_0$, we also have that $\widehat{\text{c-KSD}}(f_0\|f_0) \to 0$, but we can prove an even stronger result that follows from the theory of V-statistics.

Theorem 8 (Asymptotics under the null H_0)**.** *Let $c \in \{s,m,p\}$, and suppose that $f_X = f_0$ and that conditions a), b), and the corresponding condition c) are satisfied. Then*

$$n\left(\widehat{\text{c-KSD}}(f_X\|f_0)\right)^2 \overset{\mathcal{D}}{\to} r_c + \mathcal{Y}_c.$$

where r_c is a constant and $\mathcal{Y}_c$ is an infinite sum of independent χ^2 random variables.

While Theorem 8 ensures the existence of a limiting null distribution, which implies that a rejection region for the test is well defined, in practice it is very hard to approximate the limit distribution and the corresponding rejection regions, for which, we rely on a wild bootstrap approach.

We remark that we can obtain concentrations bounds for the test-statistics under the null hypothesis if we assume that the kernels $h^{(s)}$ and $h^{(m)}$ are bounded, by using standard methods. Obtaining concentration bounds for $h^{(p)}$ is harder as it is a random kernel, depending on all data points.

5.2. Wild Bootstrap Tests

To resample from the null distribution we use the wild bootstrap technique (Dehling & Mikosch, 1994). This technique is quite generic and it can be applied to any kernel.

The Wild Bootstrap estimator is given by

$$\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n W_iW_jh^{(c)}((T_i,\Delta_i),(T_j,\Delta_j)), \tag{21}$$

where $W_1,\ldots,W_n$ are independent random variables from a common distribution $\mathcal{W}$ with $\mathbb{E}(W_1) = 0$ and $\mathbb{V}ar(W_1) = 1$. In our experiments we consider W_i sampled from a Rademacher distribution, but any distribution with the properties above is suitable. Dehling & Mikosch (1994) proved that if the limit distribution exists (in the sense of Theorem 8), then the wild-bootstrap statistic also converges to the same limit distribution.

The testing procedure for goodness-of-fit is performed as follows: **1)** Set a type 1 error $\alpha \in (0,1)$. **2)** Compute $\widehat{\text{c-KSD}}^2(f_X\|f_0)$ using our n data points. **3)** Compute m-independent copies of the Wild Bootstrap estimator (21). **4)** Compute the proportion of wild bootstrap samples that are larger than $\widehat{\text{c-KSD}}^2(f_X\|f_0)$; if such a proportion is smaller than α we reject the null hypothesis, otherwise the do not reject it.

6. Experiments and Results

Proposed approaches: In our experiments, we denote by **mKSD** and by **pKSD**, the tests based on the martingale and the proportional kernel Stein discrepancies described in Section 3, implemented using the Wild bootstrap approach as

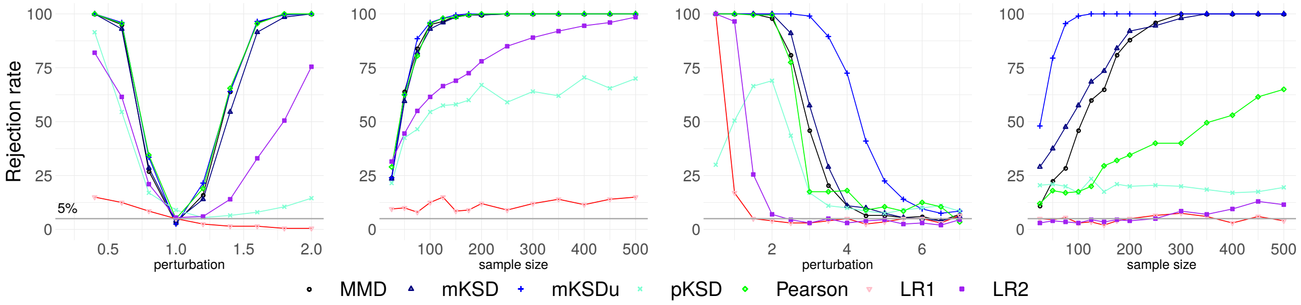

Figure 2. Rejection rate w.r.t. sample size and model perturbation. Left two for Weibull Hazard; Right two for Periodic Hazard. $\alpha = 0.01$.

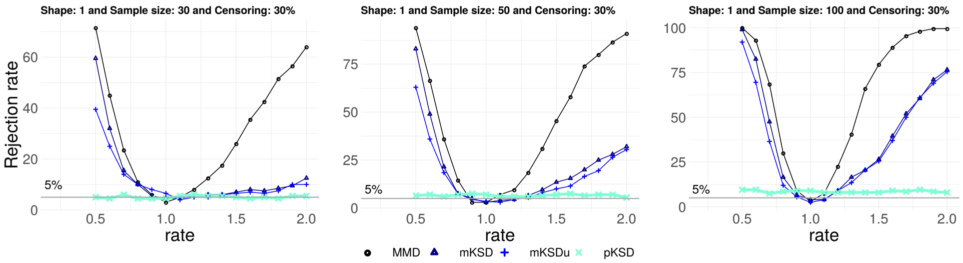

Figure 3. Rejection rate for a proportional class model. As expected, the proportional KSD test does not reject the null for different rates as all the alternatives belong to the same proportional family.

described in Section 5.2. In all our experiments we choose the null as an exponential distribution of rate 1, and in this case we can check that sKSD and mKSD coincide. Additionally, we implement **mKSDu**, which is given by the test **mKSD** applied to the transformed data $((F_0(T_i), \Delta_i))_{i=1}^n$ to test $H_0 : F_0(X) \sim \mathcal{U}(0, 1)$. Finally, for the experiments, we use an exponentiated quadratic kernel with length-scale parameter chosen by using the median-heuristic, which is the median of all the absolute differences between two different data points. We did not perform any further optimization to improve the performance of the tests.

Competing Approaches: **MMD** denotes the maximum-mean-discrepancy approach proposed by Fernandez & Gretton (2019), which provides state-of-the-art results, **Pearson** denotes the Pearson-type goodness-of-fit test proposed by Akritas (1988), which is quite competitive. **LR1** and **LR2** denote the weighted log-rank tests with respective weights functions $w_1(t) = 1$ and $w_2(t) = \sum_{i=1}^n \mathbb{1}_{\{T_i \geq t\}}$, which are classical tests, but not very competitive except for some very simple settings (e.g. testing $H_0 : \lambda_0(t) = 1$ against $\lambda_X(t) = c$, for $c \neq 1$).

Simulated experiments

Data Setting We begin by studying our method in a simulated environment where we can control all the possible parameters. We consider two data scenarios.

1. Weibull hazard functions: In our first experiment, we consider the Weibull model, which is commonly used in Survival Analysis (Bradburn et al., 2003). The Weibull distribution is characterized by the density function $f(x; k, r) = kr\,(rx)^{k-1} \exp\{-(rx)^k\}$, where k and r denote shape and rate parameters, respectively. **2. Periodic hazard functions:** A much more interesting scenario is the so-called periodic hazards, which are used to describe, for example, seasonal diseases such as Influenza. In this example, we consider the hazard function $\lambda_X(x) = 1 - \cos(\theta\pi x)$ studied in Fernandez & Gretton (2019). Note that when $\theta \to \infty$, then the distribution tends to a exponential of parameter 1. See Figure 1 for a comparison between the models.

For both models, we investigate the performance of our test in two setting: *perturbations from the null* and *increasing sample size*, which we proceed to explain. **Perturbations from the null:** In this experiment, we investigate how the power changes for perturbations of the null hypothesis. For the Weibull data, we set $H_0 : f_0(x) = f(x; 1, 1)$ and consider Weibull alternatives $f_X(x) = f(x; k, 1)$ with $k \in (0, \ldots, 2]$. Notice that we recover the null hypothesis when $k = 1$. Also, we consider a constant 30% of censored observations and a fixed sample size of $n = 100$. For the periodic experiment we set $H_0 : f_0(x) = e^{-x}$, which is recovered when we take θ tending to infinity. In this case, we consider alternatives $\theta \in \{1, 2, \ldots, 8\}$. We consider, again, a constant 30% of censoring, and a fixed sample size of $n = 100$. **Increasing sample size:** In this scenario, we investigate how the rejection rate of our test increases as the sample size increases. In the Weibull setting we set the null $H_0 : f_0(x) = f(x; 1, 1)$, the alternative as $f_X(x) = f(x; 1.5, 1)$, and in the periodic setting, we consider the null $H_0 : f_0(x) = e^{-x}$, and generate data from the alternative $\theta = 3$. In both settings we consider 30% of censored data points

Results We show our results in Figure 2. For the Weibull data (first and second plots), observe that all kernel-based methods, except the pSKD, perform very similar to the Pearson test designed to perform extremely well in these types of setting. For the Periodic data (third and fourth plots), the goodness-of-fit problem is much more challenging, and we see differences in the performances of the methods. We observe that the MMD test of Fernandez & Gretton (2019) has a better performance than the Pearson test, as was suggested by the experiments in their work. Our test mSKD performs slightly better than the the MMD test, whereas mSKDu outperform all the other methods by a huge margin. We can see that it is the most resistant to the increment in the perturbation parameter (third plot), and, for example, for $\theta = 4$, most methods cannot differentiate between null and alternative with large probability, whereas our method has power of around 75%.

Proportionality Our results show that pKSD is not a very powerful test. A possible explanation lies in the fact that, since this method tests against a model class, it must ignore all differences within this class, which affects the power of the test. Despite its lower power, it remains the only test out of the proposed methods that can test if our data was generated by a hazard proportional to λ_0. In Figure 3, we consider a Weibull hazard, given by $\lambda_X(x; k, r) = r^k k x^{k-1}$, with shape $k = 1$ and rate $r \in (0, 2)$. Note that changing the parameter r gives the same hazard up to a constant. Figure 3 shows that for a family of proportional hazards our method reaches the right type 1 error at low sample sizes, while all the other methods have non-trivial power. We observe, however, that for larger sample sizes, the test has a type 1 error that is slightly elevated over the design level. This may occur as the conditions of Theorem 8 are very hard to satisfy in general, and have yet to be proven to hold for this case. Boostraping methods with with strong theoretical guarantees under broader conditions are the subject of ongoing research.

Real Data Experiments

Data Sources We perform our tests on the following real datasets to check relevant model assumptions. **aml:** **A**cute **M**yelogenous **L**eukemia survival dataset (Miller Jr, 2011); **cgd:** **C**hronic **G**ranulotamous **D**isease dataset (Fleming & Harrington, 2011); **ovarian:** Ovarian Cancer Survival dataset (Edmonson et al., 1979); **lung:** North Central Cancer Treatment Group (NCCTG) Lung Cancer dataset (Loprinzi et al., 1994); **stanford:** Stanford Heart Transplant Data (Crowley & Hu, 1977); **nafld:** Non-alcohol fatty liver disease (NAFLD) (Allen et al., 2018).

p-value	**aml**	**cgd**	**ovarian**
Exponential	0.585	0.460	0.681
Weibull: shape=2	0.001	0.002	0.063

Table 1. Real data applications on testing hazard proportionality.

Dataset	Covarites	p-value
lung	Age	0.061
stanford	T5 mismatch score	0.057
nafld	Weight and Gender	0.108

Table 2. Real data applications on testing goodness of fit

Test Results We apply our proposed tests on real dataset for the Testing hazard proportionality and Goodness-of-fit settings. First, we check model class assumption using **pKSD** to test whether the observed data is from a desired family model without fitting model parameters. We check the exponential model class and the Weibull model with shape=2. As the results shown in Table 1, our tests does not reject the Exponential model, which is coherent with scientific domain knowledge from the literature.[1]

For the Goodness-of-fit test setting, we fit a cox proportional hazard model from the covariates provided in the datasets. The cox-proportional hazard function has the form $\lambda_X(x_i) = \lambda_b(x_i) \exp(\beta Y_i)$, where $\lambda_b(x)$ is the base hazard and Y_i is the covariate for subject i. The procedure is done via spliting the data into training set and test sets. Fitting the cox proportional-hazard model is applied on the training sets and the test sets are used to perform the goodness-of-fit tests via **mKSDu**. Results in Table 2 shows that all the models does not reject the fitted cox proportional hazard models and validate the proportional hazard assumptions for relevant fitted models, which is coherent with scientific experience stated in the literature.[2]

[1] High-grade serous ovarian carcinoma (HG-SOC) is a major cause of cancer-related death. The growth of HG-SOC acts as an indicator of survival time of ovarian cancer (Gu et al., 2019). This paper also suggests that that HG-SOC follows exponential expansion, which implies exponentially distributed survival time of ovarian patient.

[2] Chansky et al. (2016) suggests that cox proportional hazard model is a reasonable tool among practitioners for **lung** dataset. (Crowley & Hu, 1977) suggests a fit for cox proportional hazard model for **stanford** dataset. (Allen et al., 2018) states that cox proportional hazards is often used to study the impact of NAFLD on incident metabolic syndrome or death.

Acknowledgement

TF, WX, AG are grateful for the support from the Gatsby Charitable Foundation. NR is supported by Thomas Sauerwald's ERC Starting Grant 679660.

References

Aalen, O., Borgan, O., and Gjessing, H. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

Akritas, M. G. Pearson-type goodness-of-fit tests: the univariate case. *Journal of the American Statistical Association*, 83(401):222–230, 1988.

Allen, A. M., Therneau, T. M., Larson, J. J., Coward, A., Somers, V. K., and Kamath, P. S. Nonalcoholic fatty liver disease incidence and impact on metabolic burden and death: a 20 year-community study. *Hepatology*, 67(5): 1726–1736, 2018.

Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology*, 41(3): 861–870, 2012.

Barbour, A. D. and Chen, L. H. Y. *An introduction to Stein's method*, volume 4. World Scientific, 2005.

Biswas, A., Datta, S., Fine, J. P., and Segal, M. R. *Statistical advances in the biomedical science*. Wiley Online Library, 2007.

Bradburn, M. J., Clark, T. G., Love, S. B., and Altman, D. G. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3):431–436, 2003.

Brendel, M., Janssen, A., Mayer, C.-D., and Pauly, M. Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics*, 41(3):742–761, 2014.

Chansky, K., Subotic, D., Foster, N. R., and Blum, T. Survival analyses in lung cancer. *Journal of thoracic disease*, 8(11):3457, 2016.

Chen, L. H. Y., Goldstein, L., and Shao, Q. M. *Normal approximation by Stein's method*. Springer, 2010.

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

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pp. 2606–2615, 2016.

Collett, D. *Modelling survival data in medical research*. Chapman and Hall/CRC, 2015.

Crowley, J. and Hu, M. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, 72(357):27–36, 1977.

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

Edmonson, J. H., Fleming, T. R., Decker, D., Malkasian, G., Jorgensen, E., Jefferies, J., Webb, M., and Kvols, L. Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer treatment reports*, 63 (2):241–247, 1979.

Fernandez, T. and Gretton, A. A maximum-mean-discrepancy goodness-of-fit test for censored data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2966–2975, 2019.

Fernandez, T. and Rivera, N. A reproducing kernel hilbert space log-rank test for the two-sample problem. *arXiv preprint arXiv:1904.05187*, 2019.

Fernandez, T., Gretton, A., Rindt, D., and Sejdinovic, D. A kernel log-rank test of independence for right-censored data. *arXiv preprint arXiv:1912.03784*, 2019.

Fleming, T. R. and Harrington, D. P. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

Gill, R. Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist.*, 11(1): 49–58, 1983. ISSN 0090-5364. doi: 10.1214/aos/1176346055. URL https://doi.org/10.1214/aos/1176346055.

Gill, R. D. *Censoring and stochastic integrals*, volume 124 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1980. ISBN 90-6196-197-1.

Gorham, J. and Mackey, L. Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, pp. 226–234, 2015.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *ICML*, pp. 1292–1301, 2017.

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

Gu, S. S., Lheureux, S., Sayad, A., Cybulska, P., Hogen, L. B.-D., Vyarvelska, I., Tu, D., Parulekar, W., Levine, D. A., Bernardini, M. Q., et al. Computational modeling of ovarian cancer: Implications for therapy and screening. *medRxiv*, pp. 19009712, 2019.

Hollander, M. and Proschan, F. Testing to determine the underlying distribution using randomly censored data. *Biometrics*, pp. 393–401, 1979.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pp. 262–271, 2017.

Kanagawa, H., Jitkrittum, W., Mackey, L., Fukumizu, K., and Gretton, A. A kernel stein test for comparing latent variable models. *arXiv preprint arXiv:1907.00586*, 2019.

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

Klein, J. P. and Moeschberger, M. L. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

Ley, C., Reinert, G., Swan, Y., et al. Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52, 2017.

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pp. 276–284, 2016.

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., Bartel, J., Law, M., Bateman, M., and Klatt, N. E. Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group. *Journal of Clinical Oncology*, 12(3):601–607, 1994.

Miller Jr, R. G. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

Mirabello, L., Troisi, R. J., and Savage, S. A. Osteosarcoma incidence and survival rates from 1973 to 2004: data from the surveillance, epidemiology, and end results program. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 115(7):1531–1543, 2009.

Nelson, W. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Xu, W. and Matsuda, T. A stein goodness-of-fit test for directional distributions. *The 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

Yang, J., Liu, Q., Rao, V., and Neville, J. Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pp. 5557–5566, 2018.

Yang, J., Rao, V., and Neville, J. A stein–papangelou goodness-of-fit test for point processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 226–235, 2019.

Yang, S. A central limit theorem for functionals of the Kaplan-Meier estimator. *Statist. Probab. Lett.*, 21(5): 337–345, 1994. ISSN 0167-7152. doi: 10.1016/0167-7152(94)00026-3. URL https://doi.org/10.1016/0167-7152(94)00026-3.