Nonparametric independence testing and regression for time-to-event data



David Rindt

Magdalen College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Abstract

The main goal of this thesis is to develop statistical methods for nonparametric independence testing and regression between a covariate and a right-censored event-time. First, we study tests of independence that use reproducing kernel Hilbert spaces (RKHSs) to quantify the dependence in a dataset. In particular we study permutation tests with the Hilbert Schmidt independence criterion (HSIC) as the test-statistic. We show such tests are pointswise consistent, which means that, for each fixed alternative hypothesis, the probability the null hypothesis is rejected converges to 1 as the sample size converges to infinity. This shows that kernel-based permutation tests of independence are able to detect every type of independence. Second, we propose two nonparametric tests of independence between a covariate and right-censored time. The first of these tests, termed optHSIC, consists of a two-step procedure: first the censored dataset is transformed into an uncensored one using optimal transport, after which a standard permutation test with test statistic HSIC is applied to the uncensored dataset. The second proposed test is the kernel-logrank test. This test uses a supremum of scores over an RKHS as the test-statistic, and uses wild-bootstrap to test for significance. We show that these tests are able to detect a wider range of dependencies between covariates and event-times than existing methods. Lastly, we study scoring rules for right-censored regression models. We show that several often-used scoring rules are not proper, meaning that the true distribution may score worse than inaccurate distributions. Since this does not hold for the right-censored log-likelihood scoring rule, we then propose a method, termed SuMo-net, that uses neural networks to optimize the right-censored log-likelihood. SuMo-net models the cumulative distribution function and the density of a right-censored event-time jointly using partially monotonic neural networks and achieves competitive performance on a range of datasets.

Acknowledgements

My gratitude goes to my supervisors, Dino and David, for their knowledge, kindness and trust, and for the freedom they gave me to explore my research interests.

I also want to thank my collaborators, Tamara, Arthur and Robert, without whose ideas, knowledge, enthusiasm and persistence the work in this thesis would not have been possible.

Finally and most importantly, I would like to thank my parents, Jac and Brigitte, my brothers, Peter and Saul, and my grandparents for their endless love and support and for being there through all the good and the difficult moments.

Contents

1 Introduction **1**

1.1 Topic and motivation 1

1.2 Existing methods and their limitations 3

1.2.1 Nonparametric independence testing for survival data 3

1.2.2 Nonparametric regression for survival data 5

1.3 Contribution 6

1.4 Background 8

1.4.1 Survival analysis 8

1.4.2 Kernel methods 10

1.5 Publications 14

2 Consistency of permutation tests for HSIC and dHSIC **15**

2.1 Introduction 16

2.2 Background 18

2.2.1 Reproducing Kernel Hilbert Spaces 18

2.2.2 dHSIC and HSIC 20

2.2.3 Permutation testing 22

2.3 Consistency of the permutation test for HSIC and dHSIC 24

2.3.1 Comments on the proof strategy 29

2.3.2 Enumerating all permutations 30

2.4 Influence of the number of permutations on the power 31

2.4.1 For a fixed dataset 32

2.4.2 The effect on power 33

2.4.3 How many permutations to use? 35

2.5 Conclusion 36

2.6 Supplementary material 36

2.7 Statement of authorship 38

3 A kernel- and optimal transport- based test of independence between covariates and right-censored lifetimes **39**

3.1 Introduction 40

3.2 Background Material 42

3.2.1 Right-Censored Lifetimes . . . 42
3.2.2 Independence testing using kernels . . . 43
3.2.3 Optimal transport . . . 47
3.3 Data transformation based on optimal transport . . . 48
3.3.1 Objective of the algorithm . . . 48
3.3.2 Definition of the transformation . . . 49
3.3.3 Comments on the transformation algorithm . . . 50
3.3.4 Intuition behind the transformation . . . 51
3.4 Applying HSIC to the transformed dataset: optHSIC . . . 54
3.4.1 Computational cost of optHSIC . . . 55
3.4.2 Theoretical results on optHSIC . . . 56
3.4.3 Using multiple transformations . . . 57
3.5 Alternative approaches when censoring is independent of the covariate 59
3.5.1 wHSIC . . . 61
3.5.2 zHSIC . . . 63
3.6 Numerical evaluation of the methods . . . 64
3.6.1 Type 1 error rate . . . 64
3.6.2 Comparison of power . . . 66
3.6.3 Binary covariates . . . 71
3.7 Discussion . . . 72
3.8 Supplementary material . . . 73
3.9 Statement of authorship . . . 103

4 A kernel log-rank test of independence for right-censored data 104
4.1 Introduction . . . 105
4.2 Notation . . . 107
4.2.1 Survival analysis notation . . . 107
4.2.2 Counting processes notation . . . 108
4.3 Construction of the test . . . 109
4.3.1 General score test . . . 110
4.3.2 Log-rank formulation . . . 111
4.3.3 RKHS approach . . . 113
4.3.4 Reproducing Kernel Hilbert Spaces . . . 114
4.3.5 RKHS distance . . . 114
4.4 Asymptotic Analysis . . . 115
4.4.1 Asymptotic null distribution . . . 115

4.4.2 Power under alternatives . . . 117
4.4.3 Recovering existing tests . . . 119
4.5 Wild Bootstrap . . . 120
4.5.1 Implementation . . . 122
4.5.2 Computational time . . . 122
4.6 Experiments . . . 123
4.6.1 Power for 1-dimensional covariates . . . 124
4.6.2 Power for multidimensional covariates . . . 125
4.6.3 Varying censoring rates and distributions . . . 128
4.6.4 Sensitivity to choice of bandwidth . . . 128
4.6.5 Choice of kernel . . . 128
4.7 Application to real-world datasets . . . 129
4.7.1 Biofeedback data . . . 130
4.7.2 Colon data . . . 132
4.8 Conclusions . . . 133
4.9 Supplementary material . . . 133
4.10 Statement of authorship . . . 195

5 Proper scoring rules, the right-censored log-likelihood and partially monotonic neural networks for survival regression 196
5.1 Introduction . . . 197
5.2 Background and notation . . . 199
5.3 Evaluation criteria and proper scoring rules . . . 200
5.3.1 The right-censored log-likelihood as a proper scoring rule . . . 202
5.3.2 Existing evaluation metrics are not proper scoring rules . . . 203
5.3.3 Summary . . . 207
5.4 Optimizing likelihood using partially monotonic neural networks . . . 208
5.5 Experiments . . . 210
5.5.1 Results . . . 211
5.5.2 Correlation between likelihood and existing metrics . . . 213
5.5.3 Estimation of the likelihood for models other than SuMo-net . . . 213
5.5.4 Inference timings . . . 214
5.6 Discussion . . . 215
5.7 Supplementary material . . . 215
5.8 Statement of authorship . . . 224

6 Conclusion 225

Chapter 1

Introduction

This thesis follows an integrated format and contains six chapters, with the first chapter serving as an introduction, and the last chapter as a discussion. Each of the remaining chapters is based on a research article which is either published or under review at the time of submission of this thesis. Each research chapter contains introduction, background and discussion of related work specific to the topics covered therein. A general overview of the motivation and statistical context of this thesis is given in this chapter.

1.1 Topic and motivation

The main aim of this thesis is to develop assumption-free statistical methods, in particular nonparametric independence testing and regression, for the survival analysis setting.

Survival analysis is the branch of statistics that studies time-to-event data. A classical example is the study of the survival time of patients after medical treatment (Laurie et al. 1989), but time-to-event data can be found in numerous domains. In engineering one may study the time until failure of mechanical systems (Susto et al. 2015); in economics the time until borrowers default on their loans (Dirick et al. 2017); or, in business, the time until a customer is lost from a company (Van den Poel and Lariviere 2004).

A key challenge of such data is that one often does not observe the event-time exactly, but instead only observes a lower bound. That is, for certain subjects it may be known that the event happened after a certain time, but not exactly when. Such observations are called right-censored event-times, since it is only known the unobserved time lies to the right of a certain time, which is called the the censoring time. In the example of the survival time of patients after medical treatment (Laurie et al. 1989), such censoring occurs when a patient survives until the end of the study, is lost to follow-up, or withdraws from the study before it ends.

Information about the subjects is often used to predict, or understand, the event-time associated to them. We can ask, for example, if the survival time after treatment is influenced by characteristics of the patients, or, if the probability a borrower defaults on a loan can be modeled using features of the borrower. Throughout this thesis we refer to characteristics of the subjects as covariates. The challenges in survival analysis of interest to this thesis fall broadly into two categories: *regression* problems, in which we seek to model a right-censored event-time based on covariate information, and *independence-testing* problems, in which we seek to assess if a covariate affects the event-time.

Many existing methods that model the dependency between an event-time and covariate make restrictive assumptions about their true relationship. While this can simplify the modelling and interpretation, these restrictions can be limiting when the true relationship is complex or the dataset is large. In this thesis, we therefore focus on methods that do not make assumptions about the true relationship between event-times and covariates and are able to model every type of dependence between the two. We often refer to such methods as being nonparametric.

This thesis contributes to the survival analysis literature by developing novel methods for the following two challenges.

1. Assumption-free testing of the independence between a covariate and right-censored times.

2. Assumption-free modelling of the relationship between a covariate and a right-censored event-time.

The next section gives an overview of existing methods relating to the above challenges.

1.2 Existing methods and their limitations

This section gives an overview of the existing literature on the challenges identified in the previous section. Here we paint a broad picture of this literature, and we defer more detailed introductions to the relevant research chapters.

1.2.1 Nonparametric independence testing for survival data

The first main challenge we work on is testing the independence of a covariate and a right-censored event-time. Statistical independence tests assess if two random variables, say X and Y, contain information about each other.

One of the most famous methods to model the relationship between a covariate and a right-censored event-time is the Cox proportional hazards model, proposed in 1972 in the landmark paper 'Regression models and life-tables' (Cox 1972) by Sir David Cox. To this date this is arguably the most used regression method in survival analysis. The Cox proportional hazards model is semiparametric, which means that while it

allows for some flexibility, it also makes restrictive assumptions about the true relationship between the event-time and the covariate.

The Cox model can also be used to test the independence of a covariate and a right-censored event-time. However, due to the restrictive assumptions made in the Cox model, such a test is not able to detect every type of dependence. In particular the Cox model is not able to model certain nonlinear effects between the covariate and the event-time, and certain interactions between time and the covariate.

Other tests of independence can be found in the context of uncensored data. In the case of a real-valued variable Y, one may simply perform linear regression between X and Y and test if the vector of regression coefficients significantly differs from zero. An obvious limitation is that while this method is able to detect linear relationships, it may not capture nonlinear relationships between the covariate and the event-time. Other classical methods include Spearman's and Kendall's rank correlations as well as methods given in Gretton et al. (2008), Hoeffding (1948) and references therein. Many of these methods however make assumptions about the relationship between X and Y, or require one or both of the variables to be real-valued.

Over the last decade the machine learning community has revisited the topic of independence testing. In line with the development of flexible learning algorithms, a key motivation has been to develop tests that are able to detect *every type* of dependence. In Gretton et al. (2008) a statistic called the Hilbert-Schmidt independence criterion (HSIC) was proposed as a measure of dependence of random variables. This method will play an important role throughout this thesis. We will occasionally refer to HSIC as a kernel-based measure of dependence, since it uses kernels, which can be thought of as measures of similarity between datapoints, to quantify the dependence

in a dataset. HSIC has since been shown to be equivalent to a measure developed in the statistics community, called distance covariance (DCOV) (Székely and Rizzo 2009; Sejdinovic et al. 2012). Kernel-based tests have further been extended to test mutual independence (Pfister et al. 2018) and conditional independence (Zhang et al. 2011).

Kernel-based tests of independence are particularly useful when the data has complex interactions or nonlinear relationships that are difficult to detect using standard correlations (Gretton et al. 2008). Additionally, kernel based tests can be used on random variables that take values in higher-dimension, or even in non-Euclidean spaces. Indeed, kernel-based tests can be applied to text, strings, graphs and groups (Sejdinovic et al. 2012; Fukumizu et al. 2008; Borgwardt and Kriegel 2005). As mentioned, however, such kernel-based tests were, prior to the work presented in this thesis, not extended to deal with the case in which one of the variables is a right-censored event-time.

1.2.2 Nonparametric regression for survival data

The second challenge we address in this thesis is nonparametric regression of a right-censored event-time on a covariate vector. Again, it is useful to first mention the relevant methods for regressing an uncensored target variable Y on a covariate X.

Over the last several decades a range of flexible regression methods have been proposed that are able to model complex relationships between the target variable and the covariate. Some of the most notable developments include support vector machines, kernel regression (Schölkopf and Smola 2001), Gaussian processes (Rasmussen and Williams 2005), decision trees and neural networks (Bishop 2006).

In many applications, these methods have been shown to vastly outperform linear models on tasks such as regression and classification, particularly when the underlying relationships are complex and the sample size is large.

Most of the above methods have been extended to model a right-censored time based on a covariate X. In particular methods using Gaussian processes (Fernández et al. 2016), neural networks (Katzman et al. 2018; Kvamme et al. 2019), random forests (Ishwaran et al. 2008) and kernels (Li and Luan 2002) have all been extended to deal with the case where the target variable is a right-censored event-time.

Existing methods, however, often have one of the following limitations. First, many methods, including some of the methods using modern machinery, still make restrictive assumptions about the event-time distribution, and are thus not able to model every kind of dependence. Second, many methods do not have a tractable likelihood or learn discrete distributions. As we show in Chapter 5, this makes it difficult to evaluate the accuracy of the fitted models.

1.3 Contribution

In this Thesis we make the following contributions.

In Chapter 2, we resolve an important open question regarding kernel-based permutation tests of independence. Permutation tests are a popular and practical method for implementing kernel-based independence tests. Such tests compare the test statistic computed on the original data to the test statistic computed on randomly reshuffled datasets. Despite their popularity, it has remained an open question if kernel-based

permutation tests are consistent, i.e. if for each fixed alternative hypothesis, the probability the null hypothesis is rejected converges to 1 as the sample size converges to infinity. In Chapter 2 we answer this question and show that kernel-based independence tests are indeed consistent.

In Chapters 3 and 4, we develop kernel-based methods to test for independence between a covariate and a right-censored event-time. First, in Chapter 3, we propose a testing procedure termed optHSIC. This procedure consists of a two-step procedure and first transforms the censored dataset into an uncensored one using optimal transport, and then applies a permutation test with test statistic HSIC to the transformed dataset. We demonstrate through experiments that this test is able to detect a much wider range of dependencies than the Cox likelihood ratio test. In Chapter 4, we propose a second kernel-based test of independence between a covariate and a right-censored lifetime. We name this test the kernel log-rank test. It combines a test statistic similar to HSIC with a wild-bootstrap testing procedure. This method was proposed after the optHSIC procedure of Chapter 3 and comes with improved power in certain scenarios and a better developed theoretical framework. In particular, we prove that the kernel-logrank test is point-wise consistent and has asymptotically valid level.

In Chapter 5, we first investigate scoring rules for survival regression models. A scoring rule indicates how well a learned distribution approximates the true distribution. Our first contribution is to show that several existing scoring rules have the undesirable property that inaccurate distributions may score better than the true data-generating distribution. In contrast, the right-censored log-likelihood is a proper scoring rule. Therefore, we secondly propose a novel regression method that optimizes the right-censored likelihood. The method, termed SuMo-net, uses par-

tially monotonic neural networks to model the cumulative distribution and density function jointly. Unlike existing methods, SuMo-net is both able to model every type of dependence, and has a tractable likelihood, while remaining relatively easy to implement.

1.4 Background

We now introduce the key concepts in survival analysis and kernel methods on which this thesis builds.

1.4.1 Survival analysis

We are interested in an event-time $T \in \mathbb{R}_{\geq 0}$. Typically we want to know how the event-time depends on some subject information encoded in a covariate $X \in \mathbb{R}^p$ for $p \geq 1$. Importantly, we do not observe the event-time T directly. For some subjects we only know that T exceeds some censoring time $C \in \mathbb{R}_{\geq 0}$. We thus observe a time $Z \coloneqq \min\{T, C\}$ and the event-indicator $\Delta \coloneqq 1\{Z = T\}$. We denote a sample of size n by $(X_i, Z_i, \Delta_i)_{i=1}^n$. Throughout this thesis we make the following assumption about the censoring time that is common in survival analysis:

Assumption 1.1. *We assume that $T \perp\!\!\!\perp C|X$. That is, we assume that given the covariate, the censoring time has no further information about the event time.*

We denote the cumulative distribution function of T given a covariate X by $F(t|x)$. The survival function is defined by $S(t|x) = 1 - F(t|x)$. Assuming that T has a density conditional on x, we let $f(t|x) = \partial F(t|x)/\partial t$. Under the same assumption, the hazard rate is defined by

$$\lambda(t|x) = \frac{f(t|x)}{S(t|x)}.$$

This implies that $S(t|x) = \exp\left(-\int_0^t \lambda(s|x)ds\right)$. The integrated hazard is called the cumulative hazard and denoted by

$$\Lambda(t|x) = \int_0^t \lambda(s|x)ds.$$

The most famous method for modeling the event-time based on the covariates is undoubtedly the Cox proportional hazards model introduced in Cox (1972). This model assumes that the hazard rate satisfies

$$\lambda(t|x) = \lambda(t)\exp(\beta^T x)$$

where $\lambda(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is the baseline hazard, $\beta \in \mathbb{R}^p$ is a coefficient vector and the superscript T denotes the transpose. The coefficient vector β can be estimated efficiently using the Newton-Raphson algorithm, and the cumulative baseline hazard $\lambda(t)$ can be estimated using Breslow's method (Breslow 1972).

Some of the advantages of this model are the following. Firstly the effect of the covariates on the hazard rate can be interpreted through the coefficient vector β. It can furthermore be tested whether the coefficients significantly deviate from zero. Additionally, the estimate of the baseline hazard is fully nonparametric, and is not needed to estimate β, since $\lambda(t)$ cancels out in the partial likelihood (Cox 1972). Since no assumptions are made regarding the baseline hazard, the Cox model is often referred to as a semiparametric method.

1.4.2 Kernel methods

The term *kernel methods* refers to a broad class of algorithms that includes kernel ridge regression, support vector machines and kernel principal component analysis. Using functions called kernels, linear methods are turned into nonlinear ones.

Including nonlinear effects in linear models can often be done by mapping the covariate $x \in \mathbb{R}^p$ to a higher dimensional feature vector $\phi(x) \in \mathbb{R}^q$, for $q > p$ that includes nonlinear terms of the original variable x. For example, to include a quadratic effect in the regression, we could map the covariate $x \in \mathbb{R}$ to $\phi(x) = (x, x^2)$ and then linearly regress y on the feature vector $\phi(x)$.

Next, we note that many linear methods, such as principal component analysis, linear regression and linear support vector machines, depend on the covariates x only through the inner products $\langle x, x' \rangle$ between input variables. Consequently, when we map x to a higher dimensional space, we do not need to work with the feature vectors $\phi(x)$ explicitly, but instead we only need access to $k(x, x') \coloneqq \langle \phi(x), \phi(x') \rangle$ (Schölkopf and Smola 2001). The function $k(\cdot, \cdot)$ is called the kernel function. In practice, the feature map ϕ is often not made explicit and only the kernel function k is mentioned.

Importantly, unlike the feature space $\mathbb{R}^q$ mentioned above, the feature space does not need to be finite dimensional. Many widely used kernels, including the radial basis function kernel, also known as the Gaussian kernel, as well as the Matern kernel, correspond to infinite dimensional feature spaces.

We now define reproducing kernel Hilbert spaces. These spaces provide a mathematical foundation for kernelized algorithms.

Definition 1.2. *(Reproducing Kernel Hilbert Space (Schölkopf and Smola 2001)) Let $\mathcal{X}$ be a non-empty set and H a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ endowed with dot product $\langle \cdot, \cdot \rangle$. Then H is called a reproducing kernel Hilbert (RKHS) space if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties.*

1. *k satisfies the reproducing property*

$$\langle f, k(x, \cdot) \rangle = f(x) \text{ for all } f \in H, x \in \mathcal{X}. \tag{1.1}$$

2. *k spans H, that is, $H = \overline{\text{span}\{k(x, \cdot) \, | x \in \mathcal{X}\}}$ where the bar denotes the completion of the space.*

The kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite if for each set of points $x_1, \ldots, x_n \in \mathcal{X}$, the matrix $K_{ij} = k(x_i, x_j)$ is positive semidefinite. It is easy to verify that every kernel function of an RKHS is positive definite. It also holds that for every positive definite kernel k there exists a RKHS with kernel k (see Schölkopf and Smola (2001) Section 2.2.2).

The following theorem, called the Representer Theorem, relates optimization over functions in a potentially infinite dimensional RKHS to optimization over finite-dimensional spaces. Since many statistical methods optimize an objective function, this theorem provides the foundation for kernelizing many statistical techniques.

Theorem 1.3. *(Representer theorem (Kimeldorf and Wahba 1971; Schölkopf and Smola 2001)) Let $c : \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^n \to \mathbb{R}$ be an arbitrary loss function and let $J : [0, \infty) \to \mathbb{R}$ be strictly increasing. Let $x_1, \ldots, x_n \in \mathcal{X}$, $Y \in \mathbb{R}^n$. Finally let $f \in H$ where H is an RKHS with reproducing kernel k, and let $K_{ij} = k(x_i, x_j)$,*

$i, j = 1, \ldots, n$. *Then* $\hat{f}_n$ *minimizes*

$$Q_1(f) \coloneqq c(Y, x_1, \ldots, x_n, f(x_1), \ldots, f(x_n)) + J(\|f\|_H^2)$$

over $f \in H$ *if and only if* $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(x_i, \cdot)$ *and* $\hat{\alpha}$ *minimizes* Q_2 *over* $\alpha \in \mathbb{R}^n$ *where*

$$Q_2(\alpha) \coloneqq c(Y, x_1, \ldots, x_n, K\alpha) + J(\alpha^T K \alpha).$$

Finally we briefly define mean embeddings and the Hilbert-Schmidt independence criterion, which play a key role in the independence tests developed in this thesis. A more detailed discussion is given in the research Chapters 2 and 3.

Theorem 1.4. *(Existence of the mean embedding) Let* H *be an RKHS on* $\mathcal{X}$ *with kernel* k*. Let* P *be a probability distribution on* $\mathcal{X}$ *and assume that*

$$\mathbb{E}_{X \sim P} \sqrt{k(X, X)} < \infty.$$

Then there exists a unique element μ_P *in* H *so that for every function* $f \in H$ *it holds that*

$$\mathbb{E}_{X \sim P} f(X) = \langle f, \mu_P \rangle_H.$$

The element μ_P *is called the mean embedding of* P *in* H*.*

Mean embeddings can be related to a measure of distance between probability measures called the maximum mean discrepancy (MMD). Let P and Q be probability distributions on $\mathcal{X}$. Then the MMD between P and Q is defined as

$$\mathrm{MMD}(P, Q) = \sup_{f \in H, \|f\|_H \leq 1} \mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{X \sim Q} f(X).$$

It is straightforward to show that the MMD equals the RKHS distance between the two associated mean-embeddings, i.e.

$$\mathrm{MMD}(P, Q) = \|\mu_P - \mu_Q\|_H.$$

In the setting of independence testing, we study a pair of random variables (X, Y). Let $P_{X,Y}$ denote the the joint distribution of X and Y on $\mathcal{X} \times \mathcal{Y}$ and let K be a reproducing kernel on $\mathcal{X} \times \mathcal{Y}$. In Gretton et al. (2008) it was proposed to measure the strength of the dependence between X and Y through the Hilbert-Schmidt independence criterion (HSIC), which equals the MMD between the joint distribution and the product of the marginal distributions of X and Y:

$$\mathrm{HSIC}(X, Y) = \mathrm{MMD}(P_{XY}, P_X P_Y).$$

Here, $P_X P_Y$ denotes the product of the marginal distributions P_X and P_Y.

1.5 Publications

This thesis contains the following four research articles which are either published or under review at the time of the submission of this thesis.

Chapter 2

David Rindt, Dino Sejdinovic, David Steinsaltz. Consistency of permutation tests of independence using distance covariance, HSIC and dHSIC. `https://doi.org/10.1002/sta4.364`. Stat. 2021; 10:e364.

Chapter 3

David Rindt, Dino Sejdinovic, David Steinsaltz. A kernel- and optimal transport-based test of independence between covariates and right-censored lifetimes. `https://doi.org/10.1515/ijb-2020-0022`. The International Journal of Biostatistics 2020.

Chapter 4

Tamara Fernández, Arthur Gretton, *David Rindt*, Dino Sejdinovic. A kernel log-rank test of independence for right-censored data. `https://arxiv.org/abs/1912.03784`.

Chapter 5

*David Rindt**, Robert Hu*, David Steinsaltz, Dino Sejdinovic. Proper scoring rules, the right-censored log-likelihood and partially monotonic neural networks for survival regression. `https://arxiv.org/abs/2103.14755`.

* denotes equal contribution.

Chapter 2

Consistency of permutation tests for HSIC and dHSIC

This chapter is based on the following paper.

David Rindt, Dino Sejdinovic, David Steinsaltz.
Consistency of permutation tests of independence using distance covariance, HSIC and dHSIC.
https://doi.org/10.1002/sta4.364.
Stat. 2021; 10:e364.

The Hilbert–Schmidt Independence Criterion (HSIC) and its d-variable extension dHSIC are measures of (joint) dependence between random variables. While combining these statistics with a permutation test has become a popular method of testing the null hypothesis of (joint) independence, it had thus far not been proved that this results in a consistent test. In this work we provide a simple proof that the permutation test with test statistic HSIC or dHSIC is indeed consistent when using characteristic kernels. I.e. we prove that under each alternative hypothesis, the power of these permutations tests indeed converges to 1 as the sample size converges to infinity. Since the test is consistent for each number of permutations, we further give a brief discussion of how the number of permutations relates to the power of the test and how the number of permutations may be selected in practice.

2.1 Introduction

During the last decade many new tests of independence of random variables have been developed. Among these are tests based on distance covariance (Szekély and Rizzo (2009)) and the Hilbert–Schmidt independence criterion (HSIC) (Gretton et al. (2008)). In fact, under a particular choice of kernels, HSIC has been shown to be equivalent to distance covariance (Sejdinovic et al. (2012)). In Pfister et al. (2018) a d-variable extension of HSIC, named dHSIC, was proposed, which can be used to test the hypothesis of joint independence of d random variables.

These statistics have several desirable properties. For appropriate choices of kernels, the population value of HSIC equals zero if and only if the two variables are independent (Gretton et al. (2008)). Similarly, the population value of dHSIC is zero if and only if the variables are indeed jointly independent (Pfister et al. (2018)). One thus does not need to make assumptions about the form of the relationship among the variables. Furthermore, under mild conditions, the test statistic converges in probability to the population value. Additionally, these tests may be applied to multidimensional random variables, and even to variables that do not take values in the Euclidean domains, such as graphs or text (Gretton et al. (2008)).

In practice, one does not have access to the true null distribution of these statistics. As discussed in Gretton et al. (2008) and Pfister et al. (2018), the three main methods to approximate a null distribution are using the Gamma-distribution, a bootstrap procedure and a using a permutation test.

To our knowledge Pfister et al. (2018) was the first work to study consistency the properties of these three kernel-based testing procedures. A test is called consistent if the rejection rate of the test converges to 1 as the sample size increases for each distribution for which the alternative hypothesis holds. That work proved in particular that the bootstrap procedure results in a consistent test whereas the Gamma-

approximation does not guarantee a consistent test. Consistency of the permutation procedure remained an open question and is the topic of this work.

A permutation test is a test in which the statistic on the original data is compared to the statistic on a number of randomly permuted datasets. We define this in Definition 2.7. An important property of such permutation tests is that they are well known to have valid level for each sample size and for each number of permutations (see e.g. Hoeffding (1952)).

There are several relevant recent works on the power of permutation tests. The question if a permutation test with statistic dHSIC or HSIC results in a consistent test was posed in Pfister et al. (2018): see in particular Table 1 and Section 3.2.1 of that work. In Kim et al. (2020), uniform power (rather than power for a fixed alternative) of permutation tests has been studied against classes of alternative hypotheses meeting certain conditions on the densities of the joint distribution. In Berrett et al. (2020), uniform power of permutation tests has also been studied under certain conditions on the density, although not in the specific context of HSIC. In contrast to these last two papers, we study the power for each fixed alternative hypothesis, and we do not make any assumptions on the alternative hypothesis.

The main contribution of this work is to provide a simple proof of the consistency of the permutation test for each fixed alternative hypothesis. We do not follow the proof strategy proposed in Remark 2 of Pfister et al. (2018) but use much more elementary techniques that can be traced back at least to Hoeffding (1952): as we discuss in Section 2.2, the test statistic dHSIC, with appropriate choice of kernel, converges to a positive constant for each fixed alternative hypothesis. The main observation from which consistency will follow in Section 2.3 is that the statistic on the randomly permuted datasets converges to zero in probability (Lemma 2.11).

While only elementary techniques are used in proving consistency, we do believe the proved result is of importance. The main motivation for using kernel-methods

for independence testing is the ability to detect every possible type of dependence and Theorem 2.12 shows that this is indeed the case for the permutation test using dHSIC. The analogous question of consistency of the bootstrap procedure was already answered in Pfister et al. (2018), where it was also shown that the Gamma-approximation procedure did not have a consistency guarantee and the question of consistency of the permutation test was posed. This result on the permutation test thus aims to fill this gap in the literature.

Since the permutation test is both consistent and has correct type 1 error for each number of permutations, it is natural to ask how to select the number of permutations. One may ask for example how the number of permutations influences the power of the test. Section 2.4 briefly discusses these questions provides some guidelines on how one may select the number of permutations in practice.

2.2 Background

2.2.1 Reproducing Kernel Hilbert Spaces

This section reviews some relevant information about reproducing kernel Hilbert spaces (RKHSs).

Definition 2.1. *(Reproducing Kernel Hilbert Space (Schölkopf and Smola (2001)))* *Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ endowed with dot product $\langle \cdot, \cdot \rangle$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties.*

1. *k has the reproducing property*

$$\langle f, k(x, \cdot) \rangle = f(x) \text{ for all } f \in \mathcal{H}, x \in \mathcal{X}.$$

2. *k spans H; that is,* $\mathcal{H} = \overline{\text{span}\{k(x,\cdot)\ |x \in \mathcal{X}\}}$ *where the bar denotes the completion of the space.*

Let $\mathcal{X}$ be a measurable space and $\mathcal{H}_k$ be a RKHS on $\mathcal{X}$ with kernel k. Let $\mathbb{P}$ be a probability measure on $\mathcal{X}$. If $\mathbb{E}_{\mathbb{P}}\sqrt{k(X,X)} < \infty$, then there exists an element $\mu_{\mathbb{P}} \in \mathcal{H}_k$ such that $\mathbb{E}_{\mathbb{P}} f(X) = \langle f, \mu_{\mathbb{P}} \rangle$ for all $f \in H_k$ (Gretton et al. (2012a)), where we use the notation $\mathbb{E}_{\mathbb{P}} f(X) \coloneqq \int f(x)\mathbb{P}(dx)$. The element $\mu_{\mathbb{P}}$ is called the mean embedding of $\mathbb{P}$ in $\mathcal{H}_k$. Given a sample $\{x_i\}_{i=1}^n$ and the corresponding empirical distribution, $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, the corresponding mean embedding is given by $\frac{1}{n}\sum_{i=1}^n k(x_i,\cdot)$. Given a second distribution $\mathbb{Q}$ on $\mathcal{X}$, of which a mean embedding exists, we can measure the dissimilarity of $\mathbb{P}$ and $\mathbb{Q}$ by the distance between their mean embeddings in $\mathcal{H}_k$. That is,

$$\text{MMD}(\mathbb{P},\mathbb{Q}) \coloneqq ||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||_{\mathcal{H}_k}.$$

This is also called the Maximum Mean Discrepancy (MMD). The name comes from the following equality (Gretton et al. (2012a)),

$$||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k} |\mathbb{E}_{\mathbb{P}} f(X) - \mathbb{E}_{\mathbb{Q}} f(X)|,$$

showing that MMD is an integral probability metric. Lastly we give the definitions of a characteristic kernel and a c_0-universal kernel.

Definition 2.2. *(Characteristic kernel (Sriperumbudur et al. (2011))) The kernel k is said to be characteristic when* $\text{MMD}(\mathbb{P},\mathbb{Q}) = 0$ *if and only if* $\mathbb{P} = \mathbb{Q}$.

Definition 2.3. *(c_0-Universal kernel (Sriperumbudur et al. (2011))) Let $\mathcal{X}$ be a locally compact Hausdorff space and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel on $\mathcal{X}$. Then k is said to be c_0-universal if it is continuous and its associated RKHS H_k is dense in $C_0(\mathcal{X})$, the set of continuous bounded functions, with respect to the infinity*

(also called uniform) norm.

If a kernel is c_0-universal on a domain $\mathcal{X}$, then the kernel is also characteristic on that domain (Sriperumbudur et al. (2011)). We now give some examples of c_0-universal (and thus also characteristic) kernels on $\mathbb{R}^d$.

1. Gaussian kernel: $k_\sigma(x, y) = \exp(-||x - y||^2/\sigma^2)$ with $\sigma > 0$ for $x, y \in \mathbb{R}^d$.
2. Matérn kernel: $k_{v,\rho}(x, y) = \frac{2^{1-v}}{\Gamma(v)} \left(\sqrt{2v}\frac{||x-y||}{\rho}\right)^v K_v \left(\sqrt{2v}\frac{||x-y||}{\rho}\right)$, where Γ is the gamma function, K_v the modified Bessel function of the second kind, $\rho, v \in \mathbb{R}_{>0}$.
3. Rational quadratic kernel: $k_{\alpha,\ell} = \left(1 + \frac{||x-x'||^2}{2\alpha\ell}\right)^{-\alpha}$ with $\alpha, \ell \in \mathbb{R}_{>0}$. This kernel is both characteristic and c_0-universal.
4. Inverse multiquadric kernel: $k_c(x, y) = 1/\sqrt{||x - y||^2 + c}$ for $c \in \mathbb{R}_{>0}$.
5. Covariance kernel of Brownian motion: $k(x, y) = (||x|| + ||y|| + ||x - y||)/2$. This kernel is not c_0-universal on all of $\mathbb{R}^d$. However, when we view this as a kernel on a compact domain, rather than on all of $\mathbb{R}^d$, then the kernel is characteristic and c_0-universal.

2.2.2 dHSIC and HSIC

In Pfister et al. (2018) a measure of joint independence is proposed. Consider the following setting.

Setting 1. For $j = 1, \ldots, d$, let $\mathcal{X}^j$ be a locally compact metric space equipped with the Borel sigma-algebra and let the Cartesian product $\mathcal{X} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^d$ be equipped with the product sigma-algebra. Let $X^j : \Omega \to \mathcal{X}^j$ be random variables on the shared probability space $(\Omega, \mathbb{P}, \mathcal{F})$. In this section the superscript on X^j always indexes X, and never denotes a power of the variable X. Let $k^j(\cdot, \cdot) : \mathcal{X}^j \times \mathcal{X}^j \to \mathbb{R}$ be kernels on $\mathcal{X}^j$. Finally let $k \coloneqq k^1 \otimes \cdots \otimes k^d$ be the tensor product of the d kernels. We let

$\mathcal{H}_k, \mathcal{H}_{k^j}$ be the corresponding RKHSs. Assume that the kernel k is characteristic and bounded.

In practice, one may construct a bounded characteristic product kernel k, by choosing each k^j for $j = 1, \ldots, d$ to be c_0-universal and bounded: in Szabó and Sriperumbudur (2017) it is found that the product of c_0-universal kernels is c_0-universal and thus also characteristic. Kernels 1-4 above are all c_0-universal and bounded. Kernel 5 is c_0-universal and bounded when we restrict it to be a kernel on a compact subset of $\mathbb{R}^d$.

By definition $(X^1, ..., X^d)$ are said to be jointly independent if $\mathbb{P}_{X^1,...,X^d} = \mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d}$. The main topic of this work is the test of the hypothesis $H_0 : \mathbb{P}_{X^1,...,X^d} = \mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d}$ versus $H_1 : \mathbb{P}_{X^1,...,X^d} \neq \mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d}$. We now define dHSIC:

Definition 2.4. *(dHSIC, Pfister et al. (2018)) Assume Setting 1. Then dHSIC is defined as*

$$dHSIC(X^1, ..., X^d) := ||\mu_{\mathbb{P}_{X^1,...,X^d}} - \mu_{\mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d}}||^2_{\mathcal{H}_k}$$

where we recall that for any distribution $\mathbb{P}$, the element $\mu_{\mathbb{P}} \in \mathcal{H}_k$ denotes the mean embedding of $\mathbb{P}$.

In Setting 1 we have assumed k to be characteristic, so $\mathrm{dHSIC}(X^1, \ldots, X^d) = 0$ if and only if $\mathbb{P}_{X^1,...,X^d} = \mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d}$.

As we typically do not have access to the full distribution $\mathbb{P}_X$, but only to a sample $D := (x_i)_{i=1}^n \in \mathcal{X}^n$, we study the estimator proposed in Pfister et al. (2018)

$$\begin{aligned}\widehat{\mathrm{dHSIC}}(x_1, \ldots, x_n) :=& \frac{1}{n^2} \sum_{M_2(n)} \prod_{j=1}^{d} k^j(x^j_{i_1}, x^j_{i_2}) + \frac{1}{n^{2d}} \sum_{M_{2d}(n)} \prod_{j=1}^{d} k^j(x^j_{i_{2j-1}}, x^j_{i_{2j}}) \\ & - \frac{2}{n^{d+1}} \sum_{M_{d+1}(n)} \prod_{j=1}^{d} k^j(x^j_{i_1}, x^j_{i_{j+1}}).\end{aligned}$$

Here $M_q(n) = \{1, \ldots, n\}^q$. Note that this equals the RKHS distance between the mean embedding of the empirical distribution and of the product of the marginal empirical distributions. The following Theorem shows that $\widehat{\text{dHSIC}}(D)$ converges in probability to 0 if and only if the null hypothesis is true.

Theorem 2.5. *(Convergence of dHSIC to population value (Pfister et al. (2018)))
Assume Setting 1. Then*

$$\widehat{dHSIC}(D) \to dHSIC(X^1, ..., X^d)$$

as $n \to \infty$ in probability. Furthermore, under these assumptions, the limit $dHSIC(X^1, ..., X^d) = 0$ if and only if the null hypothesis, $\mathbb{P}_{X^1,...,X^d} = \mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d}$ is true.

In the case where $d = 2$, dHSIC is better known as HSIC which is defined below.

Definition 2.6. *(HSIC, Gretton et al. (2008)) The Hilbert–Schmidt independence criterion (HSIC) of random variables $X^1 \in \mathcal{X}^1$ and $X^2 \in \mathcal{X}^2$ is defined as*

$$\text{HSIC}(X^1, X^2) \coloneqq ||\mu_{\mathbb{P}_{X^1,X^2}} - \mu_{\mathbb{P}_{X^1} \times \mathbb{P}_{X^2}}||^2_{H_k}.$$

In Sejdinovic et al. (2012) HSIC was shown to be equivalent, under a certain choice of kernel, to a generalization of distance covariance (Lyons (2013)), a statistic proposed in Szekély and Rizzo (2009). We will mainly prove statements for dHSIC, which then readily carry over to HSIC, as well as to distance covariance, by way of this equivalence.

2.2.3 Permutation testing

We follow the notation of Pfister et al. (2018) for the permutation test using dHSIC. We denote by S_n the group of permutations on n elements, which consists of $n!$

distinct permutations. Let $\psi^j \in S_n$ for $j = 2, \ldots, d$, and define the permutation vector $\psi = (\psi^2, \ldots, \psi^d)$. Then ψ maps $\mathcal{X}^n$ to itself by

$$\psi : \begin{matrix} x_1 \\ \vdots \\ x_n \end{matrix} \begin{pmatrix} x_1^1 & x_1^2 & \ldots & x_1^d \\ \vdots & & & \vdots \\ x_n^1 & x_n^2 & \ldots & x_n^d \end{pmatrix} \to \begin{matrix} x_1^\psi \\ \vdots \\ x_n^\psi \end{matrix} \begin{pmatrix} x_1^1 & x_{\psi_2(1)}^2 & \ldots & x_{\psi_d(1)}^d \\ \vdots & & & \vdots \\ x_n^1 & x_{\psi_2(n)}^2 & \ldots & x_{\psi_d(n)}^d \end{pmatrix}.$$

Note that we keep the first coordinate fixed and permute the remaining $d-1$ coordinates. Hence there are $(n!)^{d-1}$ different vectors of such permutations in total. The reason for keeping the first coordinate fixed is that the order in which the n observations appear does not affect HSIC, and thus for a fixed dataset D, the distribution of $\widehat{\text{HSIC}}(\psi D)$ is the same regardless of whether all coordinates are permuted or only the last $d-1$. We are now ready to define the permutation test.

Definition 2.7. *(Permutation test dHSIC sampling $B \times (d-1)$ permutations) Let $\alpha \in (0,1)$. For $b = 1, \ldots, B$, let $\psi_b = (\psi_b^2, \ldots, \psi_b^d)$ be i.i.d. vectors of $d-1$ permutations sampled i.i.d. uniformly from S_n. Then let R be the rank of the first entry of the vector*

$$\left(\widehat{dHSIC}(D), \widehat{dHSIC}(\psi_1 D), \ldots, \widehat{dHSIC}(\psi_B D)\right)$$

when breaking ties at random and where $R = 1$ denotes the rank of the largest elements and $R = B+1$ of the smallest element. We reject H_0 if $p_B \coloneqq R/(B+1) \leq \alpha$. The quantity p_B denotes the p-value of the permutation test enumerating B permutations and we call α the level of the test.

A well known property of permutation tests is that they have valid level for every value of B.

Definition 2.8. *(Valid level) Let $\phi(X_1 \ldots, X_n) \in \{0, 1\}$ be a hypothesis test of level α returning 1 if it rejects H_0, and 0 otherwise. The test ϕ of level α is said to have valid level if for all distributions $\mathbb{P}_X \in H_0$ (i.e., all distributions $\mathbb{P}_X$ such that $\mathbb{P}_X = \mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d}$), it holds that $\mathbb{P}(\phi(X_1, \ldots, X_n) = 1) \leq \alpha$.*

Theorem 2.9. *(Valid level permutation test. E.g. Hoeffding (1952); Pfister et al. (2018)) Assume H_0 is true, i.e. the $x_1, \ldots, x_n \in \mathcal{X}$ are sampled i.i.d. from a distribution $\mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d} \in H_0$. Then for the permutation test of Definition 2.7, it holds that*

$$\mathbb{P}(p_B \leq \alpha) = \frac{\lfloor \alpha(B+1) \rfloor}{B+1} \leq \alpha.$$

2.3 Consistency of the permutation test for HSIC and dHSIC

We now turn to the main topic of this paper, which is proving that the permutation tests of independence using HSIC and dHSIC are consistent against each fixed alternative hypothesis. We first formally define consistency of a test. Let $\phi(X_1, \ldots, X_n) = 1$ if the null hypothesis is rejected and $\phi(X_1, \ldots, X_n) = 0$ otherwise.

Definition 2.10. *(Consistency for each alternative) The test ϕ is called consistent for each alternative if for all distributions $\mathbb{P}_X \in H_1$ (i.e. for all distributions $\mathbb{P}_X$ such that $\mathbb{P}_X \neq \mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d}$), it holds that*

$$\lim_{n \to \infty} \mathbb{P}_X(\phi(X_1, \ldots, X_n) = 1) = 1.$$

That is, a test is consistent for each alternative if for any fixed alternative hypothesis, the rejection rate converges to 1 as the sample size grows to infinity.

We begin by proving the empirical dHSIC of a randomly permuted sample converges to zero in probability.

Lemma 2.11. *Let $\psi = (\psi^1, \dots, \psi^{d-1})$ be a vector of i.i.d. random permutations. Then*

$$\widehat{dHSIC}(\psi D) \to 0$$

in probability.

Proof. We note that since $\widehat{\text{dHSIC}}$ is nonnegative it suffices to show that

$$\lim_{n\to\infty} \mathbb{E}\big[\widehat{\text{dHSIC}}(\psi D)\big] = 0.$$

For the context of this proof only, we slightly change the notation of the permutation vector. Recall that, since we only permute the coordinates $j = 2, \dots, d$, the permutation vector equals $\psi = (\psi^2, \dots, \psi^d)$. For convenience we add to the notation that we apply the permutation $\psi^1 = id$ to the first coordinate. So we now define $\psi = (\psi^1, \dots, \psi^d)$, where $\psi^1 = id$ and $\psi^2, \dots, \psi^d$ are sampled i.i.d. from all permutations on n-elements. The statistic $\widehat{\text{dHSIC}}(\psi D)$ can then be seen to be

$$\begin{aligned}\widehat{\text{dHSIC}}(\psi D) =& \frac{1}{n^2} \sum_{M_2(n)} \prod_{j=1}^{d} k^j(X^j_{\psi^j(i_1)}, X^j_{\psi^j(i_2)}) + \frac{1}{n^{2d}} \sum_{M_{2d}(n)} \prod_{j=1}^{d} k^j(X^j_{\psi^j(i_{2j-1})}, X^j_{\psi^j(i_{2j})}) \\ &- \frac{2}{n^{d+1}} \sum_{M_{d+1}(n)} \prod_{j=1}^{d} k^j(X^j_{\psi^j(i_1)}, X^j_{\psi^j(i_{j+1})}),\end{aligned}$$

which we abbreviate as $A_n + B_n - 2C_n$. We now aim to show that $\lim_n \mathbb{E}A_n = \lim_n \mathbb{E}B_n = \lim_n \mathbb{E}C_n = \zeta$ where $\zeta = \prod_{j=1}^{d} \mathbb{E}\left(k^j(X^j, \tilde{X}^j)\right)$ with X^j and $\tilde{X}^j$ denoting independent copies of the same random variable with law $\mathbb{P}_{X^j}$ for $j = 1, \dots, d$.
We begin proving that $\mathbb{E}A_n \to \zeta$. First define the following random (since they depend on ψ) subsets of $M_2(n) = \{1, \dots, n\}^2$:

$$U(2,\psi,n) \coloneqq \Big\{ (i_1,i_2) \in M_2(n) : (\psi^j(i_1),\psi^j(i_2) : j = 1,\ldots,d) \\ \text{are } 2\times d \text{ distinct elements} \Big\},$$

and $R(2,\psi,n) \coloneqq M_2(n) \setminus U(2,\psi,n)$. We now show that, as the sample size increases, the expected value of the portion of tuples of $M_2(n)$ that is in $U(2,\psi,n)$ tends to 1. Note first that $n(n-1)$ is the number of pairs $(i_1,i_2) \in M_2(n)$ so that $i_1 \neq i_2$. Next assume that $i_1 \neq i_2$ and that $(\psi^j(i_1),\psi^j(i_2) : j = 1,\ldots,J-1)$ are $2(J-1)$ distinct elements. Under that assumption the fraction $\binom{n-2(J-1)}{2}/\binom{n}{2}$ corresponds with the probability that $\psi^J(i_1),\psi^J(i_2) \notin \{1,\ldots,n\} \setminus \{\psi^j(i_1),\psi^j(i_2) \ : \ j = 1,\ldots,J-1\}$. Repeating that argument, we obtain

$$\begin{aligned} &\frac{\mathbb{E}|U(2,\psi,n)|}{n^2} \\ &= \frac{n(n-1)}{n^2} \times \mathbb{P}\left((\psi^j(1),\psi^j(2) : j = 1,\ldots,d) \text{ are } 2\times d \text{ distinct elements}\right) \\ &= \frac{n(n-1)}{n^2} \frac{\binom{n-2}{2}\binom{n-4}{2}\cdots\binom{n-2(d-1)}{2}}{\binom{n}{2}\binom{n}{2}\cdots\binom{n}{2}} \\ &\to 1. \end{aligned}$$

As a result, it follows that $\mathbb{E}|R(2,\psi,n)|/n^2 \to 0$. Conditioning on ψ and using the

tower property we find

$$\begin{aligned}
\mathbb{E}\left(A_n\right) &= \mathbb{E}\left(\mathbb{E}\left(A_n|\psi\right)\right) \\
&= \mathbb{E}\left(\frac{1}{n^2}\sum_{U(2,\psi,n)}\mathbb{E}\left(\prod_{j=1}^{d}k^j(X^j_{\psi^j(i_1)},X^j_{\psi^j(i_2)})\middle|\psi\right)\right) \\
&\quad+\mathbb{E}\left(\frac{1}{n^2}\sum_{R(2,\psi,n)}\mathbb{E}\left(\prod_{j=1}^{d}k^j(X^j_{\psi^j(i_1)},X^j_{\psi^j(i_2)})\middle|\psi\right)\right) \\
&= \mathbb{E}\left(\frac{|U(2,\psi,n)|}{n^2}\right)\prod_{j=1}^{d}\mathbb{E}\left(k^j(X^j,\tilde{X}^j)\right) \\
&\quad+\mathbb{E}\left(\frac{|R(2,\psi,n)|}{n^2}\right)\mathcal{O}(1) \\
&\to \zeta.
\end{aligned}$$

Note that in the last equality, we use that in the first sum all indices in the product are distinct and the expectation factorizes, and in the second sum the estimate $\mathcal{O}(1)$ follows from the fact that in Setting 1 the kernels are bounded.

Repeating this argument for B_n, we define

$$\begin{aligned}
U(2d,\psi,n) \coloneqq \Big\{&(i_1,\ldots,i_{2d})\in M_{2d}(n):\\
&(\psi^j(i_{2j-1}),\psi^j(i_{2j}))_{j=1}^{d}\text{ are }2\times d\text{ distinct elements}\Big\}
\end{aligned}$$

and let $R(2d,\psi,n) \coloneqq M_{2d}(n)\setminus U(2d,\psi,n)$. Note that there are $n^d(n-1)^d$ index-vectors $(i_1,\ldots,i_{2d})\in M_{2d}(n)$ so that $i_{2j-1}\neq i_{2j}$ for $j=1,\ldots,d$. Combined with the same computation as for A_n we find that

$$\begin{aligned}
\frac{\mathbb{E}\left(|U(2d,\psi,n)|\right)}{n^{2d}} &= \frac{n^d(n-1)^d}{n^{2d}}\frac{\binom{n-2}{2}\binom{n-4}{2}\cdots\binom{n-2(d-1)}{2}}{\binom{n}{2}\binom{n}{2}\cdots\binom{n}{2}} \\
&\to 1.
\end{aligned}$$

Proceeding similarly as we did for A_n,

$$\begin{aligned}\mathbb{E}\left(B_n\right) &= \mathbb{E}\left(\frac{|U(2d,\psi,n)|}{n^{2d}}\right)\prod_{j=1}^{d}\mathbb{E}\left(k^j(X^j,\tilde{X}^j)\right)\\ &\quad+\mathbb{E}\left(\frac{|R(2d,\psi,n)|}{n^{2d}}\right)\mathcal{O}(1)\\ &\to \zeta.\end{aligned}$$

Lastly, turning our attention to C_n, we define

$$\begin{aligned}U(d+1,\psi,n) \coloneqq \Big\{&(i_1,\ldots,i_{d+1})\in M_{d+1}(n):\\ &(\psi^j(i_1),\psi^j(i_{j+1}))_{j=1}^{d} \text{ are } 2\times d \text{ distinct elements}\Big\}\end{aligned}$$

while defining $R(d+1,\psi,n) = M_{d+1}(n)\setminus U(d+1,\psi,n)$. Note that there are $n(n-1)^d$ tuples $(i_1,\ldots,i_{d+1})$ of $M_{d+1}(n)$ for which $i_j \neq i_1$ for all $j=2,\ldots,d+1$. Hence

$$\begin{aligned}\frac{|U(d+1,\psi,n)|}{n^{d+1}} &= \frac{n(n-1)^d}{n^{d+1}}\frac{\binom{n-2}{2}\binom{n-4}{2}\cdots\binom{n-2(d-1)}{2}}{\binom{n}{2}\cdots\binom{n}{2}}\\ &\to 1.\end{aligned}$$

Applying the same argument as before to $\mathbb{E}(C_n)$ we find that $\lim_n(C_n)\to\zeta$. It follows that $\mathbb{E}[\widehat{\mathrm{dHSIC}}(\psi D)] = A_n + B_n - 2C_n \to 0$ and hence $\widehat{\mathrm{dHSIC}}(\psi D)\to 0$ in probability.

□

It is now straightforward to prove the main result of this paper.

Theorem 2.12. *(Consistency using a finite sample of permutations) Assume we are in Setting 1 with $\mathbb{P}_X$ a distribution such that $\mathbb{P}_X \neq \mathbb{P}_{X^1}\times\cdots\times\mathbb{P}_{X^d}$. Perform a permutation test on a sample of size n using B random permutation vectors $\psi_1,\ldots,\psi_B$*

of length $d-1$*, where* $B \geq \frac{1}{\alpha} - 1$ *for* $\alpha \in (0,1)$*. Then*

$$\lim_{n\to\infty} \mathbb{P}(p_B \leq \alpha) = 1.$$

Proof. Define $C \coloneqq \mathrm{dHSIC}(X^1, \ldots, X^d) > 0$. Then

$$\begin{aligned}
\mathbb{P}(p_B \leq \alpha) &\geq \mathbb{P}(\widehat{\mathrm{dHSIC}}(D) > C/2, \quad \widehat{\mathrm{dHSIC}}(\psi_b D) < C/2 \text{ for } b = 1, \ldots, B) \\
&\geq 1 - \Big(\mathbb{P}(\widehat{\mathrm{dHSIC}}(D) < C/2) + B \times \mathbb{P}(\widehat{\mathrm{dHSIC}}(\psi_1 D) > C/2)\Big) \\
&\to 1.
\end{aligned}$$

as $n \to \infty$, which proves the Theorem. The fact the two probabilities in the last equation converge to 0 follows from Lemma 2.11 and the convergence of $\widehat{\mathrm{dHSIC}}(D) \to C$ in probability. □

2.3.1 Comments on the proof strategy

Contrary to the proof above, the proof strategy proposed in Pfister et al. (2018) requires extending a proof of Romano (1989), which studies statistics of the form $\sqrt{n}\delta_{\mathcal{V}}(\mathbb{P}_X, \mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d})$ where $\delta_{\mathcal{V}}(P, Q) = \sup\{|P(V) - P(Q)| : V \in \mathcal{V}\}$ and $\mathcal{V}$ can be an arbitrary Vapnik-Cherovenkis (VC) class. Since the VC class is arbitrary, there is no closed form expression for this statistic. While the proof of Romano (1989) may be extended from indicators of sets in a VC class, to functions in an RKHS, the fact that we have an analytic closed form of HSIC enables the simpler approach given in this paper.

In fact, Romano (1989) built on previous work by Hoeffding (1952), which describes a framework for asymptotic power of permutation tests. That work considers two conditions that together characterize asymptotic power of the permutation test. Translated

to our setting, Condition 1 is that the random quantile

$$Q^{n,B}(\alpha) \coloneqq \inf\left\{t \ : 1 + \sum_{b=1}^{B} 1\left\{\widehat{\mathrm{dHSIC}}(\psi_b D) \leq t\right\} \leq \alpha(B+1)\right\} \to \lambda$$

in probability as $n \to \infty$, for some constant $\lambda \in \mathbb{R}$. Condition 2 is that there exists a function $H : \mathbb{R} \to [0,1]$, continuous at λ, so that $\mathbb{P}\left(\widehat{\mathrm{dHSIC}}(D) \leq t\right) \to H(t)$ for each t at which H is continuous, as $n \to \infty$. Under these conditions, it can be seen that the power $\mathbb{P}(p_B \leq \alpha) \geq \mathbb{P}\left(\widehat{\mathrm{dHSIC}}(D) > Q^{n,B}(\alpha)\right) \to 1 - H(\lambda)$ (see Hoeffding (1952)). For simplicity we chose not to introduce these conditions in the proof of Theorem 2.12. However, in the case of $\widehat{\mathrm{dHSIC}}$ it is easy to see that Lemma 2.11 implies that Condition 1 is met with $Q_{n,B}(\alpha) \to 0$ (so that $\lambda = 0$ in Condition 1). Furthermore, using the convergence of $\widehat{\mathrm{dHSIC}}(D)$ to its population limit it is straightforward to verify that Condition 2 is met with the step-function $H(t) = 1\left\{\mathrm{dHSIC}(X^1, \ldots, X^d) \leq t\right\}$. Since for this H, it holds that $H(\lambda) = H(0) = 0$, corresponding to the fact that the test is consistent.

2.3.2 Enumerating all permutations

Instead of sampling $B \times (d-1)$ permutations, one may also consider a permutation test enumerating all permutation vectors. There are $(n!)^{d-1} - 1$ permutation vectors of length $d-1$ so that not all coordinates are the identity permutation. One then computes the rank of $\widehat{\mathrm{dHSIC}}(D)$ in the vector

$$\left(\widehat{\mathrm{dHSIC}}(D), \widehat{\mathrm{dHSIC}}(\psi_1 D), \ldots, \widehat{\mathrm{dHSIC}}(\psi_{(n!)^{d-1}-1} D)\right),$$

breaking ties at random and rejects H_0 if the resulting rank is not greater than $\alpha(n!)^{d-1}$. The resulting test has valid level and is consistent, so it has the same guarantees as the test sampling a finite number of permutations. In practice enumerating all possible permutation vectors is almost always infeasible. See Hoeffding (1952) for

a proof of the valid level of the test enumerating all permutations, and see Section 2.6.1 for a proof of its consistency.

2.4 Influence of the number of permutations on the power

The previous results imply that, if the kernel k is characteristic, then for every B so that $B \geq \frac{1}{\alpha(B+1)} - 1$ the dHSIC permutation test with B permutation vectors of length $d-1$ is consistent and has a valid level. Since the permutation test has both these properties for each such number of permutations, it is natural to ask how to select the number of permutations and how the number of permutations influences the power of the test. That is the topic of this section.

To answer these questions we consider the permutation test of Definition 2.7, but instead break ties conservatively, by which we mean the following. Let ψ_b i.i.d. permutation vectors for $b = 1, \ldots, B$. Then define R^{cons} by $R^{\text{cons}} := 1 + \sum_{b=1}^{B} 1\{\widehat{\text{dHSIC}}(\psi_b D) \geq \widehat{\text{dHSIC}}(D)\}$. Defining $p_B^{\text{cons}} := R_{\text{cons}}/(B+1)$ we see that $p_B^{\text{cons}} = \frac{Z+1}{B+1}$ where Z is a binomial random variable with success probability $p_D := \mathbb{P}(\widehat{\text{dHSIC}}(\psi D) \geq \widehat{\text{dHSIC}}(D) \mid D)$. Note, in particular, that $p_B^{\text{cons}} \to p_D$ as $B \to \infty$. In the case of ties between statistics on permuted data and the statistic on the original data, the test breaking ties conservatively is more conservative than the test breaking ties at random. However, when the data comes from a continuous distribution and the dataset is large, typically few ties occur and the two tests perform similarly. In particular, Theorem 2.12 and Theorem 2.9 hold also when the ties are broken conservatively. We study the test breaking ties conservatively, since it simplifies the following sections.

2.4.1 For a fixed dataset

In this section we consider a fixed dataset D with an associated quantity p_D, and study the effect of the number of permutations on the probability of rejecting H_0. The rejection probabilities for a range of values of B for different values of p_D is given in Figure 2.1. We distinguish three cases based on the value of

$$p_D \coloneqq \mathbb{P}(\widehat{\mathrm{dHSIC}}(\psi D) \geq \widehat{\mathrm{dHSIC}}(D) \mid D)$$

of the given dataset D.

1. The data is such that $p_D < \alpha$: In this case, increasing the number of permutations increases the probability of rejecting the null hypothesis for this dataset.

2. The data is such that $p_D = \alpha$: In this case the probability of rejecting the null hypothesis is approximately $1/2$, since the mean and median of a binomial distribution are very close.

3. The data is such that $p_D > \alpha$: In this case, as the number of permutations increases, p_B^{cons} gets more and more likely to be close to p_D, and we are thus less likely to reject the null hypothesis. So, using fewer permutations actually raises the probability of rejecting the null hypothesis based on this dataset.

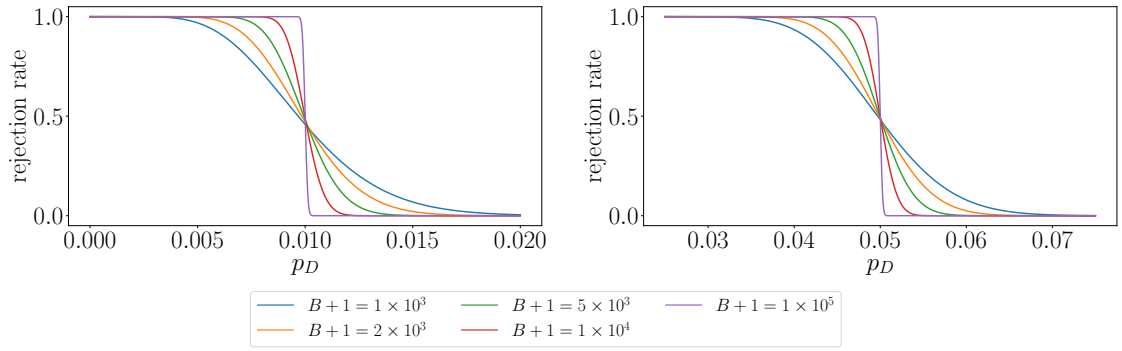


Figure 2.1: Say you are given (fixed) data D with associated p_D. When you enumerate all permutations, you reject if and only if $q \leq \alpha$. These plots plot $\mathbb{P}(p_B^{\text{cons}} \leq \alpha | p_D)$ for different values of B. In the left plot $\alpha = 0.010$ and in the right plot $\alpha = 0.05$.

2.4.2 The effect on power

While the previous subsection studied the rejection rate for a fixed dataset, we now study the rejection rate for a fixed alternative hypothesis. Note the parameter p_D itself is a random quantity too (as it is a function of the data). Say, for the sake of simplicity, that for a fixed alternative hypothesis the quantity p_D has a density $f_{p_D}(p)$ on $[0, 1]$. The total probability of rejecting the null hypothesis when using a permutation test with B permutation vectors is

$$\mathbb{P}(p_B^{\text{cons}} \leq \alpha) = \int_{[0,1]} \mathbb{P}(Z \leq \alpha(B+1) - 1 | p_D = p) f_{p_D}(p) dp.$$

While we plotted the function $\mathbb{P}(Z \leq \alpha(B+1) - 1 | p_D = p)$ in Figure 2.1, the quantity $f_{p_D}(p)$ will depend on the data generating mechanism. It will be approximately uniform under the null hypothesis, but analytic descriptions of f_{p_D} are complicated for arbitrary distributions of X. We perform two simulation studies to illustrate the relationship between power and B. We study the case where $d = 2$ and $\alpha = 0.05$ and where X^1 and X^2 are 5-dimensional random variables.

Scenario 1: Let

$$X^2 = \theta X^1 + \epsilon$$

where $X^1, \epsilon \sim \mathcal{N}(0, I_5)$ independently where I_5 is the $5-$dimensional identity matrix. When $n = 100$ we find that the power is nearly identical for all numbers of permutations. This holds for all values of θ, as shown in Figure 2.2. An explanation is that the variance in the underlying p-value p_D is large when the sample size is 100, and as a result f_{p_D} has a wide support, and the integral of the functions plotted in Figure 1 with density f_{p_D} all result in the same value.

Scenario 2: The previous scenario showed no difference in power between tests with different numbers of permutations. The explanation was the variance of p_D, or the width of the distribution f_{p_D}. It is not easy to find distributions of X^1 and X^2 such that f_{p_D} has a support only in the region where the curves in Figure 1 are separated (so near α) and this support is furthermore not symmetric around α. So in Scenario 2, for each value of θ, we choose a fixed (nonrandom) dataset. Namely:

$$X^2 = \sin(\theta X^1)$$

where $X^1_i = i2\pi/100$ for $i = 1, \ldots, 100$ (so again $n = 100$). In this case, as θ increases, the frequency of the oscillation increases and the sample looks less dependent. Indeed in Figure 2.2 we see that slow oscillations are more often rejected when using more permutations than when using fewer permutation. Similarly, fast oscillations are more often rejected when using fewer permutations than when using more permutations.

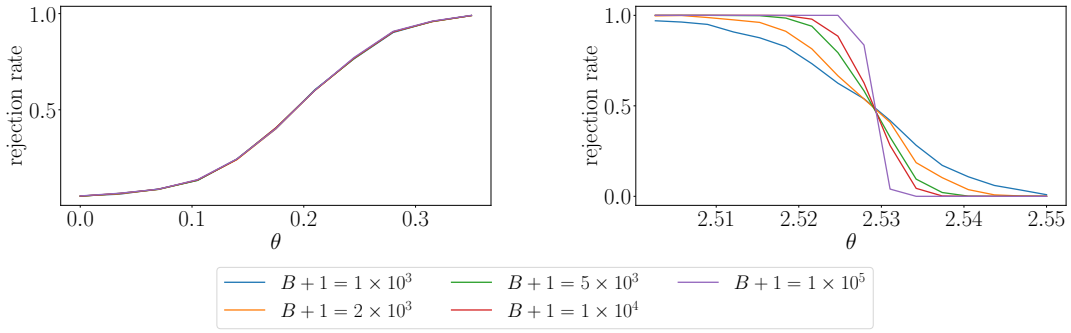


Figure 2.2: Left: The rejection rates of the permutation test with different values of B in Scenario 1. Using more permutations makes little difference to the power of test. Right: The rejection rates of the permutation test with different values of B in Scenario 2. Using more permutations makes one virtually always reject datasets with lower θ and virtually never with higher θ.

2.4.3 How many permutations to use?

In practice one is not only interested in accepting or rejecting the null hypothesis at a certain significance level, but one wants to find a reliable estimate of p_D. Therefore we recommend to choose B so large that p_B is likely to be sufficiently close to p_D for the problem at hand. We recall that, given the dataset, $p_B^{\text{cons}} = (Z+1)/(B+1)$ where $Z \sim \text{Binom}(B, p_D)$. Hence accuracy of p_B^{cons} can be described simply through confidence intervals of the binomial distribution. That is,

$$\begin{aligned}&\mathbb{P}\left(p_B^{\text{cons}} - \epsilon \leq p_D \leq p_B^{\text{cons}} + \epsilon\right)\\ &\quad = \mathbb{P}\big((p_D - \epsilon)(B+1) - 1 \leq Z \leq (p_D + \epsilon)(B+1) - 1\big).\end{aligned}$$

For a given, ϵ, p_D and confidence level $1-\lambda$, we can find B such that $p_B^{\text{cons}} \pm \epsilon$ is a confidence interval of level $1-\lambda$. As we of course do not know p_D, we could choose B so large that for *any* p_D, the interval $p_B^{\text{cons}} \pm \epsilon$ is a $(1-\lambda)-$confidence interval. However, B will always be highest for $p_D = 0.5$ as that value maximizes the variance of the binomial distribution, which means one requires very large numbers of permutations.

Noting that the accuracy of p_B is more important when p_D is small, and in particular not much larger than α, we propose a strategy to decide whether a value of B is sufficiently large, illustrating it on the example of $\alpha = 0.05$, $B = 2.3 \times 10^4$ and $C = 0.1$. We begin by selecting a $C \in (\alpha, 1)$, and consider the largest width of a 99% confidence interval in case $p_D \in [0, C]$, and in the case $p_D \in [C, 1]$. In our example, the former occurs when $p_D = C$, and the resulting confidence interval is $p_B^{\text{cons}} \pm 0.005$, and the second is $p_B^{\text{cons}} \pm 0.01$, corresponding to $p_D = 0.5$. If the accuracy in both of these two cases is sufficient for the purpose at hand, we perform the test with B permutations, otherwise we need to increase the number of permutations.

2.5 Conclusion

We have studied kernel measures of dependence and how they are combined with permutation tests to perform hypothesis testing. Our main contribution is proving the consistency of the permutation test with statistic dHSIC with a characteristic kernel. Applying this result to the case $d = 2$, it follows in particular that the permutation tests of independence with test statistics HSIC and distance covariance are also consistent for each alternative hypothesis. We further gave examples of how one may go about choosing a number of permutations in practice.

2.6 Supplementary material

2.6.1 Consistency of the test enumerating all permutations

Theorem 2.13. *(Consistency using all permutations) Let $\mathbb{P}_X$ be any distribution such that $\mathbb{P}_X \neq \mathbb{P}_{X^1} \times \cdots \times \mathbb{P}_{X^d}$. Perform the permutation test described in Section*

3.2 and let p denote the resulting p-value. Then

$$\lim_{n\to\infty} \mathbb{P}(p \leq \alpha) = 1.$$

Proof. Note that $p \leq \mathbb{P}(\widehat{\mathrm{dHSIC}}(\psi D) \geq \widehat{\mathrm{dHSIC}}(D)|D)$, which would be the p-value when ties are broken conservatively. By Markov's inequality,

$$\begin{aligned}
\mathbb{P}(p \geq \alpha) &\leq \mathbb{P}(\mathbb{P}(\widehat{\mathrm{dHSIC}}(\psi D) \geq \widehat{\mathrm{dHSIC}}(D)|D) \geq \alpha) \\
&\leq \frac{\mathbb{E}\left[\mathbb{P}(\widehat{\mathrm{dHSIC}}(\psi D) \geq \widehat{\mathrm{dHSIC}}(D)|D)\right]}{\alpha} \\
&= \frac{\mathbb{P}(\widehat{\mathrm{dHSIC}}(\psi D) \geq \widehat{\mathrm{dHSIC}}(D))}{\alpha} \\
&\to 0
\end{aligned}$$

by Theorem 2.11 and the convergence of $\widehat{\mathrm{dHSIC}}(D)$. □

2.7 Statement of authorship

Statement of authorship	
Title of paper	Consistency of permutation tests of independence using distance covariance, HSIC and dHSIC
Status	Published
Authors	David Rindt, Dino Sejdinovic and David Steinsaltz.
Details	Stat. 2021; 10:e364. `https://doi.org/10.1002/sta4.364`
Student confirmation	
Student name	David Rindt
Contribution to the paper	I am the primary author of everything in the paper. The work was completed under the supervision of Professor Dino Sejdinovic and Professor David Steinsaltz.
Date	1 June 2021
Signature	
Supervisor confirmation	
By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.	
Supervisor name and title	Dino Sejdinovic, Associate Professor of Statistics
Supervisor comments	All above seems accurate and fair.
Date	22 June 2021
Signature	

Chapter 3

A kernel- and optimal transport- based test of independence between covariates and right-censored lifetimes

This chapter is based on the following paper.

David Rindt, Dino Sejdinovic, David Steinsaltz.
A kernel- and optimal transport- based test of independence between covariates and right-censored lifetimes.
https://doi.org/10.1515/ijb-2020-0022.
The International Journal of Biostatistics 2020.

We propose a nonparametric test of independence, termed optHSIC, between a covariate and a right-censored lifetime. Because the presence of censoring creates a challenge in applying the standard permutation-based testing approaches, we use optimal transport to transform the censored dataset into an uncensored one, while preserving the relevant dependencies. We then apply a permutation test using the kernel-based dependence measure as a statistic to the transformed dataset. The type 1 error is proven to be correct in the case where censoring is independent of the covariate. Experiments indicate that optHSIC has power against a much wider class of alternatives than Cox proportional hazards regression and that it has the correct

type 1 control even in the challenging cases where censoring strongly depends on the covariate.

3.1 Introduction

We propose a nonparametric test of independence between a possibly multidimensional covariate and a right-censored lifetime. Existing approaches to this problem suffer from serious limitations: if we cluster the continuous covariate into groups, and then test for equality of lifetime distributions among the groups, the results will depend on the arbitrary choice of boundaries between the groups, and the spread of covariates within groups reduces power. Alternatively, one might fit a (semi-)parametric regression model, and test whether the regression coefficient corresponding to the covariate differs significantly from zero. The most commonly used such method is the Cox proportional hazards (CPH) model, which makes two assumptions (Cox (1972)): first, the hazard function must factorize into a function of time and a function of the covariate (the *proportional hazards* or *relative risk* condition); second, the effect of a covariate on the logarithm of the hazard function must be linear. Although this is a flexible model, in the case of complicated hazards its application is not straightforward. More complicated hazards are found for example when studying the relationship between body mass index and mortality - Zajacova and Burgard (2012) reports U- or V-shaped hazards - or between diastolic blood pressure and various health outcomes (Lip et al. (2019)). To model such effects one needs to decide which nonlinear terms should be added to the CPH model.

Since distance- and kernel-based approaches have been used successfully for independence testing on uncensored data (Gretton et al. (2008), Szekély and Rizzo (2009)), it is natural to investigate whether these methods can be extended to the case of right-

censored lifetimes. To this end we propose applying optimal transport to transform the censored dataset into an uncensored dataset in such a way that, 1) the new uncensored dataset preserves the dependencies of the original dataset, and 2) we can apply a standard permutation test to the new dataset with test statistic given by Distance Covariance (DCOV) (Székely and Rizzo (2009)) or, equivalently, the Hilbert–Schmidt Independence Criterion (HSIC) (Gretton et al. (2008)).

Progress in kernel-based independence testing for censored data seems all the more exigent when we consider how the corresponding methods have been further developed into tests of conditional independence in the simpler context of uncensored data, with crucial applications to causal inference (Zhang et al. (2011)), and in particular detection of confounders. While the present work does not propose a method for testing conditional independence, we believe independence testing is a first step toward that end. Additionally, since our method allows for multidimensional covariates, one can first test for a dependency based on the full multidimensional covariate, and then test whether the dependency remains when certain sub-dimensions are omitted from the covariate.

Section 3.2 overviews the relevant concepts in survival analysis, distance- and kernel-based independence testing, and optimal transport. Section 3.3 proposes a transformation of the data based on optimal transport. Section 3.4 introduces our testing procedure named optHSIC. Although we have not yet been able to prove control of the type 1 error rate in full generality, we do show the type 1 error rate to be correct in the case where censoring is independent of the covariate. Furthermore we obtain very promising results in simulation studies, showing correct type 1 control even under censoring that depends strongly on the covariate. Section 3.5 explores alternative kernel-based approaches under the additional assumption that censoring is independent of the covariate. These methods serve as benchmarks for the power performance of optHSIC. Section 3.6 compares the power and type 1 error of all tests

and CPH regression in simulated data.

3.2 Background Material

3.2.1 Right-Censored Lifetimes

Let $T \in \mathbb{R}_{\geq 0}$ be a lifetime subject to right-censoring, so that we do not observe T directly, but instead observe $Z \coloneqq \min\{T, C\}$ for some censoring time $C \in \mathbb{R}_{\geq 0}$, as well as the indicator $\Delta \coloneqq 1\{C > T\}$. We further observe a covariate vector $X \in \mathbb{R}^d$, where $\mathbb{R}^d$ is equipped with the Borel sigma-algebra. In total, we thus obtain the data $D \coloneqq ((x_i, z_i, \delta_i))_{i=1}^n \in (\mathbb{R}^d \times \mathbb{R}_{\geq 0} \times \{0, 1\})^n$.
The main goal of this paper is developing a test of $H_0 : X \perp\!\!\!\perp T$ versus $H_1 : X \not\perp\!\!\!\perp T$ based on the sample D. Throughout this paper we will make the following assumptions.

Assumption 1: We assume that conditional on $\{X_i\}_{i=1}^n$, the random variables $\{(T_i, C_i)\}_{i=1}^n$ are mutually independent.

Assumption 2: We assume that $C \perp\!\!\!\perp T|X$.

Denote $F_{T|X}(t|x) = P(T \leq t|X = x)$ and $F_{C|X}(t|c) = P(C \leq t|X = x)$. We assume $F_{T|X}$ has a density $f_{T|X}$. Let $S_{T|X}(t|x) = 1 - F_{T|X}(t|X)$. Then we define the hazard rate of an individual with covariate x to be $\lambda_{T|X}(t|x) = f_{T|X}(t|x)/S_{T|X}(t|x)$. The Cox proportional hazards model (CPH) assumes that the hazard rate can be written as $\lambda_{T|X}(t|x) = \lambda(t)\exp(\beta^T x)$ for some baseline hazard $\lambda(t)$ and a vector $\beta \in \mathbb{R}^d$. This model enables estimation of β and testing the significance of the difference of entries of β from zero. The CPH model is one of the most commonly used regression methods in survival analysis.

A final important concept is that of the 'individuals at risk at a time t'. By this we mean the set $\{i : Z_i \geq t\}$. We also use the notation $[a_1, \ldots, a_k]$ for a multiset (a set with potentially repeated elements) with elements $a_1, \ldots, a_k$. We refer, for example, to the multiset $\mathrm{AR}_t = [x_i : i \in \{1, \ldots, n\}, Z_i \geq t]$, the multiset of 'covariates at risk at time t'.

3.2.2 Independence testing using kernels

Kernel methods have been successfully used for nonparametric independence- and two-sample testing (Gretton et al. (2008), Gretton et al. (2012a)). We now give some of the relevant background in kernel methods.

Definition 3.1. *(Reproducing Kernel Hilbert Space)(Schölkopf and Smola (2001)) Let $\mathcal{X}$ be a non-empty set and H a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ endowed with dot product $\langle \cdot, \cdot \rangle$. Then H is called a reproducing kernel Hilbert (RKHS) space if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties.*

1. *k satisfies the reproducing property*

$$\langle f, k(x, \cdot) \rangle = f(x) \text{ for all } f \in H, x \in \mathcal{X}. \tag{3.1}$$

2. *k spans H, that is, $H = \overline{\mathrm{span}\{k(x, \cdot) \ | x \in \mathcal{X}\}}$ where the bar denotes the completion of the space.*

Let $\mathcal{X}$ together with a sigma-algebra be a measurable space and let $H_{\mathcal{X}}$ be an RKHS on $\mathcal{X}$ with reproducing kernel k. Let P be a probability measure on $\mathcal{X}$. If $E_P \sqrt{k(X, X)} < \infty$, then there exists an element $\mu_P \in H_{\mathcal{X}}$ such that $E_P f(X) = \langle f, \mu_P \rangle$ for all $f \in H_{\mathcal{X}}$ (Gretton et al. (2012a)), where the notation $E_P f(X)$ is defined to be $\int f(x) P(dx)$. The element μ_P is called the mean embedding of P in $H_{\mathcal{X}}$. Given a second distribution Q on $\mathcal{X}$, for which a mean embedding exists, we can measure the similarity of P and

Q by the distance between their mean embeddings in $H_{\mathcal{X}}$:

$$\mathrm{MMD}(P, Q) \coloneqq ||\mu_P - \mu_Q||_{H_{\mathcal{X}}},$$

which is also called the Maximum Mean Discrepancy (MMD). The name comes from the following equality (Gretton et al. (2012a)),

$$||\mu_P - \mu_Q||_{H_{\mathcal{X}}} = \sup_{f \in H_{\mathcal{X}} : \|f\| \leq 1} E_P f(X) - E_Q f(X)$$

showing that MMD is an integral probability metric. Given a sample $\{x_i\}_{i=1}^n$ and the empirical distribution, $\sum_{i=1}^n \delta(x_i)/n$, where $\delta(x)$ denotes the Dirac measure at x, the corresponding mean embedding is given by $\sum_{i=1}^n k(x_i, \cdot)/n$.

Suppose now that $\mathcal{Y}$ together with some sigma-algebra is another measurable space, and let $H_{\mathcal{Y}}$ be an RKHS on $\mathcal{Y}$ with kernel l. Let X be a random variable in $\mathcal{X}$ with law P_X and similarly let Y be a random variable in $\mathcal{Y}$ with law P_Y. Finally let P_{XY} denote the joint distribution on $\mathcal{X} \times \mathcal{Y}$ equipped with the product sigma-algebra. We let H denote the RKHS on $\mathcal{X} \times \mathcal{Y}$ with kernel

$$K((x, y), (x', y')) \coloneqq k(x, x')l(y, y').$$

In Gretton et al. (2008) it was proposed that the dependence of X and Y could be quantified by the following measure:

Definition 3.2. *The Hilbert–Schmidt independence criterion (HSIC) of X and Y is defined by*

$$\mathrm{HSIC}(X, Y) \coloneqq ||\mu_{P_{XY}} - \mu_{P_X P_Y}||_H^2$$

where $P_X P_Y$ denotes the product measure of P_X and P_Y.

Let (X, Y),(X', Y') and (X'', Y'') be three mutually independent copies of the same random variable with law P_{XY}. Using the reproducing property and the definition of mean embeddings, it can be shown that

$$\begin{aligned}\text{HSIC}(X, Y) = E_{XY}E_{X'Y'}k(X, X')l(Y, Y') + E_{XX'}k(X, X')E_{YY'}l(Y, Y') \\ - 2E_{XY}E_{X'Y''}k(X, X')l(Y, Y'').\end{aligned}$$

Now assume we are given a sample $D = ((x_i, y_i))_{i=1}^n$ of independent observations of the random pairs (X, Y). An empirical estimate of $\text{HSIC}(X, Y)$ can be obtained by measuring the distance between the embedding of the empirical distribution of the data and the embedding of the product of the marginal empirical distributions. That is, we define $\text{HSIC}(D)$ by

$$\text{HSIC}(D) \coloneqq \left|\left|\frac{1}{n}\sum_{i=1}^{n} K((x_i, y_i), \cdot) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K((x_i, y_j), \cdot)\right|\right|_H^2.$$

Using the reproducing property of the kernel and the definition of K in terms of k and l, $\text{HSIC}(D)$ can be shown to equal

$$\begin{aligned}\text{HSIC}(D) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j)l(y_i, y_j) + \frac{1}{n^2}\sum_{i,j=1} k(x_i, x_j)\frac{1}{n^2}\sum_{q,r=1}^{n} l(y_q, y_r) \\ - \frac{2}{n^3}\sum_{i,j,r=1}^{n} k(x_i, x_j)l(y_i, y_r).\end{aligned}$$

While the biased $\text{HSIC}(D)$ defined above is one of the most commonly used estimator (Szekély and Rizzo (2009),Gretton et al. (2008)), unbiased estimators of $\text{HSIC}(P)$ exist too (Song et al. (2012).) The bias of $\text{HSIC}(D)$ is $O(n^{-1})$, (Gretton et al. (2008)) and for appropriate choices of kernels, the permutation test with the biased statistic and the permutation test with the unbiased statistic are consistent tests - see section 3.2.2.1 for details. Both tests also have correct type 1 error rate. Because the biased $\text{HSIC}(D)$ is more common and has a slightly easier analytic form, we use $\text{HSIC}(D)$

throughout our paper. The following algorithm shows how HSIC is commonly combined with a permutation test for independence testing.

Algorithm 1: Independence testing using HSIC and a permutation test

Input : Observed data $D = ((x_i, y_i))_{i=1}^n$, significance level α, number of permutations B.

1 Sample $\pi_j,\ 1 \leq j \leq B$ distributed uniformly and independently from the symmetric group S_n of all permutations on n elements. Denote $\pi D \coloneqq ((x_{\pi(i)}, y_i))_{i=1}^n$.;

2 Breaking ties at random, compute the rank R of $\mathrm{HSIC}(D)$ in the vector

$$(\mathrm{HSIC}(D), \mathrm{HSIC}(\pi_1 D), \mathrm{HSIC}(\pi_2(D)), \ldots, \mathrm{HSIC}(\pi_B D))$$

where $R = 1$ is the rank of the largest element and $R = B + 1$ is the rank of the smallest element. ;

Output: Reject H_0 if $p \coloneqq R/(B+1) \leq \alpha$, otherwise accept H_0.

Figure 3.1: The algorithm for independence testing using HSIC and a permutation test, resulting in a p-value and a rejection decision.

3.2.2.1 Choice of kernel

Throughout this paper we assume the covariates take values in the Euclidean space $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$. Note that in our case $\mathcal{Y} = \mathbb{R}_{\geq 0}$. We let both k and l be instances of the covariance kernel of Brownian motion. That is,

$$k(x, x') = \left(\|x\| + \|x'\| - \|x - x'\|\right),$$

and

$$l(t, t') = \left(|t| + |t'| - |t - t'|\right),$$

where $\|\cdot\|$ denotes the Euclidean norm. See Sejdinovic et al. (2012) for a discussion of this kernel.

The reason for this choice is three-fold. Firstly, under this choice of kernels, HSIC

coincides with Distance Covariance (DCOV) (Szekély and Rizzo (2009)). The equivalence between HSIC and DCOV was proved in Sejdinovic et al. (2012). DCOV is a well studied measure of dependence and if X, T are random variables with compact support, then $\text{DCOV}(X,T) = 0$ if and only if $X \perp\!\!\!\perp T$. Furthermore, the permutation test with test statistic DCOV is consistent in the sense that, for each distribution P_{XY} with compact support and such that $X \not\perp\!\!\!\perp Y$, the probability that the permutation test rejects the null hypothesis converges to 1, as the sample size converges to infinity (see Rindt et al. (2021)). Secondly, unlike with potential kernels, such as the Gaussian kernel, the Brownian kernel spares us the need to select a bandwidth. Thirdly, our test relies on optimal transport to transform the original data into a transformed dataset. The optimal transport procedure, discussed in the next section, requires a metric with respect to which similarity is preserved. It appears natural to measure the dependency in the transformed dataset using the same metric as the one underlying the transformation.

3.2.3 Optimal transport

Let $A = [a_1, \ldots, a_{|A|}]$ and $B = [b_1, \ldots, b_{|B|}]$ be multisets of length $|A|$ and $|B|$ with $a_i, b_j \in \mathbb{R}^d$. Let v^A be a distribution vector on a, by which we mean: $v^A = (v_1^a, \ldots, v_{|a|}^a) \in \mathbb{R}^{|a|}$ so that $v_i^A \geq 0$ and $\sum_{i=1}^{|A|} v_i^A = 1$. Let v^B be a distribution vector on B. Let $m(a,b)$ be a metric on $\mathbb{R}^d$. Then the earth mover's distance (EMD) between v^A and v^B is defined as

$$\min_{P_{ij} \in \mathbb{R}^{|A| \times |B|}} \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} P_{ij} m(a_i, b_j) \tag{3.2}$$

subject to

$$
\begin{aligned}
P_{ij} &\geq 0 && i = 1, \ldots, |A|,\ j = 1, \ldots, |B|, \\
\sum_{j=1}^{|B|} P_{ij} &= v_i^A && i = 1, \ldots, |A|, \\
\sum_{i=1}^{|A|} P_{ij} &= v_j^B && j = 1, \ldots, |B|.
\end{aligned}
$$

Let U be a random variable taking values in A with distribution v^A, and U' a random variable taking values in B with distribution v^B. In the next section we will write 'let P be the joint distribution that solves the optimal transport problem between U and U''. By that we mean $P(U = a_i, U' = b_j) \coloneqq P_{ij}$ for $i = 1, \ldots, |A|, j = 1, \ldots, |B|$, where P_{ij} is the matrix that minimizes Equation 3.2. Note that P is a valid joint distribution of U and U'. Furthermore, for any i with $v_i^A > 0$ it holds that $P(U' = b_j | U = a_i) = P_{ij}/v_i^A$.

3.3 Data transformation based on optimal transport

3.3.1 Objective of the algorithm

Kernel-based independence tests start by calculating the kernel between every pair of observations, so there is no straightforward way to apply them to right-censored data. In this section we propose a transformation of the original, censored dataset, into a synthetic dataset consisting of n observed events. The algorithm uses optimal transport and its goal is twofold: first, it should return a dataset to which we can apply a permutation test with test-statistic HSIC, and obtain correct p-values under

the null hypothesis; second, it should return a dataset in which the dependence between X and T is similar to the dependence in the original dataset.

We believe that the main value of the transformation lies in how it can be straightforwardly combined with permutation testing on transformed data and we expect it may be difficult to use the transformation for other purposes. Finally, while in this work we use the test statistic HSIC/DCOV, other statistics that measure the strength of the dependence could be considered too.

3.3.2 Definition of the transformation

To define the algorithm, we use the following notation. Let $A = [a_1, \ldots, a_{|A|}]$ where $a_i \in \mathbb{R}^d$ for $i = 1, \ldots, |A|$. We define $\text{Uniform}(A) = \sum_{a \in A} \delta(a)/|A|$, to be the uniform distribution over the elements, where $\delta(a)$ is the Dirac measure at a. By $B = \text{Remove}(a, A)$ we set B to be the multiset that remains when one instance of a is removed from A. Assume that we are given a sample $D = ((x_i, z_i, \delta_i))_{i=1}^n$ so that $z_1 < z_2 < \cdots < z_n$: that is, we have n distinct event times and they are labeled in increasing order. For a variable a we use $a \leftarrow b$ to change the value of a to b. In computing the optimal transport coupling as defined in Section 3.2.3, we use the Euclidean metric $\|x - y\| = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$ for $x \in \mathbb{R}^d$. The proposed transformation is defined in Algorithm 2.

Algorithm 2: Optimal transport based transformation of D to $\tilde{D}$.

Input : $D = ((x_i, z_i, \delta_i))_{i=1}^n$, with $z_1 < \cdots < z_n$.

Define the multisets $\mathrm{L} = \mathrm{AR} = [x_1, \ldots, x_n]$ and $\tilde{D} = [\,]$. (Note $\tilde{D}$ is an empty multiset). Set $d = 0$;

for $i = 1, \ldots, n-1$ **do**

  if $\delta_i = 1$ **then**

    $d \leftarrow d + 1$;

    Compute the joint distribution matrix $P \in \mathbb{R}^{|\,\mathrm{AR}\,| \times |\,\mathrm{L}\,|}$ that solves the optimal transport problem between $U \sim \mathrm{Uniform(AR)}$ and $U' \sim \mathrm{Uniform(L)}$;

    Condition P on $U = x_i$ to get distribution vector $v \in \mathbb{R}^{|\,\mathrm{L}\,|}$ (see Section 3.2.3);

    Sample from L with distribution v. Denote the chosen element by $\tilde{x}$;

    $\tilde{D} \leftarrow \mathrm{Add}((\tilde{x}, z_i), \tilde{D})$;

    $\mathrm{L} \leftarrow \mathrm{Remove}(\tilde{x}, \mathrm{L})$;

  end

  $\mathrm{AR} \leftarrow \mathrm{Remove}(x_i, \mathrm{AR})$;

end

for $i = d+1, \ldots, n$ **do**

  Let $\tilde{x}$ be any (it does not matter which) element of L.;

  $\tilde{D} \leftarrow \mathrm{Add}((\tilde{x}, z_n), \tilde{D})$;

  $\mathrm{L} \leftarrow \mathrm{Remove}(\tilde{x}, \mathrm{L})$;

end

Output: The transformed dataset $\tilde{D}$, consisting of n points.

Figure 3.2: The optHSIC algorithm, resulting in a transformed dataset $\tilde{D}$

3.3.3 Comments on the transformation algorithm

We now give a verbal explanation of Algorithm 2.

Initialization: The input is simply $D = ((x_i, z_i, \delta_i))_{i=1}^n$ where $z_1 < \cdots < z_n$. We initialize an empty transformed dataset $\tilde{D}$, to which we will add n observations of the form (x_i, t_j) for some $i, j \in \{1, \ldots, n\}$ in the following way. We will loop over the times z_i from $i = 1, \ldots, n$. At each time, the multiset AR lists the covariates at risk in the dataset D and AR is initialized as the multiset containing all covariates. The multiset L will list all covariates that have not been added to $\tilde{D}$ yet. Indeed, as $\tilde{D}$ is initialized empty, L initially contains all n covariates. The variable d will count the

number of observed $\delta = 1$ events and is initialized 0.

First loop from $i = 1, \ldots, n-1$: At time z_i we distinguish two cases: if $\delta_i = 0$ we leave L and $\tilde{D}$ unchanged, and simply remove one instance of x_i from AR. If $\delta_i = 1$ we add 1 to d (to count the number of observed events). We also select a covariate $\tilde{x}$ from L as follows: First a joint distribution of Uniform(AR) and Uniform(L) is computed using optimal transport. This is a matrix P of size $|\,\mathrm{AR}\,| \times |\,\mathrm{L}\,|$. This distribution is then conditioned on the event that Uniform(AR) $= x_i$, yielding a distribution over L. We sample from this distribution to obtain the covariate $\tilde{x}$. The pair $(\tilde{x}, z_i)$ is then added to $\tilde{D}$, and the covariate $\tilde{x}$ is removed from L. There are now d observations in $\tilde{D}$ and there are $n - d$ covariates left in L. We also remove x_i from AR.

When we finish the first loop: It is now the case that AR $= [x_n]$ (note the previous loop went up to $i = n-1$) and the multiset L contains $n-d$ covariates. In the second loop, that runs from $d+1$ to n, each remaining covariate in L is combined with the time z_n and added to $\tilde{D}$. After this loop, L is empty, and $\tilde{D}$ contains all n covariates $x_1, \ldots, x_n$, each associated with a time. The transformed dataset $\tilde{D}$ is returned.

Consider the special case in which there is no censoring and $\delta_i = 1$ for $i = 1, \ldots, n$. Then it is easy to verify that in the first loop AR = L at each step i, so that the optimal transport algorithm returns a scaled identity matrix. Consequently, at step i the covariate $\tilde{x}$ is equal to x_i with probability 1, and the final output is $\tilde{D} = D$. A nontrivial example is worked out in Section 3.8.1.

3.3.4 Intuition behind the transformation

Before we prove properties of the proposed transformation, we briefly comment on the intuition behind the transformation. To this end, first consider a permutation test in the absence of censoring, when we simply observe $D = ((x_i, t_i))_{i=1}^n$. The permuted datasets can be generated as follows: loop through the events in order of time, and to each time, associate a covariate that you have not associated to

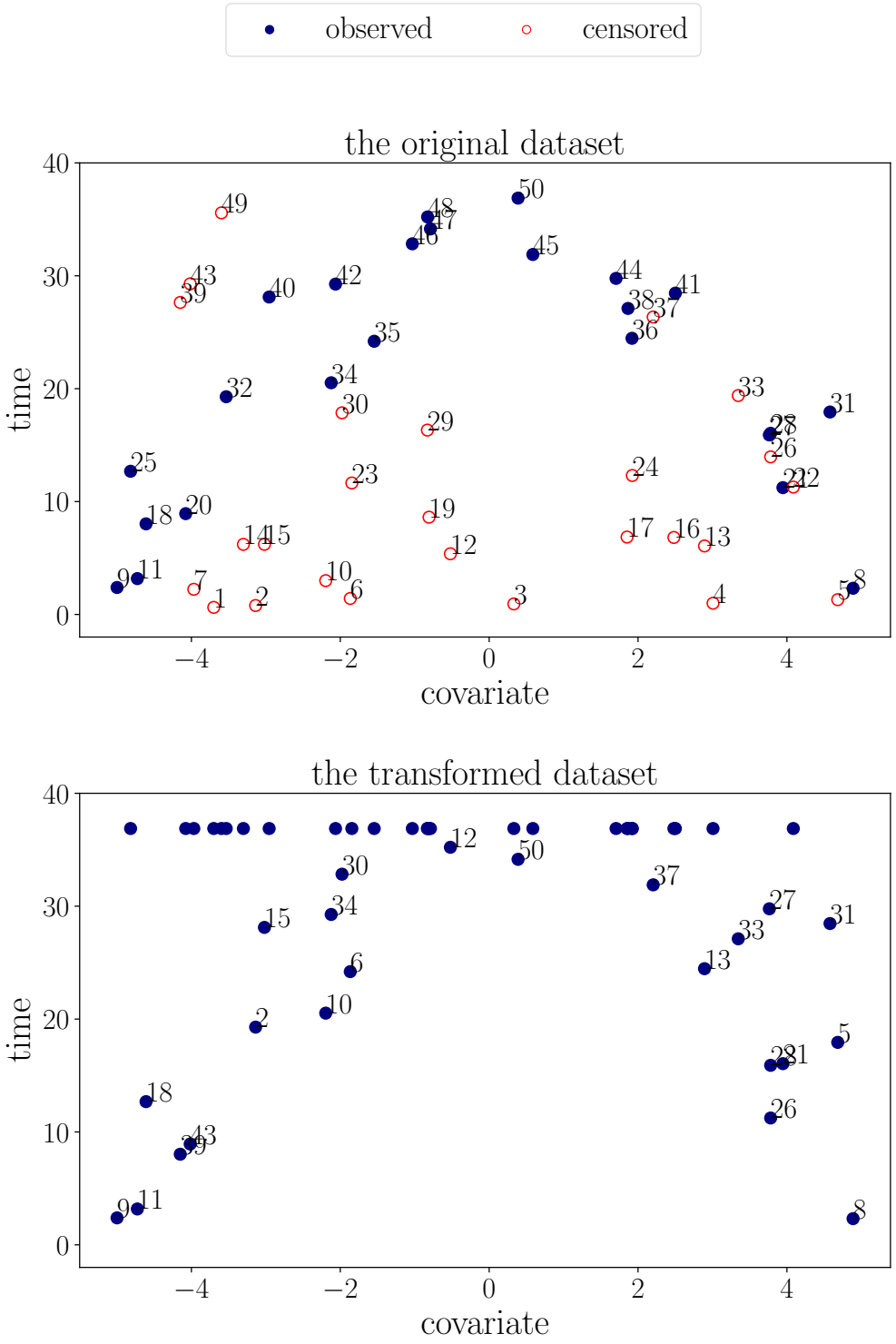


Figure 3.3: Above: The dataset as originally observed. Below: the synthetic dataset after applying the OPT-based transformation. The labels indicate which individual the observation corresponds to in each dataset. The lifetimes are sampled from a parabolic relationship between covariates and lifetimes.

any earlier time. Comparing the original dataset with the datasets obtained in this manner is justified for the following reason: *Under the null hypothesis* the original dataset can be generated by firstly sampling t_i for $i = 1, \dots n$ i.i.d. and independently sampling x_i for $i = 1, \dots n$ i.i.d. and secondly looping through the times in order, and associating to each time a covariate that has not yet been chosen at a previous time. The original dataset and the permuted datasets are thus equal in distribution: intuitively, the permutation test checks whether the dataset looks as if, at each time, a covariate is picked uniformly from those not chosen before.

It is not obvious how to translate this to censored data. Due to censoring, it may not be true that the i-th event covariate is chosen uniformly from the covariates that have not had an observed event time before. For this reason survival analysis often compares the i–th event covariate with the covariates at risk (not failed and not censored) just before the i–th event. If the null hypothesis holds, then intuitively it holds that the i-th event covariate is chosen uniformly *from the covariates at risk* at time z_i. That is, if AR denotes the multiset of covariates at risk and x_i is the event covariate, we would like to test if x_i was chosen uniformly from AR.

To do so using a permutation test, our algorithm couples Uniform(AR) to a uniform choice from those that have not yet been chosen in the synthetic dataset: uniform(L) (see algorithm 2). In this manner, when the null hypothesis is true, (intuitively) it holds that in the synthetic dataset the covariates are chosen uniformly from those not chosen before. But this last statement is exactly our intuition behind a permutation test: namely, we use a permutation test to see if the dataset looks as if, at each time, a covariate is picked uniformly from those not chosen before.

The intuition discussed thus far related to the workings of our transformation under the null hypothesis, and had at its core that Uniform(AR) is coupled to Uniform(L) in Algorithm 2. However there are many couplings between these two distributions. The reason we choose the coupling that solves the optimal transport problem is

the following: assume the alternative hypothesis holds and that given AR_i, certain covariates have a higher hazard rate than others. Then we would like this bias towards these covariates to be visible in $\tilde{D}$. In other words we would like $\tilde{x}$ to be close to $\tilde{x}_i$ in Algorithm 2. That is precisely what the optimal transport coupling achieves.

Finally, including the remaining covariates in the transformed dataset at the last time ensures the permuted datasets correctly reflect all alternative choices that can be made when covariates are chosen at random from those not chosen before. The association of these remaining covariates to the last event time reflects they have not been selected at each of the earlier times, which may indicate they have had less risk of having an event up to that point.

The goal of the transformation is thus twofold: firstly, ensuring that under the null hypothesis a permutation test is appropriate on the transformed dataset, which motivates the coupling of Uniform(AR) and Uniform(L); secondly, given the first goal, ensure covariates in $\tilde{D}$ associated to times z_i for $\delta_i = 1$ are as close as possible to the covariates associated to z_i in D, which motivates the chosen coupling to be the coupling that solves the optimal transport problem defined in Section 3.2.3.

An interesting alternative approach would be to compare x_i with $\text{Uniform}(\text{AR}_i)$ directly, without first transforming the data. We do not see, however, how to combine that approach with a permutation test, or with another procedure to test for significance.

3.4 Applying HSIC to the transformed dataset: optHSIC

We have thus far described how to transform the dataset. For the hypothesis test of $H_0 : X \perp\!\!\!\perp T$, we propose to apply a permutation test with test statistic DCOV/HSIC to the transformed dataset. This approach is summarized in Algorithm 3.

Algorithm 3: optHSIC

Input : Observed data $D = ((x_i, z_i, \delta_i))_{i=1}^n$, significance level α, number of permutations B.

1 Apply Algorithm 2 to D to obtain the transformed (and uncensored) dataset $\tilde{D}$.;

2 Apply the standard HSIC permutation test of Algorithm 1 to $\tilde{D}$ with level α and number of permutations B, resulting in a p-value. ;

Output: The p-value resulting from the permutation test. Reject H_0 if $p \leq \alpha$.

Figure 3.4: The optHSIC algorithm, resulting in a p-value and rejection decision of $H_0 : X \perp\!\!\!\perp T$.

3.4.1 Computational cost of optHSIC

The algorithm optHSIC consists of two parts: First, the dataset is transformed. Second, a permutation test with test statistic HSIC/DCOV is performed. The computational cost of the transformation may be high: the solution of the earth mover's distance has $n^3 \log(n)$ complexity (Shirdhonkar and Jacobs (2008)). Since the earth mover's distance is computed for each $i = 1, \ldots, n$ such that $\delta_i = 1$, complexity of the transformation is $O(n^4 \log(n))$. Luckily, fast approximations of the earth mover's distance exist that are linear in time (e.g. Shirdhonkar and Jacobs (2008)), making the transformation $O(n^2)$. Also in the case where the covariates are 1-dimensional, the optimal transport problem has a simple solution (Cohen (1999)). Different algorithms and/or approximations also exist for the EMD with other metrics (Shirdhonkar and Jacobs (2008)). The second step, computation of HSIC is also an $O(n^2)$-operation for which large-scale approximations can be made (Zhang et al. (2018)). Furthermore, both the computation of HSIC and the permutation test computations can be easily parallelized. In our simulations we did not use approximations of the earth mover's distance nor of HSIC, as for samples up to about 500 the optHSIC test can be performed in a few seconds on an ordinary PC.

3.4.2 Theoretical results on optHSIC

We say a test with p-value p has correct type 1 error rate if under the null hypothesis $P_{H_0}(p \leq \alpha) \leq \alpha$. The main theoretical result we obtained for optHSIC is that the type 1 error rate is correct when $C \perp\!\!\!\perp X$. Unfortunately, we have not been able to prove other important results such as (asymptotically) correct type 1 error rate when $C \not\perp\!\!\!\perp X$, or consistency (power converging to 1 for each alternative hypothesis). However, as Section 3.6 details, extensive simulations demonstrate that optHSIC achieves the correct type 1 error rate also when $C \not\perp\!\!\!\perp X$. Simulations further show that optHSIC is able to detect a wide range of dependencies between X and T without losing much power, relative to the CPH likelihood ratio test, even when the CPH model assumptions holds. Obtaining further theoretical results is an important future challenge. In the uncensored case a permutation test with Brownian kernel is consistent for distributions with compact support and has correct type 1 error (See for example Rindt et al. (2021)). The difficulty of extending these proofs to optHSIC is that optHSIC requires sequential analysis of the optimal transport distributions, breaking independence between observations in the transformed dataset.

We first prove an auxiliary result: namely, although we propose to permute the transformed dataset, this is equivalent to permuting the original dataset, and then transforming the permuted datasets.

Lemma 3.3. *Let $\pi_1, \ldots, \pi_B$ be independent uniform random permutations, and let $T, T_1, \ldots, T_B$ be independent optHSIC transformations.*

$$[T(D), \pi_1(T(D)), \ldots, \pi_B(T(D))] \stackrel{d}{=} [T(D), T_1(\pi_1(D)), \ldots, T_B(\pi_B(D))]$$

Proof. See Appendix. □

Note that this implies that in the definition of optHSIC we could have equally per-

muted D first, and then applied the transformation to each of the permuted datasets. However this is computationally more expensive. Lemma 3.3 enables us to show that optHSIC has correct type 1 error rate when $C \perp\!\!\!\perp X$.

Theorem 3.4. *Let $D = ((X_i, Z_i, \Delta_i))_{i=1}^n$ be an i.i.d. sample. Assume that $C \perp\!\!\!\perp X$. Let p be the p-value resulting from applying optHSIC (Algorithm 3) to D with B permutations and level $\alpha \in [0, 1]$. If the null hypothesis holds, i.e. $T \perp\!\!\!\perp X$, then $p \sim$ Uniform$[\frac{1}{B+1}, \ldots, \frac{B+1}{B+1}]$ and in particular it holds that*

$$P_{H_0}\big(p \leq \alpha) \leq \alpha.$$

Proof. See Appendix. □

Table 3.1 summarizes our findings of the type 1 error of optHSIC.

	Correct type 1 error rate in theory: $P_{H_0}(p \leq \alpha) \leq \alpha$.	**Distribution p under H_0**	**Correct type 1 error rate in simulations**
$C \perp\!\!\!\perp X$	Yes (Theorem 3.4)	Uniform$[\frac{1}{B+1}, \ldots, \frac{B+1}{B+1}]$ (Theorem 3.4)	Yes (Sec 3.6.1)
$C \not\perp\!\!\!\perp X$	Unknown	Unknown, but experiments suggest approximately Uniform$[\frac{1}{B+1}, \ldots, \frac{B+1}{B+1}]$ (Figure 3.10)	Yes (Sec 3.6.1)

Table 3.1: An overview of the type 1 error results obtained for optHSIC in the case when $C \perp\!\!\!\perp X$ and when $C \not\perp\!\!\!\perp X$. B denotes the number of permutations. The p-value is defined in Algorithm 3.

3.4.3 Using multiple transformations

Algorithm 2 defines the transformation used in optHSIC. In line 7, a covariate is sampled from the multiset L with distribution v. The sampled element may differ for different iterations of the algorithm. Consequently, given a fixed dataset D, the

transformed dataset will look different for different iterations of the algorithm. The optHSIC algorithm proposes to use only one transformation, resulting in a single p-value.

Say that instead of applying the optHSIC algorithm once, we repeat the optHSIC algorithm m times, resulting in m p-values $p_1, \ldots, p_m$. An example of the distribution of p-values for a fixed dataset is given in Figure 3.5.

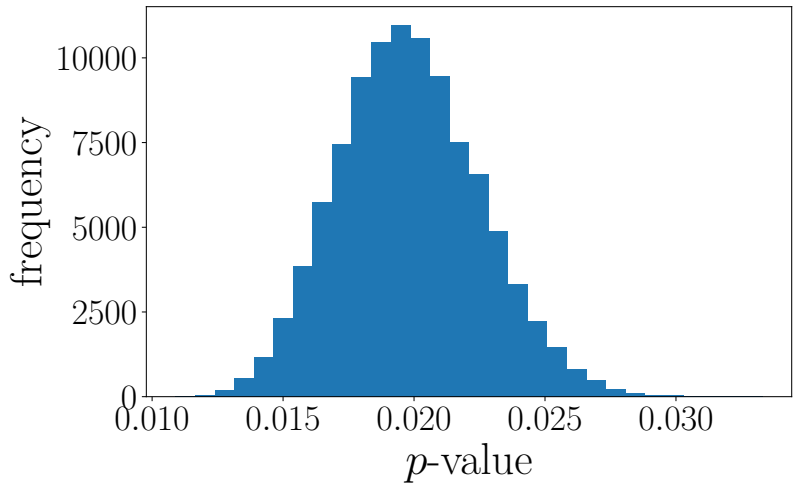


Figure 3.5: A histogram of the p-values obtained from optHSIC for a fixed dataset D sampled from D.1 of Table 3.4. The sample was of size $n = 300$ and 61% of the individuals was observed ($\delta = 1$). The variability in p comes from the fact that the transformation in optHSIC is random.

Figure 3.5 shows that in the used dataset, where 39% of observations are censored and $n = 300$, the variance in the obtained p-values is not excessively high, and in particular all p-values would have led to the decision to reject the null-hypothesis.

We also studied several techniques one may use to combine p-values, typically at the cost of being more conservative when there is little variance among p-values, like in the example above. This is discussed in Section 3.8.9.

3.5 Alternative approaches when censoring is independent of the covariate

We are not aware of fully nonparametric methods to test independence between right-censored times and continuous covariates. When C is independent of X the challenge is mitigated, since in that case any statistic can be combined with a permutation test that permutes the covariates (see Theorem 3.6). Using that approach, this Section proposes some alternative approaches in the case when $C \perp\!\!\!\perp X$ that will serve as benchmarks for studying power performance of optHSIC, additionally to the Cox proportional hazards likelihood ratio test. In Section 3.6 we compare the performance of these methods with optHSIC.

Before we propose alternative test statistics, we state formally why standard permutation tests can be used when $C \perp\!\!\!\perp X$. We first state a Theorem on the correctness of the level of permutation tests without censoring. We include the proof following Berrett and Samworth (2019) in the Appendix for completeness.

Theorem 3.5. *Let $D = (X_i, Y_i)_{i=1}^n$ be an i.i.d. sample of size n from distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$. Denote $\pi D = (X_{\pi(i)}, Y_i)_{i=1}^n$. Let $\pi_1, \dots \pi_B$ be permutations sampled uniformly and independently from S^n. Let H be any statistic of the data. Let R be the rank of the first coordinate in the vector:*

$$\big(H(D), H(\pi_1 D), \dots, H(\pi_B D)\big)$$

when ties are broken at random and where $R = 1$ is the rank of the largest element and $R = B + 1$ the rank of the smallest element. Define $p \coloneqq R/(B+1)$. Then under the null hypothesis $H_0 : X \perp\!\!\!\perp T$ it holds that: $p \sim$ Uniform$[\frac{1}{B+1}, \dots, \frac{B+1}{B+1}]$. So in

particular for $\alpha \in [0,1]$

$$P_{H_0}(p \leq \alpha) \leq \alpha.$$

Proof. See Appendix. □

In survival analysis we apply the above, setting $Y = (Z, \Delta)$, where $Z = \min\{T, C\}$ and $\Delta = 1\{T \leq C\}$.

Theorem 3.6. *Assume* $C \perp\!\!\!\perp X$ *and* $T \perp\!\!\!\perp C|X$. *Write* $D = (X_i, Z_i, \Delta_i)_{i=1}^n$, *and let* $\pi_1, \ldots, \pi_B$ *be sampled i.i.d. uniformly from* S_n. *Write* $\pi D := (X_{\pi(i)}, Z_i, \Delta_i)_{i=1}^n$. *Let* $H(D)$ *be any statistic of the data. Let* R *be the rank of the first coordinate in the vector:*

$$\big(H(D), H(\pi_1 D), \ldots, H(\pi_B D)\big)$$

where ties are broken at random and $R = 1$ *is the rank of the largest element and* $R = B+1$ *is the rank of the smallest element. Define* $p \coloneqq R/(B+1)$. *Then under the null hypothesis* $H_0 : X \perp\!\!\!\perp T$ *it holds that* $p \sim$ *Uniform*$[\frac{1}{B+1}, \ldots, \frac{B+1}{B+1}]$. *So in particular for* $\alpha \in [0,1]$

$$P_{H_0}(p \leq \alpha) \leq \alpha.$$

Proof. See Appendix. □

Having shown that we can use a permutation test on any test statistic when $C \perp\!\!\!\perp X$, the next two sections explore meaningful measures of dependency on right-censored data. Indeed such methods will lead to an excessive rate of type 1 errors when $C \not\perp\!\!\!\perp X$.

3.5.1 wHSIC

HSIC relies on estimating the mean embedding of the joint distribution P_{XT}. The estimated embedding is then compared to the embedding of the product of marginal distributions. In the uncensored case the empirical distribution equals $\sum_{i=1}^{n} \delta_{(x_i,t_i)}/n$ which has corresponding mean embedding $\sum_{i=1}^{n} K((x_i,t_i),\cdot)/n$.

Since we do not observe the t_i's we could consider replacing the empirical distribution by a weighted version $\sum_{i=1}^{n} w_i\delta_{(x_i,z_i)}/n$ where we try to find weights w_i such that $\sum_{i=1}^{n} w_i f(x_i,z_i) \approx Ef(X,T)$ for every measurable function f such that the expectation exists. A natural idea is to give an observed point (x,z,δ) a weight of zero if it is censored, and a weight equal to the inverse probability of being uncensored otherwise. This can be motivated by the following lemma, that applies also if $C \not\perp\!\!\!\perp X$. We first define

$$g(t,x) = P(C > t|X = x).$$

The following lemma proposes a weight function W in terms of g.

Lemma 3.7. *Let f be an (X,T)- measurable function such that $Ef(X,T) < \infty$ and define W by*

$$W = \begin{cases} 0 & \text{if } \Delta = 0 \\ \frac{1}{g(Z,X)} & \text{if } \Delta = 1. \end{cases}$$

Then $EWf(X,Z) = Ef(X,T)$.

Proof. See Appendix. □

We would thus like to use weights $w_i = 0$ if $\delta_i = 0$ and $1/ng(x_i,z_i)$ if $\delta = 1$. As we will be working under the assumption $C \perp\!\!\!\perp X$ we may estimate $1/P(C > t|X) = 1/P(C > t)$ using Kaplan–Meier weights, which we define now. Assume that there are no ties in the data and that $z_i < z_{i+1}$ for $i = 1,\ldots,n$. Then the Kaplan–Meier

survival curve is given by:

$$\hat{S}_T(z_k) = \prod_{i=1}^{k} \left(\frac{n-i}{n-i+1} \right)^{\delta_i}.$$

Kaplan–Meier weights are then defined by

$$w_k = \hat{P}(T = z_k) = \hat{S}_T(z_{k-1}) - \hat{S}_T(z_k).$$

One can also estimate the survival probability of the censoring time using Kaplan–Meier weights (simply replace δ_i by $1-\delta_i$ in the above formula). The following lemma shows that the weights w_k defined above, correspond up to a constant of $1/n$, to the inverse of the estimated probability of being uncensored.

Lemma 3.8. *Let $\hat{S}_C(z_k) = \hat{P}(C > z_k)$ denote the Kaplan–Meier estimate of the survival probability of the censoring distribution. Then:*

$$w_k = \frac{1}{n} \frac{1}{\hat{P}(C > z_k)}$$

Proof. See Appendix. □

We use these weights to define the statistic wHSIC as the RKHS distance between the embedding of $\sum_{i=1}^{n} w_i \delta_{(x_i,z_i)}$ and the embedding of the product of the marginals, $\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \delta_{x_i} \delta_{z_j}$. That is:

$$\text{wHSIC}(D) \coloneqq \left\| \sum_{i=1}^{n} w_i K\big((x_i, z_i), \cdot\big) - \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j K\big((x_i, y_j), \cdot\big) \right\|_H^2.$$

Here $K\big((x,y),(x',y')\big) = k(x,x')l(y,y')$. The following theorem shows how to compute wHSIC efficiently.

Algorithm 4: wHSIC

Input : $D = ((x_i, z_i, \delta_i))_{i=1}^n$, significance level α, number of permutations B.

1 Sample permutations $\pi_1, \ldots, \pi_B$ i.i.d. uniformly from S_n.;

2 Breaking ties at random, compute the rank R of wHSIC(D) in the vector

$$(\text{wHSIC}(D), \text{wHSIC}(\pi_1 D), \text{wHSIC}(\pi_2 D), \ldots, \text{wHSIC}(\pi_B D))$$

where $R = 1$ is the rank of the largest element. ;

Output: The p-value $p := R/(B+1)$. Reject H_0 if $p \leq \alpha$.

Figure 3.6: The wHSIC algorithm, resulting in a p-value and rejection decision of $H_0 : X \perp\!\!\!\perp T$.

Theorem 3.9. *(Computation of wHSIC) Given a dataset $D = ((x_i, y_i))_{i=1}^n$ with a weight vector $w \in \mathbb{R}^n$,*

$$\text{wHSIC}(D) = \text{Tr}\left(H_w K H_w L\right),$$

where $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$ and $H_w = D_w - ww^\top$, where $D_w = diag(w)$, a diagonal matrix with $(D_w)_{ii} = w_i$.

Proof. See Appendix. □

It is not difficult to see wHSIC has the same computational time as (uncensored) HSIC.

3.5.2 zHSIC

If $C \perp\!\!\!\perp X$, then any dependence between X and Z must be due to dependence between X and T. Hence we may estimate HSIC(X, Z) to measure the strength of the dependence. This test is, indeed, expected to yield false rejections if $C \perp\!\!\!\perp X$ fails to hold, and to lack power when a large portion of the events is censored. We include this test mainly to see how the power of optHSIC and wHSIC compare with this approach.

Algorithm 5: zHSIC

Input : $D = ((x_i, z_i, \delta_i))_{i=1}^n$, significance level α, number of permutations B.

1 Denote $\tilde{D} = ((x_i, z_i))_{i=1}^n$ the sample as if there was no censoring. Sample permutations $\pi_1, \ldots, \pi_B$ i.i.d. uniformly from S_n.;

2 Breaking ties at random, compute the rank R of HSIC(D) in the vector

$$\Big(\text{HSIC}(\tilde{D}), \text{HSIC}(\pi_1\tilde{D}), \text{HSIC}(\pi_2\tilde{D}), \ldots, \text{HSIC}(\pi_B\tilde{D})\Big)$$

where $R = 1$ is the rank of the largest element. ;

Output: The p-value $p \coloneqq R/(B+1)$. Reject if H_0 if $p \leq \alpha$.

Figure 3.7: The zHSIC algorithm, resulting in a p-value and rejection decision of $H_0 : X \perp\!\!\!\perp T$.

3.6 Numerical evaluation of the methods

We generate data from various distributions of X, T and C to compare the power and type 1 error rate of optHSIC, wHSIC, zHSIC and CPH. CPH stands for the Cox proportional hazards likelihood ratio test. In each scenario, we let the sample size range from $n = 40$ to $n = 400$ in intervals of 40. To obtain p-values in the three HSIC-based methods we use a permutation test with 1999 permutations. We reject the null hypothesis if our obtained p-value is less than 0.05. In the main paper, we present results of rejection rates under distributions that are chosen such that approximately 60% of the observations are observed ($\delta = 1$). We investigate the rejection rates under varying censoring regimes in Section 3.8.10.2 of the Appendix.

3.6.1 Type 1 error rate

We begin by investigating the type 1 error rate. The distributions in which we test the type 1 error rate are found in Table 3.2. In these distribution the null hypothesis holds, i.e. $X \perp\!\!\!\perp T$ and we consider both cases where $C \perp\!\!\!\perp X$ and $C \not\perp\!\!\!\perp X$, as well as the case of multidimensional covariates. We estimate the rejection rates by sampling

5000 times from each distribution for each sample size and applying the different tests to the samples. The obtained rejection rates of optHSIC are found in Table 3.3. The type 1 error rate of the remaining methods is found in Section 3.8.10.1. The most important finding is that optHSIC is found to have correct type 1 error rate of $\alpha = 0.05$ both when $C \perp\!\!\!\perp X$ and when $C \not\perp\!\!\!\perp X$. In contrast, wHSIC and zHSIC yield many false rejections when $C \not\perp\!\!\!\perp X$, as expected, but have the correct type 1 error rate when $C \perp\!\!\!\perp X$. Investigation of the p-values of optHSIC furthermore showed that p-values are distributed approximately according to Uniform$[\frac{1}{B+1}, \ldots, \frac{B+1}{B+1}]$ under the null hypothesis, even when $C \not\perp\!\!\!\perp X$ (see Figure 3.10 in Section 3.8.10.1).

D.	X	$T\|X$	$C\|X$	% $\delta = 1$
1	Unif$[-1, 1]$	Unif$[0, 1]$	Unif$[0, 1.5]$	66%
2	Unif$[-1, 1]$	Exp(mean $= 5/2$)	Exp(mean $= 5/3$)	40%
3	Unif$[-1, 1]$	Exp(mean $= 2/3$)	Exp(mean $= \exp(X)$)	60%
4	Unif$[-1, 1]$	Exp(mean $= 1.6$)	Exp(mean $= \exp(9X^2)$)	60%
5	Unif$[-1, 1]$	Exp(mean $= 0.9$)	Weib(shape $= 1.75X + 3.25$)	60%
6	Unif$[-1, 1]$	Exp(mean $= 0.9$)	$1 + X$	60%
7	$N_{10}(0, \Sigma)$	Exp(mean $= 0.6$)	Exp(mean $= \exp(1^T X)$)	60%
8	$N_{10}(0, \Sigma)$	Exp(mean $= 0.6$)	Exp(mean $= \exp(X_1/8)$)	60%

Table 3.2: The 8 scenarios in which we simulate the type 1 error rate. $N_{10}(0, \Sigma)$ is a 10-dimensional Gaussian random variable, with 0 denoting a vector of zeros of length 10, and $\Sigma = MM^T$ where M is a 10×10 matrix of i.i.d. $N(0, 1)$ entries. Note that C depends on X in D.3-8.

$n =$	40	80	120	160	200	240	280	320	360	400
D.1	0.047	0.053	0.051	0.050	0.054	0.051	0.045	0.056	0.049	0.048
D.2	0.049	0.051	0.057	0.044	0.053	0.050	0.047	0.049	0.052	0.048
D.3	0.052	0.052	0.048	0.049	0.044	0.050	0.048	0.054	0.054	0.048
D.4	0.050	0.054	0.049	0.054	0.054	0.054	0.050	0.053	0.050	0.046
D.5	0.050	0.056	0.056	0.051	0.053	0.051	0.046	0.055	0.050	0.049
D.6	0.050	0.049	0.048	0.047	0.048	0.057	0.050	0.050	0.048	0.050
D.7	0.050	0.055	0.047	0.048	0.056	0.054	0.054	0.052	0.051	0.048
D.8	0.041	0.050	0.051	0.048	0.056	0.049	0.054	0.050	0.054	0.052

Table 3.3: The rejection rate of optHSIC for the distributions D.1-8 of Table 3.2. Note the type 1 error rate is very close to the level $\alpha = 0.05$ as desired, also in the scenarios where $C \not\perp\!\!\!\perp X$.

3.6.2 Comparison of power

To compare the power of the four methods we consider a number of distributions in which $T \not\perp\!\!\!\perp X$. For each distribution we let the sample size range from $n = 40$ to $n = 400$ in steps of 40. At each sample size we take 1000 samples to estimate the rejection rate (the random covariance matrices are resampled before each sample is taken). In these distributions censoring is independent of the covariate such that the methods wHSIC and zHSIC do not have inflated rejection rate due to dependencies between C and X. Distributions with varying censoring rates and dependent censoring are investigated in Section 3.8.10.2. Parameters are chosen so that that 60% of the observations are uncensored (60% has $\delta_i = 1$), and rejection ranges take a range of values (i.e. to exclude trivial distributions so that each method rejects with probability 1 for each sample size.).

D.	X	$T\|X$	$C\|X$
1	$N(0,1)$	Exp(mean $= \exp(X/6)$)	Exp(mean $= 1.5$)
2	$N(0,1)$	Exp(mean $= \exp(X^2/5)$	Exp(mean $= 2.25$)
3	Unif$[-1,1]$	Weib(shape $= 1.75X + 3.25$)	Exp(mean $= 1.75$)
4	Unif$[-1,1]$	$N(100 - X, 2X + 5.5)$	$82 +$ Exp(mean $= 35$)
5	$N_{10}(0,\Sigma)$	Exp(mean $= \exp(1^T X/30)$)	Exp(mean $= 1.5$)
6	$N_{10}(0,\Sigma)$	Exp(mean $= \exp(X_1/8)$)	Exp(mean $= 1.5$)
7	$N_{10}(0,\Sigma)$	Exp(mean $= \exp(X^2/4)$	Exp(mean $= 10$)
8	$N_{10}(0,\Sigma)$	Exp(mean $= \exp(X_1^2/4 + X_2/7)$)	Exp(mean $= 8$)

Table 3.4: The 8 scenarios in which we test the power in the main text. $N_{10}(0,\Sigma)$ is a 10-dimensional Gaussian random variable, with 0 denoting a vector of zeros of length 10, and $\Sigma = MM^T$ where M is a 10×10 matrix of i.i.d. $N(0,1)$ entries. In these distributions about 60% of the individuals is observed. Distributions with dependent censoring and varying censoring percentages are discussed in Section 3.8.10.2.

3.6.2.1 Power for 1-dimensional covariates

In distributions 1-4 of Table 3.4 the covariate is 1-dimensional. Scatterplots of the samples and rejection rates are displayed in Figure 3.8. Note how in D.1 the CPH assumption holds, so the CPH method suits this example very well. We find however that the rejection rate of optHSIC is very similar to that of the CPH likelihood ratio test (first row, right of Figure 3.8). D.2 features a case in which hazard is highest in the middle, and lower for extreme values of the covariate. The CPH likelihood ratio test does not have power to detect this relationship, and optHSIC is the most powerful method. In D.3 and D.4 the hazard functions of different covariates cross each other. In this case, wHSIC is the top-performing method, but optHSIC is also able to detect the more complicated relationship, while CPH is not able to do so.

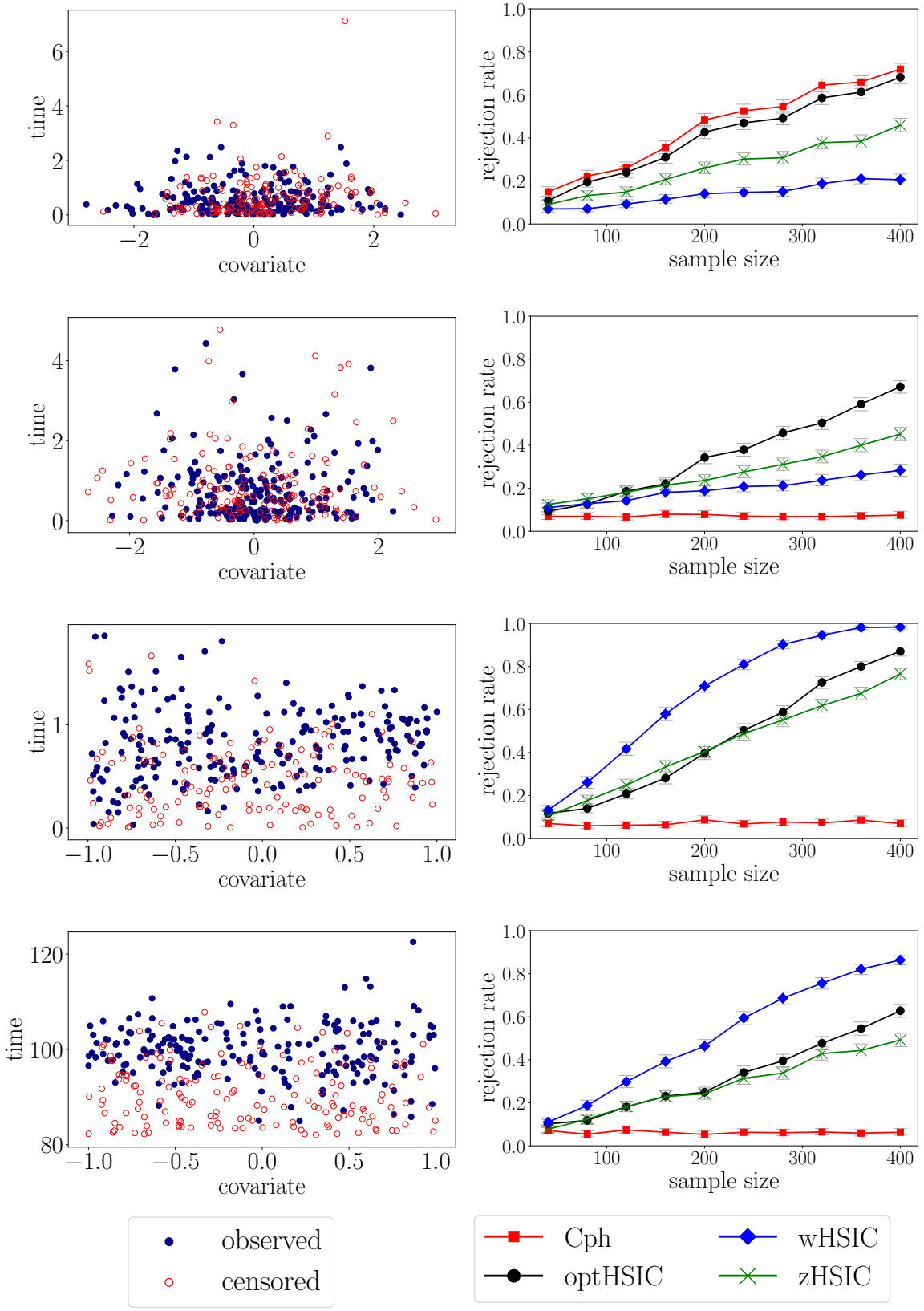


Figure 3.8: Scatterplots and rejection rates for D.1(top row)-D.4(bottom row) of Table 3.4.

3.6.2.2 Power for higher-dimensional covariates

In D.5-8 of Table 3.4 the covariates are multidimensional. Figure 3.9 shows the rejection rates of the four methods, both as the dimension is fixed at 10, and the sample size increases (left column) and when the sample size is fixed at 200, and the dimension increases (right column). In D.5 and D.6 the CPH assumption holds. Again we see that the power of optHSIC is relatively close to the power of the CPH likelihood ratio test, which is specifically designed for these assumptions. This holds both when the dependence is on a single sub-dimension of the covariate (D.6) and when the dependence is on all covariates (D.5). In D.7 and D.8 a non-linear term is present in the hazard rate. Together with zHSIC, optHSIC is now the best performing method. Note that as dimension increases, the dependence in D.6-D.8 becomes harder to detect, whereas the dependence in D.5 becomes easier to detect since T depends on all covariates.

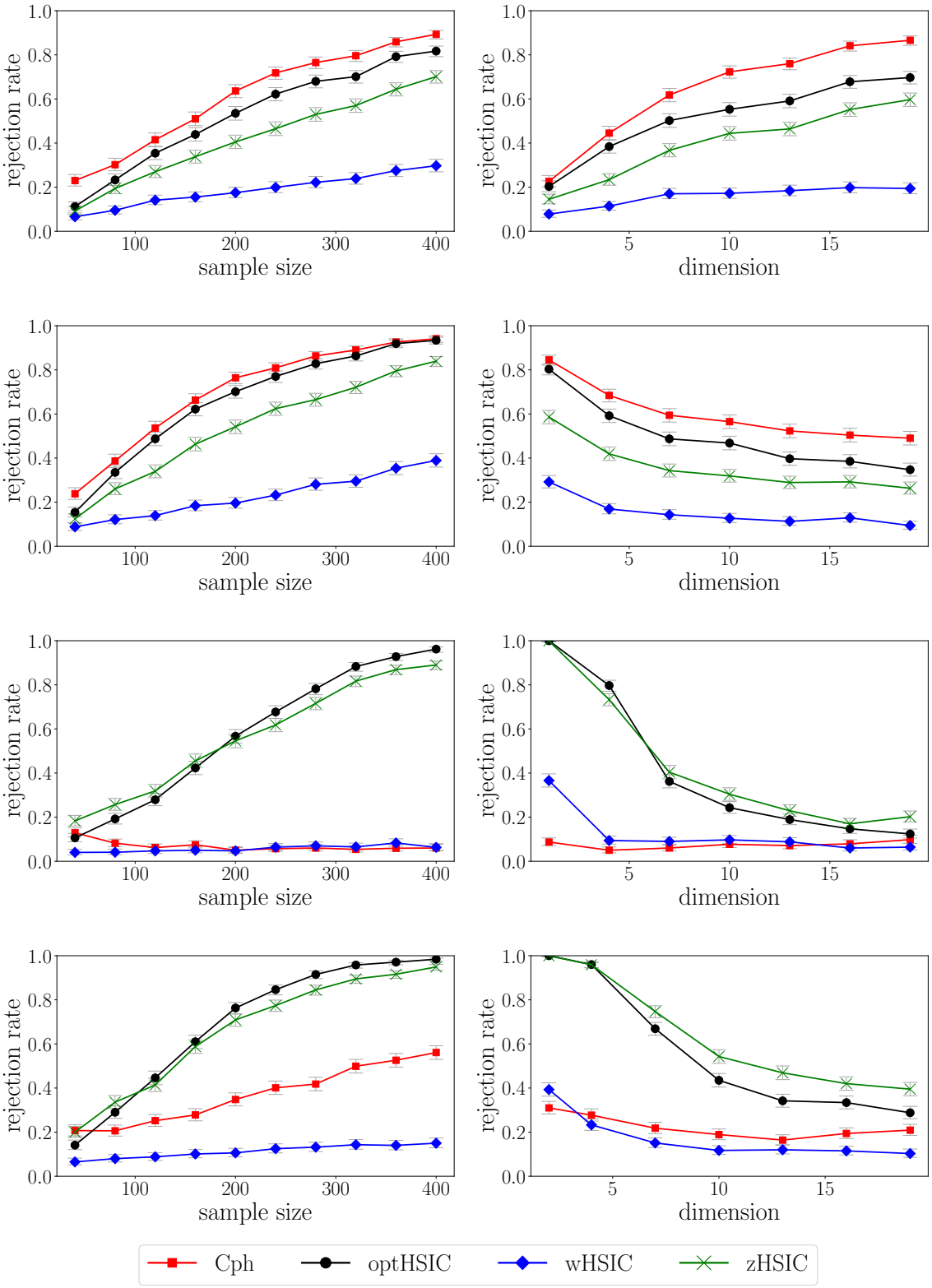


Figure 3.9: Rejection rates of the four methods for D.5 (top row)- D.8(bottom row) of Table 3.4. Left column: dimension fixed at 10, and the sample size increases. Right column: sample size fixed at 200, and the dimension increases.

We also investigated the power when $C \not\perp\!\!\!\perp X$ and as the censoring percentage varies from 20% to 80%. Distributions and rejection rates are presented in Section 3.8.10.2. The main observations are that, although all methods lose power as the censoring rate increases, the relative performance of the four methods remains similar to the results presented in the main text.

We thus find that for continuous covariates optHSIC is able to detect a wider range of dependencies that the CPH likelihood ratio test, while not losing much power when the CPH assumptions hold. This is true both when the covariate is 1-dimensional and when the covariate is multidimensional, even when the dependence arises from a lower-dimensional subspace of the covariate. Importantly, unlike the methods wHSIC and zHSIC, we find that optHSIC has very reliable type 1 error rate: that is, we find the type 1 error rate of optHSIC is found to be of the correct level both when $C \perp\!\!\!\perp X$ and when $C \not\perp\!\!\!\perp X$.

3.6.3 Binary covariates

Lastly, we did experiments with binary covariates, in which independence testing is equivalent to two-sample testing. For two-sample testing there are other alternative approaches, including Ditzhaus and Friedrich (2018) and Matabuena (2019). Furthermore, the wHSIC approach can be modified firstly to allow for the computation of weights within groups, and secondly a permutation test can be constructed that is valid also when censoring differs between groups. This permutation test was proposed by Wang et al. (2010). This special case of wHSIC coincides with the approach discussed in Matabuena (2019) for two-sample testing.

Our main finding is that optHSIC has a decent all-round performance for two-sample testing. However, especially when the distributions meet additional assumptions used in semi-parametric tests, these semi-parametric tests outperform optHSIC. We also note that, as optHSIC relies on choosing 'similar covariates' during the transformation

of Algorithm 2, it is not ideally suited for the two-sample case, as there is a chance of choosing the opposite covariate.

We also provide a 'failure mode' of a scenario in which optHSIC performs much worse than the CPH test. In this example, covariates are binary (i.e. there are two groups), group sizes are very unequal, 90% of observations is censored, censoring appears only in the large group, and survival in the two groups is identical until beyond the censoring has occurred. See 3.8.11.4 for a detailed description. We thank a reviewer for this example.

A full presentation of the experiments and methods used for the case of binary covariates is given in Section 3.8.11. We conclude that the main value of optHSIC lies in independence testing with continuous covariates.

3.7 Discussion

The main contribution of this paper is the proposal of optHSIC, combining a novel way of using optimal transport to transform right-censored datasets, with nonparametric permutation-based independence testing using HSIC/DCOV. We have shown optHSIC has power against a wider range of alternatives than the commonly used CPH model, while forfeiting little power when the CPH assumptions are satisfied exactly. Extensive numerical simulations suggest that the approach is well calibrated, yielding reliable p-values even when censoring strongly depends on the covariate. Under the assumption that censoring does not depend on the covariate, we have proven correct type 1 error rate of optHSIC. Theoretical guarantees for the type 1 error under dependent censoring are a topic of future work. We furthermore proposed reweighting the original dataset, and measuring the distance between the resulting weighted mean embeddings. While these methods showed some promise, they do rely on the very strong assumption that C and X are independent. An interesting future challenge

is to extend such tests to mutual independence testing and conditional independence testing, tests which are used in causal inference, and which are relevant to many problems biostatistics.

3.8 Supplementary material

Supplementary materials contain a worked out example of the transformation proposed in Algorithm 2, proofs of all results and details on additional experiments. They also contain methods of combining p-values from multiple transformation of the same dataset. Code to replicate the experiments and of the tests is available on `www.github.com/davidrindt/opthsic`.

3.8.1 Example of transformation

Let D be the following dataset consisting of $n = 5$ individuals:

i	x_i	z_i	δ_i
1	1.3	13	1
2	0.5	22	0
3	0.3	24	1
4	-1.1	45	1
5	-0.9	81	0

Table 3.5: An example dataset D for which we will demonstrate the transformation.

We initialize $\tilde{D} \leftarrow [\,]$ to be an empty multiset and set $\mathrm{L} \leftarrow [1.3, 0.5, 0.3, -1.1, -0.9]$ and $\mathrm{AR} \leftarrow [1.3, 0.5, 0.3, -1.1, -0.9]$. We loop over the events $i = 1, \ldots, 4$.
At the first time $z_1 = 13$ it holds that $\delta_1 = 1$. We compute the joint distribution $P_{UU'}$ that solves the optimal transport problem between $U \sim \mathrm{Uniform(AR)}$ and $U' = \mathrm{Uniform(L)}$. Since it holds that $\mathrm{AR} = \mathrm{L}$, it follows that $P_{UU'}$ is the matrix:

$$P \leftarrow \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{matrix} \\ \begin{pmatrix} 0.2 & 0. & 0. & 0. & 0. \\ 0. & 0.2 & 0. & 0. & 0. \\ 0. & 0. & 0.2 & 0. & 0. \\ 0. & 0. & 0. & 0.2 & 0. \\ 0. & 0. & 0. & 0. & 0.2 \end{pmatrix} \end{array} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix}$$

Conditioning P on $U = x_1$ yields $v \leftarrow [1, 0, 0, 0, 0]$, corresponding to the first row of P. Sampling from L with distribution v yields $\tilde{x} \leftarrow 1.3$ with probability 1. We update $\tilde{D} \leftarrow [(1.3, 13)]$. We also replace L $\leftarrow [0.5, 0.3, -1.1, -0.9]$ and AR $\leftarrow [0.5, 0.3, -1.1, -0.9]$. We move to the next event time.

At $z_2 = 22$ we note that $\delta_2 = 0$, so we only update AR $\leftarrow [0.3, -1.1, -0.9]$ and leave L and $\tilde{D}$ unchanged.

At the third event $z_3 = 24$ it holds that $\delta_3 = 1$ and AR $= [0.3, -1.1, -0.9]$ and L $= [0.5, 0.3, -1.1, -0.9]$. We couple a random variable $U \sim$ Uniform(AR) and $U' =$ Uniform(L) using optimal transport. The resulting distribution equals:

$$P \leftarrow \begin{array}{c} \begin{matrix} x_2 & x_3 & x_4 & x_5 \end{matrix} \\ \begin{pmatrix} 0.25 & 0.083 & 0. & 0. \\ 0. & 0. & 0.25 & 0.083 \\ 0. & 0.167 & 0. & 0.167 \end{pmatrix} \end{array} \begin{matrix} x_3 \\ x_4 \\ x_5 \end{matrix}$$

We condition this distribution on $U = x_3 = 0.3$. This corresponds to the first row of P, and the resulting distribution over L equals: $v \leftarrow [0.75, 0.25, 0, 0]$. We now sample a point from this distribution and, suppose, it turns out to be $\tilde{x} \leftarrow 0.5 = x_2$, which has 75% chance. We update $\tilde{D} \leftarrow [(1.3, 13), (0.5, 24)]$. We also replace L $\leftarrow [0.3, -1.1, -0.9]$ and AR $\leftarrow [-1.1, -0.9]$ before moving to the next event.

At $i = 4$ it holds that $z_4 = 45$ and $\delta_4 = 1$. We note AR $= [-1.1, -0.9]$ and L $= [0.3, -1.1, -0.9]$. We couple a random variable $U \sim$ Uniform(AR) and $U' =$ Uniform(L) using optimal transport. The resulting distribution equals:

$$P \leftarrow \begin{array}{c} \begin{matrix} x_3 & x_4 & x_5 \end{matrix} \\ \begin{pmatrix} 0. & 0.333 & 0.167 \\ 0.333 & 0. & 0.167 \end{pmatrix} \begin{matrix} x_4 \\ x_5 \end{matrix} \end{array}$$

We condition P on $U = x_4 = -1.1$, resulting in $v \leftarrow [0, 0.67, 0.33]$. In this case our sample turns out to be $\tilde{x} \leftarrow -1.1 = x_4$, which happens with probability 0.67. Hence $\tilde{D} \leftarrow [(1.3, 13), (0.5, 24), (-1.1, 45)]$. We also replace L $\leftarrow [0.3, -0.9]$ and AR $\leftarrow [-0.9]$.

We have now finished the loop $i = 1, \ldots, 4$. Since $z_5 = 81$ and L $\leftarrow [0.3, -0.9]$ we add the two datapoints $(0.3, 81)$ and $(-0.9, 81)$ to $\tilde{D}$. The finalized transformed dataset equals

$$\tilde{D} \leftarrow [(1.3, 13), (0.5, 24), (1.1, 45), (0.3, 81), (-0.9, 81)].$$

3.8.2 Proof of Lemma 3.3

Let $D = ((x_i, z_i, \delta_i))_{i=1}^n$ where z_i is increasing, and assume for convenience there are no ties in z. Denote by $k \coloneqq |\{i : \delta_i = 1\}|$ the number of observed events. We do not view D as random in this section. Applying the optimal transport algorithm results in a random, transformed dataset, which we denote by $T(D)$. Note that the times and covariates in $T(D)$ are not random, although the way in which they are paired up is. The same set of times and covariates is obtained in $\pi(T(D))$ and $T(\pi(D))$ for any $\pi \in S_n$. Denote the times in $T(D)$ by $t_1 \leq \cdots \leq t_n$ and define a standard pairing $\tilde{D} = ((x_i, t_i))_{i=1}^n$,. We will often use that $T(D), \pi(T(D)), T(\pi(D))$ are all permutations (possibly random) of $\tilde{D}$. Finally, define $h : \{1, \ldots, k\} \to \{1, \ldots, n\}$, so

that $t_i = z_{h(i)}$, which says that the i-th observed event is the $h(i)$-th overall event. As a last piece of notation, we will use Π to denote a uniform random permutation, and π to be a specific instance of a permutation. In particular we denote $\Pi_i = \Pi(i)$ and $\Pi_{1:h(i)-1} = [\Pi_1, \ldots, \Pi_{h(i)-1}]$. This corresponds to the covariates in the permuted dataset $\Pi(D)$ until just before the time of the i-th observed event.

We prove the theorem by showing that the left- and right-hand sides of Lemma 3.3 are both equal in distribution to

$$\left[T(D), \Pi^1(\tilde{D}), \ldots, \Pi^B(\tilde{D})\right].$$

This is done in separate lemmas.

Lemma 3.10.

$$\left[T(D), \Pi^1(T(D)), \ldots, \Pi^B(T(D))\right] \stackrel{d}{=} \left[T(D), \Pi^1(\tilde{D}), \ldots, \Pi^B(\tilde{D})\right]$$

Proof. By the above remarks we see that $T(D) = \Pi^D(\tilde{D})$ for some random permutation Π^D. (Note: The randomness in Π^D comes from the transformation T, not from the dataset D, which is fixed.) It suffices to show that

$$\left[\Pi^D, \Pi^1 \circ \Pi^D, \ldots, \Pi^B \circ \Pi^D\right] \stackrel{d}{=} \left[\Pi^D, \Pi^1, \ldots, \Pi^B\right].$$

This is easy to see by conditioning on Π^D. Let $\pi^0, \ldots \pi^B$ be arbitrary permutations. Then

$$\begin{aligned}
&P\big(\Pi^D = \pi^0, \Pi^1 \circ \Pi^D = \pi^1, \ldots, \Pi^B \circ \Pi^D = \pi^B\big) \\
=&P\big(\Pi^1 \circ \pi^0 = \pi^1, \ldots, \Pi^B \circ \pi^0 = \pi^B \;\big|\; \Pi^D = \pi^0\big) P(\Pi^D = \pi^0) \\
=&P\big(\Pi^1 = \pi^1 \circ (\pi^0)^{-1}, \ldots, \Pi^B = \pi^B \circ (\pi^0)^{-1}\big) P(\Pi_D = \pi^0).
\end{aligned}$$

Since $(\Pi^1, \ldots, \Pi^B)$ are independent uniform permutations, this is the same as

$$P\big(\Pi^D = \pi^0, \Pi^1 = \pi^1, \ldots, \Pi^B = \pi^B\big).$$

□

We now consider the effect of first permuting and then transforming the data.

Lemma 3.11. *Let Π be a uniformly chosen permutation of S_n and let T be defined through optHSIC. It holds that*

$$T(\Pi(D)) \stackrel{d}{=} \Pi(\tilde{D}).$$

Proof. By the comments above, we can define a random permutation Σ by $\Sigma(\tilde{D}) \coloneqq T(\Pi(D))$. We wish to show that $P(\Sigma = \sigma) = 1/n!$ for all $\sigma \in S_n$. To do so, we will condition on events of the form

$$\{\Pi_{1:h(i)-1} = \pi_{1:h(i)-1}\},$$

which determines the covariates in the permuted dataset up to (just before) the time of the i-th observed event. We also condition on $\Sigma_{1:i-1}$, fixing the covariates in the transformed dataset, up to the i-th observed event. Note that this conditioning fixes the coupling defined in the optimal transport algorithm. Namely, we let $\tilde{Y}$ and $\tilde{X}$ be the coupled random variables resulting from optimal transport between choosing uniformly from the covariates indexed by $[n]\backslash\{\sigma_{1:i-1}\}$ and choosing uniformly from the covariates indexed by $[n] \setminus \{\pi_{1:h(i)-1}\}$ respectively. Then, the transformation samples U' conditional on $U = x_{\Pi_{h(i)}}$. Because Π is a uniformly chosen permutation, given the events we conditioned on so far, U is uniformly chosen from the covariates indexed by $[n] \setminus \{\pi_{1:h(i)-1}\}$. By the definition of the coupling, U' is thus uniform on the covariates indexed by $[n]\backslash\sigma_{1:i-1}$. That is, Σ_i is chosen uniformly from $[n]\backslash\sigma_{1:i-1}$. Mathematically,

for any $\sigma, \pi \in S_n$ so that the conditioning event has nonzero probability, it holds that

$$\begin{aligned} &P(\Sigma_i = \sigma_i \mid \Pi_{1:h(i)-1} = \pi_{1:h(i)-1}, \Sigma_{1:i-1} = \sigma_{1:i-1}) \\ =&\frac{1}{n-i+1} \end{aligned}$$

Having shown that, conditioned on what happened in both the permuted dataset, and the synthetic dataset, the new synthetic covariate is chosen uniformly from those not chosen before, we aim to derive a recurrence relation so as to apply this result at each successive time. To this end note that

$$\begin{aligned} &P(\Sigma_{i:k} = \sigma_{i:k} \mid \Pi_{1:h(i)-1} = \pi_{1:h(i)-1}, \Sigma_{1:i-1} = \sigma_{1:i-1}) \\ =&P(\Sigma_{i+1:k} = \sigma_{i+1:k} \mid \Pi_{1:h(i)-1} = \pi_{h(i)-1}, \Sigma_{1:i} = \sigma_{1:i}) \\ &\times P(\Sigma_i = \sigma_i \mid \Pi_{1:h(i)-1} = \pi_{1:h(i)-1}, \Sigma_{1:i-1} = \sigma_{1:i-1}) \\ =&\frac{1}{n-i+1} P(\Sigma_{i+1:k} = \sigma_{i+1:k} \mid \Pi_{1:h(i)-1} = \pi_{h(i)-1}, \Sigma_{1:i} = \sigma_{1:i}) \\ =&\frac{1}{n-i+1} \sum_{\pi_{h(i):h(i+1)-1}} P(\Sigma_{i+1:k} = \sigma_{i+1:k} \mid \Pi_{1:h(i+1)-1} = \pi_{1:h(i+1)-1}, \Sigma_{1:i} = \sigma_{1:i}) \\ &\times P(\Pi_{h(i):h(i+1)-1} = \pi_{h(i):h(i+1)-1} \mid \Pi_{1:h(i)-1} = \pi_{1:h(i)-1}, \Sigma_{1:i} = \sigma_{1:i}) \end{aligned}$$

where we use the previously established equality in the first equality. This allows us

to compute

$$
\begin{aligned}
&P(\Sigma_{1:k} = \sigma_{1:k}) \\
=&\sum_{\pi_{1:h(1)-1}} P(\Sigma_{1:k} = \sigma_{1:k} \;\big|\; \Pi_{1:h(1)-1} = \pi_{1:h(1)-1}) \\
&\times P(\Pi_{1:h(1)-1} = \pi_{1:h(1)-1}) \\
=&\frac{1}{n} \sum_{\pi_{1:h(1)-1}} \sum_{\pi_{h(1):h(2)-1}} P(\Sigma_{2:k} = \sigma_{2:k} \;\big|\; \Pi_{1:h(2)-1} = \pi_{1:h(2)-1}, \Sigma_1 = \sigma_1) \\
&\times P(\Pi_{h(1):h(2)-1} = \pi_{h(1):h(2)-1} \;\big|\; \Pi_{1:h(1)-1} = \pi_{1:h(1)-1}, \Sigma_1 = \sigma_1) \\
&\times P(\Pi_{1:h(1)-1} = \pi_{1:h(1)-1}) \\
=&\frac{1}{n(n-1)\cdots(n-k+2)} \\
&\times \sum_{\pi_{1:h(1)-1}} \cdots \sum_{\pi_{h(k-1):h(k)-1}} P(\Sigma_k = \sigma_k \;\big|\; \Pi_{1:h(k)-1} = \pi_{1:h(k)-1}, \Sigma_{1:k-1} = \sigma_{1:k-1}) \\
&\times P(\Pi_{h(k-1):h(k)-1} = \pi_{h(k-1):h(k)-1} \;\big|\; \Pi_{1:h(k-1)-1} = \pi_{1:h(k-1)-1}, \Sigma_{1:k-1} = \sigma_{1:k-1}) \\
&\times \cdots \\
&\times P(\Pi_{1:h(1)-1} = \pi_{1:h(1)-1}) \\
=&\frac{1}{n(n-1)\cdots(n-k+1)}
\end{aligned}
$$

Since the indices $\Sigma_{(k+1):n}$ are added in uniform random order by definition of the transformation algorithm, this concludes the lemma. □

Lemma 3.12.

$$\left[T(D), T_1(\Pi^1(D)), \ldots, T_B(\Pi^B(D))\right] \stackrel{d}{=} \left[T(D), \Pi^1(\tilde{D}), \ldots, \Pi^B(\tilde{D})\right]$$

Proof. The left hand side can be written as $[\Pi^D(\tilde{D}), \Sigma^1(\tilde{D}), \ldots, \Sigma^B(\tilde{D})]$. The lemma above shows that the Σ^i for $i \geq 1$ have the correct distributions. We only need to show they and Π^D are a sequence of mutually independent permutations. But Σ^i is determined completely by Π^i and T_i, and Π^D is determined by T. The proof follows

since all these variables are mutually independent. □

Lemma A.1 and A.3 together prove the theorem.

3.8.3 Proof of Theorem 3.4

The proof of Theorem 3.6 shows that, if $C \perp\!\!\!\perp X$, then

$$(D, \Pi_1(D), \ldots, \Pi_B(D))$$

is an exchangeable vector. In particular, if $T, T_1 \ldots, T_B$ are independent identically distributed transformations of the data, then also

$$(T(D), T_1(\Pi_1(D)), \ldots, T_B(\Pi_B(D)))$$

is exchangeable. We let T be the transformation of the data using OPT–transformation of the data. By Lemma 3.3 the above vector is equal in distribution to

$$(T(D), \Pi_1(T(D)), \ldots, \Pi_B(T(D))),$$

implying that the latter is also exchangeable. For an arbitrary statistic H,

$$[H(T(D)), H(\Pi_1(T(D))), \ldots, H(\Pi_B(T(D)))]$$

is thus exchangeable too. In particular, the rank of the first entry is uniformly distributed on $1, \ldots, B+1$, which proves the theorem.

3.8.4 Proof of Theorem 3.5

Proof. This proof is based on the proof of Lemma 3 of Berrett and Samworth (2019). Since $H_0 : X \perp\!\!\!\perp Y$ implies that $(X_i, Y_j) \stackrel{d}{=} (X_i, Y_i)$ it is easy to see that $\pi D \stackrel{d}{=} \mathrm{D}$, for any permutation π. Writing $\Pi^0 = id$, and $\Pi^1, \ldots, \Pi^B$ for i.i.d. uniform permutations, we aim to show that, for any permutation σ of $\{0, 1, \ldots, B\}$

$$\left(\Pi^0(D), \Pi^1(D), \ldots, \Pi^B(D)\right) \stackrel{d}{=} \left(\Pi^{\sigma_0}(D), \Pi^{\sigma_1}(D), \ldots, \Pi^{\sigma_B}(D)\right);$$

that is, that the random vector on the left is exchangeable. We observed above that the first entries are equal in distribution. It remains to show that the other entries of the right-hand side are uniform and independently chosen permutations of the first entry. Indeed, writing $\Pi^{\sigma_0}(D) = \tilde{D}$, we can rewrite the right-hand side as:

$$\left(\tilde{D}, \Pi^{\sigma_1}(\Pi^{\sigma_0})^{-1}(\tilde{D}), \ldots, \Pi^{\sigma_B}(\Pi^{\sigma_0})^{-1}(\tilde{D})\right).$$

So it remains to show that $\left(\Pi^{\sigma_j}(\Pi^{\sigma_0})^{-1}, 1 \leq j \leq B\right)$ are independent uniformly chosen permutations of S_n. If $\sigma_0 = 0$, then $\Pi^{\sigma_0} = id$ and $\tilde{D} = D$ and the result is obvious. Now assume that $\sigma_i = 0$ for $i \geq 1$.

$$\begin{aligned}
&P\left(\Pi^{\sigma_1}(\Pi^{\sigma_0})^{-1} = \pi^1, \ldots, (\Pi^{\sigma_0})^{-1} = \pi^i, \ldots, \Pi^{\sigma_B}(\Pi^{\sigma_0})^{-1} = \pi^B\right) \\
&\quad = P\left(\Pi^{\sigma_1}\pi^i = \pi^1, \ldots, \Pi^{\sigma_B}\pi^i = \pi^B \;\middle|\; (\Pi^{\sigma_0})^{-1} = \pi^i\right) P\left((\Pi^{\sigma_0})^{-1} = \pi^i\right) \\
&\quad = P\left(\Pi^{\sigma(1)} = \pi^1(\pi^i)^{-1}\right) \ldots P\left(\Pi^{\sigma(B)} = \pi^B(\pi^i)^{-1}\right) P\left(\Pi^{\sigma_0})^{-1} = \pi^i\right) \\
&\quad = (n!)^{-B}
\end{aligned}$$

It follows that the vector

$$\left(D, \Pi^1(D), \ldots, \Pi^B(D)\right)$$

is indeed exchangeable. Letting H denote any arbitrary function on data, it follows that:

$$\big(H(D), H(\Pi^1(D)), \ldots, H(\Pi^B(D))\big)$$

is also exchangeable. If we break ties at random, this implies that every ordering of the $B+1$ elements is equally likely. In particular, the rank of an individual element is uniformly distributed on $\{1, \ldots, B+1\}$, and the result follows immediately.

□

3.8.5 Proof of Theorem 3.6

Proof. When we assume that $C \perp\!\!\!\perp X$ then, under the null hypothesis $H_0 : T \perp\!\!\!\perp X$, it follows that the pair $(T, C) \perp\!\!\!\perp X$. As (Z, D) is (T, C)–measurable, also $(Z, D) \perp\!\!\!\perp X$. If we write $Y = (Z, D)$, then Theorem 3.5 applies.

□

3.8.6 Proof of Lemma 3.7

The following computation shows that $EWf(X, Z) = Ef(X, T)$ for all functions f. We denote the distribution of (X, T, C) on $\mathcal{X} \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ by μ_{XTC}. As we are assuming independence of T and C given X we can decompose $\mu_{XTC} = \mu_{XT} \times \mu_{C|X}$.

$$
\begin{aligned}
EWf(X,Z) &= E1\{W \neq 0\}f(X,Z) \\
&= \int_{\mathcal{X}\times\mathbb{R}_{\geq 0}\times\mathbb{R}_{\geq 0}} 1\{t \geq c\}\frac{1}{g(t,x)}f(x,t)\mu_{XTC}(dx,dt,dc) \\
&= \int_{\mathcal{X}\times\mathbb{R}_{\geq 0}}\int_t^\infty \frac{1}{g(t,x)}f(x,t)\mu_{C|x}(dc)\mu_{XT}(dx,dt) \\
&= \int_{\mathcal{X}\times\mathbb{R}_{\geq 0}} \frac{1}{g(t,x)}f(x,t)\int_t^\infty \mu_{C|x}(dc)\mu_{XT}(dx,dt) \\
&= \int_{\mathcal{X}\times\mathbb{R}_{\geq 0}} f(x,t)\mu_{XT}(dx,dt) \\
&= Ef(X,T).
\end{aligned}
$$

where the penultimate equality follows because $\int_t^\infty \mu_{C|x}(dc) = \mathbb{P}(C > t|X = x) = g(t,x)$.

3.8.7 Proof of Lemma 3.8

Estimating the survival of the censoring distribution amounts to replacing δ by $1-\delta$ in the Kaplan Meier Survival curve. This yields:

$$
\hat{P}(C > z_k) = \prod_{i=1}^{k}\left(\frac{n-i}{n-i+1}\right)^{1-\delta_i}
$$

Thus the probability of being uncensored by time z_k equals:

$$
\begin{aligned}
\hat{P}(C \geq z_k) &= \prod_{i=1}^{k-1}\left(\frac{n-i}{n-i+1}\right)\prod_{i=1}^{k-1}\left(\frac{n-i}{n-i+1}\right)^{-\delta_i} \\
&= \frac{n-k+1}{n}\prod_{i=1}^{k-1}\left(\frac{n-i}{n-i+1}\right)^{-\delta_i}
\end{aligned}
$$

Note now that

$$\begin{aligned}\frac{1}{\hat{P}(C \geq z_k)} &= n \times \prod_{i=1}^{k-1} \left(\frac{n-i}{n-i+1}\right)^{\delta_i} \left(\frac{1}{n-k+1}\right) \\ &= n \times w_k\end{aligned}$$

for points that are uncensored. That is, Kaplan–Meier weights equal a re-scaled inverse of the probability of being uncensored by that time.

3.8.8 Proof of Theorem 3.9

Proof. The squared norm, written as the inner product with itself, can be expanded into three terms $a_1 + a_2 - 2a_3$ that we compute in turn. We denote by $A \circ B$ the entrywise product of the matrices A and B. Using the Hadamard product property $\alpha^\top (A \circ B)\, \beta = \operatorname{Tr}\left(D_\alpha A D_\beta B^\top\right)$ where $D_\alpha = \operatorname{diag}(\alpha)$, $D_\beta = \operatorname{diag}(\beta)$, we have the following identities:

$$\begin{aligned}a_1 &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j k(x_i, x_j) l(z_i, z_j) \\ &= w^\top (K \circ L)\, w \\ &= \operatorname{Tr}\left(D_w K D_w L\right);\end{aligned}$$

$$\begin{aligned}a_2 &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n}\sum_{s=1}^{n} w_i w_j w_r w_s k(x_i, x_j) l(z_r, z_s) \\ &= w^\top K w w^\top L w \\ &= \operatorname{Tr}\left(w w^\top K w w^\top L\right);\end{aligned}$$

$$
\begin{aligned}
a_3 &= \left\langle \sum_{i=1}^{n} w_i K((x_i, z_i), \cdot), \sum_{r=1}^{n}\sum_{s=1}^{n} w_r w_s K((z_r, z_s), \cdot) \right\rangle \\
&= \sum_{i=1}^{n} w_i \left(\sum_{j=1}^{n} w_j k\left(x_i, x_j\right) \right) \left(\sum_{r=1}^{n} w_r l\left(z_i, z_r\right) \right) \\
&= w^\top \left(Kw \circ Lw\right) \\
&= \operatorname{Tr}\left(D_w K w w^\top L\right).
\end{aligned}
$$

As the entrywise product is symmetric in its arguments, we see that also

$$
a_3 = \operatorname{Tr}\left(D_w L w w^\top K\right) = \operatorname{Tr}\left(w w^\top K D_w L\right).
$$

Thus the weighted HSIC is

$$
\begin{aligned}
a_1 + a_2 - 2a_3 &= \operatorname{Tr}(D_w K D_w L) - \operatorname{Tr}(D_w K w w^\top L) - \operatorname{Tr}(w w^\top K D_w L) \\
&\quad + \operatorname{Tr}(w w^\top K w w^\top L) \\
&= \operatorname{Tr}\left(\left(D_w - w w^\top\right) K \left(D_w - w w^\top\right) L\right) \\
&= \operatorname{Tr}\left(H_w K H_w L\right),
\end{aligned}
$$

with $H_w = \left(D_w - w w^\top\right)$. In the standard HSIC case $w = \frac{1}{n}(1, 1, \ldots, 1) \coloneqq 1_n$ and, $D = \frac{1}{n} I$, so that $H_w = \frac{1}{n} I - 1_n 1_n^\top$ is the standard (scaled) centering matrix. □

3.8.9 Using multiple transformations

We list 4 ways of combining p-values.

Method 1: Use a Bonferroni correction and reject H_0 if for the smallest p-value, denoted by $p_{(1)}$, it holds that $p_{(1)} \leq \alpha/m$.

Method 2: Make the following (random) rejection decision: reject H_0 with probability $\sum_{i=1}^{m} 1\{p_i \leq \alpha\}/m$, and accept H_0 otherwise.

Method 3: Fix $\beta \leq \alpha$ and reject if $\sum_{i=1}^{m} 1\{p_i \leq \beta\}/m \geq \beta/\alpha$. For example, reject if $\sum_{i=1}^{m} 1\{p_i \leq 3\alpha/4\}/m \geq 3/4$.

Method 4: Reject if $2\sum_{i=1}^{m} p_i/m \leq \alpha$. This has the advantage that it results in a quantity that can be used as a p-value: $2\sum_{i=1}^{m} p_i/m$.

Throughout this section, assume that the null hypothesis holds. Let p be the p-value resulting from sampling a dataset D once, followed by running optHSIC once (so exactly one transformation and one permutation test on the transformed data). We aim to show that the methods 1 and 2 above have correct type 1 error under the assumption that $P_{H_0}(p \leq \alpha) \leq \alpha$ for $\alpha \in [0, 1]$ which we proved for $C \perp\!\!\!\perp X$ and expect to be (approximately) true for $C \not\perp\!\!\!\perp X$. We aim to show that methods 3 and 4 have asymptotically (as the number of p-values goes to infinity) correct type 1 error rate under the assumption that $p \sim \text{Unif}[\frac{1}{B+1}, \ldots, \frac{B+1}{B+1}]$, which we proved for $C \perp\!\!\!\perp X$ and expect to be (approximately) true also for $C \not\perp\!\!\!\perp X$. See Table 3.1. While method 2 is less conservative, it is a random rejection decision which is less desirable. We can imagine Method 3 being not too conservative when $\beta = 3\alpha/4$.

3.8.9.1 Method 1

Assume it holds that $P_{H_0}(p \leq \alpha) \leq \alpha$ (see comments at the start of the section). Let $p_{(1)}, \ldots, p_{(m)}$ be the p-values obtained from applying optHSIC m times to D, in ascending order. The Bonferroni correction procedure rejects H_0 if $p_{(1)} \leq \alpha/m$. This has the correct type 1 error probability because by the union bound under the null hypothesis

$$\begin{aligned} P_{H_0}(\text{reject}) &= P_{H_0}(p_{(1)} \leq \alpha/m) \\ &\leq m P_{H_0}(p_1 \leq \alpha/m) \\ &\leq \alpha. \end{aligned}$$

3.8.9.2 Method 2

Assume it holds that $P_{H_0}(p \leq \alpha) \leq \alpha$ (see comments at the start of the section). Given the p-values $p_1, \ldots, p_m$, the second method makes a random rejection decision in the following way: Reject H_0 with probability $\sum_{i=1}^{m} 1\{p_i \leq \alpha\}/m$, and accept H_0 otherwise. This has correct type 1 error because

$$\begin{aligned} P_{H_0}(\text{reject}) &= E_{H_0}[P(\text{reject}|p_1, \ldots, p_m)] \\ &= E_{H_0}[\sum_{i=1}^{m} 1\{p_i \leq \alpha\}/m] \\ &= P_{H_0}(p_1 \leq \alpha) \\ &\leq \alpha. \end{aligned}$$

3.8.9.3 Method 3

Fix $\beta \leq \alpha$ and reject if $\sum_{i=1}^{m} 1\{p_i \leq \beta\}/m \geq \beta/\alpha$. An example would be to set $\beta = \alpha/2$ in which case we reject if $\sum_{i=1}^{m} 1\{p_i \leq \alpha/2\} \geq \frac{1}{2}$. Assume $P_{H_0}(p \leq \alpha) \leq \alpha$. This is an approximate method. The 'ideal' and practically impossible method is to reject if $P(p \leq \beta|D) \geq \beta/\alpha$. We show that this 'ideal' method has the correct type 1 error:

$$P_{H_0}(\text{reject}) = P_{H_0}(A)$$

where A is the event that

$$A = \{D : P(p \leq \beta|D) \geq \beta/\alpha\}.$$

Assume by contradiction that $P_{H_0}(A) > \alpha$. Then it must hold that

$$\begin{aligned} P_{H_0}(p \leq \beta) &= E_{H_0}[P(p \leq \beta|D)] \\ &\geq E_{H_0}[1_A P(p \leq \beta|D)] \\ &> \alpha\beta/\alpha \\ &= \beta. \end{aligned}$$

which contradicts that $P_{H_0}(p \leq \beta) \leq \beta$. Hence it must hold that $P_{H_0}(A) \leq \alpha$. Because in practice $P(p \leq \beta|D)$ is unknown, we can estimate it by $\sum_{i=1}^{m} 1\{p_i \leq \beta\}/m$ and reject if $\sum_{i=1}^{m} 1\{p_i \leq \beta\}/m \geq \beta/\alpha$. Since $\sum_{i=1}^{m} 1\{p_i \leq \beta\}/m \to P(p \leq \beta|D)$ as $m \to \infty$ it is easy to see the approximate method is asymptotically correct.

3.8.9.4 Method 4

Method 4 is to reject if $2\sum_{i=1}^{m} p_i/m \leq \alpha$. This is an approximation of the 'ideal' and practically impossible method of rejecting H_0 if D is such that $E(p|D) \leq \alpha/2$. We assume it holds that $p \sim \text{Uniform}[0, 1]$: if we prove it under the assumption $p \sim \text{Uniform}[0, 1]$ the result also follows under the assumption $p \sim \text{Uniform}[\frac{1}{B+1}, \ldots, \frac{1}{B+1}]$ since the latter distribution corresponds to a more conservative test. We now show that this 'ideal' method has the correct type 1 error rate. Note

$$P_{H_0}(\text{reject}) = P_{H_0}(A)$$

where A is the event that

$$A = \{D : E(p|D) \leq \alpha/2\}.$$

Define the following family of distributions:

$$M_A = (\mu_D)_{D \in A}$$

where

$$\mu_D([a,b]) \coloneqq P(p \in [a,b]|D).$$

We verify that the family μ_D and the set A satisfy three conditions:

Condition 1: For all $\mu_D \in M_A$ it holds that

$$\mu_D([0,1]) = 1.$$

Condition 2: For all $\mu_D \in M_A$ it holds that

$$\int_{[0,1]} x\mu_D(dx) \leq \alpha/2$$

by definition of A.

Condition 3: For all $0 \leq a \leq b \leq 1$

$$\begin{aligned} E_{H_0}[1\{D \in A\}\mu_D([a,b])] &= E_{H_0}[1\{D \in A\}P(p \in [a,b]|D)] \\ &\leq E_{H_0}[P(p \in [a,b]|D)] \\ &= (b-a) \end{aligned}$$

We now define the ν_A to be an 'average' of the distributions in M_A:

$$\nu_A([a,b]) = E_{H_0}[1\{D \in A\}\mu_D([a,b])]/P_{H_0}(A)$$

It is easy to see ν_A satisfies condition 1:

$$\begin{aligned}\nu_A[0,1] &= E_{H_0}[1\{D \in A\}\mu_D([0,1])]/P_{H_0}(A) \\ &= 1.\end{aligned}$$

To see ν_A satisfies condition 2 note that

$$\begin{aligned}\int_{[0,1]} x\nu_A(dx) &= \int_{[0,1]} xE_{H_0}[1\{D \in A\}\mu_D(dx)]/P_{H_0}(A) \\ &= E_{H_0}[1\{D \in A\}\int_{[0,1]} x\mu_D(dx)]/P_{H_0}(A) \\ &\leq E_{H_0}[1\{D \in A\}\alpha/2]/P_{H_0}(A) \\ &= \alpha/2.\end{aligned}$$

To see ν_A satisfies condition 3 note that

$$\begin{aligned}E'_{H_0}(1\{D' \in A\}\nu_A[a,b]) &= P_{H_0}(A)\nu_A[a,b] \\ &= E'_{H_0}(1\{D' \in A\}E_{H_0}[1\{D \in A\}\mu_D([a,b])])/P_{H_0}(A) \\ &\leq E_{H_0}(1\{D' \in A\}(b-a))/P_{H_0}(A) \\ &= b-a\end{aligned}$$

Here E'_{H_0} denotes expectation with respect to D' and E_{H_0} with respect to D. Note condition 1 says that ν_A is a probability measure, condition 2 says its expectation is less than $\alpha/2$ and condition 3 that v_A is dominated by the measure defined by the uniform density $1/P_{H_0}(A)$.

Thus, if $P_{H_0}(A) = \beta$, then ν_A and A satisfy the three conditions above. We now show that there is a maximum value $\beta^\star$ so that if $P_{H_0}(A) = \beta > \beta^\star$, then it is impossible for any distribution ν to satisfy the three conditions above.

We first show $\beta^\star \geq \alpha$. Assume that $P_{H_0}(A) = \alpha$. If we let ν_α be the uniform

probability measure on $[0, \alpha]$, then it is clear that the first two conditions are met: it is a valid probability distribution (condition 1), the expectation is exactly $\alpha/2$ (condition 2). The third condition is met since for $0 \leq a \leq b \leq \alpha$ it holds that

$$E_{H_0}[1\{D \in A\}\nu_\alpha([a, b])] = P_{H_0}(A)(b - a)/\alpha = b - a$$

because by assumption $P_{H_0}(A) = \alpha$.

We now need to show that if $\beta > \alpha$ there does not exist a distribution ν that satisfies the three conditions. To that end, note first that if ν satisfies the three conditions for β, then it also satisfies the conditions for any β' such that $\beta' < \beta$ (note $P_{H_0}(A)$ only appears in the third condition). So in particular such ν would have to satisfy the conditions also with $\beta = \alpha$. However, to change the distribution ν_α defined above, one cannot place more mass in the region $[0, \alpha]$ by condition 3, which says ν needs to be dominated by the measure defined by the uniform density $1/P_{H_0}(A)$. On the other hand, if one removes mass from $[0, \alpha]$ then one automatically increases the mean of the distribution, which violates condition 2, since the mean of $\nu_\alpha = \alpha/2$. We conclude that $\beta^\star = \alpha$. The type 1 error of the 'ideal' method is thus at most α.

Since $\sum_{i=1}^m p_i/m \to E(p|D)$ as $m \to \infty$ it is easy to see the approximate method has asymptotically correct type 1 error rate. This method has the advantage that it results in a combined p-value: $2\sum_{i=1}^m p_i/m$, whereas the other methods only lead to rejection decisions. The p-value will be conservative if there is little randomness in the dataset. In the case there is no censoring, the p-value is a factor 2 bigger than necessary. However, the Bonferroni correction would result in a p-value that is a factor m bigger than necessary (where m, the number of transformations used, which may be much larger than 2).

3.8.10 Tables

3.8.10.1 Type 1 error rates

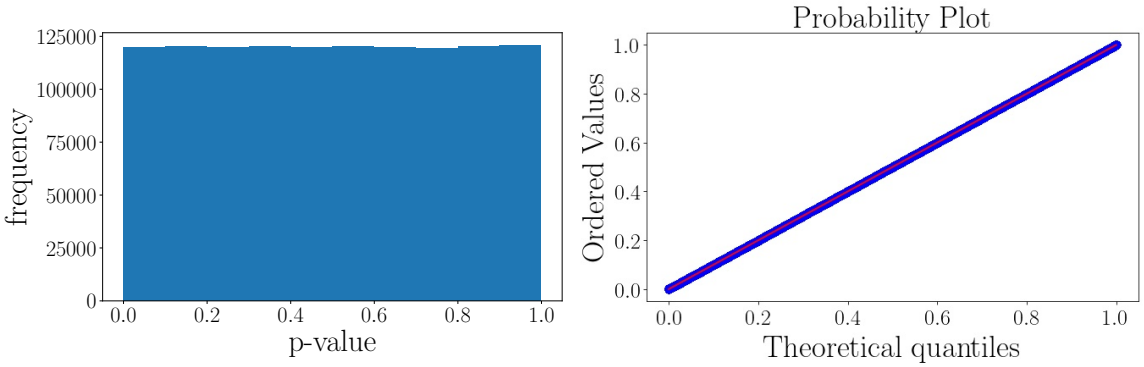


Figure 3.10: A histogram and a qq-plot of the p-values obtained from optHSIC for distribution D.6, in which $C \not\perp\!\!\!\perp X$, with a sample size of $n = 200$. Data was sampled 1.2 million times, each sample resulting in one p-value. These plots indicate that despite the fact that there was a strong dependence between C and X, the p-values returned by optHSIC are approximately Uniform$[0, 1]$.

$n =$	40	80	120	160	200	240	280	320	360	400
D.1	0.048	0.050	0.051	0.052	0.052	0.049	0.047	0.054	0.051	0.049
D.2	0.048	0.051	0.047	0.047	0.049	0.051	0.048	0.054	0.050	0.047
D.3	0.243	0.461	0.630	0.774	0.860	0.909	0.953	0.970	0.985	0.991
D.4	0.142	0.232	0.343	0.487	0.610	0.734	0.812	0.880	0.932	0.959
D.5	0.075	0.116	0.142	0.168	0.210	0.243	0.278	0.316	0.357	0.397
D.6	0.932	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
D.7	0.308	0.594	0.761	0.856	0.906	0.937	0.960	0.971	0.980	0.984
D.8	0.078	0.122	0.152	0.211	0.264	0.307	0.355	0.388	0.439	0.467

Table 3.6: The rejection rate of zHSIC in against the distributions D.1-8 of Table 3.2

$n =$	40	80	120	160	200	240	280	320	360	400
D.1	0.045	0.047	0.049	0.050	0.049	0.051	0.049	0.054	0.048	0.049
D.2	0.050	0.047	0.048	0.047	0.050	0.048	0.047	0.045	0.048	0.049
D.3	0.079	0.166	0.235	0.326	0.410	0.466	0.540	0.597	0.658	0.700
D.4	0.161	0.204	0.232	0.270	0.309	0.350	0.394	0.456	0.506	0.549
D.5	0.057	0.084	0.110	0.131	0.163	0.191	0.216	0.258	0.274	0.299
D.6	0.267	0.672	0.900	0.971	0.990	0.998	0.999	0.999	1.000	1.000
D.7	0.071	0.107	0.143	0.192	0.260	0.305	0.369	0.412	0.480	0.527
D.8	0.057	0.078	0.081	0.095	0.120	0.142	0.153	0.162	0.177	0.190

Table 3.7: The rejection rate of wHSIC in against the distributions D.1-8 of Table 3.2

$n =$	40	80	120	160	200	240	280	320	360	400
D.1	0.056	0.054	0.050	0.047	0.055	0.048	0.048	0.055	0.050	0.051
D.2	0.055	0.056	0.055	0.050	0.055	0.051	0.049	0.050	0.052	0.050
D.3	0.057	0.056	0.051	0.053	0.054	0.051	0.046	0.057	0.048	0.050
D.4	0.057	0.061	0.050	0.051	0.054	0.058	0.049	0.053	0.051	0.051
D.5	0.058	0.058	0.055	0.052	0.053	0.053	0.048	0.054	0.048	0.049
D.6	0.053	0.051	0.051	0.050	0.046	0.048	0.057	0.053	0.054	0.055
D.7	0.148	0.094	0.084	0.062	0.063	0.061	0.058	0.055	0.060	0.058
D.8	0.143	0.086	0.074	0.064	0.067	0.058	0.059	0.058	0.061	0.060

Table 3.8: The rejection rate of the Cox proportional hazards likelihood ratio test in against the distributions D.1-8 of Table 3.2

3.8.10.2 Rejection rate under varying censoring regimes

D.	$Z\|X$	$C\|X$	X
1	$\text{Exp}(\text{mean} = \exp(X/5))$	$\text{Exp}(\text{mean} = \theta)$	$N(0,1)$
2	$\text{Exp}(\text{mean} = \exp(X^2)/5)$	$\text{Exp}(\text{mean} = \theta\exp(X))$	$N(0,1)$
3	$\text{Weib}(\text{shape} = 1.75X + 3.25)$	$\text{Exp}(\text{mean} = \theta X^2)$	$\text{Unif}[-1,1]$
4	$N(\text{mean} = 100 - X, \text{var} = 2X + 5.5)$	$82 + \text{Exp}(\text{mean} = \theta)$	$\text{Unif}[-1,1]$
5	$\text{Exp}(\text{mean} = \exp(1^T X/30))$	$\text{Exp}(\text{mean} = \theta)$	$N_{10}(0, \text{cov} = \Sigma_{10})$
6	$\text{Exp}(\text{mean} = \exp(X_4/7))$	$\text{Exp}(\text{mean} = \theta\exp(1^T X/30))$	$N_{10}(0, \text{cov} = \Sigma_{10})$
7	$\text{Exp}(\text{mean} = \exp(X_4^2/20))$	$\text{Exp}(\text{mean} = \theta\exp(X_2^2)/20)$	$N_{10}(0, \text{cov} = \Sigma_{10})$
8	$\text{Exp}(\text{mean} = \exp(X_{10}^2 + 2X_8)/20)$	$\text{Exp}(\text{mean} = \theta\exp(X_2/7))$	$N_{10}(0, \text{cov} = \Sigma_{10})$

Table 3.9: The parametrized distributions to test the power under different censoring rates. Here $\Sigma_{10} = MM^T$ where M is a 10×10 matrix of i.i.d. standard normal entries. M is sampled once and then kept fixed. The parameter θ varies such that $15, 30, 45, 60, 75, 90, 100\%$ of the individuals are observed (i.e. $\Delta = 1$). The sample size is $n = 200$ in each case.

$\%\Delta = 1$		20%	40%	60%	80%	100%
D.1	Cph	0.243	0.422	0.565	0.705	0.753
	optHSIC	0.229	0.382	0.501	0.634	0.699
	wHSIC	0.038	0.062	0.182	0.395	0.703
	zHSIC	0.066	0.168	0.329	0.525	0.701
D.2	Cph	0.180	0.267	0.268	0.225	0.108
	optHSIC	0.087	0.171	0.258	0.378	0.686
D.3	Cph	0.073	0.056	0.107	0.223	0.288
	optHSIC	0.242	0.177	0.399	0.886	0.968
D.4	Cph	0.187	0.091	0.064	0.046	0.039
	optHSIC	0.346	0.224	0.275	0.509	0.779
	wHSIC	0.138	0.285	0.452	0.654	0.770
	zHSIC	0.105	0.172	0.274	0.410	0.759
D.5	Cph	0.315	0.487	0.610	0.705	0.836
	optHSIC	0.268	0.439	0.546	0.629	0.775
	wHSIC	0.055	0.072	0.169	0.409	0.786
	zHSIC	0.083	0.229	0.362	0.605	0.760
D.6	Cph	0.461	0.732	0.834	0.916	0.939
	optHSIC	0.396	0.681	0.801	0.876	0.952
D.7	Cph	0.055	0.068	0.077	0.078	0.107
	optHSIC	0.043	0.100	0.134	0.289	0.669
D.8	Cph	0.162	0.313	0.431	0.517	0.572
	optHSIC	0.164	0.335	0.498	0.619	0.916

Table 3.10: The rejection rates of the various methods against distributions D.1-D.8 given in Table 3.9. When $C \not\perp\!\!\!\perp X$, we only show rejection rates of the CPH test and optHSIC, because wHSIC and zHSIC have high inflated rejection rates due to the dependency of C and X. The top row shows the percentage of observed events ($\Delta = 1$).

3.8.11 Binary covariates

As a special case of independence testing we consider the case of a single binary covariate, i.e., $X \in \{0, 1\}$. If one groups the data by covariate, then testing independence of T and X is equivalent to testing equality of lifetime distribution between the two groups. This is known as two-sample testing on right censored data. Popular approaches to this challenge are the logrank test and various weighted logrank tests.

optHSIC can be applied to this problem without any adjustments, while wHSIC can be improved in this case in two ways: first, the weights can be estimated even when the censoring distribution differs between the two groups; and second, there exists an alternative permutation strategy that, experiments show, seems to control the type 1 error effectively even under dependent censoring. These adjustments are described in Section 3.8.11.1 and Section 3.8.11.2 respectively. We omit consideration of zHSIC, as it is fundamentally more limited, given the larger number of available methods.

3.8.11.1 wHSIC for two-sample testing

Let P_0 and P_1 denote the distribution of $T|X = 0$ and $T|X = 1$ respectively. Let the total sample be $D = ((x_i, z_i, \delta_i))_{i=1}^n$ as before, and write $((z_i^0, \delta_i^0))_{i=1}^{n_0}$ and $((z_i^1, \delta_i^1))_{i=1}^{n_1}$ for the event times and indicators of individuals with covariate $X = 0$ and $X = 1$ respectively. We want to asses if $P_0 = P_1$. We again use the Brownian distance kernel. If all of the n times were observed ($\delta = 1$), we could measure the difference in empirical distributions between both groups by the MMD between the two distributions:

$$\left|\left|\frac{1}{n_0}\sum_{i=1}^{n_0} k(z_i^0, \cdot) - \frac{1}{n_1}\sum_{j=1}^{n_1} k(z_j^1, \cdot)\right|\right|_H.$$

Similar to Section 3.5.1, when some observations are censored, we might reweight the empirical distributions, and instead compare the weighted empirical distributions

$$\sum_{i=1}^{n_0} w_i^0 k(z_i^0, \cdot) \qquad \text{and} \qquad \sum_{i=1}^{n_0} w_i^1 k(z_i^1, \cdot).$$

We propose that the weights w_i are computed by the Kaplan–Meier weights *within* each group. The test statistic thus becomes:

$$\text{wHSIC}(D) \coloneqq \left|\left|\sum_{i=1}^{n_0} w_i^0 k(z_i^0, \cdot) - \sum_{i=1}^{n_0} w_i^1 k(z_i^1, \cdot)\right|\right|_H^2.$$

This statistic was also, independently, proposed by Matabuena (2019), and can be seen as a special case of wHSIC in the case of binary covariates. Under the hypothesis that $C \perp\!\!\!\perp X$, one can obtain p-values using a permutation test, resulting in the following algorithm. Section 3.8.11.2 provides an alternative permutation strategy under dependent censoring, that was proposed by Wang et al. (2010). It was proposed in the context of the logrank test, but can equally be used for other statistics.

Algorithm 6: wHSIC for two-sample data

Input : $D = ((x_i, z_i, \delta_i))_{i=1}^n$, significance level α, number of permutations B.

1 Sample permutations $\pi_1, \ldots, \pi_B$ i.i.d. uniformly from S_n. ;

2 Breaking ties at random, compute the rank R of HSIC(D) in the vector

$$(\text{wHSIC}(D), \text{wHSIC}(\pi_1(D)), \text{wHSIC}(\pi_2(D)), \ldots, \text{wHSIC}(\pi_B(D)))$$

where wHSIC is as defined above. ;

Output: Reject if $p := R/(B+1) \leq \alpha$.

3.8.11.2 ipxHSIC

This subsection overviews a test we name ipxHSIC, which is a special case of the statistic wHSIC(D) defined in Section 3.8.11.1 above, but which uses a different permutation strategy that is robust against differences in the censoring distributions of both groups. The permutation strategy was proposed in Wang et al. (2010) to provide reliable p-values for the logrank statistic in the case of small or unequal sample sizes. In fact Wang et al. (2010) propose two permutation strategies: the first one, which they call 'ipz' (section 2.1.1), permutes group membership and the second, which they call 'ipt'(section 2.1.2), permutes survival times. These permutation strategies were proposed in the context of logrank tests - but can equally be applied to other statis-

tics, such as wHSIC. The first strategy, which permutes the covariates, is referred to in their work as 'ipz' since the procedure first imputes several unobserved times, and then permutes the covariate, which in their work is denoted by z. We refer to it as 'ipx', as our covariate is denoted by x. The algorithm uses the Kaplan–Meier estimator to estimate three distributions: 1) G^0, the censoring distribution in group 0, based on the data observed in group 0; 2) G^1, the censoring distribution in group 1, based on the data observed in group 1; 3) the distribution of the lifetimes F based on the pooled dataset containing both groups. With these estimates, a new dataset is constructed, consisting of n observations, each consisting of a covariate, an event time, and two censoring times, one for each censoring distribution. This larger dataset is then permuted, and transformed back to a censored dataset. Wang et al. (2010) describe the algorithm in full detail. This method thus combines the wHSIC statistic with an alternative permutation strategy. Because this method relies on explicitly estimating censoring distributions in each group, it is difficult to extend this to the continuous case, where for each covariate we only have one individual in the study with that exact covariate.

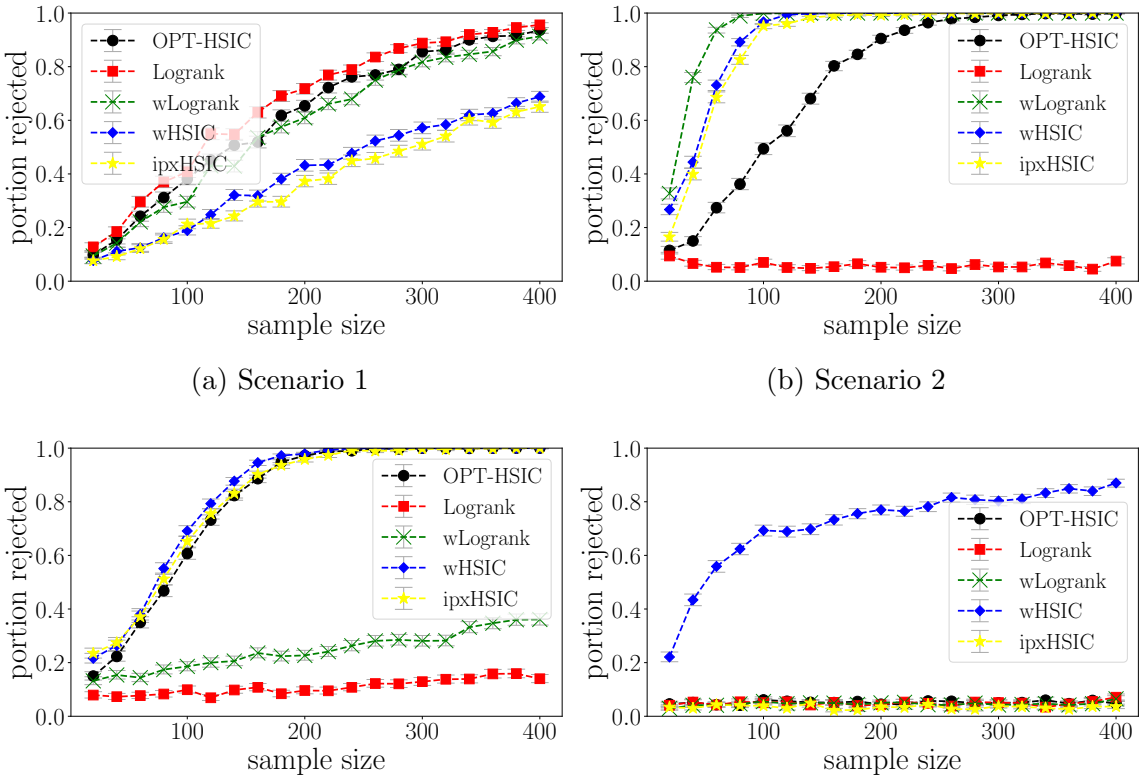
3.8.11.3 Numerical comparison of methods in the two-sample case

We generate data from four different distributions for each of X, T, and C to compare the power and type 1 error of the proposed methods optHSIC, wHSIC, ipxHSIC to the power and type 1 error of the classic logrank test and a weighted logrank test proposed by Ditzhaus and Friedrich (2018). The classical logrank test is known to have low power against certain alternatives, such as crossing survival curves. A weighted logrank test assigns weights to data, giving the logrank test power against different alternatives. In Ditzhaus and Friedrich (2018) a combination of weights is proposed, so as to achieve power against a wider class of alternatives. In particular Ditzhaus and Friedrich (2018) propose a combination of two sets of weights, corresponding

to proportional and crossing hazards. As this section mostly serves to provide an example of our methods, we simulate fewer scenarios than in Section 3.6. In each scenario we let the n values range from $n = 20$ to $n = 400$ in intervals of 20. To obtain p-values in the three HSIC based methods as well as the weighted logrank test we use a permutation test with 1999 permutations. We reject the null hypothesis if our obtained p-value is less than 0.05.

D.	T_0	T_1	C_0	C_1	% Observed
1	Exp(1)	Exp(1/1.6)	Exp(1/2)	Exp(1/2)	60 %
2	Weib$(1, 5)$	Weib$(1, 1.5)$	Exp(1/2)	Exp(1/2)	60 %
3	Exp(1)	(0.43, 1.39+Exp(1))	$1 +$ Exp(1/2)	$1 +$ Exp(1/2)	90 %
4	Exp(1)	Exp(1)	Exp(2)	None	65 %

Table 3.11: The 4 scenarios in which in which we perform two-sample tests. T_1 is 0.43 w.p. 0.75 and $1.39 +$ Exp(1) w.p. 0.25. Note that in D.4 the null hypothesis holds.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 3.11: Rejection rates of the various two-sample tests. Note that in Scenarios 1-3 the alternative hypothesis holds, implying high rejection rate is desirable. In Scenario 4, the null hypothesis holds, so a rejection rate of 0.05 is desirable. wHSIC in that case thus wrongly rejects the null: this reflects the crucial assumption of wHSIC that the groups have identical censoring distributions.

3.8.11.4 Example in which optHSIC does not perform well

Consider the following case. Group $X = 0$ contains 1050 individuals. Group $X = 1$ contains 50 individuals. Up to time $t = 50$, no events occur. At time $t = 50$, 1000 individuals of group $X = 0$ are censored. There are now 50 individuals remaining in each of the groups. The 50 individuals of group $X = 0$ have event time $100 +$ Exp(mean $= 2$) and the 50 individuals of group $X = 1$ have event time $100 +$ Exp(mean $= 1$). In this example we find the logrank test to have power of 89% and optHSIC to have power of only 12%.

What happens is the following: At time $t = 100$ there are 100 individuals at risk. The

individuals of group $X = 1$ are likely to have their event first, due to the higher rate in the corresponding exponential distribution. Because in group $X = 0$ 1000 individuals have been censored, the optimal transport map has a high chance of choosing $\tilde{x} = 0$ when $x_i = 1$. So while in the resulting dataset a slight bias will remain towards individuals in group $X = 1$ having their event first, this bias is much less clear than before the transformation. (We thank a reviewer for proposing this scenario.)

There are several characteristics that make the difference in this example so large. Firstly, as mentioned before, optimal transport relies on the ability to choose a 'similar covariate'. When covariates are binary it may happen that $\tilde{x} = 0$ while $x_i = 1$. Secondly in this case all the censoring happens in group $X = 0$, causing optimal transport to send mass from group $X = 1$ to group $X = 0$. Furthermore, the censoring rate is high (91% of all individuals). Lastly, before the censoring occurs there is no evidence of a difference in distribution.

3.8.11.5 Comments on two-sample simulations

The results show that the logrank test and the weighted logrank test have little power in scenario 2 and 3 and scenario 3 respectively, even though large differences between the samples are present. The logrank is designed to detect differences as in scenario 1, and the weighted logrank is designed to detect differences as in scenario 1 and 2, sacrificing power slightly compared to the logrank test in the first. Scenario 3 is designed to defeat the weighted logrank test, since we constructed an extreme version of an early crossing survival curve, and the test does not contain weights for early crossing. The kernel methods are fully nonparametric, but do lose power in certain scenarios, most notably in Scenario 2 and the example provided. We believe optHSIC is not ideally suited to the case of binary covariates, since optimal transport relies on choosing a 'similar' covariate. Furthermore, while there are no fully nonparametric alternatives for independence testing for continuous covariates, there

are more alternative two-sample tests. We thus believe the main value of optHSIC lies in the case of continuous covariates.

3.9 Statement of authorship

Statement of authorship	
Title of paper	A kernel- and optimal transport- based test of independence between covariates and right-censored lifetimes
Status	Published
Authors	David Rindt, Dino Sejdinovic and David Steinsaltz.
Details	The International Journal of Biostatistics 2020; `https://doi.org/10.1515/ijb-2020-0022`
Student confirmation	
Student name	David Rindt
Contribution to the paper	I am the primary author of everything in the paper. The work was completed under the supervision of Professor Dino Sejdinovic and Professor David Steinsaltz.
Date	1 June 2021
Signature	
Supervisor confirmation	
By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.	
Supervisor name and title	Dino Sejdinovic, Associate Professor of Statistics
Supervisor comments	All above seems accurate and fair.
Date	22 June 2021
Signature	

Chapter 4

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

This chapter is based on the following work.

Tamara Fernández, Arthur Gretton, *David Rindt*, Dino Sejdinovic.
A kernel log-rank test of independence for right-censored data.
`https://arxiv.org/abs/1912.03784.`

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

4.1 Introduction

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

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

The problem of testing independence has been widely studied by the statistical community. In the context of right-censored data, this problem has often been addressed through the mechanism of two-sample tests, in which the covariate takes one out of two possible values. For the two-sample problem, the main tool is the log-rank test Mantel (1966); Peto and Peto (1972) and its generalizations, namely weighted log-rank tests Harrington and Fleming (1982); Bagdonavičius et al. (2011); Fernández and Rivera (2021). The more general case, in which the covariates belong to $\mathbb{R}^d$, is much more challenging, and most of the current approaches are ad-hoc for specific semi-parametric models, e.g. Cox (1972); Gray (1992); Zucker and Karr (1990). In particular, the most popular of these approaches is the Cox proportional hazards model Cox (1972), which assumes a linear effect of the covariates on the log-hazard

function. Non-parametric approaches are more scarce, however Le et al. (1994); McKeague et al. (1995); Rindt et al. (2020). In Le et al. (1994) a nonparametric test for independence is obtained by measuring monotonic relationships between a censored survival time and an ordinal covariate; and in McKeague et al. (1995) the authors propose an omnibus test that can detect any type of association between a censored survival time and a 1-dimensional covariate. The recent non-parametric test in Rindt et al. (2020) was introduced to deal with general covariates on $\mathbb{R}^d$. This approach deals with censored data by transforming it into uncensored samples, and then applies a well-known kernel independence test based on the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. (2007), Chwialkowski and Gretton (2014)).

In this paper we propose a non-parametric test which can potentially detect any type of dependence between right-censored survival times and general covariates. Our testing procedure is based on a dependence measure between survival times and covariates which is constructed using weighted log-rank tests and the theory of reproducing kernel Hilbert spaces (RKHSs). We provide asymptotic results for our test-statistic and propose an approximation of the rejection region of our test by using a Wild Bootstrap procedure. Under mild regularity conditions, we prove that both the oracle and the testing procedure based on the Wild Bootstrap approximations are asymptotically consistent, meaning that our test can detect any type of dependence. The closest prior works to our approach are Fernández and Rivera (2021) and Rindt et al. (2020), which are both based on kernel methods. In the former, the authors specifically address the two-sample problem for right-censored data. While the present test may be seen as related, the two-sample analysis is quite different, and heavily relies on the binary nature of the covariates: the main results apply ad-hoc theory developed for log-rank tests, which is not available in our setting. As noted above, Rindt et al. (2020) bypass the problem of right-censored data by transforming it into uncensored samples, however this comes at the cost of losing considerable information

in the data. By contrast, our approach deals directly with the censored observations without loss of information, resulting in a major performance advantage in practice, as we demonstrate in our experiments.

The paper is structured as follows. In Section 4.2 we introduce relevant notation. In Section 4.3 we define the kernel log-rank test and show that it can be interpreted as i) the supremum of a collection of score tests associated to a particular family of cumulative hazard functions, and, ii) as an RKHS distance, revealing a similarity with the HSIC Gretton et al. (2012b). In Section 4.4 we study the asymptotic behavior of our statistic under both the null and alternative hypothesis, and we establish connections with known approaches such as the two-sample test proposed in Fernández and Rivera (2021), and the Cox score test. Section 4.5 shows how to effectively approximate the null distribution by using Wild Bootstrap Dehling and Mikosch (1994). Sections 4.6 and 4.7 contain extensive experiments investigating the performance of the kernel log-rank test on a range of synthetic and real datasets.

4.2 Notation

4.2.1 Survival analysis notation

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

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

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

4.2.2 Counting processes notation

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

We assume that all our random variables take values on a common filtrated probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$, where the sigma-algebra $\mathcal{F}_t$ is generated by

$$\left\{ \mathbb{1}_{\{T_i \leq s, \Delta_i = 0\}}, \mathbb{1}_{\{T_i \leq s, \Delta_i = 1\}}, X_i : s \leq t, i \in [n] \right\},$$

and the $\mathbb{P}$-null sets of $\mathcal{F}$. Under the null hypothesis, for $i \in [n]$, we define the individual and pooled $(\mathcal{F}_t)$-martingales, $M_i(t) = N_i(t) - \int_{(0,t]} Y_i(s) d\Lambda_Z(s)$ and $M(t) = N(t) - \int_{(0,t]} Y(s) d\Lambda_Z(s)$, respectively. Finally, we denote by $d\widehat{\Lambda}(t) = dN(t)/Y(t)$ the Nelson Aalen estimator of $d\Lambda_Z(t)$ under the null hypothesis. For more information about counting processes martingales, we refer the reader to Fleming and Harrington

(Fleming and Harrington 1991; Chapters 1 and 2).

In this work $\int_a^b$ means integration over $(a, b]$ unless $b = \tau$, in which case we integrate over (a, τ). Due to the simple nature of the martingales that appear in this work (which arise from counting processes), properties such as (squared-)integrability of these processes are trivial, and thus we state them without proof. Also, note that for any $t > \tau_n$, it holds that $N(t) = N(\tau_n)$ and $M(t) = M(\tau_n)$. Hence $\int_{\mathbb{R}_+} g(t)dN(t) = \int_0^\tau g(t)dN(t) = \int_0^{\tau_n} g(t)dN(t)$; the same holds for the martingale M.

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

4.3 Construction of the test

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

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

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

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

4.3.1 General score test

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

$$\Lambda_{Z|X=x}(t;\theta,\omega) = \int_0^t e^{\theta\omega(s,x)} d\Lambda_Z(s) \quad \theta \in \Theta, \tag{4.1}$$

where $\omega : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ is some non-zero fixed function, Θ is an open subset of $\mathbb{R}$ containing $\theta = 0$, and $\Lambda_Z(s)$ is the marginal (baseline) cumulative hazard function associated to the failure time Z. We note that in the above model ω can be scaled by an arbitrary constant, which can be incorporated in θ or in the baseline hazard. In this section we take ω and Λ to be a fixed and develop a test of $H_0 : \theta = 0$. That is, we first propose a test of $H_0 : \theta = 0$, for fixed ω and baseline hazard. As in practice ω is unknown, we later optimize the test statistic over functions ω with norm at most 1 and estimate the baseline hazard using the Nelson-Aalen estimator.

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

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

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

under the model in Equation (4.1) corresponds to

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

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

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

and $U_n(0,\omega)$ is the score statistic associated to the null hypothesis, $H_0 : \theta = 0$. A normalized version of $U_n(0,\omega)$ can be obtained using the variance/covariance matrix of $U_n(\theta;\omega)$, written as $\Sigma(\theta;\omega) = \mathbb{E}(-\frac{\partial^2}{\partial\theta^2}\log L_n(\theta;\omega))$, and then writing $S_n(0;\omega) = U_n(0;\omega)^\intercal\Sigma(0;\omega)^{-1}U_n(0;\omega)$. By the Neyman-Pearson Lemma (Neyman and Pearson (1933)), it follows that the test based on $S_n(0;\omega)$ is the most powerful test for small deviations from the null under the model defined in Equation (4.1).
In general the marginal hazard function $d\Lambda_Z(s)$ is unknown, and thus $U_n(0;\omega)$ cannot be evaluated in practice. However, under the null, $d\Lambda_Z(s)$ can be estimated from the data using the Nelson-Aalen estimator Aalen (1978) $d\widehat{\Lambda}_Z(t) = dN(t)/Y(t)$, yielding

$$\widehat{U}_n(0;\omega) = \sum_{i=1}^{n}\int_{\mathbb{R}_+}(\omega(t,X_i) - \bar{\omega}_n(t))dN_i(t), \tag{4.2}$$

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

4.3.2 Log-rank formulation

The expression for the un-normalized score statistic given in Equation (4.2) can be written as a discrepancy between two empirical measures with respect to the weight

function $\omega(t,x)$. In survival analysis terminology, this is known as weighted log-rank test, and, in our scenario, it takes the form

$$\mathrm{LR}_n(\omega) = \frac{1}{n}\widehat{U}_n(0;\omega) = \int_{\mathbb{R}_+}\int_{x\in\mathbb{R}^d} \omega(t,x)\left(d\nu_1^n(t,x) - d\nu_0^n(t,x)\right), \tag{4.3}$$

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

$$d\nu_1^n(t,x) = \frac{1}{n}\sum_{i=1}^n dN_i(t)\delta_{X_i}(x) = \frac{1}{n}\sum_{i=1}^n \Delta_i\delta_{T_i,X_i}(t,x) \tag{4.4}$$

and

$$d\nu_0^n(t,x) = \frac{dN(t)}{n}\sum_{i=1}^n \frac{Y_i(t)}{Y(t)}\delta_{X_i}(x) = \frac{1}{n}\sum_{j=1}^n \Delta_j\delta_{T_j}(t)\sum_{i=1}^n \frac{Y_i(t)}{Y(t)}\delta_{X_i}(x). \tag{4.5}$$

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

Theorem 4.1. *Let $\omega : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ be a bounded measurable function. Then*

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

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

Simple algebra shows that, under the null hypothesis (i.e., $H_0 : Z \perp X$), $\nu_1 = \nu_0$, and consequently $\mathrm{LR}_n(\omega) \xrightarrow{\mathbb{P}} 0$ for any weight function $\omega(t,x)$. Under some regularity conditions which we state in Assumption 4.2, we prove that $\nu_1 = \nu_0$ implies $Z \perp X$.

Assumption 4.2. *For almost all $x \in \mathbb{R}^d$, $S_{C|X=x}(t) = 0$ implies $S_{Z|X=x}(t) = 0$.*

Proposition 4.3. *Under Assumption 4.2, it holds that $\nu_1 = \nu_0$ if and only if $Z \perp X$.*

Note that $\mathrm{LR}_n(\omega) \xrightarrow{\mathbb{P}} 0$ does not necessarily imply $\nu_0 = \nu_1$, since, if we choose ω equal to the zero function, then $\mathrm{LR}_n(\omega) = 0$ trivially. Thus, when using log-rank tests, it

is very important to use a relevant weight function for the problem at hand. Instead of choosing a single weight function, we propose to optimize over a large collection of candidate functions.

4.3.3 RKHS approach

While normalized log-rank tests exhibit good statistical properties for small deviations from alternatives belonging to the model in Equation (4.1), this good behavior is only guaranteed for a single weight function ω at a time. In practice, it is very unlikely that the dependence structure of Z and X is known beforehand, and thus choosing the correct weight $\omega(t, x)$ (if it exists) seems hard.

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

$$\Psi_n^2 = \left(\sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}^2 \leq 1} \mathrm{LR}_n(\omega) \right)^2, \tag{4.6}$$

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

In particular, we choose $\mathcal{H}$ as a reproducing kernel Hilbert space (RKHS) of functions. One of the main advantages of choosing this particular space, is that it gives a simple closed-form solution for the optimization problem of Equation (4.6). For general spaces of functions, finding the function $\widehat{\omega}$ that maximizes the likelihood function, or solving the optimization problem of Equation (4.6), might be much harder problem, as it is likely that $\widehat{\omega}$ does not have a closed-form solution. We will prove that, under some mild regularity assumptions, a sufficiently rich choice of RKHS will be able to detect any type of dependencies. Comparing with works that consider a maximum among normalised log-rank statistics (i.e., divided by the standard deviation) Kössler (2010); Tarone (1981); Garès et al. (2015), our test will use the un-normalised statistic

$\mathrm{LR}_n(\omega)$. This is fundamental to our result, as the linearity in ω of $\mathrm{LR}_n(\omega)$, combined with the properties of the RKHS, leads to a simple closed formula to evaluate Ψ_n^2. This being said, note that we are indirectly normalizing by choosing ω in the unit ball of $\mathcal{H}$.

4.3.4 Reproducing Kernel Hilbert Spaces

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

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

whose existence is guaranteed by $\int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \sqrt{\mathfrak{K}((t,x),(t,x))} d\nu(t,x) < \infty$, see Berlinet and Thomas-Agnan (2004).

4.3.5 RKHS distance

We define the embeddings of the empirical measures ν_1^n and ν_0^n (introduced in Equations (4.4) and (4.5)) into $\mathcal{H}$ with reproducing kernel $\mathfrak{K}$ as

$$\phi_1^n(\cdot) = \int_0^\tau \int_{\mathbb{R}^d} \mathfrak{K}((t,x),\cdot) d\nu_1^n(t,x) \quad \text{and} \quad \phi_0^n(\cdot) = \int_0^\tau \int_{\mathbb{R}^d} \mathfrak{K}((t,x),\cdot) d\nu_0^n(t,x), \quad (4.7)$$

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

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

Theorem 4.4.

$$\Psi_n^2 = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i\Delta_j\bar{\mathfrak{K}}_n((T_i, X_i), (T_j, X_j)), \tag{4.8}$$

where

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

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

$$\Psi_n^2 = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 = \frac{1}{n^2}\operatorname{Tr}(\boldsymbol{L}^{\boldsymbol{\Delta}}(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{A})^{\intercal}) \tag{4.9}$$

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

4.4 Asymptotic Analysis

4.4.1 Asymptotic null distribution

We study the asymptotic null distribution of $n\Psi_n^2$ which is fundamental to construct a testing procedure. The key step is to show that we can rewrite Ψ_n^2 as a V-statistic plus an asymptotically negligible term. The asymptotic null distribution of $n\Psi_n^2$ then

follows from the standard theory of V-statistics. We refer to (Serfling 1980; Section 5.5.2) for a discussion of V-statistics.

Proposition 4.5. *Under the null hypothesis $H_0 : Z \perp X$, the kernel log-rank statistic can be written as*

$$\Psi_n^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} J_n((T_i, \Delta_i, X_i), (T_j, \Delta_j, X_j)), \tag{4.10}$$

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

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

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

The expression in Equation (4.10) suggests a V-statistic representation for the kernel log-rank statistic. In the next result, we prove that $n\Psi_n^2$ can, indeed, be approximated by a V-statistic, by showing that J_n can be replaced by its population version J, which follows from replacing the random kernel $\bar{\mathfrak{K}}_n$ by its corresponding population version $\bar{\mathfrak{K}}$, given by

$$\begin{aligned} &\bar{\mathfrak{K}}((t,x),(t',x')) \qquad\qquad (4.11)\\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\mathfrak{K}((t,x),(t',x')) - \mathfrak{K}((t,x),(t',y')) \frac{S_{C|X=y'}(t')}{S_C(t')} \right. \\ &\left. - \mathfrak{K}((t,y),(t',x')) \frac{S_{C|X=y}(t)}{S_C(t)} + \mathfrak{K}((t,y),(t',y')) \frac{S_{C|X=y}(t) S_{C|X=y'}(t')}{S_C(t) S_C(t')} \right) dF_X(y) dF_X(y'), \end{aligned}$$

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

Assumption 4.6. *$\mathfrak{K}((t,x),(t',x')) = L(t,t')K(x,x')$, and both K and L are bounded.*

Lemma 4.7. *Under Assumption 4.6 and the null hypothesis, it holds*

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

where

$$J((s, c, x), (s', c', x')) = \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \bar{\mathfrak{K}}((t, x), (t', x'))dm_{s,c}(t)dm_{s',c'}(t). \tag{4.12}$$

It can be easily checked that $\mathbb{E}(J((t, c, x), (T_1, \Delta_1, X_1))) = 0$ for any $(t, c, x) \in \mathbb{R}_+ \times \{0, 1\} \times \mathbb{R}^d$ under the null (since $dm_{T_i,\Delta_i}(t) = dM_i(t)$). The statistic Ψ_n^2 is then approximately a degenerate V-statistic, and thus we deduce its limit distribution from the classical theory of degenerate V-statistics (Serfling 1980; Section 5.5.2).

Theorem 4.8. *Under Assumption 4.6 and the null hypothesis, it holds that*

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

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

The next result states that if we directly replace $\bar{\mathfrak{K}}_n$ by its limit $\bar{\mathfrak{K}}$ in Equation (4.8), the resulting test-statistic has the same asymptotic null distribution as $n\Psi_n^2$.

Theorem 4.9. $n\Psi_n^2$ *and* $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i\Delta_j\bar{\mathfrak{K}}((T_i, X_i), (T_j, X_j))$ *have the same asymptotic distribution under the null hypothesis.*

4.4.2 Power under alternatives

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

Lemma 4.10. *Under Assumption 4.6, it holds* $\Psi_n^2 \xrightarrow{\mathbb{P}} \|\phi_0 - \phi_1\|_{\mathcal{H}}^2$, *where*

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

and ν_1 *and* ν_0 *are the population measures defined in Theorem 4.1.*

The next step is to ensure that $\|\phi_0 - \phi_1\|_{\mathcal{H}}^2$ is zero if and only if the null hypothesis holds. This result will follow from assuming conditions on the kernel $\mathfrak{K}$ that ensure the embeddings of the measures ν_0 and ν_1 onto $\mathcal{H}$ are injective, and from Proposition 4.3, which proves that $\nu_0 = \nu_1$ if and only if the null hypothesis holds.

Theorem 4.11. *Let* $\gamma > 0$ *be any constant. Suppose that both* L *and* K, *are bounded, continuous, characteristic Sriperumbudur et al. (2011), translation invariant, and* c_0*-kernels. Then, under the alternative hypothesis, and under Assumptions 4.2 and 4.6,* $n\Psi_n^2 \to \infty$ *as* n *grows to infinity, and thus*

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

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

Under the assumptions of Theorem 4.11, our testing procedure has asymptotic power tending to one for any alternative, and thus *it is able to detect any type of dependency* between survival times and covariates, given enough observations. Even if the kernel does not satisfy the properties stated in Theorem 4.11, however, we can guarantee the power of the test for alternatives following the model of Equation (4.1), that is, for alternatives of the form $\Lambda_{Z|X=x}(t; \theta) = \int_0^t e^{\theta\omega^\star(s,x)} d\Lambda_Z(s)$ for some $\omega^\star \in \mathcal{H}$ on the unit ball and $\theta \neq 0$, as the log-rank statistic $\mathrm{LR}_n(\omega^\star)^2 \to c > 0$, with c a positive

constant. Then,

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

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

4.4.3 Recovering existing tests

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

Example 4.12 (Two-sample weighted log-rank test)**.** *Consider $X \in \{0, 1\}$, i.e., the two-sample problem. We can recover the standard weighted log-rank test with arbitrary weight function $\tilde{\omega} : \mathbb{R}_+ \to \mathbb{R}$, by choosing $\omega(t, 1) = -\omega(t, 0)$ and $\omega(t, 0) = \tilde{\omega}(t)/2$. Then, by replacing ω into Equation* (4.3)*, we obtain*

$$LR_n(\omega) = \frac{1}{n} \int_0^\tau \tilde{\omega}(t) L(t) (d\widehat{\Lambda}^0(t) - d\widehat{\Lambda}^1(t)), \tag{4.13}$$

where $d\widehat{\Lambda}^j$ denotes the Nelson-Aalen estimator for each group $j \in \{0, 1\}$. Furthermore, $\tilde{\Psi}_n = \sup_{\omega: \|\omega\|_{\mathcal{H}}=1} LR_n(\omega)$ recovers the general test proposed in Fernández and Rivera (2021).

Example 4.13 (Cox proportional hazards model)**.** *Consider the Hilbert space of functions $\omega(t, x) = V^{1/2}\beta^\intercal x$, where $\beta \in \mathbb{R}^d$ and V is a positive-definite matrix. By using this space of functions, our kernel log-rank statistic becomes*

$$\Psi_n = \sup_{\beta \in \mathbb{R}^d: \|V^{1/2}\beta\|^2 \leq 1} LR_n(\beta), \tag{4.14}$$

and it can be computed using Equation (4.9) *with a linear kernel on the covariates,*

$K(x, x') = (V^{1/2}x)^\intercal(V^{1/2}x')$, *and a constant kernel on times,* $L(t, t') = 1$. *Then*

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

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

4.5 Wild Bootstrap

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

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

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

In this section we prove two main results. The first result establishes that, under the null hypothesis, the asymptotic distribution of $n(\Psi_n^W)^2$ coincides with the asymptotic distribution of the kernel log-rank test-statistic $n\Psi_n^2$. The second result is analogous to Theorem 4.11, but replacing γ by the $1-\alpha$ quantile obtained by the Wild Bootstrap procedure.

Lemma 4.14. *Under Assumption 4.6, it holds that*

$$(\Psi_n^W)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}W_iW_j\Delta_i\Delta_j\bar{\mathfrak{K}}((T_i, X_i), (T_j, X_j)) + o_p(n^{-1}). \tag{4.15}$$

Our first main result is given in the following Theorem.

Theorem 4.15. *Suppose that the null hypothesis and Assumption 4.6 hold true, and let $\mathcal{L}$ denote the asymptotic distribution of $n\Psi_n^2$. Then, for almost all sequences $(t_i, \delta_i, x_i)_{i\geq 1}$ sampled from $(T_i, \Delta_i, X_i)_{i\geq 1}$, $n(\Psi_n^W)^2 \xrightarrow{\mathcal{D}} \mathcal{L}$, as $n \to \infty$.*

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

Our second main result is given in the following theorem.

Theorem 4.16. *Consider Assumptions 4.2 and 4.6, and assume that both L and K, are bounded, continuous, characteristic, translation invariant, and c_0-kernels. Let $\alpha \in (0, 1)$, and let $Q_{n,M}^W$ denote the $1-\alpha$ quantile obtained from a sample of fixed size M of the Wild Bootstrap test-statistic, $n(\Psi_n^W)_1^2, \ldots, n(\Psi_n^W)_M^2$. Then, under the alternative hypothesis*

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

From the previous result, we deduce that, under the Assumptions of Theorem 4.16, the test based on the Wild Bootstrap approximation of the null distribution is able to *detect any type of dependency between survival times and covariates* asymptotically, as long as censoring does not hide the regions in which the dependence occurs.

4.5.1 Implementation

Under Assumption 4.6, the Wild Bootstrap test-statistic can be easily evaluated as follows,

$$n(\Psi_n^2)^W = \frac{1}{n^2}\operatorname{Tr}(\boldsymbol{L}^{\boldsymbol{\Delta},\boldsymbol{W}}(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{A})^\intercal), \tag{4.16}$$

where $\boldsymbol{L}^{\boldsymbol{\Delta},\boldsymbol{W}}$ is a $(n \times n)$-matrix defined as $(\boldsymbol{L}^{\boldsymbol{\Delta},\boldsymbol{W}})_{i,j} = (\Delta_i W_i)(\Delta_j W_j)L(T_i, T_j)$, and $\boldsymbol{K}, \boldsymbol{A}$ and $\boldsymbol{I}$ are defined in Theorem 4.4. Algorithm 7 below describes the implementation of our testing procedure.

4.5.2 Computational time

By the following Proposition, our algorithm has the same computational complexity as the HSIC based permutation test of Gretton et al. (2012b).

Proposition 4.17. *$n\Psi_n^2$ and $n(\Psi_n^W)^2$ can be computed in $\mathcal{O}(n^2)$ time.*

Using a simple Python implementation that does not use a GPU, running on a CPU with 4 cores at 1.6GHz, computation of the kernel log-rank statistic takes about 10 seconds for a sample of size 10000, and about 0.1 second for a sample of size 1000. If faster computation is required, we may adopt the large-scale approximations proposed in Zhang et al. (2018). Moreover, Wild Bootstrap statistics can be computed in parallel, and matrix computations can be done on a GPU.

Algorithm 7: Wild Bootstrap.

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

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

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

  Compute $(\Psi_n^W)_k^2$ as in equation (4.16)

end

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

Compute $n\Psi_n^2$ as in Equation (4.9)

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

4.6 Experiments

We study the performance of the proposed kernel log-rank test for various choices of kernels. We choose the kernels to be products of a kernel on the covariates, K, and a kernel on the times, L. We denote the product kernel by (K, L). We study the following four cases: 1. $(K = \text{Lin}, L = 1)$, 2. $(\text{Gau}, 1)$, 3. $(\text{Fis}, 1)$ and 4. (Gau, Gau), where "Lin" denotes the linear kernel, "Gau" the exponentiated quadratic (Gaussian) kernel, "Fis" the linear kernel scaled by the Fisher information (see Example 4.13) and "1" denotes the constant kernel, i.e. $L = 1$ implies $L(t, s) = 1$ for all $t, s \in \mathbb{R}_+$. In all experiments we use the *median heuristic* to select the bandwidth of the Gaussian kernel: we choose $\sigma^2 = \text{median}\{||x_i - x_j||^2 \ : i \neq j\}/2$. We discuss the sensitivity of the test to different choices of bandwidth later this section. We set the level of the test to $\alpha = 0.05$ and use Algorithm 7 of Section 4.5 to perform the test with $M = 2000$ (or $M = 5000$ to estimate type 1 error) Wild Bootstrap samples to estimate the rejection region. We compare the kernel log-rank test with the traditional Cox likelihood ratio (Cox LR) test Cox (1972), denoted by Cph in the legends, and the optHSIC test Rindt et al. (2020), denoted by Opt in the legends. As in Rindt et al. (2020), we use the Brownian covariance kernel in optHSIC. The assessment of the Type I error can be

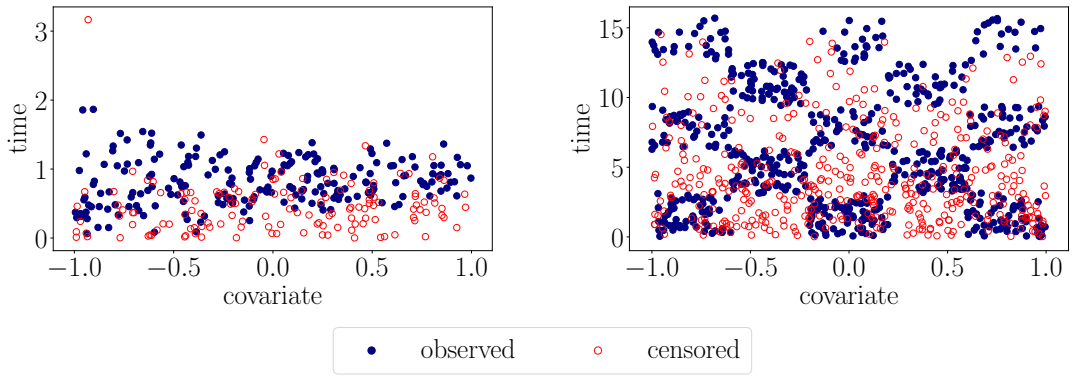


Figure 4.1: Scatterplots of samples from D.3 (left) and D.4 (right).

found in Appendix 4.9.1.1. Code to implement the kernel log-rank test and reproduce the experiments below can be found at `https://github.com/davidrindt/kernel_logrank_python_code`.

4.6.1 Power for 1-dimensional covariates

In this section we investigate the power of the different tests using data simulated from distributions in which $X \not\perp Z$. In each case, we use an exponential censoring distribution in which $C \perp X$ and the mean of C is chosen such that 60% of the events are observed. Throughout this subsection we let $X \sim \text{Unif}[-1, 1]$. Consider the following four distributions.

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

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

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

D.4 A checkerboard pattern: See Figure 4.1. Because the pattern is more complicated, we let the sample size range from 500 to 2000 in steps of 500.

The top row of Figure 4.2 displays the rejection rates of the tests for samples from D.1 and D.2. Note that in D.1 the kernel log-rank test with linear kernel performs roughly equivalently to the LR test, which is ideally suited for this distribution as the CPH assumption holds. While the kernel log-rank tests with the rich kernels $(\text{Gau}, 1)$ and (Gau, Gau) do not lose much power in detecting this CPH dependency, we do observe a small tradeoff between richness of the kernels and power: the (Gau, Gau) kernel has slightly less power than the $(\text{Gau}, 1)$ kernel, which in turn has slightly less power than the $(\text{Lin}, 1)$ kernel and CPH LR test in D.1. In D.2 the quadratic term in the log-hazard violates the CPH assumption. The top right panel of Figure 4.2 shows that the linear kernel and the CPH LR test are unable to detect the dependency. Again, we find that the least rich kernel that is still able to model the dependency has the highest power, which in this case is the $(\text{Gau}, 1)$ kernel. In D.3 and D.4 the hazard function does not factorize into a function of covariates and a function of time, and a kernel is needed on time to model the dependency. Rejection rates are displayed in the bottom of Figure 4.2, confirming that indeed the kernel log-rank test with kernel (Gau, Gau) is the only test able to reject the null hypothesis consistently at samples of size $n = 1500$.

4.6.2 Power for multidimensional covariates

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

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

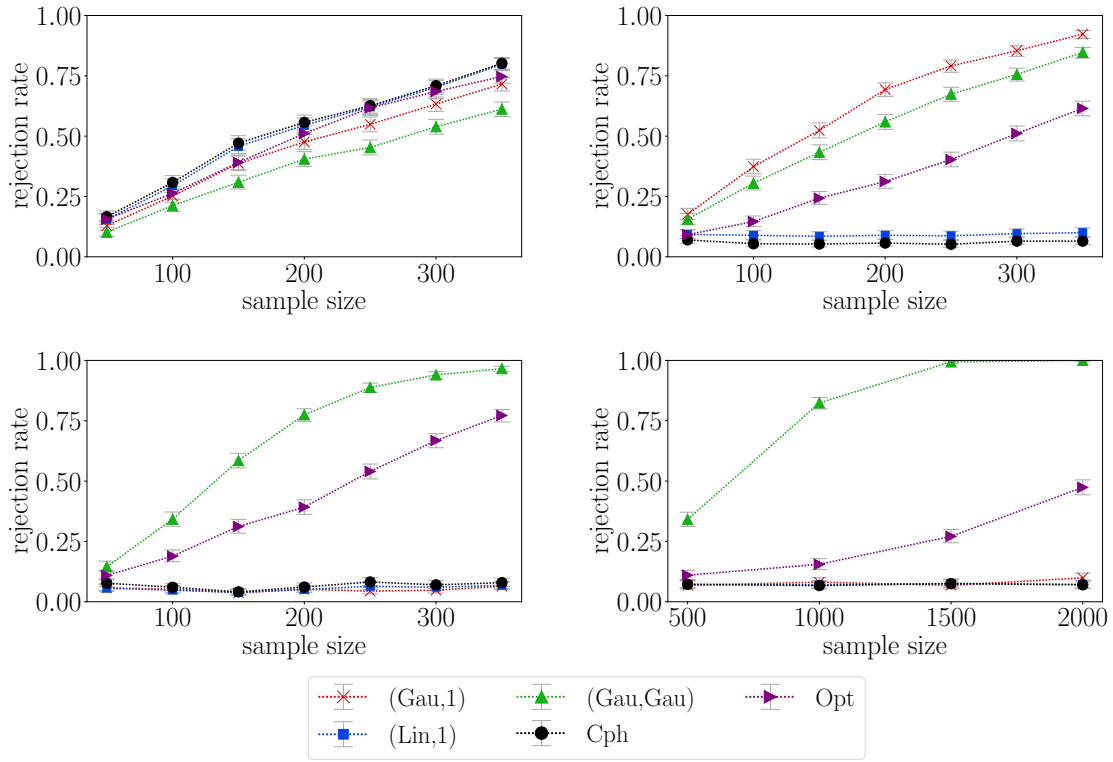


Figure 4.2: Rejection rates of the various test for D.1 (top left), D.2 (top right), D.3 (bottom left) and D.4 (bottom right).

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

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

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

Figure 4.3 displays rejection rates for both varying dimension and sample sizes. As our first finding, we observe that, while the CPH assumption holds for D.5 and D.6, the power of the kernel log-rank test is similar to that of the CPH LR test, which is an encouraging result. In D.7 and D.8 we again observe that in the presence of a non-CPH dependence, the kernel log-rank test has good power.

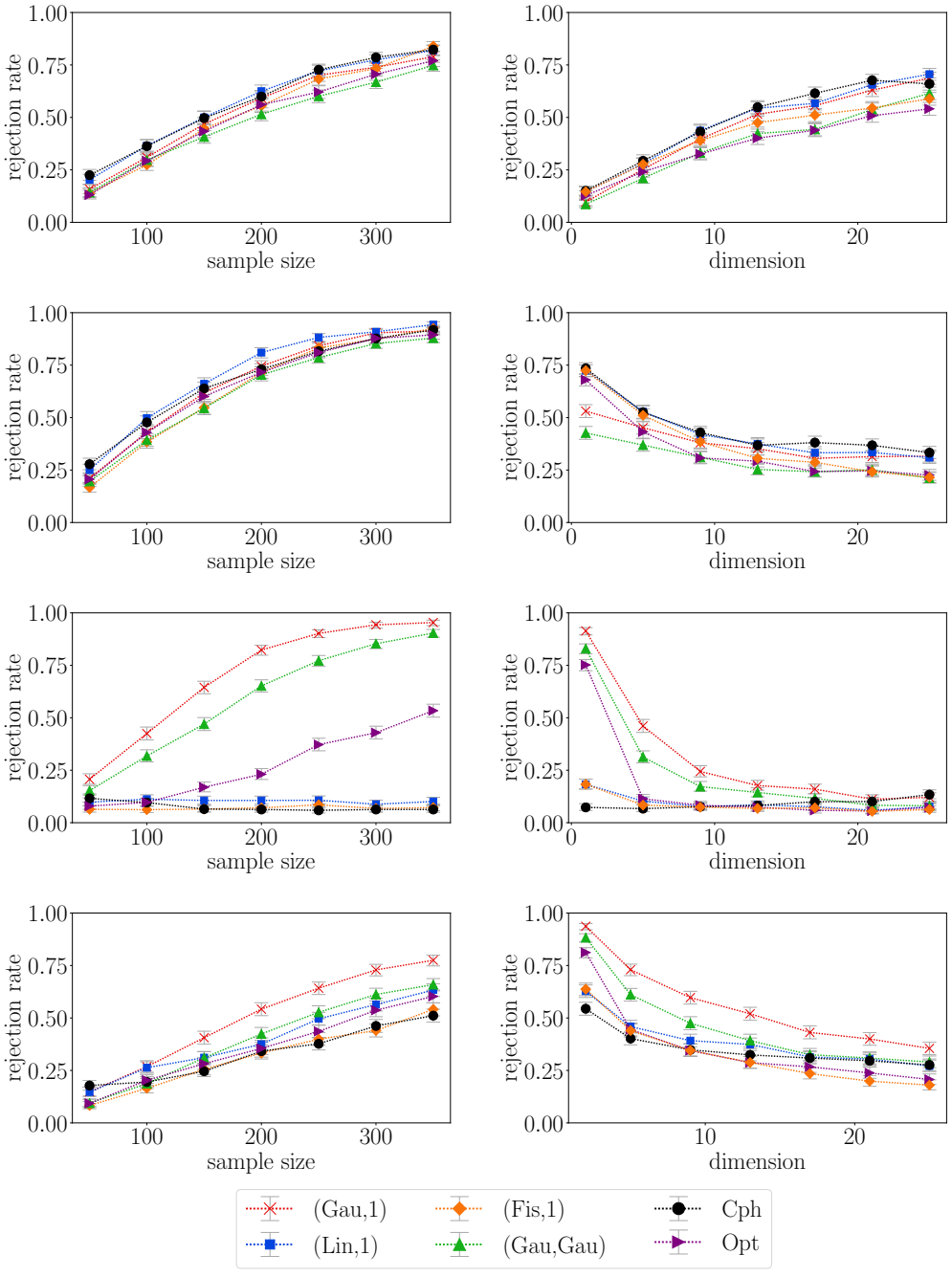


Figure 4.3: Rejection rates for D.5 (top) - D.8 (bottom). Left: rejection rates as the sample size increases. Right: rejection rates as the dimension increases.

4.6.3 Varying censoring rates and distributions

We study the rejection rates in cases where censoring depends on the covariate, and where the percentage of observed ($\Delta = 1$) events is $15, 30, 45, 60, 75, 90$ or 100%. The experiments are described in Appendix 4.9.1.3. Under these varying censoring percentages, the main findings from the previous section remain true: the kernel log-rank test shows competitive power when the CPH assumption holds, is able to detect non-CPH dependencies, and achieves correct type 1 error. A final important observation is that optHSIC loses power compared to the kernel log-rank test for higher censoring rates.

4.6.4 Sensitivity to choice of bandwidth

In the experiments presented thus far we set the bandwidth of the Gaussian kernel to be $\sigma^2 = \text{median}\{||x_i - x_j||^2 \ : i \neq j\}/2$. We now study the effect of varying the bandwidth of the Gaussian kernel on the rejection rate. Details of the experiments are in Appendix 4.9.1.6. We find that, while for most scenarios a bandwidth can be selected that slightly outperforms the median heuristic, the median heuristic has a consistently good performance across all scenarios.

4.6.5 Choice of kernel

When using the kernel log-rank test it is important to choose an appropriate kernel. While in principle we can choose $\mathfrak{K}$ to be any kernel, choosing a kernel that factorizes into a kernel on time and a kernel on the covariates has the important advantage that it gives the simple closed-form expression for our test-statistic in Theorem 4.4. In Theorems 4.11 and 4.16, we prove that our testing procedure is consistent for sufficiently expressive kernels K and L.

Our experiments show that in some scenarios there is a trade off between richness

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

We may also consider improvements on the median heuristic when selecting kernel bandwidth. In Gretton et al. (2012c); Sutherland et al. (2017); Liu et al. (2020), which deal with the case of two-sample testing in the uncensored setting, part of the data is used to select parameters that result in the lowest approximate asymptotic p-value, and the test is then performed with the selected parameters on the remaining data. In Albert et al. (2021), which addresses independence testing in the uncensored setting, an aggregation procedure is proposed over kernel bandwidths, which does not require data splitting, and is adaptive in a minimax sense over Sobolev classes of alternatives. It will be an interesting topic for future research to extend these kernel selection strategies to the censored setting.

4.7 Application to real-world datasets

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

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

Table 4.1: p-values obtained by the various tests for the Biofeedback and Colon data.

4.7.1 Biofeedback data

The biofeedback treatment data studies the time until patients suffering from aspiration after head and neck surgery achieve full swallowing rehabilitation. The study is presented in Denk and Kaider (1997) and the data was made available in the R package Coxphw Dunkler et al. (2018). In our presentation we name the event-time variable Rehab. Covariates in the dataset are a binary variable indicating biofeedback treatment, denoted Trt, and the time after the surgery until treatment could be started, denoted Recov. This dataset contains 33 individuals, of whom 70% were observed to fully rehabilitate (coded as $\delta = 1$).

In the first row of Table 4.1, we observe that the kernel log-rank test results in the p-values 0.014 for the kernel (Gau, 1) and 0.023 for the kernel (Gau, Gau) when including both covariates. In contrast, the CPH likelihood ratio test results in the higher p-value of 0.496. This suggests the possibility of a non-linear relationship between the event-time of interest, Rehab, and the covariates Trt and Recov. This interpretation is strengthened by the following two observations. First, when we apply a logarithm transformation to the covariate Recov, as suggested in Denk and Kaider (1997), the CPH likelihood ratio test (shown in the fourth row of Table 4.1) results in a p-value of 0.027. The main advantage of the kernel log-rank test is that it does not require us to manually transform the data to detect potential non-linear dependencies. Second, we

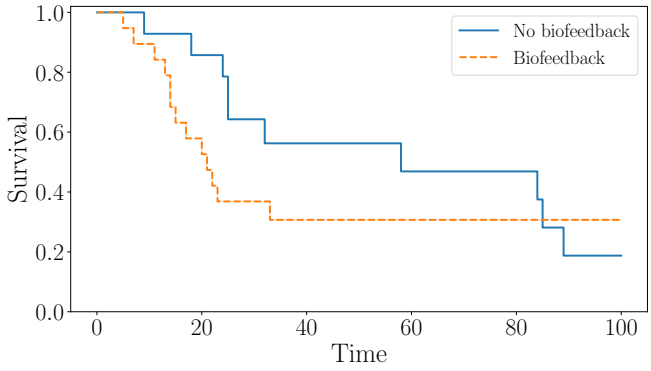


Figure 4.4: Kaplan-Meier estimates of the survival curves associated to the group that receives the Biofeedback treatment and one that does not receive the treatment. We observe that both curves cross which violates the assumption of proportional hazards.

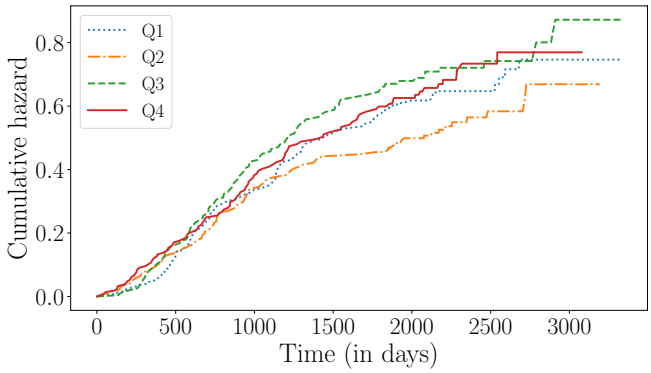


Figure 4.5: Nelson-Aalen estimates of the cumulative hazard for the four quartiles of age. Q1 denotes the youngest quartile and Q4 the oldest. Note that there is no monotonic relationship between the age quartile and the cumulative hazard.

observe that the kernel log-rank test using the (Gau, Gau) kernel is the only test that rejects the null hypothesis of independence between Rehab and Trt (see the second row of Table 4.1) at a significance level of 0.05. This result may be attributed to the fact that the survival curves associated to the two treatment groups cross (see Figure 4.4).

4.7.2 Colon data

The Colon dataset studies the recurrence of tumors and survival in patients undergoing treatment for stage B/C colon cancer. The study is presented in Laurie et al. (1989) and Moertel et al. (1995). Data from the 929 patients in the study is available in the R package Survival Therneau and Lumley (2015). Each individual has 11 covariates. We consider the time of death in our analysis as the event-time, which was observed, i.e. $\delta = 1$, for 49% of the individuals. We focus our analysis on the variables Age, Perfor (a binary variable indicating the perforation of the colon) and Adhere (a binary variable indicating adherence of the tumor to nearby organs).

The p-values of different tests of independence are given in Table 4.1. Testing for independence between the age covariate and the event-time, the kernel log-rank test results in p-values of 0.080 and 0.097 for the $(\text{Gau}, 1)$ and (Gau, Gau) kernels respectively. In contrast, the CPH likelihood ratio test produces a much higher p-value of 0.627. This suggests the possibility of a non-monotonic effect of the age variable on the event-time of interest. This interpretation is strengthened by Figure 4.5, which displays cumulative hazard functions for different age quartiles. Indeed, these curves could suggest a non-monotonic relationship between age and (cumulative) hazard. In particular we observe that first quartile has a higher cumulative hazard function than the second quartile, but lower than the third. Finally, in the last row of Table 4.1, we observe differences in the conclusions of the kernel log-rank test (for both kernels) and the CPH likelihood ratio test at a significance of $\alpha = 0.05$ when we include the binary variables Perfor and Adhere to the model. This difference in p-values suggests there may be a non-linear relationship between the event-time of interest and these covariates.

4.8 Conclusions

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

4.9 Supplementary material

1. A. Additional experiments. In Section 4.9.1.1 we study the type 1 error, in Section 4.9.1.2 we show additional experiments regarding the power of tests, in Section 4.9.1.3 we show experiments for varying censoring percentages, and in Section 4.9.1.6 we show experiments for varying bandwidths of the Gaussian kernel.

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

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

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

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

4.9.1 Additional experiments

Here we present some experiments that were not described in detail in the main text.

4.9.1.1 Type 1 error rate

To estimate the false rejection rate of our proposed test, we repeatedly take samples from distributions in which $Z \perp X$ and we count the proportion of experiments in which the kernel log-rank test rejects the null hypothesis. We set the significance level at $\alpha = 0.05$ and let the sample sizes range from 50 to 350 in steps of 50. For each distribution and sample size, we take 5000 samples. Table 4.2 shows the distributions we sample from. Figure 4.6 shows scatterplots of samples from distributions 1-4.

D.	$Z\|X$	$C\|X$	X
1	Exp(mean $= 0.66$)	Exp(mean $= \exp(X/3)$)	Unif[-1,1]
2	Exp(mean $= 0.9$)	Exp(mean $= \exp(X^2)$)	Unif[-1,1]
3	Exp(mean $= 0.9$)	Weib(mean $= 3.25 + 1.75X$)	Unif[-1,1]
4	Exp(mean $= 0.9$)	$1 + X$	Unif[-1,1]
5	Exp(mean $= 0.6$)	Exp(mean $= \exp(1^T X)$)	$\mathcal{N}_{10}(0, \text{cov} = \Sigma_{10})$
6	Exp(mean $= 0.6$)	Exp(mean $= \exp(X_1)$)	$\mathcal{N}_{10}(0, \text{cov} = \Sigma_{10})$

Table 4.2: The distributions used to estimate the type 1 error rate. Exp(mean) $= \mu$ denotes the exponential distribution with mean μ. We define $\Sigma_{10} = MM^T$ where M is a 10×10 matrix of i.i.d. standard normal entries. Parameters are chosen such that approximately 60% of events are observed ($\Delta = 1$).

We estimate the bandwidth of the Gaussian kernel from the same data to which we apply the test. Similarly we estimate the Fisher information matrix on the same data to which we apply the kernel log-rank test. Due to this dependence of the kernel on the data, the arguments of Section 4.5 do not apply directly. The obtained rejection rates show that this does not lead to increased type 1 error rate. The resulting rejection rates are displayed in Tables 4.3 and 4.4.

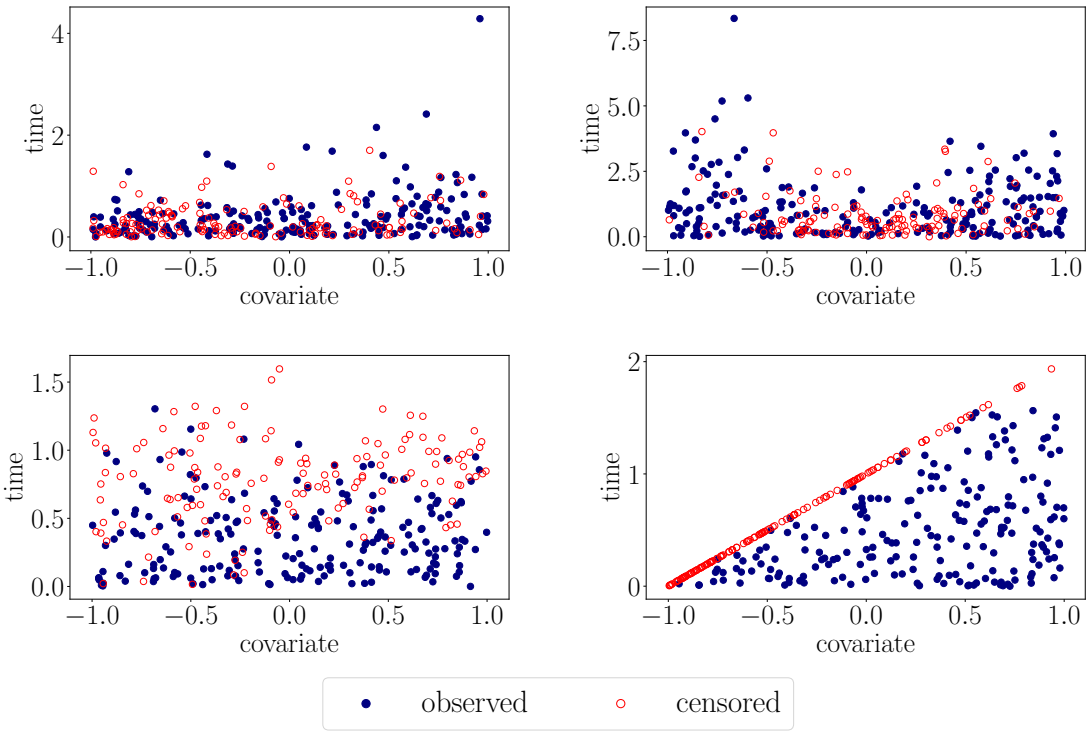


Figure 4.6: Top: Scatterplots of samples from D.1 (top left), D.2 (top right) D.3 (bottom left) and D.4 (bottom right) defined in Table 4.2.

		Sample size $n =$						
D.	Method	50	100	150	200	250	300	350
1	Cph	0.053	0.051	0.049	0.051	0.055	0.048	0.053
	(Lin,1)	0.037	0.049	0.050	0.047	0.055	0.048	0.044
	(Gau,1)	0.046	0.046	0.045	0.050	0.045	0.052	0.051
	(Gau,Gau)	0.047	0.044	0.044	0.048	0.045	0.054	0.052
	Opt	0.050	0.051	0.055	0.046	0.051	0.053	0.054
2	Cph	0.049	0.060	0.051	0.049	0.056	0.046	0.050
	(Lin,1)	0.044	0.044	0.050	0.047	0.052	0.045	0.047
	(Gau,1)	0.038	0.044	0.048	0.047	0.052	0.057	0.049
	(Gau,Gau)	0.043	0.045	0.051	0.049	0.047	0.050	0.051
	Opt	0.053	0.050	0.051	0.052	0.047	0.050	0.054
3	Cph	0.045	0.052	0.052	0.052	0.049	0.052	0.056
	(Lin,1)	0.039	0.047	0.048	0.048	0.053	0.046	0.046
	(Gau,1)	0.039	0.047	0.041	0.052	0.052	0.051	0.051
	(Gau,Gau)	0.048	0.046	0.051	0.049	0.051	0.052	0.049
	Opt	0.057	0.051	0.055	0.049	0.046	0.053	0.050
4	Cph	0.050	0.051	0.051	0.049	0.052	0.052	0.047
	(Lin,1)	0.047	0.049	0.052	0.049	0.049	0.051	0.052
	(Gau,1)	0.045	0.046	0.048	0.045	0.044	0.053	0.047
	(Gau,Gau)	0.043	0.051	0.049	0.045	0.049	0.052	0.048
	Opt	0.050	0.050	0.046	0.046	0.046	0.047	0.052

Table 4.3: The type 1 error rates of the various methods for D.1-4 defined in Table 4.2. The covariates are 1-dimensional. Rejection rates above 0.057 are displayed in italics.

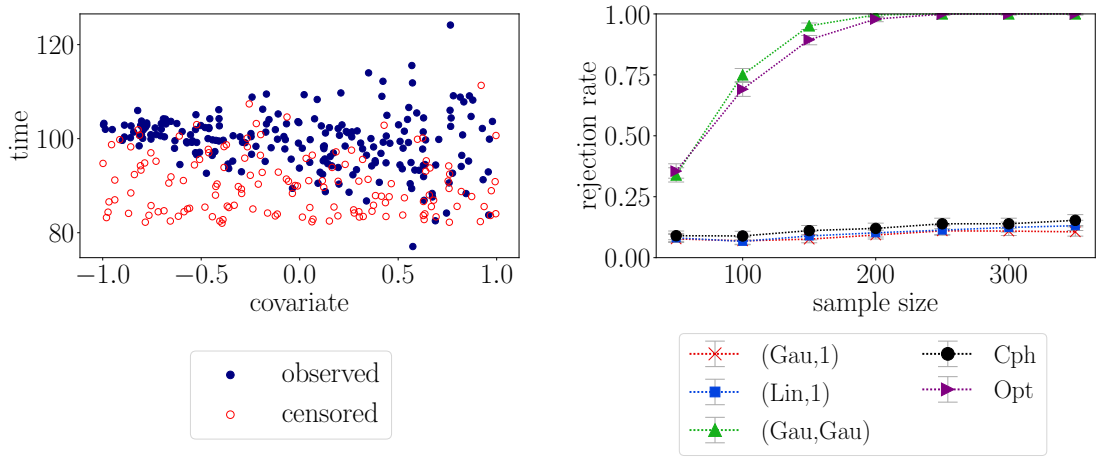


Figure 4.7: A scatterplot and the rejection rates for the normal distribution defined in Section 4.9.1.2.

		Sample size $n =$						
D.	Method	50	100	150	200	250	300	350
5	Cph	*0.122*	*0.080*	*0.064*	*0.068*	*0.067*	*0.060*	*0.060*
	(Lin,1)	0.050	0.051	0.053	0.056	0.050	0.054	*0.058*
	(Fis,1)	0.054	0.045	0.051	0.051	0.050	0.048	0.045
	(Gau,1)	0.046	0.041	0.047	0.053	0.053	0.054	0.054
	(Gau,Gau)	0.044	0.043	0.040	0.049	0.053	0.048	0.049
	Opt	0.052	0.054	0.054	0.047	0.049	0.052	0.054
6	Cph	*0.123*	*0.076*	*0.065*	*0.065*	*0.060*	*0.061*	0.056
	(Lin,1)	0.054	0.051	0.051	0.053	0.046	0.054	*0.058*
	(Fis,1)	0.050	0.048	0.054	0.050	0.050	0.049	0.047
	(Gau,1)	0.044	0.046	0.049	0.051	0.049	0.054	0.049
	(Gau,Gau)	0.047	0.045	0.044	0.053	0.048	0.045	0.054
	Opt	0.049	0.049	0.053	0.050	0.054	0.056	0.047

Table 4.4: The type 1 error rates of the various methods for D.5-6 defined in Table 4.2. The covariates are 10-dimensional. Rejection rates above 0.057 are in italics.

4.9.1.2 Additional experiments on test power

This section contains some experiments that were omitted from Section 4.6.1. First we investigate the power of the tests for a family of normal distributions. Second, we visualize the hazard rates of those normal distributions, and of the Weibull and Checkerboard patterns considered in the main paper. Third, we investigate the power

of the different tests against deviations from the null hypothesis, and lastly we give an example of the effect of the Fisher kernel.

Firstly we consider the following distribution.

> *Normal distributions with different variances:* $Z|X = x \sim \text{Normal}(\text{mean} = 100 - 2.25 \cdot x, \text{var} = 5.5 + x \cdot 4.5)$ and $C|X = x \sim 82 + \text{Exp}(\text{mean} = 35)$.

A scatter plot and rejection rates are displayed in Figure 4.7. The hazard rates of this distribution, as well as of the Weibull and Checkerboard distributions in the main text, are given in Figure 4.8.

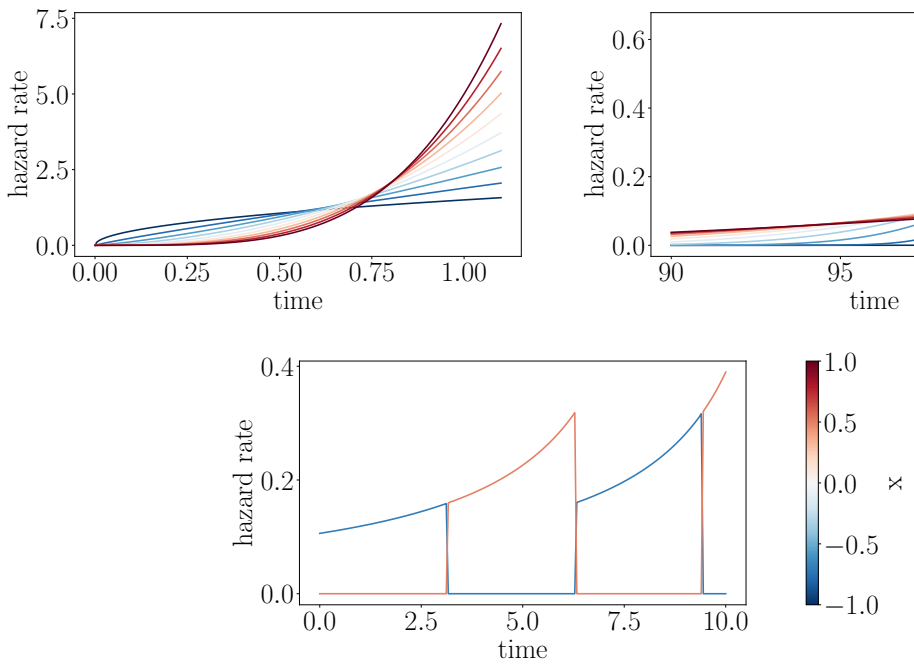


Figure 4.8: Top row left: the hazard rate as a function of time for the Weibull distribution (D.3 of the main text). Top row right: the hazard rate of the normal distributions with varying bandwidths presented in Figure 4.7. Bottom row: the hazard rate of the checkerboard pattern (D.4 of the main text).

While in the main text we investigated the power performance at different sample sizes, we now fix the sample size at 100 and study the power of the various tests for small deviations from the null hypothesis. To do so, we sample $X \sim \text{Normal}(\mu = 0, \sigma^2 = 1)$ and consider two different settings in which θ varies from -0.4 to 0.4 in

steps of 0.1. Notice that, for the two settings described below, $\theta = 0$ recovers the null hypothesis.

Cph distributions: $Z|X = x \sim \text{Exp}(\text{mean} = \exp\{\theta \cdot x\})$ and $C|X = x \sim \text{Exp}(\text{mean} = 1.5)$

Non-linear log-hazard distributions: $Z|X = x \sim \text{Exp}(\text{mean} = \exp\{\theta \cdot x^2\})$ and $C|X = x \sim \text{Exp}(\text{mean} = 1.5)$

Figure 4.9 below shows the rejection rates against deviations from the null for each method.

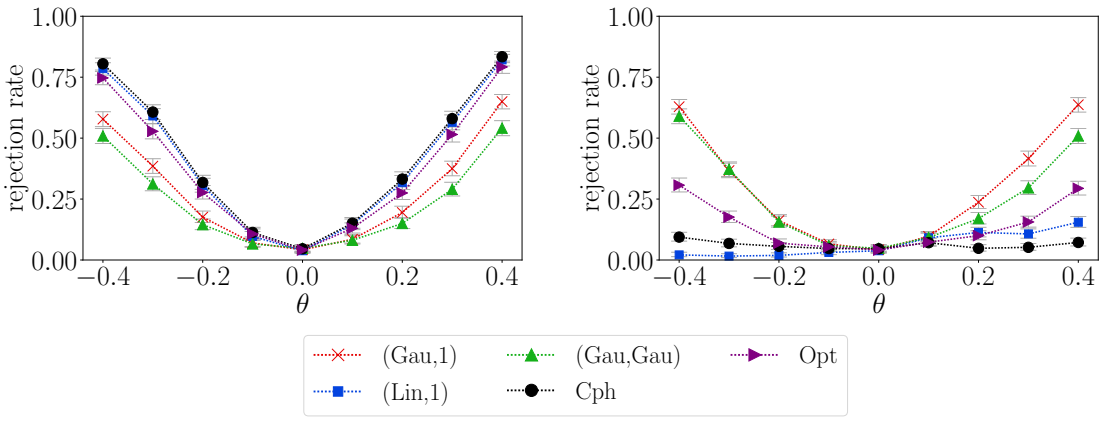


Figure 4.9: Left: the rejection rates of the various tests against the Cph distributions. Right: the rejection rates of the various tests against the non-linear log-hazard distributions.

We now provide an example of the usefulness of the Fisher kernel. In the experiments in the main text, the kernels $(\text{Lin}, 1)$, $(\text{Gau}, 1)$ and (Gau, Gau) all performed consistently well for high dimensional covariates, and thus the usefulness of the Fisher kernel was not obvious. We now show the Fisher kernel has the ability to 'standardize' the data in cases where simple centering and scaling of the individual covariate dimensions is not sufficient. To illustrate this, we consider the following example. Let $\Sigma_{11} = 1/10$, $\Sigma_{ii} = 1$ for $i > 1$ and $\Sigma_{ij} = 0$ otherwise, and let R be an orthogonal rotation matrix. Let $X \sim \text{Normal}(0, R\Sigma R^T)$ and let $v = (1, 0, ..., 0)^T$.

A distribution in which the Fisher information helps uncover the dependency:

Let $Z|X = x \sim \text{Exp}(\text{mean} = \exp\{v^T(R^T x)\})$ and $C|X = x \sim \text{Exp}(\text{mean} = 1.5)$

Figure 4.10 shows that the kernel log-rank tests with kernels (Lin,1) and (Gau,1) do not detect the dependency. The Fisher kernel and CPH LR tests have power, because the inverse information matrix has the effect of standardizing the data.

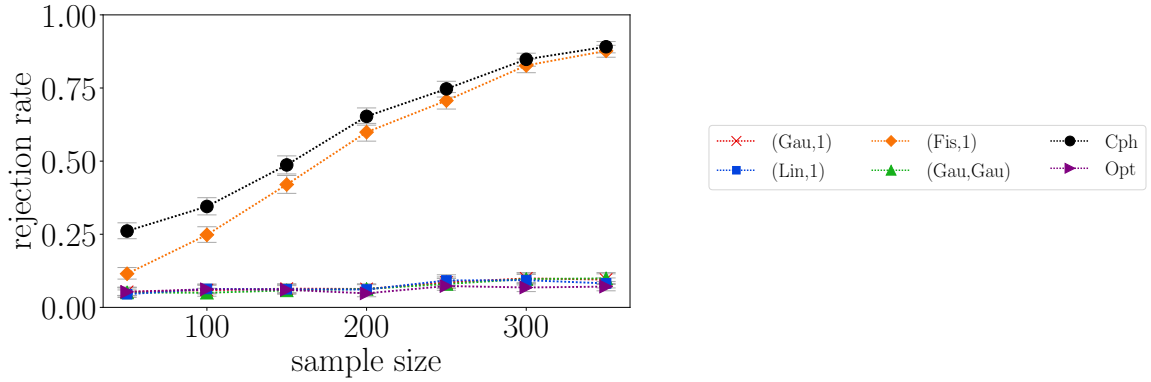


Figure 4.10: The rejection rate of the different methods for the distribution in which the Fisher information helps uncover the dependency.

4.9.1.3 Dependent censoring and varying censoring rates

We now illustrate the performance of the wild bootstrap test using the kernel log-rank statistic when the censoring rate varies, and when the censoring time depends on the covariate. In particular, we parametrize the censoring distributions and vary the parameters to estimate the power of each method when 15, 30, 45, 60, 75, 90 or 100% of the observations are uncensored. We study the type 1 error rate in Section 4.9.1.4 and the power in Section 4.9.1.5.

4.9.1.4 Type 1 error under varying censoring rates

The distributions from which we sample data in which $X \perp Z$ are given in Table 4.5. In each case the sample size is 200. To estimate the rejection rates we sampled 5000 times, and counted the number of times we rejected the null. Obtained rejection rates

are displayed in Tables 4.6 and 4.7. In each of the scenarios we find the type 1 error rate to be correct for each censoring percentage and for each kernel.

D.	$Z\|X$	$C\|X$	X
1	Exp(mean = 1)	Exp(mean = θ)	Unif[-1,1]
2	Exp(mean = 2.2)	Exp(mean = $\theta \exp(X)$)	Unif[-1,1]
3	Weib(shape = 3.25)	Exp(mean = θX^2)	Unif[-1,1]
4	$\mathcal{N}$(mean = 99, var = 5.5)	Exp(mean = $82 + \theta$)	Unif[-1,1]
5	Exp(mean = 1)	Exp(mean = θ)	$\mathcal{N}_{10}(0, \text{cov} = \Sigma_{10})$
6	Exp(mean = 1)	Exp(mean = $\theta \exp(1^T X/30)$)	$\mathcal{N}_{10}(0, \text{cov} = \Sigma_{10})$
7	Exp(mean = exp(0.5)	Exp(mean = $\theta \exp(X_2^2)/20$)	$\mathcal{N}_{10}(0, \text{cov} = \Sigma_{10})$

Table 4.5: The parametrized distributions to test the type 1 error rate under different censoring rates. Here $\Sigma_{10} = MM^T$ where M is a 10×10 matrix of i.i.d. standard normal entries. The parameter θ is varies such that $15, 30, 45, 60, 75, 90, 100\%$ of the individuals are observed (i.e. $\Delta = 1$). The sample size is 200 in each case.

		% Observed						
D.	Method	15%	30%	45%	60%	75%	90%	100%
1	Cph	0.053	0.050	0.047	0.049	0.052	0.050	0.046
	(Lin,1)	0.051	0.046	0.046	0.046	0.049	0.048	0.044
	(Gau,1)	0.046	0.045	0.047	0.050	0.053	0.048	0.044
	(Gau,Gau)	0.045	0.046	0.044	0.046	0.056	0.053	0.045
	Opt	0.048	0.047	0.052	0.045	0.053	0.046	0.045
2	Cph	0.050	0.050	0.049	0.055	0.050	0.057	0.046
	(Lin,1)	0.049	0.049	0.047	0.053	0.050	0.055	0.046
	(Gau,1)	0.050	0.050	0.047	0.045	0.050	0.053	0.045
	(Gau,Gau)	0.046	0.048	0.044	0.044	0.048	0.052	0.049
	Opt	0.053	0.055	0.050	0.051	0.050	0.053	0.049
3	Cph	0.054	0.051	0.050	0.055	0.053	0.054	0.046
	(Lin,1)	0.044	0.048	0.047	0.052	0.051	0.054	0.046
	(Gau,1)	0.043	0.048	0.044	0.049	0.054	0.050	0.046
	(Gau,Gau)	0.042	0.047	0.040	0.048	0.053	0.050	0.047
	Opt	0.054	0.052	0.050	0.053	0.051	0.052	0.050
4	Cph	0.055	0.051	0.054	0.050	0.054	0.052	0.050
	(Lin,1)	0.053	0.047	0.049	0.047	0.051	0.050	0.049
	(Gau,1)	0.049	0.045	0.048	0.047	0.050	0.051	0.051
	(Gau,Gau)	0.044	0.047	0.046	0.044	0.052	0.050	0.048
	Opt	0.050	0.044	0.049	0.047	0.055	0.053	0.049

Table 4.6: The type 1 error rates of the various methods under different censoring rates. D.1-4 are defined in Table 4.5. The covariates are 1-dimensional. There are no rejection rates larger than 0.057.

		% Observed						
D.	Method	15%	30%	45%	60%	75%	90%	100%
5	Cph	*0.063*	*0.061*	*0.068*	0.056	*0.062*	*0.058*	*0.062*
	(Lin,1)	0.046	0.043	0.056	0.047	0.048	0.048	0.049
	(Fis,1)	0.048	0.047	0.057	0.047	0.049	0.046	0.050
	(Gau,1)	0.044	0.044	0.056	0.046	0.046	0.048	0.050
	(Gau,Gau)	0.040	0.047	0.053	0.048	0.046	0.046	0.049
	Opt	0.053	0.045	*0.059*	0.048	0.048	0.049	0.054
6	Cph	*0.061*	*0.066*	*0.066*	*0.064*	*0.063*	*0.066*	*0.063*
	(Lin,1)	0.051	0.054	0.056	0.048	0.050	0.052	0.050
	(Fis,1)	0.049	0.050	0.051	0.049	0.053	0.052	0.051
	(Gau,1)	0.048	0.049	0.055	0.050	0.052	0.052	0.052
	(Gau,Gau)	0.050	0.048	0.057	0.048	0.050	0.048	0.053
	Opt	0.055	0.051	*0.058*	0.049	0.048	0.052	0.053
7	Cph	*0.068*	*0.068*	*0.067*	*0.064*	*0.064*	*0.064*	*0.063*
	(Lin,1)	0.048	0.054	0.055	0.048	0.049	0.050	0.050
	(Fis,1)	0.051	0.055	0.055	0.046	0.050	0.053	0.051
	(Gau,1)	0.047	0.049	0.053	0.047	0.050	0.050	0.052
	(Gau,Gau)	0.047	0.048	*0.059*	0.048	0.051	0.049	0.053
	Opt	0.047	0.049	0.053	0.054	0.051	0.051	0.053

Table 4.7: The type 1 error rate of the various methods under different censoring rates. D.5-7 are found in Table 4.5. The covariates are 10-dimensional. Rejection rates above 0.057 are displayed in italics.

4.9.1.5 Power under varying censoring rates

We now estimate the power when the alternative hypothesis holds. To estimate power we sample 1000 times from each distribution of Table 4.8 and count the number of rejections. In each case the sample size equals 200. Obtained rejection rates are displayed in Tables 4.9 and 4.10.

D.	$Z\|X$	$C\|X$	X
8	Exp(mean $= \exp(X/3)$)	Exp(mean $= \theta$)	Unif[-1,1]
9	Exp(mean $= \exp(X^2)$)	Exp(mean $= \theta \exp(X)$)	Unif[-1,1]
10	Weib(shape $= 1.5X + 3.25$)	Exp(mean $= \theta X^2$)	Unif[-1,1]
11	$\mathcal{N}$(mean $= 100 - X$, var $= 1.5X + 5.5$)	Exp(mean $= 82 + \theta$)	Unif[-1,1]
12	Exp(mean $= \exp(1^T X/30)$)	Exp(mean $= \theta$)	$\mathcal{N}_{10}(0, \text{cov} = \Sigma_{10})$
13	Exp(mean $= \exp(X_4/7)$)	Exp(mean $= \theta \exp(1^T X/30)$)	$\mathcal{N}_{10}(0, \text{cov} = \Sigma_{10})$
14	Exp(mean $= \exp(X_4^2/20)$)	Exp(mean $= \theta \exp(X_2^2)/20$)	$\mathcal{N}_{10}(0, \text{cov} = \Sigma_{10})$

Table 4.8: The parametrized distributions to test the power under different censoring rates. Here $\Sigma_{10} = MM^T$ where M is a 10×10 matrix of i.i.d. standard normal entries. The parameter θ is varies such that $15, 30, 45, 60, 75, 90, 100\%$ of the individuals are observed (i.e. $\Delta = 1$). The sample size is 200 in each case.

		% Observed						
D.	Method	15%	30%	45%	60%	75%	90%	100%
8	Cph	**0.16**	**0.30**	**0.44**	**0.56**	**0.67**	**0.72**	**0.76**
	(Lin,1)	**0.16**	0.28	**0.44**	0.55	0.66	0.71	**0.76**
	(Gau,1)	0.14	0.24	0.36	0.48	0.58	0.64	0.68
	(Gau,Gau)	0.12	0.20	0.31	0.40	0.48	0.54	0.57
	Opt	**0.16**	0.29	0.41	0.52	0.62	0.68	0.71
9	Cph	0.06	0.10	0.10	0.10	0.08	0.08	0.06
	(Lin,1)	0.09	0.13	0.14	0.14	0.12	0.11	0.09
	(Gau,1)	**0.22**	**0.41**	**0.56**	**0.68**	**0.79**	**0.85**	**0.90**
	(Gau,Gau)	0.16	0.31	0.42	0.54	0.66	0.77	0.84
	Opt	0.08	0.13	0.19	0.26	0.40	0.52	0.63
10	Cph	0.11	0.06	0.08	0.12	0.17	0.21	0.24
	(Lin,1)	0.10	0.05	0.06	0.09	0.14	0.19	0.22
	(Gau,1)	0.10	0.04	0.05	0.08	0.12	0.16	0.16
	(Gau,Gau)	**0.38**	**0.54**	**0.68**	**0.76**	**0.84**	0.88	**0.87**
	Opt	0.22	0.16	0.18	0.34	0.63	**0.89**	0.86
11	Cph	0.29	0.12	0.08	0.05	0.06	0.12	0.14
	(Lin,1)	0.27	0.11	0.06	0.04	0.06	0.10	0.12
	(Gau,1)	0.22	0.09	0.07	0.04	0.05	0.09	0.11
	(Gau,Gau)	0.42	**0.51**	**0.62**	**0.76**	**0.86**	0.94	0.95
	Opt	**0.54**	0.42	0.39	0.57	0.85	**0.98**	**0.99**

Table 4.9: The power of the various methods under different censoring rates. The Distributions 8-11 are found in Table 4.8. The covariates are 1-dimensional. The highest rejection rate for each scenario and percentage of observed events is in bold. If an elevated type 1 error rate was found for a method under the censoring distribution in question - see Table 4.6, then the rejection rate is displayed in italics.

		% Observed						
D.	Method	15%	30%	45%	60%	75%	90%	100%
12	Cph	***0.22***	*0.40*	***0.58***	*0.68*	*0.83*	*0.88*	*0.92*
	(Lin,1)	0.21	**0.44**	**0.58**	**0.70**	**0.84**	**0.90**	**0.94**
	(Fis,1)	0.17	0.34	0.53	0.64	0.79	0.86	0.90
	(Gau,1)	0.20	0.40	0.55	0.67	0.82	0.88	0.91
	(Gau,Gau)	0.16	0.30	0.46	0.56	0.73	0.78	0.83
	Opt	0.18	0.33	0.49	0.59	0.74	0.82	0.91
13	Cph	*0.34*	*0.60*	*0.81*	*0.92*	*0.96*	*0.98*	***1.00***
	(Lin,1)	**0.39**	**0.64**	**0.87**	**0.95**	**0.98**	**0.99**	**1.00**
	(Fis,1)	0.29	0.54	0.76	0.90	0.96	0.97	0.99
	(Gau,1)	0.33	0.59	0.81	0.92	0.97	0.97	0.99
	(Gau,Gau)	0.24	0.49	0.73	0.84	0.93	0.95	0.98
	Opt	0.30	0.53	0.74	0.86	0.95	0.96	0.99
14	Cph	*0.06*	*0.08*	*0.07*	*0.08*	*0.09*	*0.08*	*0.11*
	(Lin,1)	0.06	0.06	0.08	0.10	0.15	0.15	0.18
	(Fis,1)	0.04	0.07	0.06	0.07	0.08	0.08	0.10
	(Gau,1)	**0.08**	**0.14**	**0.28**	**0.48**	**0.76**	**0.91**	**0.96**
	(Gau,Gau)	0.07	0.11	0.18	0.29	0.52	0.65	0.75
	Opt	0.05	0.06	0.08	0.14	0.25	0.40	0.81

Table 4.10: The power of the various methods under different censoring rates. The Distributions 12-14 are found in Table 4.8. The covariates are 10-dimensional. The highest rejection rate for each scenario and percentage of observed events is in bold. If an elevated type 1 error rate was found for a method under the censoring distribution in question (see Table 4.7), then the rejection rate is displayed in italics.

4.9.1.6 Varying bandwidths of the Gaussian Kernel

We now study how the rejection rate of the kernel log-rank test with kernel $(\text{Gau}, 1)$ or (Gau, Gau) is affected by the choice of the bandwidth. Note that data is scaled so that each component has mean zero and unit variance. In this section $n = 200$ and 60 or 75% of the observations are observed. In Sections 4.9.1.7 and 4.9.1.8 we study the type 1 error rate and power of $(\text{Gau}, 1)$ respectively, and in Sections 4.9.1.9 and 4.9.1.10 we study the type 1 error rate and power of (Gau, Gau) respectively. For estimation of the type 1 error rate we sample 5000 times from each distribution, and for estimation of the power we sample 1000 times from each distribution.

4.9.1.7 Type 1 error rate of the kernel (Gau,1) for different bandwidths

Obtained rejection rates for the distributions in Table 4.5 are displayed in Table 4.11. We find that when the bandwidth is too small relative to the median heuristic, the type 1 error rate is inflated.

		$\sigma_X =$								
D.	$\Delta = 1$	σ_{med}	0.1	0.2	0.5	1	2	5	10	20
1	60%	0.049	0.044	0.044	0.051	0.050	0.047	0.050	0.048	0.045
2	60%	0.050	0.047	0.045	0.047	0.053	0.046	0.055	0.050	0.047
3	75%	0.050	0.051	0.049	0.043	0.049	0.046	0.056	0.048	0.053
4	75%	0.050	0.047	0.050	0.046	0.053	0.053	0.048	0.053	0.042
5	75%	0.048	*0.068*	*0.108*	0.049	0.043	0.050	0.050	0.046	0.048
6	60%	0.049	*0.079*	*0.154*	0.054	0.048	0.045	0.052	0.056	0.052
7	75%	0.046	*0.075*	*0.127*	0.046	0.043	0.052	0.050	0.051	0.050

Table 4.11: The type 1 error rate of the kernel log-rank test with kernel $(\text{Gau}, 1)$ with various bandwidths. Distributions can be found in Table 4.5. For distributions D.1-4 it holds that $\sigma_{\text{med}} = 0.72$ and for D.5-7 it holds that $\sigma_{\text{med}} = 2.96$. The third column lists the rejection rate when the sigma median heuristic is used as bandwidth. The remaining columns list rejection rates when the bandwidths $0.1 - 20$ are used. Rejection rates above 0.57 are displayed in italics.

4.9.1.8 Power of the kernel (Gau,1) for different bandwidths

		$\sigma_X =$								
D.	$\Delta = 1$	σ_{med}	0.1	0.2	0.5	1	2	5	10	20
8	60%	0.48	0.21	0.32	0.44	0.51	**0.54**	0.53	**0.54**	0.53
9	60%	0.68	0.50	0.62	**0.69**	0.60	0.29	0.15	0.14	0.15
10	75%	0.12	0.10	0.11	0.11	0.13	0.14	0.14	**0.16**	0.14
11	75%	0.05	**0.07**	0.06	0.07	0.05	0.06	0.05	0.05	0.05
12	75%	0.59	*0.09*	*0.17*	0.07	0.27	0.52	**0.66**	0.61	0.63
13	60%	0.85	*0.09*	*0.17*	0.09	0.42	0.79	0.86	**0.89**	0.87
14	75%	0.67	*0.03*	*0.07*	0.09	0.59	**0.75**	0.40	0.19	0.13

Table 4.12: The power of the kernel log-rank test with kernel $(\text{Gau}, 1)$ with various bandwidths. Distributions are found in Table 4.8. For distributions D.8-11 it holds that $\sigma_{\text{med}} = 0.72$ and for D.12-13 it holds that $\sigma_{\text{med}} = 2.96$. The third column lists the rejection rate when the median heuristic is used as bandwidth. The remaining columns list rejection rates when the bandwidths $0.1 - 20$ are used. The highest rejection rate is in boldface. In the event of an elevated type 1 error rate (above 0.057) in Table 4.11, the rejection rate is displayed in italics.

4.9.1.9 Type 1 error rate of the kernel (Gau, Gau) for different bandwidths

		$\sigma_X =$							
		0.72	0.10	0.20	0.50	1.00	2.00	5.00	10.00
$\sigma_Z =$	0.16	0.050	0.044	0.050	0.049	0.050	0.043	0.045	0.046
	0.10	0.051	0.045	0.050	0.051	0.045	0.044	0.049	0.052
	0.20	0.050	0.045	0.052	0.050	0.051	0.051	0.054	0.050
	0.50	0.045	0.049	0.053	0.051	0.052	0.044	0.045	0.047
	1.00	0.048	0.052	0.047	0.052	0.045	0.048	0.051	0.054
	2.00	0.045	0.045	0.046	0.048	0.052	0.049	0.046	0.052
	5.00	0.051	0.043	0.044	0.049	0.050	0.047	0.051	0.050
	10.00	0.046	0.049	0.049	0.046	0.049	0.047	0.050	0.045

Table 4.13: Type 1 error rate of the kernel log-rank test with various combinations of bandwidths for D.3 of Table 4.5 when 75% of events is observed. The median heuristic yields $\sigma_X = 0.72$ (first column) and $\sigma_Z = 0.16$ (first row) resulting in a rejection rate of 0.050. Each combination of bandwidths has the correct type 1 error rate of around $\alpha = 0.05$.

		$\sigma_X =$							
		0.72	0.10	0.20	0.50	1.00	2.00	5.00	10.00
$\sigma_Z =$	0.69	0.046	0.046	0.049	0.042	0.049	0.050	0.048	0.047
	0.10	0.047	0.039	0.044	0.047	0.052	0.043	0.045	0.052
	0.20	0.050	0.038	0.045	0.049	0.049	0.055	0.051	0.049
	0.00	0.044	0.047	0.047	0.050	0.049	0.050	0.049	0.048
	1.00	0.052	0.046	0.043	0.050	0.047	0.050	0.052	0.050
	2.00	0.050	0.046	0.050	0.052	0.047	0.052	0.050	0.054
	5.00	0.042	0.046	0.044	0.046	0.045	0.048	0.047	0.051
	10.000	0.049	0.045	0.048	0.045	0.047	0.051	0.052	0.057

Table 4.14: Type 1 error rate of the kernel log-rank test with various combinations of bandwidths for D.4 of Table 4.5 when 75% of events are observed. The median heuristic yields $\sigma_X = 0.72$ (first column) and $\sigma_Z = 0.69$ (first row) resulting in a rejection rate of 0.050. Each combination of bandwidths has the correct type 1 error rate of around $\alpha = 0.05$ - none are above 0.057.

		$\sigma_X =$							
		2.94	0.10	0.20	0.50	1.00	2.00	5.00	10.00
$\sigma_Z =$	0.49	0.047	*0.097*	*0.183*	0.047	0.044	0.046	0.046	0.047
	0.10	0.048	*0.120*	*0.206*	0.018	0.034	0.054	0.051	0.047
	0.20	0.048	*0.108*	*0.192*	0.028	0.038	0.045	0.053	0.046
	0.50	0.047	*0.102*	*0.172*	0.046	0.043	0.046	0.049	0.052
	1.00	0.045	*0.085*	*0.150*	0.049	0.046	0.051	0.052	0.049
	2.00	0.049	*0.078*	*0.131*	0.050	0.044	0.049	0.050	0.052
	5.00	0.052	*0.070*	*0.129*	0.053	0.047	0.055	0.053	0.053
	10.00	0.048	*0.079*	*0.120*	0.047	0.049	0.046	0.046	0.053

Table 4.15: Type 1 error rate of the kernel log-rank test with various combinations of bandwidths for D.7 of Table 4.5 when 75% of events are observed. The median heuristic yields $\sigma_X = 2.94$ (first column) and $\sigma_Z = 0.49$ (first row) resulting in a rejection rate of 0.053. We note that bandwidths σ_X, which are small compared to the median heuristic of 2.94, lead to an inflated type 1 error rate. Type 1 error rates above 0.057 are displayed in italics.

4.9.1.10 Power of the kernel (Gau,Gau) for different bandwidths

		$\sigma_X =$							
		0.72	0.10	0.20	0.50	1.00	2.00	5.00	10.00
$\sigma_Z =$	0.69	0.86	0.44	0.65	0.80	0.86	0.90	0.89	0.89
	0.10	0.64	0.27	0.39	0.62	0.68	0.71	0.70	0.73
	0.20	0.77	0.37	0.55	0.74	0.81	0.84	0.83	0.82
	0.50	0.86	0.48	0.67	0.84	0.88	0.88	**0.91**	**0.91**
	1.00	0.76	0.37	0.50	0.70	0.81	0.83	0.82	0.82
	2.00	0.29	0.15	0.22	0.25	0.31	0.35	0.37	0.40
	5.00	0.15	0.10	0.10	0.13	0.13	0.17	0.16	0.18
	10.00	0.11	0.09	0.10	0.11	0.12	0.14	0.15	0.15

Table 4.16: Power of the kernel log-rank test with various combinations of bandwidths for D.10 of Table 4.8 when 75% of events are observed. The median heuristic yields $\sigma_X = 0.72$ (first column) and $\sigma_Z = 0.69$ (first row) resulting in a rejection rate of 0.86. The highest rejection rate 0.91 (in bold) is achieved when $(\sigma_Z, \sigma_X) = (0.5, 5), (0.5, 10)$. In Table 4.13 we found that for this censoring distribution, each of the bandwidths led to a correct Type 1 error rate.

		$\sigma_X =$							
		0.72	0.10	0.20	0.50	1.00	2.00	5.00	10.00
$\sigma_Z =$	0.67	0.87	0.43	0.62	0.82	0.91	**0.95**	**0.95**	0.92
	0.10	0.68	0.30	0.41	0.64	0.74	0.76	0.79	0.78
	0.20	0.82	0.43	0.58	0.77	0.87	0.88	0.90	0.88
	0.50	0.90	0.47	0.69	0.85	0.92	0.94	**0.95**	0.94
	1.00	0.77	0.31	0.48	0.70	0.82	0.89	0.88	0.88
	2.00	0.29	0.11	0.14	0.24	0.34	0.39	0.36	0.37
	5.00	0.09	0.07	0.08	0.07	0.09	0.07	0.10	0.09
	10.00	0.06	0.06	0.07	0.06	0.06	0.08	0.07	0.07

Table 4.17: Power of the kernel log-rank test with various combinations of bandwidths for D.11 of Table 4.8 when 75% of events are observed. The median heuristic yields $\sigma_X = 0.72$ (first column) and $\sigma_Z = 0.67$ (first row) resulting in a rejection rate of 0.87. The highest rejection rate 0.94 (in bold) is achieved when $(\sigma_Z, \sigma_X) = (0.67, 2), (0.67, 5), (0.5, 5)$. In Table 4.14 we found that for this censoring distribution, each of the bandwidths led to a correct type 1 error rate.

		$\sigma_X =$							
		2.96	0.10	0.20	0.50	1.00	2.00	5.00	10.00
$\sigma_Z =$	0.32	0.48	*0.07*	*0.12*	0.08	0.42	0.57	0.26	0.13
	0.10	0.34	*0.08*	*0.15*	0.03	0.28	0.41	0.20	0.10
	0.20	0.43	*0.08*	*0.11*	0.06	0.42	0.49	0.24	0.13
	0.50	0.55	*0.07*	*0.09*	0.06	0.50	0.64	0.31	0.14
	1.00	0.62	*0.07*	*0.10*	0.10	0.55	0.70	0.35	0.17
	2.00	0.63	*0.04*	*0.07*	0.08	0.58	0.73	0.37	0.19
	5.00	0.67	*0.05*	*0.07*	0.08	0.60	0.75	0.43	0.17
	10.00	0.67	*0.05*	*0.07*	0.09	0.60	**0.76**	0.41	0.19

Table 4.18: Power of the kernel log-rank test with various combinations of bandwidths for D.14 4.8 when 75% of events are observed. The median heuristic yields $\sigma_X = 2.96$ (first column) and $\sigma_Z = 0.32$ (first row) corresponding to a rejection rate of 0.48. The highest rejection rate 0.76 is achieved when $(\sigma_Z, \sigma_X) = (10, 2)$. Rejection rates for bandwidths that led to an increased type 1 error rate in Table 4.15, which featured the same censoring distribution, are displayed in italics.

4.9.2 Preliminary results

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

4.9.2.1 Counting processes

We state some results from the counting process literature that are frequently used in our paper. Recall that $Y(t) = \sum_{i=1}^{n} Y_i(t)$ denotes the pooled risk function, where $Y_i(t) = \mathbb{1}_{\{T_i \geq t\}}$, $\tau = \sup\{t : S_T(t) > 0\}$, $\tau_n = \max\{T_1, \ldots, T_n\}$, and that Z and C are continuous random variables.

Proposition 4.18. *The following holds a.s.*

1. $\lim_{n\to\infty} \sup_{t\leq\tau} |Y(t)/n - S_T(t)| = 0$
2. $\sup_{t\leq t^\star} \left| \frac{n}{Y(t)} - \frac{1}{S_T(t)} \right| \to 0$. *for any* $t^\star \in (0, \tau)$

The proof of Part 1 is due to the Glivenko Cantelli's Theorem, and Part 2 is a consequence of Part 1. Also, notice that under the null hypothesis $S_T(t) = S_Z(t) \int_{\mathbb{R}^d} S_{C|X=x}(t) dF_X(x)$

since we assume non-informative censoring, i.e. $Z \perp C|X$.

Proposition 4.19. *Let $\beta \in (0,1)$, then*

1. $\mathbb{P}\left(Y(t)/n \leq \beta^{-1} S_T(t), \quad \forall t \leq \tau_n\right) \geq 1-\beta$,

2. $\mathbb{P}\left(Y(t)/n \geq \beta S_T(t), \quad \forall t \leq \tau_n\right) \geq 1 - e(1/\beta)e^{-1/\beta}$.

The proof of Part 1 follows from (Gill 1980; Theorem 3.2.1.) and Part 2 is due to Gill (1983).

4.9.2.2 Dominated convergence in probability

Lemma 4.20. *Let $(\mathcal{X}, \mathcal{B}, \mu)$ be a measurable space and $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Consider a sequence of non-negative stochastic processes $R_n : \Omega \times \mathcal{X} \to \mathbb{R}$. Suppose that*

i) For each $\alpha \in (0,1)$, there exists an event A_α with $\mathbb{P}(A_\alpha) \geq 1-\alpha$, such that $R_n(\omega, x) \to 0$ for all $(\omega, x) \in A_\alpha \times \mathcal{X}$.

ii) For each $\beta \in (0,1)$, there exists a non-negative function $R_\beta \in L_1(\mathcal{X}, \mathcal{B}, \mu)$ and N_0 large enough such that for each $n \geq N_0$ there exists an event $B_{n,\beta}$ with $\mathbb{P}(B_{n,\beta}) \geq 1-\beta$ and

$$R_n(\omega, x) \leq R_\beta(x)$$

for $(\omega, x) \in B_{n,\beta} \times \mathcal{X}$.

Then $I_n(\omega) = \int_{\mathcal{X}} R_n(\omega, x)\mu(dx) \to 0$ in $\mathbb{P}$-probability.

The proof follows by noticing that in a set of probability tending to one, we can apply dominated convergence.

4.9.2.3 Lenglart-Rebolledo Inequality

Let $X(t)$ and $X'(t)$ be right-continuous adapted processes. We say that X is majorised by X' if for all bounded stopping times T, it holds that $\mathbb{E}(|X(T)|) \leq \mathbb{E}(X'(T))$.

Lemma 4.21 (Theorem 3.4.1. of Fleming and Harrington (1991))**.** *Let X be a right-continuous adapted process, and X' a non-decreasing predictable process with $X'(0) = 0$ such that X is majorised by X'. For any stopping time T (even unbounded), and any $\varepsilon, \delta > 0$,*

$$\mathbb{P}\left(\sup_{t \leq T} |X(t)| > \varepsilon\right) \leq \frac{\delta}{\varepsilon} + \mathbb{P}(X'(T) \geq \delta).$$

In our setting $X(t)$ is a sub-martingale $W_n(t)$ (based on n data points) and $X'(t)$ is the corresponding compensator $A_n(t)$. We consider the stopping time $\tau_n = \max\{T_1, \ldots, T_n\}$, and apply the previous Lemma to prove

$$A_n(\tau_n) = o_p(1) \Rightarrow \sup_{t \leq \tau_n} W_n(t) = o_p(1). \tag{4.17}$$

4.9.2.4 Double martingale integrals

Next we state results regarding double integrals with respect to a $\mathcal{F}_t$- counting process martingale $M(t)$, introduced in Fernández and Rivera (2020). Consider $W(t) = \int_{C_t} h(x,y) dM(x) dM(y)$, where $C_t = \{(x,y) : 0 < x < y \leq t\}$. The following results state conditions under which $W(t)$ defines a proper martingale with respect the filtration $(\mathcal{F}_t)_{t \geq 0}$.

Definition 4.22. *We define the predictable σ- algebra $\mathcal{P}$ as the σ-algebra generated by the sets of the form*

$$\sigma\{\{(a_1, b_1] \times (a_2, b_2] \times X : 0 \leq a_1 \leq b_1 < a_2 \leq b_2, X \in \mathcal{F}_{a_2}\} \cup \{(0,0) \times X : X \in \mathcal{F}_0\}\}.$$

Let $C = \{(x, y) : 0 < x < y < \infty\}$. A function $h : C \times \Omega \to \mathbb{R}$ is called *"elementary predictable"* if it can be written as a finite sum of indicator functions of sets belonging to the predictable σ-algebra $\mathcal{P}$. On the other hand, if a function h is $\mathcal{P}$-measurable, then it is the limit of elementary predictable functions.

Proposition 4.23. *Any deterministic measurable function is $\mathcal{P}$-measurable. Additionally the functions $C \times \Omega \to \mathbb{R}$: $Y(x)Y(y)$ and $\mathbb{1}_{\{x<y\leq\tau_n\}}$ are $\mathcal{P}$-measurable.*

Remark. Note that any deterministic function of the covariates $X_1, \ldots, X_n$, individual risk functions $Y_i(t)$, and indicator function $\mathbb{1}_{\{t\leq\tau_n\}}$ are $\mathcal{P}$-measurable as in Definition 4.22.

Theorem 4.24. *Let $h(x, y)$ be a $\mathcal{P}$-measurable function (see Definition 4.22) and define $C_t = \{(x, y) : 0 < x < y \leq t\}$ with $t \geq 0$. Let $M(t)$ and $M'(t)$ be $\mathcal{F}_t$-martingales, and suppose that for all $t \geq 0$, it holds*

$$\mathbb{E}\left(\int_{C_t} |h||dM(x)dM'(y)|\right) < \infty. \tag{4.18}$$

Then, the process

$$W(t) = \int_{C_t} h(x, y)dM(x)dM'(y)$$

is a martingale on $t \geq 0$ with respect the filtration $\mathcal{F}_t$.

4.9.2.5 Forward and Backward operators

Consider the forward and backward operators introduced in Efron and Johnstone (1990), $A : \mathcal{L}_2(Z\times C\times X) \to \mathcal{L}_2(Z\times C\times X)$, and $B : \mathcal{L}_2(Z\times C\times X) \to \mathcal{L}_2(Z\times C\times X)$, respectively, defined as

$$(Af)(z, c, x) = f(z, c, x) - \frac{1}{S_Z(z)} \int_z^\infty f(s, c, x)dF_Z(s)$$

and

$$(Bf)(z,c,x) = f(z,c,x) - \int_0^z f(s,c,x)d\Lambda_Z(s).$$

It was proved in Efron and Johnstone (1990) that the previous operators satisfy the following properties. For any $f, g \in \mathcal{L}_2(Z \times C \times X)$:

1. $\mathbb{E}((Af)g|C_1, X_1) = \mathbb{E}(f(Bg)|C_1, X_1)$
2. $ABf = f$
3. $BAf = f - \mathbb{E}(f|C_1, X_1)$
4. $\mathbb{E}(Bf|C_1, X_1) = 0$
5. For any $f(z,c,x) = \mathbb{1}_{\{z \leq c\}} g(z,c,x)$,

$$(Bf)(z,c,x) = \int_{\mathbb{R}_+} g(s,c,x)dm_{z,c}(s),$$

where $dm_{z,c}(s) = \mathbb{1}_{z \leq c}\delta_z(s) - \mathbb{1}_{\{\min\{z,c\} \geq s\}} d\Lambda_Z(s)$.

4.9.2.6 Kernel results

We state the definition of a characteristic kernel, which is defined in Section 2 in Sriperumbudur et al. (2011).

Definition 4.25 (characteristic kernel)**.** *Given the set of all Borel probability measures defined on the topological space $\mathcal{X}$, a measurable and bounded kernel K is said to be characteristic if*

$$\mathbb{P} \to \int_{\mathcal{X}} K(\cdot, x)d\mathbb{P}(x)$$

is injective, that is, the probability measure $\mathbb{P}$ is embedded to a unique element $\int_{\mathcal{X}} K(\cdot, x)d\mathbb{P}(x) \in \mathcal{F}$, where $\mathcal{F}$ is the RKHS induced by the kernel K.

An example of a characteristic kernel is the exponentiated quadratic kernel, which is defined as $K(x, y) = \exp\{-(x-y)^\intercal(x-y)/\sigma^2\}$ for any $x, y \in \mathbb{R}^d$ and $\sigma > 0$.
We state the definition of c_0-universality, which is introduced in Proposition 2 in Sriperumbudur et al. (2011).

Proposition 4.26. *(c_0-universality and RKHS embedding of measures) Suppose that $\mathcal{X}$ is a locally compact Hausdorff space and the kernel K is a c_0-kernel. Then K is c_0- universal if and only if the embedding*

$$\mu \to \int_{\mathcal{X}} K(\cdot, x) d\mu(x), \quad \mu \in M_b(\mathcal{X})$$

is injective, where $M_b(\mathcal{X})$ denotes the space of all finite signed Radon measures on $\mathcal{X}$.

Remark. We say that K is a c_0-kernel if K is such that $K(\cdot, x) \in \mathcal{C}_0$ for any $x \in \mathcal{X}$, where $C_0(\mathcal{X})$ denotes the class of continuous functions on $\mathcal{X}$ that vanish at infinity.
Finally, we state the result which relates translation invariant kernels with the notion of c_0-universality. This result is presented in Section 3.2 in Sriperumbudur et al. (2011)

Proposition 4.27. *Bounded, continuous translation invariant, c_0-kernels are c_0-universal if and only if they are characteristic.*

Note that by combining the two previous results, under the same assumptions, we obtain that characteristic, translation invariant c_0-kernels provide injective embeddings of finite signed Radon measures.

4.9.2.7 Wild-Bootstrap for degenerate V-statistics

The following result appears in Theorem 3.1 in Dehling and Mikosch (1994).

Theorem 4.28. *Let $X_1, \ldots, X_n$ be a collection of i.i.d. random variables, and let $W_1, \ldots, W_n$ be a collection of i.i.d. Rademacher random variables which are independent of $X_1, \ldots, X_n$. Suppose h is a symmetric degenerate kernel satisfying*

$\mathbb{E}(|h(X_1, X_1)|) < \infty$ *and* $\mathbb{E}(h^2(X_1, X_2)) < \infty$. *Then there exists a probability distribution* $\mathcal{L}$ *such that*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h(X_i, X_j) \xrightarrow{\mathcal{D}} \mathcal{L},$$

and for almost every sequence $(x_i)_{i\geq 0}$ *sampled from* $(X_i)_{i\geq 0}$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h(x_i, x_j)W_iW_j \xrightarrow{\mathcal{D}} \mathcal{L},$$

as n *grows to infinity.*

4.9.3 Auxiliary results

In this section we introduce some auxiliary results that will be used in the proofs of our main results. The proofs of these results are given in Section 4.9.5.

- Proposition 4.29 is a technical result from which we deduce $\tau_x \leq \tau$ for almost all $x \in \mathbb{R}^d$.
- Proposition 4.30 is used in the proof of Theorem 4.1.
- Theorem 4.31 is used in the proof of Theorem 4.9
- Lemma 4.32 is used in the proof of Lemma 4.10.

Proposition 4.29. *Let τ' be the essential supremum defined as $\tau' = ess\ \sup \tau_x = \inf\{t : F_X(\{x : \tau_x > t\}) = 0\}$. Then $\tau = \tau'$.*

Proposition 4.30. *Let $\omega : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ be a bounded measurable function. Then*

$$\int_{\mathbb{R}^d} \int_0^\tau \omega(s,x) d\nu_0^n(s,x) = \frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n \Delta_j \omega(T_j, X_i) \frac{Y_i(T_j)}{S_T(T_j)} + o_p(1).$$

Theorem 4.31. *Let $Q : (\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^d)^2 \to \mathbb{R}$, $T^Q : \mathcal{L}_2(Z \times C \times X) \to \mathcal{L}_2(Z \times C \times X)$ and $T^J : \mathcal{L}_2(T \times \Delta \times X) \to \mathcal{L}_2(T \times \Delta \times X)$ be given by*

$$Q((z_1, c_1, x_1), (z_2, c_2, x_2)) = \mathbb{1}_{\{z_1 \leq c_1\}} \mathbb{1}_{\{z_2 \leq c_2\}} \bar{\mathfrak{K}}((z_1, x_1), (z_2, x_2)),$$

$$(T^Q f)(Z_1, C_1, X_1) = \mathbb{E}_2 \left(Q((Z_1, C_1, X_1), (Z_2, C_2, X_2)) f(Z_2, C_2, X_2) \right),$$

and

$$(T^J f)(T_1, \Delta_1, X_1) = \mathbb{E}_2 \left(J((T_1, \Delta_1, X_1), (T_2, \Delta_2, X_2)) f(T_2, \Delta_2, X_2) \right), \tag{4.19}$$

respectively, where J is the function defined in Equation (4.12) *in the main document,* $\mathbb{E}_2(\cdot) = \mathbb{E}(\cdot|(T_1, \Delta_1, X_1))$*, and* (Z_1, C_1, X_1) *and* (Z_2, C_2, X_2) *are independent copies of our random variables in the augmented space.*

Then, $T^J = BT^Q A$*, and* T^J *and* T^Q *have the same set of non-zero eigenvalues, including multiplicities.*

Lemma 4.32.

$$\frac{1}{n}\int_0^\tau \left|1 - \frac{Y(t)/n}{S_T(t)}\right|^k dN(t) = o_p(1), \tag{4.20}$$

for $k = 1$ *and* $k = 2$*.*

4.9.4 Main proofs

In this section we prove the main results of our paper.

4.9.4.1 Proofs of Section 4.3

4.9.4.2 Proof of Theorem 3.1

The result follows from proving

$$\int_{\mathbb{R}^d}\int_0^\tau \omega(s,x)d\nu_k^n(s,x) \xrightarrow{\mathbb{P}} \int_{\mathbb{R}^d}\int_0^\tau \omega(s,x)d\nu_k(t,x)$$

for $k \in \{0,1\}$.

Note that the result for $k=1$ follows from a direct application of the law of large numbers. We continue proving the result for $k=0$. By Proposition 4.30, it holds

$$\int_{\mathbb{R}^d}\int_0^\tau \omega(s,x)d\nu_0^n(s,x) = \frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n \Delta_j\omega(T_j,X_i)\frac{Y_i(T_j)}{S_T(T_j)} + o_p(1).$$

The sum in the left-hand-side of the previous equation can be decomposed into two sums, one over indices such that $\{i \neq j\}$ and the other over indices such that $\{i=j\}$. Notice that the sum over $\{i \neq j\}$ satisfies

$$\frac{1}{n^2}\sum_{i\neq j}^n \Delta_j\omega(T_j,X_i)\frac{Y_i(T_j)}{S_T(T_j)} \xrightarrow{\mathbb{P}} \mathbb{E}\left(\Delta_2\omega(T_2,X_1)\frac{Y_1(T_2)}{S_T(T_2)}\right) = \int_{\mathbb{R}^d}\int_0^\tau \omega(s,x)d\nu_0(s,x)$$

by the law of large numbers for U-statistics. Thus, to conclude the proof, we just need to prove that the sum over indices $\{i=j\}$ converges to zero in probability. To see this, notice that $U_i = S_T(T_i)$ are i.i.d. uniform random variables, and that

$$\frac{1}{n^2}\sum_{i=1}^n \Delta_i\omega(T_i,X_i)\frac{1}{S_T(T_i)} \leq \frac{C}{n^2}\sum_{i=1}^n \frac{1}{S_T(T_i)} \leq \frac{C}{n^2}\sum_{i=1}^n \frac{1}{U_i} = o_p(1),$$

where the first inequality follows from the assumption that ω is bounded by some constant C, and the last equality follows from $n^{-2}\sum_{i=1}^{n} U_i^{-1} = o_p(1)$, which we now proceed to prove.

Define the event $\mathcal{B}_n = \cup_{i=1}^{n} \left\{1/U_i \geq n^{3/2}\right\}$, and note that $\mathbb{P}(\mathcal{B}_n) \leq n\mathbb{P}(U_i \leq 1/n^{3/2}) = n^{-1/2}$. Then, for any $\epsilon > 0$, it holds that

$$\begin{aligned}
\mathbb{P}\left(\frac{C}{n^2}\sum_{i=1}^{n}\frac{1}{U_i} > \epsilon\right) &= \mathbb{P}\left(\left\{\frac{C}{n^2}\sum_{i=1}^{n}\frac{1}{U_i} > \epsilon\right\} \cap \mathcal{B}_n\right) + \mathbb{P}\left(\left\{\frac{C}{n^2}\sum_{i=1}^{n}\frac{1}{U_i} > \epsilon\right\} \cap \mathcal{B}_n^c\right) \\
&\leq \mathbb{P}\left(\mathcal{B}_n\right) + \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}_{\{1/U_i < n^{3/2}\}}}{U_i} > \frac{n\epsilon}{C}\right) \\
&\leq \frac{1}{n^{1/2}} + \frac{C}{n\epsilon}\mathbb{E}\left(\frac{\mathbb{1}_{\{1/U_i < n^{3/2}\}}}{U_i}\right) \\
&\leq \frac{1}{n^{1/2}} + \frac{C}{n\epsilon}\int_{n^{-3/2}}^{1}\frac{1}{u}du \\
&\leq \frac{1}{n^{1/2}} + \frac{3C}{2\epsilon}\frac{\log(n)}{n} \to 0,
\end{aligned}$$

for any $\epsilon > 0$ as n grows to infinity.

4.9.4.3 Proof of Proposition 4.3

($\Longleftarrow$) Assume $Z \perp X$. Note that $\tau = \tau_x$ for almost all $x \in \mathbb{R}^d$, since if this did not hold, then $Z \not\perp X$. Consider arbitrary $t \leq \tau$ and measurable $A \subseteq \mathbb{R}^d$, then it holds that

$$\nu_1((0,t] \times A) = \int_0^t \int_{x\in A} S_{C|X=x}(s) dF_Z(s) dF_X(x),$$

and

$$\begin{aligned}
\nu_0((0,t] \times A) &= \int_0^t \int_{x\in\mathbb{R}^d} S_{C|X=x}(s) \frac{\int_{x'\in A} S_Z(s) S_{C|X=x'}(s) dF_X(x')}{\int_{x''\in\mathbb{R}^d} S_Z(s) S_{C|X=x''}(s) dF_X(x'')} dF_Z(s) dF_X(x) \\
(Z \perp X) &= \int_0^t \int_{x\in A} S_{C|X=x}(s) dF_Z(s) dF_X(x).
\end{aligned}$$

Because ν_1 and ν_0 agree on generating sets, they agree on any measurable set in $\mathbb{R}_+ \times \mathbb{R}^d$.

($\Longrightarrow$) We first prove that $\nu_1 = \nu_0$ implies that $\tau = \tau_x$ for almost all $x \in \mathbb{R}^d$. For this result, we observe that Assumption 4.2 implies $\tau_x = \sup\{t : S_{T|X=x}(t) > 0\} = \sup\{t : S_{Z|X=x}(t) > 0\}$, and by Proposition 4.29, $\tau_x \leq \tau$ for almost all $x \in \mathbb{R}^d$. We now proceed by contradiction. Assume there exists $\tau' < \tau$ such that the set $E = \{x \in \mathbb{R}^d : \tau_x \leq \tau'\}$ satisfies $\mathbb{P}(X \in E) > 0$, otherwise $\tau_x = \tau$ for almost all $x \in \mathbb{R}^d$. Let $\omega(s,x) = \mathbb{1}_{\{s<\tau_x\}} S_{T|X=x}(s)^{-1}$, and note that $\omega(s,x) < \infty$ for all $(s,x) \in (0,\tau) \times \mathbb{R}^d$. By the hypothesis $\nu_0 = \nu_1$,

$$\int_A \int_0^t \omega(s,x) S_{C|X=x}(s) dF_{ZX}(s,x) = \int_A \int_0^t \omega(s,x) S_{T|X=x}(s) d\alpha(s) dF_X(x) \tag{4.21}$$

for any $t \leq \tau$, measurable set $A \subseteq \mathbb{R}^d$, and measurable function $\omega(s,x)$. This leads to a contradiction, however. Indeed, on one hand, the left-hand-side of equation (4.21), when evaluated at $\omega(s,x) = \mathbb{1}_{\{s<\tau_x\}} S_{T|X=x}(s)^{-1}$, $A = E$ and $t = \tau'$ satisfies

$$\begin{aligned}
\int_E \int_0^{\tau'} \omega(s,x) S_{C|X=x}(s) dF_{ZX}(s,x) &= \int_E \int_0^{\tau'} \mathbb{1}_{\{s<\tau_x\}} d\Lambda_{Z|X=x}(s) dF_X(x) \\
(\tau_x < \tau' \text{ in } E) &= \int_E \Lambda_{Z|X=x}(\tau_x) dF_X(x) \\
&= \infty,
\end{aligned}$$

since $\mathbb{P}(X \in E) > 0$ and the cumulative hazard function $\Lambda_{Z|X=x}(t)$ diverges as t tends to τ_x.

On the other hand, the right-hand-side of equation (4.21), when evaluated at $\omega(s,x) = \mathbb{1}_{\{s<\tau_x\}} S_{T|X=x}(s)^{-1}$, $A = E$ and $t = \tau'$, satisfies

$$\begin{aligned}
\int_E \int_0^{\tau'} \omega(s,x) S_{T|X=x}(s) d\alpha(s) dF_X(x) &= \int_E \int_0^{\tau'} \mathbb{1}_{\{s\leq\tau_x\}} d\alpha(s) dF_X(x) \\
&\leq \mathbb{P}(X \in E) \alpha((0,\tau')) < \infty,
\end{aligned}$$

since

$$\alpha((0,\tau')) = \int_0^{\tau'} \frac{\int_{x\in\mathbb{R}^d} S_{C|X=x}(s)dF_{ZX}(s,x)}{S_T(s)} \leq \frac{1}{S_T(\tau')} < \infty,$$

due to the fact that $S_T(t) > 0$ for all $t < \tau$ and $\tau' < \tau$. This finally leads to a contradiction, meaning that such a τ' cannot exist, and thus $\tau_x = \tau$ for almost all $x \in \mathbb{R}^d$.

We proceed to prove that $\nu_0 = \nu_1$ implies $Z \perp X$. Let $\omega(s,x) = S_{T|X=x}(s)^{-1}$, let $A \subseteq \mathbb{R}^d$ be a measurable set, and let $t < \tau$. Note that the function ω is well-defined F_X-almost surely for $t < \tau$ as $\mathbb{P}(X \in \{x : \mathbb{R}^d : \tau_x \neq \tau\}) = 0$. By substituting ω in equation (4.21), we obtain

$$\int_0^t \int_A \frac{S_{C|X=x}(s)}{S_{T|X=x}(s)} dF_{ZX}(s,x) = \int_0^t \int_A d\alpha(s) dF_X(x), \tag{4.22}$$

and by further choosing $A = \mathbb{R}^d$, we obtain

$$\int_{\mathbb{R}^d} \int_0^t \frac{dF_{Z|X=x}(s)}{S_{Z|X=x}(s)} dF_X(x) = \int_{\mathbb{R}^d} \Lambda_{Z|X=x}(t) dF_X(x) = \Lambda_Z(t) = \int_0^t d\alpha(t),$$

which implies $d\alpha(s) = d\Lambda_Z(s)$ for almost all $s \in \mathbb{R}_+$. Substituting the previous equality in equation (4.22), we obtain for arbitrary measurable A,

$$\int_0^t \int_A \frac{S_{C|X=x}(s)}{S_{T|X=x}(s)} dF_{ZX}(s,x) = \int_A \Lambda_{Z|X=x}(t) dF_X(x) = \int_A \Lambda_Z(t) dF_X(x),$$

from which we conclude that $d\Lambda_{Z|X}(s) = d\Lambda_Z(s)$ F_X-almost surely for all $s < \tau$. Note that Assumption 4.2 implies that $\tau = \sup\{t : S_Z(t) > 0\}$ from which we deduce $Z \perp X$.

4.9.4.4 Proof of Theorem 4.4

Let $\omega^\star = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^{-1}(\phi_0^n - \phi_1^n)$, clearly $\|\omega^\star\|_{\mathcal{H}}^2 = 1$ and

$$\langle \omega^\star, \phi_0^n - \phi_1^n \rangle_{\mathcal{H}} = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}} \leq \Psi_n.$$

On the other hand

$$\Psi_n = \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}^2 \leq 1} \langle \omega, \phi_0^n - \phi_1^n \rangle_{\mathcal{H}} \leq \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}^2 \leq 1} \|\omega\|_{\mathcal{H}} \|\phi_0^n - \phi_1^n\|_{\mathcal{H}} = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}},$$

from which we deduce the desired result $\Psi_n = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}}$.

Observe

$$\begin{aligned}
\|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 &= \left\| \int_{\mathbb{R}^d} \int_{\mathbb{R}_+} L(t,\cdot) K(x,\cdot) (d\nu_1^n(t,x) - d\nu_0^n(t,x)) \right\|_{\mathcal{H}}^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}_+} L(t,\cdot) \left(K(X_i,\cdot) - \sum_{j=1}^n K(X_j,\cdot) \frac{Y_j(t)}{Y(t)} \right) dN_i(t) \right\|_{\mathcal{H}}^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \Delta_i L(T_i,\cdot) \left(K(X_i,\cdot) - \sum_{j=1}^n K(X_j,\cdot) \frac{Y_j(T_i)}{Y(T_i)} \right) \right\|_{\mathcal{H}}^2 .
\end{aligned}$$

Define $\boldsymbol{\Phi^X} = (K(X_1,\cdot), \ldots, K(X_n,\cdot))^\intercal$ and $\boldsymbol{\Phi^T} = (\Delta_1 L(T_1,\cdot), \ldots, \Delta_n L(T_n,\cdot))^\intercal$. Then

$$\begin{aligned}
\|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 &= \left\| \frac{1}{n} \sum_{i=1}^n \Phi_i^T ((\boldsymbol{I} - \boldsymbol{A}) \boldsymbol{\Phi^X})_i \right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n \left\langle \Phi_i^T ((\boldsymbol{I} - \boldsymbol{A}) \boldsymbol{\Phi^X})_i, \Phi_{i'}^T ((\boldsymbol{I} - \boldsymbol{A}) \boldsymbol{\Phi^X})_{i'} \right\rangle_{\mathcal{H}} \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n (\boldsymbol{L^\Delta})_{i,i'} ((\boldsymbol{I} - \boldsymbol{A}) \boldsymbol{K} (\boldsymbol{I} - \boldsymbol{A})^\intercal)_{i,i'} \\
&= \frac{1}{n^2} \operatorname{Tr}(\boldsymbol{L^\Delta} (\boldsymbol{I} - \boldsymbol{A}) \boldsymbol{K} (\boldsymbol{I} - \boldsymbol{A})^\intercal).
\end{aligned}$$

4.9.4.5 Proofs of Section 4.4

4.9.4.6 Proof of Proposition 4.5

Observe that

$$
\begin{aligned}
(\phi_0^n - \phi_1^n)(\cdot) &= \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \mathfrak{K}((t,x),\cdot)(d\nu_1^n(t,x) - d\nu_0^n(t,x)) \\
&= \sum_{i=1}^n \int_{\mathbb{R}_+} \left(\mathfrak{K}((t,X_i),\cdot) - \sum_{j=1}^n \mathfrak{K}((t,X_j),\cdot) \frac{Y_j(t)}{Y(t)} \right) dN_i(t)
\end{aligned}
$$

follows from the definition of the embeddings, ϕ_0^n and ϕ_1^n, and of the empirical measures, ν_1^n and ν_0^n.

Assume that the process N_i can be replaced by the martingale M_i in the previous equation, that is,

$$
(\phi_0^n - \phi_1^n)(\cdot) = \sum_{i=1}^n \int_{\mathbb{R}_+} \left(\mathfrak{K}((t,X_i),\cdot) - \sum_{j=1}^n \mathfrak{K}((t,X_j),\cdot) \frac{Y_j(t)}{Y(t)} \right) dM_i(t). \tag{4.23}
$$

Then, using Theorem 4.4 and Equation (4.23), it holds that

$$
\Psi_n^2 = \|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 = \left\| \sum_{i=1}^n \int_{\mathbb{R}_+} \left(\mathfrak{K}((t,X_i),\cdot) - \sum_{j=1}^n \mathfrak{K}((t,X_j),\cdot) \frac{Y_j(t)}{Y(t)} \right) dM_i(t) \right\|_{\mathcal{H}}^2,
$$

where the main result is deduced by computing the previous term using the reproducing property.

To end the proof, we just need to prove Equation (4.23). Since $dM_i(t) = dN_i(t) - Y_i(t)d\Lambda_Z(t)$, the result follows from proving

$$
\sum_{i=1}^n \int_{\mathbb{R}_+} \left(\mathfrak{K}((t,X_i),\cdot) - \sum_{j=1}^n \mathfrak{K}((t,X_j),\cdot) \frac{Y_j(t)}{Y(t)} \right) Y_i(t) d\Lambda_Z(t) = 0.
$$

Observe

$$\begin{aligned}
&\sum_{i=1}^{n}\int_{\mathbb{R}_+}\left(\mathfrak{K}((t,X_i),\cdot)-\sum_{j=1}^{n}\mathfrak{K}((t,X_j),\cdot)\frac{Y_j(t)}{Y(t)}\right)Y_i(t)d\Lambda_Z(t)\\
&=\sum_{i=1}^{n}\int_{\mathbb{R}_+}\mathfrak{K}((t,X_i),\cdot)Y_i(t)d\Lambda_Z(t)-\sum_{i=1}^{n}\int_{\mathbb{R}_+}\sum_{j=1}^{n}\mathfrak{K}((t,X_j),\cdot)\frac{Y_j(t)}{Y(t)}Y_i(t)d\Lambda_Z(t)\\
&=\sum_{i=1}^{n}\int_{\mathbb{R}_+}\mathfrak{K}((t,X_i),\cdot)Y_i(t)d\Lambda_Z(t)-\sum_{j=1}^{n}\int_{\mathbb{R}_+}\mathfrak{K}((t,X_j),\cdot)Y_j(t)d\Lambda_Z(t)\\
&=0,
\end{aligned}$$

which completes the result.

4.9.4.7 Proof of Lemma 4.7

Recall the definition of the log-rank test statistic $\mathrm{LR}_n(\omega)$ given in Equation (4.3) in the main document,

$$\mathrm{LR}_n(\omega)=\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}_+}(\omega(t,X_i)-\bar{\omega}_n(t))dN_i(t),$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}_+}(\omega(t,X_i)-\bar{\omega}_n(t))Y_i(t)d\Lambda_Z(t)=0.$$

Then, by substracting the previous term to $\mathrm{LR}_n(\omega)$, we obtain

$$\mathrm{LR}_n(\omega)=\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}_+}(\omega(t,X_i)-\bar{\omega}_n(t))dM_i(t),$$

and thus, by the definition of the log-rank test-statistic, it follows that

$$\Psi_n=\sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}^2\leq 1}\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}_+}(\omega(t,X_i)-\bar{\omega}_n(t))\,dM_i(t). \tag{4.24}$$

Let $\{(T_i', \Delta_i', X_i')\}_{i=1}^n$ be an independent copy of the data $\mathcal{D} = \{(T_i, \Delta_i, X_i)\}_{i=1}^n$, let $Y_i'(t) = \mathbb{1}_{\{T_i' \geq t\}}$ and let $\mathbb{E}'(\cdot) = \mathbb{E}(\cdot|\mathcal{D})$. The desired result is then obtained by showing that we can replace $\bar{\omega}_n(t)$ in Equation (4.24) by its population limit (under the null hypothesis), given by

$$\frac{\mathbb{E}'(\omega(t, X_1')Y_1'(t))}{S_T(t)} = \frac{\int_{\mathbb{R}^d} \omega(t,x) S_{C|X=x}(t) dF_X(x)}{S_C(t)},$$

up to an error of order $o_p(n^{-1/2})$. This result is not obvious because of the supremum. Let's us make this replacement, and define

$$\begin{aligned}
\bar{\Psi}_n &= \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}^2 \leq 1} \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}_+} \left(\omega(t, X_i) - \frac{\mathbb{E}'(\omega(t, X_1')Y_1'(t))}{S_T(t)} \right) dM_i(t) \\
&= \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}^2 \leq 1} \left\langle \omega, \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}_+} \left(\mathfrak{K}((t, X_i), \cdot) - \frac{\mathbb{E}'(\mathfrak{K}((t, X_1'), \cdot)Y_1'(t))}{S_T(t)} \right) dM_i(t) \right\rangle.
\end{aligned}$$

Note that since the supremum is taken over the unit ball of a reproducing kernel Hilbert space, it is straightforward that

$$\begin{aligned}
\bar{\Psi}_n^2 &= \left\| \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}_+} \left(\mathfrak{K}((t, X_i), \cdot) - \frac{\mathbb{E}'(\mathfrak{K}((t, X_1'), \cdot)Y_1'(t))}{S_T(t)} \right) dM_i(t) \right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{l=1}^n \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} \bar{\mathfrak{K}}((t, X_i), (t', X_l)) dM_i(t) dM_l(t'),
\end{aligned}$$

where $\bar{\mathfrak{K}}$ is the population limit of $\bar{\mathfrak{K}}_n$ (under the null) given in Equation (4.11) in the main document. Thus, the desired result follows from proving $\Psi_n^2 = \bar{\Psi}_n^2 + o_p(n^{-1})$. Define the error term

$$\Upsilon_n = \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}^2 \leq 1} \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}_+} \left(\frac{\mathbb{E}'(\omega(t, X_1')Y_1'(t))}{nS_T(t)} - \bar{\omega}_n(t) \right) dM_i(t),$$

and note that

$$\begin{aligned}\Upsilon_n &= \sup_{\omega\in\mathcal{H}:\|\omega\|^2_{\mathcal{H}}\leq 1} \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}_+}\left(\frac{\mathbb{E}'(\omega(t,X_1')Y_1'(t))}{nS_T(t)} - \bar{\omega}_n(t)\right)dM_i(t) \\ &= \sup_{\omega\in\mathcal{H}:\|\omega\|^2_{\mathcal{H}}\leq 1} \frac{1}{n}\int_{\mathbb{R}_+}\left(\frac{\mathbb{E}'(\omega(t,X_1')Y_1'(t))}{nS_T(t)} - \sum_{j=1}^{n}\omega(t,X_j)\frac{Y_j(t)}{Y(t)}\right)\sum_{i=1}^{n}dM_i(t) \\ &= \sup_{\omega\in\mathcal{H}:\|\omega\|^2_{\mathcal{H}}\leq 1} \frac{1}{n}\int_{\mathbb{R}_+}\sum_{j=1}^{n}\left(\frac{\mathbb{E}'(\omega(t,X_j')Y_j'(t))}{nS_T(t)} - \omega(t,X_j)\frac{Y_j(t)}{Y(t)}\right)dM(t).\end{aligned}$$

Also, note that by the triangular inequality,

$$\Psi_n \leq \bar{\Psi}_n + \Upsilon_n \quad \text{and} \quad \Psi_n \geq \bar{\Psi}_n - \Upsilon_n.$$

Then, if we assume that $\Upsilon_n = o_p(n^{-1/2})$, the result follows by taking square in the previous inequalities,

$$n\Psi_n^2 = (n^{1/2}\bar{\Psi}_n + o_p(1))^2 = n\bar{\Psi}_n^2 + o_p(1),$$

where we note that $n^{1/2}\bar{\Psi}_n o_p(1) = o_p(1)$ since $n^{1/2}\bar{\Psi}_n$ converges in distribution to some random variable (which is proved in Theorem 4.8).

The rest of this proof is concerned with showing that $\Upsilon_n = o_p(n^{-1/2})$ holds under Assumption 4.6. Again, observe that since the supremum in Υ_n is taken over the unit ball of a reproducing kernel Hilbert space, Υ_n^2 satisfies

$$\Upsilon_n^2 = \left\|\frac{1}{n}\int_{\mathbb{R}_+}\sum_{j=1}^{n}\left(\frac{\mathbb{E}'(\mathfrak{K}((t,X_j'),\cdot)Y_j'(t))}{nS_T(t)} - \mathfrak{K}((t,X_j),\cdot)\frac{Y_j(t)}{Y(t)}\right)dM(t)\right\|^2_{\mathcal{H}}. \tag{4.25}$$

Under Assumption 4.6, $\mathfrak{K}((t,x),(t',x')) = L(t,t')K(x,x')$, and by using the reproduc-

ing property, we get

$$\Upsilon_n^2 = \frac{1}{n^2}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+} L(t,s)\sum_{j=1}^{n}\sum_{i=1}^{n}\kappa_{i,j}(t,s)dM(t)dM(s),$$

where

$$\kappa_{i,j}(t,s) = \frac{\mathbb{E}'(K(X_i', X_j')Y_i'(t)Y_j'(s))}{n^2 S_T(t)S_T(s)} - \frac{\mathbb{E}'(K(X_i', X_j)Y_i'(t))Y_j(s)}{nS_T(t)Y(s)} - \frac{\mathbb{E}'(K(X_i, X_j')Y_j'(s))Y_i(t)}{nS_T(s)Y(t)} + \frac{K(X_i, X_j)Y_i(t)Y_j(s)}{Y(t)Y(s)}.$$

Define the process $(R(z))_{z\geq 0}$ by

$$R(z) = \frac{1}{n}\int_0^z\int_0^z L(t,s)\sum_{j=1}^{n}\sum_{i=1}^{n}\kappa_{i,j}(t,s)dM(t)dM(s),$$

and note that

$$n\Upsilon_n^2 = R(\tau_n).$$

We will prove that there exists an increasing predictable process $(A(z))_{z\geq 0}$ with $A(0) = 0$ such that $\mathbb{E}(R(T)) \leq \mathbb{E}(A(T))$ for any finite stopping time, i.e., the process A majorises R. Then, we will apply the Lenglart-Rebolledo inequality, given in Lemma 4.21, to show that $A(\tau_n) = o_p(1)$ implies $\sup_{z\leq\tau_n} R(z) = o_p(1)$, and thus we will deduce the desired result from $n\Upsilon_n^2 = R(\tau_n)$.

To find the process A, note that

$$R(z) = Z_D(z) + Z_O(z),$$

where $(Z_D(z))_{z\geq 0}$ and $(Z_O(z))_{z\geq 0}$ are the processes defined by

$$Z_D(z) = \frac{1}{n}\int_0^z L(t,t)\sum_{i=1}^n\sum_{j=1}^n \kappa_{i,j}(t,t)dM(t)dM(t)$$

and

$$Z_O(z) = \frac{2}{n}\int_0^z \int_{(0,t)} L(t,s)\sum_{i=1}^n\sum_{j=1}^n \kappa_{i,j}(t,s)dM(s)dM(t).$$

By Theorem 4.24, since the functions $L(t,s)$ and $\kappa_{i,j}(t,s)$ (for any $i,j \in \{1,\ldots,n\}$) are both predictable functions in the sense of definition 4.22, the process $Z_O(z)$ is a zero-mean square-integrable $(\mathcal{F}_z)$-martingale. By the Optional Stopping Theorem, $\mathbb{E}(Z_O(T)) = 0$ for any bounded stopping time T, and thus

$$\mathbb{E}(R(T)) = \mathbb{E}(Z_D(T)), \tag{4.26}$$

for any bounded stopping time T.

Note that

$$\begin{aligned} Z_D(z) &= \frac{1}{n}\int_0^z L(t,t)\sum_{i=1}^n\sum_{j=1}^n \kappa_{i,j}(t,t)dM(t)dM(t) \\ &= \frac{1}{n}\int_0^z L(t,t)\sum_{i=1}^n\sum_{j=1}^n \kappa_{i,j}(t,t)dN(t), \end{aligned}$$

where the second equality holds since $dM(t)^2 = dN(t)$, which follows from the fact the the T_i's are continuous random variables. Additionally, note that $Z_D(0) = 0$, and that $Z_D(z) \geq 0$, since

$$L(t,t)\sum_{i=1}^n\sum_{j=1}^n \kappa_{i,j}(t,t) = \left\| \sum_{j=1}^n \left(\frac{\mathbb{E}(\mathfrak{K}((t,X_j'),\cdot)Y_j'(t))}{nS_T(t)} - \mathfrak{K}((t,X_j),\cdot)\frac{Y_j(t)}{Y(t)} \right) \right\|_{\mathcal{H}}^2 \geq 0.$$

The process $Z_D(z)$ is adapted and increasing, and it can be compensated by

$$\langle Z_D(z)\rangle = \frac{1}{n}\int_0^z L(t,t)\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_{i,j}(t,t)Y(t)d\Lambda_Z(t).$$

Since $Z_D - \langle Z_D\rangle$ is a martingale, then $\mathbb{E}(Z_D(T) - \langle Z_D(T)\rangle) = 0$ for any finite stopping time T. Using Equation (4.26), we conclude that

$$E(R(T)) = \mathbb{E}(\langle Z_D(T)\rangle),$$

for any finite stopping time. That is, the process $\langle Z_D(z)\rangle$ is a predictable increasing process with $\langle Z_D(0)\rangle = 0$, and $\langle Z_D(z)\rangle$ majorizes $R(z)$.

The final step is to prove that $\langle Z_D(\tau_n)\rangle = o_p(1)$, and to use the Lenglart-Rebolledo inequality to conclude the main result. Note that

$$\langle Z_D(\tau_n)\rangle = O_p(1)\int_0^z L(t,t)\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_{i,j}(t,t)S_C(t)dF_Z(t),$$

since by by Proposition 4.19.1, $Y(t)/n = O_p(1)S_T(t)$ uniformly for all $t \leq \tau_n$. Recall that

$$L(t,t)\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_{i,j}(t,t) = \left\|\sum_{j=1}^{n}\left(\frac{\mathbb{E}(\mathfrak{K}((t,X_j'),\cdot)Y_j'(t))}{nS_T(t)} - \mathfrak{K}((t,X_j),\cdot)\frac{Y_j(t)}{Y(t)}\right)\right\|_{\mathcal{H}}^2.$$

and note that

$$
\begin{aligned}
&L(t,t)\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_{i,j}(t,t)\\
&\leq \sum_{j=1}^{n}\left\|\frac{\mathbb{E}(\mathfrak{K}((t,X_j'),\cdot)Y_j'(t))}{nS_T(t)}-\mathfrak{K}((t,X_j),\cdot)\frac{Y_j(t)}{Y(t)}\right\|_{\mathcal{H}}^2\\
&= L(t,t)\sum_{j=1}^{n}\left(\frac{\mathbb{E}(K(X_j',X_j')Y_j'(t))}{n^2S_T(t)^2}-\frac{2\mathbb{E}'(K(X_j',X_j)Y_j'(t))Y_j(t)}{nS_T(t)Y(t)}+\frac{K(X_j,X_j)Y_j(t)}{Y(t)^2}\right)\\
&\leq C\sum_{j=1}^{n}\left(\frac{1}{n^2S_T(t)}+\frac{2Y_j(t)}{nY(t)}+\frac{Y_j(t)}{Y(t)^2}\right)\\
&\leq C\left(\frac{1}{nS_T(t)}+\frac{2}{n}+\frac{1}{Y(t)}\right),
\end{aligned}
$$

where the first equality is due to the reproducing property and Assumption 4.6, and the second inequality follows also from Assumption 4.6, in which we assume the kernel is bounded by some constant $C>0$.

Using the previous upper bound, we obtain

$$
\begin{aligned}
\langle Z_D(\tau_n)\rangle &= O_p(1)\int_0^{\tau_n}\left(\frac{1}{nS_T(t)}+\frac{2}{n}+\frac{1}{Y(t)}\right)S_C(t)dF_Z(t)\\
&= O_p(1)\int_0^{\tau_n}\frac{1}{nS_T(t)}S_C(t)dF_Z(t)\\
&= O_p(1)\frac{1}{n}\Lambda_Z(\tau_n).
\end{aligned}
$$

Note that $\tau_n=\max\{T_1,\ldots,T_n\}\leq\max\{Z_1,\ldots,Z_n\}$, and thus

$$
\frac{1}{n}\Lambda_Z(\tau_n)\leq\frac{1}{n}\max_{i\in\{1,\ldots,n\}}\Lambda_Z(Z_i).
$$

Note also that the random variables $\Lambda_Z(Z_i)$ are i.i.d. exponential random variables with mean 1, as

$$
\mathbb{P}(\Lambda_Z(Z_1)>x)=\mathbb{P}(Z_1>\Lambda_Z^{-1}(x))=S_Z(\Lambda_Z^{-1}(x))=\exp\{-\Lambda_Z(\Lambda_Z^{-1}(x))\}=\exp\{-x\}.
$$

We finally deduce the result in consequence of

$$P\left(\max_{i\in\{1,\dots,n\}}\Lambda_Z(Z_i)\geq 2\log(n)\right)\leq nP\left(\Lambda_Z(Z_1)\geq 2\log(n)\right)=n\exp\{\log(n^{-2})\}=\frac{1}{n}.$$

Thus in a set of probability tending to 1, $\max_{i\in\{1,\dots,n\}}\Lambda_Z(Z_i)\leq 2\log(n)$, and thus

$$\langle Z_D(\tau_n)\rangle = O_p(1)\frac{1}{n}\Lambda_Z(\tau_n)=o_p(1).$$

4.9.4.8 Proof of Theorem 4.8

Proof. By Lemma 4.7,

$$\Psi_n^2=\frac{1}{n^2}\sum_{i=1}^n\sum_{l=1}^n J((T_i,\Delta_i,X_i),(T_l,\Delta_l,X_l))+o_p(n^{-1})$$

under the null hypothesis. The non-negligible part in the right-hand side of the previous equation is a degenerate V-statistic of order 2.

The degeneracy property can be easily verified, since

$$\begin{aligned}J((T_i,\Delta_i,X_i),(t,r,x)) &= \int_{\mathbb{R}_+}\left(\int_{\mathbb{R}_+}\bar{\mathfrak{K}}((s,X_i),(s',x))dm_{t,r}(s')\right)dM_i(s)\\ &= \int_{\mathbb{R}_+} f_i(s,t,r,x)dM_i(s),\end{aligned}$$

where $f_i(s,t,r,x)=\int_{\mathbb{R}_+}\bar{\mathfrak{K}}((s,X_i),(s',x))dm_{t,r}(s')$. Then, note that for any fixed triple $(t,r,x)\in\mathbb{R}_+\times\{0,1\}\times\mathbb{R}^d$, the process $(f_i(s,t,r,x))_{s\geq 0}$ is $(\mathcal{F}_t)$-predictable, and thus $\mathbb{E}(J(T_i,\Delta_i,X_i),(t,r,x))=\mathbb{E}\left(\int_{\mathbb{R}_+} f_i(s,t,r,x)dM_i(s)\right)=0$ for any $(t,r,x)\in\mathbb{R}_+\times\{0,1\}\times\mathbb{R}^d$, since $\int_0^z f_i(s,t,r,x)dM_i(s)$, $z\geq 0$, is a $\mathcal{F}_z$-martingale.

Using the previous observation, and from the classical theory of V-statistics Serfling (1980), we obtain

$$n\Psi^2\xrightarrow{\mathcal{D}}E(J((T_i,\Delta_i,X_i),(T_i,\Delta_i,X_i)))+\mathcal{Y},\tag{4.27}$$

where $\mathcal{Y} = \sum_{i=1}^{n} \lambda_i(\xi_i^2 - 1)$, $\xi_1, \xi_2, \ldots$ are independent standard normal random variables, and $\lambda_1, \lambda_2, \ldots$ are the eigenvalues associated to the integral operator of J, T^J, defined in Equation (4.19).

Finally, note that

$$\begin{aligned}
&\mathbb{E}(J((T_i, \Delta_i, X_i), (T_i, \Delta_i, X_i))) \\
&= \mathbb{E}\left(\int_0^{T_i}\int_0^{T_i} \mathbb{1}_{\{t \neq t'\}} \bar{\mathfrak{K}}((t, X_i), (t', X_i)) dM_i(t) dM_i(t')\right) + \mathbb{E}\left(\bar{\mathfrak{K}}((T_i, X_i), (T_i, X_i))\Delta_i\right) \\
&= \int_{x \in \mathbb{R}^d} \int_0^{\tau} \bar{\mathfrak{K}}((t, x), (t, x)) S_{C|X=x}(t) dF_Z(t) dF_X(x),
\end{aligned}$$

where the expectation of the first term is equal to zero by the double martingale Theorem 4.24. □

4.9.4.9 Proof of Theorem 4.9

Proof. Let $Q : (\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^d)^2 \to \mathbb{R}$ be given by

$$Q((z_1, c_1, x_1), (z_2, c_2, x_2)) = \mathbb{1}_{\{z_1 \leq c_1\}} \mathbb{1}_{\{z_2 \leq c_2\}} \bar{\mathfrak{K}}((z_1, x_1), (z_2, x_2)).$$

Note that $\Delta_i \Delta_j \bar{\mathfrak{K}}((T_i, X_i), (T_j, X_j)) = Q((Z_i, C_i, X_i), (Z_j, C_j, X_j))$, and thus the asymptotic distribution of $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i \Delta_j \bar{\mathfrak{K}}((T_i, X_i), (T_j, X_j))$ can be found by studying the asymptotic distribution of the degenerate V-statistic,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} Q((Z_i, C_i, X_i), (Z_j, C_j, X_j)).$$

Let $Q_{ij} = Q((Z_i, C_i, X_i), (Z_j, C_j, X_j))$. Then $\mathbb{E}(|Q_{ii}|) < \infty$ and $\mathbb{E}(Q_{ij}^2) < \infty$ for $i \neq j$, since, by Assumption 4.6, the kernel $\mathfrak{K}$ is bounded. By the classical theory of

U/V-statistics Serfling (1980),

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij} \xrightarrow{\mathcal{D}} \int_{\mathbb{R}^d}\int_0^\tau \bar{\mathfrak{K}}((t,x),(t,x))S_{C|X=x}(t)dF_Z(t)dF_X(t) + \mathcal{Y}',$$

where $\mathcal{Y}' = \sum_{i=1}^{\infty}\lambda_i'(\xi_i^2 - 1)$, $\xi_1, \xi_2, \ldots$ are independent standard normal random variables and $\lambda_1,', \lambda_2', \ldots$ are the eigenvalues associated to the integral operator T^Q. By Theorem 4.31, T^J (the integral operator of J) and T^Q share the same set of non-zero eigenvalues (including multiplicities), and thus the result holds. □

4.9.4.10 Proof of lemma 4.10

Observe that

$$\|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 = V_{0,0} - 2V_{0,1} + V_{1,1},$$

where

$$V_{0,0} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\mathfrak{K}((T_i,X_j),(T_l,X_k))\frac{Y_j(T_i)}{Y(T_i)}\frac{Y_k(T_l)}{Y(T_l)}\Delta_i\Delta_l, \tag{4.28}$$

$$V_{0,1} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\mathfrak{K}((T_i,X_j),(T_l,X_l))\Delta_i\Delta_l\frac{Y_j(T_i)}{Y(T_i)}, \tag{4.29}$$

$$V_{1,1} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n}\mathfrak{K}((T_i,X_i),(T_l,X_l))\Delta_i\Delta_l, \tag{4.30}$$

hold by the reproducing property. The desired result

$$\|\phi_0^n - \phi_1^n\|_{\mathcal{H}}^2 \xrightarrow{\mathbb{P}} \|\phi_0 - \phi_1\|_{\mathcal{H}}^2,$$

follows by proving

i) $V_{0,0} \xrightarrow{\mathbb{P}} \|\phi_0\|_{\mathcal{H}}^2$,

ii) $V_{0,1} \xrightarrow{\mathbb{P}} \langle\phi_0, \phi_1\rangle_{\mathcal{H}}$,

iii) $V_{1,1} \xrightarrow{\mathbb{P}} \|\phi_1\|^2_{\mathcal{H}}$.

We start with iii). Note that $V_{1,1}$ is a V-statistic of order 2, and that, by Assumption 4.6, $\mathbb{E}(|\mathfrak{K}((T_i, X_i), (T_i, X_i))\Delta_i|) < \infty$ and $\mathbb{E}(|\mathfrak{K}((T_i, X_i), (T_j, X_j))\Delta_i\Delta_j|) < \infty$ for $i \neq j$. Then, by the classical theory of V-statistics, it holds

$$\begin{aligned} V_{1,1} &\xrightarrow{\mathbb{P}} \mathbb{E}(\mathfrak{K}((T_1, X_1), (T_2, X_2))\Delta_1\Delta_2) \\ &= \int_0^\tau \int_{x\in\mathbb{R}^d} \int_0^\tau \int_{x'\in\mathbb{R}^d} \mathfrak{K}((t, x), (t', x'))S_{C|X=x}(t)S_{C|X=x'}(t')dF_{ZX}(t', x')dF_{ZX}(t, x) \\ &= \|\phi_1\|^2_{\mathcal{H}}. \end{aligned}$$

We next prove i) and ii). Since the proofs of $i)$ and $ii)$ use the same arguments, it sufficesto prove the result for $i)$. Rewrite $V_{0,0}$ as

$$V_{0,0} = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n\sum_{k=1}^n f(i, j, l, k),$$

where

$$f(i, j, l, k) = \mathfrak{K}((T_i, X_j), (T_l, X_k))\frac{Y_j(T_i)}{Y(T_i)}\frac{Y_k(T_l)}{Y(T_l)}\Delta_i\Delta_l.$$

Note that

$$f(i, j, l, k) \leq C\frac{Y_j(T_i)}{Y(T_i)}\frac{Y_k(T_l)}{Y(T_l)},$$

for some constant $C \geq 0$ due to Asssumption 4.6.

We begin by proving that the sum over elements in $V_{0,0}$ that contain the repetition of one index (i, j, l, k) converge to zero. Define the set $\mathcal{S} = \{(i, j, l, k) \in [n]^4 : |\{i, j, l, k\}| \leq 3\}$, e.g. $(1, 1, 2, 3) \in \mathcal{S}$, and let $\mathcal{S}_{i=l}$ be the set containing all the indices (i, j, l, k) such that $i = l$. It is not hard to see that $\mathcal{S} \subseteq \mathcal{S}_{i=l} \cup \mathcal{S}_{i=j} \cup \mathcal{S}_{i=k} \cup \mathcal{S}_{l=k} \cup \mathcal{S}_{l=j} \cup \mathcal{S}_{j=k}$ (note that the intersection of these sets may not be empty).

Consider the set $\mathcal{S}_{i=l}$, and note that the sum over the elements of $\mathcal{S}_{i=l}$ satisfies

$$\frac{1}{n^2}\sum_{(i,j,k,l)\in\mathcal{S}_{i=l}} f(i,j,l,k) \leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} C\frac{Y_j(T_i)}{Y(T_i)}\frac{Y_k(T_i)}{Y(T_i)} \leq \frac{1}{n^2}\sum_{i=1}^{n} C \to 0.$$

as n tends to infinity.

The sum over the elements in $\mathcal{S}_{i=j}$ (or, by symmetry, over the elements $\mathcal{S}_{l=k}$) satisfies

$$\begin{aligned}\frac{1}{n^2}\sum_{(i,j,k,l)\in\mathcal{S}_{i=j}} f(i,j,l,k) &\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n} C\frac{Y_i(T_i)}{Y(T_i)}\frac{Y_k(T_l)}{Y(T_l)} \leq \frac{C}{n}\sum_{i=1}^{n}\frac{1}{Y(T_i)} \\ &\leq \frac{C}{n}\sum_{i=1}^{n}\frac{1}{n-i+1} \leq \frac{C}{n}\sum_{k=1}^{n}\frac{1}{k} \leq C\frac{\log(n)+1}{n} \to 0,\end{aligned}$$

as n grows to infinity.

The sum over the elements of $\mathcal{S}_{i=k}$ (or, by symmetry, over $\mathcal{S}_{l=j}$) satisfies

$$\begin{aligned}\frac{1}{n^2}\sum_{(i,j,k,l)\in\mathcal{S}_{i=k}} f(i,j,l,k) &\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} C\frac{Y_j(T_i)}{Y(T_i)}\frac{Y_i(T_l)}{Y(T_l)} \\ &\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n} C\frac{Y(T_i)}{Y(T_i)}\frac{Y_i(T_l)}{Y(T_l)} \leq \frac{1}{n^2}\sum_{l=1}^{n} C \to 0.\end{aligned}$$

Finally, the sum over the elemnts of $\mathcal{S}_{j=k}$, satisfies

$$\frac{1}{n^2}\sum_{(i,j,k,l)\in\mathcal{S}_{j=k}} f(i,j,l,k) \leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n} C\frac{Y_k(T_i)}{Y(T_i)}\frac{Y_k(T_l)}{Y(T_l)} \leq \frac{C}{n}\sum_{l=1}^{n}\frac{1}{Y(T_l)} \to 0,$$

as n grows to infinity.

The previous results imply

$$V_{0,0} = \frac{1}{n^2}\sum_{(i,j,k,l)\in\mathcal{S}^c} \mathfrak{K}((T_i,X_j),(T_l,X_k))\frac{Y_j(T_i)}{Y(T_i)}\frac{Y_k(T_l)}{Y(T_l)}\Delta_i\Delta_l + o(1),$$

where $\mathcal{S}^c = \{(i,j,l,k) \in [n]^4 : |\{i,j,l,k\}| = 4\}$.

We continue by assuming that $Y(t)/n$ can be replaced by its limit, $S_T(t)$, in the previous equation. By doing this, we obtain

$$V_{0,0} = \frac{1}{n^4} \sum_{\mathcal{S}^c} \mathfrak{K}((T_i, X_j), (T_l, X_k)) \frac{Y_j(T_i)}{S_T(T_i)} \frac{Y_k(T_l)}{S_T(T_l)} \Delta_i \Delta_l + o_p(1), \tag{4.31}$$

which is a U-statistic of order 4. Then, by using the law of large numbers for U-statistics Serfling (1980), it follows that

$$\begin{aligned} V_{0,0} &= \mathbb{E}\left(\mathfrak{K}((T_1, X_3), (T_2, X_4)) \frac{Y_3(T_1)}{S_T(T_1)} \frac{Y_4(T_2)}{S_T(T_2)} \Delta_1 \Delta_2 \right) + o_p(1) \\ &= \int_0^\tau \int_{x \in \mathbb{R}^d} \int_0^\tau \int_{x' \in \mathbb{R}^d} K((t,x),(t',x')) d\nu_0(t',x') d\nu_0(t,x) + o_p(1), \\ &= \|\phi_0\|_{\mathcal{H}}^2 + o_p(1) \end{aligned}$$

from which we conclude the desired result.

We finalise our proof by justifying the replacement of $Y(t)$ by $nS_T(t)$ made in equation (4.31). By the triangular inequality, it holds

$$\begin{aligned} &\left| \frac{1}{n^2} \sum_{\mathcal{S}^c} \mathfrak{K}((T_i, X_j), (T_l, X_k)) \left(\frac{Y_j(T_i)}{Y(T_i)} \frac{Y_k(T_l)}{Y(T_l)} - \frac{Y_j(T_i)}{nS_T(T_i)} \frac{Y_k(T_l)}{nS_T(T_l)} \right) \Delta_i \Delta_l \right| \\ &\quad \leq \left| \frac{1}{n^2} \sum_{\mathcal{S}^c} \mathfrak{K}((T_i, X_j), (T_l, X_k)) \frac{Y_j(T_i)}{Y(T_i)} \left(\frac{Y_k(T_l)}{Y(T_l)} - \frac{Y_k(T_l)}{nS_T(T_l)} \right) \Delta_i \Delta_l \right| \qquad (4.32) \\ &\qquad + \left| \frac{1}{n^2} \sum_{\mathcal{S}^c} \mathfrak{K}((T_i, X_j), (T_l, X_k)) \frac{Y_k(T_l)}{S_T(T_l)} \left(\frac{Y_j(T_i)}{Y(T_i)} - \frac{Y_j(T_i)}{nS_T(T_i)} \right) \Delta_i \Delta_l \right|. \qquad (4.33) \end{aligned}$$

We prove that equations (4.32) and (4.33) are both $o_p(1)$ under Assumption 4.6. Since the proofs use the same arguments, we only show the result for equation (4.32).

Note that under Assumption 4.6, it holds

$$\begin{aligned}
&\left|\frac{1}{n^2}\sum_{(i,j,k,l)\in\mathcal{S}^c}\mathfrak{K}((T_i,X_j),(T_l,X_k))\frac{Y_j(T_i)}{Y(T_i)}\left(\frac{Y_k(T_l)}{Y(T_l)}-\frac{Y_k(T_l)}{nS_T(T_l)}\right)\Delta_i\Delta_l\right| \\
&\quad\leq \frac{C}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{Y_j(T_i)}{Y(T_i)}\left|\frac{Y_k(T_l)}{Y(T_l)}-\frac{Y_k(T_l)}{nS_T(T_l)}\right|\Delta_l \\
&\quad\leq \frac{C}{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\left|\frac{Y_k(T_l)}{Y(T_l)}-\frac{Y_k(T_l)}{nS_T(T_l)}\right|\Delta_l \\
&\quad\leq \frac{C}{n}\int_0^\tau\sum_{k=1}^{n}\frac{Y_k(t)}{Y(t)}\left|1-\frac{Y(t)/n}{S_T(t)}\right|dN(t) \\
&\quad\leq \frac{C}{n}\int_0^\tau\left|1-\frac{Y(t)/n}{S_T(t)}\right|dN(t)=o_p(1),
\end{aligned}$$

where the last equality follows from Lemma 4.32.

4.9.4.11 Proof of Theorem 4.11

By Lemma 4.10 under Assumption 4.6, we have

$$\Psi_n^2 \xrightarrow{\mathbb{P}} \|\phi_1-\phi_0\|_{\mathcal{H}}^2,$$

where

$$\phi_i(s,y)=\int_0^\tau\int_{x\in\mathbb{R}^d}L(s,t)K(y,x)d\nu_i(t,x),\quad i\in\{0,1\}.$$

Then, the result follows from proving $\|\phi_1-\phi_0\|_{\mathcal{H}}^2=0$ if and only if the null hypothesis holds, since if $\|\phi_1-\phi_0\|_{\mathcal{H}}^2>0$, then $n\Psi_n^2=n\|\phi_1-\phi_0\|_{\mathcal{H}}^2\to\infty$ as $n\to\infty$.

Suppose that the null hypothesis holds. Then, under Assumption 4.2, Proposition 4.3 implies that $\nu_1=\nu_0$, and thus $\phi_0=\phi_1$.

We proceed to show that $\phi_1-\phi_0=0$ implies $\nu_1=\nu_0$. Note that by Proposition 4.3, the latter implies that the null hypothesis holds.

To show this result, we repeat the proof in Gretton (2015). Recall that

$$(\phi_1 - \phi_0)(s, y) = \int_0^\tau \int_{x \in \mathbb{R}^d} K(y, x) L(s, t)(d\nu_1(t, x) - d\nu_0(t, x)). \tag{4.34}$$

Note that by our assumptions both kernels K and L are characteristic, translation invariant, c_0-kernels. Thus, an application of proposition 4.27 allows us to deduce that both K and L are c_0-universal kernels.

Let $\mathcal{H}_K$ and $\mathcal{H}_L$ be the RKHS induced by the kernels K and L, respectively. For every $\kappa \in \mathcal{H}_K$ define the signed measure γ_κ given by

$$\gamma_\kappa(B) = \int_B \int_{\mathbb{R}^d} \langle K(\cdot, x), \kappa \rangle_{\mathcal{H}_K} (d\nu_1(t, x) - d\nu_0(t, x)),$$

for any measurable $B \subseteq (0, \tau)$. Consider the embedding of the above measure onto $\mathcal{H}_L$, μ_{γ_κ} given by

$$\mu_{\gamma_\kappa}(s) = \int_0^\tau L(s, t) d\gamma_\kappa(t).$$

Following some computations, we obtain

$$\begin{aligned}
\mu_{\gamma_\kappa}(s) &= \int_0^\tau L(s, t) \int_{\mathbb{R}^d} \langle K(\cdot, x), \kappa \rangle_{\mathcal{H}_K} (d\nu_1(t, x) - d\nu_0(t, x)) \\
&= \left\langle \int_0^\tau \int_{\mathbb{R}^d} L(s, t) K(\cdot, x)(d\nu_1(t, x) - d\nu_0(t, x)), \kappa \right\rangle_{\mathcal{H}_K} \\
&= \langle (\phi_1 - \phi_0)(s, \cdot), \kappa \rangle_{\mathcal{H}_K} \\
&= 0,
\end{aligned}$$

since $(\phi_1 - \phi_0)(s, y) = 0$ for all $(s, y) \in \mathbb{R}_+ \times \mathbb{R}_d$ by hypothesis.

By Proposition 4.26, since L is c_0-universal, the embedding μ_{γ_κ} is injective, and thus $\gamma_\kappa(B) = 0$ for all B and $\kappa \in \mathcal{H}_K$. Since this is true for all $\kappa \in \mathcal{H}_K$, it is true in

particular for $K(\cdot, y) \in \mathcal{H}_K$, and thus for $\kappa(\cdot) = K(\cdot, y)$, we have

$$\begin{aligned}\gamma_\kappa(B) &= \int_B \int_{\mathbb{R}^d} \langle K(\cdot, x), K(\cdot, y)\rangle_{\mathcal{H}_K} (d\nu_1(t, x) - d\nu_0(t, x)) \\ &= \int_0^\tau \int_{\mathbb{R}^d} K(y, x) \mathbb{1}_B(t)(d\nu_1(t, x) - d\nu_0(t, x)) = 0.\end{aligned}$$

for all measurable $B \subseteq \mathbb{R}$. For any fixed measurable $B \subseteq (0, \tau)$, define the signed measure $\tilde{\gamma}_B(A) = \int_0^\tau \int_A \mathbb{1}_B(t)(d\nu_1(t, x) - d\nu_0(t, x))$ for any measurable $A \subseteq \mathbb{R}^d$. The above equation can be interpreted as the embeding of $\tilde{\gamma}_B$ to $\mathcal{H}_K$, that is,

$$\mu_{\tilde{\gamma}_B}(y) = \int_0^\tau \int_{\mathbb{R}^d} K(y, x) \mathbb{1}_B(t)(d\nu_1(t, x) - d\nu_0(t, x)) = 0.$$

which is also injective (by Proposition 4.26, since K is c_0-universal). We conclude that

$$\tilde{\gamma}_B(A) = \int_0^\tau \int_A \mathbb{1}_B(t)(d\nu_1(t, x) - d\nu_0(t, x)) = 0$$

for all Borel sets A, B, and hence $\nu_1 - \nu_0 = 0$.

4.9.4.12 Proofs of Section 4.5

4.9.4.13 Proof of Lemma 4.14

The desired result is obtained by showing that

$$\sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}} \leq 1} \frac{1}{n} \sum_{i=1}^n \int_0^\tau W_i \left(\bar{\omega}_n(t) - \bar{\omega}(t)\right) dN_i(t) = o_p(n^{-1/2}),$$

where $\bar{\omega}(t)$, given by $\bar{\omega}(t) = \int_{\mathbb{R}^d} \omega(t, x) \frac{S_{T|X=x}(t)}{S_T(t)} dF_X(x)$, is the population version of $\bar{\omega}_n(t) = \sum_{j=1}^n \omega(t, X_j) \frac{Y_j(t)}{Y(t)}$.

Let $\omega^\star(t) = \frac{1}{nS_T(t)} \sum_{j=1}^n \omega(t, X_j) Y_j(t)$. By the triangular inequality,

$$\sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}} \leq 1} \frac{1}{n} \sum_{i=1}^n \int_0^\tau W_i \left(\bar{\omega}_n(t) - \bar{\omega}(t) \right) dN_i(t) \leq E_1 + E_2,$$

where

$$\begin{aligned} E_1 &= \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}} \leq 1} \frac{1}{n} \sum_{i=1}^n \int_0^\tau W_i \left(\bar{\omega}_n(t) - \omega^\star(t) \right) dN_i(t) \\ &= \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}^2 \leq 1} \frac{1}{n} \sum_{i=1}^n \int_0^\tau W_i \sum_{j=1}^n \omega(t, X_j) Y_j(t) R_n(t) dN_i(t), \end{aligned} \tag{4.35}$$

where $R_n(t) = \frac{1}{Y(t)} - \frac{1}{nS_T(t)}$, and

$$\begin{aligned} E_2 &= \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}} \leq 1} \frac{1}{n} \sum_{i=1}^n \int_0^\tau W_i \left(\bar{\omega}_n(t) - \omega^\star(t) \right) dN_i(t) \\ &= \sup_{\omega \in \mathcal{H}, \|\omega\|^2 \leq 1} \frac{1}{n} \sum_{i=1}^n \int_0^\tau \frac{W_i}{nS_T(t)} \left(\sum_{j=1}^n \omega(t, X_j) Y_j(t) - g(t) \right) dN_i(t), \end{aligned} \tag{4.36}$$

where $g(t) = \int_{\mathbb{R}^d} \omega(t, x) S_{T|X=x}(t) dF_X(x)$. The result follows by proving that both E_1 and E_2 are $o_p(n^{-1/2})$.

We proceed to prove $E_1 = o_p(n^{-1/2})$. Let $\beta \in (0, 1)$, and define the set $B_{n,\beta}$ by

$$B_{n,\beta} = \left\{ \frac{Y(t)/n}{S_T(t)} \leq \beta^{-1}, \quad \forall t \leq \tau_n \right\}. \tag{4.37}$$

By Proposition 4.19.1, the set $B_{n,\beta}$ satisfies $\mathbb{P}(B_{n,\beta}) \geq 1 - \beta$. Note that we only need to prove that $n^{1/2} E_1$ converges to zero in probability in the set $B_{n,\beta}$, as β can be

chosen arbitrarily small. Observe

$$
\begin{aligned}
nE_1^2 &= \left(\sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}^2 \leq 1} \frac{1}{n^{1/2}} \sum_{i=1}^{n} \int_0^\tau W_i \sum_{j=1}^{n} \omega(t, X_j) Y_j(t) R_n(t) dN_i(t) \right)^2 \\
&= \frac{1}{n} \sum_{i,l=1}^{n} W_i W_l \int_0^\tau \int_0^\tau \sum_{j,k=1}^{n} \mathfrak{K}((t, X_j), (s, X_k)) Y_j(t) Y_k(s) R_n(t) R_n(s) dN_i(t) dN_l(s) \\
&\leq \frac{C}{n} \sum_{i,l=1}^{n} W_i W_l \int_0^\tau \int_0^\tau Y(t) Y(s) |R_n(t)| |R_n(s)| dN_i(t) dN_l(s),
\end{aligned}
$$

where the first equality is due to the fact that we are taking supremum in the unit ball of a RKHS, and the inequality in the second line follows from Assumption 4.6; that is, the kernel $\mathfrak{K}$ is bounded by some constant $C > 0$.

Note that

$$
\mathbb{E}(nE_1^2 \mathbb{1}_{\{B_{n,\beta}\}}) \leq \mathbb{E}\left(\frac{C}{n} \sum_{i=1}^{n} \left(\int_0^\tau Y(t) |R_n(t)| dN_i(t) \right)^2 \mathbb{1}_{\{B_{n,\beta}\}} \right), \tag{4.38}
$$

since $W_1, \ldots, W_n$ are i.i.d. Rademacher random variables which are independent of our data $((T_i, \Delta_i, X_i))_{i=1}^n$. We will prove that the expectation in the right-hand side of Equation (4.38) converges to zero. To do this, we use dominated convergence in sets of arbitrarily large probability. First, note that

$$
Y(t)|R_n(t)| = \left| 1 - \frac{Y(t)/n}{S_T(t)} \right| \leq 1 + \beta^{-1}
$$

uniformly for all $t \leq \tau_n$ in the set $B_{n,\beta}$, and thus

$$
\frac{C}{n} \sum_{i=1}^{n} \left(\int_0^\tau Y(t) |R_n(t)| dN_i(t) \right)^2 \mathbb{1}_{\{B_{n,\beta}\}} \leq C(1 + \beta^{-1})^2. \tag{4.39}
$$

Second, note that

$$\frac{C}{n}\sum_{i=1}^{n}\left(\int_0^\tau Y(t)|R_n(t)|dN_i(t)\right)^2 = \frac{C}{n}\int_0^\tau \left|1-\frac{Y(t)/n}{S_T(t)}\right|^2 dN(t) = o_p(1),$$

by Lemma 4.32. By the dominated convergence theorem we deduce $\mathbb{E}(nE_1^2\mathbb{1}_{\{B_{n,\beta}\}}) \to 0$ as n grows to infinity, for which follows the desired result $E_1 = o_p(n^{-1/2})$.

We continue by proving that $E_2 = o_p(n^{-1/2})$. Note that it is enough to prove the result in the set $B_{n,\beta}$ defined in Equation (4.37), as β can be chosen arbitrarily small. Observe that

$$\begin{aligned} nE_2^2 &= \left(\sup_{\omega\in\mathcal{H},\|\omega\|^2\leq 1}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\int_0^\tau \frac{W_i}{nS_T(t)}\left(\sum_{j=1}^{n}\omega(t,X_j)Y_j(t)-g(t)\right)dN_i(t)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{n}\int_0^\tau\int_0^\tau \frac{W_iW_l}{n^2S_T(t)S_T(s)}\mathfrak{K}^\star(t,s)dN_i(t)dN_l(s), \end{aligned}$$

where

$$\begin{aligned} \mathfrak{K}^\star(t,s) = &\sum_{j=1}^{n}\sum_{k=1}^{n}(\mathfrak{K}((t,X_j),(s,X_k))Y_j(t)Y_k(s) \\ &-\int_{x'\in\mathbb{R}^d}\mathfrak{K}((t,X_j),(s,x'))S_{T|X=x'}(s)dF_X(s)Y_j(t) \\ &-\int_{x\in\mathbb{R}^d}\mathfrak{K}((t,x),(s,X_k))S_{T|X=x}(t)dF_X(x)Y_k(s) \\ &+\int_{x'\in\mathbb{R}^d}\int_{x\in\mathbb{R}^d}\mathfrak{K}((t,x),(s,x'))S_{T|X=x}(t)S_{T|X=x'}(s)dF_X(x)dF_X(x')\Bigg), \end{aligned} \tag{4.40}$$

follows from the fact we are taking supremum in the unit ball of an RKHS, and from

the definition of the function $g(t)$. Note that

$$\begin{aligned}\mathbb{E}(nE_2^2\mathbb{1}\{B_{n,\beta}\}) &= \mathbb{E}\left(\frac{1}{n^3}\sum_{i=1}^{n}\int_0^\tau\int_0^\tau \frac{1}{S_T(t)S_T(s)}\mathfrak{K}^\star(t,s)dN_i(t)dN_i(s)\mathbb{1}_{\{B_{n,\beta}\}}\right)\\ &= \mathbb{E}\left(\frac{1}{n^3}\sum_{i=1}^{n}\int_0^\tau \frac{1}{S_T(t)^2}\mathfrak{K}^\star(t,t)dN_i(t)\mathbb{1}_{\{B_{n,\beta}\}}\right), \qquad (4.41)\end{aligned}$$

where the first line follows since $W_1,\ldots,W_n$ are i.i.d. Rademacher random variables independent from the data, and the second line follows from the fact the process $N_i(t)$ only jumps once.

We use dominated convergence in sets of large probability to show that the expectation in Equation (4.41) converges to 0. To verify the conditions of the dominated convergence theorem, we verify that the term

$$\Gamma_n = \frac{1}{n^3}\sum_{i=1}^{n}\int_0^\tau \frac{1}{S_T(t)^2}\mathfrak{K}^\star(t,t)dN_i(t)\mathbb{1}_{\{B_{n,\beta}\}} \qquad (4.42)$$

is bounded, and that it converges to zero in probability.

We first check that Γ_n is bounded. Note that by Assumption 4.6,

$$|\mathfrak{K}^\star(t,t)| \le C(Y(t)+nS_T(t))^2, \qquad (4.43)$$

where $C>0$ is some constant. Thus,

$$\Gamma_n \le \frac{C}{n}\sum_{i=1}^{n}\int_0^\tau \left(\frac{Y(t)/n}{S_T(t)}+1\right)^2 dN_i(t)\mathbb{1}_{\{B_{n,\beta}\}} \le C(\beta^{-1}+1)^2.$$

Finally, we need to check that $\Gamma_n = o_p(1)$. Let $\epsilon>0$, and let $t_\epsilon>0$ be such that $S_T(t_\epsilon)=\epsilon$. Observe that

$$\Gamma_n = \frac{1}{n^3}\sum_{i=1}^{n}\int_0^{t_\epsilon} \frac{1}{S_T(t)^2}\mathfrak{K}^\star(t,t)dN_i(t)\mathbb{1}_{\{B_{n,\beta}\}} + \frac{1}{n^3}\sum_{i=1}^{n}\int_{t_\epsilon}^\tau \frac{1}{S_T(t)^2}\mathfrak{K}^\star(t,t)dN_i(t)\mathbb{1}_{\{B_{n,\beta}\}}.$$

We prove that each of the sums in the right-hand side of the previous equation converges to zero in probability. For the first sum, observe that

$$\begin{aligned}\frac{1}{n^3}\sum_{i=1}^{n}\int_0^{t_\epsilon}\frac{1}{S_T(t)^2}\mathfrak{K}^\star(t,t)dN_i(t)\mathbb{1}_{\{B_n\}} &\leq \frac{1}{\epsilon^2}\frac{1}{n^3}\sum_{i=1}^{n}\int_0^{t_\epsilon}\mathfrak{K}^\star(t,t)dN_i(t)\mathbb{1}_{\{B_n\}}\\ &\leq \frac{1}{\epsilon^2}\frac{1}{n^3}\sum_{i=1}^{n}\mathfrak{K}^\star(T_i,T_i)\\ &= o_p(1),\end{aligned}$$

where the first inequality follows from the definition of t_ϵ, and by

$$\mathfrak{K}^\star(t,t)=\left(\sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}^2\leq 1}\sum_{j=1}^{n}\left(\omega(t,X_j)-\int_{\mathbb{R}^d}\omega(t,x)S_{T|X=x}(t)dF_X(x)\right)\right)^2\geq 0,$$

and the first equality follows from $\frac{1}{n^3}\sum_{i=1}^{n}\mathfrak{K}^\star(T_i,T_i)$ being a bounded V-statistic of order 3 (see Equation (4.40)) whose diagonal part converges to zero, and with zero mean on the off-diagonal part.

For the second integral, by Equation (4.43) and the definition of the set $B_{n,\beta}$, observe

$$\begin{aligned}\frac{1}{n^3}\sum_{i=1}^{n}\int_{t_\epsilon}^{\tau}\frac{1}{S_T(t)^2}\mathfrak{K}^\star(t,t)dN_i(t)\mathbb{1}_{\{B_{n,\beta}\}} &\leq \frac{C}{n}\sum_{i=1}^{n}\int_{t_\epsilon}^{\tau}\left(\frac{Y(t)/n}{S_T(t)}+1\right)^2 dN_i(t)\mathbb{1}_{\{B_{n,\beta}\}}\\ &\leq C(\beta^{-1}+1)^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{T_i\geq t_\epsilon\}}\\ &= O_p(1)C(\beta^{-1}+1)^2\epsilon,\end{aligned}$$

where the last equality follows from the definition of t_ϵ. Since $\epsilon>0$ can be chosen arbitrarily small, we conclude that the second integral converges to zero in probability. Thus, by the dominated convergence theorem, we conclude the desired result $E_2 = o_p(n^{-1/2})$.

4.9.4.14 Proof of Theorem 4.15

By Lemma 4.14, under Assumption 4.6,

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

where

$$h((T_i, \Delta_i, X_i), (T_j, \Delta_j, X_j)) = \Delta_i \Delta_j \bar{\mathfrak{K}}((T_i, X_i), (T_j, X_j)),$$

and $\bar{\mathfrak{K}}$ is defined in Equation (4.11) in the main document.

A simple computation shows that the kernel h is degenerate under the null hypothesis: that is,

$$\mathbb{E}(h((T_1, \Delta_1, X_1), (t, \delta, x))) = \delta \mathbb{E}(\Delta_1 \bar{\mathfrak{K}}((T_1, X_1), (t, x))) = 0$$

holds for all $(t, \delta, x) \in (0, \tau) \times \{0, 1\} \times \mathbb{R}^d$ under the null. Also, note that under Assumption 4.6, $\bar{\mathfrak{K}}$ is bounded, and thus $\mathbb{E}(|h((T_1, \Delta_1, X_1), (T_1, \Delta_1, X_1))|) < \infty$ and $\mathbb{E}(h((T_1, \Delta_1, X_1), (T_2, \Delta_2, X_2))^2) < \infty$. Then, by Theorem 4.28, there exists a distribution $\mathcal{L}$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i \Delta_j \bar{\mathfrak{K}}((T_i, X_i), (T_j, X_j)) \xrightarrow{\mathcal{D}} \mathcal{L}$$

under the null hypothesis, and for almost every sequence $(t_i, \delta_i, x_i)_{i \geq 1}$ sampled from $(T_i, \Delta_i, X_i)_{i \geq 1}$,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} W_i W_j \delta_i \delta_j \bar{\mathfrak{K}}((t_i, x_i), (t_j, x_j)) \xrightarrow{\mathcal{D}} \mathcal{L}.$$

Finally, note that by Theorem 4.9, $n\Psi_n^2$ and

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

converge to the same asymptotic distribution under the null hypothesis, which implies the desired result.

4.9.4.15 Proof of Theorem 4.16

This proof is similar to the proof of consistency of HSIC in Rindt et al. (2021). Given any $\gamma > 0$

$$\begin{aligned}\mathbb{P}(n\Psi_n^2 < Q_{n,M}^W) &= \mathbb{P}(n\Psi_n^2 < Q_{n,M}^W, n\Psi_n^2 \leq \gamma) + \mathbb{P}(n\Psi_n^2 < Q_{n,M}^W, n\Psi_n^2 > \gamma)\\ &\leq \mathbb{P}(n\Psi_n^2 \leq \gamma) + \mathbb{P}(\gamma < Q_{n,M}^W).\end{aligned}$$

The first term is dealt with by Theorem 4.11, from which we deduce $\limsup_{n\to\infty}\mathbb{P}(n\Psi_n^2 \leq \gamma) \to 0$ (since $n\Psi_n^2 \to \infty$).

For the second term, observe

$$\mathbb{P}(\gamma < Q_{n,M}^W) \leq \mathbb{P}\left(\bigcup_{m\in\{1,\ldots,M\}}\left\{\gamma < (n(\Psi_n^W)^2)_m\right\}\right) \leq M\mathbb{P}\left(\gamma < n(\Psi_n^W)^2\right).$$

By Markov's inequality

$$\mathbb{P}\left(\gamma < n(\Psi_n^W)^2\right) \leq \frac{\mathbb{E}(n(\Psi_n^W)^2)}{\gamma}.$$

Recall that

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

where $\bar{\mathfrak{K}}_n$ is the kernel defined in Theorem 4.4 in the main document. Then,

$$\begin{aligned}
&\mathbb{P}\left(\gamma < n(\Psi_n^W)^2\right) \\
&\leq \frac{1}{\gamma}\mathbb{E}(n(\Psi_n^W)^2) \\
&= \frac{1}{\gamma}\mathbb{E}\left(\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} W_i W_j \Delta_i \Delta_j \bar{\mathfrak{K}}_n((T_i, X_i), (T_j, X_j))\middle| (T_i, \Delta_i, X_i)_{i=1}^n\right)\right) \\
&= \frac{1}{\gamma}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \Delta_i \bar{\mathfrak{K}}_n((T_i, X_i), (T_i, X_i))\right),
\end{aligned}$$

since $W_1, \ldots, W_n$ are Rademacher random variables. By Assumption 4.6, we have that $\bar{\mathfrak{K}}_n((t,x),(t',x')) \leq C$ for all $(t,x),(t',x') \in \mathbb{R} \times \mathbb{R}^d$ for some constant $C > 0$, and thus

$$\mathbb{P}\left(\gamma < n(\Psi_n^W)^2\right) \leq \frac{C}{\gamma}.$$

Let $\delta > 0$. By choosing $\gamma = MC/\delta$, it holds

$$\mathbb{P}(\gamma < Q_{n,M}^W) \leq M\mathbb{P}\left(\gamma < n(\Psi_n^W)^2\right) \leq \frac{C\delta}{MC} = \delta.$$

Since δ can be chosen arbitrarily small, and since $\limsup_{n\to\infty} \mathbb{P}(n\Psi_n^2 \leq \gamma) = 0$ for any $\gamma > 0$, the result holds.

4.9.4.16 Proof of Proposition 4.17

As to the computational cost of $n\Psi_n^2$, note first that

$$\mathrm{Tr}(\boldsymbol{L}^{\Delta}(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{K}(\boldsymbol{I}-\boldsymbol{A})^{\intercal}) = \sum_{i,j=1}^{n}[\{\boldsymbol{L}^{\Delta}(\boldsymbol{I}-\boldsymbol{A})\} \circ \{(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{K}]_{ij},$$

where $\circ$ denotes the elementswise product. The matrix $\boldsymbol{AK}$ can be computed in $\mathcal{O}(n^2)$ because, after sorting the data, $\boldsymbol{A}$ is upper triangular, with the nonzero entries

equal along each row. The same observation allows for fast computation of the matrix products in $n\Psi_n^2$ and $(n\Psi_n^2)^W$.

4.9.5 Proofs auxiliary results

4.9.5.1 Proof of Proposition 4.29

Recall $\tau_x = \sup\{t : S_{T|X=x}(t) > 0\}$ and $\tau = \sup\{t : S_T(t) > 0\}$. We start by proving $\tau' \leq \tau$. Assume otherwise, $\tau < \tau'$, then there exists $\epsilon > 0$ such that $B = \{x : \tau_x > \tau + \epsilon\}$ and $F_X(B) > 0$. Let

$$B_{1/n} = \left\{ x : S_{T|X=x}(\tau + \epsilon) > \frac{1}{n} \right\},$$

and observe $B = \bigcup_{n\geq 1} B_{1/n}$. Since $F_X(B) > 0$, by the union bound, we deduce that there exists $n \geq 1$ such that $F_X(B_{1/n}) > 0$. We then arrive at the contradiction

$$0 = \int_{\mathbb{R}^d} S_{T|X=x}(\tau + \epsilon) dF_X(x) \geq \int_{B_{1/n}} \frac{1}{n} dF_X(x) \geq \frac{1}{n} F_X(B_{1/n}) > 0,$$

from which we deduce that $\tau' \leq \tau$.

We continue by proving $\tau' \geq \tau$. Assume that $\tau' < \tau$, then

$$0 < S_T(\tau') = \int_{\mathbb{R}^d} S_{T|X=x}(\tau') dF_X(x) = 0,$$

from which we deduce $\tau \leq \tau'$, and thus $\tau' = \tau$.

4.9.5.2 Proof of Proposition 4.30

Proof. Recall $d\nu_0^n(s,x) = \frac{1}{n}\sum_{j=1}^n \sum_{i=1}^n \Delta_j \delta_{T_j}(s) \frac{Y_i(s)}{Y(s)} \delta_{X_i}(x)$. Then, the result follows from proving

$$D_n = \frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n \Delta_j \omega(T_j, X_i) Y_i(T_j) \left(\frac{n}{Y(T_j)} - \frac{1}{S_T(T_j)} \right) = o_p(1).$$

Notice that D_n can be written as

$$D_n = \frac{1}{n^2} \int_0^{\tau_n} \sum_{i=1}^{n} \omega(s, X_i) Y_i(s) \left(\frac{n}{Y(s)} - \frac{1}{S_T(s)} \right) dN(s).$$

Let $\epsilon > 0$, and define $t_\epsilon \geq 0$ such that $S_T(t_\epsilon) = \epsilon$ (notice that $t_\epsilon \geq 0$ exists since the distribution over the times is continuous). We decompose D_n into two integrals, one considering integration over $\{s \leq t_\epsilon\}$ and the other one over $\{s > t_\epsilon\}$, and we prove that both integrals converge to zero in probability.

For the first integral, observe that

$$\begin{aligned} &\frac{1}{n^2} \int_0^{t_\epsilon} \sum_{i=1}^{n} \omega(s, X_i) Y_i(s) \left(\frac{n}{Y(s)} - \frac{1}{S_T(s)} \right) dN(s) \\ &\quad \leq \sup_{s \leq t_\epsilon} \left| \frac{n}{Y(s)} - \frac{1}{S_T(s)} \right| \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} |\omega(T_j, X_i)| = o_p(1), \end{aligned}$$

where the last equality follows by noticing the supremum converges to zero by Proposition 4.18.2., and that $Y_i(s) \leq 1$ and ω are bounded.

For the second integral, it holds

$$\begin{aligned} &\frac{1}{n^2} \int_{t_\epsilon}^{\tau_n} \sum_{i=1}^{n} \omega(s, X_i) Y_i(s) \left(\frac{n}{Y(s)} - \frac{1}{S_T(s)} \right) dN(s) \\ &\leq \frac{C}{n} \int_{t_\epsilon}^{\tau_n} \sum_{i=1}^{n} \frac{Y_i(s)}{Y(s)} \left| 1 - \frac{Y(s)/n}{S_T(s)} \right| dN(s) \\ &\quad = O_p(1) \frac{1}{n} \int_{t_\epsilon}^{\tau_n} dN(s) = O_p(1) \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{t_\epsilon \geq T_j\}} = O_p(1)\epsilon, \end{aligned}$$

where the second line follows from the assumption $|\omega| \leq C$ for some constant $C > 0$, the first equality in the third line is due to Proposition 4.19.1, and the last equality is due to the definition of t_ϵ. Since $\epsilon > 0$ can be chosen arbitrarily small, the result holds. □

4.9.5.3 Proof of Theorem 4.31

Proof. Consider a re-parametrisation of the kernel J (given in Equation (4.12) in the main document) in terms of the augmented space (Z, C, X), as $J : (\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^d)^2 \to \mathbb{R}$ given by

$$J((Z_1, C_1, X_1), (Z_2, C_2, X_2)) = \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} \bar{\mathfrak{K}}((s_1, X_1), (s_2, X_2)) dM_1(s_1) dM_2(s_2),$$

where $dM_1(s_1) = \mathbb{1}_{\{Z_1 \leq C_1\}} \delta_{Z_1}(s_1) - \mathbb{1}_{\{\min\{Z_1, C_1\}\} \geq s_1} d\Lambda_Z(s_1)$. Notice that to evaluate J in the previous definition we do not need extra information, it only needs (T_1, Δ_1, X_1) and (T_2, Δ_2, X_2) to be evaluated, but it is more convenient theoretically.
From the definition of the backward operator B, and Property 4.9.2.5.5 (see Section 4.9.2.5), we deduce

$$J((Z_1, C_1, X_1), (Z_2, C_2, X_2)) = (B_1 B_2 Q)((Z_1, C_1, X_1), (Z_2, C_2, X_2)), \tag{4.44}$$

where B_1 and B_2 denote the operator B applied to (Z_1, C_1, X_1) and (Z_2, C_2, X_2), respectively; the same definition holds for A_1 and A_2. Thus,

$$(T^J f)(Z_1, C_1, X_1) = \mathbb{E}_2 \left((B_1 B_2 Q)((Z_1, C_1, X_1), (Z_2, C_2, X_2)) f(Z_2, C_2, X_2) \right).$$

Using property 4.9.2.5.1, and the linearity of B_1, we find that

$$\begin{aligned}
(T^J f)(Z_1, C_1, X_1) &= \mathbb{E}_2 \left((B_1 Q)((Z_1, C_1, X_1), (Z_2, C_2, X_2)) (Af)(Z_2, C_2, X_2) \right) \\
&= B \mathbb{E}_2 \left(Q((Z_1, C_1, X_1), (Z_2, C_2, X_2)) (Af)(Z_2, C_2, X_2) \right) \\
&= B(T^Q (Af)),
\end{aligned}$$

where T^Q is the integral operator associated to the kernel Q.
Let (λ_i, f_i), with $\lambda_i \neq 0$, be an eigenpair of T^J, then we claim that (λ_i, Af_i) is a

eigenpair of T^Q. Indeed,

$$T^Q Af_i = ABT^Q Af_i = AT^J f_i = \lambda_i Af_i,$$

where in the first equality we use property 4.9.2.5.2. Also, if (λ_j, f_j) is another eigenpair such that $\langle f_i, f_j \rangle_{\mathcal{L}_2} = 0$, then

$$\langle Af_i, Af_j \rangle_{\mathcal{L}_2} = \langle f_i, BAf_j \rangle_{\mathcal{L}_2} = \langle f_i, f_j - \mathbb{E}(f_j|C_1, X_1) \rangle_{\mathcal{L}_2} = 0,$$

since $\mathbb{E}(f_j|C_1, X_1) = \frac{1}{\lambda_j}\mathbb{E}(T^J f_j|C_1, X_1) = \frac{1}{\lambda_j}\mathbb{E}(B(T^Q Af_j)|C_1, X_1) = 0$ by property 4.9.2.5.4. Hence, we conclude that for every different eigenpair (λ_i, f_i) belonging to T^J, there exists a corresponding pair (λ_i, Af_i) associated to T^Q.
Conversely for any eigenpair (λ_i', g_i) associated to T^Q, we claim that (λ_i', Bg_i) is a eigenpair for T^J. To see this, observe

$$T^J Bg_i = BT^Q ABg_i = BT^Q g_i = \lambda_i' Bg_i.$$

Similarly, if (λ_j', g_j) is another eigenpair with $\lambda_j \neq 0$, and such that $\langle g_i, g_j \rangle_{\mathcal{L}_2} = 0$, then

$$\langle Bg_i, Bg_j \rangle_{\mathcal{L}_2} = \langle g_i, ABg_j \rangle_{\mathcal{L}_2} = \langle g_i, g_j \rangle_{\mathcal{L}_2} = 0.$$

Hence, we conclude that for every different eigenpair (λ_i', g_i) belonging to T^Q, there exists a corresponding pair (λ_i', Bg_i) associated to T^Q. We conclude that T^Q and T^J have the same set of non-zero eigenvalues including multiplicities. □

4.9.5.4 Proof of Lemma 4.32

Proof. We only prove the result for $k = 1$ as the other result follows by using the same arguments.
Let $\epsilon > 0$ and choose $t_\epsilon \geq 0$ such that $S_T(t_\epsilon) = \epsilon$. We split the integral into two

integrals, one over $\{t \leq t_\epsilon\}$, and other $\{t > t_\epsilon\}$. For the integral, observe that

$$\frac{1}{n}\int_0^{t_\epsilon} \left|1 - \frac{Y(t)/n}{S_T(t)}\right| dN(t) \leq \sup_{t \leq t_\epsilon} \left|\frac{n}{Y(t)} - \frac{1}{S_T(t)}\right| \frac{1}{n}\int_0^{t_\epsilon} dN(t) = o_p(1),$$

since $\frac{1}{n}\int_0^{t_\epsilon} dN(t) \leq 1$ for all $n \geq 1$, and $\sup_{t \leq t_\epsilon} \left|\frac{n}{Y(t)} - \frac{1}{S_T(t)}\right| \to 0$ by Proposition 4.18.2. For the second integral, observe that

$$\begin{aligned}
\frac{1}{n}\int_{t_\epsilon}^{\tau} \left|1 - \frac{Y(t)/n}{S_T(t)}\right| dN(t) &= O_p(1)\frac{1}{n}\int_0^{\tau_n} \mathbb{1}_{\{t > t_\epsilon\}} dN(t) \\
&= O_p(1)\frac{1}{n}\sum_{i=1}^{n} \Delta_i \mathbb{1}_{\{T_i > t_\epsilon\}} \\
&= O_p(1)\epsilon,
\end{aligned}$$

where the first equality is due to $Y(t)/n = O_p(1)S_T(t)$ uniformly for all $t \leq \tau_n$ by Proposition 4.19.1, and the third equality follows from Markov's inequality and the definition of t_ϵ. Since $\epsilon > 0$ is arbitrary, we deduce equation (4.20). □

4.10 Statement of authorship

Statement of authorship	
Title of paper	A kernel log-rank test of independence for right-censored data
Status	Submitted and under review.
Authors	Tamara Fernández, Arthur Gretton, David Rindt, Dino Sejdinovic.
Details	`https://arxiv.org/abs/1912.03784`
Student confirmation	
Student name	David Rindt
Contribution to the paper	This work was done with Tamara Fernández, Arthur Gretton and Dino Sejdinovic. I proved convergence of the unscaled statistic Lemma 4.10, computational time of the test statistic Proposition 4.17 and consistency of the Wild-Bootstrap procedure Theorem 4.16. Proofs were later edited by Tamara Fernández. I also performed all experimental analysis, wrote all code and wrote all of Sections 4.6, 4.7 and Appendix A Additional experiments, as well as parts of Section 4.5 and smaller parts throughout the paper. The colon dataset was proposed by Tamara Fernández.
Date	1 June 2021
Signature	
Supervisor confirmation	
By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.	
Supervisor name and title	Dino Sejdinovic, Associate Professor of Statistics
Supervisor comments	All above seems accurate and fair.
Date	22 June 2021
Signature	

Chapter 5

Proper scoring rules, the right-censored log-likelihood and partially monotonic neural networks for survival regression

This chapter is based on the following work.

*David Rindt**, Robert Hu*, David Steinsaltz, Dino Sejdinovic.
Proper scoring rules, the right-censored log-likelihood and partially monotonic neural networks for survival regression.
`https://arxiv.org/abs/2103.14755`.

* denotes equal contribution.

The contribution of this paper is two-fold. First, we show that popular scoring rules for right-censored regression models, including time-dependent concordance and (integrated) Brier score, have the undesirable property that the true distribution may score worse than less accurate distributions. In contrast to these scores, for the right-censored log-likelihood the true distribution achieves the optimal expected value. Unfortunately, existing neural-network based regression models do not have a tractable

log-likelihood, and this can thus not be used to assess model fit. Our second contribution is therefore to propose a novel regression method which, unlike the existing methods, optimizes a tractable right-censored likelihood. The method, termed SuMo-net, models the survival function and the density of the event-time jointly using partially monotonic neural networks. SuMo-net furthermore evaluates the survival function through a simple forward pass through a neural network, achieves a speedup factor of $20 - 100$ compared to the most similar competing method and, achieves competitive performance in terms of likelihood and existing metrics.

5.1 Introduction

Survival regression is the task of learning a time-to-event distribution from a sample of covariates and associated times. In this work, we consider the case in which the event-times are right-censored. This means that for some individuals it is known their event happened after a certain time, but not exactly when. Applications of survival regression are numerous. A classical example is the study of the recovery time from medical treatment (Laurie et al. 1989). Applications in other contexts include predicting the failure times of mechanical systems (Susto et al. 2015), modeling the time until the default of a business (Dirick et al. 2017), and churn prediction, where the time a customer is lost to the business is modeled (Van den Poel and Lariviere 2004). Right-censoring would then mean that a patient may not have recovered by the time the study ended, or that a business may not have defaulted during the period of observation.

Classical models for survival regression include the Cox proportional hazards model (Cox 1972) and the accelerated failure time (Kalbfleisch and Prentice 2011). Over the last decade many machine learning methods for survival data have been proposed. While an exhaustive list is impossible to give, we mention the most relevant methods

which allow more flexible survival functions than the Cox model. In Katzman et al. (2018); Kvamme et al. (2019), the Cox model is extended by replacing the linear term of the Cox model with a neural network. In Ishwaran et al. (2008) a random-forest approach is taken, in which cumulative hazard functions are computed using the Nelson-Aalen estimate. In Lee et al. (2018) a discrete distribution is learned, by optimizing a mix of the discrete likelihood and a rank-based score. Ranganath et al. (2016) and Miscouridou et al. (2018) first learn a latent representation of the covariates and then learn a survival function using the latent representation. In particular, the latter method can deal with missing data in the context of electronic health records. In Chapfuwa et al. (2018) an adversarial model is used for survival regression.

None of the modern methods mentioned above have a tractable likelihood. In particular, Kvamme et al. (2019); Katzman et al. (2018); Ishwaran et al. (2008) all learn discrete distributions, consisting of point masses at fixed times. In Lee et al. (2018) time is discretized arbitrarily into bins and in Ranganath et al. (2016); Miscouridou et al. (2018) the posterior distribution is intractable. In each case computing the likelihood of new observations is problematic.

Consequently, whereas for uncensored data the likelihood is a standard measure of model fit, for right-censored data one typically resorts to alternative measures. The most commonly used scores to evaluate the fit of a right-censored regression model are the time-dependent concordance (Antolini et al. 2005), the Brier score for right-censored data (Graf et al. 1999) and the binomial log-likelihood. More recently, in Avati et al. (2020), an alternative scoring rule, the survival-CRPS was proposed.

The contribution of this work is two-fold. First, we illustrate the limitations of existing measures of model fit, other than the right-censored log-likelihood, by showing that inaccurate distributions may achieve better scores than the true distribution. This shows that these measures, which include time-dependent concordance, (inte-

grated) Brier score and binomial log-likelihood, are not proper scoring rules. For time-dependent concordance and the survival-CRPS, this is a new result.

In contrast to the above scores, the right-censored log-likelihood is a proper score. Therefore our second contribution is to propose SuMo-net, a novel survival regression method that optimizes the right-censored likelihood. SuMo-net uses partially monotonic neural networks to jointly model the survival function and the density of the event-time. We show this technique is 20-100 times faster than the most similar method, evaluates the survival function through a single forward pass, and performs well, both in terms of the existing scores and in terms of likelihood.

5.2 Background and notation

We are interested in an event-time $T \in \mathbb{R}_{\geq 0}$ and how it depends on a covariate $X \in \mathbb{R}^p$ for $p \geq 1$. For simplicity, we focus on vector inputs here, but the framework can be straightforwardly extended to non-Euclidean or structured domains, such as images or graphs, as long as these can be fed into a neural network. Importantly, we consider the case in which T is subject to right-censoring. In such data, the event-time T is not known for every individual in the sample. Instead of observing T directly, we observe $Z = \min\{T, C\}$ where $C \in \mathbb{R}_{\geq 0}$ is a censoring time, as well as the indicator $D = 1\{Z = T\}$ indicating whether we observe the event-time. A sample of size n can thus be denoted by $((X_i, Z_i, D_i))_{i=1}^n$, where we use the uppercase letters when we treat the dataset as random and lowercase letters otherwise. We let $S(t|x) \coloneqq 1 - F(t|x) = \mathbb{P}(T > t|X = x)$ denote the (conditional) survival distribution, and, assuming T has a density, we let $f(t|x) \coloneqq \frac{d}{dt}F(t|x)$ denote the density of the event-time. In some equations we will write F_x for $F(\cdot|x)$, and similarly for S and f. The goal of this work is the following: If we let $S(t|x)$ denote the survival function of T, our aim is to estimate S with $\hat{S}(t|x)$ based on the sample $((X_i, Z_i, D_i))_{i=1}^n$. We

also look at ways to assess if $\hat{S}$ is an accurate estimate of the true distribution S.
In survival analysis one typically assumes that censoring is not informative about the event-time conditional on the covariate, which we formally state in the following assumption.

Assumption 5.1. *(Independent censoring) We assume that, conditionally on the covariate, the censoring- and event-time are independent. That is, we assume* $T \perp\!\!\!\perp C|X$.

Let $S(t|x)$ be a survival function and let $f(t|x) = -\partial S(t|x)/\partial t$. Then under Assumption 5.1 the right-censored log-likelihood is defined as:

$$\log L_n = \sum_{i=1}^{n} d_i \log f(z_i|x_i) + (1 - d_i) \log S(z_i|x_i). \tag{5.1}$$

The appearance of both the density- and the survival function imply that, if one models the density, then the likelihood contains $\mathcal{O}(n)$ integrals, making it a difficult optimization objective (see e.g. p. 288 of Gu (2013)). For that reason, many methods, in particular all methods based on the Cox model, optimize the partial likelihood instead of the right-censored likelihood ((Cox 1972)). Finally, we point out that for unconstrained densities, the maximum likelihood on the training data is infinite (i.e. the nonparametric MLE does not exist). For large scale parametric models (such as neural network parametrizations of the density) on the other hand, the maximum likelihood is still well defined.

5.3 Evaluation criteria and proper scoring rules

A scoring rule $\mathcal{S}$ takes a distribution S over some set $\mathcal{Y}$ and an observed value $y \in \mathcal{Y}$ and returns a score $\mathcal{S}(S, y)$, assessing how well the model predicts the observed value. For a positive scoring rule, a higher score indicates a better model fit. The scoring

rule is called proper if the true distribution achieves the optimal expected score, i.e. if

$$\mathbb{E}_{y\sim S}\mathcal{S}(S,y) \geq \mathbb{E}_{y\sim S}\mathcal{S}(\hat{S},y) \qquad \text{for all distributions} \quad \hat{S}. \tag{5.2}$$

In the context of survival regression we call the scoring rule $\mathcal{S}$ proper if for every true distribution S, every censoring distribution C and for every covariate x it holds that

$$\mathbb{E}_{T,C|X=x}\mathcal{S}(S_x,(Z,D)) \geq \mathbb{E}_{T,C|X=x}\mathcal{S}(\hat{S}_x,(Z,D)) \tag{5.3}$$

for every distribution $\hat{S}_x$, where in the above expression $T|X=x \sim S_x$. By taking the expectation with respect to X in the above expression, we find that if $\mathcal{S}$ is a proper score, then also

$$\mathbb{E}_{T,C,X}\mathcal{S}(S_X,(Z,D)) \geq \mathbb{E}_{T,C,X}\mathcal{S}(\hat{S}_X,(Z,D)) \tag{5.4}$$

for every family of distributions $\hat{S}_x$ with $x \in \mathbb{R}^p$. In Section 5.3.1, we show that the right-censored log-likelihood is a proper score. However, since most existing regression models do not have a tractable likelihood, the three most popular measures of model-fit are the integrated Brier score, integrated binomial log-likelihood and time-dependent concordance. In Section 5.3.2 we show that the true model is not guaranteed to achieve the optimal score for any of these measures, and hence none of them is a proper scoring rule.

5.3.1 The right-censored log-likelihood as a proper scoring rule

The right-censored log-likelihood score of an observation (z, d), given a covariate x and distribution S_x is simply the likelihood given in Equation 5.1, i.e. :

$$\text{log-likelihood}(S_x, (z, d)) = d \log f(z|x) + (1 - d) \log S(z|x). \tag{5.5}$$

Let $\hat{S}$ denote any continuous survival curve, and let S denote the true survival distribution of the event-time T, and make the following assumption.

Assumption 5.2. *Assume that $KL(S||\hat{S}) < \infty$, where KL denotes the Kullback-Leibler divergence.*

Under this assumption, we prove the following lemma.

Lemma 5.3 (See also Diks et al. (2011))**.** *Let S be the true survival distribution of the event-time T. For every distribution $\hat{S}$ that satisfies Assumption 5.2 and for every $x \in \mathbb{R}^p$, it holds that*

$$\mathbb{E}\left[\textit{log-likelihood}(S_x, (Z, D))|X = x\right] \geq \mathbb{E}[\textit{log-likelihood}(\hat{S}_x, (Z, D))|X = x] \tag{5.6}$$

Hence, under Assumption 5.2, the right-censored log-likelihood is a proper scoring rule. We provide a proof in the Appendix, as a simple extension of Lemma 1 of Diks et al. (2011), by conditioning on the censoring variable and the covariate. This result is also stated in Section 3.5 of Dawid and Musio (2014), when one uses the function $\psi(p) = p \log p - p$.

5.3.2 Existing evaluation metrics are not proper scoring rules

We now show that time-dependent concordance (Antolini et al. 2005), the Brier score for right-censored data (Graf et al. 1999), the binomial log-likelihood and the survival-CRPS (Avati et al. 2020) are not proper scoring rules. For time-dependent concordance and the survival-CRPS, this is a novel result.

Time-dependent concordance Let (X, Z, D) and (X', Z', D') be i.i.d. covariates and right-censored event-times. Time-dependent concordance C^{td} (Antolini et al. 2005) is defined by

$$\mathbb{P}(\hat{S}(Z|X) < \hat{S}(Z|X')|Z \leq Z', D = 1) + \frac{1}{2}\mathbb{P}(\hat{S}(Z|X) = \hat{S}(Z|X')|Z \leq Z', D = 1)$$

where ties in the survival probabilities are dealt with as proposed in Ishwaran et al. (2008) (Step 3 of Section 5.1). It has been remarked that poorly calibrated models can still have a high time-dependent concordance (Kvamme et al. 2019). In Rizopoulos et al. (2014); Antolini et al. (2005) it has been pointed out that censoring affects time-dependent concordance. We now show that the concordance of the true distribution can be much lower than the concordance of less accurate distributions, a property we have not found to be reported explicitly in previous literature. We demonstrate this by giving an explicit example of a true distribution and a false distribution, so that the false distribution has higher concordance when estimated on a sample from the true distribution.

Example: optimizing concordance for a binary covariate. Let $X, X' \sim \text{Bernoulli}(1/2)$

independently. Then $\hat{S}(\cdot|\cdot)$ maximizes C^{td} if and only if it maximizes

$$\begin{aligned}
&\mathbb{P}(\hat{S}(Z|X) < \hat{S}(Z|X'), Z \leq Z', D = 1|X = 1, X' = 0) \\
&+ \mathbb{P}(\hat{S}(Z|X) < \hat{S}(Z|X'), Z \leq Z', D = 1|X = 0, X' = 1) \\
&= \int_0^\infty \big(1\{\hat{S}(t|1) < \hat{S}(t|0)\} f_T(t|1) S_C(t|1) S_T(t|0) S_C(t|0) \\
&\qquad + 1\{\hat{S}(t|0) < \hat{S}(t|1)\} f_T(t|0) S_C(t|0) S_T(t|1) S_C(t|1)\big) dt.
\end{aligned}$$

It is clear that this expression is optimized by constructing $\hat{S}(\cdot|\cdot)$ such that $\hat{S}(t|1) < \hat{S}(t|0)$ if and only if

$$\begin{aligned}
f_T(t|1) S_C(t|1) S_T(t|0) S_C(t|0) &\geq f_T(t|0) S_C(t|0) S_T(t|1) S_C(t|1) \\
\iff \lambda_T(t|1) &\geq \lambda_T(t|0)
\end{aligned}$$

where $\lambda_T(t|x) = f_T(t|x)/S_T(t|x)$ is called the hazard rate. Hence, in this example, to maximize concordance we need the ordering of $\hat{S}(t|1)$ and $\hat{S}(t|0)$ to be the reversed order of the true hazard rates. In Figure 5.1, we use this insight to generate a false distribution with concordance much higher than the concordance of the true distribution.

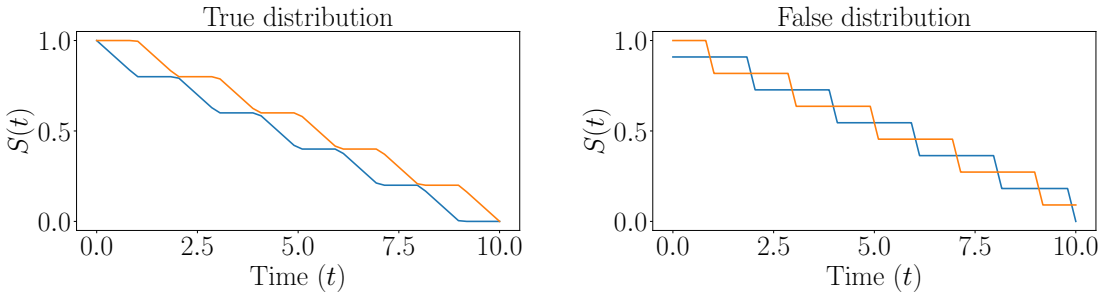


Figure 5.1: The survival curves for the true distribution on the left and a false distribution, designed to optimize concordance, on the right. The two colors, orange and blue, represent the survival distribution given $X = 0$ and $X = 1$ respectively. By drawing a sample of size 1000 from the true distribution, we find that the concordance score of the true distribution is 0.57 and of the false distribution is 0.75, confirming that an inaccurate distribution can have high concordance.

Brier score The Brier score at time t for the estimated survival distribution $\hat{S}(\cdot|\cdot)$ is typically defined as (Graf et al. 1999)

$$\text{BS}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{S}(t|X_i)^2 1\{Z_i \leq t, D_i = 1\}}{\hat{G}(Z_i)} + \frac{(1 - \hat{S}(t|X_i))^2 1\{Z_i > t\}}{\hat{G}(t)}$$

where $\hat{G}$ is the Kaplan-Meier estimate of the censoring distribution. Assuming that the estimated survival function of the censoring time C, $\hat{G}(\cdot)$ is the exact marginal distribution $G(\cdot) = \mathbb{P}(C > \cdot)$, then the expectation, conditional on the X_i equals

$$\begin{aligned}
&\mathbb{E}\left(\text{BS}(t)|X_1, \ldots X_n\right) \\
=&\frac{1}{n}\sum_{i=1}^{n} \hat{S}(t|X_i)^2 \int_0^t \frac{G(z|X_i)}{G(z)} f_T(z|X_i)dz + (1 - \hat{S}(t|X_i))^2 \frac{G(t|X_i)}{G(t)} S_T(t|X_i). \qquad (5.7)
\end{aligned}$$

If $C \perp\!\!\!\perp X$, and $G(\cdot|x) = G(\cdot)$, for all x, then it is easy to see the above equals

$$\frac{1}{n}\sum_{i=1}^{n} \hat{S}(t|X_i)^2(1 - S_T(t|X_i)) + (1 - \hat{S}(t|X_i))^2 S_T(t|X_i).$$

By setting $a_i = \hat{S}(t|X_i)$ and optimizing with respect to each of the a_i, we find that the expected Brier score is minimized for $\hat{S}(t|X_i) = S_T(t|X_i)$, i.e. when the estimated survival probabilities equal the true survival probabilities. *Under the assumption of independent censoring and a perfectly estimated censoring distribution,* the Brier score is thus a proper score. However, as Equation 5.7 makes clear, the Brier score may not be a proper score when $C \not\perp\!\!\!\perp X$. The problematic assumption on the censoring distribution has, for example, been reported in Kvamme and Borgan (2019). We give an example of a scenario in which an inaccurate distribution achieves a lower (i.e. better) Brier score than the true distribution in the Appendix.

Integrated Brier score and integrated binomial log-likelihood The Brier score above was defined for a fixed time t. To this end, the integrated Brier score (IBS) has been defined as a measure of goodness-of-fit of the estimate $\hat{S}(\cdot|\cdot)$.

$$\text{IBS} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \text{BS}(t) dt$$

where typically $t_1 = 0$ and $t_2 = z_{\max} = \max\{z_1, \ldots, z_n\}$. In experiments, the integral is approximated by a sum over an equally spaced grid of times.

The binomial log-likelihood at time t, $\text{BLL}(t)$ is defined by:

$$\text{BLL}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{\log(1 - \hat{S}(t|X_i)) 1\{Z_i \leq t, D_i = 1\}}{\hat{G}(Z_i)} + \frac{S(t|X_i) 1\{Z_i \geq t\}}{\hat{G}(t)}.$$

Analogously to the definition of IBS, the integrated binomial log-likelihood, IBLL, is defined as

$$\text{IBLL} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \text{BLL}(t) dt.$$

The exact analysis of the Brier score can be applied to the BLL. In particular, the BLL is a proper score under the assumption that $C \perp\!\!\!\perp X$, but it need not be proper when this assumption is violated.

Survival-CRPS The survival continuous ranked probability score (survival-CRPS) is defined by

$$\mathcal{S}_{\text{CRPS}}(\hat{F}_X, (Z, D)) = \int_0^Z \hat{F}(t|X)^2 dt + D \int_Z^\infty (1 - \hat{F}(t|X))^2 dt$$

and was proposed in Avati et al. (2020) as an extension of the continuous ranked probability score (CRPS) (Gneiting et al. 2007). This is equivalent to the integrated Brier score without inverse probability weighting. In Avati et al. (2020), the score was also claimed to be a proper score. The proof strategy given there argues that the survival-CRPS is simply the weighted CRPS of Gneiting and Ranjan (2011) with the weight function set to the indicator of uncensored regions. As we show in the Appendix, however, this weight function does not recover the survival-CRPS. This suggests that the survival-CRPS is not a proper score. If the censoring time C is deterministic, i.e. $C = c$ w.p. 1 for $c \in \mathbb{R}_{\geq 0}$, then it is clear that to minimize the survival-CRPS, one would set $\hat{F}(c) = 1$, even if the true CDF is less than 1. This principle remains in the case of random censoring, where one can construct false CDFs which yield a lower (i.e. better) survival-CRPS. In the Appendix we give a simple example where this is the case.

5.3.3 Summary

In summary, we believe that optimizing time-dependent concordance and survival-CRPS could lead to inaccurate models, as these are not proper scoring rules. Although we believe IBS and IBLL to give reasonable approximations of model accuracy, there is no guarantee the most accurate distribution achieves the optimal score when cen-

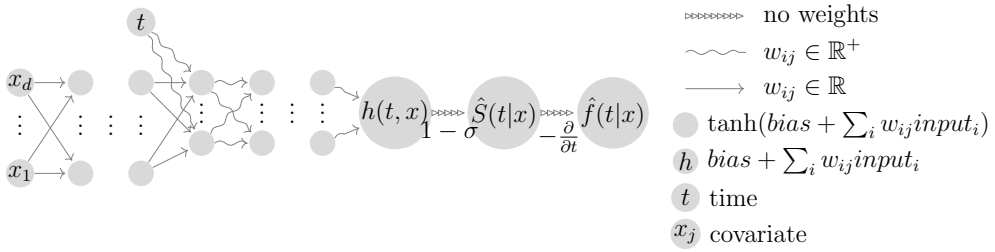


Figure 5.2: [Figure adapted from Figure 1 of Chilinski and Silva (2020)] The graph representing SuMo-net, modelling the survival curve. The last edge symbolizes the operation of differentiating the conditional survival distribution function $\hat{S}(t|x)$ with respect to the input t, resulting in the conditional density of the survival time. The second-to-last edge represents passing $h(t,x)$ through $1-\text{sigmoid}$. If one models the cumulative hazard instead, we replace $1-\text{sigmoid}$ by SoftRelu and use $\frac{\partial}{\partial t}$ in the last edge instead of $-\frac{\partial}{\partial t}$.

soring depends on the covariate. Since for the right-censored likelihood the true distribution achieves the optimal score, regardless of censoring, we believe it would be advantageous to have a model with a tractable likelihood, and, in particular, a natural thought is to find the model that optimizes the right-censored likelihood.

5.4 Optimizing likelihood using partially monotonic neural networks

To optimize the right-censored log-likelihood, one needs to jointly model the survival distribution and the density. To do so we adapt the monotonic neural density estimator (MONDE) of Chilinski and Silva (2020) in the case of a univariate response variable to the survival context. The MONDE was originally proposed to estimate densities in the case of uncensored data. While the MONDE architecture is not a novel contribution, adapting it to jointly model the survival function and the density function of a right-censored event-time is, to the best of our knowledge, new. We now provide a brief overview of how we adapt the architecture of the MONDE estimator to the survival context.

Figure 5.2, adapted from Figure 1 of Chilinski and Silva (2020), pictures the SuMo-network structure as a directed graph with two types of edges. The network can be defined in terms of two fully connected feed forward sub-networks. A covariate x is first passed through a network $h^{\text{cov}}(x)$ with L^{cov} layers in which each layer is fully connected to the next, and no other edges are present. The output of the covariate network, together with the time input t forms the input of a second network $h^{\text{mixed}}(t, u)$, where u has the same dimension as the output of the covariate network $h^{\text{cov}}(x)$. In this network, too, all edges are present from layer l to $l + 1$ for $l = 1, \ldots, L^{\text{mixed}} - 1$ and no additional edges are added. The output layer of the mixed network consists of a single neuron with value $h(t, x)$, thus returning a scalar. The total network h thus equals:

$$h(t, x) = h^{\text{mixed}}(t, h^{\text{cov}}(x)). \tag{5.8}$$

We then model $S(t|x)$ through

$$\hat{S}(t|x) = 1 - \text{sigmoid}\left(h(t, x)\right). \tag{5.9}$$

We also investigated modeling the cumulative hazard function $\Lambda(t|x)$ through $\hat{\Lambda}(t|x) = \text{SoftReLU}(h(t, x))$ where $\text{SoftReLU}(u) = \log\left(1 + \exp(u)\right)$ but found that this did not improve performance.

After passing $h(t, x)$ through the sigmoid function, the final edge in Figure 5.2 represents differentiation of $\text{sigmoid}(h(t, x))$ with respect to t and we set $\hat{f}(t|x) = -\partial\hat{S}(t|x)/\partial t$. To compute the derivative, we use automatic differentiation as implemented in PyTorch (Paszke et al. 2019), in line with the approach taken in Chilinski and Silva (2020). One can also approximate the derivative by applying the chain rule

to sigmoid$(h(t,x))$, yielding that

$$f(t|x) = -\operatorname{sigmoid}(h(t,x))(1-\operatorname{sigmoid}(h(t,x))\left(\frac{h(t+\epsilon,x)-h(t,x)}{\epsilon}\right)+o(1).$$

We implemented SuMo-net both using this approximation and using auto-grad and found the implementations performed comparably during the experiments. We found further that both of these implementations improved performance over simply evaluating $-\left(S(t+\epsilon|x)-S(t|x)\right)/\epsilon$, which may be due to the vanishing gradient of the sigmoid function as one moves away from 0.

Since the survival function and the cumulative hazard function are non-increasing and non-decreasing in t respectively, the output $h(t,x)$ is enforced to be non-decreasing in t. As discussed in Lang (2005), non-decreasing behaviour in t is guaranteed if all weights that are descendants of t in Figure 5.2 are non-negative. In Lang (2005) it is furthermore shown that if there are at least 2 hidden layers in the mixed network $h_{\text{mixed}}(t,u)$, then, given enough nodes in each layer, the network can universally approximate any true survival function. In our implementation, we ensure that weights are non-negative by defining them as the square of unrestricted weights.

5.5 Experiments

We apply SuMo-net to five real datasets SUPPORT, METABRIC, GBSG, FLCHAIN, and KKBOX presented in Kvamme et al. (2019). We compare the results with the methods DeepHit (Lee et al. 2018), Cox-Time and Cox-CC (Kvamme et al. 2019), DeepSurv (Katzman et al. 2018) and the classical Cox model (Cox 1972). We choose these methods since easy-to-use code was available in the PyCox library. We use a 5-fold cross-validation procedure in which the model is trained on three folds, one fold is used for validation, and one fold is reserved for testing. We then average the evaluation criteria calculated on five different test folds. For datasets FLCHAIN,

Dataset	Size	Covariates	Unique Durations	Prop. Censored
SUPPORT	8,873	14	1,714	0.32
METABRIC	1,904	9	1,686	0.42
Rot. & GBSG	2,232	7	1,230	0.43
FLCHAIN	6,524	8	2,715	0.70
KKBOX	2,646,746	15	2,344,004	0.28
Weibull	25,000	1	25000	0.55
Checkerboard	25,000	1	25000	0.75
Normal	25,000	1	25000	0.50

Table 5.1: Summary of all datasets used in the experiments. Weibull, checkerboard and normal are synthetic datasets and the others are real datasets. See Kvamme et al. (2019) for more details on the real datasets.

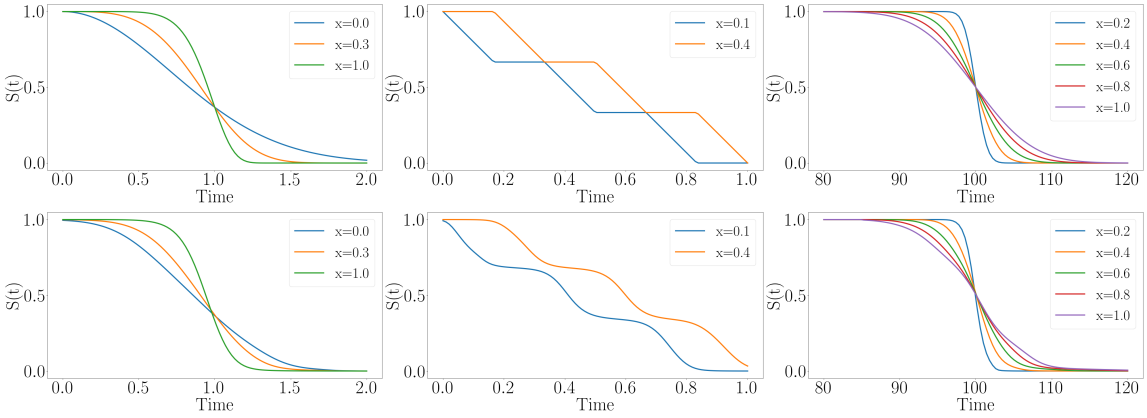


Figure 5.3: True distributions (top row) and learned distributions using SuMo-net (bottom row). From left to right, the distributions are Weibull, Checkboard, and Normal.

GBSG, METABRIC, and SUPPORT, we optimize SuMo-net and baselines by using the hyperopt library (Bergstra et al. 2013) for 300 iterations. For the baselines, we replicate the setup used in Kvamme et al. (2019). For the KKBOX dataset, we do the same but for 100 hyperopt iterations due to computational constraints. All experiments were run on a GPU cluster consisting of 8 Nvidia 1080 GTX TI cards.

5.5.1 Results

In Table 5.2 we compare the methods in terms of right-censored likelihood. As the benchmarks do not have tractable right-censored likelihood, we approximate it by in-

terpolating their survival curves (See Section 5.5.3 for details of this interpolation). In Table 5.3 we compare the binomial log-likelihood of the models. While, as discussed in this paper, there is no guarantee the most accurate model has the highest score, an advantage is that we do not need to approximate the score for the benchmarks. Results for IBS and time-dependent concordance are given in the Appendix. Notably, we find that DeepHit, which optimizes a concordance-like loss function, generally has the best performance in terms of time-dependent concordance, yet achieves the worst scores in terms of likelihood. This agrees with our findings that time-dependent concordance is not a proper scoring rule and that optimizing time-dependent concordance may lead to inaccurate distributions.

Dataset Method	Likelihood FLCHAIN	GBSG	KKBOX	METABRIC	SUPPORT
CoxCC	-0.507 ± 0.014	-0.501 ± 0.038	0.797 ± 0.013	-0.305 ± 0.037	0.533 ± 0.047
Cox-Linear	-0.513 ± 0.014	-0.498 ± 0.04	0.706 ± 0.01	-0.316 ± 0.031	0.545 ± 0.03
Cox-Time	-0.584 ± 0.239	-0.454 ± 0.034	1.411 ± 0.014	-0.233 ± 0.066	0.57 ± 0.038
DeepHit	-0.85 ± 0.493	-0.719 ± 0.19	-4.174 ± 4.128	-0.565 ± 0.348	0.098 ± 0.141
DeepSurv	-0.529 ± 0.015	-0.502 ± 0.04	0.758 ± 0.036	-0.318 ± 0.029	0.549 ± 0.045
SuMo-net	$\mathbf{-0.378} \pm 0.003$	$\mathbf{-0.391} \pm 0.027$	$\mathbf{1.553} \pm 0.132$	$\mathbf{-0.163} \pm 0.048$	$\mathbf{0.64} \pm 0.046$

Table 5.2: Likelihood scores and standard deviations over the 5 folds. Higher is better. Only for SuMo-net we have the exact likelihood. For other methods we have approximated the likelihood. In particular, we approximate the density by interpolating the survival curve. See Section 5.5.3 for details.

Dataset Method	IBLL FLCHAIN	GBSG	KKBOX	METABRIC	SUPPORT
CoxCC	-0.464 ± 0.194	-0.546 ± 0.02	-0.382 ± 0.005	-0.53 ± 0.035	-0.572 ± 0.008
Cox-Linear	-0.336 ± 0.012	-0.549 ± 0.011	-0.429 ± 0.001	-0.522 ± 0.012	-0.577 ± 0.005
Cox-Time	-0.47 ± 0.241	-0.541 ± 0.021	$\mathbf{-0.33} \pm 0.004$	-0.526 ± 0.012	-0.577 ± 0.009
DeepHit	-0.646 ± 0.462	-1.157 ± 1.287	-0.745 ± 0.513	-0.656 ± 0.156	-0.701 ± 0.065
DeepSurv	-0.339 ± 0.01	-0.539 ± 0.013	-0.368 ± 0.002	-0.52 ± 0.013	$\mathbf{-0.569} \pm 0.01$
SuMo-net	$\mathbf{-0.333} \pm 0.011$	$\mathbf{-0.528} \pm 0.012$	-0.402 ± 0.059	$\mathbf{-0.49} \pm 0.014$	-0.573 ± 0.008

Table 5.3: IBLL scores and standard deviations over the 5 folds. Higher is better

5.5.2 Correlation between likelihood and existing metrics

As shown in Section 3 likelihood is the only proper score. In this section, we investigate how the other scores compare to likelihood by calculating the rank correlation between all scores. In particular, for each of the datasets FLCHAIN, GBSG, METABRIC, and SUPPORT and each hyperparameter run of SuMo-net, we compute all scores on the validation set and turn them into ranks, which are then concatenated for the 4 datasets. The correlation matrix between these ranks is displayed in table 5.4. Sample sizes are given in Table 5.1. The validation set is one fifth of the total dataset.

	Likelihood	Concordance	IBS	IBLL
Likelihood	1.000	0.031	0.683	0.873
Concordance	0.031	1.000	0.031	0.041
IBS	0.683	0.031	1.000	0.837
IBLL	0.873	0.041	0.837	1.000

Table 5.4: Rank correlation matrix between likelihood, concordance, IBS and IBLL.

We see that concordance is almost independent of likelihood while IBS is correlated with likelihood. IBLL exhibits the strongest correlation, suggesting that, despite it not being a proper score, it serves as a good proxy when the likelihood is not available.

5.5.3 Estimation of the likelihood for models other than SuMo-net

All the baseline models proposed in Kvamme et al. (2019), learn discrete distributions and hence have no density and the likelihood cannot be computed. For those methods we use interpolation of the survival curve to approximate the density. Let the survival curve of the discrete distribution be denoted by S and denote the points at which the survival curve makes jumps by $t_0 \leq \cdots \leq t_m$. Then, given a time z, we find i so that

$t_i \leq z < t_{i+1}$ and then estimate

$$\hat{f}(t) = -\frac{S(t_{i+k}) - S(t_{i-k+1})}{t_{i+k} - t_{i-k+1}}$$

where k determines the width of the interval we use to approximate the derivative. We found that setting $k = 1$ resulted in a highly variable density and low likelihood, and hence we use $k = 2$ in the main text. SuMo-net has tractable likelihood, so this approximation does not have to be made.

5.5.4 Inference timings

Here we comment on the timing of predictions of the survival probability $S(t|x)$ using a trained model. We compare the time of SuMo-net with the time of Cox-Time presented in Kvamme et al. (2019), which is arguably the most similar to SuMo-net: it can model every type of dependence between time and covariates, and, although it learns a discrete distribution, it treats the input times as continuous. To compute the survival probability, Cox-Time integrates the learned hazard function over time, which is computationally costly. In contrast, SuMo-net evaluates S through a single forward pass. To compare the timings we initialize SuMo-net and Cox-Time using an identical architecture and compare inference timings on the test-fold of all datasets in Table 5.5. The findings indicate SuMo-net does indeed achieve its anticipated speedup.

	Cox-Time	SuMo-net	Speedup factor
SUPPORT	2.97 ± 0.126	0.047 ± 0.004	63
METABRIC	1.573 ± 0.126	0.016 ± 0.001	100
GBSG	1.46 ± 0.112	0.021 ± 0.002	69
FLCHAIN	4.199 ± 0.246	0.044 ± 0.003	96
KKBOX	140.505 ± 7.976	7.034 ± 1.092	20

Table 5.5: Inference timings on 20% of the data. SuMo-net achieves a speedup factor of between 20 and 100.

5.6 Discussion

We have shown that several existing scoring rules for right-censored regression models are not proper, i.e. they do not necessarily reflect the accuracy of a fitted distribution. We also proved that the right-censored log-likelihood is a proper scoring rule. While other scoring rules can still be useful, their limitations should be taken into consideration when used. Since a tractable likelihood is attractive to assess model fit, we proposed a simple, fast and competitive regression method for survival data that uses partially monotonic neural networks and optimizes the right-censored likelihood.

5.7 Supplementary material

5.7.1 Proof of Lemma 5.3

We need to prove that for every $x \in \mathbb{R}^p$, for every distribution S_x

$$\mathbb{E}\left[\mathcal{S}(S_x, (Z, D))|X = x\right] \geq \mathbb{E}[\mathcal{S}(\hat{S}_x, (Z, D))|X = x]$$

where $T \sim S_x$, the censoring time C follows an arbitrary distribution and it is assumed that S_x and $\hat{S}_x$ satisfy Assumption 1. We can safely omit x from the notation and prove that

$$\mathbb{E}\mathcal{S}(S, (Z, D)) \geq \mathbb{E}\mathcal{S}(\hat{S}, (Z, D))$$

where again $T \sim S$, the censoring time C follows an arbitrary distribution and it is assumed that S and $\hat{S}$ satisfy Assumption 1.

This has been proved in a slightly different context in Diks et al. (2011) Lemma 1.

In that context the censoring variable is assumed known. The same argument can be applied to the case of random censoring, by first conditioning on the censoring variable. Namely, if we can prove that

$$\mathbb{E}\left[\mathcal{S}(S,(Z,D))|C\right] \geq \mathbb{E}[\mathcal{S}(\hat{S},(Z,D))|C]$$

then it follows, by taking the expectation over C, that also

$$\mathbb{E}\mathcal{S}(S,(Z,D)) \geq \mathbb{E}\mathcal{S}(\hat{S},(Z,D)).$$

We now prove the first inequality following the proof of Lemma 1 in Diks et al. (2011). Let f denote the density of S, and let $\hat{f}$ denote the density of $\hat{S}$. Note that the expected score can be written as follows:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{S}(\hat{S},(Z,D))|C\right] &= \mathbb{E}\left[D\log\hat{f}(Z) + (1-D)\log\hat{S}(Z)\middle|C\right] \\
&= \int_0^\infty 1\{t \leq C\}f(t)\log\hat{f}(t) + 1\{t > C\}f(t)\log\hat{S}(C)dt \\
&= \int_0^C f(t)\log\hat{f}(t)dt + S(C)\log\hat{S}(C).
\end{aligned}$$

Hence,

$$
\begin{aligned}
&\mathbb{E}\big[\mathcal{S}(S,(Z,D)) - \mathcal{S}(\hat{S},(Z,D)) \;\big|\; C\big] \\
&\qquad = \int_0^C f(t)\log\frac{f(t)}{\hat{f}(t)}dt + S(C)\log\frac{S(C)}{\hat{S}(C)} \\
&\qquad = F(C)\int_0^C \frac{f(t)}{F(C)}\log\frac{f(t)/F(C)}{\hat{f}(t)/\hat{F}(C)}dt \\
&\qquad\quad + S(C)\log\frac{S(C)}{\hat{S}(C)} - F(C)\int_0^C \frac{f(t)}{F(C)}\log\frac{1/F(C)}{1/\hat{F}(C)}dt \\
&\qquad = F(C)\int_0^C \frac{f(t)}{F(C)}\log\frac{f(t)/F(C)}{\hat{f}(t)/\hat{F}(C)}dt \\
&\qquad\quad + S(C)\log\frac{S(C)}{\hat{S}(C)} + F(C)\log\frac{F(C)}{\hat{F}(C)} \\
&\qquad = F(C)\mathrm{KL}\big(f(t)/F(C)\,\|\,\hat{f}(t)/\hat{F}(C)\big) + \mathrm{KL}\big(\mathrm{Ber}(S(C))\,\|\,\mathrm{Ber}(\hat{S}(C))\big) \\
&\qquad \geq 0.
\end{aligned}
$$

Here $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. The distributions $f(t)/F(C)$ and $\hat{f}(t)/\hat{F}(C)$ are the densities of f and $\hat{f}$ conditioned on the event $T \leq C$. The first KL divergence exists by Assumption 1. The second term is the KL divergence between Bernoulli random variables with success probabilities $S(C)$ and $\hat{S}(C)$ respectively.

5.7.2 Details on time-dependent concordance example

Let $X \in \{0,1\}$. Let $K \sim \mathrm{Uniform}\{0,2,4,6,8\}$ and let $U \sim \mathrm{Uniform}[0,1]$. Then $T|X=0 \sim K+U$ and $T|X=1 \sim K+U+1$. We further set $C \sim \mathrm{Exponential}(\mathrm{mean} = 20)$. Note that the hazard rates of the two distributions are nonzero on disjoint sets. The False distribution in Figure 5.1, is constructed such that at each time, the distribution with the nonzero hazard rate has the lowest survival probability.

5.7.3 Details on Brier score example

Let $X \sim \text{Bernoulli}(0.1)$. We create an extreme example of unequal censoring in which $C \sim \text{Exp}(1)$ if $X = 0$ and $C = \infty$ if $X = 1$. Let $T \sim \text{Exponential}(10)$ (independent of X), so that $F_{\text{True}}(t) = 1 - \exp(-t/10)$. Finally create a fake distribution $\hat{F}$ so that $\hat{F}(t|X = 1) = F_{\text{true}}(t)$ and $\hat{F}(t|X = 0) = 1 - \exp(-t/4)$. Then based on a sample of size $n = 10^6$, the Brier score at time $t = 4$ is 0.22 for the true distribution and 0.19 for the false distribution $\hat{F}$. The integrated Brier score, from $t = 0$ to $t = 40$ equals 0.044 for the true distribution and 0.041 for the false distribution.

5.7.4 Discussion of survival-CRPS

5.7.4.1 Discussion of proof given in Avati et al

In Gneiting and Ranjan (2011) it is stated that the following score is proper (see Matheson and Winkler (1976) for a proof),

$$\begin{aligned}\mathcal{S}(F, z) &= \int_0^\infty (F(r) - 1\{r \geq z\})^2 u(r) dr \\ &= \int_0^z F(r)^2 u(r) dz + \int_z^\infty (1 - F(r))^2 u(r) dr.\end{aligned}$$

for non-negative weight functions u. In Avati et al. (2020) it is argued that the survival-CRPS is simply the above score, with weight function equal to the uncensored region. It is not immediately clear how this is meant to be interpreted, as the weight function needs to be a non-random function. If it were a proper score when conditioned on a known censoring time $C = c$, for every value of c, then we could proceed as in Section 5.3.1, using the conditional independence of C and T. Assuming $C = c$, the uncensored region equals $u(r) = 1\{r \leq c\}$. Proceeding with this as the

weight function we would obtain

$$\begin{aligned} S(F,z) &= \int_0^\infty (F(r) - 1\{r \geq z\})^2 1\{r \leq c\} dr \\ &= \int_0^z F(r)^2 1\{r \leq c\} dz + \int_z^\infty (1 - F(r))^2 1\{r \leq c\} dr. \end{aligned}$$

Note that $z \leq c$ since $z = \min(t, c)$ and therefore the first indicator is always true and can be removed, to arrive at:

$$\begin{aligned} S(F,z) &= \int_0^z F(r)^2 dr + \int_z^\infty (1 - F(r))^2 1\{r \leq c\} dr \\ &= \int_0^z F(r)^2 dr + \int_z^c (1 - F(r))^2 dr. \end{aligned}$$

Hence, the survival-CRPS has an additional term

$$d \int_c^\infty (1 - F(r))^2 dr \tag{5.10}$$

which is missing from the above expression. This shows, at least, that this interpretation of the proof strategy would not succeed. Furthermore, assuming that c is a fixed censoring time, then $F(r)$ is only evaluated for $r > c$ in Equation 5.10. Clearly to minimize the survival-CRPS, one then sets $F(r) = 1$ for $r \geq c$, even if the true distribution satisfies $F_{\text{true}}(r) < 1$ for $r \geq c$.

5.7.4.2 An example

Assume for simplicity that there are no covariates. Let $T \sim \text{Exp}(\text{mean} = 100)$ with every individual set to fail at $t = 200$ so that $F_{True}(t) = 1 - \exp(-t/100)$ for $t < 200$ and $F_{\text{true}}(200) = 1$. Let $C \sim \text{Exp}(\text{mean} = 10)$. Observations are thus likely to be censored. Let a fake distribution be $F_{\text{False}} = \text{Exp}(\text{mean} = 25)$. If we take a sample of size 10^6, from F_{True} under censoring C, and compute the survival-CRPS in Python,

we find the survival-CRPS of the true distribution is 3.87, while the survival-CRPS of the false distribution is 1.67. When the inverse probability weighting is added and the integrated Brier score is computed we find that the true distribution achieves the optimal score in this example.

5.7.5 Likelihood scores for $d = 1, 4, 8$

dataset Method	likelihood FLCHAIN	GBSG	KKBOX	METABRIC	SUPPORT
CoxCC	-0.583 ± 0.014	-0.632 ± 0.032	0.156 ± 0.075	-0.451 ± 0.06	0.437 ± 0.06
Cox-Linear	-0.586 ± 0.026	-0.603 ± 0.032	0.05 ± 0.111	-0.444 ± 0.03	0.451 ± 0.053
Cox-Time	-0.565 ± 0.071	-0.569 ± 0.063	0.926 ± 0.09	-0.346 ± 0.056	0.481 ± 0.047
DeepHit	-0.878 ± 0.496	-0.805 ± 0.212	-1.432 ± 0.226	-0.686 ± 0.5	0.04 ± 0.159
DeepSurv	-0.605 ± 0.01	-0.609 ± 0.04	0.179 ± 0.073	-0.455 ± 0.027	0.446 ± 0.036
SuMo-net	$\mathbf{-0.378} \pm 0.003$	$\mathbf{-0.391} \pm 0.027$	$\mathbf{1.553} \pm 0.132$	$\mathbf{-0.163} \pm 0.048$	$\mathbf{0.64} \pm 0.046$

Table 5.6: Likelihood scores for $d = 1$

Dataset Method	Likelihood FLCHAIN	GBSG	KKBOX	METABRIC	SUPPORT
CoxCC	-0.482 ± 0.01	-0.451 ± 0.033	0.181 ± 0.088	-0.286 ± 0.054	0.564 ± 0.04
Cox-Linear	-0.49 ± 0.017	-0.465 ± 0.043	0.095 ± 0.113	-0.265 ± 0.02	0.565 ± 0.032
Cox-Time	-0.496 ± 0.166	-0.409 ± 0.043	0.93 ± 0.089	-0.189 ± 0.061	0.6 ± 0.041
DeepHit	-0.846 ± 0.497	-0.94 ± 0.415	-4.454 ± 4.441	-0.56 ± 0.338	0.113 ± 0.126
DeepSurv	-0.498 ± 0.015	-0.457 ± 0.021	0.158 ± 0.117	-0.282 ± 0.027	0.562 ± 0.029
SuMo-net	$\mathbf{-0.378} \pm 0.003$	$\mathbf{-0.391} \pm 0.027$	$\mathbf{1.553} \pm 0.132$	$\mathbf{-0.163} \pm 0.048$	$\mathbf{0.64} \pm 0.046$

Table 5.7: Likelihood scores for $d = 4$

Dataset Method	Likelihood FLCHAIN	GBSG	KKBOX	METABRIC	SUPPORT
CoxCC	-0.485 ± 0.014	-0.451 ± 0.033	0.42 ± 0.021	-0.243 ± 0.023	0.568 ± 0.036
Cox-Linear	-0.477 ± 0.012	-0.442 ± 0.033	0.344 ± 0.037	-0.253 ± 0.012	0.572 ± 0.037
Cox-Time	-0.55 ± 0.178	-0.398 ± 0.036	0.985 ± 0.037	-0.179 ± 0.064	0.599 ± 0.043
DeepHit	-0.835 ± 0.491	-0.637 ± 0.138	-3.622 ± 3.38	-0.552 ± 0.299	0.086 ± 0.134
DeepSurv	-0.495 ± 0.01	-0.435 ± 0.025	0.414 ± 0.04	-0.249 ± 0.018	0.573 ± 0.042
SuMo-net	$\mathbf{-0.378} \pm 0.003$	$\mathbf{-0.391} \pm 0.027$	$\mathbf{1.553} \pm 0.132$	$\mathbf{-0.163} \pm 0.048$	$\mathbf{0.64} \pm 0.046$

Table 5.8: Likelihood scores for $d = 8$

5.7.6 Other experiments

Dataset Method	C^{td} FLCHAIN	GBSG	KKBOX	METABRIC	SUPPORT
CoxCC	0.792 ± 0.005	0.672 ± 0.014	0.832 ± 0.001	0.642 ± 0.03	0.608 ± 0.011
Cox-Linear	0.792 ± 0.006	0.66 ± 0.014	0.795 ± 0.001	0.633 ± 0.022	0.596 ± 0.012
Cox-Time	0.792 ± 0.005	0.678 ± 0.017	$\mathbf{0.848} \pm 0.008$	0.649 ± 0.026	0.613 ± 0.007
DeepHit	$\mathbf{0.796} \pm 0.004$	$\mathbf{0.683} \pm 0.013$	0.799 ± 0.005	$\mathbf{0.692} \pm 0.022$	$\mathbf{0.642} \pm 0.003$
DeepSurv	0.791 ± 0.004	0.671 ± 0.014	0.828 ± 0.001	0.645 ± 0.02	0.609 ± 0.011
SuMo-net	0.79 ± 0.004	0.671 ± 0.013	0.76 ± 0.045	0.658 ± 0.029	0.603 ± 0.004

Table 5.9: Concordance scores. Higher is better

Dataset Method	IBS FLCHAIN	GBSG	KKBOX	METABRIC	SUPPORT
CoxCC	0.112 ± 0.014	0.185 ± 0.006	0.118 ± 0.001	0.179 ± 0.013	$\mathbf{0.194} \pm 0.003$
Cox-Linear	0.102 ± 0.004	0.188 ± 0.004	0.138 ± 0.0	0.176 ± 0.005	0.197 ± 0.002
Cox-Time	0.114 ± 0.017	0.183 ± 0.008	$\mathbf{0.104} \pm 0.001$	0.178 ± 0.005	0.196 ± 0.004
DeepHit	0.162 ± 0.079	0.227 ± 0.064	0.16 ± 0.012	0.21 ± 0.037	0.227 ± 0.019
DeepSurv	0.103 ± 0.004	0.183 ± 0.005	0.116 ± 0.001	0.176 ± 0.005	$\mathbf{0.194} \pm 0.004$
SuMo-net	$\mathbf{0.101} \pm 0.004$	$\mathbf{0.179} \pm 0.005$	0.13 ± 0.023	$\mathbf{0.164} \pm 0.005$	0.196 ± 0.003

Table 5.10: IBS scores. Lower is better

5.7.7 Implementation Details

We present the implementation details to SuMo-net, including network architecture parameters, optimizer parameters and the code base.

5.7.7.1 Hyperparameter space

Hyperparameter	Values
Layers	{1, 2}
Layers (Covariate part)	{1, 2}
Nodes per layer	{8,16,32}
Nodes per layer (Covariate part)	{8,16,32}
Dropout	[0.0,0.1,0.2,0.3,0.4,0.5]
Weigh decay	{0.4, 0.2, 0.1, 0.05, 0.02, 0.01, 0}
Batch size	{5,10,25,50,100,250}

Table 5.11: Hyperparameter search space for experiments on Rot. & GBSG, SUPPORT, METABRIC, and FLCHAIN.

Hyperparameter	Values
Layers	{1, 2,4}
Layers (Covariate part)	{4, 6, 8}
Nodes per layer	{8, 16, 32,64}
Nodes per layer (Covariate part)	{128, 256, 512}
Dropout	[0, 0.7]
Weigh decay	{0.4, 0.2, 0.1, 0.05, 0.02, 0.01, 0}
Batch size	{1000,2500,5000}

Table 5.12: KKBox hyperparameter configurations. (*) denotes parameters found with a two layer network with 128 nodes.

For exact details, we refer to the code base attached.

5.7.7.2 Toy datasets

We provide here the distributions of the toy datasets. In each of the distributions $X \sim \text{Unif}[0, 1]$.

Dataset	$T\|X$	$C\|X$
Weibull	Weib(shape $= 2 + 6X$)	Exp(mean $= 1.5$)
Normal	N$(100, 6X)$	N$(100, 6)$
Checkerboard	CB$(4, 6\|X)$	Exp(mean $= 1.5$)

Table 5.13: Distributions of the toy datasets Weibull, Normal and Checkerboard. Weib(shape $= s$) refers to the Weibull distribution with shape s and scale 1. For the definition of CB$(4, 6)$ see below. Parameters are chosen such that appropriate numbers of individuals is censored (see data description main text) and the survival curves of the Normal and Weibull distributions cross (see plots main text).

The distribution CB$(4, 6|X)$ is defined as follows. Let X be on the horizontal axis and let T be on the vertical axis. We split the square $[0, 1]^2$ up in a grid of 6×4 equally large rectangles, where there are 4 columns and 6 rows. If X is in the 1st or 3rd column, then T is distributed uniformly over the 1st, 3rd, or 5th row. If X is in the 2nd or 4th column, then T is distributed uniformly over the 2nd, 4th or 6th row.

5.7.7.3 Code Base

An anonymized version of the code base can be found here[1].

[1]https://anonymous.4open.science/r/Sumo-Net-9833/README.md

5.8 Statement of authorship

Statement of authorship	
Title of paper	Proper scoring rules, the right-censored log-likelihood and partially monotonic neural networks for survival regression
Status	Submitted and under review.
Authors	David Rindt, Robert Hu, David Steinsaltz, Dino Sejdinovic.
Details	`arxiv.org/abs/2103.14755`
Student confirmation	
Student name	David Rindt
Contribution to the paper	I am the primary author of Sections 5.1-4 and 5.6. I initiated the project and proposed to use partially monotonic neural networks for survival regression. I also performed the analysis of the scores. I only had a minor role in the implementation of the method, the benchmarks, or the running of all experiments, which was mainly done by Robert Hu.
Date	1 June 2021
Signature	
Supervisor confirmation	
By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.	
Supervisor name and title	Dino Sejdinovic, Associate Professor of Statistics
Supervisor comments	All above seems accurate and fair.
Date	22 June 2021
Signature	

Chapter 6

Conclusion

In this thesis we have explored assumption-free methods for independence testing and regression between a covariate and a right-censored event-time. In Chapter 2, we studied the permutation test with the Hilbert-Schmidt independence criterion (HSIC) as the test statistic, as proposed in Gretton et al. (2008). We showed that, when the kernel used is characteristic, this testing procedure is point-wise consistent. That is, for each fixed alternative hypothesis, the probability that the test rejects the null hypothesis converges to 1 as the sample size converges to infinity. Since this is true for any fixed number of permutations, we also investigated how one may choose the number of permutations in practice.

In Chapter 3, we proposed optHSIC, a kernel-based test of independence between a covariate and a right-censored time. The test consists of two steps: first, the original data, which includes both censored and observed event-times, is transformed into a dataset in which all times are observed. Second, a permutation test with test statistic HSIC is applied to the resulting uncensored dataset. We showed in a range of experiments that this testing procedure is able to detect a much wider range of dependencies than the Cox likelihood ratio test, without losing much power when the Cox proportional hazards assumption holds true. We also showed experimentally that the type 1 error rate achieves the correct level, even in the challenging case in which the censoring time depends on the covariate.

In Chapter 4, we proposed the kernel log-rank test. Similar to optHSIC, the kernel log-

rank test procedure tests the independence between a covariate and a right-censored event-time. The test comes with a more developed theoretical framework than the optHSIC test of Chapter 3. In particular the kernel log-rank statistic can be understood in two ways: either as a supremum of scores over functions in a Reproducing Kernel Hilbert Space (RKHS), or as the RKHS distance between the mean embeddings of two appropriately defined probability distributions. The test is implemented using a wild-bootstrap procedure, which we showed to be point-wise consistent and to have point-wise asymptotically correct level.

In Chapter 5, we discussed scores for right-censored regression models and proposed SuMo-net, a method for learning the survival distribution conditional on a covariate. Our contribution in this chapter was two-fold. First, we showed that several existing scoring rules are not proper, which means they have the undesirable property that inaccurate distributions may score better than the true distribution. Since by contrast the right-censored log-likelihood is a proper score, we secondly proposed a method termed SuMo-net to optimize the right-censored log-likelihood. SuMo-net uses partially monotonic neural networks to model the survival function and the density jointly. This regression procedure is relatively easy to implement and is able to model every type of relationship between covariates and the event-time. SuMo-net furthermore achieves competitive performance in a range of experiments. Finally, the method also allows for non-Euclidean covariates, as long as these can be entered into a neural network.

A key feature of the independence tests developed in this thesis is that they have the ability to detect a wider range of dependencies than the Cox proportional hazards likelihood ratio test. In practice this could help to test for dependencies in which the effect of the covariate on the event-time is nonlinear, or in which the hazard rate does

not factorize into a function of covariates and a function of time. Additionally, the proposed tests enable testing of independence of the event-time and non-Euclidean covariates, such as images, strings or graphs. Using the proposed tests in this setting would be an exciting application of the work done in this thesis.

A further interesting application of the kernel log-rank test of Chapter 4 lies in feature selection for supervised learning. Feature selection is the task of selecting the subdimensions of the covariates that are most relevant to the target variable of interest. For the case of an uncensored target variable, in Song et al. (2012) it was proposed to select features by maximizing the dependence between features and the response variable as measured through the unbiased HSIC estimate. This can be done either through forward or backward selection, which iteratively add features to or remove features from the set of selected features. While this idea could be readily applied when the response variable is a right-censored event-time by replacing HSIC with the kernel log-rank statistic, it would need to be studied whether the bias of the kernel log-rank statistic has to be accounted for and what the effect of censoring is on the selected features.

Finally, it is not uncommon to make repeated measurements of the covariates until the event of interest occurs. For example, medical tests may be done regularly over time, and one may test if the measurement results affect event-times associated to a patient. Such longitudinal covariates, however, cannot currently be included in the methods proposed in this thesis, as there would be a trivial dependence between the number of measurements and the event-time and censoring-time. Likewise, while in this thesis we considered only right-censored event-times, other types of censoring exist. In particular for left- and interval-censored event-times, it is known the event happened before a fixed time-point and within two time-points respectively. Alter-

natively, the covariates may not be observed exactly, or there may be missing entries in the data. Further developing the proposed methods to allow for such types of missing information, and for longitudinal covariates would be an interesting direction for future work.

Bibliography

O. Aalen. Nonparametric estimation of partial transition probabilities in multiple decrement models. *Ann. Statist.*, 6(3):534–545, 1978. ISSN 0090-5364. URL http://links.jstor.org/sici?sici=0090-5364(197805)6:3<534:NEOPTP>2.0.CO;2-1&origin=MSN.

M. Albert, B. Laurent, A. Marrel, and A. Meynaoui. Adaptive test of independence based on HSIC measures. *arXiv preprint arXiv:1902.06441*, 2021.

L. Antolini, P. Boracchi, and E. Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

A. Avati, T. Duan, S. Zhou, K. Jung, N. H. Shah, and A. Y. Ng. Countdown regression: sharp and calibrated survival predictions. In *Uncertainty in Artificial Intelligence*, pages 145–155. PMLR, 2020.

V. Bagdonavičius, J. Kruopis, and M. S. Nikulin. *Non-parametric tests for censored data*. ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, 2011. ISBN 978-1-84821-289-3.

J. Bergstra, D. Yamins, and D. D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page I–115–I–123. JMLR.org, 2013.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004. ISBN 1-4020-7679-7. doi: 10.1007/978-1-4419-9096-9. URL https://doi.org/10.1007/978-1-4419-9096-9. With a preface by Persi Diaconis.

T. Berrett and R. Samworth. Nonparametric independence testing via mutual information. *Biometrika*, 106(3):547–566, 2019.

T. B. Berrett, I. Kontoyiannis, and R. J. Samworth. Optimal rates for independence testing via u-statistic permutation tests. *arXiv preprint arXiv:2001.05513*, 2020.

C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

K. M. Borgwardt and H.-P. Kriegel. Shortest-path kernels on graphs. In *Fifth IEEE international conference on data mining (ICDM'05)*, pages 8–pp. IEEE, 2005.

N. E. Breslow. Discussion of the paper by dr cox. *Journal of the Royal Statistical Society, Series B*, 34:216–217, 1972.

P. Chapfuwa, C. Tao, C. Li, C. Page, B. Goldstein, L. C. Duke, and R. Henao. Adversarial time-to-event modeling. In *International Conference on Machine Learning*, pages 735–744. PMLR, 2018.

P. Chilinski and R. Silva. Neural likelihoods via cumulative distribution functions. In *Conference on Uncertainty in Artificial Intelligence*, pages 420–429. PMLR, 2020.

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

S. Cohen. *Finding color and shape patterns in images*. Number 1620. Stanford University, Department of Computer Science, 1999.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

A. P. Dawid and M. Musio. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014.

H. Dehling and T. Mikosch. Random quadratic forms and the bootstrap for U-statistics. *J. Multivariate Anal.*, 51(2):392–413, 1994. ISSN 0047-259X. doi: 10.1006/jmva.1994.1069. URL `https://doi.org/10.1006/jmva.1994.1069`.

D.-M. Denk and A. Kaider. Videoendoscopic biofeedback: a simple method to improve the efficacy of swallowing rehabilitation of patients after head and neck surgery. *ORL*, 59(2):100–105, 1997.

C. Diks, V. Panchenko, and D. Van Dijk. Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230, 2011.

L. Dirick, G. Claeskens, and B. Baesens. Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, 68(6):652–665, 2017.

M. Ditzhaus and S. Friedrich. More powerful logrank permutation tests for two-sample survival data. *arXiv preprint arXiv:1807.05504*, 2018.

D. Dunkler, M. Ploner, M. Schemper, and G. Heinze. Weighted cox regression using the R package coxphw. *Journal of Statistical Software, Articles*, 84(2):1–26, 2018. ISSN 1548-7660. doi: 10.18637/jss.v084.i02. URL `https://www.jstatsoft.org/v084/i02`.

B. Efron and I. M. Johnstone. Fisher's information in terms of the hazard rate. *Ann. Statist.*, 18(1):38–62, 1990. ISSN 0090-5364. doi: 10.1214/aos/1176347492. URL `https://doi.org/10.1214/aos/1176347492`.

T. Fernández and N. Rivera. Kaplan-Meier V- and U-statistics. *Electron. J. Stat.*, 14 (1):1872–1916, 2020. doi: 10.1214/20-EJS1704. URL `https://doi.org/10.1214/20-EJS1704`.

T. Fernández, N. Rivera, and Y. W. Teh. Gaussian processes for survival analysis. *Advances in Neural Information Processing Systems*, 29:5021–5029, 2016.

T. Fernández and N. Rivera. A reproducing kernel hilbert space log-rank test for the two-sample problem. *Scandinavian Journal of Statistics*, 2021. doi: https://doi.org/10.1111/sjos.12496. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12496`.

T. R. Fleming and D. 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. ISBN 0-471-52218-X.

K. Fukumizu, B. K. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic kernels on groups and semigroups. In *NIPS*, pages 473–480, 2008.

V. Garès, S. Andrieu, J.-F. Dupuy, and N. Savy. An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials. *Stat. Med.*, 34(4): 541–557, 2015. ISSN 0277-6715. doi: 10.1002/sim.6366. URL `https://doi.org/10.1002/sim.6366`.

R. Gill. 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`.

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

T. Gneiting and R. Ranjan. Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29 (3):411–422, 2011.

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951, 1992. ISSN 01621459. URL `http://www.jstor.org/stable/2290630`.

A. Gretton. A simpler condition for consistency of a kernel independence test. *arXiv preprint arXiv:1501.06103*, 2015.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 585–592, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, pages 585–592, 2008.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two–sample test. *Journal of Machine Learning Research*, 12:723–773, 2012a.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012b. ISSN 1532-4435.

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances*

in Neural Information Processing Systems, volume 25, pages 1205–1213. Curran Associates, Inc., 2012c. URL https://proceedings.neurips.cc/paper/2012/file/dbe272bab69f8e13f14b405e038deb64-Paper.pdf.

C. Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

D. P. Harrington and T. R. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566, 1982. ISSN 0006-3444. doi: 10.1093/biomet/69.3.553. URL https://doi.org/10.1093/biomet/69.3.553.

W. Hoeffding. A non-parametric test of independence. *The annals of mathematical statistics*, pages 546–557, 1948.

W. Hoeffding. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, pages 169–192, 1952.

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

J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

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

I. Kim, S. Balakrishnan, and L. Wasserman. Minimax optimality of permutation tests. *arXiv preprint arXiv:2003.13208*, 2020.

G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

W. Kössler. Max-type rank tests, U-tests, and adaptive tests for the two-sample location problem—an asymptotic power study. *Comput. Statist. Data Anal.*, 54(9): 2053–2065, 2010. ISSN 0167-9473. doi: 10.1016/j.csda.2010.03.014. URL https://doi.org/10.1016/j.csda.2010.03.014.

H. Kvamme and Ø. Borgan. The brier score under administrative censoring: Problems and solutions. *arXiv preprint arXiv:1912.08581*, 2019.

H. Kvamme, Ø. Borgan, and I. Scheel. Time-to-event prediction with neural networks and Cox regression. *Journal of machine learning research*, 20(129):1–30, 2019.

B. Lang. Monotonic multi-layer perceptron networks as universal approximators. In *International conference on artificial neural networks*, pages 31–37. Springer, 2005.

J. A. Laurie, C. G. Moertel, T. R. Fleming, H. S. Wieand, J. E. Leigh, J. Rubin, G. W. McCormack, J. B. Gerstner, J. E. Krook, and J. Malliard. Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. the north central cancer treatment group and the mayo clinic. *Journal of Clinical Oncology*, 7(10):1447–1456, 1989.

C. T. Le, P. M. Grambsch, and T. A. Louis. Association between survival time and ordinal covariates. *Biometrics*, 50(1):213–219, 1994. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/2533211`.

C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *AAAI*, pages 2314–2321, 2018.

H. Li and Y. Luan. Kernel cox regression models for linking gene expression profiles to censored survival data. In *Biocomputing 2003*, pages 65–76. World Scientific, 2002.

S. Lip, L. E. Tan, P. Jeemon, L. McCallum, A. F. Dominiczak, and S. Padmanabhan. Diastolic blood pressure j-curve phenomenon in a tertiary-care hypertension clinic. *Hypertension*, 74(4):767–775, 2019.

F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *ICML '20: Proceedings of the 37th International Conference on Machine Learning*, pages 6316–6326, 2020.

R. Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5): 3284–3305, 2013.

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

M. Matabuena. Energy distance and kernel mean embedding for two sample survival test. *arXiv preprint arXiv:1901.00833*, 2019.

J. E. Matheson and R. L. Winkler. Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096, 1976.

I. W. McKeague, A. M. Nikabadze, and Y. Q. Sun. An omnibus test for independence of a survival time from a covariate. *Ann. Statist.*, 23(2):450–475, 1995. ISSN 0090-5364. doi: 10.1214/aos/1176324530. URL https://doi.org/10.1214/aos/1176324530.

X. Miscouridou, A. Perotte, N. Elhadad, and R. Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pages 244–256. PMLR, 2018.

C. G. Moertel, T. R. Fleming, J. S. Macdonald, D. G. Haller, J. A. Laurie, C. M. Tangen, J. S. Ungerleider, W. A. Emerson, D. C. Tormey, J. H. Glick, et al. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage iii colon carcinoma: a final report. *Annals of internal medicine*, 122(5):321–326, 1995.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. ISSN 02643952. URL http://www.jstor.org/stable/91247.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135(2):185–207, 1972. ISSN 00359238. URL http://www.jstor.org/stable/2344317.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

R. Ranganath, A. Perotte, N. Elhadad, and D. Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pages 101–114. PMLR, 2016.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

D. Rindt, D. Sejdinovic, and D. Steinsaltz. A kernel-and optimal transport-based test of independence between covariates and right-censored lifetimes. *The International Journal of Biostatistics*, 2020.

D. Rindt, D. Sejdinovic, and D. Steinsaltz. Consistency of permutation tests of independence using distance covariance, hsic and dhsic. *Stat*, page e364, 2021.

D. Rizopoulos, L. A. Hatfield, B. P. Carlin, and J. J. Takkenberg. Combining dynamic predictions from joint models for longitudinal and time-to-event data using bayesian model averaging. *Journal of the American Statistical Association*, 109(508):1385–1397, 2014.

J. P. Romano. Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, pages 141–159, 1989.

B. Schölkopf and A. Smola. *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, Massachusetts, 1 edition, 2001.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals Of Statistics*, 41:2263–2291, 2012.

R. J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, Inc., New York, 1980. ISBN 0-471-02403-1. Wiley Series in Probability and Mathematical Statistics.

S. Shirdhonkar and D. W. Jacobs. Approximate earth mover's distance in linear time. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *The Journal of Machine Learning Research*, 13(1):1393–1434, 2012.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, 12: 2389–2410, 2011. ISSN 1532-4435.

G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi. Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, 11(3):812–820, 2015.

D. J. 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 *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Z. Szabó and B. K. Sriperumbudur. Characteristic and universal tensor product kernels. *The Journal of Machine Learning Research*, 18(1):8724–8752, 2017.

G. Szekély and M. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3:1236–1265, 2009.

R. E. Tarone. On the distribution of the maximum of the logrank statistic and the modified wilcoxon statistic. *Biometrics*, 37(1):79–85, 1981. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/2530524`.

T. M. Therneau and T. Lumley. Package 'survival'. *R Top Doc*, 128(10):28–33, 2015.

D. Van den Poel and B. Lariviere. Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, 157 (1):196–217, 2004.

R. Wang, S. Lagakos, and R. Gray. Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring. *Biostatistics*, 11.4:676–692, 2010.

A. Zajacova and S. A. Burgard. Shape of the bmi-mortality association by cause of death, using generalized additive models: Nhis 1986–2006. *Journal of aging and health*, 24(2):191–211, 2012.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Stat. Comput.*, 28(1):113–130, 2018. ISSN 0960-3174. doi: 10.1007/s11222-016-9721-7. URL `https://doi.org/10.1007/s11222-016-9721-7`.

D. M. Zucker and A. F. Karr. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Statist.*, 18(1):329–353, 1990. ISSN 0090-5364. doi: 10.1214/aos/1176347503. URL `https://doi.org/10.1214/aos/1176347503`.