A Appendix

A.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 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 $U \sim \mathrm{Uniform}(\mathrm{AR})$ and $U' = \mathrm{Uniform}(\mathrm{L})$. Since it holds that $\mathrm{AR} = \mathrm{L}$, it follows that $P_{UU'}$ is the matrix:

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

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 $\tilde{x} \leftarrow 1.3 = x_1$ with probability 1. We update $\tilde{D} \leftarrow [(1.3, 13)]$. We also replace $\mathrm{L} \leftarrow [0.5, 0.3, -1.1, -0.9]$ and $\mathrm{AR} \leftarrow [0.5, 0.3, -1.1, -0.9]$. We move to the next event time.

At $z_2 = 22$ we note that $\delta_2 = 0$, so we only remove $x_2 = 0.5$ from AR and update $\mathrm{AR} \leftarrow [0.3, -1.1, -0.9]$, while leaving L and $\tilde{D}$ unchanged.

At the third event $z_3 = 24$ it holds that $\delta_3 = 1$ and $\mathrm{AR} = [0.3, -1.1, -0.9]$ and $\mathrm{L} = [0.5, 0.3, -1.1, -0.9]$. We couple a random variable $U \sim \mathrm{Uniform}(\mathrm{AR})$ and $U' = \mathrm{Uniform}(\mathrm{L})$ 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 $\mathrm{L} \leftarrow [0.3, -1.1, -0.9]$ and $\mathrm{AR} \leftarrow [-1.1, -0.9]$ before moving to the next event.

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

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

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 $\mathrm{L} \leftarrow [0.3, -0.9]$ and $\mathrm{AR} \leftarrow [-0.9]$.

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

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

A.2 Proof of Lemma 4.1

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 $k := |\{i : \delta_i = 1\}|$ 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, since they are determined by D, but the way in which they are paired up in the transformation T may be random. 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 $h : \{1,\dots,k\} \to \{1,\dots,n\}$, 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 $\Pi_{1:h(i)-1} = [\Pi_1,\dots,\Pi_{h(i)-1}]$. 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 4.1 are both equal in distribution to

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

This is done in separate lemmas.

Lemma A.1.

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

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

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

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

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

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

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

□

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

Lemma A.2. *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 $\Sigma(\tilde{D}) := T(\Pi(D))$. 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 $[n] \setminus \{\sigma_{1:i-1}\}$ 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 $[n] \setminus \{\sigma_{1:i-1}\}$. That is, Σ_i is chosen uniformly from $[n] \setminus \{\sigma_{1:i-1}\}$. Mathematically, for any $\sigma, \pi \in S_n$ so that the conditioning event has nonzero probability, it holds that

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

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

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

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

Lemma A.3.

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

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.

A.3 Proof of Theorem 4.1

The proof of Theorem 5.2 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 the transformation of the data. By Lemma 4.1 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.

A.4 Proof of Theorem 5.1

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 $\{0, 1, \ldots, B\}$

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

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 $i \geq 1$.

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

It follows that the vector

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

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.

□

A.5 Proof of Theorem 5.2

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 5.1 applies.

□

A.6 Proof of Lemma 5.1

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

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

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

A.7 Proof of Lemma 5.2

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.

A.8 Proof of Theorem 5.3

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:

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

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

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

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. □

A.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 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 $\beta = 3\alpha/4$.

A.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

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

A.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

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

A.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 unkown, 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.

A.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 M_A 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 ν_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 satisfies the three conditions above, with β in the third condition. 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).

A.10 Tables

A.10.1 Type 1 error rates

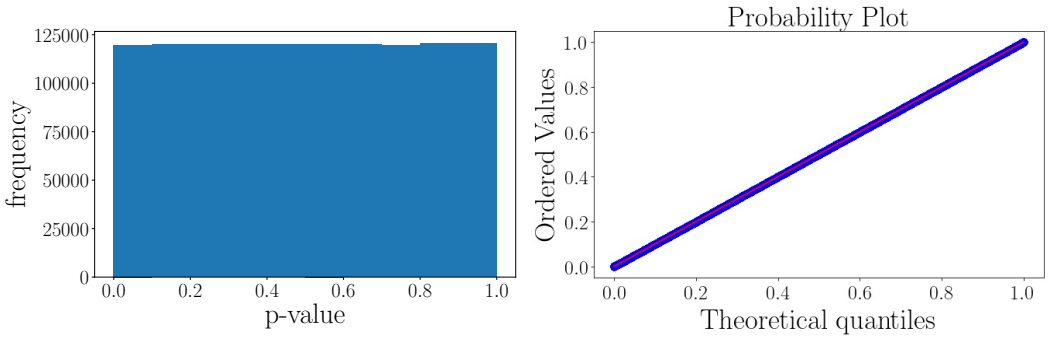


Figure 11: A histogram and a qq-plot of the p-values obtained from optHSIC for distribution D.6 of Table 2, in which $C \not\perp\!\!\!\perp X$, with a sample size of $n = 200$. Data was sampled 1.2 million times and on each sample the optHSIC test was performed, resulting in 1.2 million p-values. These 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 6: The rejection rate of zHSIC in against the distributions D.1-8 of Table 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 7: The rejection rate of wHSIC in against the distributions D.1-8 of Table 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 8: The rejection rate of the Cox proportional hazards likelihood ratio test in against the distributions D.1-8 of Table 2.

A.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 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 $20, 40, 60, 80, 100\%$ 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 10: The rejection rates of the various methods against distributions D.1-D.8 given in Table 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 $(\Delta = 1)$.

A.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. Pop-

ular 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 A.11.1 and Section A.11.2 respectively. We omit consideration of zHSIC, as it is fundamentally more limited, given the larger number of available methods.

A.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 $\left((z_i^0, \delta_i^0)\right)_{i=1}^{n_0}$ and $\left((z_i^1, \delta_i^1)\right)_{i=1}^{n_1}$ 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 covariance kernel of Brownian motion. 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 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:

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

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 A.11.2 provides an alternative permuation strategy under dependent censoring, that was proposed by Wang, Lagakos, and Gray (2010). It was proposed in the context of the logrank test, but can equally be used

for other statistics.

Algorithm 1: 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 $\text{wHSIC}(D)$ in the vector

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

where wHSIC is as defined above. ;

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

A.11.2 ipxHSIC

This subsection overviews a test we name ipxHSIC, which uses the same statistic $\text{wHSIC}(D)$ defined in Section A.11.1 above, but 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 statistics, 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

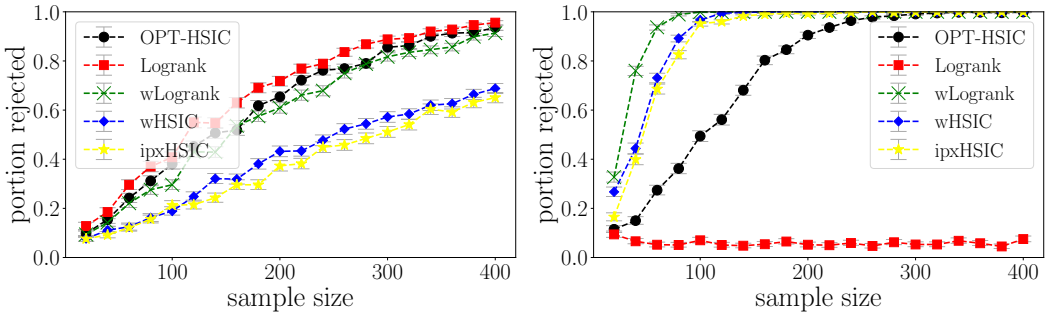
to the continuous case, where for each covariate we only have one individual in the study with that exact covariate.

A.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 (2020). 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 (2020) a combination of weights is proposed, so as to achieve power against a wider class of alternatives. In particular Ditzhaus and Friedrich (2020) 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 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.

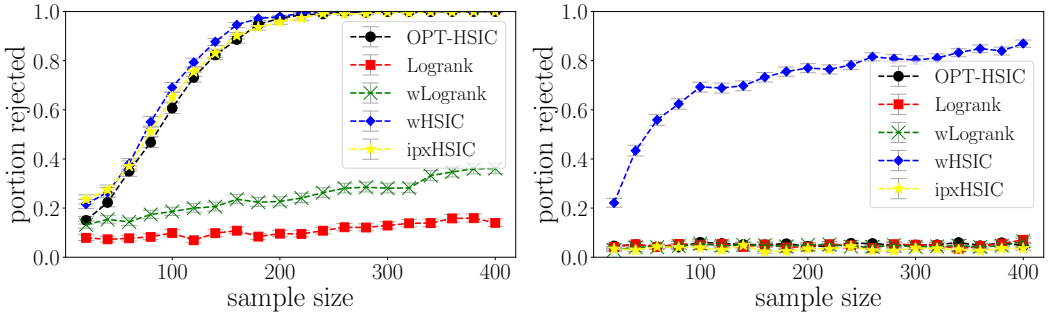
D.	T_0	T_1	C_0	C_1	% Observed
1	$\text{Exp}(1)$	$\text{Exp}(1/1.6)$	$\text{Exp}(1/2)$	$\text{Exp}(1/2)$	60 %
2	$\text{Weib}(1,5)$	$\text{Weib}(1,1.5)$	$\text{Exp}(1/2)$	$\text{Exp}(1/2)$	60 %
3	$\text{Exp}(1)$	(0.43, 1.39+Exp(1))	$1+\text{Exp}(1/2)$	$1+\text{Exp}(1/2)$	90 %
4	$\text{Exp}(1)$	$\text{Exp}(1)$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

Table 11: The 4 scenarios in which in which we perform two-sample tests. T_1 is 0.43 w.p. 0.75 and $1.39+\text{Exp}(1)$ 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 12: 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.

A.12 Example of data with binary covariates 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 $100 + \text{Exp}(\text{mean} = 2)$ and the 50 individuals of group $X = 1$ have event time

$100 + \text{Exp}(\text{mean} = 1)$. 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.

A.12.1 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.

References

Berrett, T. B. and R. J. Samworth (2019): "Nonparametric independence testing via mutual information," *Biometrika*, 106, 547–566.

Ditzhaus, M. and S. Friedrich (2020): "More powerful logrank permutation tests for two-sample survival data," *Journal of Statistical Computation and Simulation*, 90, 2209–2227, URL `https://doi.org/10.1080/00949655.2020.1773463`.

Matabuena, M. (2019): "Energy distance and kernel mean embedding for two sample survival test," *arXiv preprint arXiv:1901.00833*.

Wang, R., S. Lagakos, and R. Gray (2010): "Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring," *Biostatistics*, 11.4, 676–692.