Copula-Based Deep Survival Models for Dependent Censoring (Supplementary Material)

Ali Hossein Gharari Foomani[*,1,2] **Michael Cooper**[*,3,5] **Russell Greiner**[1,2] **Rahul G. Krishnan**[3,4,5]

[1]Department of Computing Science, University of Alberta
[2]Alberta Machine Intelligence Institute
[3]Department of Computer Science, University of Toronto
[4]Department of Laboratory Medicine and Pathobiology, University of Toronto
[5]Vector Institute

CONTENTS

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

A TABLE OF NOTATION

$\mathbf{1}^N$	N-vector filled with 1's.
$\mathbb{1}[\cdot]$	Indicator function.
$\mathcal{L}(\cdot)$	Likelihood function.
$\ell(\cdot)$	Log-likelihood function.
$X \in \mathcal{X}$	Covariates of one instance (as elements of the covariate space, $\mathcal{X}$).
$T_E \in \mathbb{R}_+$	Event time.
$T_C \in \mathbb{R}_+$	Censorship time.
$T_{\text{obs}} \in \mathbb{R}_+$	Time of last observation; the minimum of T_E, T_C.
$T \in \mathbb{R}_+$	Either event or censoring time; used in contexts where a quantity may refer to either.
$\delta \in \{0, 1\}$	Event indicator. Equal to 1 if the observed time is the event time; 0 otherwise.
$\mathcal{D} \subset \mathcal{X} \times \mathbb{R}_+ \times \{0, 1\}$	Survival dataset of the form $\{(X^{(i)}, T_{\text{obs}}^{(i)}, \delta^{(i)})\}_{i=1}^N$.
$S_T \in \mathcal{S}$	Survival function, $S : \mathbb{R}_+ \to [0, 1]$, and space of survival functions, $\mathcal{S}$.
f_T	Probability density function over time, representing $\Pr(T = t)$.
F_T	Cumulative density function over time, representing $\Pr(T < t)$.
C	A copula. If written as C_θ, this denotes a copula parameterized by the dependence parameter θ.
u_1, u_2	Inputs to a copula function. It is assumed that these are uniformly distributed.

B COPULA FORMULAE AND ALGORITHMS

B.1 TABLE OF PRELIMINARIES

Copula	$C_\theta(u_1, u_2)$	Θ	$\frac{\partial}{\partial u_1} C_\theta(u_1, u_2)$
Independence Copula	$u_1 u_2$	N/A	u_2
Clayton Copula	$\left(\max\left(u_1^{-\theta} + u_2^{-\theta} - 1, 0\right)\right)^{-1/\theta}$	$[-1, \infty) \backslash \{0\}$	$\begin{cases} \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{\frac{-\theta-1}{\theta}} u_2^{-\theta-1} & u_1^{-\theta} + u_2^{-\theta} > 1 \\ 0 & \text{otherwise} \end{cases}$
Frank Copula	$\frac{-1}{\theta} \log\left(1 + \frac{(\exp(-\theta u_1) - 1)(\exp(-\theta u_2) - 1)}{\exp(-\theta) - 1}\right)$	$\mathbb{R} \backslash \{0\}$	$\frac{\exp(-\theta u_1)(\exp(-\theta u_2) - 1)}{\exp(-\theta) - 1}$

Table 1: A table of formulas representing different classes of bivariate copulas used in our experiments. This table provides $C_\theta(u_1, u_2)$, the formula for the cumulative distribution function of the copula; Θ, representing the family Θ from which valid θ may be drawn; and $\frac{\partial}{\partial u_1} C_\theta(u_1, u_2)$, representing the partial derivative of the copula with respect to its first parameter. Due to the symmetric nature of these copulas, one can readily find $\frac{\partial}{\partial u_2} C_\theta(u_1, u_2)$ from $\frac{\partial}{\partial u_1} C_\theta(u_1, u_2)$ by simply interchanging u_1, u_2 (hence, we only provide $\frac{\partial}{\partial u_1} C_\theta(u_1, u_2)$).

B.2 SAMPLING FROM A COPULA

Algorithm 2 requires that we draw samples from the Clayton and Frank copulas. To do so, we implement the copula sampling scheme from in the Python `statsmodels` package [Seabold and Perktold, 2010].

B.3 QUANTILE DENSITY VISUALIZATIONS

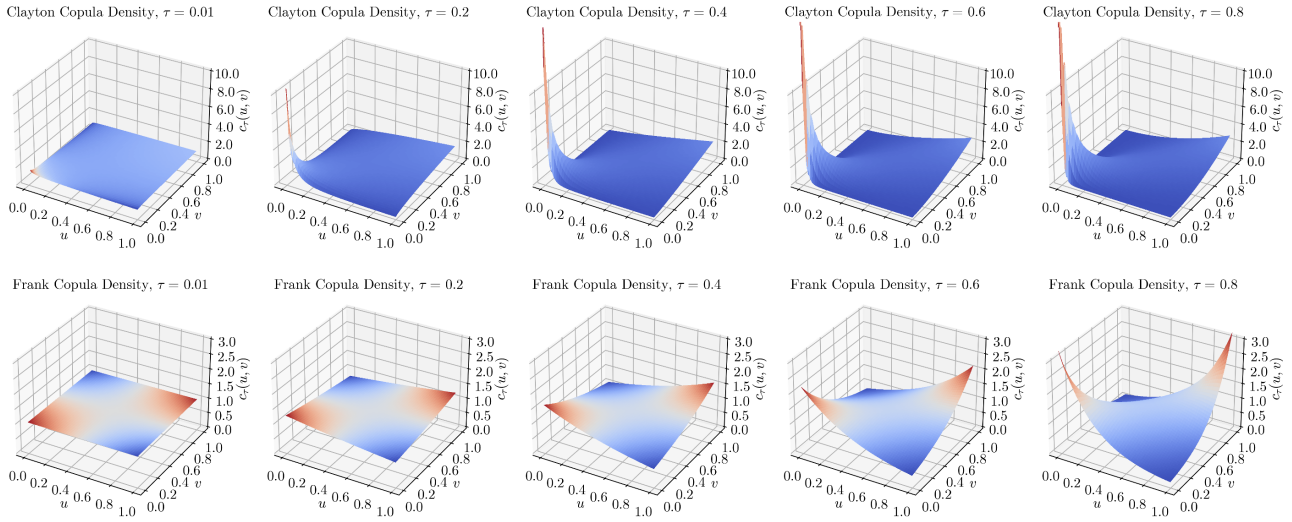

Figure 1: Plots of the densities for the Clayton (top row) and Frank (bottom row) copulas, under different degrees of dependence. These plots are functions of each of the copula's margins, u and v. In practice, u and v are quantiles of the event and censoring distributions. Observe that, as the dependence increases, the difference in density between the on-diagonal points (points where $u \approx v$) and the off-diagonal points increases. Note also that, while the Clayton copula concentrates density around low quantiles (points where $u \approx v \approx 0$) as dependence increases, the Frank copula concentrates density more uniformly around the on-diagonal.

B.4 INTUITION FOR COPULA SELECTION

In Section 7, we discussed three different cases that can be used to build intuition around the forms of dependence induced by various copulas. In Figure 2, we visualize these cases, and relate them to the quantile density plots in Appendix B.3. The point of this section is to build intuition regarding the *a priori* selection of a copula, so we will necessarily make a few simplifications. For example, although the three cases we discuss are not exhaustive – it is possible that the event and censoring survival curves cross (*e.g.* if the event and censoring distributions have different baseline hazards) – they present clean intuition relating the choice of copula to the structure of the joint density it produces.

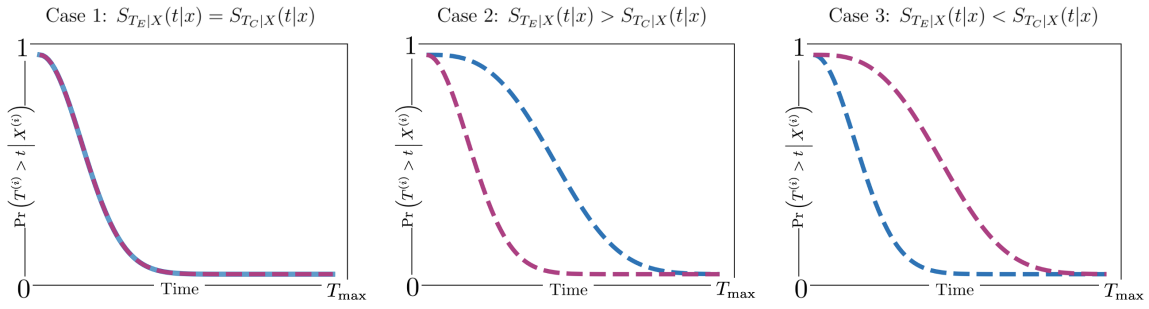

Figure 2: Three survival functions highlighting the three cases we presented in Section 7 of the main body. **Left**: the case where the conditional survival and censoring functions are the same. **Center**: the censoring survival function decays faster than the event survival function. **Right**: the event survival function decays faster than the censoring survival function.

The key intuition for selecting a copula from domain knowledge can be drawn from Sklar's Theorem (Survival), which states that a joint distribution over event and censoring times can be modelled as two independent event and censoring

distributions the quantiles of which are linked by a copula. When the event and censoring distributions are the same (left), the event quantile of a given time is the same as the censoring quantile for that same time. Thus, an increased dependence between event and censoring quantiles is directly reflected in a positive dependence between event and censoring times. When the censoring survival curve decays more quickly than the event survival curve, the event quantile of a given event time is higher than the censoring quantile for that same time. Therefore, increasing the dependence between event and censoring quantiles increases the likelihood that the censoring time precedes the event time. By symmetry, the opposite is true when the event survival curve decays more quickly than the censoring survival curve. An increase in dependence beteween quantiles in this setting increases the likelihood that the event itme precedes the censoring time under the model.

C DERIVATIONS

C.1 THE RIGHT-CENSORED LIKELIHOOD

As a starting point for the subsequent derivations, we discuss the intuition behind the general likelihood for right-censored survival data, and present its formulation in Equation 12.

Recall that a survival dataset $\mathcal{D}$ consists of N i.i.d. samples of the form $\{(X^{(i)}, T_{\text{obs}}^{(i)}, \delta^{(i)})\}_{i=1}^{N} \subset \mathcal{X} \times \mathbb{R}_+ \times \{0, 1\}$. The likelihood expressed in Equation 12 uses the $\delta^{(i)}$ terms in the exponent as a conditional binary filter: raising a term to the power of $\delta^{(i)}$ ensures it is non-degenerate only when the patient experiences an event; raising a term to the power of $1 - \delta^{(i)}$ ensures it is non-degenerate only when the patient is censored.

Let $f_{T_E, T_C|X}$ represent the joint density function of the event and censoring times, respectively, conditional on the patients' covariates. There are two mutually-exclusive, collectively-exhaustive into which we can decompose the right-censored likelihood for a given patient i:

1. **Case 1** ($\delta^{(i)} = 1$): If $\delta^{(i)} = 1$, the likelihood term should express that $T_E^{(i)} = T_{\text{obs}}^{(i)}$, and $T_C^{(i)} > T_{\text{obs}}^{(i)}$. This corresponds to the observation that the patient experienced the event at time $T_{\text{obs}}^{(i)}$, and was not censored prior to experiencing the event. The probability of this event under our density function is $\int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E, T_C|X}(T_{\text{obs}}^{(i)}, t_c|X^{(i)})dt_c$.
2. **Case 2** ($\delta^{(i)} = 0$): If $\delta^{(i)} = 0$, the likelihood term should express that $T_C^{(i)} = T_{\text{obs}}^{(i)}$, and $T_E^{(i)} > T_{\text{obs}}^{(i)}$. This corresponds to the observation that the patient is censored at time $T_{\text{obs}}^{(i)}$, and did not experience an event prior to being censored. The probability of this event under our density function is $\int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E, T_C|X}(t_e, T_{\text{obs}}^{(i)}|X^{(i)})dt_e$.

Combining these two cases, and applying the assumption that our data is i.i.d., yields the general likelihood function for right-censored data.

$$\mathcal{L}(\mathcal{D}) = \prod_{i=1}^{N} \underbrace{\left[\int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E, T_C|X}(T_{\text{obs}}^{(i)},\, t_c \,|\, X^{(i)})\, dt_c \right]}_{\Pr\left(T_E = T_{\text{obs}}^{(i)},\, T_C > T_{\text{obs}}^{(i)} \,|\, X^{(i)}\right)}^{\delta^{(i)}} \underbrace{\left[\int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E, T_C|X}(t_e,\, T_{\text{obs}}^{(i)} \,|\, X^{(i)})\, dt_e \right]}_{\Pr\left(T_C = T_{\text{obs}}^{(i)},\, T_E > T_{\text{obs}}^{(i)} \,|\, X^{(i)}\right)}^{1-\delta^{(i)}} \tag{12}$$

C.2 THE RIGHT-CENSORED LOG-LIKELIHOOD UNDER CONDITIONAL INDEPENDENCE

Under the assumption that $T_E \perp T_C|X$, we can factorize the conditional density distributions in Equation 12. $f_{T_E,T_C|X}$ factorizes into $f_{T_E|X} f_{T_C|X}$.

$$\mathcal{L}(\mathcal{D}) = \prod_{i=1}^{N} \left[f_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_C|X}(t_c|X^{(i)}) dt_c \right]^{\delta^{(i)}} \left[f_{T_C|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E|X}(t_e|X^{(i)}) dt_e \right]^{1-\delta^{(i)}} \tag{13}$$

$$= \prod_{i=1}^{N} \left[f_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \left(1 - F_{T_C|X}(T_{\text{obs}}^{(i)}|X^{(i)})\right) \right]^{\delta^{(i)}} \left[f_{T_C|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \left(1 - F_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)})\right) \right]^{1-\delta^{(i)}} \tag{14}$$

$$= \prod_{i=1}^{N} \left[f_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) S_{T_C|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \right]^{\delta^{(i)}} \left[f_{T_C|X}(T_{\text{obs}}^{(i)}|X^{(i)}) S_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \right]^{1-\delta^{(i)}} \tag{15}$$

$$\therefore \quad \ell(\mathcal{D}) = \sum_{i=1}^{N} \delta^{(i)} \log \left[f_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \right] + \delta^{(i)} \log \left[S_{T_C|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \right] + (1-\delta^{(i)}) \log \left[f_{T_C|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \right] + (1-\delta^{(i)}) \left[S_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \right] \tag{16}$$

C.3 THE RIGHT-CENSORED LOG-LIKELIHOOD UNDER DEPENDENCE DEFINED BY A COPULA

C.3.1 Proof of Lemma 1

Lemma 1 (Conditional Survival Function Under Sklar's Theorem (Survival))**.** If $S_{T_E,T_C|X}(t_e, t_c|x) = C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E|X}(t_e|x) \\ u_2=S_{T_C|X}(t_c|x)}}$, then,

$$\int_{t_c}^{\infty} f_{T_C|T_E,X}(t_c|t_e, x) = \frac{\partial}{\partial u_1} C(u_1, u_2)\Big|_{\substack{u_1=S_{T_E|X}(t_e|x) \\ u_2=S_{T_C|X}(t_c|x)}} \tag{17}$$

Proof.

$$\int_{t_c}^{\infty} f_{T_C|T_E,X}(t_c|t_e,x) = \frac{\int_{t_c}^{\infty} f_{T_C,T_E|X}(t_c,t_e|x)dt_c}{f_{T_E|X}(t_e|x)} \quad \text{(Def'n of Cond. Prob.)} \tag{18}$$

$$= \frac{\frac{-\partial}{\partial T_E}\int_{t_e}^{\infty}\int_{t_c}^{\infty} f_{T_C,T_E|X}(t_c,t_e|x)dt_c dt_e}{f_{T_E|X}(t_e|x)} \tag{19}$$

$$= \frac{\frac{-\partial}{\partial T_E} S_{T_C,T_E|X}(t_c,t_e|x)}{f_{T_E|X}(t_e|x)} \quad \text{(Def'n of Survival Function)} \tag{20}$$

$$= \frac{\frac{-\partial}{\partial T_E}\left(C(u_1,u_2)\Big|_{\substack{u_1=S_{T_E|X}(t_e|x)\\u_2=S_{T_C|X}(t_c|x)}}\right)}{f_{T_E|X}(t_e|x)} \quad \text{(Sklar's Theorem)} \tag{21}$$

$$= \frac{\frac{-\partial}{\partial u_1}\left(C(u_1,u_2)\Big|_{\substack{u_1=S_{T_E|X}(t_e|x)\\u_2=S_{T_C|X}(t_c|x)}}\right)\frac{\partial}{\partial T_E}S_{T_E|X}(t_e|x)}{f_{T_E|X}(t_e|x)} \quad \text{(Chain Rule)} \tag{22}$$

$$= \frac{-\partial}{\partial u_1}\left(C(u_1,u_2)\Big|_{\substack{u_1=S_{T_E|X}(t_e|x)\\u_2=S_{T_C|X}(t_c|x)}}\right)\frac{-\cancel{f_{T_E|X}(t_e|x)}}{\cancel{f_{T_E|X}(t_e|x)}}^{-1} \tag{23}$$

$$= \frac{\partial}{\partial u_1}\left(C(u_1,u_2)\Big|_{\substack{u_1=S_{T_E|X}(t_e|x)\\u_2=S_{T_C|X}(t_c|x)}}\right) \tag{24}$$

□

Corollary. We can symmetrically apply this lemma to the converse case, $f_{T_E|T_C,X}$, to obtain:

$$\int_{t_e}^{\infty} f_{T_E|T_C,X}(t_e|t_c,x) = \frac{\partial}{\partial u_2}\left(C(u_1,u_2)\Big|_{\substack{u_1=S_{T_E|X}(t_e|x)\\u_2=S_{T_C|X}(t_c|x)}}\right) \tag{25}$$

C.3.2 Derivation of the Right-Censored Log Likelihood Under a Copula

Having now proven Lemma 1, we apply it to derive a likelihood function for survival prediction under dependent censoring. We use Equation 12 as the starting point for our derivation.

$$\mathcal{L}(\mathcal{D}) = \prod_{i=1}^{N} \left[\int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E,T_C|X}(T_{\text{obs}}^{(i)}, t_c|X^{(i)})\, dt_c \right]^{\delta^{(i)}} \left[\int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E,T_C|X}(t_e, T_{\text{obs}}^{(i)}|X^{(i)})\, dt_e \right]^{1-\delta^{(i)}} \tag{26}$$

$$= \prod_{i=1}^{N} \left[f_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_C|T_E,X}(t_c|T_{\text{obs}}^{(i)}, X^{(i)})\, dt_c \right]^{\delta^{(i)}} \times \quad \text{(Chain Rule)} \tag{27}$$

$$\left[f_{T_C|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \int_{T_{\text{obs}}^{(i)}}^{\infty} f_{T_E|T_C,X}(t_e|T_{\text{obs}}^{(i)}, X^{(i)})\, dt_e \right]^{1-\delta^{(i)}}$$

$$= \prod_{i=1}^{N} \left[f_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \frac{\partial}{\partial u_1} \left(C(u_1,u_2)\Big|_{\substack{u_1=S_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \\ u_2=S_{T_C|X}(T_{\text{obs}}^{(i)})|X^{(i)}}} \right) \right]^{\delta^{(i)}} \times \quad \text{(Lemma 1)} \tag{28}$$

$$\left[f_{T_C|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \frac{\partial}{\partial u_2} \left(C(u_1,u_2)\Big|_{\substack{u_1=S_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \\ u_2=S_{T_C|X}(T_{\text{obs}}^{(i)})|X^{(i)}}} \right) \right]^{1-\delta^{(i)}}$$

$$\therefore \quad \ell(\mathcal{D}) = \sum_{i=1}^{N} \delta^{(i)} \log \left[f_{T_E|X}\left(T_{\text{obs}}^{(i)}|X^{(i)}\right) \right] + \delta^{(i)} \log \left[\frac{\partial}{\partial u_1} C(u_1,u_2)\Big|_{\substack{u_1=S_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \\ u_2=S_{T_C|X}(T_{\text{obs}}^{(i)})|X^{(i)}}} \right] + \tag{29}$$

$$(1-\delta^{(i)}) \log \left[f_{T_C|X}\left(T_{\text{obs}}^{(i)}|X^{(i)}\right) \right] +$$

$$(1-\delta^{(i)}) \log \left[\frac{\partial}{\partial u_2} C(u_1,u_2)\Big|_{\substack{u_1=S_{T_E|X}(T_{\text{obs}}^{(i)}|X^{(i)}) \\ u_2=S_{T_C|X}(T_{\text{obs}}^{(i)})|X^{(i)}}} \right]$$

C.4 THE WEIBULL COXPH MODEL

Recall that the Weibull CoxPH model is defined in terms of its hazard, as follows.

$$h_{T|X}(\,t|X\,) \;=\; \left(\frac{\nu}{\rho}\right)\left(\frac{t}{\rho}\right)^{\nu-1} \exp\left(\,g_\psi(X)\,\right) \tag{30}$$

Our method, however, relies on the ability to extract additional quantities – the density ($\hat{f}_{T|X}$) and survival functions ($\hat{S}_{T|X}$) – from the model, as these are essential to computing our likelihood function. In this section, we derive the closed-form expressions for these two quantities that are present in the main body of our work.

C.4.1 The Survival Function

The survival function under our model can be derived via its cumulative hazard.

Definition 1 (Cumulative Hazard)**.** The *cumulative hazard*

$$\hat{H}_{T|X}(t|X) \;\triangleq\; \int_0^t \hat{h}_{T|X}(u|X)du \tag{31}$$

represents the integral of the hazard function over all time prior to a specified time, t.

The cumulative hazard of the Weibull CoxPH can be expressed in closed form as follows:

$$\hat{H}_{T|X}(\,t|X\,) \;=\; \int_0^t \left(\frac{\nu}{\rho}\right)\left(\frac{u}{\rho}\right)^{\nu-1} \exp\left(\,g_\psi(X)\,\right)\, du \tag{32}$$

$$\;=\; \left(\frac{t}{\rho}\right)^{\nu} \exp\left(\,g_\psi(X)\,\right) \tag{33}$$

One alternative formulation of the survival function expresses $S_{T|X}$ in terms of the hazard function, as follows.

$$S_{T|X}(t|X) \triangleq \exp(-H_{T|X}(t|X)) \tag{34}$$

We can apply this identity to Equation 33 to obtain the following expression for $\hat{S}_{T|X}$ under the Weibull CoxPH model:

$$\hat{S}_{T|X}(t|X) = \exp\left(-\left(\frac{t}{\rho}\right)^{\nu} \exp\left(\, g_\psi(X)\,\right)\right) \tag{35}$$

C.4.2 The Density Function

From Equation 3, we know that the density of an event can be calculated as follows.

$$f_{T|X}(t|X) \;=\; S_{T|X}(t|X) h_{T|X}(t|X) \tag{36}$$

C.5 A STABLE IMPLEMENTATION

In order to optimize a Weibull model in a stable way we used another representation of Weibull distribution. This new representation is derived by applying log transformation to the cumulative hazard function of Weibull distribution.

$$\begin{aligned} H_{T|X}(\,t|X\,) &= \exp(\log(H_{T|X}(t|X))) \\ &= \exp\left(\log\left(\left(\frac{t}{\rho}\right)^{\nu} \exp(g_\psi(X))\right)\right) \\ &= \exp(\nu\log(t) - \nu\log(\rho) + g_\psi(X)) \end{aligned} \tag{37}$$

Setting $\sigma = \frac{1}{\nu}$, $\mu = \log(\rho)$,and $f(x) = -\frac{g_\psi(X)}{\nu}$, gives us a long-cumulative hazard function of the following form.

$$H_{T|X}(\,t|X\,) \;=\; \exp\left(\frac{\log(t) - \mu - f(x)}{\sigma}\right) \tag{38}$$

C.5.1 Hazard function

Given the formula for the cumulative hazard function we can derive the hazard function in the new format by taking the derivative of cumulative hazard with respect to t.

$$h_{T|X}(\,t|X\,) = \frac{\partial H_{T|X}(\,t|X\,)}{\partial t} = \frac{H_{T|X}(\,t|X\,)}{t\sigma} \tag{39}$$

D ALGORITHMS

D.1 COMPUTING THE SURVIVAL-ℓ_1

Here, we expand on the computation of the Survival-ℓ_1 metric from the main paper by providing an algorithm for the explicit computation of the inner term of the Survival-ℓ_1 metric, as well as the value T_{max} for the given pair of survival curves, $S, \hat{S}$:

$$\mathcal{C}_{\textit{Survival-}\ell_1}(S, \hat{S}) \quad = \quad \sum_{i=1}^{N} \frac{1}{N \times T_{\text{max}}^{(i)}} \underbrace{\int_0^{\infty} \left| S_{T\,|\,X}(t\,|\,X^{(i)}) - \hat{S}_{T\,|\,X}(t\,|\,X^{(i)}) \right| dt}_{\text{Inner Term}}$$

Although the integral in the $\mathcal{C}_{\textit{Survival-}\ell_1}$ is over an infinite domain, in this approximation, we consider only the simplified case wherein the upper bound of integration is T_{max}.

Algorithm 3: Discrete Approximation of the Inner Term of the Survival-ℓ_1

Input:

1. S_1, S_2: Survival curves to compare under the Survival-ℓ_1 metric. Here, we assume S_1 is the ground-truth survival curve, and S_2 is the estimated curve.
2. $Q_{\|\cdot\|}$: Normalizing quantile.
3. N_{steps}: Number of discretization steps.

Result:

1. Δ_{total}: a discretized approximation of the integral $\int_0^{T_{\max}} \left|S_1(t \mid X^{(i)}) - S_2(t \mid X^{(i)})\right| dt$.
2. $T_{\max}$: This is used as a normalization weight when computing the full expression for the Survival-ℓ_1 metric.

$T_{\max} \leftarrow S_1^{-1}\left(Q_{\|\cdot\|}\right)$;
$\Delta_{\text{total}} \leftarrow 0$
for $i = 1, \ldots, N_{steps}$ **do**
  $\Delta_{i;S_1,S_2} \leftarrow \frac{T_{\max}}{N_{\text{steps}}} \times \ell_1 \left[S_1\left(\frac{i \times T_{\max}}{N_{\text{steps}}}\right), S_2\left(\frac{i \times T_{\max}}{N_{\text{steps}}}\right)\right]$;
  $\Delta_{\text{total}} \leftarrow \Delta_{\text{total}} + \Delta_{i;S_1,S_2}$;
end
return Δ_{total}, T_{max}

D.2 CREATING A SEMI-SYNTHETIC DATASET WITH DEPENDENT CENSORING

We convert a regression dataset to a survival dataset with dependent censoring using the following algorithm.

Algorithm 4: Semi-Synthetic Dataset Construction with Dependent Censoring

Input:

1. $\mathcal{D}_{\text{reg}} = \left\{X^{(i)}, Y^{(i)}\right\}_{i=1}^{N} \subseteq \mathcal{X} \times \mathbb{R}_+$. Regression dataset consisting of covariates and labels.
2. $C_\theta : [0,1] \times [0,1] \to [0,1]$. A bivariate, uniparametric copula.

Result:

1. $\mathcal{D}_{C,\theta} \subseteq \mathcal{X} \times \mathbb{R}_+\{0,1\}$. Artificially censored version of D_{reg} in which the joint distribution between Y and T_C is governed by the application of Sklar's Theorem to the copula C_θ.

```
# Learn a Weibull CoxPH model based on the outcomes of the train set without any
  censoring
```
$\hat{W}_E \leftarrow$ `Weibull-Linear`$(Y, X, \mathbf{1}^N)$;
$W_C \leftarrow W_E$;
$W_C.\nu \leftarrow W_C.\nu / 0.6$ `# Decreases the variance of the censoring distribution`
$T_C \leftarrow \mathbf{0}^N$;
$\mathcal{D}_{C,\theta} = \emptyset$;
for $i = 1, \ldots, N$ **do**
  $u_1^{(i)} \leftarrow \hat{S}_{W_E}(Y^{(i)})$; `# Obtain event quantile`
  $u_2^{(i)} \sim C_\theta(\cdot \mid u_1^{(i)})$; `# Sample censoring quantile conditionally from the copula`
  $T_C^{(i)} \leftarrow \hat{S}_{W_C}^{-1}(u_2^{(i)})$; `# Obtain censoring time via inv. censoring survival function`
  $\mathcal{D}_{C,\theta} \leftarrow \mathcal{D}_{C,\theta} \cup \{(X^{(i)}, \min\left(Y^{(i)}, T_C^{(i)}\right), \mathbb{1}[Y^{(i)} \leq T_C^{(i)}])\}$;
end
return $\mathcal{D}_{C,\theta}$;

E ADDITIONAL EXPERIMENTAL DETAILS

E.1 EVALUATION METRIC BIAS UNDER DEPENDENCE

For this experiment, we sampled 10,000 data points according to Algorithm 2 with $X \in \mathbb{R}^{N\times 10} \sim \mathcal{U}_{[0,1]}$, $\nu_E^* = 4, \rho_E^* = 17, \psi_E^*(X) = X_1^2 + X_2^2$, $\nu_C^* = 3, \rho_C^* = 16, \psi_C^*(X) = \sum_{i=1}^3 \beta_{C_i} X_i^2$, where $\beta_C \in [0,1]^{10} \sim \mathcal{U}_{[0,1]}$.

E.2 IMPLEMENTATION DETAILS

We halted the learning algorithms if the validation loss failed to improve for a consecutive 3000 epochs. The `Linear-Risk` experiments were conducted without any form of regularization, whereas the `Nonlinear-Risk` experiments employed ℓ_2 regularization with a coefficient of $\lambda = 0.001$. For all experiments, the learning rate remained constant at 0.001.

F DATASETS AND PROCESSING

F.1 STEEL INDUSTRY ENERGY CONSUMPTION (`STEEL`) DATASET

The `STEEL` dataset [VE et al., 2021, Sathishkumar et al., 2020a,b] is a regression dataset from the UCI Machine Learning Repository [Asuncion and Newman, 2007], comprising 35,040 observations of of the power consumption of plants run by DAEWOO Steel Co. Ltd in Gwangyang, South Korea. The data includes 9 covariates (including day of the week, type of load (light/medium/heavy), CO_2 measurements in PPM, and leading/lagging reactive power measurements), and one outcome variable (the industry energy consumption, measured in kWh). For our semi-synthetic experiment, we used 70% of the data as the train set, 15% as the validation set, and 15% as the test set.

F.2 AIRFOIL SELF-NOISE (`AIRFOIL`) DATASET

The `Airfoil` dataset [Dua and Graff, 2017] is another regression dataset from the UCI Machine Learning Repository [Asuncion and Newman, 2007]. It comprises 1,503 observations obtained from aerodynamic and acoustic tests of two and three-dimensional airfoil blade sections conducted in an anechoic wind tunnel. The data includes 6 covariates (including frequency, angle of attack, chord length, free-stream velocity, suction side displacement thickness) and one outcome variable (scaled sound pressure level). For our semi-synthetic experiment, we used 70% of the data as the train set, 15% as the validation set, and 15% as the test set.

G ADDITIONAL RESULTS

For the experiments in this section we used a Clayton copula to censor the dataset as described in Algorithm 4.

G.1 SEMI-SYNTHETIC SURVIVAL REGRESSION ON THE `STEEL` DATASET

Below, we present the results of our survival regression on the test set of the `STEEL` dataset.

	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.6$	$\tau = 0.8$
Weibull CoxPH (No Censoring)	0.513	0.513	0.513	0.513
Weibull CoxPH (Independence Assuming)	0.333	0.309	0.324	0.341
Weibull CoxPH (Dependent, **ours**)	0.371	0.442	0.512	0.508

Table 2: A table of R^2 values given by performing survival regression on the `STEEL` dataset under various degrees of dependence induced by Algorithm 4. A higher R^2 indicates a better performing algorithm. The top row represents the performance of a Weibull CoxPH model trained on the regression data without censoring; this should indicate an upper bound on the performance of any survival model under censoring. We find that the performance of our approach, though below the theoretical upper bound, lies substantially above that of the independence-assuming approach.

G.2 SEMI-SYNTHETIC SURVIVAL REGRESSION ON THE AIRFOIL DATASET

Below, we present the results of our survival regression on the test set of the AIRFOIL dataset.

	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.6$	$\tau = 0.8$
Weibull CoxPH (No Censoring)	0.572	0.572	0.572	0.572
Weibull CoxPH (Independence Assuming)	0.583	0.549	0.465	0.330
Weibull CoxPH (Dependent, **ours**)	0.580	0.564	0.507	0.484

Table 3: A table of R^2 values given by performing survival regression on the AIRFOIL dataset under various degrees of dependence induced by Algorithm 4. The top row represents the performance of a Weibull CoxPH model trained on the regression data without censoring; this should indicate an upper bound on the performance of any survival model under censoring. While performance of both methods degrades as dependence increases, we find that our method is better able to obtain higher values of R^2 than the independence-assuming model under greater degrees of dependence.