Copula-based multi-event modeling and prediction using fleet service records

Akash Deep, Shiyu Zhou, Dharmaraj Veeramani

Abstract

Due to the fast development of communication and information technology, the event sequences collected from a fleet, which consists of potentially a large number of similar units become readily available. Each event sequence is collected from a specific unit and may consists of multiple types of events (e.g., different types of failure event). In this paper, we present a novel method for modeling and prediction using those event sequences. Conventional approaches to model and predict event data involve regression methods, such as Cox Proportional Hazards. The proposed method essentially uses Copula to approximate the joint distribution of time-to-event variables corresponding to each type of events. The marginal distributions of the time-to-event variables that are needed for Copula function is obtained through the Cox PH regression models. The proposed model is more flexible and efficient in modeling the relationships among multiple events. With simulations and real-world case study, we show that proposed method outperforms the base regression model in prediction accuracy.

Keywords: Event Prediction, Correlated events, Copula, Survival Models, Association

1 Introduction

Technological advancements in communications and information technology is rapidly catalyzing data collection, sharing and processing. As a direct consequence, the service record for a fleet consisting potentially a large number of similar units becomes readily available. For example, many Internet-of-things enabled tele-service systems are available in practices now. In such a system, the products (e.g., automotive, forklift) in the field are brought online and linked to the cloud. The system failures and related service actions for the whole fleet are recorded, aggregated on the cloud and readily accessible. The fleet service record provides valuable information regarding various system failures and their relationships. This work concerns the modeling and prediction of failure events using fleet service records.

As shown in Figure 1, we view fleet service record as a collection of event sequences consisting multiple event types and their time stamps.

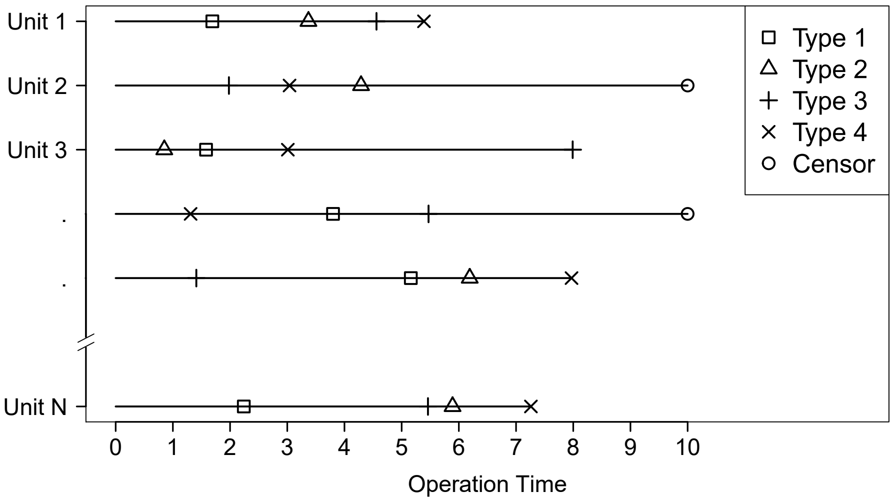


Figure 1: Illustration of multi-type event data, $K = 4$

Each event type represents one type of failure and each event sequence represents the event record for one unit. A typical engineering product consists of multiple components and can experience multiple types of failure events. The occurrence of different types of failure may be correlated. The rationale of the correlation is that first, a common environmental factor may lead to different failure (e.g., very hot environment may lead to motor overheat and motor brush failure); second, the operation of the components within the same unit is inter-related and one failure may lead to another failure (e.g., a bad alternator in a car will cause the battery failure).

In engineering practice, it is highly desirable to predict the occurrence of failures for efficient and safe system operations. In principle, failure prediction can be achieved by a very comprehensive physical analysis on the relationships among different failures and environmental factors. However, due to the system complexity and the facts that many impacting factors for system failures are not observable, this approach is often not feasible. On the other hand, the availability of fleet service records provides great opportunities for developing data-driven statistical methods to take advantage of the potential correlation among failure events and predict the future failure events.

In this paper, our goal is to use a data driven model to effectively learn the dependence structure among multiple events from historical fleet service records. Then the learned structure and the observed event sequence from the unit under study are used to predict the expected time to next event for the unit under study.

One popular school of thoughts on event prediction using historical event sequences is temporal mining. The basic idea of temporal mining is to first identify the temporal patterns heuristically, i.e., the sequences of events that frequently occur, and then prediction rules are developed based on these patterns. Mitsa (2010) covers theoretical and application aspects of several methods and algorithms for event prediction using temporal mining method. However, the temporal mining approaches are heuristic in nature. The prediction rules once built are rigid and not very flexible.

Within statistical literature, survival modeling and analysis focuses on event or time-to-event data (Hougaard (2000)). Several suitable approaches are available for event modeling and prediction problem. One obvious strategy is regression. We can regress an event of interest with rest of the events (see Li *et al.* (2007); Yuan *et al.* (2011)). Cox PH model is a popular regression model in survival and reliability literature which models the event hazard rate using available covariates (also known as predictors) (Cox, 1972). After the model is estimated, for prediction, we can plug in the observed predictor variables to obtain the event hazard function, with which we can get the mean time-to-event (Klein & Moeschberger, 2005). Although Cox PH is a very powerful modeling technique and has been extended in several ways, the available Cox PH model based prediction methods cannot differentiate the order of past events. In other words, different order of event occurrences would yield the same predictor set for the Cox PH, thus resulting in the same prediction – a limitation we will tackle in this paper.

Another approach to model different types of time-to-event variables together is through construction of a multi-variate joint distribution. Prediction from a joint distribution can be obtained by conditioning on observed events and then marginalizing the distribution of event of interest. Expected value of obtained marginal distribution provides the mean time-to-event. In literature, there are available methods, particularly multi-variate Weibull distribution which can be utilized (Hougaard, 1986; Lee & Wen, 2006; Marshall & Olkin, 1967). But, the model is inherently limited by the assumption of a parametric form which may not be true in many real-world applications. Another quite popular way of constructing multivariate joint distributions is through Copula. Copulas, however, under general formulation are limited if the data is censored (Meeker & Escobar, 1998; Aalen *et al.*, 2008). This limitation has been approached in survival analysis literature where Copulas have been adapted to study association between censored survival times (Hougaard, 2000, see chap 13). The resulting model is called survival Copula. Several studies have used the survival Copula formulation, however, majority of them are limited to bivariate cases (Wang & Ding, 2000; Chen *et al.*, 2010; Schemper *et al.*, 2013). There are limited work on extending the model to more than two variables case, such as, Othus & Li (2010); Barthel (2015). To the best of our knowledge, the existing literature on survival Copula focuses on describing the relationship among events and there is no literature in this domain which deals with event prediction. Furthermore, in the available joint distribution based approach, it is not clear how to incorporate external predictors such as the product type, known environmental factors into consideration.

Accounting for these limitations, in this paper we propose an integrated approach to model the event sequences and predict the next event. In a nutshell, our idea is to construct multivariate joint distribution of the time-to-event variables using Copulas. The marginal distributions of each time-to-event variable that is needed for Copula construction is obtained through the corresponding Cox PH regression model for each event. We call the Cox PH regression models *base models*. During prediction, we obtain conditional density of the event we try to predict using the multivariate joint distribution. Copulas, particularly Gaussian Copula is employed in this paper to create joint distribution from base models in the presence of censoring. The advantages of the proposed framework are – (i). Copula is a very flexible method, making the joint distribution quite flexible to fit the data. (ii). The model can easily incorporate various predictors and the order of the observed events is taken into consideration as well. These features result in a tighter probability space for the predicted future event and thus lead to a more accurate prediction.

The rest of the paper is organized as follows: in Section 2, we introduce the model formulation and the conventional regression based event prediction approach is reviewed. In Section 3, the Copula construction based on Cox PH regression models is presented. The Copula based event prediction and the intuition are described. In Section 4, we conduct several numerical experiments and discuss the results. Finally, in Sections 5 and 6 we demonstrate the effectiveness of our proposed model on a real-world dataset obtained from an industrial equipment, and conclude this paper with a brief discussion on several future directions.

2 Problem Formulation and Review of Regression Based Event Prediction

2.1 Data description

We consider a fleet of N independent units available in the historical dataset, where each unit is prone to possibly experience K different kinds of events during it operation (we also use $\mathbf{K}$ to denote the set of events). We assume that for a unit a particular event-type only occurs once during the observation period. The rationale of this assumption is that the first occurrence of an event often indicates the onset/existence of an underlying cause that leads to other events. Therefore, we can consider only the first time-to-events for all event-types. For any unit i, let the period of observation be $[0, \tau_i]$. We represent the available information for this unit using a set of two vectors: (i) event occurrence times as $\mathbf{t}_i = t_{1i}, t_{2i}, \ldots, t_{Ki}$, and (ii) censoring indicators $\boldsymbol{\delta}_i = \delta_{1i}, \delta_{2i}, \ldots, \delta_{Ki}$ which takes value 1 if the unit experienced this event or 0 otherwise. This gives, $t_{ki} = \tau_i, \quad \forall \delta_{ki} = 0$. Overall dataset can be summarized as $\mathbf{D} = \{\mathbf{t}_i, \boldsymbol{\delta}_i\} \quad \forall i \in N$.

As an example, we provide a sample dataset obtained from a industrial material handling machine in Table 1. The data was collected from a operation period of 250 time units. In total, 40 units were studied and 4 different kinds of events were recorded. Due to confidentiality reasons, we have renamed the event names simply as $E_1, \ldots, E_4$. Besides, for each event occurrence time, we also have the corresponding censoring indicator. If a unit (for example, unit 1) does not experience an event until the end of observation period (for example, E_2 for unit 1), the corresponding censoring indicator is 0, else it is 1 when the respective event had occurred.

Our aim is to use multi-type event data available in the historical dataset, $\mathbf{D}$, to make predictions for a new unit m. In other words, if we denote T_i as the time-to-event random variable for event type i, we aims at establishing a probabilistic model for $T_1, ..., T_K$ and estimate the model parameters using $\mathbf{D}$. Then for a unit that is not in the historical dataset,

Table 1: Sample dataset illustrating recorded measurements for the units, $N = 40, K = 4$

Unit	Event times and Censoring Indicator							
	E_1	δ_1	E_2	δ_2	E_3	δ_3	E_4	δ_4
1	32.81	1	250.00	0	60.86	1	80.46	1
2	250.00	0	59.00	1	250.00	0	250.00	0
3	7.78	1	250.00	0	30.56	1	71.88	1
..	..	..	..	..	..	..	..	..
38	11.00	1	1.80	1	1.86	1	250.00	0
39	18.89	1	250.00	0	250.00	0	250.00	0
40	250.00	0	250.00	0	250.00	0	250.00	0

we want predict the next event using the model and the event observations from the unit until current time instance t.

In the rest, we use $f(.)$ and $F(.)$ to denote the density and distribution functions of the time-to-event variable(s). The survival function for a time-to-event variable is denoted using $S(.) = 1 - F(.)$. In survival analysis, we also often use hazard function $\lambda(t)$ to describe the time-to-event random variable distribution. By definition, hazard function is the rate of probability of experiencing an event and it can be computed as $\lambda(t) = \frac{f(t)}{S(t)}$. From $\lambda(t)$ we can obtain $S(t)$ as

$$S(t) = \exp\left(-\int \lambda(t)dt\right) \tag{1}$$

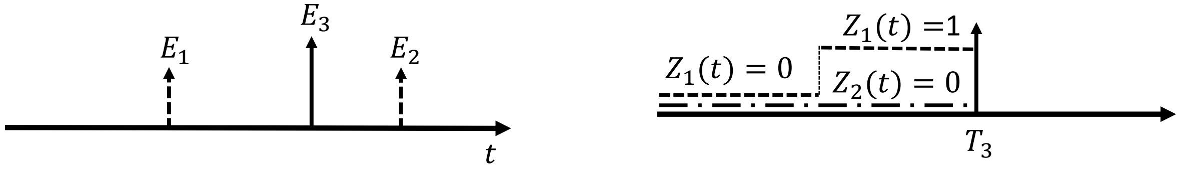
2.2 Review of Cox PH based event prediction

Cox PH model has been used for event prediction using multi-event sequences Li *et al.* (2007). Here we provide a brief review of this approach.

For event type k, we can establish a Cox PH model as

$$\lambda_k(t) = \lambda_{k0}(t)\exp(\boldsymbol{\beta}_k^T \mathbf{Z}_{E_k}(t)) \tag{2}$$

where $\lambda_k(t)$ is the hazard function of the kth event E_k, $\lambda_{k0}(t)$ is a non-parametric baseline hazard, $\boldsymbol{\beta}_k$ is the coefficient vector, and $\mathbf{Z}_{E_k}(t)$ is the predictor vector. Cox PH regression



(a) A sequence of three events for a unit

(b) Encoded predictors $Z_1(t)$ and $Z_2(t)$

Figure 2: An example of event sequence for a unit

model is a quite flexible model as the predictors $\mathbf{Z}_{E_k}(t)$ could include time fixed or time varying external factors such as "type of equipment", "location", or "workloads". The predictors could also include the other event types. Indeed, we can encode event type j, $j \neq k$, as

$$Z_j(t) = \begin{cases} 0, & 0 \leq t < t_j \\ 1, & t_j \leq t < t_k \end{cases}$$

In other words, $Z_j(t)$ is a function taking binary values: it is zero before E_j and jumps to 1 at the time instance when E_j occurs. We can encode all the rest event type in the same way and include them as predictors in $\mathbf{Z}_{E_k}(t)$. As mentioned above, $\mathbf{Z}_{E_k}(t)$ could also include other system factors. For the sake of simplicity, we only consider $Z_j(t)$s, $j \neq k$ as predictors in the Cox PH model. However, all the results still hold if we have additional external predictors in $\mathbf{Z}_{E_k}(t)$.

As an example, consider a unit which experiences three types of events E_1, E_2 and E_3 as shown in figure 2a, and let event-type E_3 be the response event with events E_1 and E_2 being predictors. Z_1 changes from 0 to 1 once event-type E_1 is observed, and since event-type E_2 is not experienced by the unit at least until event-type E_3, the corresponding predictor remains as 0 (see figure 2b).

After specifying the model, we use $\Theta_k = \{\lambda_{k0}, \boldsymbol{\beta}_k\}$ to summarize regression parameters for k-th event-type regression model. We estimate the model parameters for this regression model by maximizing the likelihood, L_k (see Klein & Moeschberger (2005)). Below we express L_k

$$L_k(\Theta_k;\mathbf{D}) = \prod_{i=1}^{N} \Bigg[\lambda_{k0}(t)\exp(\boldsymbol{\beta}_k^T\mathbf{Z}_{E_k}(t))\Bigg]^{\delta_{ik}} \exp\Bigg\{-\Lambda_{k0}(t)\exp(\boldsymbol{\beta}_k^T\mathbf{Z}_{E_k}(t))\Bigg\} \tag{3}$$

Let $\hat{\Theta}_k$ denote the estimated parameters. Please note that we can construct regression model for each event-type and thus we have K separate Cox PH regression models if we have K types of event. We let $\hat{\Theta}$ denote regression function parameters across all these K models.

During prediction, for a unit under service, at any time t^*, we can use the expected time-to-event as the predicted time of event. In reliability literature, the expected time-to-event is also often called as *mean remaining life* (MRL). We partition the events in two sets – the observed events from the unit under study until t^* are contained in Ψ, and, the remaining events are in set $\Xi = \mathbf{K}\backslash\Psi$. The estimated survival function of the event $p \in \Xi$, $\hat{S}_p(\cdot)$ can be obtained through Eq. (1) with the estimated hazard function $\hat{\lambda}_p(\cdot)$ from the corresponding Cox PH model.

$$\hat{S}_p(t) = \exp\Big(\int -\hat{\lambda}_{p0}(t)\exp(\hat{\boldsymbol{\beta}}_p^T\mathbf{Z}_{E_p}(t))dt\Big) = \hat{S}_{p0}(t)^{\exp(\hat{\boldsymbol{\beta}}_p^T\mathbf{Z}_{E_p}(t))} \tag{4}$$

where, $\hat{S}_{p0}(t)$ is estimated baseline survival function. The MRL denoted as $\mathrm{mrl}_p(t^*)$, is obtained as

$$\mathrm{mrl}_p(t^*) = \int_0^\infty \hat{S}_p(t > t^*|\mathbf{Z}_{E_p}(t^*) = \mathbf{z}_{E_p}(t^*))dt = \frac{\int_{t^*}^\infty \hat{S}_p(t|\mathbf{Z}_{E_p}(t^*) = \mathbf{z}_{E_p}(t^*))dt}{\hat{S}_p(t^*|\mathbf{Z}_{E_p}(t^*) = \mathbf{z}_{E_p}(t^*))} \tag{5}$$

where, $\mathbf{z}_{E_p}(t^*) = \{z_{j\in\Psi} = 1, z_{j\in\Xi\backslash p} = 0\}$.

Figure 3 illustrates two instances of evaluated MRL with respect to event p. Please note that as operation time passes, the predictor set may change, and at any t^*, the MRL is the area towards the right and under the survival function (shaded area in Fig. 3).

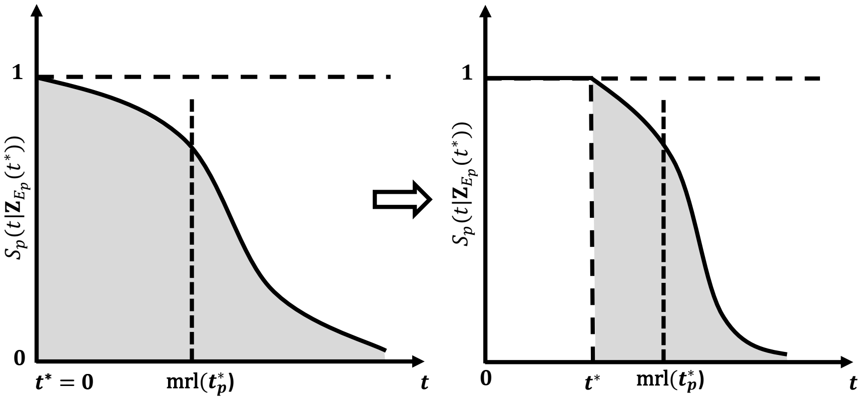


Figure 3: MRL for event p at unit's commission at $t^* = 0$ and at $t^* > 0$

As seen above, due to the binary nature of predictors, the Cox PH sets all predictors in Ψ to 1 and the remaining ones to 0, therefore, the actual observed sequence of events in Ψ is not captured. As a result, this modeling and prediction procedure is unable to distinguish the order of events in Ψ. As an example, consider if we have two sequences of events both leading to event-type E_3 as $\mathcal{T}^a : E_1 \prec E_2 \prec E_3$ and $\mathcal{T}^b : E_2 \prec E_1 \prec E_3$, then $Z_1 = Z_2 = 1$ for both of these sequences.

In a typical scenario, for any unit which experiences K events, there are 2^{K-1} combinations of predictor variables, where as, there are a total of $K!$ possible permutations of event occurrence for the unit, suggesting that there is under-utilization of available information in above described Cox PH model based prediction. We propose the following Copula based approach to address this issue.

3 Copula-based events modeling and prediction

3.1 General joint distribution fitting using Copula

Consider we have K different random variables as $X_1, X_2, \ldots X_K$ with the kth random variable having density $X_k \sim f_k(x; \boldsymbol{\theta}_k)$. Then, Sklar's theorem allows us to create joint distribution using a specified Copula function $C(\boldsymbol{\alpha})$ with parameter $\boldsymbol{\alpha}$, and marginal distribution functions, $F(x_k; \boldsymbol{\theta}_k)$ (Sklar, 1959). The variables are transformed to probability space (or $u-$space) through their respective distribution function as

$$u_k = F_k(x_k; \boldsymbol{\theta}_k) \tag{6}$$

where, $u_k \in \mathcal{U}[0, 1]$. Thereafter, a K-dimensional joint distribution using Copula $C(\boldsymbol{\alpha})$ can be obtained as

$$F(x_1, x_2, \ldots, x_K; \boldsymbol{\alpha}) = C(u_1, u_1, \ldots, u_K; \boldsymbol{\alpha}) \tag{7}$$

We also obtain the joint density function as

$$f(x_1, x_2, \ldots, x_K; \boldsymbol{\alpha}) = c(u_1, u_2, \ldots, u_K; \boldsymbol{\alpha}) \prod_{k=1}^{K} f_k(x_k) \tag{8}$$

where, $c(u_1, u_2, \ldots, u_K; \boldsymbol{\alpha}) = \frac{\partial C(u_1, u_2, \ldots, u_K)}{\partial u_1 \partial u_2 \ldots \partial u_K}$ is Copula density function.

The parameters in the Copula model to be estimated contain individual density function parameters $\boldsymbol{\theta}_k$s and the Copula parameter $\boldsymbol{\alpha}$. The likelihood across set of N independent observations can be written as follows:

$$L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\alpha}; \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K) = \prod_{i=1}^{N} f(x_{1i}, x_{2i}, \ldots, x_{Ki}; \boldsymbol{\alpha}) \tag{9}$$

The above likelihood can be maximized using EM algorithm to obtain the parameters, however, if the variables are time-to-event random variables with censoring, then parameter estimation becomes challenging (Leung *et al.*, 1997). Shih & Louis (1995) present Copula estimation method for time-to-event variables with the presence of censoring for a bivariate case. The essential idea is to use $1 - S(\cdot)$ to estimate the marginal distribution $F(\cdot)$, where $S(\cdot)$ can be estimated with the presence of censoring. Alternatively, there exists concept of *survival Copula* which directly uses $S(\cdot)$ as the input to the Copula function. Survival

Copulas, $\tilde{C}(.)$ can be obtained from the conventional Copula function through variable transformation, $C(.)$ and vice versa (see section 2.6 in Nelsen (2007)).

3.2 Copula construction based on Cox PH regression

The key idea of our proposed model is to use Cox PH regression models fitted from the historical multi-event sequences to obtain the marginal distribution of the time-to-event variable for each event type. Then these marginal distributions are used to construct the Copula distribution. Through this integrated framework we are able to take advantage of flexibility of Cox PH model while the Copula model can capture the subtle relationships among the events (e.g., the actual event times, not only an encoded binary variable are considered). The idea of using Cox PH model to obtain marginal distribution of time-to-event variables has been used in medical applications (Massonnet *et al.* (2009); Othus & Li (2010)). Specifically, a variable X_k in Eq. (7) becomes the time-to-event variable T_k for event-type E_k and Eq. (7) becomes

$$F_{\boldsymbol{T}}(t_1, t_2, \ldots, t_K) = C(u_1, u_2, \ldots, u_K; \boldsymbol{\alpha}) \tag{10}$$

where $u_k = F_k(t_k) = 1 - S_k(t_k)$ and $S_k(t_k)$ is obtained from the Cox PH model for event-type k as described in the previous section.

Many choices for Copula function are available in literature (Nelsen, 2007) and it has also been well-studied by researchers (see Genest *et al.* (2009) for review). However, we limit our choice to Gaussian Copulas in this work, for two main reasons - (i) we can estimate the Copula parameter in presence of censoring with the help of a quick heuristic (later in this section), and (ii) for practitioners, Gaussian Copula is easy to interpret. Please note any other Copula functions, or vine-Copula structure can also be used instead of Gaussian Copula, however their modeling is often complex and difficult to interpret. The parameter for the Gaussian Copula is the correlation matrix which we denote using $\Sigma_{\mathbf{K}}$ with $\{\Sigma_{\mathbf{K}}\}_{kl} =$

$\rho_{kl}, \{k,l\} \in \mathbf{K}$. The joint distribution from Eq. (10) is given as:

$$\begin{aligned} F_{\mathbf{T}}(t_1, t_2, \ldots, t_K) &= C(u_1, u_2, \ldots, u_K; \Sigma_{\mathbf{K}}) \\ &= \Phi_{\Sigma_{\mathbf{K}}}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_K)) \end{aligned} \tag{11}$$

where, $\Phi_{\Sigma_{\mathbf{K}}}$ is multivariate Gaussian distribution with parameter $(\mathbf{0}, \Sigma_{\mathbf{K}})$, and $\Phi^{-1}(\cdot)$ is inverse of standard Gaussian distribution.

Since we use Cox PH regression functions as the marginals, the overall likelihood to be maximized contains the marginal parameters (Θ from section 2.2) and Copula parameter ($\Sigma_{\mathbf{K}}$). We can write out the overall likelihood as a product of likelihood for each unit present in historical dataset. Further, for the unit i, let $\Psi_i \subseteq \mathbf{K}$ represent the set of events which has been observed, and Ξ_i is the set of unobserved/censored events, then the contribution of this unit towards the likelihood will depend on these sets, given below:

$$L(\Theta, \Sigma_{\mathbf{K}}; \mathbf{D}) \propto \prod_{i=1}^{N} \frac{\partial^{|\Psi_i|}}{\partial \boldsymbol{u}_{\Psi_i}} \tilde{C}(1 - u_{1i}, 1 - u_{2i}, \ldots, 1 - u_{Ki}; \Sigma_{\mathbf{K}}) \tag{12}$$

where the function $\tilde{C}(\cdot)$ is the survival Copula function that can be obtained from the conventional Copula function through a variable transformation. The derivation of (12) can be found in appendix A. The Copula function parameters $\Sigma_{\mathbf{K}}$ are explicit in the likelihood function (12). The Cox PH model parameters Θ are implicit in the likelihood function as when we transform the raw time-to-event data t_{ki} to u_{ki}, we will need those parameters Θ.

It is generally very difficult to simultaneously estimate all the parameters in (12) due to the large number of unknown parameters. Here we adopt the method of two-step estimation or Inference Functions for Margins (IFM) (Shih & Louis, 1995; Joe & Xu, 1996). Although the method may cause bias in the estimation, the computational benefits make the method highly attractive. The choice of IFM is also supported by the literature (Georges *et al.*, 2001; Othus & Li, 2010). The first step of IFM is to estimate marginal parameters (Θ in our case) using any procedure of choice. Once $\hat{\Theta}$ is obtained, in the second step, variables are transformed from $T-$space to $u-$space and plugged into Copula density and subsequently optimized to obtain the Copula parameters. The steps in detail is as follows

1. Obtain $\hat{\Theta}$ by maximizing $L_k \quad \forall k \in \mathbf{K}$ for the base Cox PH model using $\mathbf{D}$ (see section 2.2).

2. Transform observed event-times present in the historical dataset to probability space using the corresponding estimated survival function and predictors. For a unit i and event-type k, the event-time t_{ki} is mapped as $\hat{u}_{ki} = 1 - \hat{S}_k(t_{ki})$, with predictors for the Cox PH model as $\mathbf{z}_{E_k i} = \{\underset{j:t_{ji}<t_{ki}}{z_j = 1},\ \underset{j':t_{j'i}>t_{ki}}{z_{j'} = 0}\}$.

3. Substitute the transformed variables, $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_K$ in Eq. (12) and then estimate the Copula parameter $\hat{\Sigma}_{\mathbf{K}}$. The likelihood to estimate Copula is

$$L(\Sigma_{\mathbf{K}}; \hat{u}_1, \hat{u}_2, \ldots, \hat{u}_K) \propto \prod_{i=1}^{N} \frac{\partial^{|\Psi_i|}}{\partial \boldsymbol{u}_{\Psi_i}} \tilde{C}(1 - \hat{u}_{1i}, 1 - \hat{u}_{2i}, \ldots, 1 - \hat{u}_{Ki}; \Sigma_{\mathbf{K}}). \qquad (13)$$

The above likelihood can be maximized directly to estimate $\hat{\Sigma}_{\mathbf{K}}$. However, if K is large then estimation can be computationally expensive. Alternatively, we can use a pair-wise approach as a heuristic to obtain the correlation matrix ($\hat{\Sigma}_{\mathbf{K}}^{PW}$). Compared to full likelihood, the pair-wise approach is quick and easy to implement. The pair-wise approach can also be found in literature (for example, Mehrotra (1995) in missing data, and Pesonen *et al.* (2015) in left censored data). In the current work, we use estimated $\hat{\Sigma}_{\mathbf{K}}^{PW}$ as a starting point for optimizing the full likelihood. The details of estimating $\Sigma_{\mathbf{K}}^{PW}$ is provided in appendix B.

3.3 Event prediction using the Copula based model

We can use the multivariate joint distributions estimated from Copula to obtain event predictions. The basic idea is to obtain the conditional-marginal distribution of the event-type we want to predict from the joint distribution and the observed predictors up to now. Then the mean of the conditional-marginal distribution can be used as the event-time's prediction. The steps for prediction using integrated Copula and Cox PH model are as follows.

1. Assume we want to make prediction for a specific unit. For all the observed events $\psi \in \Psi$ from this unit, convert each event-time T_ψ to probability space as $u_\psi = 1 - \hat{S}_\psi(t_\psi)$, where $\hat{S}_\psi(t_\psi)$ is obtained from the corresponding Cox PH model. We collectively write them as $\boldsymbol{u}_\Psi$.

2. For event $p \in \Xi$ to be predicted for this unit, condition the joint Copula distribution by $\boldsymbol{u}_\Psi$ and obtain the marginal of p in probability space. To do so, we first obtain the conditional distribution as (derivation in appendix C)

$$C(\boldsymbol{u}_\Xi; \hat{\Sigma}_{\Xi|\Psi}) = \Phi_{\hat{\Sigma}_{\Xi|\Psi}}\Big\{\Big((\Phi^{-1}(\boldsymbol{u}_\Xi) - \hat{\Sigma}_{\Xi\Psi}\hat{\Sigma}_{\Psi\Psi}^{-1}\Phi^{-1}(\boldsymbol{u}_\Psi))\backslash\sqrt{\mathcal{D}(\hat{\Sigma}_{\Xi|\Psi})}\Big); \hat{\Sigma}_{\Xi|\Psi}\backslash\mathcal{D}(\hat{\Sigma}_{\Xi|\Psi})\Big\} \tag{14}$$

Then, conditional-marginal density for p denoted as $u_p|\boldsymbol{u}_\Psi$ can be worked out from $C(\boldsymbol{u}_\Xi; \hat{\Sigma}_{\Xi|\Psi})$ as follows

$$\begin{aligned} u_p|\boldsymbol{u}_\Psi &= \operatorname*{marg}_p C(\boldsymbol{u}_\Xi; \hat{\Sigma}_{\Xi|\Psi}) \\ &= \Phi_{\hat{\Sigma}_{p|\Psi}}\Big((\Phi^{-1}(u_p) - \hat{\Sigma}_{p\Psi}\hat{\Sigma}_{\Psi\Psi}^{-1}\Phi^{-1}(\boldsymbol{u}_\Psi))/\sqrt{(\hat{\Sigma}_{\Xi|\Psi})_{\{pp\}}}\Big) \end{aligned} \tag{15}$$

We can see in (15) that if there are no events to condition upon, then $u_p|\boldsymbol{u}_\Psi$ is simply equal to $\Phi\Big(\Phi^{-1}(u_p)\Big) = u_p \sim \mathcal{U}[0,1]$.

3. We can transform the distribution $u_p|\boldsymbol{u}_\Psi$ back to the $T-$space as following. First the survival function to be used from Eq. 4 is

$$u_p|\boldsymbol{u}_\Psi = 1 - \hat{S}_p(t > t^*) \tag{16}$$

with predictors, $\mathbf{z}_{E_p}(t^*) = \{z_{j\in\Psi} = 1, z_{j\in\Xi\backslash p} = 0\}$.

4. By inverting the survival function we get

$$t_p = \hat{S}_p^{-1}(1 - u_p | \boldsymbol{u}_\Psi), \tag{17}$$

The expected time of event is, $\text{mrl}_p(t^*) = \mathbf{E}t_p$.

We use a simple example to illustrate the prediction process. Consider three event-types: E_1, E_2, E_3, where E_1 shares mild association with the E_2 and E_3, and E_2, E_3 are strongly associated. Let us further assume that we have obtained the model parameters from historical dataset and we are particularly interested in the prediction of E_3. If E_3 is the last event being experienced, then the two possible sequences leading to occurrence of the third event are: $\mathcal{T}^a : E_1 \prec E_2 \prec E_3$ and $\mathcal{T}^b : E_2 \prec E_1 \prec E_3$. The prediction begins at $t^* = 0$ and Fig. 4 depicts the marginal distributions of variables in $u-$space as the unit continues to experience the events.

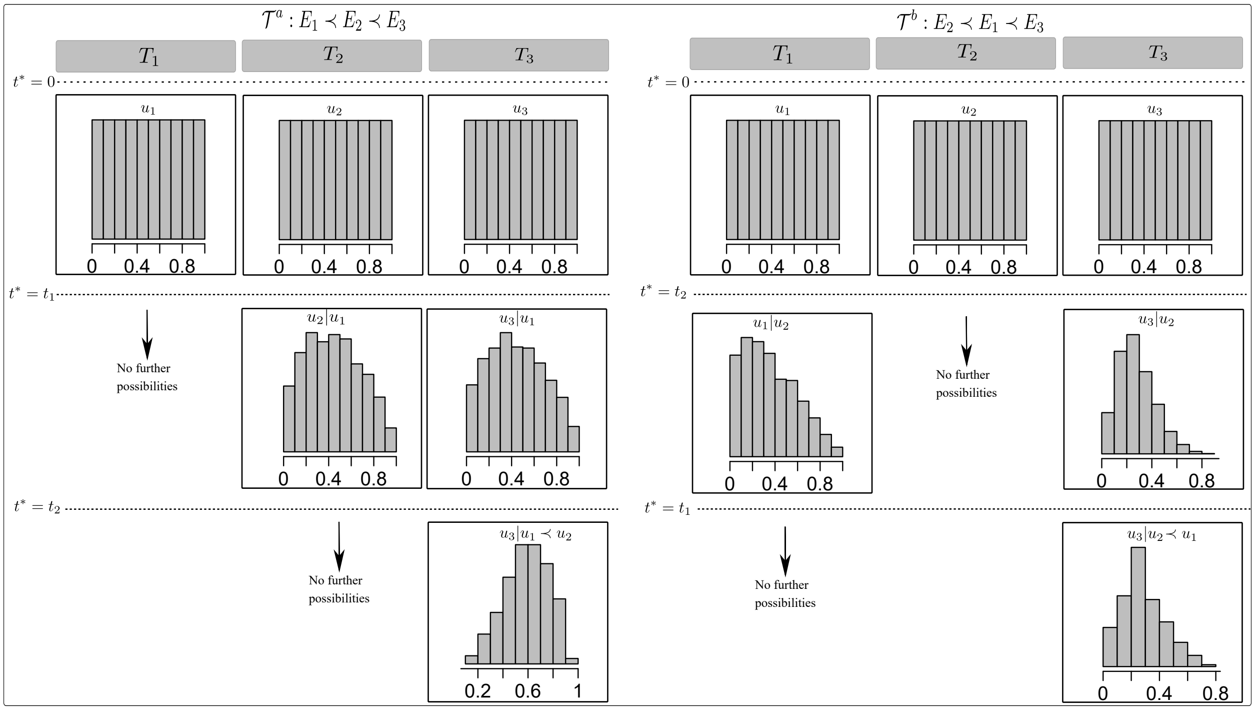


Figure 4: Marginal distributions of variables in $u-$space upon conditioning for two sequences.

As shown in Fig. 4, at $t^* = 0$ when the unit is fresh and has not experienced any event, then there is no event to condition on. The obtained marginals will be standard uniform. When these marginals are transformed back to original time-to-event space, the expected event times will be exactly identical to that from the corresponding base Cox PH model (In fact, these expected event times will be determined by the baseline hazard function of the corresponding Cox PH model). As the unit begins to experience events, the conditional marginals of the events that have not occurred may not be standard uniform. In Fig. 4, consider sequence $\mathcal{T}^a$ where first event experienced is E_1 which shares the same level of correlation with E_2 and E_3, the marginal densities of u_2 and u_3 are very similar. On the other hand, for sequence $\mathcal{T}^b$, the first event experienced is E_2 and it shares different level of correlation with E_1 and E_3, the corresponding densities are substantially different. The above features lead to the distinct predictions for these two sequences. Please note that the prediction only using the base Cox PH model does not have this feature.

4 Numerical Study based on Simulated Data

In this section, we use simulated data to evaluate the performance of the proposed approach. One important issue is how to generate correlated multi-event data. We adopted two simulation settings as described in the following sub-section.

4.1 Different experiment settings and data generation process

In setting (I), we generate sequence of multiple types of events through regression. These event-types are manifestations of some underlying root causes or factors. For this experiment setting we can draw an analogy, where these factors, in general depict working conditions of the industrial equipment (temperature, operation-hours etc.). And, as the equipment operates certain conditions or their combinations can have different effect on the event-times being observed. Mathematically, let n_f be the number of underlying factors which affect the occurrence of event-times. Corresponding to each factor, we further have the coefficient vector as $\boldsymbol{\gamma}$. These factors are associated with event's hazard in the formulation same as Cox PH (2) but with a specified and common baseline, i.e.

$$\lambda_k(t) = \lambda_{0k}(t)\exp(\boldsymbol{\gamma}_k^T\mathbf{M}) \tag{18}$$

where, $\mathbf{M}$ is set of common predictor variables sampled from standard uniform distribution, and the baseline follows Weibull distribution with parameters (b, ν). Thus, the event time vector of k-th event-type can be generated using following equation (Bender *et al.*, 2005):

$$T_k = \left[\frac{-\log \mathcal{U}}{b\exp(\boldsymbol{\gamma}_k^T\mathbf{M})}\right]^{1/\nu} \tag{19}$$

where, $\mathcal{U}$ is standard uniform distribution.

Next, we generate K different event-types and to distinguish them, we change the coefficient vectors resulting in different magnitude of effect of the underlying factors. In detail, we consider $n_f = 3$ and generate eight different event-types ($K = 8$). The covariates considered are enlisted in table 2. Since we set some covariates' coefficients as zero, the result can be treated as absence of the respective factors. We provide the algorithm for data generation for setting I in procedure 1.

Table 2: Covariate coefficient values for setting I

$\boldsymbol{\gamma}$	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8
γ_1	0	-1	0	0	-1	0	-1	-1
γ_2	0	0	-1	0	-1	-1	0	-1
γ_3	0	0	0	-1	0	-1	-1	-1

Algorithm 1 Data generation for setting I

```
Event time generation:
Initialize number of units $N = 100$, baseline parameters $b = 2$ and $\nu = 1.2$, model matrix $\mathbf{M}$ of dimension $N \times n_f$, and covariate coefficient vectors as in Table 2.
For each event-type, generate event-time vector using equation 19.
Censoring:
For each $i \in N$, generate censoring/follow-up times ($\tau_i$) of each unit. Let $\tau_i \sim Weib(.)$. We set $\tau_i \sim Weib(5, 5)$.
For each $i \in N, \delta_{ki} = 0$ if $T_{ki} > \tau_i,$ else $\delta_{ki} = 1, \forall k \in \mathbf{K}$.
Update $T_{ki} = \min(\tau_i, T_{ki})$.
```

As we can see, that each unit shares same observation values, but each event-type is output of different level of coefficient values. Therefore, we expect that while these event-types are different, they also share some commonality. Since our motivation is to highlight the association in event-types, we do not use the observation matrix ($\mathbf{X}$) during modeling and prediction in the experiments.

Moving to experiment setting (II), in this case we generate event-times from multivariate weibull distribution as presented in (Hougaard, 1986). We perform this experiment with four event-types. Interestingly, under different parametrization, it is equivalent to consider a Gumbel-Hougaard Copula (Hutchinson, 1990) with association parameter α. To generate the event times under this setting, we need to specify the association parameter and individual marginal distributions of event-times. We consider three sub-cases by varying the level of α. We keep $\alpha = 1$ denoting independence and other two values are 1.1 and 1.25. For each event-type we set marginals as $Weib(1, 5)$. In this setting, only one parameter controls the association across all variables.

The steps involved in data generation for setting (II) are:

Algorithm 2 Data generation for setting II

```
    *Event time generation:*
1:  Initialize number of units $N = 50$, specify $K = 4$ event types, underlying association structure through Copula $C(\alpha)$ and marginal distributions.
2:  Follows steps in Nelsen (2007) section 2.9 to generate observations in $u-$space.
3:  Transform variables from $u-$space to $T-$space through respective marginal distributions.
    *Censoring:*
4:  Follow steps 3 through 5 from algorithm 1. We set $\tau_i \sim Weib(1, 50)$.
```

4.2 Numerical Experiment

The overall study is in two steps: first, we use some portion of data to estimate the model parameters. This known data is analogous to historical dataset available in teleservice systems. It is assumed that the estimated parameters are very close to real estimates, therefore, for a new test-unit, which belong to the same family share the same governing parameters. In the second step, for each test-unit, we begin the prediction procedure from $t^* = 0$, which can be treated as the unit's commission to service. At any given point of time, we predict the estimated time for each event which this test-unit has not experienced yet. We record the absolute error (AE) in prediction times. For event p, if T_p is the realized event time and $\text{mrl}_p(t^*)$ is the predicted time of event, then $AE(t^*)_p = |\text{mrl}_p(t^*) - T_p|$. Due to censoring, it may however happen that any particular event is not realized for the test-unit, in that case we do not assess the prediction performance for this particular event.

We choose leave-one-out cross-validation (LOO-CV) method as our approach to evaluate and compare the prediction errors (Stone, 1974). Algorithm 3 provides the broader outline of the adopted cross validation procedure for a Cox PH regression based modeling and prediction model. The cross-validation procedure is same for the proposed model except step 10 is now obtained from section 3.3. The performances are compared for each base model against its respective integrated version. Please note that in regression model we only use other event-types as covariates. In the next section we discuss the results obtained.

Table 3: Prediction performance of competing models under setting I; Mean Absolute Error (standard deviation)

Prediction methodology	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8
Cox PH	0.215	0.295	0.296	0.306	0.419	0.456	0.495	0.738
	(0.196)	(0.256)	(0.304)	(0.232)	(0.409)	(0.416)	(0.367)	(0.591)
Cox PH with Copula	**0.182**	**0.249**	**0.197**	**0.259**	**0.289**	**0.344**	**0.334**	**0.279**
	(0.162)	(0.266)	(0.141)	(0.259)	(0.445)	(0.406)	(0.362)	(0.211)
Percentage improvement	9%	10%	25%	18%	33%	47%	42 %	54%

Algorithm 3 LOO-CV for Cox PH regression based prediction

```
for $m$ in $1 : N$ do
  Offline stage
    $m \leftarrow$ test unit
    $Hist \leftarrow N \setminus m$
    Estimate $\hat{\Theta}$ using section 2.2 and $Hist$
  Online stage
    while $|\Psi| < K$ do
        $t^* \leftarrow \max(0, \max(T_\Psi))$
        Estimate $\text{mrl}_p(t^*) = \int \hat{S}_p(t > t^* | \mathbf{z}_\Psi = 1, \mathbf{z}_{\Xi \setminus p} = 0) dt_p$ for each $p \in \Xi$
        $p \leftarrow$ event realized
        For $p$: $AE_{pm} = |T_{pm} - \text{mrl}_p(t^*)|$
        $\Xi \leftarrow \Xi \setminus p$
        $\Psi \leftarrow \Psi \cup p$
    end while
end for
```

4.3 Results and Discussion

In the two experiment settings that have been considered, the proposed model outperforms the respective regression model. The trends for all of the experiment cases justify that the proposed method not only enjoys the information present in the individual regression models, but it also adequately utilizes the information offered from the joint distribution. Moreover, the performance also highlights that a sophisticated regression model can be topped with extra information from the joint and improve the overall prediction. Below we discuss the results and insights of each experiment setting one-by-one.

For the setting (I), time variables are generated through regression functions. From this

Table 4: Prediction performance of competing models under setting II; Mean Absolute Error (standard deviation)

Case	Prediction methodology	E_1	E_2	E_3	E_4
$\alpha = 1$	Cox PH	1.982 (2.362)	1.340 (1.579)	1.179 (1.068)	2.148 (2.088)
	Cox PH with Copula	**1.703** (1.906)	**0.995** (1.155)	**1.045** (1.126)	**1.623** (1.650)
	Percentage improvement	14%	26%	11%	24%
$\alpha = 1.1$	Cox PH	1.569 (1.248)	2.112 (2.168)	1.765 (2.068)	3.256 (3.812)
	Cox PH with Copula	**1.389** (1.347)	**1.652** (1.968)	**1.335** (1.392)	**2.992** (4.037)
	Percentage improvement	12%	22%	24%	8%
$\alpha = 1.25$	Cox PH	1.606 (1.812)	1.552 (1.249)	2.004 (2.413)	1.764 (1.975)
	Cox PH with Copula	**1.518** (1.454)	**1.504** (1.276)	**1.489** (1.270)	**1.434** (1.432)
	Percentage improvement	5%	3%	26%	19%

setup, we gather several insights related to prediction performance of different predictive models. First, it is worth noting that as we observe from event-type 1 through 8, the several factors alter the overall event-time densities. This alteration is also reflected in the error distributions as they increase (see table 3).

Second, in general, table 3 show that for any regression model, our method provides considerable decrement in error values. The observation can be explained by the fact that Cox PH models are indeed limited by the binary nature of the predictors, which renders the real observation times to be masked. On the other hand, in our proposed framework, we retain this information when we build the joint (since u is continuous). Consequently, during prediction we are able to condition upon the observed event times.

Third, the variation of error distribution is generally less for the joint model against the respective regression model which further suggests that the proposed model performs better in presence of extreme values.

The second setting assumes multivariate weibull distribution as the true underlying model. Despite this fact, we find that the employed Gaussian Copula sufficiently captures the dependence among the time variables and provide improved predictions (see table 4). Here, parameter α characterizes the measure of dependence among the variables. For the case where event-times are independent to each other ($\alpha = 1$), the correlation parameters in Gaussian Copula estimated from historical dataset are non-zero. The underlying reason is that we transform variables through regression models which share several common covariates. Thus, the $u-$s obtained are correlated and the parameter is not necessarily zero. As a result, during prediction, the Copula density when conditioned does not have standard uniform distribution – thus giving better prediction estimates.

5 Real case study

5.1 Data description

In this section, we demonstrate the efficacy of our proposed method on a real world data collected from an industrial equipment. There are multiple sensors onboard the equipment which monitor the unit's operation. The units are connected wirelessly and they immediately send out warning data whenever an event has occurred. Due to confidentiality we do not describe the warning signals in current text. The dataset is collected from 40 units and four different kinds of events are recorded. Table 1 presents a part of the dataset. Please note that the event times are not the calendar times, rather it is the time since the unit began operating. As is the characteristic of the event data, not all of the units experienced all of the events and certain amount of censoring was present. Particularly, E_1 was censored for roughly a quarter of the units, E_2 and E_4 were censored for a fifth of the units and E_3 did not occur for roughly a third of the units in operation.

5.2 Performance evaluation

We fit Cox PH regression models for each individual event-type. The summary of model fits is reported in Table 5. Thereafter, we convert observed event times through respective Cox

PH models and estimate Copula parameter while adjusting for censoring which is mentioned in Table 6. The non-zero correlation parameters indicate the presence of association between the transformed variables.

Table 5: Obtained coefficient values with respective standard error for different Cox PH regression models for the real case data

Model for	β_1	β_2	β_3	β_4
$k = 1$	-	0.89 (0.30)	0.15 (0.40)	-0.14 (0.33)
$k = 2$	0.34 (0.31)	-	-0.47 (0.44)	0.31 (0.31)
$k = 3$	-0.01 (0.29)	0.36 (0.3)	-	-0.27 (0.3)
$k = 4$	0.14 (0.29)	-0.07 (0.29)	-0.16 (0.34)	-

Table 6: Correlation among transformed event-times through Cox PH regression models

	E_1	E_2	E_3	E_4
E_1	1	0.25	0.39	0.01
E_2	0.25	1	0.23	0.09
E_3	0.39	0.23	1	-0.30
E_4	0.01	0.09	-0.30	1

We apply the LOO-CV on the real case dataset and compare the proposed framework against Cox PH. The results are reported in Table 7 and we find that our proposed method outperforms the respective regression models for all event-types.

Table 7: Prediction performance of competing models on real case data; Mean Absolute Error (standard deviation)

Prediction methodology	E_1	E_2	E_3	E_4
Cox PH	46.604 (27.715)	39.377 (29.186)	44.255 (27.580)	44.602 (23.709)
Cox PH with Copula	**38.824** (25.980)	**36.957** (22.939)	**41.633** (31.961)	**41.247** (26.128)
Percentage improvement	17%	6%	6%	8%

6 Conclusion and future directions

To conclude, in this paper, we presented a framework to make prediction in a multi-type event scenario. Cox PH regression functions although very flexible and robust in modeling, are strongly limited by the binary formulation of the predictors during prediction. The framework presented in this paper creates a joint with marginals being Cox PH regression functions using Gaussian Copula, and, during prediction, an extra level of conditioning extracts more information from the unit's history. As a result the method adequately distinguishes between the event sequence and provides a curtailed and better prediction estimates.

There are two major future research directions for the current work: (i) extend the work by generalizing it to accommodate event's recurrence. One straightforward way to achieve this is by creating dummy variables, however, there are two main challenges: first, a future recurring event can only occur after the latest event, thus the risk period will be different, and second challenge, mainly from the practice perspective is that during the test period the event count may exceed the limited dummy variables created in the training dataset. Another way to solve this challenge is by using a *gap-time* approach – which resets the time after event's occurrence – however, in an engineering context where equipments age with operation such assumption might not be right. (ii) Often it happens that events or warnings are ambiguous indicators of several underlying root-cause. In this scenario, using a fixed Copula structure might not offer the best prediction estimates. This requirement can be met by adopting a suitable Bayesian approach and making the Copula parameter a random variable.

References

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

Barthel, Nicole. 2015. Multivariate Survival Analysis using Vine-Copulas.

Bender, Ralf, Augustin, Thomas, & Blettner, Maria. 2005. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, **24**(11), 1713–1723.

Chen, Xiaohong, Fan, Yanqin, Pouzo, Demian, & Ying, Zhiliang. 2010. Estimation and model selection of semiparametric multivariate survival functions under general censorship. *Journal of econometrics*, **157**(1), 129–142.

Cox, D. R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**(2), 187–220.

Genest, Christian, Rémillard, Bruno, & Beaudoin, David. 2009. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and economics*, **44**(2), 199–213.

Georges, Pierre, Lamy, Arnaud-Guilhem, Nicolas, Emeric, Quibel, Guillaume, & Roncalli, Thierry. 2001. Multivariate survival modelling: a unified approach with copulas.

Hougaard, P. 2000. *Analysis of Multivariate Survival Data*. Statistics for Biology and Health. Springer New York.

Hougaard, Philip. 1986. A class of multivanate failure time distributions. *Biometrika*, **73**(3), 671–678.

Hutchinson, Timothy P. 1990. *Continuous bivariate distributions emphasising applications*. Tech. rept.

Joe, Harry, & Xu, James Jianmeng. 1996. The estimation method of inference functions for margins for multivariate models.

Klein, John P, & Moeschberger, Melvin L. 2005. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Lee, Cheng K, & Wen, Miin-Jye. 2006. A multivariate weibull disitribution. *arXiv preprint math/0609585*.

Leung, Kwan-Moon, Elashoff, Robert M, & Afifi, Abdelmonem A. 1997. Censoring issues in survival analysis. *Annual review of public health*, **18**(1), 83–104.

Li, Zhiguo, Zhou, Shiyu, Choubey, Suresh, & Sievenpiper, Crispian. 2007. Failure event prediction using the Cox proportional hazard model driven by frequent failure signatures. *IIE transactions*, **39**(3), 303–315.

Marshall, Albert W, & Olkin, Ingram. 1967. A multivariate exponential distribution. *Journal of the American Statistical Association*, **62**(317), 30–44.

Massonnet, Goele, Janssen, Paul, & Duchateau, Luc. 2009. Modelling udder infection data using copula models for quadruples. *Journal of Statistical Planning and Inference*, **139**(11), 3865–3877.

Meeker, William Q, & Escobar, Luis A. 1998. *Statistical methods for reliability data*. John Wiley & Sons.

Mehrotra, Devan V. 1995. Robust elementwise estimation of a dispersion matrix. *Biometrics*, 1344–1351.

Mitsa, Theophano. 2010. *Temporal data mining*. Chapman and Hall/CRC.

Nelsen, Roger B. 2007. *An introduction to copulas*. Springer Science & Business Media.

Othus, Megan, & Li, Yi. 2010. A gaussian copula model for multivariate survival data. *Statistics in biosciences*, **2**(2), 154–179.

Pesonen, Maiju, Pesonen, Henri, & Nevalainen, Jaakko. 2015. Covariance matrix estimation for left-censored data. *Computational Statistics & Data Analysis*, **92**, 13–25.

Schemper, Michael, Kaider, Alexandra, Wakounig, Samo, & Heinze, Georg. 2013. Estimating the correlation of bivariate failure times under censoring. *Statistics in medicine*, **32**(27), 4781–4790.

Shih, Joanna H, & Louis, Thomas A. 1995. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 1384–1399.

Sklar, M. 1959. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, **8**, 229–231.

Stone, Mervyn. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, 111–147.

Wang, Weijing, & Ding, A Adam. 2000. On assessing the association for bivariate current status data. *Biometrika*, **87**(4), 879–893.

Yuan, Yuan, Zhou, Shiyu, Sievenpiper, Crispian, Mannar, Kamal, & Zheng, Yibin. 2011. Event log modeling and analysis for system failure prediction. *IIE Transactions*, **43**(9), 647–660.

A Likelihood construction

List of assumptions and notations

- $\mathbf{S}(\boldsymbol{s})$ represents a joint survival function. $\mathbf{S}(\boldsymbol{s}) = 1$ if $\mathbf{s} = \mathbf{0}$, and $\boldsymbol{S}(\boldsymbol{s}) = 0$ if $\boldsymbol{s} = \boldsymbol{\infty}$
- Censoring is independent and non-informative of failures. Censoring time for a unit is τ.
- Let g_c denote density function of censoring with $g_c(C \leq 0) = 0$ indicating that probability of unit being censored at the time of commission is zero.
- Bold characters denote the multivariate versions of variables and integrations.
- Using Sklar's representation:

$$f_{\mathbf{K}}(\boldsymbol{s}) = (-1)^K \frac{\partial \mathbf{S}(\mathbf{s_K})}{\partial \boldsymbol{s}_{\mathbf{K}}} = (-1)^K \frac{\partial \tilde{C}(S_1(s_1), S_2(s_2), \ldots, S_K(s_k))}{\partial \boldsymbol{s}_{\mathbf{K}}} = (-1)^K \frac{\partial \tilde{C}(S_{\mathbf{K}}(\boldsymbol{s}_{\mathbf{K}}))}{\partial \boldsymbol{s}_{\mathbf{K}}} \tag{20}$$

  where, $\tilde{C}(\cdot)$ is the survival Copula which can be obtained from Copula $C(\cdot)$ (Georges *et al.*, 2001, see Theorem 2 and 3).

We start by evaluating the joint probability of observed/censored times for one unit. Please note that for neatness we drop the unit's index. Also, $C = \{T, \delta = 0\}$ represent censored time and $T = \{T, \delta = 1\}$ represent the observed event time.

$$P[\boldsymbol{C}_\Xi \leq \boldsymbol{\tau}_\Xi, \boldsymbol{T}_\Psi \leq \boldsymbol{t}_\Psi, \boldsymbol{T}_\Xi > \boldsymbol{c}_\Xi] = \int\limits_{\mathbf{0}}^{\boldsymbol{\tau}} \Bigg[\int\limits_{\boldsymbol{c}}^{\boldsymbol{\infty}} \int\limits_{\mathbf{0}}^{\boldsymbol{t}_\Psi} f_{\mathbf{K}}(s_{\mathbf{K}}) \boldsymbol{ds}_{\mathbf{K}} \Bigg] \boldsymbol{g_c(c) dc_\Xi} \tag{21}$$

Working out the integrand inside the square brackets and using equation 20

$$\begin{align}
\int\limits_{\boldsymbol{c}}^{\boldsymbol{\infty}} \int\limits_{\mathbf{0}}^{\boldsymbol{t}_\Psi} f_{\mathbf{K}}(s_{\mathbf{K}}) \boldsymbol{ds}_{\mathbf{K}} &= \int\limits_{\boldsymbol{c}}^{\boldsymbol{\infty}} \int\limits_{\mathbf{0}}^{\boldsymbol{t}_\Psi} (-1)^K \frac{\partial \tilde{C}(S_{\mathbf{K}}(\boldsymbol{s}_{\mathbf{K}}))}{\partial \boldsymbol{s}_{\mathbf{K}}} \boldsymbol{ds}_{\mathbf{K}} \tag{22} \\
&\propto \int\limits_{\boldsymbol{c}}^{\boldsymbol{\infty}} \left. \frac{\partial \tilde{C}(S_{\mathbf{K}}(\boldsymbol{s}_{\mathbf{K}}))}{\partial \boldsymbol{s}_\Xi} \right|_{\mathbf{0}}^{\boldsymbol{t}_\Psi} \boldsymbol{ds}_\Xi \tag{23} \\
&\propto \int\limits_{\boldsymbol{c}}^{\boldsymbol{\infty}} \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{s}_\Xi), S_\Psi(\boldsymbol{t}_\Psi))}{\partial \boldsymbol{s}_\Xi} - \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{s}_\Xi), S_\Psi(\mathbf{0}))}{\partial \boldsymbol{s}_\Xi} \boldsymbol{ds}_\Xi \tag{24} \\
&\propto \int\limits_{\boldsymbol{c}}^{\boldsymbol{\infty}} \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{s}_\Xi), S_\Psi(\boldsymbol{t}_\Psi))}{\partial \boldsymbol{s}_\Xi} - \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{s}_\Xi), \mathbf{1})}{\partial \boldsymbol{s}_\Xi} \boldsymbol{ds}_\Xi \tag{25} \\
&\propto \left\{ \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{\infty}), S_\Psi(\boldsymbol{t}_\Psi))}{\partial \boldsymbol{s}_\Xi} - \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{c}), S_\Psi(\boldsymbol{t}_\Psi))}{\partial \boldsymbol{s}_\Xi} \right\} \tag{26} \\
&\quad - \left\{ \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{\infty}), \mathbf{1})}{\partial \boldsymbol{s}_\Xi} - \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{c}), \mathbf{1})}{\partial \boldsymbol{s}_\Xi} \right\} \notag
\end{align}$$

Using first assumption

$$\propto \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{c}), \mathbf{1})}{\partial \boldsymbol{s}_\Xi} - \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{c}), S_\Psi(\boldsymbol{t}_\Psi))}{\partial \boldsymbol{s}_\Xi} \tag{27}$$

Going back to equation 21

$$P[\boldsymbol{C}_\Xi \leq \boldsymbol{\tau}_\Xi, \boldsymbol{T}_\Psi \leq \boldsymbol{t}_\Psi, \boldsymbol{T}_\Xi > \boldsymbol{c}_\Xi] \propto \int\limits_{\mathbf{0}}^{\boldsymbol{\tau}} \Big[\frac{\partial \tilde{C}(S_\Xi(\boldsymbol{c}), \mathbf{1})}{\partial \boldsymbol{s}_\Xi} - \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{c}), S_\Psi(\boldsymbol{t}_\Psi))}{\partial \boldsymbol{s}_\Xi} \Big] \boldsymbol{g_c(c) dc_\Xi} \tag{28}$$

Next, to obtain the likelihood, we partially differentiate the above equation wrt K variables to get

$$L \propto \frac{\partial^K}{\partial \boldsymbol{t}_{\mathbf{K}}} \left\{ \int\limits_{\mathbf{0}}^{\boldsymbol{\tau}} \Big[\frac{\partial \tilde{C}(S_\Xi(\boldsymbol{c}), \mathbf{1})}{\partial \boldsymbol{s}_\Xi} - \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{c}), S_\Psi(\boldsymbol{t}_\Psi))}{\partial \boldsymbol{s}_\Xi} \Big] \boldsymbol{g_c(c) dc_\Xi} \right\} \tag{29}$$

$$\propto \frac{\partial^{|\Psi|}}{\partial \boldsymbol{t}_\Psi} \left\{ \frac{\partial^{|\Xi|}}{\partial \boldsymbol{t}_\Xi} \int\limits_{\mathbf{0}}^{\boldsymbol{\tau}} \Big[\frac{\partial \tilde{C}(S_\Xi(\boldsymbol{c}), \mathbf{1})}{\partial \boldsymbol{s}_\Xi} - \frac{\partial \tilde{C}(S_\Xi(\boldsymbol{c}), S_\Psi(\boldsymbol{t}_\Psi))}{\partial \boldsymbol{s}_\Xi} \Big] \boldsymbol{g_c(c) dc_\Xi} \right\} \tag{30}$$

$$\propto \frac{\partial^{|\Psi|}}{\partial \boldsymbol{t}_\Psi} \left\{ \Big[\tilde{C}(S_\Xi(\boldsymbol{\tau}), \mathbf{1}) - \tilde{C}(S_\Xi(\boldsymbol{\tau}), S_\Psi(\boldsymbol{t}_\Psi)) \Big] \boldsymbol{g_c}(\boldsymbol{\tau}) - \right. \tag{31}$$

$$\left. \Big[\tilde{C}(S_\Xi(\mathbf{0}), \mathbf{1}) - \tilde{C}(S_\Xi(\mathbf{0}), S_\Psi(\boldsymbol{t}_\Psi)) \Big] \boldsymbol{g_c}(\mathbf{0}) \boldsymbol{dc_\Xi} \right\}$$

$$\propto \frac{\partial^{|\Psi|}}{\partial \boldsymbol{t}_\Psi} \tilde{C}(S_\Xi(\boldsymbol{\tau}), S_\Psi(\boldsymbol{t}_\Psi)) = \frac{\partial^{|\Psi|}}{\partial \boldsymbol{t}_\Psi} \tilde{C}(S_{\mathbf{K}}(\boldsymbol{t}_{\mathbf{K}})) \tag{32}$$

Converting to $u-$notation using $u = 1 - S(t|\mathbf{Z})$

$$L \propto \frac{\partial^{|\Psi|}}{\partial \boldsymbol{u}_\Psi} \tilde{C}(1 - u_1, 1 - u_2, \ldots, 1 - u_K; \Sigma_{\mathbf{K}}) \tag{33}$$

The obtained likelihood also has a nice interpretation. As can be observed that we partially differentiate Copula function with respect to observed event-types – this means that events in Ψ contribute proportionally to their density function, whereas for the unobserved events the contribution towards final likelihood is proportional to their cumulative density function beyond the censored time.

B Deriving pair-wise (PW) correlation parameter of correlation matrix

We want to estimate the Copula parameter pair-wise $\Sigma_{\mathbf{K}}^{PW}$ which is:

$$\Sigma_{\mathbf{K}}^{PW} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots \\ \vdots & \ddots & \\ \rho_{K1} & & \rho_{KK} \end{bmatrix} \tag{34}$$

where, ρ_{kl} is the correlation between event-type k and l with $\rho_{kk} = 1$. Here, we estimate each correlation parameter separately. In other words, any pair of event-type $\{kl\}$ forms a two-dimensional Copula with parameter ρ_{kl}. The relation between Copula and survival Copula in bivariate case is: $\tilde{C}(1-u_k, 1-u_l) = 1 - u_k - u_l + C(u_k, u_l)$. Thus, the following likelihood can be derived using Eq. 13 as follows:

- Case 1: The unit i experienced both events k and l, then, $\delta_{ki} = \delta_{li} = 1, \Psi_i = \{k, l\}$. The likelihood contribution is

$$L \propto \frac{\partial^2}{\partial u_k \partial u_l} \tilde{C}(1 - \hat{u}_{ki}, 1 - \hat{u}_{li}; \rho_{kl}) \tag{35}$$

$$\propto c(\hat{u}_{ki}, \hat{u}_{li}; \rho_{kl}) \tag{36}$$

- Case 2: If unit i experienced only event k, then, $\delta_{ki} = 1, \delta_{li} = 0, \Psi_i = \{k\}$. The likelihood contribution using Eq. 13 is:

$$L \propto \frac{\partial}{\partial u_k} \tilde{C}(1 - \hat{u}_{ki}, 1 - \hat{u}_{li}; \rho_{kl}) \tag{37}$$

$$\propto \Big[1 - \frac{\partial C(\hat{u}_{ki}, \hat{u}_{li}; \rho_{kl})}{\partial u_k}\Big] \tag{38}$$

  Similarly we can work out the case when unit experienced only event l.

- Case 3: If the unit i did not experienced any event, then, $\delta_{ki} = \delta_{li} = 0$. The likelihood contribution is:

$$L \propto \tilde{C}(1 - \hat{u}_{ki}, 1 - \hat{u}_{li}; \rho_{kl}) \tag{39}$$

$$\propto \Big[1 - \hat{u}_k - \hat{u}_l + C(\hat{u}_{ki}, \hat{u}_{li}; \rho_{kl})\Big] \tag{40}$$

Combining these cases, the likelihood we obtain is:

$$\begin{aligned} L(\rho_{kl}; \hat{u}_k, \hat{u}_l) = \prod_{i=1}^{N} \Bigg[& c(\hat{u}_{ki}, \hat{u}_{li}; \rho_{kl})^{\delta_{ki}\delta_{li}} \\ & \times \left[1 - \frac{\partial C(\hat{u}_{ki}, \hat{u}_{li}; \rho_{kl})}{\partial u_k}\right]^{\delta_{ki}(1-\delta_{li})} \\ & \times \left[1 - \frac{\partial C(\hat{u}_{ki}, \hat{u}_{li}; \rho_{kl})}{\partial u_l}\right]^{(1-\delta_{ki})\delta_{li}} \\ & \times \Big[1 - \hat{u}_k - \hat{u}_l + C(\hat{u}_{ki}, \hat{u}_{li}; \rho_{kl})\Big]^{(1-\delta_{ki})(1-\delta_{li})} \Bigg] \end{aligned} \tag{41}$$

The above likelihood can be quickly maximized by parallelizing the estimation process. Once we have the estimates of the correlation parameters, $\hat{\Sigma}_{\mathbf{K}}^{PW}$ can be easily constructed giving us a joint distribution of transformed event-times.

C Conditional Gaussian Copula distribution

First, lets denote $T_k^{'} = \Phi^{-1}(1 - S(T_k))$, thus, $\{T'_1, T'_2, \ldots, T'_K\} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{K}})$. Let at any time of prediction t^* the set of observed events is Ψ and the remaining events are in Ξ. Then, correlation matrix can be split as

$$\Sigma_{\mathbf{K}} = \begin{bmatrix} \Sigma_{\Xi\Xi} & \Sigma_{\Xi\Psi} \\ \Sigma_{\Psi\Xi} & \Sigma_{\Psi\Psi} \end{bmatrix} \tag{42}$$

Next, we know that normal distribution remains closed under conditioning, thus

$$\{\boldsymbol{T'}_{\Xi}|\boldsymbol{T'}_{\Psi} = \boldsymbol{t'}_{\Psi}\} \sim \mathcal{N}(\boldsymbol{\mu}_{\Xi|\Psi}, \Sigma_{\Xi|\Psi}) \tag{43}$$

where, $\boldsymbol{\mu}_{\Xi|\Psi} = \mathbf{0} + \Sigma_{\Xi\Psi}\Sigma_{\Psi\Psi}^{-1}(\boldsymbol{t'}_{\Psi} - \mathbf{0})$ and $\Sigma_{\Xi|\Psi} = \Sigma_{\Xi\Xi} - \Sigma_{\Xi\Psi}\Sigma_{\Psi\Psi}^{-1}\Sigma_{\Psi\Xi}$. Further, let the variance (or diagonal) components of $\Sigma_{\Xi|\Psi}$ be denoted as $\mathcal{D}(\Sigma_{\Xi|\Psi})$. Now, to write the conditional distribution in form of Copula, first we normalize the conditional $\boldsymbol{T'}_{\Xi}|\boldsymbol{T'}_{\Psi} = \boldsymbol{t'}_{\Psi}$ as follows

$$(\boldsymbol{T'}_{\Xi} - \Sigma_{\Xi\Psi}\Sigma_{\Psi\Psi}^{-1}\boldsymbol{t'}_{\Psi}) \backslash \sqrt{\mathcal{D}(\Sigma_{\Xi|\Psi})} \tag{44}$$

where, symbol '\' denotes element wise division.

The $|\Xi|$ dimensional Copula then obtained is

$$C(\Sigma_{\Xi|\Psi}) = \Phi_{\Sigma_{\Xi|\Psi}}(\boldsymbol{T'}_{\Xi}|\boldsymbol{T'}_{\Psi} = \boldsymbol{t'}_{\Psi}; \Sigma_{\Xi|\Psi} \backslash \mathcal{D}(\Sigma_{\Xi|\Psi})) \tag{45}$$

In terms of $u-$notation

$$C(\Sigma_{\Xi|\Psi}) = \Phi_{\Sigma_{\Xi|\Psi}}\Big\{\Big((\Phi^{-1}(\boldsymbol{u}_{\Xi}) - \Sigma_{\Xi\Psi}\Sigma_{\Psi\Psi}^{-1}\Phi^{-1}(\boldsymbol{u}_{\Psi})) \backslash \sqrt{\mathcal{D}(\Sigma_{\Xi|\Psi})}\Big); \Sigma_{\Xi|\Psi} \backslash \mathcal{D}(\Sigma_{\Xi|\Psi})\Big\} \tag{46}$$