Statistical Modelling Seminar

Research Papers Summary: Deep Survival Models

Yedidia Agnimo, C. Yann Eric Choho

April 2023

Survival analysis is a field of statistics that aims to analyse the time it takes for an event of interest to occur, such as the time until a patient dies from a disease or the time until a mechanical component fails. In such studies, it is common to encounter *competing risks*, where multiple outcomes can prevent the event of interest from occurring. For example, in a study of the time until a patient dies from cancer, the patient may die from another cause before dying from cancer. Furthermore, during the observation period, it is possible that no events occurred. This is a common characteristic of survival data called *(right-)censoring*. We can also have *left-censoring*: when the true timing of the event for an individual is known to occur before a certain point in time. Censorship has to be accounted for because it still provides important information.

Main prior works model this survival time as the first hitting time of a stochastic process. The objective is then to estimate a survival function $\mathbb{S}(t|X) := \mathbb{P}(T > t|X)$ that estimate the probability that the event of interest or the risk has not occurred at a given time t, possibly conditional on some covariates X. With competing risk, the survival function is a joint probability called cumulative incidence function (CIF), $F_{k_0}(t|x) = \mathbb{P}(T \leq t, k = k_0|X = x)$ that expresses the probability that a particular event k_0 occurs on or before time t conditional to covariate x. Another fundamental concept is the hazard rate, denoted by $\lambda(t|X)$. It represents the instantaneous rate of occurrence of the event of interest or the risk at time t, given that the event has not yet occurred up to time t and conditional on covariates X. Mathematically, the hazard rate is defined as the limit of the conditional probability of the event occurring in a very short time-interval, given that it has not occurred up to time t, divided by the length of that interval, as the interval length approaches zero. That is,

$$\lambda(t|X) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t|T \geq t, X)}{\Delta t}$$

where T is the random variable denoting the time of occurrence of the event of interest or the risk. The hazard rate can be used to estimate the survival function, as $\mathbb{S}(t|X) = \exp\left(-\int_0^t \lambda(u|X)du\right)$. Also, it can be used to compare the risk of the event of interest or the risk between different groups, as it allows for the estimation of hazard ratios between two groups defined by different covariate values.

Different approaches have been developed based on different assumptions, which are their main limitation. Historically, the first remarkable attempt was the [Kaplan and Meier, 1958] estimator. It is a non-parametric method that involves constructing a step function that estimates the probability of survival at different time points based on the available data. The estimator calculates the probability of surviving to the next time point as the number of individuals still at risk who do not experience the event at that time point, divided by the total number of individuals still at risk. It is a powerful and flexible tool for analysing censored data, but it has limitations, including its inability to model time-dependent covariates and to account for the effects of competing risks.

Parametric and semi-parametric models make assumptions about the distribution of the underlying stochastic and about the form of the relationship between the covariates and the assumed process. The Cox proportional hazard model (CPH) [Cox, 1972] is the most popular reference that models the logarithm of the hazard rate as a linear function of covariates, assuming furthermore that it is constant over time. This means that the relative risk between two individuals is constant over time (hence proportional rate), which is sometimes a too strong assumption. Besides, CPH is not able to handle competing risks. Other models rely on different assumptions about the underlying stochastic process, such as the use of a Weibull distribution for the fat tail (mainly in finance), see [Lee et al., 2010] for more examples. An alternative to this approach is the Fine-Gray model used for modelling competing risks that focuses on the cumulative incidence function by extending the proportional hazards model to sub-distribution [Fine and Gray, 1999]. It is nevertheless limited by the way in which parameters depend on covariates.

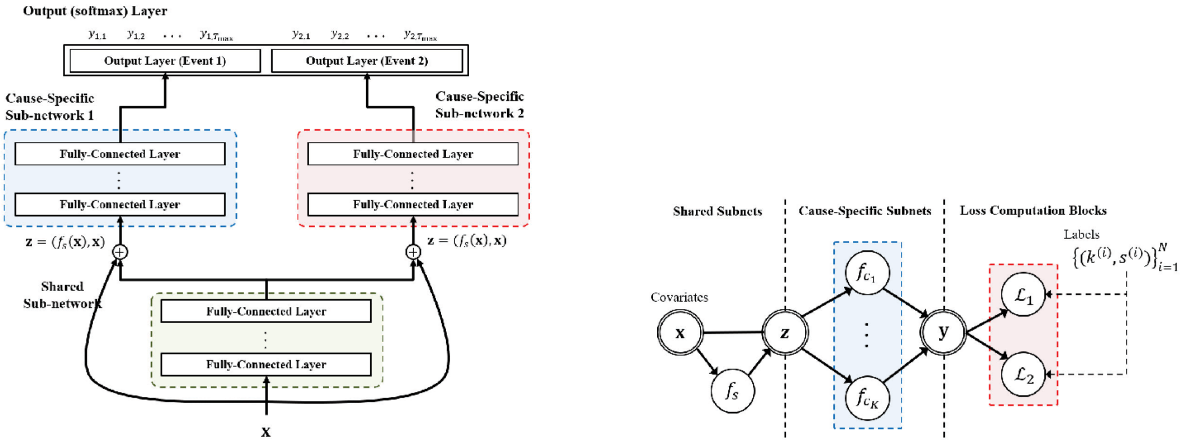
Besides, non-parametric approaches are mostly improvement of the Kaplan-Meier (KM) estimator, which is able to learn a more flexible survival curve but not able to incorporate covariates. In recent years, there has been considerable interest in applying machine learning techniques to the field of survival analysis. Several new models have been developed, including random survival forests [Ishwaran et al., 2008], deep exponential families [Ranganath et al., 2016], dependent logistic regressors [Yu et al., 2011], and semi-parametric Bayesian models based on Gaussian processes [Fernandez et al., 2016]. While these models are capable of incorporating individual patient covariates, none of them has yet tackled the challenge of competing risks. It is possible to treat all but one event as right-censored, but this approach is limited because competing risks are often not independent. A recent breakthrough was the development of a non-parametric Bayesian model for survival analysis with competing risks using a deep multi Gaussian multitasks by [Alaa and van der Schaar, 2017]. However, this model still assumes that the latent stochastic process follows a Gaussian process.

Since [Faraggi and Simon, 1995], a couple of papers have applied neural networks to survival analysis. *DeepSurv* from [Katzman et al., 2018] or [Luck et al., 2017] extends CPH to include non-linearity, replacing linear interaction in CPH by deep neural network and estimating directly the survival function instead of the hazard rate. However, this approaches improves upon the CPH model by relaxing its assumptions while still assuming that the hazard rate is constant over time. Despite this improvement, they have not yet harnessed the capacity of deep neural networks to capture complex representations of risk, including the time-varying effects of covariates on survival time.

To assess the predictive accuracy of a survival model, one common measure is the *concordance index* (a.k.a. c-index or Harrell's C). This index evaluates to what extent the predicted order of survival times agrees with the observed order of survival time, at period 1 (assuming a constant effect of covariates over time).

Three recent approaches aim to exploit the full potential of neural networks to estimate the survival function. Here is a summary of their main features. In the following, we are given a dataset $D := \{(x_i, t_i, k_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ are the covariates, $t_i \in \mathcal{T}$ is the survival time over a set of discrete and fixed horizons, and $k_i \in \mathcal{K} := \{\varnothing, 1, \dots, K\}$ are the K competing (exclusive) events, where $\varnothing$ denotes censoring (no event).

DeepHit [Lee et al., 2018] propose a flexible method called DeepHit, which uses a multi-task neural network to estimate the joint distribution of survival times and competing risks. They assume a discrete fixed horizon time $\mathcal{T} = \{0, \ldots, T_{\max}\}$, and that right-censoring occurs completely at random.



(a) The architecture of DeepHit with two competing events

(b) Computational graph to compute the training loss of DeepHit.

Figure 1: Illustration of the architecture and the training loss of DeepHit, source: [Lee et al., 2018]

The key innovation of DeepHit is its ability to handle both single and competing risks without making any assumption on the form of the relationship. This is achieved through a multi-task neural network that consists of a shared sub-network and risk-specific sub-networks. The shared sub-network captures the common latent representation of the covariates for all risks, while the risk-specific sub-networks estimate the probability of the first hitting of each event type. By directly learning the joint distribution, DeepHit allows for varying hazard rates and handles smoothly single and competing risks. Then output distribution is $\mathbf{y} = [y_{1,1}, \cdots, y_{1,T_{\max}}, \cdots, y_{K,1}, \cdots, y_{K,T_{\max}}]$ where $y_{k,t} = \hat{\mathbb{P}}(t, k|X = x)$ is the estimated probability that an individual will experience the event k at time t given its covariates x. Then, the cause-specific CIF is estimated by $\hat{F}_{k_0}(t|x) = \sum_{s=0}^{t} y_{k,s}$.

To train the model, the authors use a loss function that combines the log-likelihood of the joint distribution with a cause-specific ranking loss function $\mathcal{L}_{Total} = \mathcal{L}_1 + \mathcal{L}_2$, where

$$\mathcal{L}_1 = -\sum_{i=1}^{N} \left[\mathbf{1}(k_i \neq \varnothing) \cdot \log(y_{k_i,t_i}) + \mathbf{1}(k_i = \varnothing) \cdot \log\left(1 - \sum_{k=1}^{K} \hat{F}_k(t_i|x_i)\right)\right]$$

$$\mathcal{L}_2 = \sum_{k=1}^{K} \alpha_k \sum_{i \neq j} \mathbf{1}(k_i = k, s_i < s_j) \cdot \eta\left(\hat{F}_k(t_i|x_i), \hat{F}_k(t_i|x_j)\right),$$
$$\text{with } \eta(x, y) = \exp\frac{-(x-y)}{\sigma}$$

The weights α_k can be seen as a measure of the importance of each risk k, that can be adjusted depending on how critical each risk is. The parameters α_k and σ are learnt based on a discriminative performance on a validation set.

The performance metric used is the concordance index, modified to account for time dependence

(C^{td}-index). Specifically, at each time point t, the C^{td}-index calculates the proportion of pairs of individuals that are correctly ordered based on their predicted survival probabilities up to time t, using the predicted probabilities of survival at that time point. Experiments on both real data (SEER, UNOS, METABRIC)[1] and synthetic[2] dataset show that DeepHit outperforms state-of-the-art methods (Cox, Threshold Regression, Random Survival Forest 100 trees, DeepSurv, Mortality prediction combined with Random Forest, LogitR or AdaBoost) in terms of the concordance index, demonstrating the effectiveness of the proposed method in handling survival analysis with competing risks.

Deep Survival Machine (DSM) This paper [Nagpal et al., 2021b] is the first work involving fully parametric estimation of survival times with single event and competing risks in the presence of censoring. Indeed, the presented model, DSM, doens't make the strong assumption of proportional hazards and enables learning with time-varying risks. This method estimates the conditional survival function as a mixture of individual parametric survival distributions.

First, the input features are passed through a deep multilayer perceptron [Rumelhart et al., 1986] to learn both the covariate representation and the common representation for multiple risks. This representation is then used to model the conditional distribution as a weighted mixture over K well-defined, parametric distributions called primitive distributions drawn from some prior, and the K weights are a softmax over the output of a neural network.

Moreover, two distributions: Weibull and Log-Normal, in the paper, were chosen because they have a closed-form solution for the cumulative distribution function (CDF) which is desirable for Maximum Likelihood Estimation. Furthermore, the DSM loss function is designed to handle both censored and uncensored data and include the strength of the prior. A combined loss on all 3 elements is processed.

The performance of DSM against methods subject to proportional hazard assumption is demonstrated in an experiment, where the time-dependent Concordance Index is used to compare models. Deep Survival Machines have also the advantage to manage computational resources and can incorporate multiple sources of data, including complex modalities such as images.

Recurrent Deep Survival Machines (RDSM) [Nagpal et al., 2021a]

RDSM is an extension of DSM which focuses on adding time-varying covariates in survival modelling. Indeed, the event-time distribution may exhibit temporal dependencies at different time scales, which classical survival models that assume independence between training data points may not be able to capture effectively.

The proposed approach involves using Recurrent Neural Networks (RNN) like LSTMs and GRUs [Cho et al., 2014] to learn representations of input temporal data (instead of simple MLP in DSM), followed by describing the event distribution as a fixed mixture of parametric distributions. The recurrent neural networks can model long-term dependencies in the input data while estimating the Time-to-Event. Like the original DSM model, this approach assumes that the time of events is distributed as a mixture of parametric (Weibull or Log-normal) distributions. The parameters of these

[1] The SEER dataset contains cancer incidence and survival data, the UNOS dataset contains organ transplant data, and the METABRIC dataset contains breast cancer genomics data. SEER and SYNTHETIC datasets have two events while UNOS and METABRIC have a single event.

[2] SYNTHETIC dataset: 3 independant standard Gaussian covariates (x_1, x_2, x_3) and 2 exponential event times (T_1, T_2). x_1 and x_2 only affect T_1 and T_2, respectively, while x_3 affects both.

distributions are assumed to be functions of the learned representations and are jointly learned with the recurrent neural architectures. Here, the use of RNNs enables the learning of representations that retain knowledge from previous temporal steps.

To allow the RDSM to integrate streaming data, this approach takes into account the distribution of the time remaining until the event at each time step as a function of time. This contrasts with standard survival settings, where the distribution is treated as static. Modelling the distribution as a function of time allows capturing the time-varying effects of the input covariates. The paper introduces two assumptions. The first assumption is about independent censoring, which is required for identifiability. It assumes that the remaining time-to-event distribution is independent of censoring time given the observed covariates. The second assumption is about statefulness, which assumes that the distribution of the remaining time-to-event at time j is completely characterized by the data at time steps preceding j. This assumption is needed to enable the inference of event risks dynamically in a streaming fashion.

The log-likelihood function is modified to allow streaming data and is represented as a combination of the probability of the individual experiencing the event or being censored. The loss is then factorized over each time-step, leveraging the statefulness assumption, and rewritten as a sum over each time-step. The probability of the event or censoring is a function of the input representation, which is computed using a softmax function. The parameters of the RDSM are learned by maximizing the log-likelihood function using gradient descent.

The performance of the RDSM was compared to the standard Deep Survival Machines, the DeepSurv model and the DeepHit model on the MIMIC III dataset.[3]. The performance of the models is evaluated using two metrics: Area under the Receiver Operating Characteristic curve (AuROC) and Brier score. The authors performed a grid search to optimize the performance of the RDSM model, using various hyperparameters such as the type of RNN cell, batch size, number of hidden layers, the dimensionality of the hidden layer, etc. All models were trained using the Adam optimizer, and early stopping was employed to evaluate the likelihood on a 10% subset of the training set.

The experimental results show that incorporating recurrent neural networks enhances the performance of Deep Survival Machines, leading to competitive results compared to other deep learning-based survival models. Additionally, this approach does not require the discretization of event times, enabling making predictions efficiently and quickly, and better-calibrated risk estimates.

Perspective for Safran How these models could be applied to the Safran use cases?

At Safran, survival analysis is a powerful tool currently used in aircraft engineering, particularly for assessing the reliability of critical components such as aircraft engines. By monitoring the time-to-failure of multiple engines and analysing relevant data, survival analysis can provide insights into the probability of engine failure over time, as well as identify factors that influence the likelihood of failure. This information can be used to identify maintenance needs, pinpoint areas of concern that require additional attention, and ultimately improve the safety and performance of aircraft engines.

[3] This is a widely used dataset in healthcare. It includes data from more than 40,000 patients admitted to intensive care at Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001 and 2012. The data include the amount of time a patient spends in an intensive care unit, the mortality as well as clinical and physiological measures, electronic medical records, medications, procedures, and demographics.

Table 1: Models comparison

Aspect	**DeepHit**	**DSM**	**RDSM**
Model type	Non-parametric	Fully-parametric survival regression model	Fully-parametric survival regression model with time-varying covariates
Architecture	Deep neural network with separate output for each event type	Deep neural network with fully-connected layers	Recurrent neural network (e.g. LSTM or GRU)
Advantages	Flexible modelling of survival distribution	Robust to missing data, More interpretable	Captures dynamic changes in time-varying covariates, Robust to missing data
Limitations	May have limitations in modelling long-term event horizons and require numerous parameters for arity discrete output spaces	May not be well-suited for handling time-varying covariates or capturing dynamic changes in covariate values over time	May struggles with capturing long-range dependencies in modelling complex interactions between covariates

Although Weibull and Gaussian parametric models are currently used to achieve the goal of risk prevention, they rely on strong assumptions about the shape of the stochastic process, assuming invariance of risk across time and independence of component reliabilities. Deep survival models, on the other hand, have the potential to leverage the vast amount of data available and improve performance in risk prevention.

We suggest using DSM for analyzing the Safran case data because it is a parametric model and therefore easier to interpret than non-parametric models like DeepHit. Additionally, DSM is less tied to the data and can be more robust in the presence of outliers or missing data. Compared to RDSM, DSM is less demanding in terms of computing time, making it a more practical choice for this analysis. However, it is important to ensure that the underlying process assumed by DSM is a good match for the Safran case data, so careful evaluation of the model fit is necessary.

DeepHit could be useful in scenarios involving competing risks, as it enables the capture of potential dependencies across components and provides flexibility to adjust the importance of each component based on its level of criticality. However, it assumes a discrete time horizon, which may result in lower time efficiency when the maximum potential survival time varies significantly across components. In such cases, it may be necessary to increase the number of time periods considered, which in turn increases the number of parameters that need to be estimated. In addition, the non-parametric nature of DeepHit means that it is not inherently self-interpretable, which is a necessary quality in practical applications. While the learned weights through the loss function, α_k, can be used as a measure of relative importance for each component, this alone may not be sufficient. One possible approach to interpretability is to use methods based on feature permutation, which involves systematically shuffling the values of a feature and measuring its impact on the model's performance to gain insights into its importance. But, the latter methods can be computationally expensive and may not be applicable to large datasets.

References

[Alaa and van der Schaar, 2017] Alaa, A. M. and van der Schaar, M. (2017). Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12:1031–1046.

[Cho et al., 2014] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

[Cox, 1972] Cox, D. R. (1972). Regression models and life-tables. *Journal of the royal statistical society series b-methodological*, 34:187–220.

[Faraggi and Simon, 1995] Faraggi, D. and Simon, R. M. (1995). A neural network model for survival data. *Statistics in medicine*, 14 1:73–82.

[Fernandez et al., 2016] Fernandez, T., Rivera, N., and Teh, Y. W. (2016). Gaussian processes for survival analysis. In *NIPS*.

[Fine and Gray, 1999] Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94:496–509.

[Ishwaran et al., 2008] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Wiley StatsRef: Statistics Reference Online*.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

[Katzman et al., 2018] Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18.

[Lee et al., 2018] Lee, C., Zame, W., Yoon, J., and Van Der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

[Lee et al., 2010] Lee, M.-L. T., Whitmore, G. A., and Rosner, B. A. (2010). Threshold regression for survival data with time-varying covariates. *Statistics in Medicine*, 29.

[Luck et al., 2017] Luck, M., Sylvain, T., Cardinal, H., Lodi, A., and Bengio, Y. (2017). Deep learning for patient-specific kidney graft survival analysis. *ArXiv*, abs/1705.10245.

[Nagpal et al., 2021a] Nagpal, C., Jeanselme, V., and Dubrawski, A. (2021a). Deep parametric time-to-event regression with time-varying covariates. In *Survival Prediction-Algorithms, Challenges and Applications*, pages 184–193. PMLR.

[Nagpal et al., 2021b] Nagpal, C., Li, X., and Dubrawski, A. (2021b). Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175.

[Ranganath et al., 2016] Ranganath, R., Perotte, A. J., Elhadad, N., and Blei, D. M. (2016). Deep survival analysis. *ArXiv*, abs/1608.02158.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[Yu et al., 2011] Yu, C.-N., Greiner, R., Lin, H.-C., and Baracos, V. E. (2011). Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *NIPS*.