Deep Survival Experts:
A Fully Parametric Survival Regression Model

Xinyu Li[*]
Auton Lab
Carnegie Mellon University
Pittsburgh, PA 15213
xinyul2@andrew.cmu.edu

Chirag Nagpal[*]
Auton Lab
Carnegie Mellon University
Pittsburgh, PA 15213
chiragn@cs.cmu.edu

Artur Dubrawski
Auton Lab
Carnegie Mellon University
Pittsburgh, PA 15213
awd@cs.cmu.edu

Abstract

Survival Analysis is extensively used in the medical domain, typically involving the estimation of patients' survival times after certain treatment or before death. In this paper we propose *Deep Survival Experts*, a Bayesian hierarchical model to estimate time-to-event in case of right-censored data. By leveraging deep neural networks, we are able to model the non-linear interactions in covariates and estimate survival time in a fully parametric manner. We do not require to make the common strong assumption of constant baseline hazard of underlying survival distribution as in the Cox proportional hazard model (CPH), which removes the need of using non-parametric approaches such as the Breslow's estimator to estimate the survival time. We demonstrate the superiority of our approach at estimating both relative risks and time-to-event through extensive experiments with datasets from breast cancer studies. **To the best of our knowledge, this is the first work involving fully parametric estimation of survival distributions in the presence of censoring**.

1 Introduction

Survival Analysis is a field of statistics and machine learning that focuses on estimating the risk of an event of interest taking place beyond a certain time in the future. In healthcare, the tasks typically involve estimating the distributions of patients' survival times before the onset of certain conditions or death using the patient covariates, such as demographics, medical history, or test results.

The Cox proportional hazards model (CPH) (Cox, 1972) is the most broadly used model for medical prognosis, and researchers have been successfully employing various techniques for CPH. (Rosen and Tanner, 1999) proposed using a mixture of linear experts for CPH. With the recent trend of deep and non-linear representation learning, (Xiang et al., 2000) evaluated different methods of combining neural networks with CPH to model right-censored survival data, and (Katzman et al., 2018) achieved state-of-the-art results with *DeepSurv* network. However, these approaches are still constrained by the strong assumption of CPH that the relative hazard ratio between any two individuals is constant over time, which may not hold in many practical scenarios. Besides, for medical applications which need the predictions of not only the risk but also the actual time-to-event, models subject to the CPH assumption normally have to require the non-parametric estimation of survival times using the Breslow's estimator (Breslow, 1972) based on the estimated relative risks.

Lee et al. (2018) proposed *DeepHit*, a deep learning approach capable of estimating the risk when multiple events could lead to failure. However, their method can only deal with survival times discretized into a finite set, and the sizes of both output space and parameters could become intractable with large amounts of complex training data. In this paper, we propose *Deep Survival Experts*, a Bayesian hierarchical model to directly estimate time-to-event, allowing full parameterization with

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

deep neural networks to capture non-linear interactions of patient covariates. To our knowledge, this is the first fully parametric estimation of survival distributions in the presence of censoring.

2 Approach

2.1 Survival Data

We assume that the survival data we consider is right-censored. This implies that the data, $\mathcal{D}$ is a set of tuples $\{(\mathbf{x}_i, t_i, \delta_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ are the covariates of a patient i, t_i is time-to-event, and δ_i is an indicator denoting whether t_i is actual event time or right-censoring time. For one individual, we observe either the actual event or censoring time, but not both. For simplicity, we assume that in the true data generating process, the censoring process is independent of the actual time-to-event.

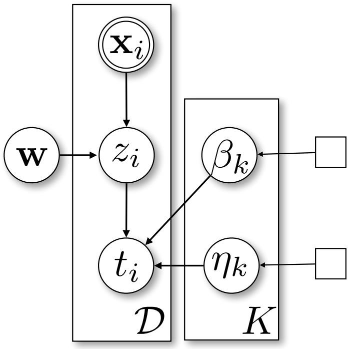
2.2 Deep Survival Experts

To accommodate patient heterogeneity arising in data, similar to (Rosen and Tanner, 1999), we propose to model the survival distribution of each patient as a fixed-size mixture of survival experts, the commonly used distribution primitives in parametric survival analysis. As shown in Figure 1, the covariates $\mathbf{x}$ are passed through a deep neural network (DNN) followed by a Softmax over K. The conditional distribution of the survival time T, $\mathbf{P}(T|X = \mathbf{x})$, is then described as a weighted mixture of K estimates by survival experts, with the outputs of the Softmax over K as the weights.

We choose Weibull distribution as the survival expert for our approach. This distribution is a widely used parametric model in survival analysis, with survival function $S(T) = \exp\left(-(\frac{T}{\beta})^{\eta}\right)$. At training time, the parameters of the deep neural network and the K Weibull distributions are learned jointly. At test time, the survival time of the held-out individual is predicted as a weighted mixture of the medians of all Weibull(β_k, η_k), for $k = 1...K$. As survival distributions with positive support tend to have long tails, we choose to use the median for survival time prediction instead of the expectation.

1. $\mathbf{x}_i \sim \mathcal{D}$
   We draw the covariates of a patient, $\mathbf{x}_i$.
2. $z_i \sim \text{Discrete}(\text{SOFTMAX}(\text{DNN}(\mathbf{x}_i, \mathbf{w})))$
   Conditioned on the covariates, $\mathbf{x}_i$ and the parameters, $\mathbf{w}$, we draw the latent z_i.
3. $\beta_k \sim \text{Log-Normal}(\beta, \sigma)$
   $\eta_k \sim \text{Log-Normal}(\eta, \sigma)$
   The set of parameters $\{\beta_k\}_{k=1}^{K}$ and $\{\eta_k\}_{k=1}^{K}$ are drawn from the prior β and η.
4. $t_i \sim \text{Weibull}(\beta_k, \eta_k)$
   Finally, the event time t_i is drawn conditioned on β_{z_i} and η_{z_i}.

a)



b)

Figure 1: The generative story and a plot in plate notation of the proposed model.

2.3 Loss Function

To handle both the uncensored data, $\mathcal{D}_U$ (patients with actual time-to-event available) and the censored data, $\mathcal{D}_C$, we train the *Deep Survival Experts* model by minimizing a total loss function $\mathcal{L}_{\text{total}} = \mathcal{L}_{D_U} + \mathcal{L}_{D_C} + \mathcal{L}_{\text{prior}}$. $\mathcal{L}_{\text{prior}}$ imposes the strength of the priors β and η on the β_k, η_k of each of the K experts. The priors are fit without conditioning on the covariates. To adjust for the positive bias brought by the long tails of survival distributions when predicting event times, we include another term, $\mathcal{L}_{\text{bias}}$, that explicitly penalizes median survival times of the uncensored data under the distribution imposed by the model to be close to the true survival time in L_2 norm. Using the maximum likelihood estimators ($\mathcal{LL}$), we have $\mathcal{L}_{D_U} = \alpha \cdot \mathcal{L}_{\text{bias}} + (1 - \alpha) \cdot (-\mathcal{LL}_{D_U})$, where α is chosen to trade off the likelihood and bias loss, and $\mathcal{L}_{D_C} = -\mathcal{LL}_{D_C}$.

The maximum likelihood estimator and its evidence lower bound (ELBO) for the uncensored data $\mathcal{D}_U$ can be written as

$$
\begin{aligned}
\mathcal{LL}_{D_U} &= \ln \mathbf{P}(\mathcal{D}_U|\mathbf{\Theta}) = \ln \left(\prod_{i=1}^{|D_U|} \mathbf{P}(T = t_i | X = \mathbf{x}_i, \mathbf{\Theta}) \right) \\
&= \sum_{i=1}^{|D_U|} \ln \left(\sum_{k=1}^{K} \mathbf{P}(T = t_i | Z, \beta_k, \eta_k) \mathbf{P}(Z | X = \mathbf{x}_i, \mathbf{w}) \right) \\
&= \sum_{i=1}^{|D_U|} \ln \left(\mathop{\mathbb{E}}_{Z \sim (\cdot|\mathbf{x}_i, \mathbf{w})} [\mathbf{P}(T = t_i | Z, \beta_k, \eta_k)] \right) \\
&\text{(Applying Jensen's Inequality)} \\
&\geq \sum_{i=1}^{|D_U|} \left(\mathop{\mathbb{E}}_{Z \sim (\cdot|\mathbf{x}_i, \mathbf{w})} [\ln \mathbf{P}(T = t_i | Z, \beta_k, \eta_k)] \right) \triangleq \mathbf{ELBO}_U(\Theta)
\end{aligned}
$$

Proceeding as above, for the censored data $\mathcal{D}_C$ as

$$
\begin{aligned}
\mathcal{LL}_{D_C} = \ln \mathbf{P}(\mathcal{D}_C|\Theta) &= \ln \left(\prod_{i=1}^{|D_C|} \mathbf{P}(T > t_i | X = \mathbf{x}_i, \Theta) \right) \\
&\geq \sum_{i=1}^{|D_C|} \left(\mathop{\mathbb{E}}_{Z \sim (\cdot|\mathbf{x}_i, w)} [\ln \mathbf{P}(T > t_i | Z, \beta_k, \eta_k)] \right) \triangleq \mathbf{ELBO}_C(\Theta)
\end{aligned}
$$

The strength of the priors on the β_k, η_k is included in L_2 norm as

$$
\begin{aligned}
\mathcal{L}_{\text{prior}} &= \ln \left(\prod_{k=1}^{K} \mathbf{P}(\beta_k, \eta_k | \beta, \eta) \right) = \sum_{k=1}^{K} \ln \mathbf{P}(\beta_k|\beta) + \ln \mathbf{P}(\eta_k|\eta) \\
&= \lambda \sum_{k=1}^{K} ||\beta_k - \beta||_{\mathbf{2}}^2 + ||\eta_k - \eta||_{\mathbf{2}}^2
\end{aligned}
$$

And the $\mathcal{L}_{\text{bias}}$ can be written as below, for Weibull distribution, the median $\hat{\mathbf{t}}[k] = \beta_k (\ln 2)^{\frac{1}{\eta_k}}$

$$
\begin{aligned}
\mathcal{L}_{\text{bias}} = \sum_{i=1}^{|\mathcal{D}_U|} || \mathop{\mathbb{E}}_{Z \sim (\cdot|\mathbf{x}_i, w)} [\text{Median}(T | X = \mathbf{x}_i, \mathbf{\Theta}, Z)] - t_i||_{\mathbf{2}}^2 = \sum_{i=1}^{|\mathcal{D}_U|} ||1^\top (\hat{\mathbf{t}} \odot \tilde{\mathbf{o}}) - t_i||_{\mathbf{2}}^2, \\
\text{where } \tilde{\mathbf{o}} = \textsc{softmax}(\text{DNN}(\mathbf{x}_i, \mathbf{w})) \text{ and } \hat{\mathbf{t}} \in \mathbb{R}^k.
\end{aligned}
$$

3 Experiments and Results

We evaluate our approach by both assessing the ordering of pairwise relative risks using Concordance-Index (C-Index) (Harrell, 1982), and by measuring the actual time-to-event detection using Root Mean Square Error (RMSE) around the predicted event times for the uncensored observations. We compare our performance against CPH, and two other state-of-the-art non-linear survival models - *DeepSurv* (Katzman et al., 2018) and Random Survival Forests (RSF) (Ishwaran et al., 2008).

We evaluated the performances on two real-world medical datasets: Rotterdam & German Breast Cancer Study Group (GBSG) (Schumacher et al., 1994) and The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012). Both datasets were downloaded from (Katzman et al., 2018), [1] and have already been partitioned for training and testing. GBSG contains records of patients with primary node positive breast cancer taken from Rotterdam tumour bank and clinical trials to study the effects of chemotherapy and hormonal treatment on survival rate (Royston and Altman, 2013). METABRIC contains the gene expressions and clinical features of patients to determine breast cancer subgroups and facilitate treatment improvement. The sizes and censoring ratios of these datasets are described in Appendix A.1.

We reserved 10% of the training set as validation set (the ratio of censoring was maintained). We tuned the hyper-parameters and selected the best model that achieved the lowest RMSE among the models that achieved the top 5 C-Index, both on the validation set. We bootstrapped the test set to obtain the confidence intervals as (Katzman et al., 2018). The hyper-parameters we used for each dataset and the implementation details of the baseline models are shown in Appendix A.2 and A.3.

The performances of *Deep Survival Experts* and the baseline models are shown in Table 1 and Table 2. Our approach achieved the best performance both in C-Index and RMSE for the GBSG dataset, and is only inferior to *DeepSurv* (not statistically significantly) for METABRIC dataset in Concordance-Index, yet still outperforms all three baseline models in RMSE. The results demonstrate the superiority of our approach in estimating time-to-event compared with other models which have to rely on the Breslow's estimator for non-parametric estimations of event times, while remaining very competitive in estimating the relative risks.

Table 1: Concordance-Index of all models (95% Confidence Interval)

Model	Dataset			
	CPH	DeepSurv	RSF	Deep Survival Experts
GBSG	0.658 (0.654, 0.661)	0.668 (0.665, 0.671)	0.651 (0.648, 0.654)	**0.685** **(0.682, 0.689)**
METABRIC	0.631 (0.627, 0.635)	0.643 (0.639, 0.647)	0.624 (0.620, 0.629)	0.638 (0.634, 0.642)

Table 2: RMSE of all models (95% Confidence Interval)

Model	Dataset			
	CPH	DeepSurv	RSF	Deep Survival Experts
GBSG	26.293 (26.168, 26.418)	25.504 (25.391, 25.617)	25.918 (25.786, 26.049)	**18.791** **(18.68, 18.902)**
METABRIC	92.184 (91.446, 92.923)	89.816 (89.216, 90.417)	94.586 (93.858, 95.314)	**73.826** **(73.119, 74.534)**

4 Conclusion

In this paper, we proposed *Deep Survival Experts*, a novel approach of estimating survival time in a fully parametric manner, by using the mixture of survival experts and deep neural networks. We have demonstrated on real-world medical data that our model can outperform state-of-the-art baselines at estimating time-to-event and can achieve competitive results in estimating the relative risks. The resulting models are readily interpretable due to their parametric nature. Our method is universally applicable to a wide range of survival modeling tasks in healthcare applications. Our future work includes incorporating other types of survival experts, such as log-logistic or log-normal distributions into our current framework, and testing the method on more and larger real-world survival datasets.

References

Norman E Breslow. Discussion of the paper by D.R. Cox. *J R Statist Soc B*, 34:216–217, 1972.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[1] https://github.com/jaredleekatzman/DeepSurv

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, and Yinyin et al. Yuan. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012. doi: 10.1038/nature10983.

Frank E. Harrell. Evaluating the yield of medical tests. *JAMA: The Journal of the American Medical Association*, 247(18):2543, 1982. doi: 10.1001/jama.1982.03320430047030.

H. Ishwaran and U.B. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2019. URL https://cran.r-project.org/package=randomForestSRC. R package version 2.9.1.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Ori Rosen and Martin Tanner. Mixtures of proportional hazards regression models. *Statistics in Medicine*, 18(9): 1119–1131, 1999.

Patrick Royston and Douglas G Altman. External validation of a cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13(1), 2013. doi: 10.1186/1471-2288-13-33.

M Schumacher, G Bastert, H Bojar, K Hübner, M Olschewski, W Sauerbrei, C Schmoor, C Beyerle, R L Neumann, and H F Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994. doi: 10.1200/jco.1994.12.10.2086.

Anny Xiang, Pablo Lapuerta, Alex Ryutov, Jonathan Buckley, and Stanley Azen. Comparison of the performance of neural network methods and cox regression for censored survival data. *Computational statistics & data analysis*, 34(2):243–257, 2000.

A Appendix

A.1 Datasets

The number of patients, number of covariates, and ratio of censoring (proportion of patients whose follow-up were lost before death) of each dataset is show in Table 3.

Table 3: Detailed information of each dataset.

	Dataset			
	GBSG - train	GBSG - test	METABRIC - train	METABRIC - test
Number of patients	$1,546$	686	$1,523$	381
Number of covariates	7	7	9	9
Number of events	$968(62.61\%)$	$299(43.59\%)$	$887(58.24\%)$	$216(56.69\%)$

A.2 Hyper-parameter Tuning for *Deep Survival Experts*

Deep Survival Experts was implemented in PyTorch environment. The hyper-parameters include: number of survival experts (N), deep neural network (DNN) structure, learning rate (LR), λ for $\mathcal{L}_{\text{prior}}$, α for $\mathcal{L}_{\text{bias}}$, and whether using the log-likelihood or its ELBO in loss function. The neural network structure is described in the format [number of input covariates, number of nodes in layer $1...n$, number of survival experts], if there is only [number of input covariates, number of survival experts], it means that a linear network is chosen by the performance on validation set. The numbers of input covariates for GBSG and METABRIC are 7 and 9. We used ReLu activation function for the hidden layers. The network was trained by back-propagation with Adam optimizer. The hyper-parameters we used for each dataset is shown in Table 4.

Table 4: Hyper-parameters of *Deep Survival Experts* for each dataset.

HYPER-PARAMETERS	DATASET	
	GBSG	METABRIC
N	3	3
DNN structure	$[7, 7, 3]$	$[9, 3]$
LR	0.001	0.001
λ	0.01	0.01
α	0.001	0.0001
ELBO	Yes	No

A.3 Implementation of Baseline Models

We reported the C-Index with reference to the results reported in (Katzman et al., 2018). We then tried to reproduce their experiments to estimate the survival times. We ran CPH and C-index statistics using the Lifelines Python library. We ran *DeepSurv* experiments using code and instructions provided in `https://github.com/jaredleekatzman/DeepSurv`. We ran RSF experiments with the R package randomForestSRC (Ishwaran and Kogalur, 2019). The survival curves of CPH and *DeepSurv* were then estimated using the Breslow's estimator (RSF has an estimation of the survival curves), and then the expectations of survival times were used as the predicted event times. We did not use the median of survival times as the empirical survival curves do not guarantee to cross 0.5.