Deep Survival: The Deep Cox Proportional Hazards Network

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, Yuval Kluger
Yale University

Abstract

Different neural network architectures have been effective in modeling survival datasets in which patients’ death times are unknown, i.e. right-censored. However, these neural networks have rarely been shown to outperform their linear counterparts, such as the Cox proportional hazards model. In this paper, we run simulated experiments and use real survival data to build upon the risk-regression architecture proposed by Faraggi-Simon. We demonstrate that a Deep Cox proportional hazards model not only works as well, but also outperforms the standard linear Cox proportional hazards model in predictive ability on survival datasets with non-linear proportional risks. We then show that the neural network can also function as a recommender system by including a categorical variable representing a treatment group. The results of our study suggest, that with modern techniques, neural networks can successfully model Cox regression and, furthermore, explain the interactions of a patient’s factors with their calculated risk.

1 Introduction

Medical researchers rely on survival models to evaluate the significance of prognostic variables in outcomes such as death or cancer recurrence and to inform treatment options [1, 2, 3, 4]. One standard survival model is the Cox proportional hazards model (CPH), in which a risk function associated with the event time is estimated [5]. Although several neural network survival models have been proposed, Faraggi and Simon [6] were the first to apply a neural network architecture directly to the CPH model. However, to the best of our knowledge, no research has shown that the Faraggi-Simon method outperforms the standard linear CPH [7, 8, 9]. This in conjunction with neural network’s (NN) lack of interpretability has led to the slow adoption of the Faraggi-Simon method in survival risk analysis.

In this paper, we apply the Faraggi-Simon approach to a deep architecture. We use modern deep learning techniques to improve performance and provide methods to better interpret results in medical applications. We perform experiments on both simulated data and on real survival data from the Worcester Heart Attack Survey (WHAS) [10]. We then view the network as a recommender system by including an additional categorical variable, representing a patient’s treatment group. The network allows researchers to model and interpret the interactions of a patient’s baseline data with treatment, and thus be able to provide clinical decision support. An open source Python module will soon be made publicly available to increase adoption of these methods and further research and application of deep learning in survival analysis.

2 Deep Linear Cox Proportional Hazards Model (Deep CPH)

Survival data is comprised of three elements: baseline data, an event time, and an event indicator. If the event (i.e. death) is observed, then the time corresponds to the time of the event since collecting baseline data, and the event indicator is 1. If the event is not observed, then the time corresponds to the last time the individual was observed (e.g. end of study) since collecting the baseline data, and the event indicator is 0; in this case, the patient is said to be right-censored. Right-censoring distinguishes survival data and demands a specialized model, or else it must be discarded in order to use traditional methods.

Linear Cox regression is a specialized model for survival data that relies on the assumption of proportional hazards. The proportional hazard condition states that an individual’s risk \( h(x_i) \) is directly proportional to their covariates \( (x_{i0}, ..., x_{id}) \), and thus Cox regression models \( h(x_i) \) as the linear combination of \( x_i \).
However, this poses limitations because the data may require higher-level interaction terms between covariates and more complex feature engineering.

The Faraggi-Simon method maximizes a penalized Cox likelihood \[5\] as the neural network’s loss function. This provides the basis for a non-linear proportional hazards model. While this should allow for the modeling of more complex interactions and an increase in predictive ability, most reported results show that the Faraggi-Simon method does not outperform the linear CPH model. One reason could be that the field of neural networks was not as developed at the time of the their study.

Hence, we introduce the Deep Cox proportional hazards (DCPH) model. We leverage recent deep learning techniques, unavailable to Faraggi and Simon, and improve upon the their network. We use a deep architecture and apply modern techniques such as: \(\ell_2\) regularization, Rectified Linear Units (ReLU) activations \[11\], Batch Normalization \[12\], dropout \[13\], stochastic gradient descent with Nesterov momentum \[14\] and Power-schedule learning rate decay \[15\].

**Objective Function**

The DCPH network’s loss function is the negative log partial likelihood:

\[
l(\theta) = \sum_{i : E_i = 1} \left( \hat{h}(x_i, \theta) - \log \sum_{j \in \mathbb{R}(t_i)} e^{\hat{h}(x_j, \theta)} \right) - \sum_{i : E_i = 1} \log \sum_{j \in \mathbb{R}(t_i)} e^{\hat{h}(x_j, \theta)}
\]

The partial likelihood measures the joint probability of an observed event at a given time, conditioned on the set of all possible events occurring at that time. Where \(x_i\) is the \(i\)th observation, \(T_i\) and \(E_i\) are the event time and event indicator respectively. The output of the network, \(\hat{h}(x_i, \theta)\), is an estimation of the risk of an patient \(x_i\) parameterized by network parameters \(\theta\). The risk set \(\mathbb{R}(t) = \{ i : T_i \geq t \}\) is the set of patients still at risk of death at time \(t\).

**Evaluation**

To evaluate the predictive accuracy of the DCPH and compare it to the linear Cox regression, we measure the concordance-index (C-index) \(c\) as outlined by Harrell et. al \[16\]. For context, a \(c = 0.5\) represents a the average C-index of a random model, while a \(c = 1\) is a perfect ranking of event times. We perform bootstrapping to obtain confidence intervals for each model’s C-index \[17\].

**Model Selection**

The hyper-parameters of the network include: \(\ell_2\) regularization coefficient, learning rate, learning rate decay constant, dropout rate, momentum, and the size and depth of the NN. These were chosen by performing the Random hyper-parameter optimization search proposed by Bergstra and Bengio \[18\] and evaluating models’ C-index on a cross-validation set.

### 3 Simulated Experiments

In each experiment, we generate a set of \(N = 10000\) observations, such that each observation represents a patient vector with \(d = 10\) covariates drawn from a uniform distribution on \([-1, 1]\). We generate an individual’s death time \(T_i\) according to an exponential Cox model \[19\]:

\[
T_i \sim \text{Exp}(\lambda_i(t; x_i)) = \text{Exp}(\lambda_0 \cdot e^{h(x_i)})
\]

In both experiments, the risk function \(h(x_i)\) only depends on two of the ten covariates. We then choose a censoring time to represent the ‘end of study’, such that an average of 30% of the patients have an observed event in the dataset.

**Linear Risk Function**

We first simulate patients to have a linear risk function \(h(x) = x_0 + 2x_1\) for \(x \in \mathbb{R}^d\), so that the proportional hazard condition holds true. Both the linear and deep CPH models should be able to calculate the correct coefficients; thus, we expect them to have equivalent performance.

The linear CPH model is able to achieve a C-index of 0.7954 (95% CI: 0.7942 - 0.7984) on the test set, in addition to estimating the correct coefficients of \(x_0\) and \(x_1\) within \(\pm 0.1\) accuracy. The DCPH network has a C-index of 0.7953 (95% CI: 0.7931 - 0.7974) on the test set and is also able to reconstruct the surface of the risk function \(h(x_i)\) within the 0.01 level. With a comparable C-index and successful coefficient estimation, the DCPH network performs as well as a standard linear Cox regression.
Nonlinear Risk Function

We set the risk function to be a Gaussian on the first two covariates with $\lambda_{max} = 5.0$ and a scale factor of $r = 0.5$:

$$h(x) = -\log(\lambda_{max}) \exp\left(-\frac{x_0^2 + x_1^2}{2r^2}\right)$$

The surface of the risk function is depicted in (a). Because this risk function does not satisfy the proportional hazard condition, we do not expect CPH regression to model the risk function properly without adding interaction terms between the covariates. We expect the DCPH network to be more successful in reconstructing the Gaussian risk function and accurately predict patients’ risks.

Shown in Figure (b), the linear CPH regression fails to determine the first two covariates as significant and results in a test C-index of 0.4983 (95% CI: 0.4948 - 0.5017). In this case, when the patients’ risk is non-linearly dependent on their covariates, the CPH model performs as well as randomly predicting death times. The DCPH network, however, has a higher predictive accuracy of 0.6049 (95% CI: 0.6016 - 0.6082). Figure (c) shows that the DCPH network reconstructs the Gaussian relationship between the first two covariates and a patient’s risk. Not only does the DCPH outperform the linear CPH in predictive ability, but is also able to accurately compute the relationships between a patient’s covariates and their risk.

4 Worcester Heart Attack Study (WHAS)

The goal of the Worcester Heart Attack Study (WHAS) was to study the effects of patient’s factors on acute myocardial infarction (MI) survival; subsets of their data are common case studies in survival analysis textbooks [20]. The dataset we use consists of 1,638 observations and 5 features: age, body-mass-index, left heart failure complications, and order of MI (first or recurrent MI). A total of 42.12% of patients died during the survey with a median death rate of 516.0 days.

When we run a linear CPH on the dataset we achieve a test C-index of 0.7599 (95% CI: 0.7576 - 0.7622) and 3 of the 5 factors have strong predictive significance (p-value $< 10^{-6}$). Therefore, the linear Cox regression is fairly successful in modeling the relationships between covariates and survival rate. However, the DCPH outperforms the CPH and achieves a C-index of 0.7889 (95% CI: 0.7868 - 0.7909) on the test set. The DCPH is able to improve the predictive ability of the model and can be subjected to further research to explore the interactions between patient’s covariates and MI survival rate.

5 Recommender System

One of the key characteristics of NNs is their lack of interpretability. This poses a problem in medical applications, as determining the significance of patient’s factors on their risk is common practice in survival analysis [9, 3, 21]. We show that by including an additional categorical variable, i.e. treatment group, to the simulated data from Section 3 we can view the DCPH network as a recommender system by measuring the predicted risk of a patient being prescribed each treatment. Therefore, we can recommend the least risky treatment and provide interpretation and application of the DCPH’s results.
We randomly assign a treatment group \( \tau_i \) (0 or 1) to the simulated patients in the dataset. All of the patients in group \( \tau = 0 \) were ‘unaffected’ (e.g., given a placebo) and have a constant risk function regardless of their covariates. The other group, \( \tau = 1 \), is prescribed a treatment with nonlinear effects as in Equation 3. Collectively, the hazard function becomes:

\[
\lambda_i(t; x_i) = \begin{cases} 
\lambda_0(t) & \text{if treatment group } \tau_i = 0 \\
\lambda_0(t) \cdot e^{h(x_i)} & \text{if treatment group } \tau_i = 1
\end{cases}
\]

(4)

Our goal is to evaluate whether the model can generalize how the treatment is affected by an individual’s covariates. For any patient, the network should be able to accurately predict the risk of being prescribed a given treatment. If we pass a patient through the network, once in treatment group 0 and again in treatment group 1, we can take the log of their hazards ratio to calculate the personal risk of the prescribed treatment:

\[
\log \left( \frac{\lambda_i(t; x_i | \tau_i = 1)}{\lambda_i(t; x_i | \tau_i = 0)} \right) = \log \left( \frac{\lambda_0(t) \cdot e^{h(x_i)}}{\lambda_0(t)} \right) = h(x)
\]

(5)

Thus, this is equivalent to calculating the objective risk function \( h(x) \). In practice, if a patient \( i \) has a positive risk \( h(x) \), they will be harmed by a treatment (greater risk of death); similarly, a negative risk corresponds to being helped by the treatment. If the DCPH is successful in modeling these interactions, we can be confident in the calculations of a treatment’s predicted risk and use it to recommend treatment options.

Figure 2 shows how successful the network is in predicting the risk function \( h(x) \), plotted in Figure 2(a), for patients in the test set. As expected, Figure 2(b) shows that the network models a constant risk for a patient in treatment 0, independent of a patient’s covariates. Meanwhile, Figure 2(c) shows how the network models the Gaussian effects of a patient’s covariates on their treatment risk. With the network accurately reconstructing the risk surfaces, we can be confident in our calculations of the true risk \( h(x) \) and provide treatment recommendations for new patients.

6 Conclusion

We built upon Faraggi and Simon’s risk-regression neural network and showed how applying modern machine learning techniques results in improved performance and more generalized predictive ability. The DCPH network is a nonlinear proportional hazards model and can be applied to larger-feature datasets with highly non-linear and complex interactions between covariates. Switching to a DCPH allows the network to discover nonlinear features instead of relying on researchers to manually experiment with higher-level interactions of factors.

Additionally, we show how to view the DCPH network as recommender system by including categorical variables representing an individual’s treatment group. The DCPH network is able to model multiple risk distributions; by comparing a network’s predicted risk for each treatment, we can determine which course of action will help or harm a patient.

A full length manuscript describing this work is in final preparation, and will be uploaded to arXiv within a few weeks. Additionally, an open source Python package using Theano will be made publicly available.
References

[1] Yeh RW, Secemsky EA, Kereiakes DJ, and et al. Development and validation of a prediction rule for benefit and harm of dual antiplatelet therapy beyond 1 year after percutaneous coronary intervention. JAMA, 315(16):1735–1749, 2016.

[2] Patrick Royston and Douglas G Altman. External validation of a cox prognostic model: principles and methods. BMC medical research methodology, 13(1):1, 2013.

[3] Eric Bair and Robert Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol, 2(4):e108, 2004.

[4] Wei-Yi Cheng, Tai-Hsien Ou Yang, and Dimitris Anastassiou. Development of a prognostic model for breast cancer survival in an open challenge environment. Science translational medicine, 5(181):181ra50–181ra50, 2013.

[5] David R Cox. Regression models and life-tables. In Breakthroughs in statistics, pages 527–541. Springer, 1992.

[6] David Faraggi and Richard Simon. A neural network model for survival data. Statistics in medicine, 14(1):73–82, 1995.

[7] Daniel J Sargent. Comparison of artificial neural networks with other statistical approaches. Cancer, 91(8):1636–1642, 2001.

[8] 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.

[9] L Mariani, D Coradini, E Biganzoli, P Boracchi, E Marubini, S Pilotti, B Salvadori, R Silvestrini, U Veronesi, R Zucali, et al. Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear cox regression model and its artificial neural network extension. Breast cancer research and treatment, 44(2):167–178, 1997.

[10] Saczynski JS, Spencer FA, Gore JM, and et al. Twenty-year trends in the incidence of stroke complicating acute myocardial infarction: Worcester heart attack study. Archives of Internal Medicine, 168(19):2104–2110, 2008.

[11] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 807–814, 2010.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014.

[14] Yuri Nesterov et al. Gradient methods for minimizing composite objective function. Technical report, UCL, 2007.

[15] Alan Senior, Georg Heigold, Marc’ Aurelio Ranzato, and Ke Yang. An empirical study of learning rates in deep neural networks for speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 6724–6728. IEEE, 2013.

[16] Frank E Harrell, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. Statistics in medicine, 3(2):143–152, 1984.

[17] Bradley Efron and Robert J Tibshirani. An introduction to the bootstrap. CRC press, 1994.

[18] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. The Journal of Machine Learning Research, 13(1):281–305, 2012.

[19] Peter C Austin. Generating survival times to simulate cox proportional hazards models with time-varying covariates. Statistics in medicine, 31(29):3946–3958, 2012.

[20] David W. Hosmer Jr., Stanley Lemeshow, and Susanne May. Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley-Interscience, 2008.

[21] Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. Nature, 474(7353):609–615, 2011.