Neural Frailty Machine: Beyond proportional hazard assumption in neural survival regressions

Ruofan Wu*, Jiawei Qiao*, Mingzhe Wu, Wen Yu, Ming Zheng, Tengfei Liu, Tianyi Zhang, Weiqiang Wang

Ant Group, Fudan University and Coupang

Introduction

- We study regression analysis with time-to-event data under **right censoring**.
- Event time $\tilde{T} \geq 0$ with survival function $S(t)$, density $f(t)$ and hazard function $\lambda(t)$.
- Feature vector Z.
- Censoring time C which satisfies $\tilde{T} \perp\!\!\!\perp C|Z$, resulting in observed tuple $T = \tilde{T} \wedge C, \delta = I(\tilde{T} \leq C)$.

CoxPH model [Cox72]

In its linear form, the widely used CoxPH model assumes the proportional hazard (PH) assumption

$$\lambda(t|Z) = \lambda_0(t)e^{\langle Z,\theta \rangle} \tag{1}$$

Extending CoxPH

- Due to limited expressivity, there have been attempts that generalizes CoxPH using more elegant function approximators like neural networks [FS95, KSC+18].
- Nonlinear CoxPH models still assumes proportional hazard assumption which has limitations in modeling specific phenomenons like crossing hazards [Ben83].
- A more general solution: Using the idea of **frailty**.

Frailty model [Hou86]

In its linear form, (multiplicative) frailty model introduces an additional unobserved heterogeneity ω into the hazard/intensity formulation:

$$\lambda(t|Z, \omega) = \omega \lambda_0(t) e^{\langle Z, \theta \rangle} \tag{2}$$

Neural Frailty Machine

To further enhance the capability of frailty model, we equip neural function approximations. The resulting modeling framework is called Neural Frailty Machine (NFM) which we propose to approaches.

Proportional frailty scheme (PF)

The PF scheme directly replaces linear term in ordinary frailty model with a neural network $m(Z)$ that depends only on the features.

$$\lambda(t|Z,\omega) = \omega e^{h(t)+m(Z)} \tag{3}$$

Fully neural scheme (FN)

The FN scheme further relaxes the separation between baseline hazard and feature dependence, using a neural network $h(t, Z)$ for approximation.

$$\lambda(t|Z,\omega) = \omega e^{\nu(t,Z)} \tag{4}$$

Parameter learning

We use observed log-likelihood as the learning objective.
PF-Scheme We use two MLPs $\widehat{h} = \widehat{h}(t; \mathbf{W}^h, \mathbf{b}^h)$ and $\widehat{m} = \widehat{m}(Z; \mathbf{W}^m, \mathbf{b}^m)$ as function approximators to h and m.

$$\mathcal{L}(\mathbf{W}^h, \mathbf{b}^h, \mathbf{W}^m, \mathbf{b}^m, \theta)$$
$$= \frac{1}{n}\left[\sum_{i\in[n]} \delta_i \log g_\theta \left(e^{\widehat{m}(Z_i)} \int_0^{T_i} e^{\widehat{h}(s)} ds\right) + \delta_i \widehat{h}(T_i) + \delta_i \widehat{m}(Z_i) - G_\theta \left(e^{\widehat{m}(Z_i)} \int_0^{T_i} e^{\widehat{h}(s)} ds\right)\right].$$

FN-Scheme We use $\widehat{\nu} = \widehat{\nu}(t, Z; \mathbf{W}^\nu, \mathbf{b}^\nu)$ to approximate $\nu(t, Z)$

$$\mathcal{L}(\mathbf{W}^\nu, \mathbf{b}^\nu, \theta)$$
$$= \frac{1}{n}\left[\sum_{i\in[n]} \delta_i \log g_\theta \left(\int_0^{T_i} e^{\widehat{\nu}(s, Z_i; \mathbf{W}^\nu, \mathbf{b}^\nu)} ds\right) + \delta_i \widehat{\nu}(T_i, Z_i; \mathbf{W}^\nu, \mathbf{b}^\nu) - G_\theta \left(\int_0^{T_i} e^{\widehat{\nu}(s, Z_i; \mathbf{W}^\nu, \mathbf{b}^\nu)} ds\right)\right].$$

Here G_θ is defined as the negative of the logarithm of the Laplace transform of the frailty distribution, with g_θ being its derivative w.r.t. t.

We study the rates of convergence for estimated parameters with the true parameters lying inside a Hölder ball with radius M and smoothness parameter β. Under the following two distance metrics

$$d_{\mathrm{PF}}\left(\widehat{\phi}_n, \phi_0\right) = \sqrt{\mathbb{E}_{z\sim\mathbb{P}_Z}\left[H^2(\mathbb{P}_{\widehat{\phi}_n, Z=z} \,\|\, \mathbb{P}_{\phi_0, Z=z})\right]}, d_{\mathrm{FN}}\left(\widehat{\psi}_n, \psi_0\right) = \sqrt{\mathbb{E}_{z\sim\mathbb{P}_Z}\left[H^2(\mathbb{P}_{\widehat{\psi}_n, Z=z} \,\|\, \mathbb{P}_{\psi_0, Z=z})\right]}$$

where $\widehat{\phi}_n$ and $\widehat{\psi}_n$ are bundled parameter updates and H denotes Hellinger distance, we have the following statistical guarantee:

Theorem

Under some regularity conditions, we have

$$d_{PF}\left(\widehat{\phi}_n, \phi_0\right) = \widetilde{O}_{\mathbb{P}}\left(n^{-\frac{\beta}{2\beta+2d}}\right), d_{FN}\left(\widehat{\psi}_n, \psi_0\right) = \widetilde{O}_{\mathbb{P}}\left(n^{-\frac{\beta}{2\beta+2d+2}}\right)$$

where d is the feature dimension and logarithmic factors are hided.

Empirical evaluations

Empirical evaluations over 4 relatively small scale datasets.

Model	METABRIC		RotGBSG		FLCHAIN		SUPPORT	
	IBS	INBLL	IBS	INBLL	IBS	INBLL	IBS	INBLL
CoxPH	16.46$_{\pm 0.90}$	49.57$_{\pm 2.66}$	18.25$_{\pm 0.44}$	53.76$_{\pm 1.11}$	10.05$_{\pm 0.38}$	33.18$_{\pm 1.16}$	20.54$_{\pm 0.38}$	59.58$_{\pm 0.86}$
GBM	16.61$_{\pm 0.82}$	49.87$_{\pm 2.44}$	17.83$_{\pm 0.44}$	52.78$_{\pm 1.11}$	9.98$_{\pm 0.37}$	32.88$_{\pm 1.05}$	19.18$_{\pm 0.39}$	56.46$_{\pm 0.10}$
RSF	16.62$_{\pm 0.64}$	49.61$_{\pm 1.54}$	17.89$_{\pm 0.42}$	52.77$_{\pm 1.01}$	**9.96**$_{\pm 0.37}$	32.92$_{\pm 1.05}$	19.11$_{\pm 0.40}$	56.28$_{\pm 1.00}$
DeepSurv	16.55$_{\pm 0.93}$	49.85$_{\pm 3.02}$	17.80$_{\pm 0.49}$	52.62$_{\pm 1.25}$	10.09$_{\pm 0.38}$	33.28$_{\pm 1.15}$	19.20$_{\pm 0.41}$	56.48$_{\pm 1.08}$
CoxTime	16.54$_{\pm 0.83}$	49.67$_{\pm 2.67}$	17.80$_{\pm 0.58}$	52.56$_{\pm 1.47}$	10.28$_{\pm 0.45}$	34.18$_{\pm 1.53}$	19.17$_{\pm 0.40}$	56.45$_{\pm 1.10}$
DeepHit	17.50$_{\pm 0.83}$	52.10$_{\pm 2.16}$	19.61$_{\pm 0.38}$	56.67$_{\pm 1.10}$	11.83$_{\pm 0.39}$	37.72$_{\pm 1.02}$	20.66$_{\pm 0.32}$	60.06$_{\pm 0.72}$
DeepEH	16.56$_{\pm 0.65}$	49.42$_{\pm 1.53}$	17.62$_{\pm 0.52}$	52.08$_{\pm 1.27}$	10.11$_{\pm 0.37}$	33.30$_{\pm 1.10}$	19.30$_{\pm 0.39}$	56.67$_{\pm 0.94}$
SuMo-net	16.49$_{\pm 0.83}$	49.74$_{\pm 2.21}$	17.77$_{\pm 0.47}$	52.62$_{\pm 1.11}$	10.07$_{\pm 0.40}$	33.20$_{\pm 1.10}$	19.40$_{\pm 0.38}$	56.87$_{\pm 0.96}$
SODEN	16.52$_{\pm 0.63}$	49.39$_{\pm 1.97}$	**17.05**$_{\pm 0.63}$	**50.45**$_{\pm 1.97}$	10.13$_{\pm 0.24}$	33.37$_{\pm 0.57}$	19.07$_{\pm 0.50}$	56.15$_{\pm 1.35}$
SurvNode	16.67$_{\pm 1.32}$	49.73$_{\pm 3.89}$	17.42$_{\pm 0.53}$	51.70$_{\pm 1.16}$	10.40$_{\pm 0.29}$	34.37$_{\pm 1.03}$	19.58$_{\pm 0.34}$	57.49$_{\pm 0.84}$
DCM	16.58$_{\pm 0.87}$	49.48$_{\pm 2.23}$	17.66$_{\pm 0.54}$	52.26$_{\pm 1.23}$	10.13$_{\pm 0.50}$	33.40$_{\pm 1.38}$	19.29$_{\pm 0.42}$	56.68$_{\pm 1.09}$
DeSurv	16.71$_{\pm 0.75}$	49.61$_{\pm 2.15}$	17.98$_{\pm 0.46}$	53.23$_{\pm 1.15}$	10.06$_{\pm 0.62}$	33.18$_{\pm 1.93}$	19.50$_{\pm 0.40}$	57.28$_{\pm 0.89}$
NFM-PF	16.33$_{\pm 0.75}$	49.07$_{\pm 1.96}$	17.60$_{\pm 0.55}$	52.12$_{\pm 1.34}$	**9.96**$_{\pm 0.39}$	**32.84**$_{\pm 1.15}$	19.14$_{\pm 0.39}$	56.35$_{\pm 1.00}$
NFM-FN	**16.11**$_{\pm 0.81}$	**48.21**$_{\pm 2.04}$	17.66$_{\pm 0.52}$	52.41$_{\pm 1.22}$	10.05$_{\pm 0.39}$	33.11$_{\pm 1.10}$	**18.97**$_{\pm 0.60}$	**55.87**$_{\pm 1.50}$

Empirical evaluations over 2 datasets with larger scale

Model	MIMIC-III		KKBOX	
	IBS	INBLL	IBS	INBLL
CoxPH	$20.40_{\pm 0.00}$	$60.02_{\pm 0.00}$	$12.60_{\pm 0.00}$	$39.40_{\pm 0.00}$
GBM	$17.70_{\pm 0.00}$	$52.30_{\pm 0.00}$	$11.81_{\pm 0.00}$	$38.15_{\pm 0.00}$
RSF	$17.79_{\pm 0.19}$	$53.34_{\pm 0.41}$	$14.46_{\pm 0.00}$	$44.39_{\pm 0.00}$
DeepSurv	$18.58_{\pm 0.92}$	$55.98_{\pm 2.43}$	$11.31_{\pm 0.05}$	$35.28_{\pm 0.15}$
CoxTime	$17.68_{\pm 1.36}$	$52.08_{\pm 3.06}$	$\underline{10.70}_{\pm 0.06}$	$\underline{33.10}_{\pm 0.21}$
DeepHit	$19.80_{\pm 1.31}$	$59.03_{\pm 4.20}$	$16.00_{\pm 0.34}$	$48.64_{\pm 1.04}$
SuMo-net	$18.62_{\pm 1.23}$	$54.51_{\pm 2.97}$	$11.58_{\pm 0.11}$	$36.61_{\pm 0.28}$
DCM	$18.02_{\pm 0.49}$	$52.83_{\pm 0.94}$	$10.71_{\pm 0.11}$	$33.24_{\pm 0.06}$
DeSurv	$18.19_{\pm 0.65}$	$54.69_{\pm 2.83}$	$10.77_{\pm 0.21}$	$33.22_{\pm 0.10}$
NFM-PF	$\mathbf{16.28}_{\pm 0.36}$	$\mathbf{49.18}_{\pm 0.92}$	$11.02_{\pm 0.11}$	$35.10_{\pm 0.22}$
NFM-FN	$\underline{17.47}_{\pm 0.45}$	$\underline{51.48}_{\pm 1.23}$	$\mathbf{10.63}_{\pm 0.08}$	$\mathbf{32.81}_{\pm 0.14}$

Conclusion and future works

- We have introduced NFM as a flexible and powerful neural modeling framework for survival analysis.
- NFM is shown to be both statistically correct in theory, and empirically effective in predictive tasks.
- Future directions: Establishing theory guarantees toward more realistic predictive metrics instead of nonparametric parameter estimation.

Steve Bennett.
Analysis of survival data by the proportional odds model.
Statistics in medicine, 2(2):273–277, 1983.

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

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

Philip Hougaard.
Survival models for heterogeneous populations derived from stable distributions.
Biometrika, 73(2):387–396, 1986.

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):1–12, 2018.