Neural Frailty Machine

Beyond proportional hazard assumption in neural survival regressions

NeurIPS 2023 Poster accepted

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

[Reviewer] Chanmoo Park
January 8, 2024

Neural survival regressions

- **Cox PH model** (Cox 1972)

$$\lambda(t \mid Z) = \lambda_0(t) \exp\left(\beta^\top Z\right)$$

- **1-layer NN** instead of $\beta^\top Z$ (Faraggi and Simon 1995)
  - No significant improvement observed due to the shallowness and computing limits.
- **Deep NN** ; DeepSurv (Katzman et al. 2018)
  - Remarkable results achieved in applications using the multilayer NN.
- and a lot of variants...
- **Deep Partially Linear Cox Model ; DPLCM** (Zhong et al. 2022)
  - "The first theoretical analysis of neural survival regression."

$$\lambda(t \mid X, Z) = \lambda_0(t) \exp\left\{\beta^\top Z + g(X)\right\}$$

- **Neural Frailty Machine** (Ruofan, et al. 2023)
  - Extended to include frailty model and theory

Frailty Model : Beyond the CoxPH model and PH assumption

- Cox PH model (with the covariate vector Z)

$$\lambda(t \mid Z) = \lambda_0(t) \exp\left(\beta^\top Z\right)$$

- PH assumption gives **time-independent harzard ratio.**

$$\frac{\lambda\left(t \mid Z^*\right)}{\lambda(t \mid \widetilde{Z})} = \exp\left(\beta^\top\left(Z^* - \widetilde{Z}\right)\right)$$

- Frailty models extend CoxPH model via **multiplicative random effect** to capture unobserved heterogeneity

$$\lambda(t \mid Z, \omega) = \omega \widetilde{\nu}(t, Z)$$

$$\omega \sim f_\theta(\omega)$$

; usually 1-dim parametrized

and positive r.v. (e.g., Gamma)

Neural Frailty Machine : Two frameworks

- Frailty model

$$\lambda(t \mid Z, \omega) = \omega \widetilde{\nu}(t, Z)$$

if $\omega = 1$ (degenerated) and $\widetilde{\nu}(t, Z) = \lambda_0(t)\exp(\beta^\top Z)$, it becomes CoxPH model.

- The **proportional frailty** scheme (PF)

$$\begin{aligned}\widetilde{\nu}(t, Z) &= \widetilde{h}(t)\widetilde{m}(Z) \\ &= \exp(h(t) + m(Z)) \quad ; h \text{ and } m \text{ are approximated by DNN}\end{aligned}$$

- The **fully neural** scheme (FN)

$$\widetilde{\nu}(t, Z) = \exp(\nu(t, Z)) \quad ; \nu \text{ are approximated by DNN}$$

* Both schemes use $\omega \sim \mathrm{Gamma}(1, \theta)$

Estimating strategy : Maximizing log partial likelihood

- Using the frailty transform (negative log Laplace transform), one can construct **the log partial likelihood.**

$$G_\theta(x) = -\log\left(\mathbb{E}_{\omega\sim f_\theta}\left[e^{-\omega x}\right]\right)$$

- **PF** scheme:

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

- **FN** scheme:

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

* Integrals of an exponentially transformed DNN's are evaluated using numerical integration.

Asymptotic theory

- **True function** : β - Hölder class
- **DNN structure** : $O(\log n)$ layer and $O\left(n^{\frac{d}{\beta+d}} \log n\right)$ sparsity
- **Metric**

  Hellinger distance of conditional distribution: (L2 in Zhong et al. 2022)

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

  - PF scheme : $\phi_0 = (h_0, m_0, \theta_0)$ and $\widehat{\phi}_n = \left(\widehat{h}_n, \widehat{m}_n, \widehat{\theta}_n\right)$

  - FN scheme : $\phi_0 = (\nu_0, \theta_0)$ and $\widehat{\phi}_n = \left(\widehat{\nu}_n, \widehat{\theta}_n\right)$

  Remark 1. Because of the frailty transform, likelihood can not be well-controlled by the L2 metric.

  Remark 2. To develop L2 theory, need additional curvature assumption on likelihood.

- **Theorem : Convergence Rate**

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

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

Experiments

- Evaluation metrics :

$$\mathcal{S}(\ell, t_0, t_{\max}) = \int_{t_0}^{t_{\max}} \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\ell\left(0, \widehat{S}(t \mid Z_i)\right) I(T_i \leq t, \delta_i = 1)}{\widehat{S}_C(T_i)} + \frac{\ell\left(1, \widehat{S}(t \mid Z_i)\right) I(T_i > t)}{\widehat{S}_C(t)} \right] dt$$

  - $\ell : \{0,1\} \times [0,1] \to \mathbb{R}^+$; binary classification loss function
  - Integrated Brier score (IBS)

$$\ell(D, \widehat{D}) = \{D - \widehat{D}\}^2$$

  - Integrated negative binomial log-likelihood (INBLL)

$$\ell(D, \widehat{D}) = (1 - D) \log \widehat{D} + D \log(1 - \widehat{D})$$

- Details of the metric (Graf et al. 1999)

Experiments

- Real world data Results (IBS, INBLL)

  * **boldfaced** : the best result / underlined : the second-best result

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	$\underline{17.47}_{\pm 0.45}$	$\underline{51.48}_{\pm 1.23}$	$\mathbf{10.63}_{\pm 0.08}$	$\mathbf{32.81}_{\pm 0.14}$

- Author's remark.
  - *Not so much significant improvements*
  - Lack of open-to-public large-scale survival datasets.
  - No authoritative train-test splits.

- Real world data Results (C-index)

  * **boldfaced** : the best result / underlined : the second-best result

Model	METABRIC	RotGBSG	FLCHAIN	SUPPORT	MIMIC-III	KKBOX	Ave. Rank
CoxPH	$63.42_{\pm 1.81}$	$66.14_{\pm 1.46}$	$79.09_{\pm 1.11}$	$56.89_{\pm 0.91}$	$74.91_{\pm 0.00}$	$83.01_{\pm 0.00}$	11.33
GBM	$64.02_{\pm 1.79}$	$67.35_{\pm 1.16}$	$\mathbf{79.47}_{\pm 1.08}$	$61.46_{\pm 0.80}$	$75.20_{\pm 0.00}$	$85.84_{\pm 0.00}$	7.17
RSF	$64.47_{\pm 1.82}$	$67.33_{\pm 1.34}$	$78.75_{\pm 1.07}$	$61.63_{\pm 0.84}$	$75.47_{\pm 0.17}$	$85.79_{\pm 0.00}$	8.00
DeepSurv	$63.95_{\pm 2.12}$	$67.20_{\pm 1.22}$	$79.04_{\pm 1.14}$	$60.91_{\pm 0.85}$	$80.08_{\pm 0.44}$	$85.59_{\pm 0.08}$	8.50
CoxTime	$66.22_{\pm 1.69}$	$67.41_{\pm 1.35}$	$78.95_{\pm 1.01}$	$61.54_{\pm 0.87}$	$78.78_{\pm 0.62}$	$\mathbf{87.31}_{\pm 0.24}$	5.00
DeepHit	$66.33_{\pm 1.61}$	$66.38_{\pm 1.07}$	$78.48_{\pm 1.09}$	$\mathbf{63.20}_{\pm 0.85}$	$79.16_{\pm 0.59}$	$86.12_{\pm 0.26}$	6.50
DeepEH	$66.59_{\pm 2.00}$	$\mathbf{67.93}_{\pm 1.28}$	$78.71_{\pm 1.44}$	$61.51_{\pm 1.04}$	–	–	6.33
SuMo-net	$64.82_{\pm 1.80}$	$67.20_{\pm 1.31}$	$79.28_{\pm 1.02}$	$62.18_{\pm 0.78}$	$76.23_{\pm 1.06}$	$84.77_{\pm 0.02}$	7.00
SODEN	$64.82_{\pm 1.05}$	$66.97_{\pm 0.75}$	$79.00_{\pm 0.96}$	$61.10_{\pm 0.59}$	–	–	10.17
SurvNode	$64.64_{\pm 4.91}$	$67.30_{\pm 1.65}$	$76.11_{\pm 0.98}$	$55.37_{\pm 0.77}$	–	–	11.50
DCM	$65.76_{\pm 1.25}$	$66.75_{\pm 1.35}$	$78.61_{\pm 0.79}$	$62.19_{\pm 0.95}$	$76.45_{\pm 0.34}$	$83.48_{\pm 0.07}$	8.33
DeSurv	$65.88_{\pm 2.02}$	$67.30_{\pm 1.45}$	$78.97_{\pm 1.64}$	$61.47_{\pm 0.97}$	$\mathbf{80.97}_{\pm 0.30}$	$86.11_{\pm 0.05}$	5.67
NFM-PF	$64.98_{\pm 1.87}$	$\underline{67.77}_{\pm 1.35}$	$\underline{79.45}_{\pm 1.03}$	$61.33_{\pm 0.83}$	$79.56_{\pm 0.15}$	$86.23_{\pm 0.01}$	$\underline{4.67}$
NFM-FN	$\mathbf{66.63}_{\pm 1.82}$	$67.73_{\pm 1.29}$	$79.29_{\pm 0.93}$	$\underline{62.21}_{\pm 0.41}$	$\underline{80.18}_{\pm 0.20}$	$\underline{86.61}_{\pm 0.01}$	**2.16**

- Author's remark.
  - (Rindt et el., 2022) observed loose correlation between the C-index and the likelihood-based learning objective.

References

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

Faraggi, D., Simon, R. (1995). A neural network model for survival data. Statistics in medicine, 14(1), 73-82.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology, 18(1), 1-12.

Zhong, Q., Mueller, J., Wang, J. L. (2022). Deep learning for the partially linear Cox model. The Annals of Statistics, 50(3), 1348-1375.

Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine, 18(17-18), 2529-2545.

Rindt, D., Hu, R., Steinsaltz, D., Sejdinovic, D. (2022, May). Survival regression with proper scoring rules and monotonic neural networks. In International Conference on Artificial Intelligence and Statistics (pp. 1190-1205). PMLR.