Supplementary Material: DAGSurv: Directed Ayclic Graph Based Survival Analysis Using Deep Neural Networks

Ansh Kumar Sharma* ANSH18130@IIITD.AC.IN
Rahul Kukreja* RAHUL18254@IIITD.AC.IN *
Ranjitha Prasad RANJITHA@IIITD.AC.IN
ECE dept., IIIT Delhi
Shilpa Rao SHILPA@IIITG.AC.IN
ECE dept., IIIT Guwahati

Editors: Vineeth N Balasubramanian and Ivor Tsang

1. Proof of Propositions

In the appendix we present the proofs of propositions presented in the paper.

1.1. Proof of Proposition 1

The adjacency matrix $\mathbf{A}$ consists of entries such that $A(i,j) \neq 0$, if there exists a directed edge starting at node i and terminating at node j, while $A(i,j) = 0$ indicates that there is no directed edge between i and j. The adjacency matrix can be used to factorize the joint probability distribution function as (Koller and Friedman, 2009)

$$p(\mathbf{t}, \mathbf{X}|K_{\mathbf{A}}) = p(\mathbf{t}|\mathbf{X}, K_{\mathbf{A}})p(\mathbf{X}|K_{\mathbf{A}}) \tag{1}$$

$$= p(\mathbf{t}|\mathcal{X}_{pa(L+1)}) \prod_{i\in V, i\neq L+1} p(X_i|\mathcal{X}_{pa(i)}). \tag{2}$$

In the above, we use $p(\mathbf{X}|K_{\mathbf{A}}) = \prod_{i\in V, i\neq L+1} p(X_i|\mathcal{X}_{pa(i)})$ (Koller and Friedman, 2009), where $\mathbf{X}_{pa(i)}$ specifies the parents of the i-th vertex, as specified by the directed edge coming into the node i. In particular, $\mathcal{X}_{pa(L+1)}$ specifies the parent nodes of the target. We see from above that the adjacency matrix factorizes the joint distribution, and hence characterizes it.

1.2. Proof of Proposition 2

First we prove the sufficiency, i.e., if the i-th term in the factorization of $p(\mathbf{t}, \mathbf{X}|K_{\mathbf{A}})$ given by $p(X_i|\mathcal{X}_{pa(i)})$ (as given in (2)) is not equal to $p(X_i)$, for any i. If $A(j,i) \neq 0$ for any j where $i \neq j$, it implies that $\mathcal{X}_{pa(i)} = \{X_j\}$, and hence, the i-th term in the factorization is given by $p(X_i|\{X_j\})$.

In order to prove the necessary condition, we use the contradiction argument, where we first assume that if $p(X_i|\mathcal{X}_{pa(i)}) \neq p(X_i)$ for any i, then $\mathbf{A} = 0$. However, if $p(X_i|\mathcal{X}_{pa(i)}) \neq$

* * indicates equal contribution

$p(X_i)$, it implies that $\mathcal{X}_{pa(i)} \neq \{\}$, which further implies that node i has incoming edges, i.e., $A(j,i) \neq 0$ for some j. Hence, by contradiction, we prove the necessary condition.

1.3. Proof of Proposition 3

Using the chain rule (Cover, 1999) for entropy,

$$H(\mathbf{X}) = \sum_{i=1}^{L} H(X_i|\mathcal{X}_{pa(i)}), \tag{3}$$

where the k-th entry in the factorization of $p(\mathbf{t}, \mathbf{X}|K_{\mathbf{A}})$, given by $p(X_k|\mathcal{X}_{pa(k)})$ is not equal to $p(X_k)$. Hence, we rewrite the above expression as

$$H(\mathbf{X}) \leq H(X_k|\mathcal{X}_{pa(k)}) + \sum_{i=1, i\neq k}^{L} H(X_i) < \sum_{i=1}^{L} H(X_i), \tag{4}$$

since we know that $H(X_k|\mathcal{X}_{pa(k)}) < H(X_k)$. This completes the proof.

2. Additional Experimental Results

In this section, we present additional results that help us better understand the proposed `DAGSurv` framework.

2.1. KKBox Dataset

We performed experiments with the high-dimensional KKBox dataset. This dataset is from KKBOX, which is Asia's music streaming service that consists of Asia-Pop music library with more than 30 million tracks. This dataset was the part of a data challenge at WSDM 2018, where the key questions involved design of algorithms to predict users' preference regarding a new song or a new artist, so that appropriate recommendations were made to new users. This dataset has also been employed for survival analysis (Kvamme et al., 2019).

KKbox	
Algorithms	C_{td} (95% C_I)
`DAGSurv`	0.8635 ± 0.0002
`DeepHit`	0.9003 ± 0.0002
`DeepSurv`	0.8124 ± 0.0002
`CoxTime`	0.8229 ± 0.0002

Table 1: C_{td} for KKbox

We see that `Deephit` performs marginally better as compared to `DAGSurv`. In particular, `Deephit` performs very well if there is abundant data to learn from, especially because it does not consider any model assumptions. KKBox dataset consists of more than 10^5 observations, and is well-suited for `DeepHit`. However, we see that performance of `DAGSurv`

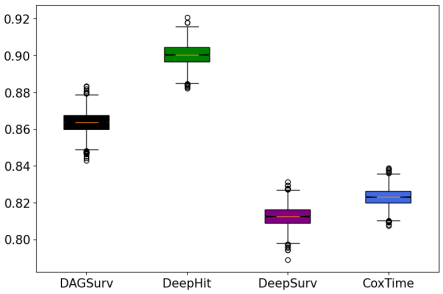

Figure 1: Box-plot: C_{td} for KKbox

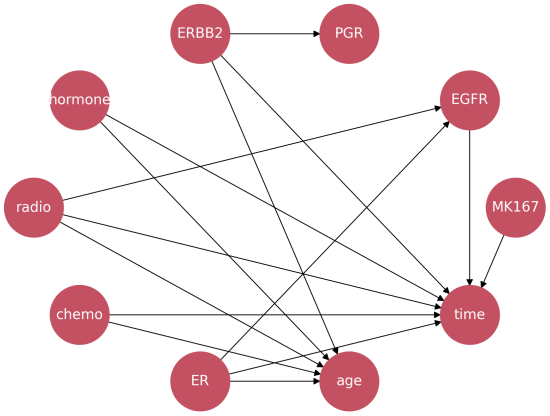

Figure 2: DAG: METABRIC (NOTEARS)

is is better compared to other baselines. Hence, we were able to confirm that `DAGSurv` performs well even for high-dimensional datasets. From the box-plot in Fig. 2, we see that `DAGSurv`, like all other methods, has little variability around its mean value.

2.2. Experiment with `DAG-NOTEARS`

Another popular tool for DAG structure learning is `DAG-NOTEARS` (Zheng et al., 2018). We experimented with this tool for obtaining an alternate graph for the Metabric dataset. The results are as given in Table 2. We see that the DAG structure obtained using DAG-GNN performs marginally better compared to the one using DAG-NOTEARS. This helps us infer that the graph obtained using DAG-GNN suits better in the context of survival analysis.

References

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, 20(129):1–30, 2019.

METABRIC	
Algorithms	C_{td} (95% C_I)
`DAGSurv (DAG-GNN)`	**0.7323** ± 0.0056
`DAGSurv (NOTEARS)`	**0.7233** ± 0.0034
`DeepHit`	0.7309 ± 0.0047
`DeepSurv`	0.6575 ± 0.0021
`CoxTime`	0.6679 ± 0.0020

Table 2: C_{td} for METABRIC

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.