Exploring the Wasserstein metric for survival analysis

Tristan Sylvain* TRISTAN.SYLVAIN@GMAIL.COM
Mila, University of Montreal

Margaux Luck* MARGAUX.LUCK@GMAIL.COM
Mila

Joseph Paul Cohen JOSEPH@JOSEPHPCOHEN.COM
Mila, University of Montreal

Heloise Cardinal HELOISE.CARDINAL.CHUM@SSSS.GOUV.QC.CA
Centre Hospitalier de l'Université de Montréal

Andrea Lodi ANDREA.LODI@POLYMTL.CA
Polytechnique Montréal

Yoshua Bengio YOSHUA.BENGIO@MILA.QUEBEC
Mila, University of Montreal

Abstract

Survival analysis is a type of semi-supervised task where the target output (the survival time) is often right-censored. Utilizing this information is a challenge because it is not obvious how to correctly incorporate these censored examples into a model. We study how three categories of loss functions can take advantage of this information: partial likelihood methods, rank methods, and our own classification method based on a Wasserstein metric (WM) and the non-parametric Kaplan Meier (KM) estimate of the probability density to impute the labels of censored examples. The proposed method predicts the probability distribution of an event, letting us compute survival curves and expected times of survival that are easier to interpret than the rank. We also demonstrate that this approach directly optimizes the expected C-index which is the most common evaluation metric for survival models.

Introduction

Survival analysis aims to predict the first occurrence of a stochastic event, conditioned on a set of features. Cases where the sample time wasn't recorded because the event in question wasn't observed can be framed as a particular type of semi-supervised learning where part of the target values are referred to as right-censored. In formal terms, we can say that for some examples we do not have the time of event T, but rather a time T_0 (censoring time) such that we know $T > T_0$. The classical approach to survival analysis, the Cox proportional hazards model, Cox (1972) takes into account censored samples. Ranking approaches Raykar et al. (2007) is another way to take these censored samples into account by incorporating them into the training using pairwise ranking loss. Although the exact time of the event is not known, the pairwise relationship with respect to a censoring date is known for an event occurring before the censored event. We would thus like to predict the probability distribution of an event.

[*] equal contribution

In this study, we propose to use the Wasserstein metric (WM) to learn the probability distribution of the event time. This approach not only provides a more easily interpreted prediction but allows us to impute the distribution of censored samples given global survival statistics with the non-parametric Kaplan Meier (KM) estimate. Our intuition is that training with the KM estimate provides a richer signal during training than a rank loss. We find that this method produces useful predictions despite a high percentage of censored samples. We also find that this approach directly optimizes the C-index Harrell et al. (1982) which is the most common evaluation metric for survival models. We compared our proposed loss to a set of common ranking-specific losses on several reference survival datasets. Our method is competitive with ranking and likelihood methods that take into account pairwise interactions, and has the added advantage of providing a good estimation of the event time and the survival curve, yielding more easily interpreted predictions.

Survival data

In what follows, we will use the following notation (summarized in Table 1). Let $\mathbf{x}^{(i)}$ be the feature vector of the i-th example and let $\mathbf{y}_t^{(i)}$ take value 1 if event i happened at time t and 0 otherwise. Moreover, let $\hat{\mathbf{y}}_t^{(i)}$ be the estimated probability of event i happening at time t and let $t^{(i)}$ be the (scalar) actual time of event i. We denote by $\mathbf{z}_t^{(i)}$ and $\hat{\mathbf{z}}_t^{(i)}$ the true and estimated cumulative probability distribution of y. Namely, $\mathbf{z}_{t_0}^{(i)} = \sum_{t<t_0} \mathbf{y}_t^{(i)}$. Finally, let $c^{(i)}$ be 1 if example i is observed (non-censored) and 0 otherwise.

Notation	Meaning
T	random variable for time of event
$\mathbf{x}^{(i)}$	feature vector
$\mathbf{y}^{(i)}$	$\mathbf{y}_t^{(i)}$ is 1 if event occurred at time t, 0 otherwise
$t^{(i)}$	time of the event (real-valued)
$c^{(i)}$	indicator of whether example is right-censored
$\mathbf{z}_t^{(i)}$	true CDF ($\sum_{t'<t} \mathbf{y}_{t'}^{(i)}$) at time t
$\hat{\mathbf{z}}_t^{(i)}$	estimated CDF at time t

Table 1: Notations used in this section, with superscript $^{(i)}$ indicating the i-th example is concerned.

Ties and censored data

Survival datasets describe events that can have a low temporal resolution (time scale) causing ties between samples. A given *unique time* (at a given resolution, e.g. one day) can correspond to multiple events. Such events are *tied* which implies that more precise predictions are not relevant. However, they must be given special attention in constructing loss functions.

As mentioned earlier, another characteristic of survival data is that they are right-censored. We can still use these examples but only in comparison with samples that had an event before the date of censorship or by imputing the event time based on statistics over the data.

Metric of evaluation

The concordance index or C-index Harrell et al. (1982) is the standard evaluation metric for ranking survival models. It corresponds to the probability of correctly ranking a pair of event times sampled from the data distribution, and to the normalized Kendall tau metric between the true and predicted distribution Kendall (1938). It can be seen as a generalization of the Area Under the Receiver Operating Characteristic Curve (AUROC) that can handle right-censored data Raykar et al. (2007).

We define an *acceptable* pair as one for which we are sure that the first event occurs before the second. These are the pairs for which the first element is non-censored, and for which the censoring or event time of the second element is strictly greater than the first. Let $\mathcal{A}$ be the set of acceptable pairs. Then, the C-index to be maximized can be written as:

$$\frac{1}{|\mathcal{A}|}\sum_{(\mathbf{x}^{(i)},\mathbf{x}^{(j)})\in\mathcal{A}} \mathbb{1}\Big(f(\mathbf{x}^{(i)})<f(\mathbf{x}^{(j)})\Big)+\tfrac{1}{2}\mathbb{1}\Big(f(\mathbf{x}^{(i)})=f(\mathbf{x}^{(j)})\Big).$$

Loss functions for censored data

The aim of this paper is to explore the benefits of using a WM as a loss function for survival data analysis. As a result, in this section we present loss functions in the context of survival prediction for censored data. We divide these loss functions into three categories: partial likelihood methods, rank methods, and our classification method based on the WM.

As we focused on comparable neural network architectures, we exclude from our benchmark more classical methods such as Random Survival Forests Ishwaran et al. (2008) (shown in Katzman et al. (2016); Luck et al. (2017) to perform worse than deep learning models trained with a Cox Loss). We also did not attempt a comparison with Deep Exponential families Ranganath et al. (2016), which use a generative approach, making a direct comparison harder.

Cox Model

Cox introduced a general conditional log-likelihood to fit survival models, in which the probability of observations is maximized Cox (1972). It was demonstrated by Raykar et al. (2007) that Cox's partial likelihood is approximately equivalent to maximizing the C-index. We present the general formula, with a real-valued score prediction function f_θ estimating the probability of the event at a particular time, given input features $\mathbf{x}^{(i)}$.

Denoting the predicted score $f_\theta(\mathbf{x}^{(i)}) = \exp\theta\cdot\mathbf{x}^{(i)}$ the loss is:

$$\ell(\theta) = \sum_{i:c^{(i)}=1}\Big(\log f_\theta(\mathbf{x}^{(i)}) - \log\sum_{j:t^{(j)}\geq t^{(i)}} f_\theta(\mathbf{x}^{(j)})\Big).$$

We also consider a variant of this loss, Efron's approximation Efron (1977) that commonly improves performance when there are many tied event times. In that case, the left-hand term of the log-likelihood remains the same, but the right-hand sum $-\sum_{i:c^{(i)}=1}\log\sum_{j:t^{(j)}\geq t^{(i)}} f_\theta(\mathbf{x}^{(i)})$ is replaced by

$$-\sum_{\tau}\sum_{l=0}^{|H_\tau|-1}\log\Big(\sum_{j:t^{(j)}\geq\tau} f_\theta(\mathbf{x}^{(j)}) - \frac{l}{|H_\tau|}\sum_{k\in H_\tau} f_\theta(\mathbf{x}^{(k)})\Big),$$

where τ denotes the unique times, H_τ the set of indices i such that $t^{(i)}=\tau$ and $c^{(i)}=1$.

In our experiments, the Cox variant refers to a multi-layer perceptron (MLP) f_θ trained with the normal Cox loss or with Efron's approximation loss, as in Katzman et al. (2016); Luck et al. (2017).

Ranking

Many methods attempt to directly predict the rank of the different examples. This is done by learning the following objective:

$$\underset{\theta}{\operatorname{argmax}} \frac{1}{|\mathcal{A}|} \sum_{(\mathbf{x}^{(i)},\mathbf{x}^{(j)})\in\mathcal{A}} \phi(f_\theta(\mathbf{x}^{(i)}) - f_\theta(\mathbf{x}^{(j)}))$$

where ϕ is a function that relaxes the non-differentiable $\mathbb{1}$ of the C-index Raykar et al. (2007). We evaluated the functions used in Raykar et al. (2007), Ranking SVM Herbrich et al. (2000), Rankboost Freund et al. (2003) and RankNet Burges et al. (2005). These functions have been shown in Kalbfleisch (1978) to correspond to lower bounds on the C-index. Some of the expressions mentioned in those works are identical up to a constant, which would have no impact on the final result. We use σ to denote the Sigmoid function $z \to \frac{1}{1+\exp(-z)}$.

Wasserstein metric

The Wasserstein metric was considered among others as part of a tree-based algorithm for survival analysis Crowley et al. (1995). The work proposes the metric as one of many possible choices, and does not propose a theoretical justification for its use. While there have to our knowledge been no other previous attempts to use the Wasserstein metric on survival data or ranking problems, Frogner et al. (2015) used a Wasserstein loss for image classification and tag prediction. Hou et al. (2016) and Beckham and Pal (2017) apply a Wasserstein metric for the more restrictive case of ordinal classification. Recently, Mena et al. (2018) used the Sinkhorn algorithm, which is commonly used in optimal transport applications, as an analogy to the Softmax for permutations.

The WM is the minimum cost to transport the mass from one probability distribution to another. In the case of distributions of discrete supports (histograms of class probabilities), this is computed by moving probability mass from one class to another, according to the ground distance matrix specifying the cost to transport probability mass to and from different classes. Thus, the WM takes advantage of knowledge of the structure of the space of values considered, e.g., the 1-dimensional real-valued time axis, so that some errors (e.g. between neighboring events) are appropriately penalized less than others.

The WM is particularly adapted to a survival context. We denote p_r the true data distribution, and p_θ the distribution estimated by the model. We write Π the set of joint distributions $p(\cdot,\cdot)$ with left and right marginals p_θ and p_r respectively. Given an example $\mathbf{x}$ and corresponding real time of event T, we can write:

$$W(p_\theta,p_r) = \inf_{p(\cdot,\cdot|\mathbf{x})\in\Pi} \mathbb{E}_{T_1,T_2\sim p(\cdot,\cdot|\mathbf{x})}\left[d(T_1,T_2)\right]$$

As p_r is a Dirac, we have that:

$$\mathbb{E}_{T_1,T_2\sim p(\cdot,\cdot|\mathbf{x})}\left[d(T_1,T_2)\right] = \mathbb{E}_{T_1\sim p(\cdot,T|\mathbf{x})}\left[d(T_1,T)\right]$$

In all that follows, $d(T_1,T_2)$ is chosen to be proportional to the number of train set elements having events between T_1 and T_2. The term is therefore $\mathbb{E}_{T\sim p_\theta(\cdot,T|\mathbf{x})}\left[1-\text{Cindex}\right]$.

USE AS A LEARNING OBJECTIVE

Levina and Bickel (2001) notes that under certain conditions satisfied in the case of ordinal classification, the WM takes the following expression:

$$\text{WM}(p,q) = \left(\frac{1}{T}\right)^{1/l} ||CDF(p) - CDF(q)||_l,$$

where T is the size of the Softmax layer and $CDF(.)$ is a function that returns the cumulative density function of its input density. Here, p and q are two probability distributions with discrete supports.

We write $f_\theta(\mathbf{x}^{(i)}) = \hat{\mathbf{z}}_\theta^{(i)}$ to highlight the dependency on θ. The objective can be written as:

$$\underset{\theta}{\operatorname{argmin}} \frac{1}{T} \sum_i w_i ||\hat{\mathbf{z}}_\theta^{(i)} - \mathbf{z}^{(i)}||^l.$$

Here w_i corresponds to the $d(T_i, T_{i+1})$. It is computed using train set data. The correspondence to the expected C-index only holds for $l = 1$. Works in ordinal classification of images considered the squared Wasserstein metric Hou et al. (2016) in addition to the $L1$ distance. We considered additional values of l, namely 1.5 and 2 in our experiments as a relaxation of the problem, and observed in practice better gradients for those higher values of l.

IMPUTING MISSING VALUES FOR CLASSIFICATION

In order to allow the WM objective to lead to good training, we have imputed the CDF of the censored data with $1. - KM$, where KM is the Kaplan-Meier non parametric estimate of the survival distribution function computed on the training set (see Figure 1). With the KM estimator, the survival distribution function $S(t)$ is estimated as a step function, where the value at time t_i is calculated as follows:

$$\hat{S}(t_i) = \hat{S}(t_{i-1})(1 - d_i/n_i),$$

with d_i denoting the number of events at t_i and n_i the number of patients alive just before t_i.

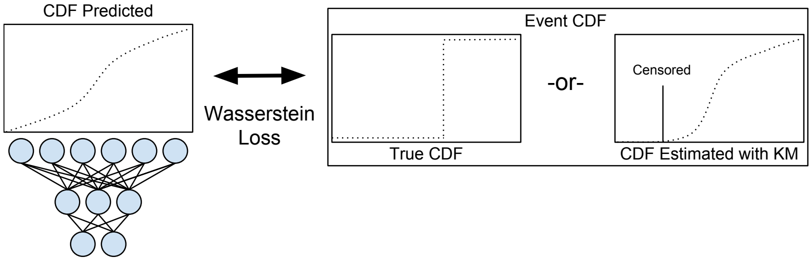


Figure 1: An overview of the proposed distribution matching loss. In the case that a sample is censored the KM estimate is used to impute the probability that should be assigned for that event.

Experiments

Datasets

We assess the presented models on a variety of publicly available datasets. The characteristics of these datasets are summarized in Table 2.

Datasets	Nb. samples	Nb. (%) censored	Nb. (%) unique times	Nb. features
SUPPORT2	9105	2904 (32.2)	1724 (19.1)	98
AIDS3	3985	2223 (55.8)	1506 (37.8)	19
COLON	929	477 (51.3)	780 (84.0)	48

Table 2: Characteristics of the datasets used in our evaluation. The datasets have different numbers of samples, percentage of censored, and tied patients. The features are typically continuous or discrete clinical attributes.

SUPPORT2[1] records the survival time for patients of the SUPPORT study. It is a medium-sized dataset with a relatively low proportion of censored patients.

AIDS3[2] corresponds to the Australian AIDS Survival Data. Data on patients diagnosed with AIDS in Australia before July 1, 1991. This data set has been slightly altered as a condition of its release, to ensure patient confidentiality. It is a small dataset with a medium proportion of censored patients.

COLON[2] consists of data from the first successful trials of adjuvant chemotherapy for colon cancer. We considered death as a target event for our study.

Data pre-processing

We used a one-hot encoding for categorical features, and unit scaling for continuous features. For features with missing values, we added an indicator function for the absence of a value.

We performed 5 fold cross validation and kept 20% of the train set as a validation set. The prediction performance was reported as mean ± standard error of the C-index over the 5 folds. Early stopping was performed on the validation C-index.

We used a multi-layer perceptron with ReLU activation functions where applicable, and used Dropout Hinton et al. (2012), Batch Normalization Ioffe and Szegedy (2015) and L2 regularization on the weights. We used the Adam optimizer. For the ranking and log-likelihood methods the output was a single unit with a linear activation function. For the methods requiring a prediction of output times, we used a Softmax function. Our code was written in PyTorch Paszke et al. (2017)

We performed hyper-parameter tuning with Bayesian optimization for each split independently. The training and validation sets were used to determine the optimal hyper-parameters, namely number of layers (1-4), layer width (10-1000), dropout fraction (0.0-1.0), learning rate (1e-6, 1e-3) and, the L_2 regularization coefficient α (1e-8, 1e-2). For WM models, we also optimized the time-granularity in the range (1-5). We use a patience of 20 for early stopping. The hyper-parameters of each model were selected with the same number of runs, in order to provide a fair comparison.

Quantitative evaluation

We aim to demonstrate that our method provides a good time prediction, while remaining competitive in terms of C-index with more specialized methods, such as ranking losses and Cox partial-likelihoods. To do so, we first report C-index values over multiple runs of our model and the baselines, on different datasets. We also introduce experiments to evaluate the effect of increasing the proportion of censored individuals, and either removing censoring, or treating censoring as an actual event. Finally, we report the absolute error relative to the event range, allowing a comparison with state-of-the-art survival models.

COMPARISON OF DIFFERENT RANKING METHODS

We study the impact of the different loss functions in Table 3. We determine how the standard Cox model performs in comparison to ranking and classification losses.

The ranking functions are plotted in Figure 2 to illustrate how they scale errors differently.

One way to study the differences between these methods is to look at the counts of how many examples were wrongly predicted at every time point. In Figure 3, we look at the number of samples over predicted at every time point using the following calculation for each non-censored patient:

$$\sum_t \frac{\mathbb{1}(\text{rank} > t)}{N - t}.$$

[1] available at http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets

[2] available at https://vincentarelbundock.github.io/Rdatasets/datasets.html

Loss Type	Variant	SUPPORT2	AIDS3	COLON DEATH
Partial likelihood	Cox	87.16 ± 0.23	55.92 ± 0.61	64.25 ± 0.27
Partial likelihood	Cox Efron's	86.86 ± 0.22	56.44 ± 0.29	62.24 ± 0.37
Ranking	$\sigma(z)$	87.21 ± 0.20	55.38 ± 0.63	63.83 ± 0.25
Ranking	Log-sigmoid	86.98 ± 0.21	$\mathbf{57.34 \pm 0.41}$	63.97 ± 0.31
Ranking	$(z-1)_+$	$\mathbf{87.22 \pm 0.21}$	57.03 ± 0.33	$\mathbf{65.16 \pm 0.30}$
Ranking	$1-\exp(-z)$	87.21 ± 0.18	55.98 ± 0.54	63.93 ± 0.45
WM (ours)	$l=1$	86.52 ± 0.12	55.14 ± 0.56	62.92 ± 0.59
WM (ours)	$l=1.5$	86.73 ± 0.21	57.25 ± 0.4	63.17 ± 0.25
WM (ours)	$l=2$	87.15 ± 0.22	56.07 ± 0.48	63.60 ± 0.58

Table 3: Performance scores of the different methods. The table reports the C-index mean ± standard error over the 5 fold. For each dataset, the best model in terms of mean score is highlighted in bold. We draw the readers attention to the classification losses which are among the losses that give the best results.

In Figure 4 we also directly plot the alignment of predicted rank against the ground truth rank for the SUPPORT2 dataset. We observe that WM captures the rank well and exposes a step in the predictions which may expose some unknown pattern in the data. The red line indicates the min and the max rank to highlight where samples are tied.

TIME PREDICTION EVALUATION

The performance of our models are also evaluated in terms of absolute error relative to the event range, i.e., $|\hat{t}-t|/t_{\text{max}}$. For censored events, the relative error is defined $\max(0,t-\hat{t})/t_{max}$, to account for the fact that no error is made as long as $t \leq \hat{t}$. Table 4 shows the mean ± standard error over the 5 fold of the median and 50% empirical intervals for relative absolute errors on non-censored events, on all test-data. Our method is competitive with the state-of-the-art DATE methods, without requiring the introduction of a generative model, or extra regularization losses. We achieve better C-index values than those reported, and for the best methods $(l=1, l=1.5)$, the 50% empirical ranges is from the same range. Interestingly, for $l=2$ our model performs worse, while still maintaining a good C-index value. This is due to the fact that the evaluation criteria relates to a $L1$ error.

Method	SUPPORT2	AIDS3	COLON DEATH
WM (l = 1)	1.545±0.059 (0.274, 16.818)	29.781±1.109 (29.904, 33.100)	48.053±3.886 (31.343, 62.771)
WM (l = 1.5)	1.938±0.272 (0.414, 18.524)	29.458±0.007 (27.430, 31.564)	31.757±2.716 (17.897, 39.596)
WM (l = 2)	3.721±0.232 (1.196, 20.612)	28.769±0.716 (27.285, 31.001)	32.314±0.502 (16.863, 38.385)
DATE	2.7 (0.4,16.1)*	-	-
DATE-AE	1.5 (0.4,19.2)*	-	-
DRAFT	2.0 (0.2,35.3)*	-	-

Table 4: Mean ± standard error over the 5 fold of that median relative absolute errors (as percentages of t_{max}), on non-censored data. Other methods are the recent state-of-the-art models for survival modeling from Chapfuwa et al. (2018), indicated by an asterisk (only one split considered in this case). Ranges in parentheses are 50% empirical ranges over (median) test-set predictions. - indicates no reported value.

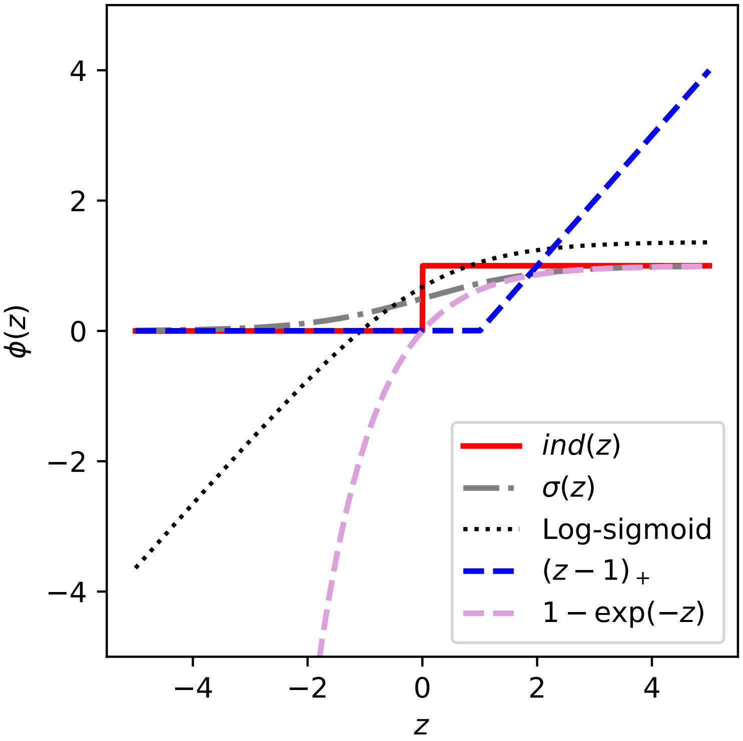
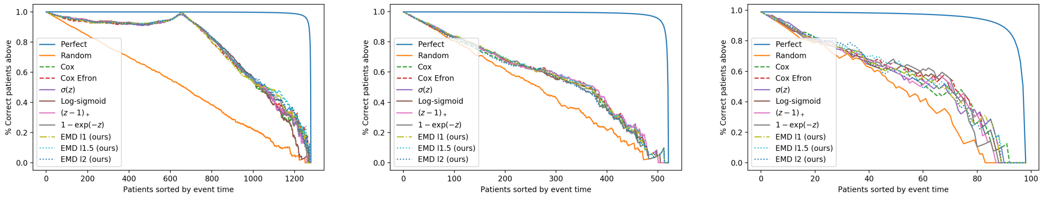


Figure 2: The different functions ϕ used in the ranking losses. Depending on this function, ranking errors will be accounted for differently which can impact overall learning.

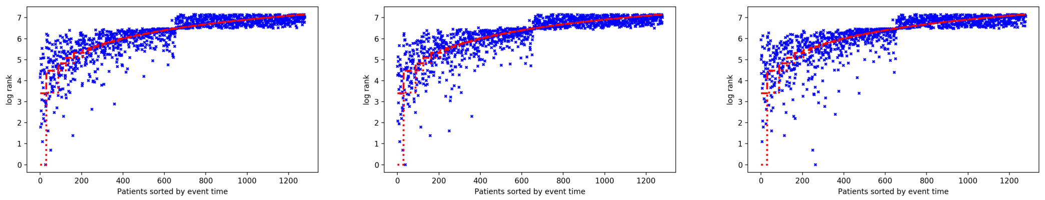


(*a*) SUPPORT2 (*b*) AIDS3 (*c*) COLON

Figure 3: We compare each method by plotting the percentage of patients that survived after their predicted rank. The ranking provided by each method fixed training/validation/test split is used for all methods and the test set is shown.

IMPACT OF USING CENSORED DATA

The purpose of this section is to demonstrate why censoring should not be ignored due to the information we can garner from it. We compare three methods to account for censored data. First, we completely removed censored examples from the training set (no censored data). Second, we also considered the censoring time to correspond to an actual event occurrence (transforming each example censored at time t into the same example with an event occurring at time t) (death at censoring). Finally,



(*a*) Cox (*b*) Ranking $\sigma(z)$ (*c*) WM l=1 (ours)

Figure 4: Actual and predicted log rank for the uncensored patients only. From left to right the Cox Efron, the log sigmoid and the WM for the SUPPORT2 dataset. The min and the max correct rank is shown in red. They overlap unless there are many ties.

we also listed results for the standard approach (with censored data). In the case of WM, the censored times are imputed with the $(1-KM)$ curve.

We run this experiment on the SUPPORT2 dataset for the three best methods of each category as it is the largest public dataset we have: Cox Efron's, σ and our method's WM. The results are presented in Table 5. Overall, the WM is equivalent to the others in the two contexts examined.

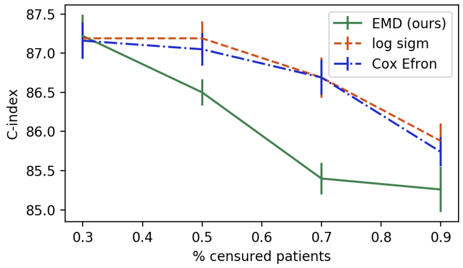


Figure 5: We study how the composition of censored and uncensored patients during training impacts the C-index mean ± standard error over the 5 fold in the SUPPORT2 dataset. The validation and test sets are fixed and the training set has censored patients introduced by marking patients as censored at random. The plot starts at 30% because the dataset has that many censored patients by default.

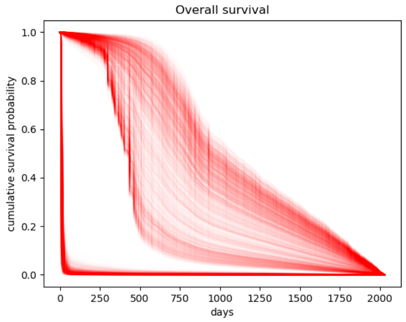


Figure 6: Predicted survival curve for all individuals in the test set of a single fold of SUPPORT2. Darker hues of red indicate higher densities of examples having similar survival curves.

Method	WM (l = 2)	Ranking (σ)	Cox
No censored data	85.322 ± 0.169	85.830 ± 0.230	85.460 ± 0.192
Death at censoring	83.714 ± 0.242	83.743 ± 0.229	83.305 ± 0.208
With censored data	87.147 ± 0.216	87.214 ± 0.198	87.162 ± 0.226

Table 5: We explore how the three categories of methods are impacted by adding censored data. The table reports the C-index mean ± standard error over the 5 fold. For "Death at censoring", we set the death event as the censored time. It is clear that censored data contains information that we can use to make better predictions.

EXPLORING THE IMPACT OF CENSORED DATA

In order to determine how much of an improvement we can obtain from incorporating censored data we can vary the composition of samples that are censored in the training data, while keeping the validation and test sets the same. In Figure 5 we show the evolution of the C-index with different percentages of censoring of the training set in the SUPPORT2 dataset.

Overall, we observe a drop in performance of our method relative to the others. This is most likely due to the fact that the KM estimator used for the WM during training benefits from having many uncensored examples.

Visual comparison of the test-set survival distribution and survival curve prediction

We present qualitative results that show particularities of the data and predictions for SUPPORT2, the dataset that gives the highest C-index.

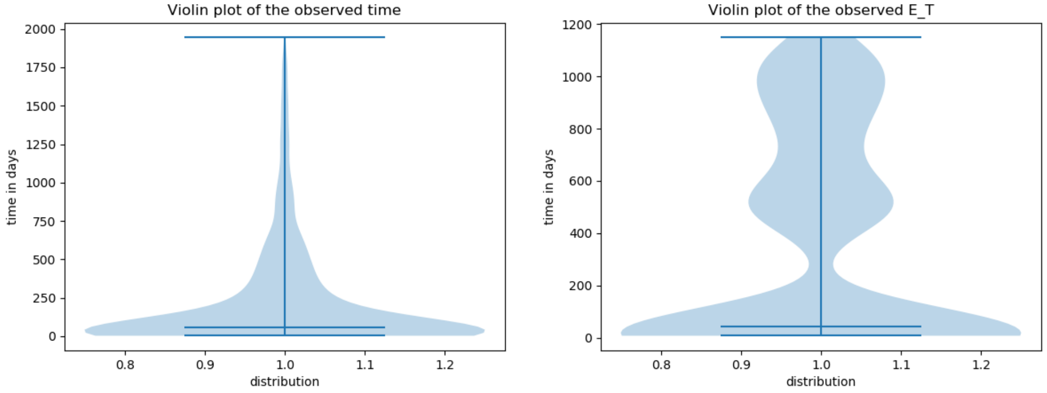
Figure 6 shows the predicted survival curves for all test set patients for SUPPORT2. Many events occur very early, making these events a separate mode. The strong penalty placed by the weighted WM on predicting an event too early explains the gap between the two groups of survival curves. Addressing this by adding a regularization term could improve the resulting error.

The model is very good at predicting events early on. Further time steps introduce greater uncertainty as the probability mass is spread out. Wrongly predicted examples tend to occur at later times. This is again a consequence of the fact that many events occur early in training, meaning that the transportation cost is highest between the early time buckets.

Figure 7 shows the violin plots for the real time and expected time (as predicted by the model). This recapitulates what was mentioned previously about the model being better at predicting earlier time steps due to the nature of the loss. The bottleneck in the right-most plot (expected time) corresponds to a region of high transportation cost due to the large quantities of events occurring at that point in time.

Conclusion

We proposed a new method based on the WM for survival data analysis. Experiments on the different datasets show that our models trained with the WM loss produce accurate predictions (evaluated in terms of C-index) compared to the more classical losses of the Cox model and ranking loss functions, which directly approximate a lower bound of the C-index. While not always state of the art, our method always yields some of the best results for each dataset.



(*a*) Cox

(*b*) Cox

Figure 7: Violin plots for the true time (left), and the expected time (right) predicted by the model. The bottleneck on the right corresponds to a part of the time scale with high transportation cost, meaning examples will be predicted further away.

Our method is resilient to a modification of the treatment of censored examples (censoring treated as an event and removing censored examples), but sees its performance reduced somewhat once the proportion of censored examples strongly increases.

We have also shown that our method can be seen as directly optimizing the expected C-index which is the most common evaluation metric for ranking survival models. Moreover, our results demonstrate that imputing the values with the KM curve for the missing times in a classification framework can increase the resulting C-index.

Finally, in addition to those advantages, our model outputs an actual, interpretable prediction for the time of event, which is very useful in e.g. a clinical setting to convey prognostic information to patients. We have identical results to that of state of the art time-oriented methods, without using any of the extra regularizers used by those methods to improve generalization.

References

Christopher Beckham and Christopher Pal. Unimodal probability distributions for deep ordinal classification. *International Conference on Machine Learning*, 2017. URL http://arxiv.org/abs/1705.05278.

Christopher J C Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to Rank using Gradient Descent. In *International Conference on Machine Learning*, pages 89–96. ACM, 2005.

Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin, and Ricardo Henao. Adversarial time-to-event modeling. *arXiv preprint arXiv:1804.03184*, 2018.

D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.

J Crowley, M LeBlanc, R Gentleman, and S Salmon. Exploratory methods in survival analysis. *Lecture Notes-Monograph Series*, pages 55–77, 1995.

Bradley Efron. The Efficiency of Cox's Likelihood for Censored Data Function. *Journal of the American Statistical Association*, 72(359):557–565, 1977.

Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*, 4(Nov):933–969, 2003. ISSN 15324435. doi: 10.1162/jmlr.2003.4.6.933.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *Journal of the American Medical Association*, 1982. ISSN 15383598. doi: 10.1001/jama.1982.03320430047030.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, 2000. ISSN 10495258. doi: 10.1002/ardp.18440880237. URL http://research.microsoft.com/apps/pubs/default.aspx?id=65610{\T1\textgreater}{%}5Cnhttp://research.microsoft.com/pubs/65610/herobergrae99.ps.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580*, 2012. ISSN 9781467394673. doi: arXiv:1207.0580. URL http://arxiv.org/abs/1207.0580.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared Earth Mover's Distance-based Loss for Training Deep Neural Networks. *arXiv:1611.05916*, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456, 2015.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 2008. ISSN 19326157. doi: 10.1214/08-AOAS169.

John Kalbfleisch. Non-Parametric Bayesian Analysis of Survival Time Data. *Journal of the Royal Statistical Society. Series B*, 40(2):214–221, 1978.

Jared 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. *International Conference of Machine Learning Computational Biology Workshop*, (i):1–11, 2016. URL http://arxiv.org/abs/1606.00931.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Elizaveta Levina and Peter Bickel. The Earth Mover's distance is the Mallows distance: Some insights from statistics. In *International Conference on Computer Vision*, volume 2, pages 251–256. IEEE, 2001. ISBN 0769511430. doi: 10.1109/ICCV.2001.937632.

Margaux Luck, Tristan Sylvain, Héloïse Cardinal, Andrea Lodi, and Yoshua Bengio. Deep Learning for Patient-Specific Kidney Graft Survival Analysis. *arXiv:1705.10245*, 2017. URL http://arxiv.org/abs/1705.10245.

Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. *arXiv preprint arXiv:1802.08665*, 2018.

Adam Paszke, Gregory Chanan, Zeming Lin, Sam Gross, Edward Yang, Luca Antiga, and Zachary Devito. Automatic differentiation in PyTorch. pages 1–4, 2017.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep Survival Analysis. *Machine Learning and Healthcare Conference*, 2016. doi: 10.1007/978-3-319-66185-8. URL http://arxiv.org/abs/1608.02158.

Vikas C Raykar, Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, and Philippe Lambin. On ranking in survival analysis: Bounds on the concordance index. In *Neural Information Processing Systems*, pages 1209–1216, 2007. ISBN 160560352X. doi: 10.1.1.121.2670. URL http://machinelearning.wustl.edu/mlpapers/paper{_}files/NIPS2007{_}535.pdf.