# Generalizing right-censored data into standard regression problem through complete ranking

Yuanfang Guan, Nov 18, 2015

(Update: Dec 9, 2017)

This document describes a method to generalize right-censored data (survival problems) into a standard regression problem via complete ranking of all training data points regardless of censoring status, allowing building in any base-learner to predict survival.

This is the winning solution for the 2015 ALS Prize4Life DREAM challenge, the 2017 Multiple Myeloma DREAM challenge, sub-challenge 3, evaluated by prediction accuracy in blind survival data. bioRxiv rejected my submission saying it does not even look like a research paper. So please directly cite: http://guanlab.ccmb.med.umich.edu/yuanfang/Guan_rank.pdf

I will review existing survival models (Cox-related, survival-random-forest, etc.) and their limitations as an incomplete ranking of training data, and the resulting limitation in the choices of base-learners. Then, I will describe an approach to give probabilistic comparison between an earlier censored data point and a later data point, which can be derived from the Kaplan-Meier curve. This process eventually gives a complete probabilistic ranking of all training data points, regardless of censoring status. This generalizes a censored data prediction problem to a standard regression problem, which allows us to build in generic base-learners.

## 1. Definition of the problem

The Table below represents a typical right-censored data. The first column is the patient ID, the second column the last date that the patient is seen. The third column is the status of the patient at last-seen. If the status is 1, the patient died at that date. If the status is 0, this patient is alive at the last-seen day and this data point is called censored. For each patient, we have a feature vector, from which we are asked to generate a model for predicting death.

| Patient ID | Time (day) | Status |
|---|---|---|
| 10253 | 185 | 0 |
| 102688 | 257 | 1 |
| 101445 | 298 | 1 |
| 10172 | 430 | 0 |
| 101154 | 506 | 1 |

Because not all patients are dead at the last-seen day, generic regression methods cannot be used on this type of censored data.

## 2. Review of the Cox model

The Cox proportional hazard model assumes that covariates are multiplicatively related to the hazard. This assumption allows us to calculate the partial maximum likelihood of the following function (Cox, 1972):

$$\prod_{A:S_A=1} \frac{\exp(X_A\beta')}{\sum_{B:T_B>T_A}\exp(X_B\beta')}$$

$X$ is the feature vector, $\beta$ is the parameters for the variables to be inferred. In the above formula, for all cases where $A$'s status is dead ($S_A$ = 1), we maximize the above log likelihood for all the $B$s, as long as $B$'s observation time is longer than A. However, we cannot consider the cases where $S_A$ = 0. This is because for a censored data point, we do not know whether at a later time $B$ ($t_B$), the patient would be dead or not.

There are two limitations in the Cox model. One is the multiplicative assumption underlying the model. The other is that the early-censored-late-uncensored pairs, and the early-censored-late-censored pairs are not considered in the partial maximum likelihood calculation. In the past decades, Cox model has been extended to include regularization parameters, but there is no fundamental improvement overcoming the above two limitations.

## 3. Review of survival random forest

Survival random forest takes advantage of the splitting behavior of decision trees. A typical way to implement survival random forest is to calculate the log-rank between the splits (Ishwaran *et al.*, 2008). The death rate is defined as the number of death in a time interval *vs.* the number at risk. The deaths during a time interval are compared against all patients whose observation times are longer. Similar to the Cox model, the censored points in a time interval are not tested against the patients that have a longer observation time, because of obvious reasons discussed above. Another limitation of this method is that it is limited to using random forest/decision tree as a base-learner. It cannot be extended to other base learners such as SVM, logistic regression, deep learning, *etc.*, as it is not formulated as a standard regression problem.
As a result, both the Cox-model and survival random forest are doing regression against relative ranks, although it is an incomplete ranking, by dropping out early censored-late censored/noncensored data point pairs.

## 4. GuanRank: Complete ranking of right censored data

Our aim is to assign each training samples with a single value through a uniformed scheme. This single value, will be the target in the regression, where any base learner can be integrated. It is not intuitive how this can be done, because originally we have two target columns: Time and Status, and direct ranking of any or a subset of any is either wrong or incomplete ranking.

However, ranking of a set of samples is not only achievable by sorting a single array of numbers (as there is nothing to be sorted in censored data), but can also be achieved by a thorough pair-wise comparison among all data points. For example, If we have 100 balls, each of a specific size, we could acquire their ranking by ordering their sizes. But, we could also acquire a specific ball's relative ranking, by comparing it against every other ball, and count how many having a size smaller than it. This simple yet interesting property of ranking allows us to give a full ranking of samples in a censored datasets, by assigning the probabilities of one sample ranking ahead of another sample, for the cases where the absolute relative ranking is ambiguous due to censoring.

Now, we are going to break down a right-censored dataset into different situations of pair-wise comparisons. Let us order the patients so that the ones that are going to die early will eventually have a higher target value. Let $t$ be the time of last seen, $S$ be the survival status at last seen: *S=1*: the patient is dead; *S=0, the patient* is alive. A pair of samples (patient A and patient B) may have the following 4 possibilities when $t_A < t_B$:

1) $S_A$ =1, $S_B$ = 1;

2) $S_A$ =1, $S_B$ = 0;
3) $S_A$ = 0, $S_B$ = 1;
4) $S_A$ = 0, $S_B$ = 0;

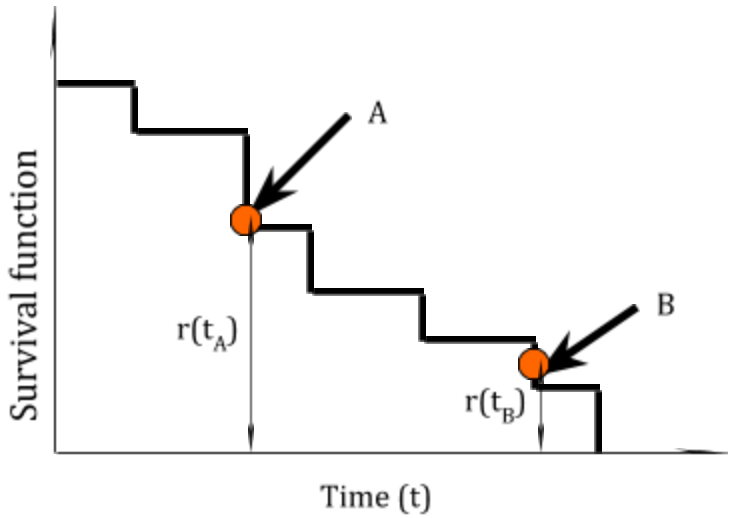
plus two cases when $t_A = t_B$:

5) $S_A = S_B$;

6) $S_A = 1$, $S_B = 0$;

**For Case 1 and Case 2**, we can easily add 1 to the rank of A, and add zero to the rank of B, since it is clear that A died before B. These cases are sufficiently considered in the Cox model and survival random forest, but Case 3 and 4 are not considered in the maximal likelihood function in the Cox model or in the log-rank test in survival random forest.

Now, we need to calculate the K-M curve from the censored data. This curve gives an estimation of the survival function across time, *r(t)*, which is the proportion of the patients that are still alive at time *t* (Figure below).




**For Case 3**, $S_A$ = 0 and $S_B$ = 1, we know that B died at $t_B$. If A dies between $t_A$ and $t_B$, A should rank above B; other wise, B should rank above A. Obviously we cannot determine a binary relative ranking of this pair. However, we can derive a probability of A ranking above B using the K-M curve:

$(r(t_A) - r(t_B))/r(t_A)$, which is the probability that A dies before B, and is the value added to the rank of A.

and

$r(t_B)/r(t_A)$, which is the probability that B dies before A, and is the value added to the rank of B.

**For Case 4**, $S_A$ = 0 and $S_B$ = 0, we first calculate the probability of A dies before the time point B (note, not before B dies, since we do not know when B would actually die):

$p^* = (r(t_A) - r(t_B))/r(t_A)$

Then the probability of A dying after time point B would be:

$1 - p^*$

Since we know that B dies after time point B, without any other information, the probability that B dying before A would be

$(1 - p^*)X0.5$, which is added to the rank of B, i.e. half-to-half chance after reaching time point B.

and the probability of A dying before B is

*p*+(1-p*)X0.5*, which is added to the rank of A, i.e. the chance that A dies before time point B, and half-to-half chance after reaching time point B.

**For Case 5**, where $t_A = t_B$ and $S_A = S_B$, we add 0.5 to the rank of A and B, respectively, because there is no other information telling us A or B would die first.

**For Case 6**, where $t_A = t_B$ and $S_A = 1$, $S_B = 0$, we add 1 to the rank of A and add 0 to the rank of B, because it is clearly A would die before B

Up until this point, we are able to complete all pair-wise comparison between all types of data point pairs in this censored data, and thus able to acquire a relative ranking of how likely to die for all individuals in the training set, this is our final target that we could build into any base learner, be it random forest, linear regression, neural network, or gaussian process regression, etc. In the final model, as a summary of the above different cases, the rank of patient A is given by:

If $S_A = 1$,

$$\sum_{\forall B: t_B > t_A} 1 + \sum_{\forall B: t_B \le t_A, S_B=0} \frac{r(t_A)}{r(t_B)} + \sum_{\forall B: t_B = t_A, S_B=1} 0.5$$

If $S_A = 0$,

$$\sum_{\forall B: t_B \ge t_A, S_B=0} (1 - \frac{0.5r(t_B)}{r(t_A)}) + \sum_{\forall B: t_B \ge t_A, S_B=1} (1 - \frac{r(t_B)}{r(t_A)}) + \sum_{\forall B: t_B < t_A, S=0} \frac{0.5r(t_A)}{r(t_B)}$$

**Discussion and future work:**

The future work is mainly to demonstrate its performance versus Cox-related or SRF related models, which is pretty straightforward: I will just directly plug it into different benchmark studies involving to survival models. This work can be adapted to left or interval-censored data, but, for me, there is no application.

**Acknowledgement:**

All models have to go through intensive real world tests to let us understand their application domains and limitations. Thus I thank the following people who re-implemented and/or tested my model on other survival datasets (including successful applications and cases where both Cox and Guan-rank failed): Matthias Kretzler, Laura Mariani, Kayvan Najarian, Zhi Li, Brahmajee Nallamothu, Hongjiu Zhang, Mi Yang, Ljubomir Buturovic, Michael Mason and Zhengnan Huang.

I thank DREAM for providing unbiased validations to the Guan-rank model and a dozen of my other models.

**References:**

1. Cox, David R (1972). "Regression Models and Life-Tables". *Journal of the Royal Statistical Society, Series B* **34** (2): 187–220. JSTOR 2985181. MR 0341758
2. Ishwaran et al. (2008) The Annals of Applied Statistics, Vol. 2, No. 3, 841–860 DOI: 10.1214/08-AOAS169