A General Framework for Censored Regression Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

For various reasons, datasets used in many prediction tasks are right-censored. The classic example is predicting patient survival under a treatment when survival time is not known for patients who are still alive at the end of the study or who have left the study. Only a lower bound on survival, the last time of study contact, is known for these patients. This censoring issue has been studied with vigor in biostatistics, giving rise to the entire field of survival analysis, but it has been less examined in machine learning. In this study we propose a general framework for censored regression prediction using an expectation maximization (EM) procedure. In previous work, the Buckley-James method has utilized a related EM procedure using linear regression with and without regularization. Unfortunately such a restricted class of models does not lend itself to the power of the rich regression algorithms available in machine learning such as regression and model trees, support vector regression, etc. Our simulations studies show that datasets generated from non-linear models can be more accurately predicted using an appropriate model with our proposed EM method compared to the Buckley-James method. A real world cancer dataset also shows the benefits of using alternative algorithms. We believe the general framework for censored prediction described in this study can expand researchers' tool sets to better predict outcomes with censored data.

1 Introduction

The machine learning community is accustomed to analyzing data with missing values. Many algorithms exist to deal with missing values such as Gibbs sampling [1] and expectation maximization [2]. A special kind of missing data occurs when the continuous-valued prediction outcome is censored, where the observed outcome is shorter than the true outcome. Examples include reliability testing and maximum measurements. In reliability testing only a certain number of products are observed to fail given the maximum total observation time, and the products that have not failed at the maximum observation time are considered censored. Maximum measurements can occur in instruments with maximum sensing capabilities in astronomy or weather detection. This type of missing data is particularly important in the biostatistics community and has been studied extensively.

In the medical arena, while classification is widely studied, many clinical outcomes are continuous-valued. Many binary or discretized outcomes are derived from raw continuous-values. Often, these continuous outcomes are related to *time-to-event* outcomes such as survival, time to hospital readmission, or time to disease recurrence. Yet, given the nature of medical data collection, much of this data is right-censored. Right censoring in this domain occurs when the true time-to-event, T, is greater than the observed time-to-event, Y. For example, if the time-to-event is survival (time-to-death), a patient may still be living at the end of the study or move away from the study location,

becoming lost to follow up. Therefore, we can only observe Y, and the patient's true survival time T is greater than Y.

Common naive methods of dealing with censored outcomes are to discard the censored data points or use them only in training. The former case throws away data and may actually be discarding the longer surviving patients if censoring is more likely the longer a patient lives. The latter case likely under-predicts the test set since censored values are considered true values in training.

More sophisticated methods have involved the Cox proportional hazards models [3] with and without regularization. In this framework, a Cox proportional hazards model is created using all the data. For censored subjects, a survival curve is estimated from the Cox proportional hazards model. Then the median survival time, for example, is chosen as the true outcome for this patient. Extensions to the Cox model that can be used include lasso [4] and kernel Cox proportional hazards models [5].

Expectation maximization procedures [2] have also been used in linear regression. Strictly speaking, an accelerated failure time (AFT) model is a linear regression model with a log outcome. Methods of AFT have been adapted with the widely used Buckley-James method [6] when specific subjects are right censored. The Buckley-James method uses the AFT with an expectation maximization procedure to update the expected outcome for censored subjects. Extensions to this include regularization with L_1 (i.e. lasso [7, 8, 9]) or L_1 with L_2 (i.e. elastic-net [10]).

Unfortunately, the Cox model and Buckley-James methods with and without regularization can be used with only specific modeling approaches. Hence, they severely restrict the models that can be used for predictions. Here we propose a methodology to perform regression with right censored data using *any* regression model with potentially *any* loss function.

2 Statistical Methods

In this section, we use the terminology of survival time regression, but the approach can be applied to any time-to-event regression task. We use the following notation to denote subject i. T_i is the true survival time if there were no censoring. However, we can only observe y_i for censored patients where $T_i > y_i$. With this notation, $T_i = y_i$ for uncensored patients. Therefore, for any model $f(x)$, $T_i = f(x_i) + \epsilon_i$, we wish to minimize $\sum_i \ell(T_i; f(x_i))$ where $\ell(T_i; f(x_i))$ is our loss function. We do not assume any distribution on the error term, ϵ_i, only that it has zero mean and finite variance. We denote the set of censored subjects as C and uncensored subjects as $\bar{C}$. From this notation we first show the expectation step of this EM procedure.

We can write the expectation as follows:

$$min \left(\sum_{i \in C} E(\ell(T_i, f(x_i)|y_i, f(x_i)) + \sum_{i \in \bar{C}} \ell(T_i, f(x_i)) \right) \tag{1}$$

For uncensored patients, we do not need to find the expected value because $T_i = y_i$. Thus we concentrate on the summation term in Equation 1 for censored subjects:

$$\sum_{i \in C} E(\ell(T_i, f(x_i)|y_i, f(x_i)) \tag{2}$$

We next must find the expectation for the following.

$$\sum_{i \in C} E(\ell(T_i, f(x_i)|y_i, f(x_i)) = \sum_{i \in C} E(\ell(T_i, f(x_i)|T_i > y_i, y_i, f(x_i)) \tag{3}$$

To obtain the expectation, we must find the probability distribution function

$$\begin{aligned} P(T_i|T_i > y_i, y_i, f(x_i)) &= P(T_i - f(x_i)|T_i - f(x_i) > y_i - f(x_i), y_i, f(x_i)) & (4) \\ &= P(\zeta_i|\zeta_i > y_i - f(x_i), y_i, f(x_i)) & (5) \end{aligned}$$

where ζ_i is the prediction error. Since y_i and $f(x_i)$ can be considered constants, c_i, we are interested in

$$P(\zeta_i|\zeta_i > y_i - f(x_i), c_i) \tag{6}$$

10 yr	22* yr	25 yr	27* yr	40 yr
1/5	1/5	1/5	1/5	1/5
1/5	1/5	1/5	0	2/5
1/5	0	1/5+1/15	0	2/5+2/15

Table 1: Each row corresponds to a redistribution of censored subject weight to find the probability of survival. Starred numbers in the first row correspond to censored subjects.

We can calculate $P(\zeta)$ from all the data using a Kaplan-Meier (KM) curve [11], the most common survival function estimator. Once we have $P(\zeta)$ using the KM curve from all the data, we can obtain $P(\zeta_i|\zeta_i > y_i - f(x_i), c_i)$.

We can create $P(\zeta)$ via the alternative representation of the KM curve using the *redistribution to the right* algorithm [12]. All censored and uncensored observations are ordered in increasing order. Each individual (both $C, \bar{C}$) is given equal mass. Starting from the longest censored observation, the mass of each censored observation is moved to each uncensored observation on the right in proportion to the uncensored observations' current weight. The survival curve from the final set of probabilities is the KM curve. An example of this method is shown in Table 1.

To obtain the expected observation for the next iteration, $j+1$, we can choose $\zeta_{i,j+1}$ by finding $E(\zeta_i|\zeta_i > y_i - f(x_i), c_i)$. With $\zeta_{i,j+1}$, we can then recover $T_{i,j+1}$. This is a "hard" EM procedure [13]. We could also use $P(\zeta_i|\zeta_i > y_i - f(x_i), c_i)$ for soft EM.

The hard censored EM procedure will work with any model that can predict a real value outcome. With soft EM, the algorithm will have to be able to weight the error of examples accordingly, which is easily integrated into many regression models. The soft EM may be especially useful when the loss function $\ell(T_i, f(x_i)|y_i, f(x_i)$ is not squared error, preventing a simple derivative when calculating the expectation.

Now for m iterations the expectation is analogous to Equation 1.

$$min\left(\sum_{i\in C}\sum_{j=1}^{m} W_{ij}\ell(T_{ij}, f(x_i)|y_i, f(x_i)) + \sum_{i\in\bar{C}}\ell(T_i, f(x_i))\right) \tag{7}$$

The maximization step is performed as standard where a new model $f(x)$ is predicted with the imputed outcome values from the censored patients and the true outcome values for uncensored patients.

3 Simulation Studies

We simulate five datasets, one linear and four non-linear, with 100 examples in the training set and 100 examples in the test set. Features for each example i, $\mathbf{X}_i$, are generated from a multivariate normal distribution with $\mu = 0, \sigma^2 = 1, \Sigma = \mathbf{1}$.

For a given $\mathbf{X}_i$, the output T_i is generated by applying the target function. To ensure that all $T_i > 0$, a constant term $1.5*|\min(T)|$ is added to each T_i. The censoring time C is generated from a uniform distribution $U(-\tau, 0)$ where τ is chosen to be $-T_i$. We assume if censored, censoring can occur at any time after enrollment begins. The observed survival time is $Y = T + C$.

The percentage of examples censored ranged from 5% to 85% in increments of 5%. At each censoring rate, simulated data and model fitting was performed 100 times to obtain a 95% confidence interval.

We use the relative prediction error (RPE) from the independent test data set where $RPE \approx (1/n)\sum_{i=1}^{n}(T_i - P_i)^2/\sigma^2$. P_i is the model's predicted score and σ^2 is the variance of the target function.

3.1 Simulated Datasets

Dataset 1 is linear while datasets 2-4 are non-linear. For all T_i simulated, an error term of $\sigma\epsilon$ is added, where $\epsilon \sim N(0,1), \sigma = 8$.

Dataset 1 (LM) is a linear model generated from 10 features and coefficients β are chosen from $N(3, 0.5)$.

Dataset 2 (Quad) is a quadratic function generated from 4 features with the equation $T = f(\bar{X}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1^2 + \beta_6 x_2 x_3 + \beta_7 x_3 x_4$.

Dataset 3 (Cubic) is a cubic polynomial function generated from 10 features with the equation $T = f(\bar{X}) = \beta_1 x_1 x_2 x_3 + \beta_2 x_4^2 x_5 + \beta_3 x_6 x_7^2 + \beta_4 x_8 x_9 x_{10}$.

Dataset 4 (RT) is a regression tree generated from 4 features. There are 4 leaves where T_i are generated from $N(0,1)$, $N(25,1)$, $N(50,1)$, $N(75,1)$, and $N(100,1)$.

Dataset 5 (MT) is simulated from a model tree, a plausible biological scenario. At each of the leaves, there is a linear regression function for a time-to-event outcome. We believe such models are common in medical domains when unknown disease subtypes constitute a heterogeneous disease definition. Such disease examples include diffuse large B-cell lymphoma [14], asthma exacerbation frequency [15], ovarian cancer [16] and Alzheimer's disease [17]. However, in many cases we do not know what these subtypes are and must create a regression model using all the data. In this simulated data we use 4 features with 5 leaves where T_i are generated from $\beta_1^T X$, $25 + \beta_2^T X$, $50 + \beta_3^T X$, $75 + \beta_4^T X$, and $100 + \beta_5^T X$. β_i are chosen from $N(3, 0.5)$.

3.2 Prediction Models

For each dataset, the prediction models used include: the mean observation in the training set (Mean-Obs), the linear model using the Buckley-James method (BJ), the "appropriate" model using the our censored subject EM method (Our method), and the "appropriate" model using the complete true non-censored training set (M-Comp). The appropriate model used for LM is a linear model, for dataset 2 (Quad) is a quadratic model chosen with the forward Aikake Information Criteria model selection where quadratic terms were included in the scope, for dataset 3 (Cubic) is a cubic kernel SVR, for dataset 4 (DT) is a decision tree, and for dataset 5 (MT) is a model tree. Code from R packages Design, kernlab, rpart, and RWeka is used for modeling.

3.3 Quadratic Convexity

Using the same conditions as explained for the previous simulation studies, we study the difference in error of models as the convexity of a quadratic function increases. We examine the models for the Buckley James method and a quadratic model chosen with the forward Aikake Information Criteria model selection where linear and quadratic terms were included in the scope. We use the latter model with the censored training set (Quad our method) and the complete true non-censored training set (Quad-Comp). We simulate data at a 35% censoring rate as previously described with one feature and the equation $T = f(\bar{X}) = \beta_1 x_1^2$ where β_1 is $[0.5, 10]$ with a 0.5 step size.

3.4 Alternative Loss Functions

The framework presented here does not necessitate a specific model nor a quadratic loss function as used in linear regression for the Buckley-James method. The Buckley-James method proposes to use $E(T_i)$ at each iteration. Our framework is related to this method when the quadratic loss function is used. However, when a different loss function is used to obtain $E(T_i)$ at each iteration, our general framework can accommodate this case.

We use an ϵ-insensitive loss function as an example motivated from a study predicting the optimal starting dose of warfarin [18], a blood thinning drug difficult to dose. A question of interest was the percentage of patients that were within $p\%$ of the true dose. To simulate such an example, we use dataset 3 with the traditional regression tree (RT) and a modified regression tree (RT-ϵ-ins). In regression trees for continuous-values, the value to split a node is chosen where the children leaves account for the highest percentage of variance compared to the variance in parent node. We modify

this procedure by calculating $\epsilon -_{ins,parent} - \sum_i \epsilon_{ins,child_i}$, the ϵ-insensitive error of the parent node minus the sum of the ϵ-insensitive error from all children nodes. The ϵ-insensitive region is $0.15y|$. Model are evaluated based on the mean absolute ϵ-insensitivity.

4 Real world data

Data from a squamous cell lung carcinoma study to predict the survival of early-stage lung cancer patients using microarray gene expression were analyzed. The study included 129 squamous cell lung carcinoma patients with log-transformed mRNA from the Affymetrix U133A microarray and a 48% censoring rate. Data were downloaded from the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) with accession number GSE4573.

Ten fold cross validation was conducted to obtain a test fold prediction value for each example. For each fold, predictors from the training set were chosen using the R Bioconductor package 'genefilter' using a Cox proportional hazards model to filter out features with $p > 0.002$. Subjects were assigned to high and low risk group using a cutoff prediction time of 3 years, a common threshold used in squamous cell lung carcinoma [19].The Buckley-James Method was compared to a linear kernel SVR using our method.

5 Results

5.1 Simulation Studies

Simulation results for the linear model are summarized in Figure 1. The Buckley-James method's performance is similar to that of our method using an EM procedure with a linear model. The censoring rate has little effect on linear model accuracy until a 65% censoring rate.

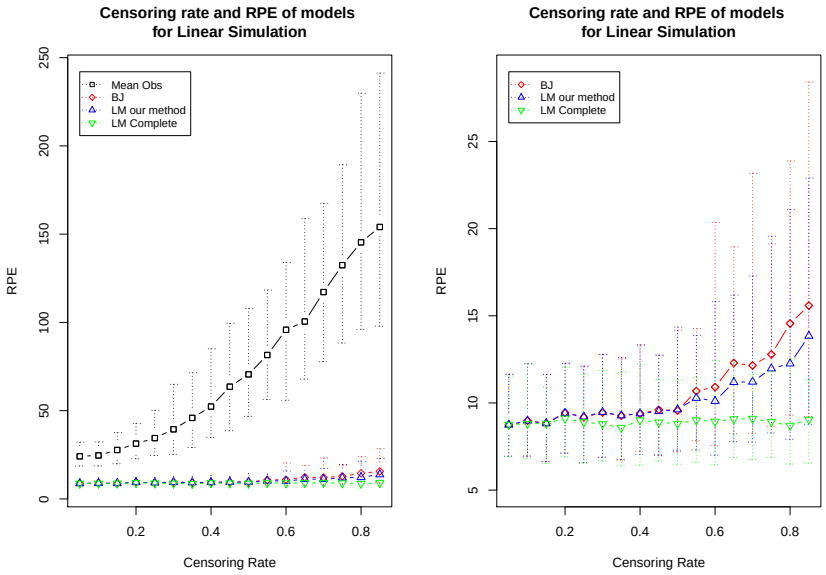

Figure 1: Simulation and prediction using a linear model. Models tested included the average observation in the training set (Mean-Obs), linear model using Buckley James method (BJ), a linear model using our method (LM our method), and the linear model using the complete true uncensored training data (LM Complete). The graph to the right is a magnified view of BJ, LM our method, and LM Complete

Results for non-linear data simulations are shown in Figure 2. The Buckley-James methods always has higher RPE on average than the "appropriate" model. The "appropriate" model track fairly well

the same model using the complete true uncensored training set until a 40-50% censoring rate. After that, the error begins to rise to the level of the Buckley-James method.

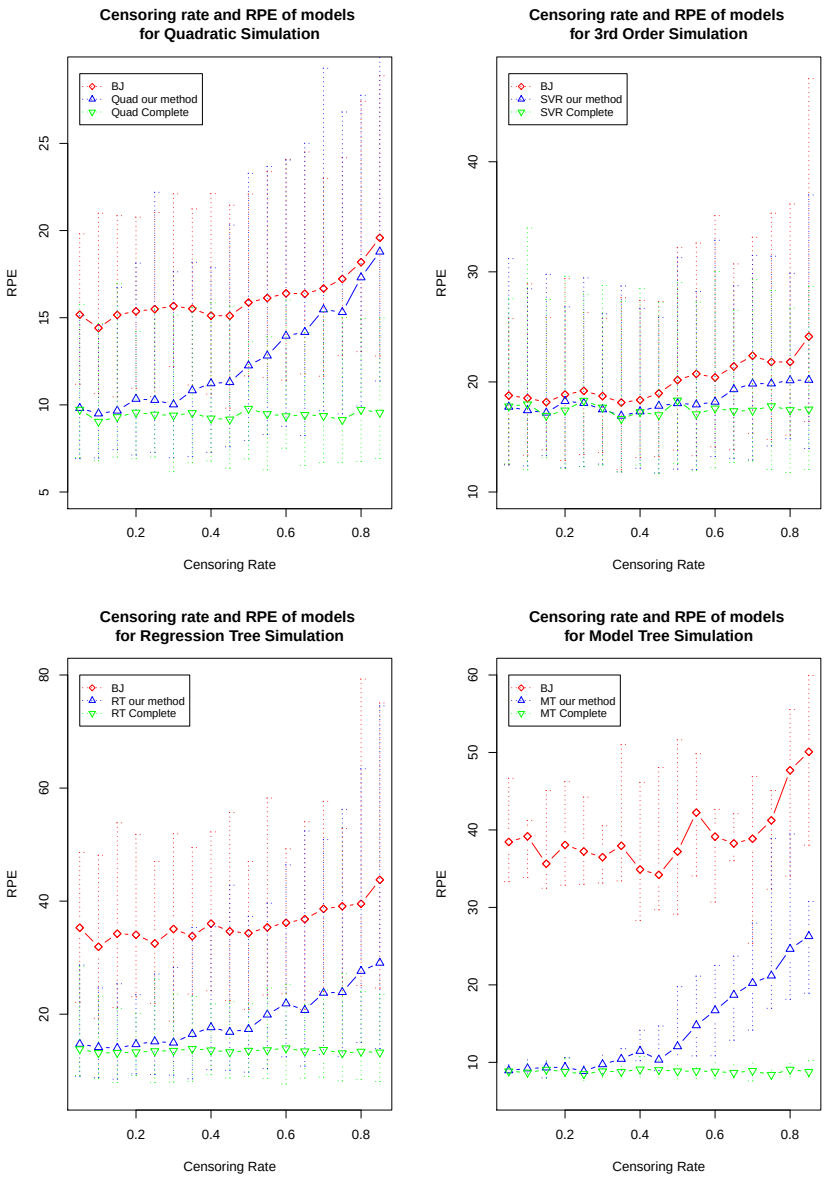

Figure 2: Simulation and prediction model using four non-linear functions. Models tested included the Buckley James method (BJ), the "appropriate" model using our method, and the "appropriate" model using the complete true uncensored training data.

5.2 Quadratic Convexity

Figure 3 shows the relationship of quadratic convexity and error for the Buckley James method and quadratic models. The error of the Buckley James method increases in a quadratic fashion as expected as the convexity of the simulated data increases. The quadratic models using training data with a 35% censoring rate or no censoring remain similar at a low error. Similar results were seen at 10% and 65% censoring rates.

Model	$\mathbf{MAE}_{\epsilon-\text{insensitive}}$
Mean-Obs	11.71 (7.95,15.94)
BJ	4.59 (2.27,7.13)
RT our method	1.72 (0.61,4.34)
RT-ϵ our method	1.67 (0.65, 4.28)

Table 2: Simulation study of mean absolute ϵ-insensitive error where $\epsilon = 0.15y$. Dataset 3 was used. The prediction models used as shown in the first column were the average observation in the training set (Mean-Obs), linear model using Buckley James method (BJ), standard regression tree using our method (RT our method), and modified ϵ-insensitive regression tree using our method (RT-ϵ our method).

5.3 Alternative Loss Functions

Results using the modified regression tree are shown in Table 2. Using the mean absolute ϵ-insensitive error as the comparator, the modification of the loss function via node splitting in regression trees decreased the error compared to Buckley-James and the traditional regression tree.

5.4 Real Data

The number of training set predictors ranged from 16-45 predictors using the Cox proportional hazards model to filter out predictors where $p > 0.002$. Kaplan-Meier survival curves are shown in Figure 4 for high and low risk patients comparing the Buckley-James method and linear kernel SVR with our method. Using a logrank test [20], the common statistical test between two survival estimators, between high and low risk groups, the p-value for Buckley-James was 0.121 and SVR-linear kernel was 0.039.

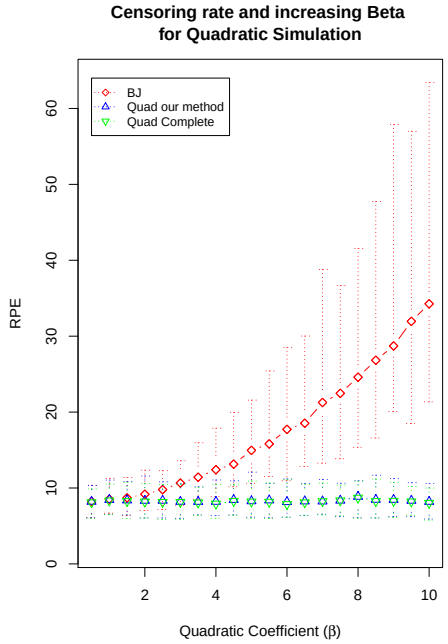

Figure 3: Relationship of quadratic convexity and error of the Buckley James method (BJ) and a quadratic model with censored data (Quad our method) or with complete training data (Quad-Comp).

6 Conclusion

The main contribution of this study is the ability to use *any* regression model and any loss function with censored data to improve model regression testing performance. The EM procedure to update the negative error associated with censored patients using the Kaplan-Meier survival curve is simple to implement.

This study has not compared the general framework EM procedure proposed here with procedures that do not use a related EM procedure. Other models that have been specifically designed or algorithmically modified to more carefully handle censored data include the previously mentioned Cox proportional hazards model and support vector regression for censored data [21]. With this method, we are no longer restricted to a select number models that have to be specifically and algorithmically altered to properly account for censored data.

The asymptotic properties of the Buckley-James method have not been proven, though it is commonly used in practice. However, closely related methods, a modified Buckley-James method [22] and a rank based estimator [23], have shown consistency and asymptotic normality. In this study, we do not attempt to prove convergence of our EM method, but this is an interesting direction for further work.

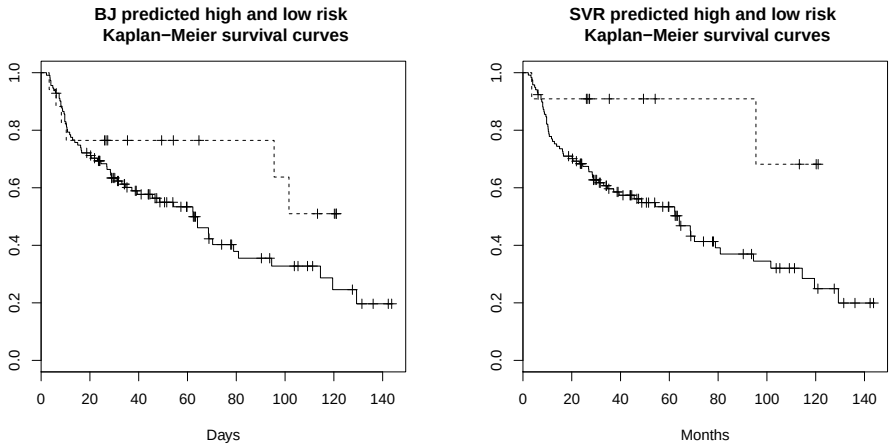

Figure 4: Kaplan-Meier curves for high and low risk patients using the Buckley-James method and a linear kernel SVR with our method. Predicted values include all patients from ten fold cross validation test sets where 3 year survival was used to determine high (solid) and low risk (dotted) patients. Crosshatches in each curve are the time of censored patients. Using a log-rank test, the p-value for Buckley James was 0.121 and linear kernel SVR was 0.038

With the growing application of machine learning to medical data, censored datasets will be of increased significance for the machine learning community. As described in the introduction, censored datasets are not only restricted to medical datasets, but also to other datasets where machine learning is applied. We believe the general framework for censored prediction described in this study help prevent discarding censored data, expanding researchers' tool sets to better predict time-to-event outcomes and identify current and novel interventions to improve patient care.

References

[1] A. E. Gelfand. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398409, 1990.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):138, 1977.

[3] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[4] J. J. Goeman. L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, 52(1):70–84, 2010.

[5] H. Li and Y. Luan. Kernel cox regression models for linking gene expression profiles to censored survival data. In *Pacific symposium on biocomputing*, volume 8, pages 65–76, 2003.

[6] J. Buckley and I. James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

[7] J. Huang, S. Ma, and H. Xie. Regularized estimation in the accelerated failure time model with High-Dimensional covariates. *Biometrics*, 62(3):813–820, 2006.

[8] T. Cai, J. Huang, and L. Tian. Regularized estimation for the accelerated failure time model. *Biometrics*, 65(2):394–404, 2009.

[9] B. A. Johnson. On lasso for censored data. *Electronic Journal of Statistics*, 3:485–506, 2009.

[10] S. Wang, B. Nan, J. Zhu, and D. G. Beer. Doubly penalized BuckleyJames method for survival data with High-Dimensional covariates. *Biometrics*, 64(1):132–140, 2008.

[11] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457481, 1958.

[12] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, December 2010.

[13] M. J. Kearns, M. Yishay, and A. Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Learning in graphical models*, pages 495–520. MIT Press, Cambridge, MA, USA, 1999.

[14] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. F. Aliferis, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, February 2000.

[15] W. C. Moore, D. A. Meyers, S. E. Wenzel, W. G. Teague, H. Li, X. Li, R. D'Agostino Jr, M. Castro, D. Curran-Everett, and A. M. Fitzpatrick. Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *American journal of respiratory and critical care medicine*, 181(4):315–323, 2010.

[16] D. S. Miller, J. A. Blessing, C. N. Krasner, R. S. Mannel, P. Hanjani, M. L. Pearl, S. E. Waggoner, and C. H. Boardman. Phase II evaluation of pemetrexed in the treatment of recurrent or persistent platinum-resistant ovarian or primary peritoneal carcinoma: a study of the gynecologic oncology group. *Journal of Clinical Oncology*, 27(16):2686–2691, 2009.

[17] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen. Mild cognitive impairment: clinical characterization and outcome. *Archives of neurology*, 56(3):303–308, 1999.

[18] M. T. Lee, Y. T. Chen, M. S. Wen, Y. Caraco, I. Achache, S. Blotnick, M. Muszkat, J. G. Shin, H. S. Kim, G. Suarez-Kurtz, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med*, 360(8):753–64, 2009.

[19] M. Raponi, Y. Zhang, J. Yu, G. Chen, G. Lee, J. M. G. Taylor, J. Macdonald, D. Thomas, C. Moskaluk, Y. Wang, and D. G. Beer. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Research*, 66(15):7466–7472, August 2006. PMID: 16885343.

[20] R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135(2):185207, 1972.

[21] F. M. Khan and F. M. Zubek. Support vector regression for censored data (svrc): A novel tool for survival analysis. In *Eighth IEEE International Conference on Data Mining*, pages 863–868, 2008.

[22] Y. Ritov. Estimation in a linear regression model with censored data. *The Annals of Statistics*, 18(1):303–328, March 1990.

[23] A. A. Tsiatis. Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, 18(1):354–372, March 1990.