Piecewise-constant parametric approximations for survival learning

Jeremy C. Weiss JEREMYWEISS@CMU.EDU
Carnegie Mellon University, Heinz College
Pittsburgh, PA 15213 USA

Abstract

Logged events occur both regularly and irregularly over time. In electronic health records, these events represent mixtures of scheduled and urgent or emergent encounters. Whereas most survival models use baseline events to estimate the rate function for an outcome, *e.g.*, Cox processes using the proportional-hazards assumption, our framework uses logged events over time to predict survival outcomes with piecewise approximations of arbitrary hazard functions. We develop a procedure to learn forests as combinations of piecewise-constant and parameterized distributions to compactly model survival distributions from data. Under this construction, the model provides a "now-time" risk that incorporates irregularly-repeated data and for health outcomes serves as a surrogate for patient disposition. We illustrate the advantages of our method in simulations and in longitudinal, intensive care unit data of individuals with diabetes admitted for ketoacidosis.

1. Introduction

Electronic health records (EHRs) have become the primary tool for storing and using patient encounter data with over 80 percent adoption across hospitals and clinics in the United States. Leveraging the EHR for clinical insight remains a challenge, with existing systems running into adoption challenges, such as clinical decision support alert fatigue, documentation overload, and diminished quality of patient-physician interactions. The machine learning community has identified the EHR as a potential resource for improving care through evidence generation, with applications ranging from risk scores and risk factor quantification and imaging classification. Despite these advances, the temporal nature of EHR data remains a substantial challenge.

Recognition of non-ignorable patient heterogeneity has led to research moving beyond classical frameworks of logistic regression and the temporal analogue Cox proportional hazards model. The presence of heterogeneity has thwarted treatment advancements, *e.g.*, leading sepsis to be considered "the graveyard for pharmaceutical companies" (Riedemann et al., 2003; Prescott et al., 2016). Advancements in heterogeneity detection and modeling while also considering the time-to-event formulation of survival analysis are of substantial interest for medical practitioners.

Our work addresses heterogeneity over time by using machine learning to learn the hazard used in survival analysis without adopting common assumptions including proportional hazards or accelerated failure times. Instead, we build upon the continuous-time point process framework, which models rates of event occurrences in event networks. One leading approach is the use of forests to model a piecewise-constant hazard (Weiss and Page, 2013). In principle, these forests enable learning approximations to arbitrary hazard functions. In practice, however, they can require large amounts of data to learn relatively simple survival distributions. This work develops a flexible framework to introduce parametric distributions into forest learning, exemplified by the integration of the log logistic distribution commonly used to capture both the timing of event occurrences and the uncertainty of said timing.

The learning procedure is outlined as follows. Given an initial piecewise-constant hazard function, a new hazard is proposed, given by the product of the piecewise-constant hazard and a parameter-estimated log logistic hazard. If the new hazard is able to better model the time-to-event distribution, subject to a complexity penalty, the new hazard is adopted. Then, after each log logistic parameter estimation step, the learned distribution is folded into the piecewise-constant approximation, so that the next iteration of learning has the same form. The procedure is illustrated in Figure 1. We illustrate the model's flexibility in both identifying regular and random time-to-event patterns in simulations and compare its performance in EHR data.

The medical task we address is the prediction of discharge of patients with diabetic ketoacidosis (DKA) from the intensive care unit. Ketoacidosis is a life-threatening state, usually occurring in patients with diabetes, where ketones build up in the blood with concomitant glucose and electrolyte abnormalities secondary to endocrine dysregulation or impairment. Treatment of diabetic ketoacidosis involves an intensive policy of monitoring and therapy including electrolyte management and intravenous insulin. A physician conducting chart review typically can identify resolution of the ketoacidotic state, which suggests that the EHR contains many of the indicators for resolution.

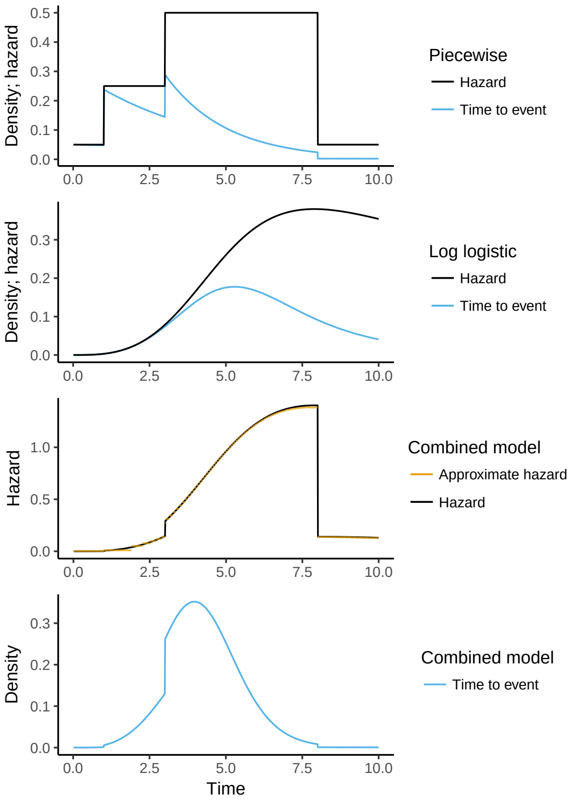

Figure 1: A piecewise-constant hazard distribution (first panel) characterizes the time-to-event distribution. A proposed, height-parameterized log logistic distribution (second panel) is proposed for multiplicative introduction into the model, resulting in a combined hazard which is approximated to produce a piecewise-constant hazard (third panel). The resulting time-to-event distribution is shown (fourth panel).

Nevertheless, there is substantial uncertainty in the resolution of ketoacidosis, illustrated by the distribution in intensive care unit length of stay–one measurable surrogate of resolution time. We posit there is substantial heterogeneity in resolution due to underlying factors and use point processes to examine intermediate indicators of resolution. Like sepsis, the treatment policy for DKA is one-size-fits-all. Therefore, better characterization of DKA resolution could lead to characterization of individuals who are high risk, high cost, or who might benefit from targeted therapy.

Regarding algorithm interpretability, point processes enable "now-time" risk prediction from medical histories, which for prediction of discharge is a strong surrogate of "is the patient ready for discharge"? Alternatively, like Cox processes, they can be learned from baseline variables or using a gap time τ to provide a forecast of when the patient will be discharged.

We demonstrate point processes in this framework. In Section 2 we review point processes in the survival analysis framework. In Section 3 we demonstrate the process of parametric piecewise-approximation. In Section 4 we describe the medical task and experimental setup. In Section 5 we demonstrate the advantages of point process parametric approximation, and illustrate the method in simulations and in characterization of the length of ICU stay in patients with diabetic ketoacidosis.

1.1 Related work

There has been extensive work developed on survival analysis in health care. Generally this work builds on frameworks reviewed in Kalbfleisch and Prentice (2011) and Aalen et al. (2008) for using time as a key factor in characterizing variable relationships for attribution and prediction. The leading model has been the Cox regression model, which uses a separation of time modeling and covariate modeling useful for risk attribution questions, but requires making the proportional hazards assumption. With this framework, time-varying Cox models account for updated or changing baseline covariates, *e.g.*, as in Therneau et al. (2017). However, even with time-varying coefficients or repeated variable measurements, the proportionality assumption is often inappropriate in the presence of heterogeneity. Furthermore, the separation of the nuisance function from covariate adjustments limits modeling of predictable time-to-events.

Related work in point processes networks has used Gaussian process priors (Lian et al., 2015; Lasko, 2014; Saul et al., 2016), piecewise constant forests (Weiss et al., 2012; Weiss and Page, 2013), and Gaussian mixture models (Goulding et al., 2016) to learn point process rates. The first two approaches are limited in their ability to efficiently model parametric survival distributions often observed in data, with the first requiring matrix inversion or variational compression and the second requiring large data sizes to capture smooth curves. The third set of methods require assumptions about the granularity of time and positioning of the mixture components. Random survival forests (RSFs) (Ishwaran et al., 2008) also adopt a forest framework where the focus is to address estimate the cumulative hazard function through bootstrapping of the samples. However RSFs neither integrate parametric models of time-to-event distributions nor utilize data beyond baseline. Other related work includes Poisson generalized additive models, *e.g.*, Tutz and Binder (2006); Lawless (1987), which model counts within time windows using a linear combination of underlying functions. These models focus on the count distribution rather than modeling the events over time.

Within machine learning, popular frameworks for time include Gaussian processes and deep learning frameworks, *e.g.*, Ghassemi et al. (2015); Lipton et al. (2015); Razavian et al. (2016). Gaussian processes use time as a dimension but, without extension, do not consider interarrival times nor rate forecasting in its formulation. Meanwhile, deep learning frameworks, *e.g.* recurrent neural networks and long short-term memory networks, formulate the model as time series with fixed bins and often introduce an artificial missing data problem.

2. Background

We first state the time-to-event distribution, survival distribution, and hazard function relationships. Let time t represent the time to an event and $f(t)$ the distribution of the time to event. Let $F(t)$ be the cumulative distribution function of $f(t)$, i.e. $F(t) = \int_{-\infty}^{t} f(\tau)d\tau$. Then the survival distribution is given by $S(t) = 1 - F(t)$. The hazard function $\lambda(t)$ is defined as the time distribution conditional on the event having not yet occurred: $\lambda(t) = \lim_{h \to 0^+} p(t \leq T < t + h | T \geq t)/h = f(t)/S(t)$.

Next we define the data and model representation. Given a finite set of event types $l \in \mathcal{L}$, an instance is an event sequence or trajectory x specified by an ordered set of $\{$time, event$\}$ pairs (t_i, l_i) for $i = 1$ to n. A history h_i is the subset of x whose times are less than t_i. Let l_0 denote the null event type, and append the null event pairs (l_0, t_0) and (l_0, t_{end}) to the beginning and end of x. Let Θ represent the model providing the specification of the hazard function for each event type l; then the likelihood of the instance is:

$$p(x|\Theta) \propto \prod_{l\in\mathcal{L}}\prod_{i=1}^{n} \lambda_l(t_i|h_i)^{\mathbb{1}(l=l_i)} e^{-\int_{-\infty}^{t_i}\lambda_l(\tau|h_i)d\tau}$$

Piecewise-constant intensity models (PCIMs) assume that the hazard functions are constant over intervals. Let S be a discrete set of states $S = \{S_l \mid l \in \mathcal{L}\}$. Let $\Lambda = \{\Lambda_{ls} \mid l \in \mathcal{L}, s \in S_l\}$ be the set of mappings from event and state to rate parameters for each event l and state s. Let $\Sigma = \{\Sigma_l \mid l \in \mathcal{L}\}$ be the set of functions that map from a time and history to $s \in S$: $\Sigma = \{\Sigma_l \mid \Sigma_l : (t, h) \rightarrow S\}$. Then, the model is specified by $\Theta = \{S, \Lambda, \Sigma\}$. Under this construction, the piecewise-constant assumption is invoked: $\lambda_l(t_i|h_i) = \lambda_{ls}$. The PCIM likelihood simplifies to:

$$p(x|\Theta) \propto \prod_{l\in\mathcal{L}}\prod_{s\in S_l} \lambda_{ls}^{M_{ls}(x)} e^{-\lambda_{ls}T_{ls}(x)}, \tag{1}$$

where $M_{ls}(x)$ is the count of events of type l while s is active in trajectory x, and $T_{ls}(x)$ is the total duration that s, for event type l, is active.

2.1 Specification of Σ and Λ with multiplicative hazard regression forests

Previous work on PCIMs uses trees and forests to define the mapping. Following Gunawardana et al. (2011), we can specify Σ_l and Λ_l represent a regression tree of the form $\Lambda_l \circ \Sigma_l$: Σ_l maps the history and time to $s \in S_l$; then Λ_l maps s to hazard λ_{ls}. To specify the form of Σ_l, let $\mathcal{B}_l$ be the set of basis state functions $f(t, x)$ that maps to a basis state set S_f. As in Weiss et al. (2012), the basis functions can be viewed as set partitions of the space over $S = S_{l_1} \times S_{l_2} \times \ldots S_{l_{|\mathcal{L}|}}$. Each interior node in the regression tree is specified by a basis function f. Each leaf holds a non-negative real value: the hazard. Thus one path ρ through the regression tree for event type l corresponds to a recursive subpartition resulting in a set S_ρ, and every $(l, s) \in S_\rho$ corresponds to leaf hazard $\lambda_{l\rho}$, i.e., we set $\lambda_{ls} = \lambda_{l\rho}$.

Forest point processes multiply leaf hazards from multiple trees. Given that each tree represents a partition, the intersection of trees, *i.e.* a forest, forms a finer partition. Then, a subpartition is given by the intersection $S_\rho = \bigcap_{j=1}^{k} S_{\rho,j}$, the intersection of sets of active paths through trees 1 to k. The hazard $\lambda_{l\rho}$ is given by the product of leaf hazards. For constant leaf hazards, maximum likelihood (or MAP) estimates for a split on ρ_j can be learned and are given by the number of observed events divided by the predicted number of events: $M_{l\rho_j}/\hat{M}_{l\rho_j}$ (Weiss et al., 2012).

3. Piecewise parametric survival distributions

Consider the situation where are interested in the likelihood function for one target event $l \in \mathcal{L}$ and we can remove outside product from Equation 1 and the subscripts l. Our intention is to modify the existing piecewise-constant hazard λ_s to $\lambda_s\lambda_p$, where λ_p is given by the log logistic distribution with scale α and shape β parameters:

$$f(t;\alpha,\beta) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{(1+(t/\alpha)^{\beta})^2} \;\; ; \;\; S(t;\alpha,\beta) = \frac{1}{1+(t/\alpha)^{\beta}} \;\; ; \;\; \lambda(t;\alpha,\beta) = \frac{f(\cdot)}{S(\cdot)} = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{1+(t/\alpha)^{\beta}}$$

Under a multiplicative hazard assumption, we combine the piecewise-constant and log logistic hazard functions, λ_s and λ_p. Note the convenience of the multiplicative hazard assumption because of the mapping into valid hazard space $\{\mathbb{R}^+, \mathbb{R}^+\} \to \mathbb{R}^+$. The log likelihood becomes:

$$\begin{aligned} \log p(x|\Theta) &\propto \sum_{i=1}^{n} \Big(\mathbb{1}(l = l_i) \log \lambda(t_i|h_i) - \int_{-\infty}^{t_i} \lambda(\tau|h_i) d\tau \Big) \\ &= \sum_{i=1}^{n} \Big(\mathbb{1}(l = l_i)(\log \lambda_s + \log \lambda_p(t_i|h_i)) - \lambda_s \int_{-\infty}^{t_i} \lambda_p(\tau|h_i) d\tau \Big) \end{aligned} \tag{2}$$

From Equation 2 we observe that, for fixed λ_s, the likelihood optimization is the same as log logistic distribution optimization with the survival component weighted by λ_s. To allow for further flexibility when introducing the log logistic hazard, we provide a third parameter γ, a height scaling factor, such that $\lambda(t|h) = \gamma \lambda_s \lambda_p(t|h)$. We can use standard gradient methods to optimize either formulation; we use L-BFGS-B with bounds described in Section 4.

3.1 Approximation of log logistic distribution

Once we have learned the parameters for the γ-scaled log logistic distribution, we piecewise approximate the learned log logistic distribution, shift it into the piecewise components, and learn another log logistic factor. Figure 1 demonstrates one iteration of this procedure. Iteratively fixing the log logistic parameters creates a simple forest building process, though joint optimization over the product of log logistic distributions with fixed piecewise-constant multipliers is also possible. We select the simpler approach in our experiments.

We adopt near-quantile, survival-matched, quantized, piecewise-constant approximations. The near-quantile property prevents creation of unnecessary piecewise approximations where the hazard is insignificant. Survival matching means that each near-quantile hazard approximation has the property that, for all near-quantile time t, the survival function to the right of t is unaffected by the approximation left of t. Quantization shifts the quantile approach to a near-quantile approach by selecting the nearest time on a fine grid. Requiring the piecewise approximation times to be on a grid sets an upper limit on the number of pieces created by the distribution approximations. We believe quantization is not too bad, for example, using a grid of 1 minute when the clinical events occur at the scale of hours to years. Note that just the piecewise approximations of the hazard occur on the grid; the events need not.

4. Experimental setup

We conduct simulations on periodic and aperiodic signals, and apply our method on the MIMIC III data set, version 1.4 (Johnson et al., 2016) for determination of length of stay in the ICU for individuals admitted with diabetic ketoacidosis.

4.1 Log logistic forest learning

The forest is learned from an initialization of a single root node with fixed hazard of 1. The model is run for d iterations with proposal to modify the forest, with equal probability of generating (1) a split based on time randomly sampled from the empirical distribution of time to events times uniform random noise times two, or (2) using a log logistic distribution who parameters are learned

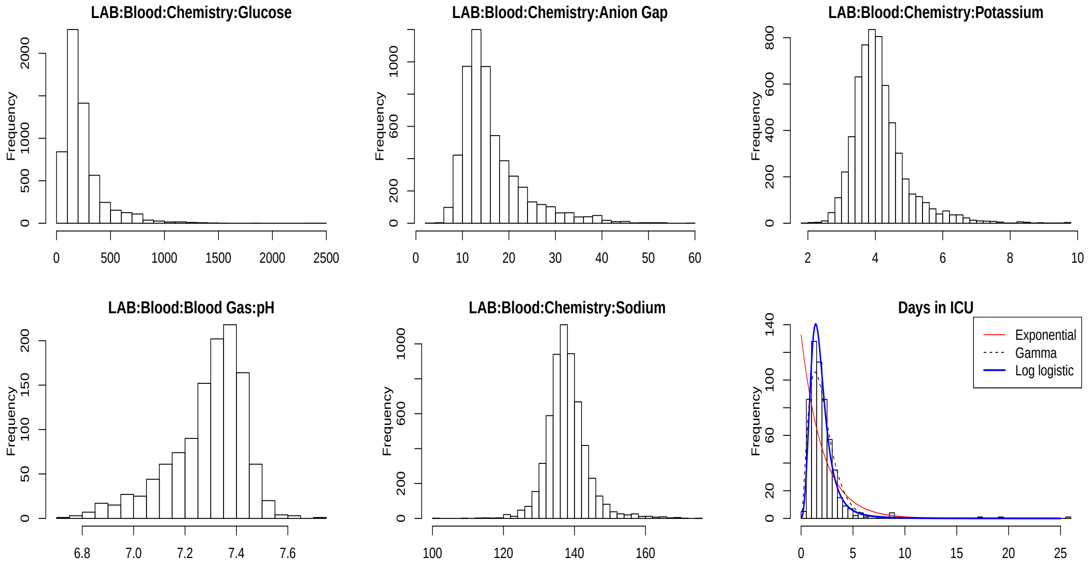

Figure 2: Histograms of key laboratory measures in patients with diabetic ketoacidosis. Indicators of resolution of ketoacidosis include (1) return to normal to mildly elevated glucose levels to below 200 (top left) and (2) closure of the anion gap to 12 (top center). Physicians also monitor potassium whose aberrant levels cause cardiac arrhythmia (top right), pH indicative of the acidotic state (bottom left), and sodium with concerns for altered mental status from fluid correction (bottom center). There is variation in the length of stay in the ICU in days captured by the log logistic distribution (bottom right).

using L-BFGS-B. The bounds are placed on the log logistic parameters: $(\log \alpha) \in [-5, 23], \beta \in [0.01, \infty), \gamma \in [-100, 5]$. The lower bound on shape parameter β prevents creating a spike log logistic function to limit the gain in the likelihood for exactly predicting the event time in a completely periodic signal. See Simulation 1 and Figure 3 for an illustrative example. The bounds on scale parameter α helps prevent double overflow and underflow, and values outside the range are not meaningful. The multiplicative hazard parameter γ bounds help prevent underflow and spike behaviors respectively. The location of each proposal is uniformly random over all existing nodes with additional weight of e^{δ} (we choose $\delta = 3$) for the blank node corresponding to creation of a new tree in the forest. Each proposal is evaluated in terms of change in training set log likelihood and is accepted if the proposal passes AIC criterion.

Once a proposal is accepted, the tree is modified to contain the split or to contain the distribution and its parameterizations. When considering the next proposal, the piecewise approximation must take into account the splits and log logistic distributions already learned. Thus the intervals in the data are split into smaller intervals on the grid where piecewise-constant approximations are specified. We use a quantile size of 0.01 and a grid step $g = 0.01$ days ≈ 14 minutes.

4.2 Simulations

We conduct two simulations for recurrent events of interest to examine the algorithms ability to model both regular and irregular timing of events. The first is the task of learning an entirely regular signal occurring at intervals of 10 time units apart. Data were generated as time $\{k, k +$

$10, \ldots, k+100\}$ for k in $\{1, \ldots, 10\}$. The second simulation task was learning an entirely irregular signal having fixed rate of 0.1, so that, on average, an events occurs at 10 time units, *i.e.*, events are drawn from a homogeneous Poisson with rate $\lambda = 0.1$. Ten samples of length 110 time units were generated to match the time length of the periodic case in Simulation 1. Note that these tasks are different from the standard survival analysis task: the task continues after each failure. For the simulations we set the number of iterations to $d = 10$.

4.3 Diabetic ketoacidosis in MIMIC III

The MIMIC III population includes over 58,000 individuals admitted to intensive care units at Beth Israel Deaconess Center between 2001 and 2012 (Johnson et al., 2016). We identified the set of admissions to intensive care units as those whose primary coded diagnosis contained the word "ketoacidosis". This cohort criterion resulted in a population of 355 individuals. Seventy individuals had multiple visits to the ICU; visits beyond the first were included.

Features included in the study comprised demographics: age, gender, ethnicity, insurance type, and marital status; admissions data: admission type and location; lab values: blood counts, electrolytes, anion gap, and pH; intravenous or oral medications: insulin, potassium, magnesium and bicarbonate; monitor data of mean arterial blood pressure; and comorbidities recorded in the coded diagnosis list. Real valued data were binned into septiles. The resulting process produced 416 features measured over time. MIMIC III contains features measured throughout the hospital encounter and these data were also used. The period of likelihood calculation was over the time spent in the ICU, defined by the timestamps of ICU entry and ICU exit.

Table 1: Baseline patient characteristics, reported with median [2.5%,97.5%] and *n* (fraction).

Feature	*n*=355
Age	42 [20, 79]
Gender	
male	162 (0.46)
female	193 (0.54)
Ethnicity	
Caucasian	206 (0.57)
African American	91 (0.26)
Hispanic or Latino	25 (0.07)
Asian	5 (0.01)
Other	28 (0.08)
WBC	12.2 [5.1,30.3]
Lactate	261 (0.74)
	2.3 [0.9, 8.1]
Co-sepsis	11 (0.03)
Deaths	4 (0.01)

Table 1 provides descriptive characteristics of the study population. Figure 2 shows the distribution across several key variables measured during the resolution process of DKA, with the bottom right showing the distribution of ICU length of stay. Note its distribution is nicely modeled by the log logistic, where as commonly used exponential and gamma distributions would not as effectively capture the distribution of time-to-event.

We compare the piecewise-constant log logistic learning algorithm (LL forest) against forest learning alone (MFPPs) and the Cox proportional hazards model. To ensure that the ICU entry is included in the model, we initialize the LL forest with a log logistic distribution initiated by ICU entry. We set the number of proposal iterations to $d = 200$. For Cox regression, we specify the nuisance function as log logistic estimated by gradient descent on negative log likelihood and use glmnet to select a sparse subset of features with parameter $s = 0.1$ (Friedman et al., 2009). To compare performance, we use paired t-tests between the LL forest method and the other method across patients.

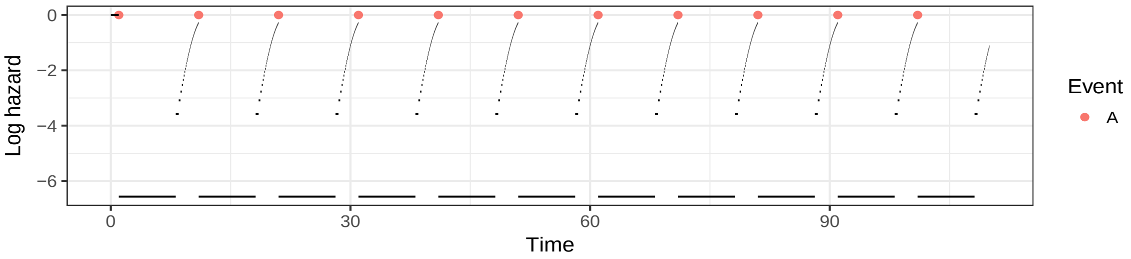

Figure 3: Periodic simulated data with average rate of 0.1. The LL forest learning algorithm identifies the periodicity maintaining a very low hazard rate until near the next periodic event.

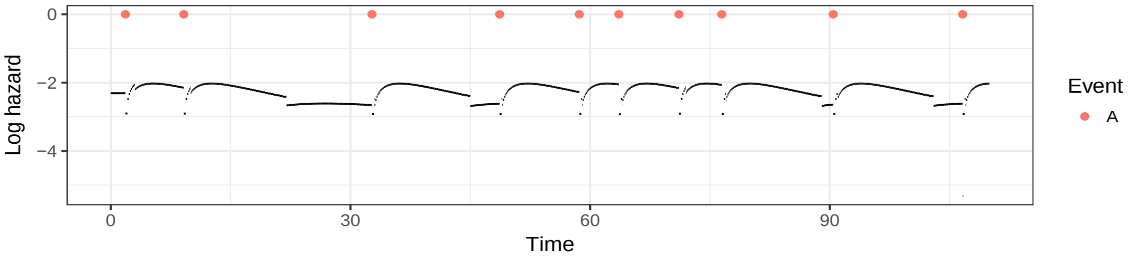

Figure 4: Homogeneous Poisson simulated data with hazard rate $\lambda = 0.1$; $\log \lambda = -2.30$. LL forests overfit yet approximate the hazard.

For forest learning, random proposals were generated with events uniformly randomly sampled and with time-to-events sampled according to the training data event distribution, subject to the time being valid for the particular experimental design. Proposals included leaf node splits and leaf modifications converting a rate factor leaf to a log logistic-parameterized leaf. To encourage forest growth, after each root modification, a blank root node is added to the forest with increased odds of being chosen for modification: $\delta = 3$, *i.e.*, $e^3 : 1$ odds compared to each modified node.

The patients are randomly split into training and test sets in a 4:1 ratio for 5-fold cross validation. Our primary evaluation endpoint is the test set average log likelihood. We additionally develop visualizations to compare and interpret the models overlaid with patient events.

5. Results

Figure 3 demonstrates the ability of the joint forest and log logistic learning algorithm to learn a periodic signal. Without the bound imposed on the shape parameter β in L-BFGS-B, the spikes would be sharper and the wells lower. However, practically speaking, this solution effectively captures the periodicity and would be robust to minor shifts in test set data time-stamps.

Figure 4 shows the joint model mildly overfitting Poisson data. The ground truth solution is a horizontal line at $\log 0.1 \approx -2.30$. While constructing an overly complicated model, the learning algorithm does not appear to memorize the data. As the training set size increases, the overfitting is reduced substantially (not shown).

Table 2 highlights the performance of LL forests in prediction of timing of discharge for ketoacidosis ICU admissions. We observe that LL forest outperforms the forest learning algorithm (MFPPs) and the log logistic Cox regression model. Because the LL forest algorithm has access

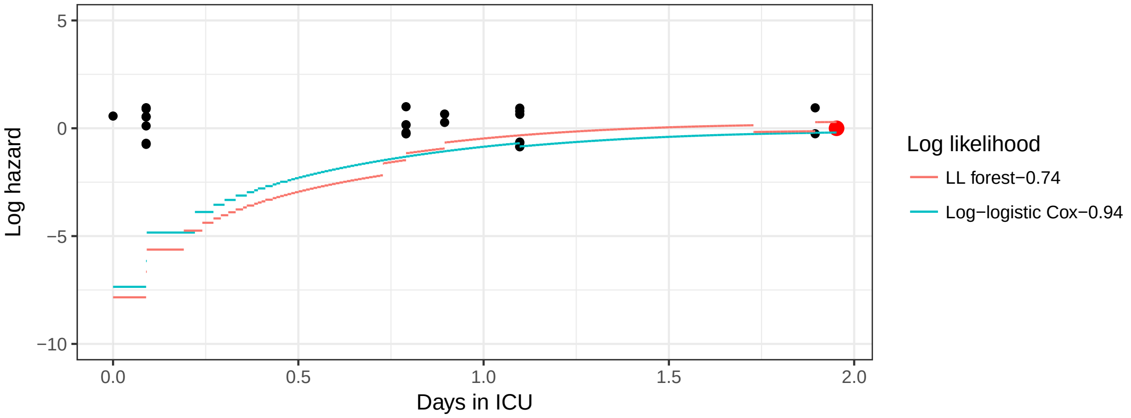

Figure 5: Patient timeline with estimated log hazard function for the joint model and the log logistic Cox regression. The large red dot indicates discharge from the ICU and jittered black dots indicate measurements of features included in the model. A higher hazard (or log hazard) is the model's suggestion that the event is more likely to occur.

to data beyond the baseline, the algorithm was rerun on baseline data only and compared to the log logistic Cox regression model. The results suggest that again the LL forest model significantly outperforms the log logistic Cox model. While the LL forest model had lower average negative low likelihood than the LL forest using only baseline features, the result was not significant.

The forest method without log logistic leaves (MFPPs) has notable high variability in its predictions, making some apt predictions and some costly mistakes. Nonetheless, the paired t-test suggests that the LL forest model significantly outperforms both MFPPs and log logistic Cox regression. Figure 5 illustrates the learned hazard for the LL forest model and the log logistic Cox regression model overlaid on the data for one patient. LL forests appear more confident about lack of early discharge and lack of discharge without recent measurement.

6. Discussion

Iterative forest and log logistic approximation learning enables the efficient modeling of the hazard function at any time, with the ability to model both regular, *i.e.* predictable or scheduled events, and irregular or entropic events. Whereas forest point processes have difficulty modeling parametric distributions often seen in real-world data, LL forests can incorporate approximations to them. LL

Table 2: Average likelihood ± 95% CI; paired t-test for LL forest, forest (MFPPs), and log logistic Cox regression; paired t-test for log logistic Cox regression versus LL forest with baseline features

Data	LL Forest	Forest	LL-Cox	LL Forest, baseline features
Test set	-1.34 ± 0.48	-2.22 ± 1.39	-1.54 ± 0.39	-1.45 ± 0.39
p-value	–	0.010	0.036	0.40
gap p-value			<0.001	–

forests also can effectively model the hazard with arbitrary time-stamping of data, instead of relying on covariates measured in panels all at once. Effectively, the LL forest point process transforms series of time-stamped events into hazard functions, which provides an interpretation for the risk of the outcome of interest at any point in time.

A difference of 0.2 in average log likelihood corresponds either to the comparison algorithm's comparative overestimates of the rate of discharge over time or underestimates when the discharges do occur, or some mixture of both. In the case of ketoacidosis length of stay prediction, the benefit seen in LL forests is likely due to two factors, first, the better ability to model survival distributions than MFPPs given limited data, and, second, violation of the proportional hazards assumption made in Cox models. Limited, heterogeneous data are commonplace in the health care setting, and LL forests may be good choice for this setting.

The comparison against the log logistic Cox regression is limited by the fact that the LL forest model has access to the intermediate findings past the admission time. However, our method substantially outperforms MFPPs which has the same access. Additionally, when using baseline features only, the LL forest significantly outperforms the log logistic Cox regression.

We included mediating events in our modeling because by doing so, we are able to provide risk assessments with as much information as possible, an analogue to what signals the physician has access to throughout the admission. This is a demonstration of "now-time" hazard estimation, depicted in Figure 5, where model belief about the "present"–namely a model for, "will the patient discharge right now"–could be interpreted as a useful surrogate for "is the patient ready for discharge?"

To achieve forecasting to time T, the features used to model the outcome event at time T_0 must simply derive from times before $T_0 - T$. The learning framework could then be extended to learn features with time gap T_f and shrinking $T_f \to 0$. By restricting the learning of trees to have specified time gaps, the learned forest can then make forecasts at arbitrary length scales between 0 and T. This model variant could be useful for screening for high risk individuals and determining schedules for the timing of follow-up.

Future work will consider modifying the objective function of bounded model likelihood to better address pertinent clinical questions, including: readiness of discharge, real-valued and structured predictions. While log logistic distributions have useful time and uncertainty representations, other forms of parametric modeling will be considered including, periodic filters to account for latent temporal regularity such as night shifts and seasonality. We will also consider risk attribution questions which can be informed by specification the length scale T_f. Instead of making risk attribution statements about features measured at T, or using the distribution of features and outcome without time to infer causation, point processes could provide early indications of risk at any time T_f.

7. Conclusion

Piecewise-constant representations can provide arbitrarily close approximations to any hazard function. The difficulty is effectively and efficiently learning those approximations from data. This work introduces an iterative piecewise-constant conversion method which enables complex hazard functions with compact parameterizations to be learned from regularly or irregularly time-stamped data. We demonstrate in real data of patients with diabetic ketoacidosis that our method outperforms forest and Cox regression algorithms at learning hazard functions from regular and irregularly time-stamped data.

References

Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4), 2009.

Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 446, 2015.

James Goulding, Simon Preston, and Gavin Smith. Event series prediction via non-homogeneous Poisson process modelling. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 161–170. IEEE, 2016.

A. Gunawardana, C. Meek, and P. Xu. A model for temporal dependencies in event streams. Advances in Neural Information Processing Systems, 2011.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3, 2016.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Thomas A Lasko. Efficient inference of gaussian-process-modulated renewal processes with application to medical event data. In *Conference on Uncertainty in Artificial Intelligence*, 2014.

Jerald F Lawless. Regression methods for Poisson process data. *Journal of the American Statistical Association*, 82(399):808–815, 1987.

Wenzhao Lian, Ricardo Henao, Vinayak Rao, Joseph Lucas, and Lawrence Carin. A multitask point process predictive model. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2030–2038, 2015.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Hallie C Prescott, Carolyn S Calfee, B Taylor Thompson, Derek C Angus, and Vincent X Liu. Toward smarter lumping and smarter splitting: rethinking strategies for sepsis and acute respiratory distress syndrome clinical trial design. *American journal of respiratory and critical care medicine*, 194(2):147–155, 2016.

Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal lab tests. *arXiv preprint arXiv:1608.00647*, 2016.

Niels C Riedemann, Ren-Feng Guo, and Peter A Ward. The enigma of sepsis. *The Journal of clinical investigation*, 112(4):460–467, 2003.

Alan D Saul, James Hensman, Aki Vehtari, and Neil D Lawrence. Chained Gaussian processes. In *Artificial Intelligence and Statistics*, pages 1431–1440, 2016.

Terry Therneau, Cindy Crowson, and Elizabeth Atkinson. Using time dependent covariates and time dependent coefficients in the Cox model. *Survival Vignettes*, 2017.

Gerhard Tutz and Harald Binder. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62(4):961–971, 2006.

Jeremy C Weiss and David Page. Forest-based point process for event prediction from electronic health records. In *Machine Learning and Knowledge Discovery in Databases*, pages 547–562. Springer, 2013.

Jeremy C Weiss, Sriraam Natarajan, and David Page. Multiplicative forests for continuous-time processes. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2012.