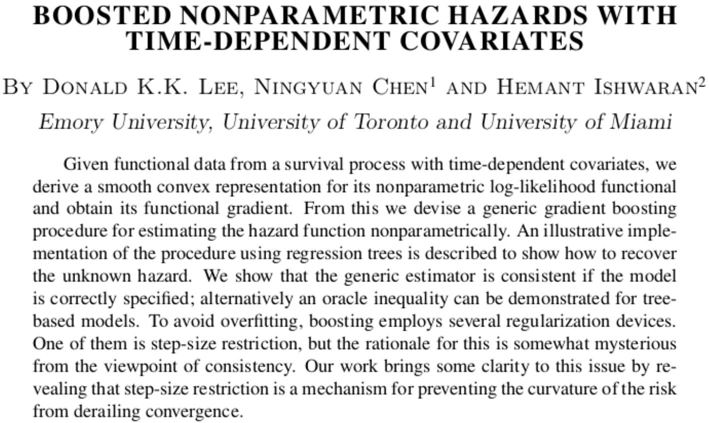
Journal club: Boosted nonparametric hazards with time-dependent covariates

Anders Munch

February 24, 2021

The article



BOOSTED NONPARAMETRIC HAZARDS WITH TIME-DEPENDENT COVARIATES

By Donald K.K. Lee, Ningyuan Chen[1] and Hemant Ishwaran[2]

Emory University, University of Toronto and University of Miami

Given functional data from a survival process with time-dependent covariates, we derive a smooth convex representation for its nonparametric log-likelihood functional and obtain its functional gradient. From this we devise a generic gradient boosting procedure for estimating the hazard function nonparametrically. An illustrative implementation of the procedure using regression trees is described to show how to recover the unknown hazard. We show that the generic estimator is consistent if the model is correctly specified; alternatively an oracle inequality can be demonstrated for tree-based models. To avoid overfitting, boosting employs several regularization devices. One of them is step-size restriction, but the rationale for this is somewhat mysterious from the viewpoint of consistency. Our work brings some clarity to this issue by revealing that step-size restriction is a mechanism for preventing the curvature of the risk from derailing convergence.

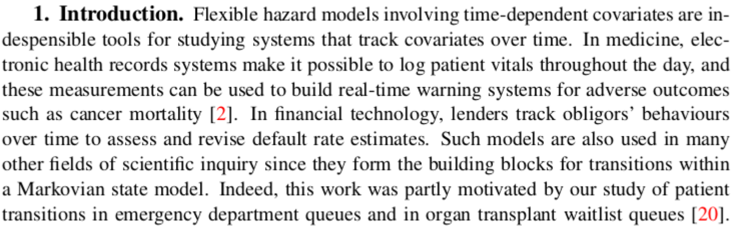
To appear in Annals of Statistics.

Outline of the problem

While there are many boosting methods for dealing with time-static covariates, the literature is far more sparse for the case of time-dependent covariates. In fact, to our knowledge there is no general nonparametric approach for dealing with this setting. This is because in order to implement a fully nonparametric estimator, one has to contend with the issue of identifying the gradient, which turns out to be a non-trivial problem due to the functional nature of the data. This is unlike most standard applications of gradient boosting where the gradient can easily be identified and calculated. (p.2)

Setting and motivation – why?

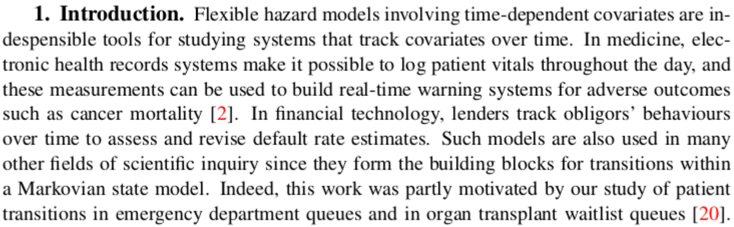
Their setting



1. Introduction. Flexible hazard models involving time-dependent covariates are indespensible tools for studying systems that track covariates over time. In medicine, electronic health records systems make it possible to log patient vitals throughout the day, and these measurements can be used to build real-time warning systems for adverse outcomes such as cancer mortality [2]. In financial technology, lenders track obligors' behaviours over time to assess and revise default rate estimates. Such models are also used in many other fields of scientific inquiry since they form the building blocks for transitions within a Markovian state model. Indeed, this work was partly motivated by our study of patient transitions in emergency department queues and in organ transplant waitlist queues [20].

Setting and motivation – why?

Their setting



1. Introduction. Flexible hazard models involving time-dependent covariates are indespensible tools for studying systems that track covariates over time. In medicine, electronic health records systems make it possible to log patient vitals throughout the day, and these measurements can be used to build real-time warning systems for adverse outcomes such as cancer mortality [2]. In financial technology, lenders track obligors' behaviours over time to assess and revise default rate estimates. Such models are also used in many other fields of scientific inquiry since they form the building blocks for transitions within a Markovian state model. Indeed, this work was partly motivated by our study of patient transitions in emergency department queues and in organ transplant waitlist queues [20].

Nuisance / plug-in estimator

- mediation analysis in continuous time
- dynamic treatment regime in continuous time
- local independence testing

Boosting in general (regression and classification)

Boosting in general (regression and classification)

Fitting residuals

Initialize $f_0 = \bar{y}_i$, then iteratively estimate f_m to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(r_{im} - f_m(x_i))^2 \quad \text{with} \quad r_{im} := y_i - f_{m-1}(x_i).$$

Boosting in general (regression and classification)

Fitting residuals

Initialize $f_0 = \bar{y}_i$, then iteratively estimate f_m to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(r_{im} - f_m(x_i))^2 \quad \text{with} \quad r_{im} := y_i - f_{m-1}(x_i).$$

Re-weighting samples

Initialize weights $w_i = 1/n$, $i = 1, \ldots, n$, then repeat

1. fit f_m to weighted data set
2. update weight w_i based on f_m's performance on sample i

Boosting in general (regression and classification)

Fitting residuals

Initialize $f_0 = \bar{y}_i$, then iteratively estimate f_m to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(r_{im} - f_m(x_i))^2 \quad \text{with} \quad r_{im} := y_i - f_{m-1}(x_i).$$

Re-weighting samples

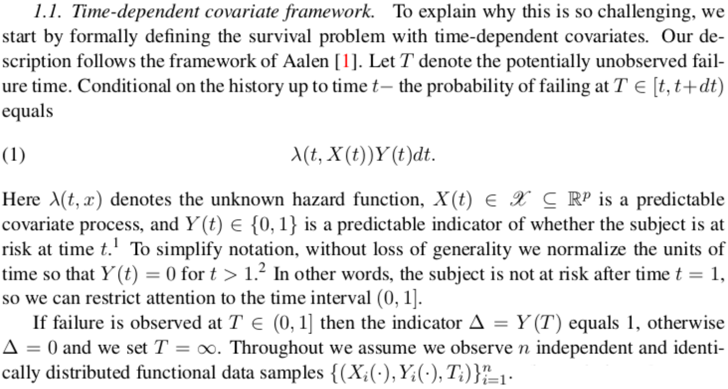
Initialize weights $w_i = 1/n$, $i = 1, ..., n$, then repeat

1. fit f_m to weighted data set
2. update weight w_i based on f_m's performance on sample i

→ Functional gradient descent of some loss function L

$$F_m = F_{m-1} - \nu f_m, \quad \text{with} \quad f_m = \underset{f}{\operatorname{argmin}} \left\| \left.\frac{\partial L}{\partial F}\right|_{F=F_{m-1}} - f \right\|$$

The loss function in this setting



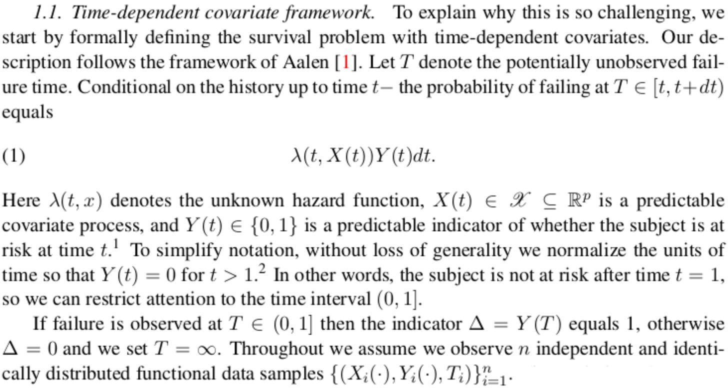
1.1. Time-dependent covariate framework. To explain why this is so challenging, we start by formally defining the survival problem with time-dependent covariates. Our description follows the framework of Aalen [1]. Let T denote the potentially unobserved failure time. Conditional on the history up to time $t-$ the probability of failing at $T \in [t, t+dt)$ equals

$$\lambda(t, X(t))Y(t)dt. \tag{1}$$

Here $\lambda(t, x)$ denotes the unknown hazard function, $X(t) \in \mathscr{X} \subseteq \mathbb{R}^p$ is a predictable covariate process, and $Y(t) \in \{0, 1\}$ is a predictable indicator of whether the subject is at risk at time t.[1] To simplify notation, without loss of generality we normalize the units of time so that $Y(t) = 0$ for $t > 1$.[2] In other words, the subject is not at risk after time $t = 1$, so we can restrict attention to the time interval $(0, 1]$.

If failure is observed at $T \in (0, 1]$ then the indicator $\Delta = Y(T)$ equals 1, otherwise $\Delta = 0$ and we set $T = \infty$. Throughout we assume we observe n independent and identically distributed functional data samples $\{(X_i(\cdot), Y_i(\cdot), T_i)\}_{i=1}^{n}$.

The loss function in this setting



1.1. Time-dependent covariate framework. To explain why this is so challenging, we start by formally defining the survival problem with time-dependent covariates. Our description follows the framework of Aalen [1]. Let T denote the potentially unobserved failure time. Conditional on the history up to time $t-$ the probability of failing at $T \in [t, t+dt)$ equals

$$\lambda(t, X(t))Y(t)dt. \tag{1}$$

Here $\lambda(t, x)$ denotes the unknown hazard function, $X(t) \in \mathscr{X} \subseteq \mathbb{R}^p$ is a predictable covariate process, and $Y(t) \in \{0, 1\}$ is a predictable indicator of whether the subject is at risk at time t.[1] To simplify notation, without loss of generality we normalize the units of time so that $Y(t) = 0$ for $t > 1$.[2] In other words, the subject is not at risk after time $t = 1$, so we can restrict attention to the time interval $(0, 1]$.

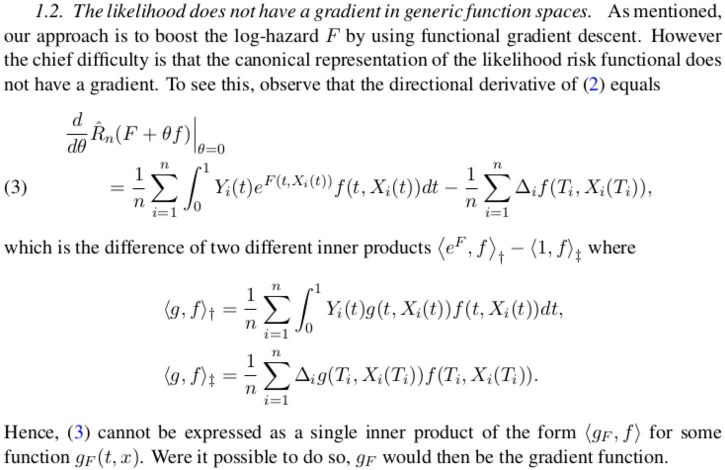
If failure is observed at $T \in (0, 1]$ then the indicator $\Delta = Y(T)$ equals 1, otherwise $\Delta = 0$ and we set $T = \infty$. Throughout we assume we observe n independent and identically distributed functional data samples $\{(X_i(\cdot), Y_i(\cdot), T_i)\}_{i=1}^n$.

The (scaled) negative log-likelihood functional

$$\hat{R}_n(F) = \frac{1}{n}\sum_{i}^{n}\int_0^1 Y_i(t)e^{F(t,X_i(t))}\,\mathrm{d}t - \frac{1}{n}\sum_{i=1}^{n}\Delta_i F(T_i, X_i(T_i)),$$

where $F(t, x) = \log(\lambda(t, x))$ and Δ_i is event indicator.

$\hat{R}_n$ does not have a gradient . . . ?



1.2. The likelihood does not have a gradient in generic function spaces. As mentioned, our approach is to boost the log-hazard F by using functional gradient descent. However the chief difficulty is that the canonical representation of the likelihood risk functional does not have a gradient. To see this, observe that the directional derivative of (2) equals

$$\begin{aligned} &\frac{d}{d\theta}\hat{R}_n(F+\theta f)\Big|_{\theta=0} \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 Y_i(t)e^{F(t,X_i(t))}f(t,X_i(t))dt - \frac{1}{n}\sum_{i=1}^{n}\Delta_i f(T_i,X_i(T_i)), \end{aligned} \tag{3}$$

which is the difference of two different inner products $\left\langle e^F, f\right\rangle_\dagger - \langle 1, f\rangle_\ddagger$ where

$$\begin{aligned} \langle g, f\rangle_\dagger &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 Y_i(t)g(t,X_i(t))f(t,X_i(t))dt, \\ \langle g, f\rangle_\ddagger &= \frac{1}{n}\sum_{i=1}^{n}\Delta_i g(T_i,X_i(T_i))f(T_i,X_i(T_i)). \end{aligned}$$

Hence, (3) cannot be expressed as a single inner product of the form $\langle g_F, f\rangle$ for some function $g_F(t,x)$. Were it possible to do so, g_F would then be the gradient function.

Finding a domain for $\hat{R}_n$

Let $\hat{\mu}_n$ be the empirical sub-probability measure on $[0,1] \times \mathscr{X} \subset [0,1] \times \mathbb{R}^p$, defined as

$$\hat{\mu}_n(B) := \frac{1}{n} \sum_{i=1}^{n} \int_0^1 Y_i(t) \cdot I[\{t, X_i(t)\} \in B] \, \mathrm{d}t,$$

let $\{\varphi_j(t,x)\}_{j=1}^{d}$ be a set of bounded functions

$$\varphi_j \colon [0,1] \times \mathscr{X} \to [-1,1],$$

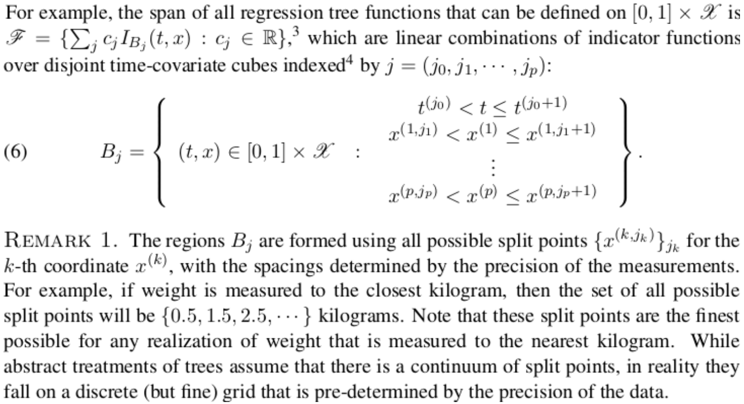
that are linearly independent in $\mathcal{L}^2(\mathrm{d}t \otimes \mathrm{d}x)$, and set

$$\mathcal{F} := \mathrm{span}\{\varphi_j \ : \ j = 1, \ldots, d\}.$$

Then the sample-dependent domain of $\hat{R}_n$ is

$$(\mathcal{F}, \langle \cdot, \cdot \rangle_{\hat{\mu}_n}) \subset \mathcal{L}^2(\hat{\mu}_n).$$

The domain when using regression trees

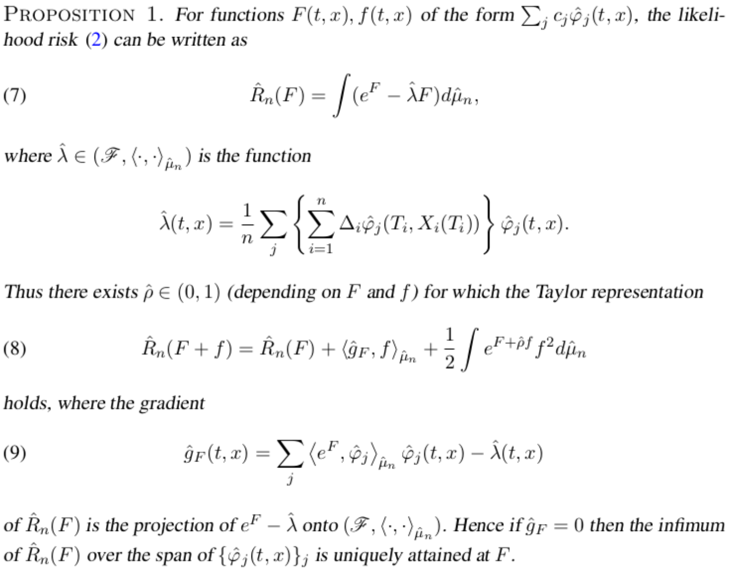


For example, the span of all regression tree functions that can be defined on $[0,1]\times\mathscr{X}$ is $\mathscr{F}=\{\sum_j c_j I_{B_j}(t,x) : c_j \in \mathbb{R}\}$,[3] which are linear combinations of indicator functions over disjoint time-covariate cubes indexed[4] by $j=(j_0,j_1,\cdots,j_p)$:

$$B_j = \left\{ (t,x)\in[0,1]\times\mathscr{X} \quad : \quad \begin{array}{c} t^{(j_0)} < t \leq t^{(j_0+1)} \\ x^{(1,j_1)} < x^{(1)} \leq x^{(1,j_1+1)} \\ \vdots \\ x^{(p,j_p)} < x^{(p)} \leq x^{(p,j_p+1)} \end{array} \right\}. \tag{6}$$

REMARK 1. The regions B_j are formed using all possible split points $\{x^{(k,j_k)}\}_{j_k}$ for the k-th coordinate $x^{(k)}$, with the spacings determined by the precision of the measurements. For example, if weight is measured to the closest kilogram, then the set of all possible split points will be $\{0.5, 1.5, 2.5, \cdots\}$ kilograms. Note that these split points are the finest possible for any realization of weight that is measured to the nearest kilogram. While abstract treatments of trees assume that there is a continuum of split points, in reality they fall on a discrete (but fine) grid that is pre-determined by the precision of the data.

Integral representation of the likelihood



Proposition 1. *For functions $F(t,x), f(t,x)$ of the form $\sum_j c_j \hat\varphi_j(t,x)$, the likelihood risk* (2) *can be written as*

$$\hat R_n(F) = \int (e^F - \hat\lambda F) d\hat\mu_n, \tag{7}$$

where $\hat\lambda \in (\mathscr{F}, \langle\cdot,\cdot\rangle_{\hat\mu_n})$ is the function

$$\hat\lambda(t,x) = \frac{1}{n} \sum_j \left\{ \sum_{i=1}^{n} \Delta_i \hat\varphi_j(T_i, X_i(T_i)) \right\} \hat\varphi_j(t,x).$$

Thus there exists $\hat\rho \in (0,1)$ (depending on F and f) for which the Taylor representation

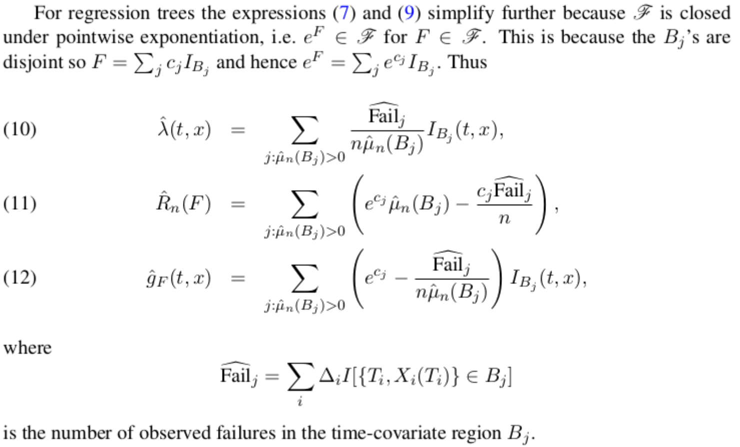
$$\hat R_n(F+f) = \hat R_n(F) + \langle \hat g_F, f \rangle_{\hat\mu_n} + \frac{1}{2} \int e^{F+\hat\rho f} f^2 d\hat\mu_n \tag{8}$$

holds, where the gradient

$$\hat g_F(t,x) = \sum_j \left\langle e^F, \hat\varphi_j \right\rangle_{\hat\mu_n} \hat\varphi_j(t,x) - \hat\lambda(t,x) \tag{9}$$

of $\hat R_n(F)$ is the projection of $e^F - \hat\lambda$ onto $(\mathscr{F}, \langle\cdot,\cdot\rangle_{\hat\mu_n})$. Hence if $\hat g_F = 0$ then the infimum of $\hat R_n(F)$ over the span of $\{\hat\varphi_j(t,x)\}_j$ is uniquely attained at F.

Representation for regression trees



For regression trees the expressions (7) and (9) simplify further because $\mathscr{F}$ is closed under pointwise exponentiation, i.e. $e^F \in \mathscr{F}$ for $F \in \mathscr{F}$. This is because the B_j's are disjoint so $F = \sum_j c_j I_{B_j}$ and hence $e^F = \sum_j e^{c_j} I_{B_j}$. Thus

$$\text{(10)} \qquad \hat{\lambda}(t,x) = \sum_{j:\hat{\mu}_n(B_j)>0} \frac{\widehat{\text{Fail}}_j}{n\hat{\mu}_n(B_j)} I_{B_j}(t,x),$$

$$\text{(11)} \qquad \hat{R}_n(F) = \sum_{j:\hat{\mu}_n(B_j)>0} \left(e^{c_j}\hat{\mu}_n(B_j) - \frac{c_j \widehat{\text{Fail}}_j}{n} \right),$$

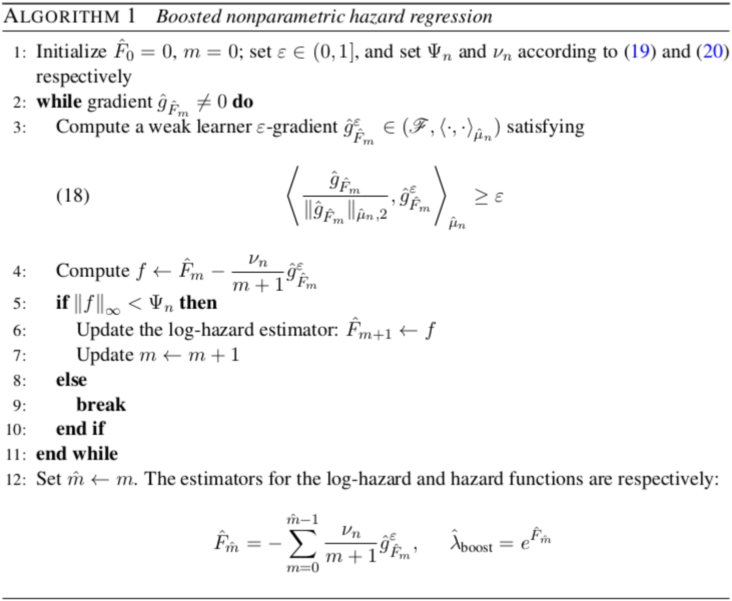
$$\text{(12)} \qquad \hat{g}_F(t,x) = \sum_{j:\hat{\mu}_n(B_j)>0} \left(e^{c_j} - \frac{\widehat{\text{Fail}}_j}{n\hat{\mu}_n(B_j)} \right) I_{B_j}(t,x),$$

where

$$\widehat{\text{Fail}}_j = \sum_i \Delta_i I[\{T_i, X_i(T_i)\} \in B_j]$$

is the number of observed failures in the time-covariate region B_j.

The boosting algorithm



ALGORITHM 1 *Boosted nonparametric hazard regression*

```
Initialize $\hat{F}_0 = 0$, $m = 0$; set $\varepsilon \in (0,1]$, and set $\Psi_n$ and $\nu_n$ according to (19) and (20) respectively
while gradient $\hat{g}_{\hat{F}_m} \neq 0$ do
    Compute a weak learner $\varepsilon$-gradient $\hat{g}^{\varepsilon}_{\hat{F}_m} \in (\mathscr{F}, \langle\cdot,\cdot\rangle_{\hat{\mu}_n})$ satisfying
    (18)    $\left\langle \frac{\hat{g}_{\hat{F}_m}}{\|\hat{g}_{\hat{F}_m}\|_{\hat{\mu}_n,2}}, \hat{g}^{\varepsilon}_{\hat{F}_m} \right\rangle_{\hat{\mu}_n} \geq \varepsilon$
    Compute $f \leftarrow \hat{F}_m - \frac{\nu_n}{m+1}\hat{g}^{\varepsilon}_{\hat{F}_m}$
    if $\|f\|_\infty < \Psi_n$ then
        Update the log-hazard estimator: $\hat{F}_{m+1} \leftarrow f$
        Update $m \leftarrow m + 1$
    else
        break
    end if
end while
Set $\hat{m} \leftarrow m$. The estimators for the log-hazard and hazard functions are respectively:
    $\hat{F}_{\hat{m}} = -\sum_{m=0}^{\hat{m}-1} \frac{\nu_n}{m+1}\hat{g}^{\varepsilon}_{\hat{F}_m}, \qquad \hat{\lambda}_{\text{boost}} = e^{\hat{F}_{\hat{m}}}$
```

Skipping some technical stuff

- Consistency
- Convergence rates
- Choice of hyper-parameters
- Better understanding of boosting in general

Some details for a tree-based implementation

At the m'th iteration approximate the gradient $\hat{g}_{\hat{F}_m}$ with a tree: Split leaf regions $A \subset [0,1] \times \mathscr{X}$ into left and right daughter sub-regions A_1 and A_2, either by splitting on a covariate k,

$$A_1 = \{(t,x) \in A \,:\, x^{(k)} \leq s\}, \quad A_2 = \{(t,x) \in A \,:\, x^{(k)} > s\},$$

or on time

$$A_1 = \{(t,x) \in A \,:\, t \leq s\}, \quad A_2 = \{(t,x) \in A \,:\, t > s\}.$$

Some details for a tree-based implementation

At the m'th iteration approximate the gradient $\hat{g}_{\hat{F}_m}$ with a tree: Split leaf regions $A \subset [0,1] \times \mathscr{X}$ into left and right daughter sub-regions A_1 and A_2, either by splitting on a covariate k,

$$A_1 = \{(t,x) \in A \;:\; x^{(k)} \leq s\}, \quad A_2 = \{(t,x) \in A \;:\; x^{(k)} > s\},$$

or on time

$$A_1 = \{(t,x) \in A \;:\; t \leq s\}, \quad A_2 = \{(t,x) \in A \;:\; t > s\}.$$

Choosing these to minimize $\mathcal{L}^2(\hat{\mu}_n)$ error is equivalent to minimizing

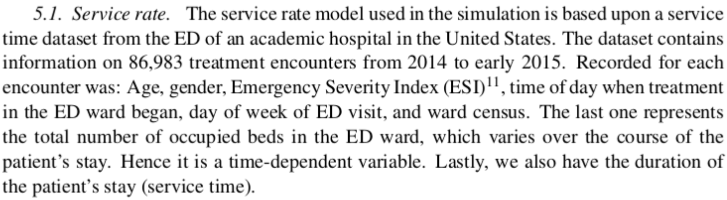
$$\min_{\gamma_1} \sum_{\substack{j:B_j \subseteq A_1, \\ w_j > 0}} w_j \cdot (\tilde{y}_j - \gamma_1)^2 + \min_{\gamma_2} \sum_{\substack{j:B_j \subseteq A_2, \\ w_j > 0}} w_j \cdot (\tilde{y}_j - \gamma_2)^2,$$

where

$$\tilde{y}_j := e^{c_{m,j}} - \frac{\hat{\mathrm{Fail}}_j}{n\hat{\mu}_n(B_j)}, \quad \text{and} \quad w_j := \hat{\mu}_n(B_j),$$

are the pseudo-response and its weight.

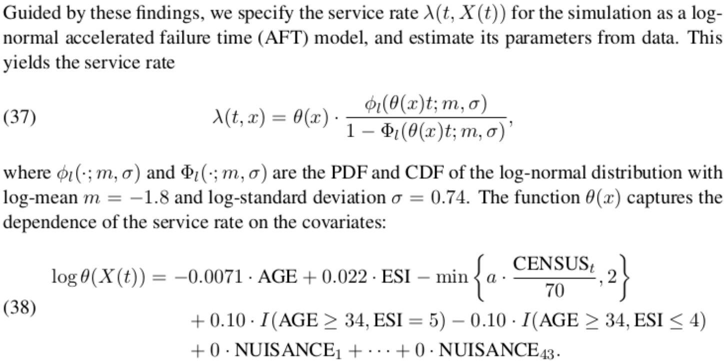
Simulations



5.1. Service rate. The service rate model used in the simulation is based upon a service time dataset from the ED of an academic hospital in the United States. The dataset contains information on 86,983 treatment encounters from 2014 to early 2015. Recorded for each encounter was: Age, gender, Emergency Severity Index (ESI)[11], time of day when treatment in the ED ward began, day of week of ED visit, and ward census. The last one represents the total number of occupied beds in the ED ward, which varies over the course of the patient's stay. Hence it is a time-dependent variable. Lastly, we also have the duration of the patient's stay (service time).

Use some exploratory analysis to find a suitable, realistic functional form of the hazard.

Simulations



Guided by these findings, we specify the service rate $\lambda(t, X(t))$ for the simulation as a log-normal accelerated failure time (AFT) model, and estimate its parameters from data. This yields the service rate

$$\lambda(t,x) = \theta(x) \cdot \frac{\phi_l(\theta(x)t; m, \sigma)}{1 - \Phi_l(\theta(x)t; m, \sigma)}, \tag{37}$$

where $\phi_l(\cdot; m, \sigma)$ and $\Phi_l(\cdot; m, \sigma)$ are the PDF and CDF of the log-normal distribution with log-mean $m = -1.8$ and log-standard deviation $\sigma = 0.74$. The function $\theta(x)$ captures the dependence of the service rate on the covariates:

$$\begin{aligned} \log\theta(X(t)) = {} & -0.0071 \cdot \text{AGE} + 0.022 \cdot \text{ESI} - \min\left\{a \cdot \frac{\text{CENSUS}_t}{70}, 2\right\} \\ & + 0.10 \cdot I(\text{AGE} \geq 34, \text{ESI} = 5) - 0.10 \cdot I(\text{AGE} \geq 34, \text{ESI} \leq 4) \\ & + 0 \cdot \text{NUISANCE}_1 + \cdots + 0 \cdot \text{NUISANCE}_{43}. \end{aligned} \tag{38}$$

Results

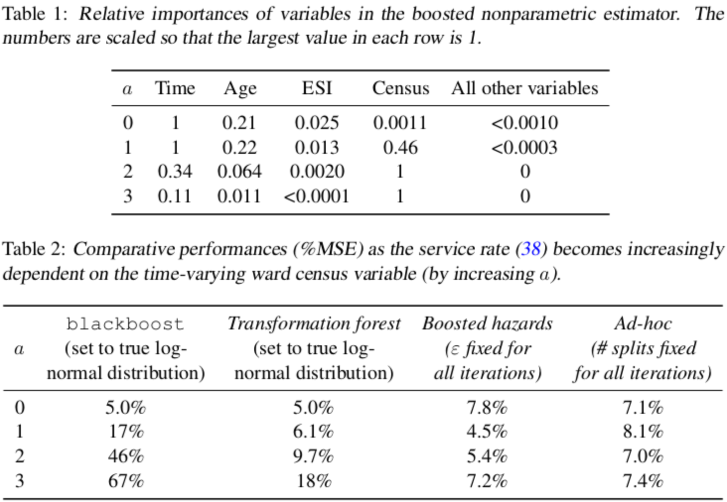


Table 1: *Relative importances of variables in the boosted nonparametric estimator. The numbers are scaled so that the largest value in each row is 1.*

a	Time	Age	ESI	Census	All other variables
0	1	0.21	0.025	0.0011	<0.0010
1	1	0.22	0.013	0.46	<0.0003
2	0.34	0.064	0.0020	1	0
3	0.11	0.011	<0.0001	1	0

Table 2: *Comparative performances (%MSE) as the service rate (38) becomes increasingly dependent on the time-varying ward census variable (by increasing a).*

a	`blackboost` (set to true log-normal distribution)	*Transformation forest* (set to true log-normal distribution)	*Boosted hazards* (*ε fixed for all iterations*)	*Ad-hoc* (*# splits fixed for all iterations*)
0	5.0%	5.0%	7.8%	7.1%
1	17%	6.1%	4.5%	8.1%
2	46%	9.7%	5.4%	7.0%
3	67%	18%	7.2%	7.4%

Summary

- First (?) general method for non-parametric conditional hazard estimation
- Finding the proper domain for the log-likelihood and an integral representation for its derivative
- Should discuss the Markov-like assumption and the time/measurement splitting