RUHR
ECONOMIC PAPERS

Matthias Kaeding

Efficient Bayesian Nonparametric Hazard Regression

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

RWI

#850

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, e-mail: W.Leininger@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Ronald Bachmann, Prof. Dr. Torsten Schmidt, Prof. Dr. Manuel Frondel, Prof. Dr. Ansgar Wübker
RWI, Phone: +49 (0) 201/81 49-213, e-mail: presse@rwi-essen.de

Editorial Office

Sabine Weiler
RWI, Phone: +49 (0) 201/81 49-213, e-mail: sabine.weiler@rwi-essen.de

Ruhr Economic Papers #850

Responsible Editor: Volker Clausen

ISSN 1864-4872 (online) – ISBN 978-3-86788-985-8

Ruhr Economic Papers #850

Matthias Kaeding

Efficient Bayesian Nonparametric Hazard Regression

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

RWI

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/86788985
ISSN 1864-4872 (online)
ISBN 978-3-86788-985-8

Matthias Kaeding[1]

Efficient Bayesian Nonparametric Hazard Regression

Abstract

We model the log-cumulative baseline hazard for the Cox model via Bayesian, monotonic P-splines. This approach permits fast computation, accounting for arbitrary censorship and the inclusion of nonparametric effects. We leverage the computational efficiency to simplify effect interpretation for metric and non-metric variables by combining the restricted mean survival time approach with partial dependence plots. This allows effect interpretation in terms of survival times. Monte Carlo simulations indicate that the proposed methods work well. We illustrate our approach using a large data set of real estate data advertisements.

JEL-Code: C11, C14, C41

Keywords: Bayesian survival analysis; nonparametric modeling; penalized spline: restricted mean survival time

May 2020

1 Matthias Kaeding, RWI and UDE. – *I gratefully acknowledge the support of Philipp Breidenbach, Rilana Decker, Alexander Haering, Christoph Hanck, Sandra Schaffner and Anna Werbeck for helpful comments.* - *All correspondence to: Matthias Kaeding, RWI, Hohenzollernstr. 1/3, 45128 Essen, Germany, e-mail: kaeding@rwi-essen.de*

1 Introduction

In economic, epidemiological and engineering applications, the Cox proportional hazards model is the benchmark for survival analysis. However, nonparametric modeling strategies for the Cox model do not scale up to large data sets. This paper aims to alleviate this problem by speeding up computation. The baseline hazard $h_0(t)$ is the key concept for the Cox model, it gives the instantaneous rate of failure at t, conditional on survival until t and covariate values of zero. We propose to model the log-integrated baseline hazard via Bayesian, monotonic penalized B-splines. As we can evaluate the likelihood analytically, and due to the benefits of Bayesian P-splines, our approach holds five key advantages: (1) Fast, automatic computation. (2) Exact likelihood calculation. (3) Accounting for arbitrary censoring. (4) Inclusion of nonparametric components. (5) Easier effect interpretation in regards to survival times, not hazard rates.

Most Bayesian non- or semiparametric approaches use a flexible model for some functional of the baseline hazard: Fernandez, Rivera, and Teh (2016) use a Gaussian process, Hennerfeind, Brezger, and Fahrmeir (2006) use P-splines for the (log) baseline hazard. Because the likelihood is usually not analytically available under this strategy, numerical integration is necessary, introducing approximation error and slowing down inference. There are approaches where this does not apply, as the likelihood is analytically available: Dykstra and Laud (1981) use the extended gamma process prior, Nieto-Barajas and Walker (2002) use a Markov increment prior. Gelfand and Mallick (1995) use a mixture of Beta densities, Kalbfleisch (1978) uses the gamma process prior, Cai, Lin, and Wang (2011) and Lin et al. (2015) use monotone regression splines for left- or right censored data. Zhou and Hanson (2018) use a Bernstein polynomial prior for arbitrary censored data. In a frequentist context, Zhang, Hua, and Huang (2010) use monotone B-splines for interval censored data, Royston and Parmar (2002) use natural cubic splines for left- or right censored data. However, these approaches are either computationally expensive, not flexible enough or only cover special cases, which does not apply to the estimation strategy proposed here.

The paper is structured as follows: section 2 gives the modeling approach, section 3 details inference. Section 4 shows a simulation study, section 5 applies the methods to real estate data. Section 6 concludes.

2 Hazard regression model

In hazard regression, the modeling of survival times is of interest, for instance unemployment durations or time until death. A non-negative random variable T with density $s(t)$ and survival function $S(t) = P(T > t)$ represents survival time. The hazard rate $h(\cdot)$ is the conditional density of T, given that $T > t$, so that

$$h(t) = s(t|T > t) = s(t)/S(t).$$

It holds that

$$S(t) = \exp(-H(t)), \text{ where } H(t) \coloneqq \int_0^t h(u)du$$

is the cumulative (or integrated) hazard, so that h uniquely determines T.

Under interval censoring, we observe data

$$\mathcal{D} = \{(y_i, \boldsymbol{x}_i), i = 1, \ldots, N\},$$

where $\boldsymbol{x}_i$ is a covariate vector and $y_i = [t_i^-, t_i^+)$ denotes the interval containing the true survival time. Left censoring is a special case with lower bound $t_i^- = 0$, right censoring is a special case with upper bound $t_i^+ = \infty$. By convention, we write $t_i^- = t_i^+ = t_i$, for an uncensored survival time.

The benchmark model for survival times is the semiparametric Cox model (Cox, 1972) with conditional survival function

$$S(t_i|\boldsymbol{z}_i, \boldsymbol{\alpha}) = \exp(-\exp\left(\log(H_0(t_i)) + \boldsymbol{z}_i^\top \boldsymbol{\alpha}\right)),$$

where $H_0(t_i)$ is the unspecified cumulative baseline hazard

$$H_0(t_i) = \int_0^{t_i} h_0(u)du,$$

with baseline hazard h_0. In a nonparametric setting, the model includes nonlinear effects. We partition each $\boldsymbol{x}_i$ into vectors $\boldsymbol{z}_i$ (linear effects) and $\boldsymbol{v}_i$ (nonlinear effects). Define the linear predictor

$$\xi_i = \xi(\boldsymbol{x}_i, t_i) \coloneqq f_0(t) + f_1(v_{i1}) + \cdots + f_R(v_{iR}) + \boldsymbol{z}_i^\top \boldsymbol{\alpha},$$

where $\boldsymbol{\alpha}$ is a vector of regression coefficients, $f_0(t_i) \coloneqq \log H_0(t_i)$ is the log-cumulative hazard and $f_1, \ldots, f_R$ are functions. Let $\boldsymbol{f}_r = (f_r(v_{1r}), \ldots, f_r(v_{N,r}))^\top$ denote the vector of function evaluations for f_r and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^\top$ denote the linear predictor vector, than we can write

$$\boldsymbol{\xi} = \boldsymbol{Z}\boldsymbol{\alpha} + \sum_{r=0}^{R} \boldsymbol{f}_r,$$

where $\boldsymbol{Z}$ is a design matrix. We can write the survival function as

$$S(t_i|\xi_i) = \exp(-\exp(\xi_i)).$$

We model the log-cumulative baseline hazard f_0 via monotonic, penalized B-splines (P-splines). As Hennerfeind, Brezger, and Fahrmeir (2006), we model $f_1, \ldots, f_R$ via P-splines. The basic idea of P-splines is to model a function f_r by a weighted sum of B-spline basis functions $B_{r1}, \ldots, B_{r,J}$, augmenting the loss function with a penalty controlling the smoothness of the estimated function. Hence,

$$f_r(v) = \sum_{j=1}^{J} \beta_{rj} B_{rj}(v) \text{ for } r = 0, \ldots, R,$$

where $\beta_{r1}, \ldots, \beta_{r,J}$ are regression coefficients associated with the function f_r, see figure 1. Given a knot vector $\boldsymbol{k}_r \in \mathbb{R}^m$, a B-spline $B_{rj}(v) = B^l_{rJ}(v)$ of order $l = 1$ is the function

$$B^1_{rj}(v) \coloneqq I[v \in [k_{r,j-1}, k_{r,j})],$$

where $I[\text{condition}]$ equals one if the condition is met and zero otherwise. See De Boor et al. (1978) for a rigorous introduction to B-splines. We assume that $\boldsymbol{k}_r$ is equally spaced from the minimum to the maximum of a covariate v_r. Then B-splines of order $l > 1$ are defined recursively[1] as

$$B^l_{rj}(v) = w^l_{rj} B^{l-1}_{rj}(v) + (1 - w^l_{r,j+1}) B^{l-1}_{r,j+1}(v), \text{ for } j = 1, \ldots, J,$$

with $J = m + l - 2$ and

$$w^l_{rj} \coloneqq \frac{x - k_{rj}}{(l-1)h},$$

where h is the spacing between the knots. We assume that J and l are equal for $f_0, \ldots, f_R$. One usually sets l to the smallest value where the smoothness of the estimated function is satisfactory, a good default is $l = 4$. Define the design matrix $\boldsymbol{B}_r \in \mathbb{R}^{N \times J}$ with i, jth element $B_{rj}(v_{ir})$. Then we can write each vector of function evaluation as $\boldsymbol{f}_r = \boldsymbol{B}_r \boldsymbol{\beta}_r$ and represent the linear predictor vector $\boldsymbol{\xi}$ compactly as

$$\boldsymbol{\xi} = \boldsymbol{Z}\boldsymbol{\alpha} + \sum_{r=0}^{R} \boldsymbol{B}_r \boldsymbol{\beta}_r.$$

Because B-splines vanish outside a domain spanned by $l - 1 + 2$ knots (see figure 1), the matrices $\boldsymbol{B}_0, \ldots, \boldsymbol{B}_R$ are sparse. Hence the computation of the linear predictor vector $\boldsymbol{\xi}$ is fast.

[1] Due to this recursive definition, for $l > 1$, the knot vector needs to be extended with additional outer knots defined analogous to $\boldsymbol{k}_r$.

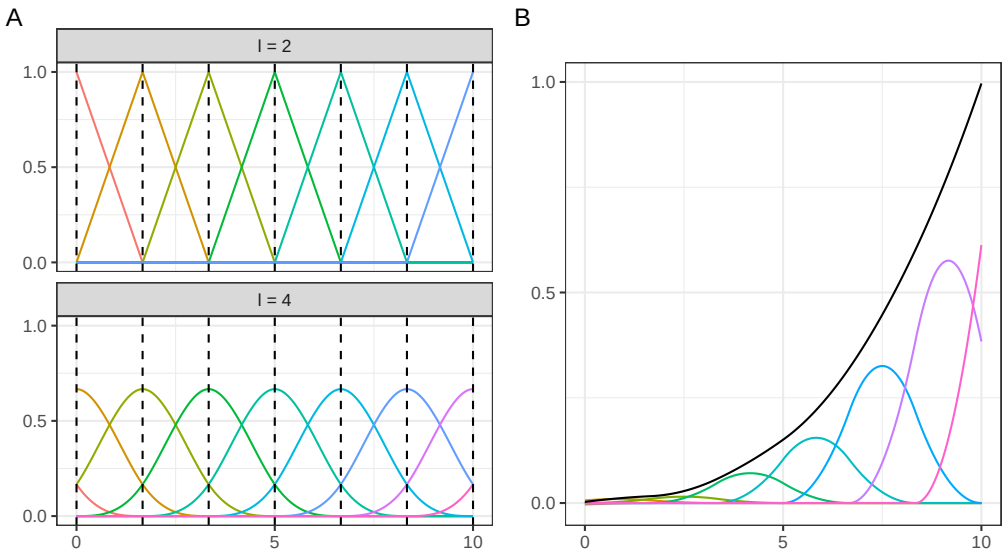

Figure 1: A: B-splines for varying order. Dotted, vertical lines mark the knots. B: Function obtained by weighted sum of B-splines. The red line is the estimated function, given by the sum of the scaled basis functions below, here giving a monotone estimate.

2.1 Priors

The flexibility of the B-splines basis increases with m, the number of knots, which determines J, the number of basis functions. For a large number of knots, the B-spline fit approaches a rough interpolation of the data which is usually undesired behaviour. Varying m on the fly changes the number of parameters, complicating inference. Using penalization, one can use fixed, large m, say $m = 30$ (so that we obtain a flexible fit) and control the smoothness of the estimated function by a single parameter penalizing unsmooth function estimates. See figure 2 for a demonstration. In a Bayesian context, this is handled by the prior distribution of $\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_R$ and the associated penalty parameters. As a result, we can directly obtain precision measures for function estimates from the posterior distribution. Furthermore, inference is automatic, in the sense that no post-processing such as cross validation is necessary. Let Δ^d be the difference operator of order d, defined recursively by

$$\begin{aligned} \Delta^1 \beta_{rg} &:= \beta_{rg} - \beta_{r,g-1}, \\ \Delta^d &:= \Delta^1 \Delta^{d-1} \text{ for } d > 1. \end{aligned}$$

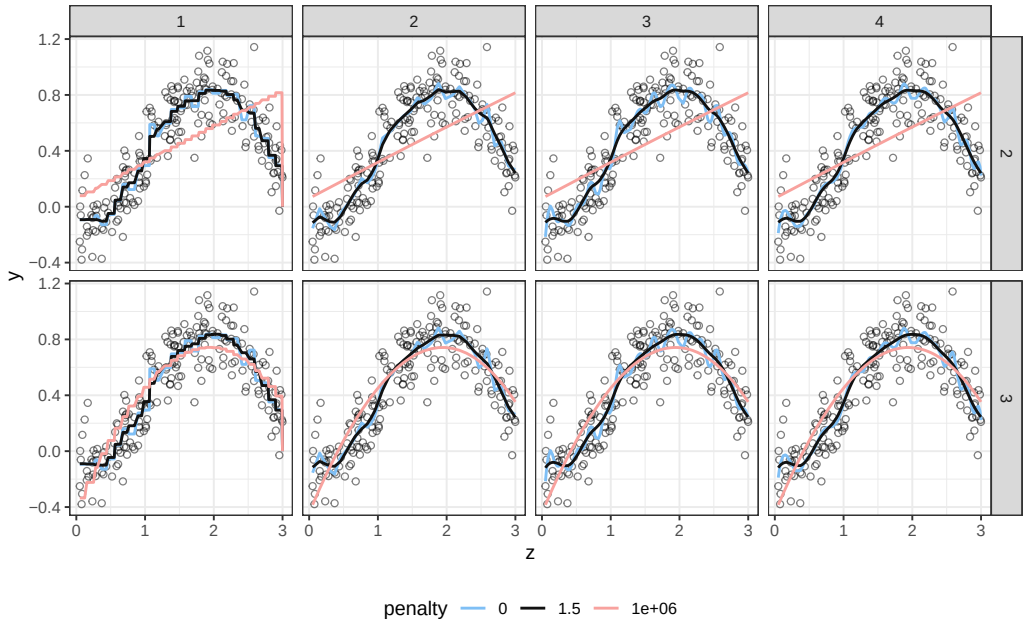

Figure 2: Influence of order l, difference order d and penalty parameter on function estimates. Data is simulated via $y_i \sim N(\sin(z_i)\log(z_i + 0.5), 0.15^2)$. Rows are varying values of d, columns are varying values of l. Lines are the estimated function under varying penalty parameter. Without a penalty, the estimated function is very unsmooth, for a large penalty the function estimate approaches a polynomial of degree $d - 1$.

The curve fitting literature uses the squared dth derivative of the estimated function as smoothness penalty. Eilers and Marx (1996) show that

$$\lambda_r \sum_{j=d}^{J} (\Delta^d \beta_{rj})^2, \tag{1}$$

approximates this smoothness penalty. As such, λ_r controls the smoothness of the estimated function $\hat{f}_r$. For $\lim_{\lambda_r \to \infty} \lambda_r$ the estimated function approaches a polynomial of degree $d - 1$. Increasing d results in smoother estimates. A value of $d > 3$ is rarely used. We use $d = 3$ as default option and assume that d is the same for $\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_R$. We use the prior distribution from Lang and Brezger (2004), who base their prior on (1). Here

$$\Delta^d \beta_{rj} \sim N(0, \tau_r^2) \text{ for } j > d,$$

so that for instance

$$\beta_{rj} = \beta_{r,j-1} + e_{rj}, \text{for a difference of order } d = 1,$$
$$\beta_{rj} = 2\beta_{r,j-1} - \beta_{r,j-2} + e_{rj} \text{ for a difference of order } d = 2,$$
$$\beta_{rj} = 3\beta_{r,j-1} - 3\beta_{r,j-2} + \beta_{r,j-3} + e_{rj} \text{ for a difference of order } d = 3.$$

with $e_{rj} \sim N(0, \tau^2)$. A high τ_r, indicating an unsmooth function, is associated with a low λ_r. Parameters $\beta_{01}, \ldots, \beta_{0d}, \beta_{11}, \ldots, \beta_{1d} \ldots \beta_{R,d}$ are assigned a flat prior $p(\cdot) \propto 1$. Let $\boldsymbol{D}_d \in \mathbb{R}^{(J-d)\times J}$ denote a matrix representation of Δ^d, so that element j of $\boldsymbol{D}_d\boldsymbol{\beta}_r$ is $\Delta^d\beta_{rj}$ and $\boldsymbol{\beta}_r^\top \boldsymbol{K}_d\boldsymbol{\beta}_r = \sum_{j=d}^{J}(\Delta^d\beta_{rj})^2$, where $\boldsymbol{K}_d = \boldsymbol{D}_d^\top \boldsymbol{D}_d$ is the penalty matrix. For instance, for $d = 2$, we have

$$\boldsymbol{D}_2 = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix}.$$

and

$$\boldsymbol{K}_2 = \begin{bmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{bmatrix}.$$

We can write the prior for $\boldsymbol{\beta}$ as

$$p(\boldsymbol{\beta}_r|\tau_r) \propto \exp(-\frac{1}{2\tau_r^2}\boldsymbol{\beta}_r^\top \boldsymbol{K}_d\boldsymbol{\beta}_r). \tag{2}$$

Because $\boldsymbol{K}_d$ is a sparse band matrix with range $d + 1$, we can exploit sparse matrix operations to compute the quadratic form $\boldsymbol{\beta}_r^\top \boldsymbol{K}_d\boldsymbol{\beta}_r$ in (2).

Some adjustments are necessary for modeling the log-cumulative baseline hazard f_0 via P-splines. Because the cumulative baseline hazard is defined on $[0, t_i]$, the knot vector is a sequence from 0 to the largest $t_i^+ < \infty$. To achieve a monotonic function estimate for the log-cumulative baseline hazard[2], we restrict the prior (2) to non-decreasing vectors, resultant in a monotonic function estimate as Brezger and Steiner (2008) show:

$$p_0(\boldsymbol{\beta}_0|\tau_0) \coloneqq p(\boldsymbol{\beta}_0|\tau_0)I[\beta_{0,1} \leq \beta_{0,2} \leq \cdots \leq \beta_{0,J}]. \tag{3}$$

[2]One might also model the cumulative baseline hazard, this involves an additional positivity restriction on $\boldsymbol{\beta}_0$. We tried this but the HMC sampler converged slowly.

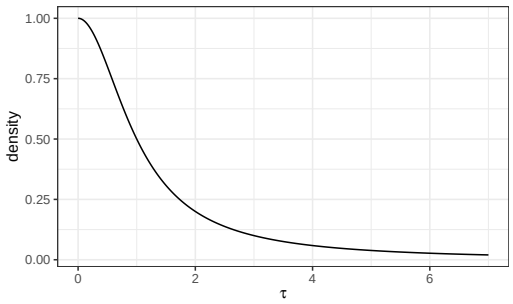

Figure 3: Cauchy prior for τ with $\phi = 1$.

We can extend this to further nonlinear effects if a monotonic function estimate for $f_1, \ldots, f_R$ is desired. We assign a flat prior to regression coefficients $\boldsymbol{\alpha}$ associated with linear effects. For positive scale parameters such as τ_r, a popular[3] choice is an inverted gamma prior, see for instance Hennerfeind, Brezger, and Fahrmeir (2006) or Kneib and Fahrmeir (2007). We follow Gelman (2006), who recommends a half Cauchy prior instead:

$$p(\tau_r|\phi_r) \propto I[\tau_r > 0](1 + (\tau_r/\phi_r)^2)^{-1} \text{ for } r = 0, \ldots, R,$$

with low scale parameter $\phi_r = 1$ as default option. This puts most prior mass on smooth functions, i.e. those with low τ_r. Due to the heavy tails of the Cauchy distribution, τ_r may be large, resultant in less smooth function estimates if the data demands it. We estimate τ_r from the data, so that the parameter adjusts to the number of B-splines.

2.2 Likelihood construction

We use P-splines to model the log-cumulative baseline hazard, so that

$$\log H_0(t) = f_0(t) = \sum_{j=1}^{J} \beta_{0j} B_{0j}(t) \text{ and}$$

$$h_0(t) = \frac{dH_0(t)}{dt} = \exp\big(f_0(t)\big) df_0(t)/dt.$$

Because the derivative of a weighted sum of B-splines is

$$\frac{d\sum_{j=1}^{J} \beta_{0j} B_{0j}^{l}(t)}{dt} = h^{-1} \sum_{j=2}^{J} B_{0j}^{l-1}(t) \Delta^1 \beta_{0j}, \tag{4}$$

[3]The popularity of the inverted gamma prior is probably due to its convenience under a Gaussian model, so that it has become somewhat of a default.

the baseline hazard is analytically available, resulting in a tractable likelihood. Because the computation of (4) involves lower order B-splines, the computational advantages of B-splines carry over to the computation of the baseline hazard. The likelihood is $L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{N} L_i$, with likelihood contributions $L_1, \ldots, L_N$ accounting for censoring. Each likelihood contribution is the probability $P(t_i \in y_i|\xi_i)$, except for uncensored survival times. Here the likelihood contribution is the density $h(t_i|\xi_i)S(t_i|\xi_i)$, so that:

$$L_i = \begin{cases} S(t_i^-|\xi_i) - S(t_i^+|\xi_i) & \text{if } t_i \text{ is interval censored,} \\ 1 - S(t_i^+|\xi_i) & \text{if } t_i \text{ is left censored,} \\ S(t_i^-|\xi_i) & \text{if } t_i \text{ is right censored,} \\ h(t_i|\xi_i)S(t_i|\xi_i) & \text{if } t_i \text{ is uncensored.} \end{cases}$$

We need to compute the log-likelihood $\mathcal{L}$ for model evaluation and to sample from the posterior distribution via Hamiltonian Monte Carlo. There are some convenient shortcuts for the computation of $\mathcal{L}$. Let $\mathcal{S}$ denote the set of all uncensored observations. Define the vectors of totals $\boldsymbol{z}^{\mathcal{S}} \coloneqq \sum_{i \in \mathcal{S}} \boldsymbol{z}_i$ and $\boldsymbol{b}_j^{\mathcal{S}} \coloneqq \sum_{i \in \mathcal{S}} \boldsymbol{b}_r(v_{ir})$, which we have to compute only once. Then we can write the log-likelihood for the uncensored observations as the sum $\xi^{\mathcal{S}} + \sum_{i \in \mathcal{S}} \eta_i$, where

$$\xi^{\mathcal{S}} \coloneqq \boldsymbol{\alpha}^\top \boldsymbol{z}^{\mathcal{S}} + \sum_{r=0}^{R} \boldsymbol{\beta}_r^\top \boldsymbol{b}_r^{\mathcal{S}} \tag{5}$$

is the sum of the linear predictor vector and η_i is defined as

$$\eta_i \coloneqq \log\left(\frac{df_0(t_i)}{dt_i}\right) - \exp(\xi_i).$$

The computational cost of $\xi^{\mathcal{S}}$ does not grow with the cardinality of $\mathcal{S}$. However, this does not hold for the computation of $\sum \eta_i$, but computation of ξ_i is fast due to the sparsity of the involved vectors. This applies to the contributions of censored observations as well: Here, the likelihood contributions depend on the value of the linear predictor, with aforementioned computational advantages. For instance, the likelihood contribution is $\log\left(S(t_i^-|\xi_i)\right) = -\exp(\xi_i)$ for a right censored survival time.

3 Inference

We use the probabilistic programming language Stan (Carpenter et al., 2017) to sample from the posterior distribution

$$p(\boldsymbol{\theta}|\mathcal{D}) = L(\boldsymbol{\theta}|\mathcal{D}) p_0(\boldsymbol{\beta}_0|\tau_0) \prod_{r=1}^{R} p(\boldsymbol{\beta}_r|\tau_r) p(\tau_r|\phi_r). \tag{6}$$

For point estimation we use the posterior mean, for interval estimation we use the 0.025 and 0.975 quantile. A nice feature of simulation based Bayesian inference is the option to obtain uncertainty measures directly for functions of parameters from the samples of the parameters, e.g. for $\exp(f_0) = H_0$.

Stan implements the No-U-Turn sampler for Hamiltonian Monte Carlo (Hoffman and Gelman, 2014). This sampler converges quickly for high dimensional posterior distribution of correlated parameters as for the problem at hand. It is fully automatic and allows easy use of non-conjugate prior distributions such as the Cauchy prior for τ_r, unlike a Gibbs sampler. Furthermore, Stan supports sparse matrix operations[4].

For models with nonparametric components the means of the function $f_0, \ldots, f_r$ are not identified. For instance

$$\xi_i = f_0(t_i) + f_1(v_{1i})$$

is equivalent to

$$\xi_i^\star = f_0^\star(t_i) + f_1^\star(v_{1i})$$

with $f_0^\star(t_i) = f_0(t_i) + c$ and $f_1^\star(v_{1i}) = f_1(v_{1i}) - c$, so that the mean of f_0 and f_1 is not identifiable. As such, we need to impose constraints or the sampler would not converge. We use the decomposition from Kneib and Fahrmeir (2007) for P-spline regression coefficients:

$$\boldsymbol{\beta}_r = \boldsymbol{X}_r^{unpenalized}\boldsymbol{\beta}_r^{unpenalized} + \boldsymbol{X}_r^{penalized}\boldsymbol{\beta}_r^{penalized}, \text{ for } r = 1 \ldots, R,$$

with priors $p(\beta_r^{unpenalized}) \propto 1$ and $\boldsymbol{\beta}_r^{penalized} \sim N(0, \tau_r^2)$. As figure 2 shows, the vector $\boldsymbol{\beta}_r^{unpenalized}$ captures the unpenalized polynomial of degree $d-1$ in f_r. The first column of $\boldsymbol{X}_r^{unpenalized}$ is a vector of ones, so that the parameter $\beta_{r,1}^{unpenalized}$ represents the mean of f_r. Deleting each vector of ones is comparable to imposing a zero mean constraint. The log-cumulative baseline hazard f_0 sets the global mean, so that we can sample $\boldsymbol{\beta}_0$ without further restrictions. The vector $\boldsymbol{\beta}_r^{penalized}$ is equivalent to a vector of random effects, allowing the use of specialized Stan routines such as the non-centered parameterization.

3.1 Model choice

Model choice in a Bayesian framework is an ongoing research area with several competing approaches. We use expected log predictive density criterion

[4]Stan also includes optimizing routines based on the automatic differentiation, so that the posterior mode (equivalent to penalized maximum likelihood) is also an option for point estimation, for example for frequentist inference.

(henceforth *elpd*), because it is a measure for the generalizability of a model to unknown data, which is usually the pertinent task. Vehtari and Ojanen (2012) derive the criterion in a Bayesian decision theoretic approach: Here we choose some model M which maximizes an utility function of our choice. Using the log score results in the *elpd*:

$$elpd_M \coloneqq \int \pi(\dot{y}) \log p_M(\dot{y}|\dot{\boldsymbol{x}}, \mathcal{D}) \; d\dot{y},$$

where

$$p_M(\dot{y}|\dot{\boldsymbol{x}}, \mathcal{D}) \coloneqq \int p_M(\dot{y}|\dot{\boldsymbol{x}}, \boldsymbol{\theta}) p_M(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$

is the posterior predictive distribution for a new observation $\dot{y}$ with covariates $\dot{\boldsymbol{x}}$, under model M and π is the distribution associated with the unknown data generating process. Using leave-one-out cross-validation, an estimator for the *elpd* is given by:

$$\widehat{elpd}_M = N^{-1} \sum_{i=1}^{N} \log p_M(y_i|\boldsymbol{x}_i, \mathcal{D}_{-i}), \tag{7}$$

where $\mathcal{D}_{-i}$ is the data without observation i. Vehtari, Gelman, and Gabry (2017) show how to compute (7) efficiently via Pareto smoothed importance sampling. Their method bypasses the need to compute N models, instead using log-likelihood evaluations from a single MCMC run. Magnusson et al. (2019) present a method to further speed up computation for large data sets based on subsampling.

3.2 Covariate effects

Let σ denote the follow up time and $\mu = \mu(\xi)$ denote the conditional expectation $E[T|\xi]$. If σ and the sample size are large enough so that we can precisely estimate the survival time where the baseline survivor function tends to zero, we can estimate μ via the identity

$$\mu = \int_0^\infty S(u|\xi) du. \tag{8}$$

In practice, this is rarely the case so that estimating the integral in equation (8) necessitates extrapolation. The restricted mean survival time (rmst), defined as

$$\mu^\sigma \coloneqq E[\min(T, \sigma)|\xi] = \int_0^\sigma S(u|\xi) du$$

is an alternative which avoids extrapolation beyond the follow up time and bypasses the need to interpret effects in terms of the hazard rate (Stensrud et al., 2018). Because researchers can interpret the restricted mean survival time as the average survival time until σ, the rmst has attracted much attention as a measure for covariate effects, see for instance Chen and Tsiatis (2001), Royston and Parmar (2011) and Zhao et al. (2012). We use numerical integration with the trapezoid rule to estimate the integral in (8), where we split up the integral at $\min(t_1^-, \ldots, t_N^-)$, to avoid extrapolation[5]:

$$\widehat{\mu}^\sigma(\xi) = \frac{t_1^\star}{2}(1 + S(t_1|\xi) + \frac{h}{2}(S(t_1|\xi) + S(\sigma|\xi)) + h\sum_{k=2}^{K-1}\big(S(t_k^\star|\xi) + S(t_{k+1}^\star|\xi)\big),$$

with spacing $h = \sigma/(K-1)$ and K control points $t_k^\star = \min(t_1^-, \ldots, t_N^-) + (k-1)h$. We can estimate the restricted mean survival time of one observation via

$$\widehat{\mu}^\sigma(\xi_i) = Q^{-1}\sum_{q=1}^{Q}\widehat{\mu}^\sigma(\xi_i^q),$$

where the superscript $q = 1 \ldots Q$ denotes the qth draw from the posterior distribution.

The most important application of the restricted survival time is the estimation of a binary treatment effect, the comparison between outcomes μ_1^σ (treatment) and μ_0^σ (control). Most commonly this the simple difference $\mu_1^\sigma - \mu_0^\sigma$. However, inference for other forms such as ratios can easily be done in a Bayesian framework (Imbens and Rubin, 2015). A unit level treatment effect W_i is the comparison of μ_1^σ and μ_0^σ for unit i. We estimate W_i via $\widehat{W_i} = Q^{-1}\sum_{q=1}^{Q} W_i^q$. An easy-to-interpret scalar measure is the average treatment effect (ATE), which we estimate via

$$\widehat{ATE} = N^{-1}\sum_{i=1}^{N}\widehat{W_i}.$$

We propose to combine partial dependence plots (Friedman, 2001) with the restricted mean survival to simplify effect interpretation for metric covariates: Say we are interested in the effect of the metric covariate v_k. The basic idea of partial dependence plots is to compute the restricted mean survival time, marginalizing over all parameters and covariates except v_k. Let

[5]This might be problematic if the observed minimum is large. In this case extrapolating H_0 or $\log H_0$ might be preferable.

$\dot{\xi}_i$ denote the linear predictor for unit i where we set the value of v_{ik} to $\dot{v}_k$. Then we create a partial dependence plot for the covariate v_k by computing

$$(NQ)^{-1} \sum_{i=1}^{N} \sum_{q=1}^{Q} \hat{\mu}^{\sigma}(\dot{\xi}_i^q)$$

over a grid of control points $\dot{v}_{k,1}, \ldots, \dot{v}_{k,C}$, plotting the result with the associated posterior interval. We can do this for variables which we model by a linear or a nonlinear component. Partial dependence plots may also be used for non-metric variables.

A related concept is the marginal survival function. Often one is interested in a global average of the survival function. For the Cox model, a simple solution is the baseline survival function, which is the survival function conditional on all covariates taking the value zero. However, this might be nonsensical or require an unwanted transformation of the covariates to achieve interpretability. The marginal survival function allows marginalization over the covariates and parameters, we compute it via

$$\hat{S}_{marginal}(t) = (NQ)^{-1} \sum_{i=1}^{N} \sum_{q=1}^{Q} S(t|\xi_i^q).$$

We can furthermore condition on specific covariates values, for instance a subgroup indicator for group differences.

To speed up computations one may use thinning, i.e. the use of every nth sample from the posterior distribution.

4 Simulation study

We investigate the performance of the presented methods under varying censoring mechanisms and sample sizes. For each censoring mechanism, we simulate 50 data sets each with sample size $N = 100, 200, 500, 1000, 2000$. Survival times are additive Weibull distributed with hazard where

$$\begin{aligned} h(t_i|\boldsymbol{x}_i, \boldsymbol{\alpha}) &= h_0(t_i) \exp(\boldsymbol{z}_i^\top \boldsymbol{\alpha} + f_1(v_{1i}) + f_2(v_{2i})), \\ h_0(t_i) &= t_i^5 + 2\sqrt{t_i}, \\ f_1(v_{i1}) &= \sin(4v_{i1}) \text{ and } f_2(v_{i2}) = 0.5[\cos(5v_{i2}) - 1.5v_{i2}]. \end{aligned}$$

Figure 4 shows the baseline hazard and involved functions. The baseline hazard is bathtub shaped, exemplifying a shape that is hard to capture by

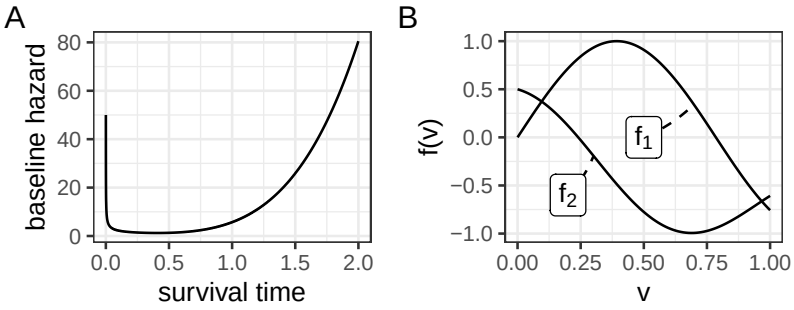

Figure 4: Functions involved in the simulation study. A: Additive Weibull baseline hazard. B: Functions f_1 and f_2.

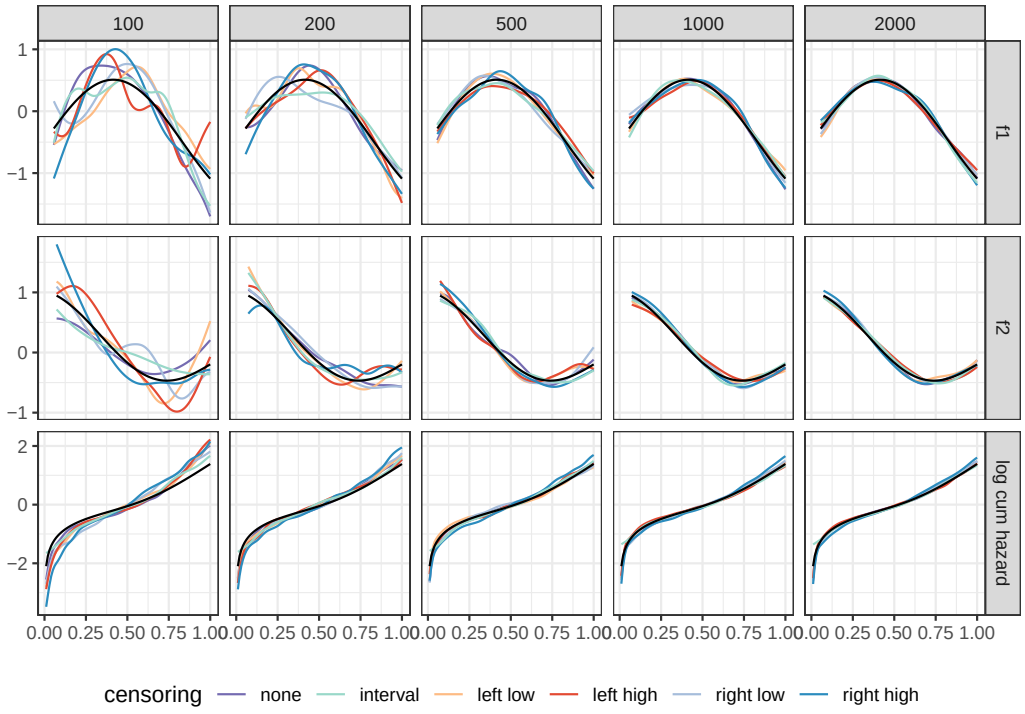

Figure 5: Representative examples by sample size (columns), variable (rows). The color of the lines is mapped to the combination of the type of censoring with the fraction of censored survival times. For instance, right low means 20% of the survival times are right censored. The x-axis is scaled to the interval $[0, 1]$ for visualization purposes. Each example is chosen to be closest to the overall mean squared error over all iterations.

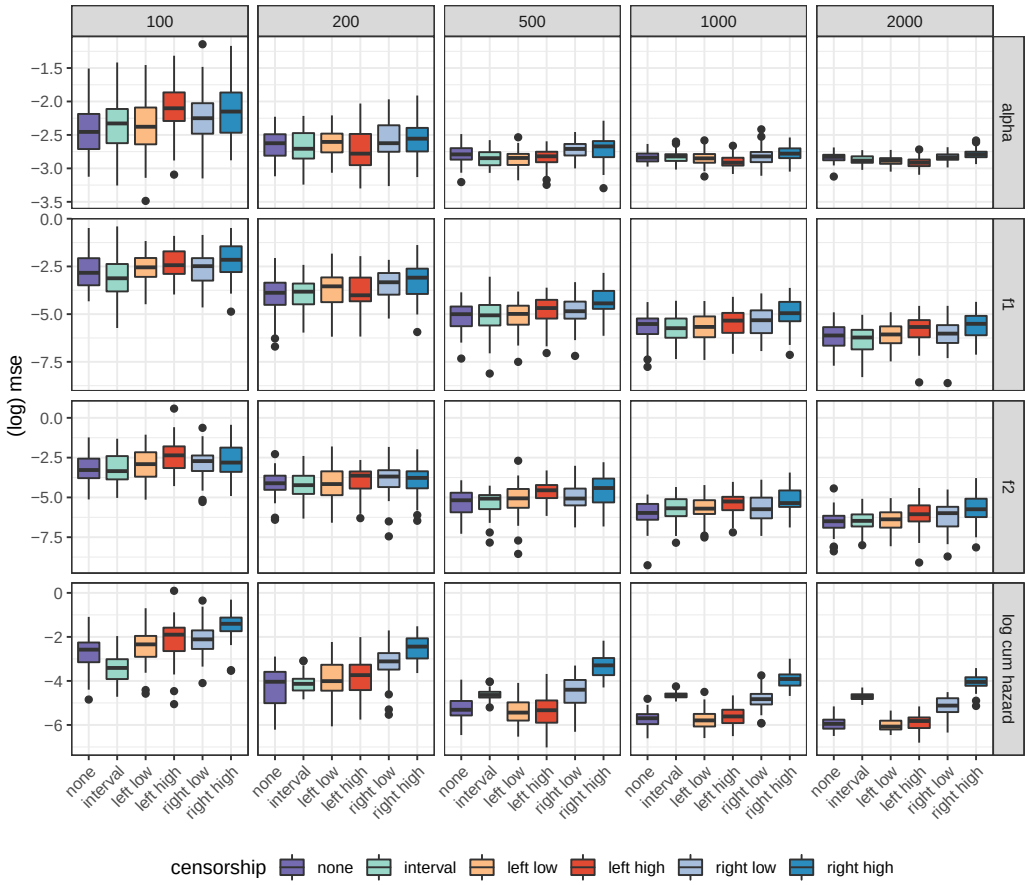

Figure 6: Boxplots showing log mean squared error by sample size (columns), variable (rows). The horizontal axis within each box corresponds to the combination of the type of censoring with the fraction of censored survival times, for instance right low means 20% of the survival times are right censored, which is furthermore mapped to the color of each boxplot.

common parametric methods yet highly relevant in practice. An example for such a mechanism is human mortality: here the hazard rate is high immediately after birth, followed by a period with low hazard, while rising in later years. Regression coefficients $\boldsymbol{\alpha}$ are equally spaced between -0.3 and 0.3, covariates $z_1, \ldots, z_5$ are standard normal, v_1 and v_2 are standard uniform. There are four variants regarding the fraction of censored observations: no censoring, low (20%) and high (40%) percent censored observations for all three censoring types, and 100% for interval censored. For all censoring types, we draw a simple random sample of the failure times and censor afterwards.

To evaluate estimates, we compute the mse of f_0, f_1 and f_2 on a grid of

C control points $e_{r,1}, \ldots, e_{r,C}$ as

$$\text{mse}(\hat{f}_r) = C^{-1} \sum_{c=1}^{C} \left(\hat{f}_r(e_{r,c}) - f_r(e_{r,c}) \right)^2$$

and the mse for $\boldsymbol{\alpha}$ as

$$\text{mse}(\hat{\boldsymbol{\alpha}}) = 5^{-1} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^\top (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}).$$

Figure 5 shows representative examples of estimates, figure 6 shows the results for the log mean squared error, henceforth log-mse. As might be expected for a complex model, estimates for f_0, f_1 and f_2 are quite imprecise for small sample sizes ($N \leq 200$). The log-mse decreases with increasing sample size. For the estimation of $\boldsymbol{\alpha}$, f_0 and f_1 the censoring mechanism does not seem to cause large differences. However, the log-mse is usually highest under a high fraction of right censored observations and lowest under no censoring and interval censoring. For estimation of the log-cumulative baseline hazard, there is a clear negative effect associated with the information loss from censoring. This holds strongest for right censoring. Under interval censoring, the log-mse for estimates of f_0 is lowest for small sample sizes, however it does not improve much for $N > 500$. Overall, the estimation strategy works as desired, given a large enough sample size. However, large fractions of right censored observations may be problematic.

5 Application: Real estate data

The data set in this application consists of survival times of real estate advertisements for flat rents. The advertisements were published on the website ImmobilienScout24 in the year 2017. For a description of the data set and data access see Boelmann and Schaffner (2018). The survival time is the number of days an advertisement is online. While there are several reasons why an advertisement may be taken offline, we assume that in the majority of cases someone rented the flat. To demonstrate inference for a treatment effect, we estimate a hazard model for the city states Bremen and Hamburg, where we create a balanced data set using coarsened exact matching (Iacus, King, and Porro, 2012). There are 16480 observations in Hamburg, of which 953 are right uncensored. In Bremen, the are 7356 observations, of which 466 are right censored. We include several covariates in the model, see table 1 for descriptive statistics: (1) The residual from a hedonic regression of rent price per square metre on a set of control variables such as age of the house.

Table 1: Descriptive statistics

	Mean		Std. deviation	
Variable	Bremen	Hamburg	Bremen	Hamburg
days online (uncens.)	21.239	16.189	36.595	26.778
days online (cens.)	42.790	23.939	89.599	45.876
censored	0.063	0.058	0.244	0.233
commission	0.018	0.023	0.132	0.151
missing entries	5.232	4.130	2.432	2.098
rent residual	-0.480	1.443	1.942	2.494

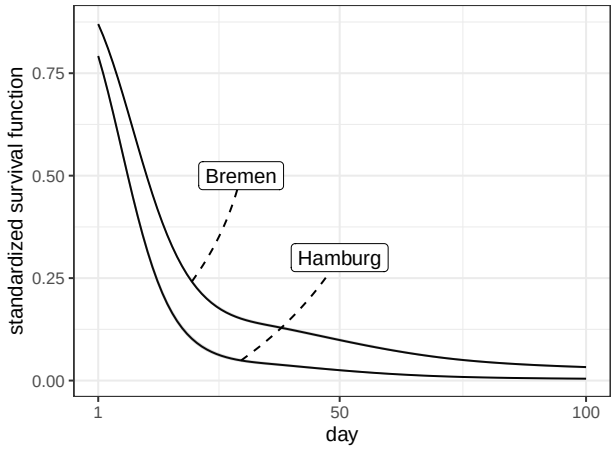

Figure 7: Standardized survival curves for Hamburg and Bremen with 0.025 and 0.975 posterior quantile.

A positive residual indicates an overpriced flat. We use a similar set of variables as Eilers (2017), extended by a spatial effect, so that the residual gives the relative price for a flat conditional covariates and the location. (2) The number of missing fields in the advertisement. (3) Binary indicators for flats requiring a broker commission and a binary indicator for Hamburg, representing a treatment effect in our analysis. We define the individual treatment W_i effect as difference between restricted mean survival times with follow up time $\sigma = 100$ days.

Figure 7 shows the marginal survival function for Hamburg and Bremen, indicating that flats in Hamburg are rented out quicker than in Bremen. Figure 8 shows the partial dependence plots. An overtly high price is associated with an increase in the restricted mean survival time. This effect is approximately linear. An increase in the number of missing entries is associated with

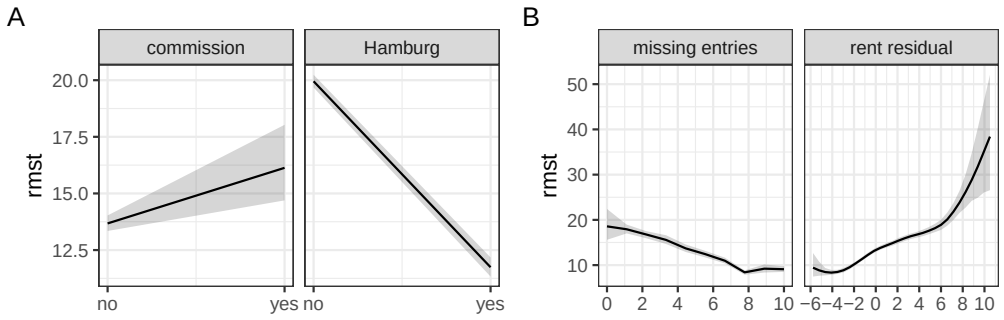

Figure 8: Partial dependence plots for restricted mean survival time, for $\sigma =$ 100 days, with 0.025 and 0.975 posterior quantiles. A: Partial dependence plots for binary covariates. B: Metric covariates.

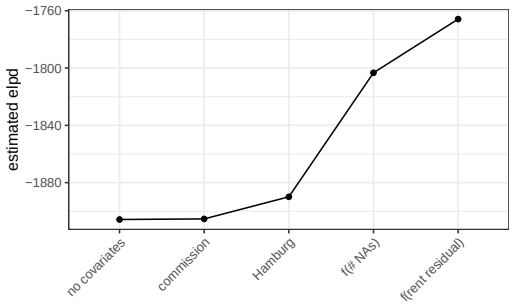

Figure 9: Estimated *elpd* for the chosen models. From left to right, the x-axis gives the element which is added to the model, starting with a model without covariates, so that the second model for the cumulative baseline hazard is $H_0(t)\exp(\beta_1 \text{commission})$, the third model $H_0(t)\exp(\beta_1 \text{commission} + \beta_2 \text{Hamburg})$ and so forth.

a decrease in restricted mean survival time. This might due to advertisement for unattractive flats, where the supplier hopes to increase attractiveness by providing more information.

For Hamburg, the estimated average treatment effect $\widehat{ATE}$ is -8.62, probably due to higher demand. The associated posterior interval is $[-8.21, -7.77]$, so that the effect is precisely estimated. Requiring a broker commission is associated with an increase in restricted mean survival time by 2.45 days. Because the share of advertisements requiring a broker commission in the data set is low (2.1 %), the associated posterior interval $[1.03, 4.19]$ is much wider.

Figure 9 shows the estimated expected log predictive density values, working up from a model containing no covariates. The effect of the covariates on

the *elpd* varies strongly between the variables. However, the order of the covariates influences this effect. For instance, including the broker commission hardly reduces the *elpd* compared to the inclusion of the number of missing entries.

6 Conclusion

This paper presents an approach for fast Bayesian hazard regression using monotonic P-splines. Because involved quantities are analytically available and we can exploit sparsity for the involved computations, this estimation strategy is computationally more efficient than existing approaches. We leverage this efficiency to simplify effect interpretation by combining partial dependence plots with the restricted mean survival time approach. We tested the proposed strategy with numerical examples: Simulations show that the approach works well, an application shows that the approach gives useful results for a large data set.

There are several extensions of this work. It might be fruitful to relax the proportional hazards assumption. This may be done by allowing interactions with survival or by allowing the baseline hazard to vary by subgroup. While using P-splines for $f_1, \ldots, f_R$ is one (very good) choice among many, we argue that monotonic P-splines are useful for $\log H_0$. For instance, one might use a Gaussian process instead, which would be feasible in our framework.

References

Boelmann, B. and S. Schaffner (2018). "FDZ Data description: Real-Estate Data for Germany (RWI-GEO-RED)-Advertisements on the Internet Platform ImmobilienScout24". In: *RWI Projektberichte*.

Brezger, A. and W. J. Steiner (2008). "Monotonic Regression Based on Bayesian P–Splines: An Application to Estimating Price Response Functions From Store-Level Scanner Data". In: *Journal of Business & Economic Statistics* 26.

Cai, B., X. Lin, and L. Wang (2011). "Bayesian proportional hazards model for current status data with monotone splines". In: *Computational Statistics and Data Analysis* 55.

Carpenter, B. et al. (2017). "Stan: A Probabilistic Programming Language". In: *Journal of Statistical Software* 76.

Chen, P.-Y. and A. A. Tsiatis (2001). "Causal Inference on the Difference of the Restricted Mean Lifetime between Two Groups". In: *Biometrics* 57.

Cox, D. R. (1972). "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.

De Boor, C. et al. (1978). *A Practical Guide to Splines*. Springer, New York.

Dykstra, R. L. and P. Laud (1981). "A Bayesian Nonparametric Approach to Reliability". In: *The Annals of Statistics* 9.

Eilers, L. (2017). "Is My Rental Price Overestimated? A Small Area Index for Germany". In: *Ruhr Economic Papers* 734.

Eilers, P. H. and B. D. Marx (1996). "Flexible smoothing with B-splines and penalties". In: *Statistical science*.

Fernandez, T., N. Rivera, and Y. W. Teh (2016). "Gaussian Processes for Survival Analysis". In: *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc.

Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine". In: *The Annals of Statistics* 29.

Gelfand, A. E. and B. K. Mallick (1995). "Bayesian Analysis of Proportional Hazards Models Built from Monotone Functions". In: *Biometrics* 51.

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models". In: *Bayesian Analysis* 1.

Hennerfeind, A., A. Brezger, and L. Fahrmeir (2006). "Geoadditive Survival Models". In: *Journal of the American Statistical Association* 101.

Hoffman, M. D. and A. Gelman (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15.

Iacus, S. M., G. King, and G. Porro (2012). "Causal Inference without Balance Checking: Coarsened Exact Matching". In: *Political Analysis* 20.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. University Press, Cambridge.

Kalbfleisch, J. D. (1978). "Non-parametric Bayesian Analysis of Survival Time Data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 40.

Kneib, T. and L. Fahrmeir (2007). "A Mixed Model Approach for Geoadditive Hazard Regression". In: *Scandinavian Journal of Statistics* 34.

Lang, S. and A. Brezger (2004). "Bayesian P-splines". In: *Journal of Computational and Graphical Statistics* 13.

Lin, X. et al. (2015). "A Bayesian proportional hazards model for general interval-censored data". In: *Lifetime Data Analysis* 21.

Magnusson, M. et al. (2019). "Bayesian leave-one-out cross-validation for large data". In: *International Conference on Machine Learning* 97.

Nieto-Barajas, L. E. and S. G. Walker (2002). "Markov Beta and Gamma Processes for Modelling Hazard Rates". In: *Scandinavian Journal of Statistics* 29.

Royston, P. and M. K. Parmar (2002). "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects". In: *Statistics in Medicine* 21.

— (2011). "The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt". In: *Statistics in Medicine* 30.

Stensrud, M. J. et al. (2018). "Limitations of hazard ratios in clinical trials". In: *European Heart Journal* 40.

Vehtari, A., A. Gelman, and J. Gabry (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and Computing* 27.

Vehtari, A. and J. Ojanen (2012). "A survey of Bayesian predictive methods for model assessment , selection and comparison". In: *Statistics Surveys* 6.

Zhang, Y., L. E. I. Hua, and J. Huang (2010). "A Spline-Based Semiparametric Maximum Likelihood Estimation Method for the Cox Model with Interval-Censored Data". In: *Scandinavian Journal of Statistics* 37.

Zhao, L. et al. (2012). "Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study". In: *Clinical Trials* 9.

Zhou, H. and T. Hanson (2018). "A Unified Framework for Fitting Bayesian Semiparametric Models to Arbitrarily Censored Survival Data, Includ-

ing Spatially Referenced Data". In: *Journal of the American Statistical Association* 113.