Conditional hazard rate estimation for right censored data

SAM EFROMOVICH[1,a]

[1] *Department of Mathematical Sciences, UTDallas, Richardson, USA,* [a] *efrom@utdallas.edu*

Theory and methodology of nonparametric sharp minimax estimation of the conditional hazard rate function of a right censored lifetime given a continuous covariate are developed. The theory, using an oracle's approach, shows how the conditional hazard and nuisance functions affect rate and constant of the mean integrated squared error (MISE) convergence. The methodology suggests a data-driven estimator matching performance of the oracle. Further, if the lifetime is independent of the covariate, the estimator recognizes that and the MISE converges with the univariate rate. Then the setting is extended to a vector of continuous and ordinal/nominal categorical predictors, and an estimator performing adaptation to smoothness and dimensionality of conditional hazard is suggested. Practical examples devoted to reducing potent greenhouse gas emissions are presented.

Keywords: Adaptation; dimension reduction; nonparametric; nuisance function; sharp minimax

1. Introduction

We begin with a familiar problem of statistical analysis of a relationship between an explanatory continuous variable (predictor) X and a continuous lifetime of interest (response) T based on a sample of size n from (X,T). Classical examples are age X at the time of a cancer surgery and how it affects time T until cancer relapse, or how credit score X affects time T until mortgage default. A traditional approach is to study regression $\mathbb{E}\{T|X=x\} := \int_0^\infty t f^{T|X}(t|x)dt$ where $f^{T|X}(t|x) := f^{X,T}(x,t)/f^X(x)$, $f^{X,T}$ is the joint density of (X,T), and $f^X(x) := \int_0^\infty f^{X,T}(x,t)dt$ is the marginal (design) density of the predictor. In this paper we use a more complicated and simultaneously more appealing approach for analysis of the relationship via the conditional hazard rate function (also referred to as the conditional failure rate in reliability theory, conditional force of mortality in actuarial science and sociology, or conditional age-specific rate in different fields of engineering and medical statistics)

$$h^{T|X}(t|x) := \frac{f^{T|X}(t|x)}{S^{T|X}(t|x)}, \quad S^{T|X}(t|x) := \int_t^\infty f^{T|X}(u|x)du. \tag{1.1}$$

If one thinks about T as a time to an event-of-interest, then given the covariate $X = x$ the quantity $h^{T|X}(t|x)dt$ represents the instantaneous likelihood that the event occurs within the interval $(t, t+dt)$ given that the event has not occurred at time t. In what follows, whenever no confusion can occur, we may refer to the conditional hazard rate function $h^{T|X}$ as conditional hazard rate, conditional hazard or simply hazard, while $H^{T|X}(t|x) := \int_0^t h^{T|X}(u|x)du$ is called the conditional cumulative hazard.

There is one important issue to be mentioned that distinguishes nonparametric estimation of the conditional hazard from other classical nonparametric problems like regression or density estimation. In classical problems an estimated function is assumed to be integrable, while a conditional hazard is not integrable, that is $\int_0^\infty h^{T|X}(t|x)dt = \infty$. Moreover, if T is supported on a finite interval $[0,w]$ then $\lim_{t\to w} h^{T|X}(t|x) = \infty$. This creates the curse of right tail.

Another traditional complication in survival analysis is that the lifetime of interest T is right censored, and then instead of observing a direct sample from (X,T) we observe a sample from triplet

$(X, V, \Delta) := (X, \min(T, C), I(T \le C))$. Here C is a censoring random variable (another lifetime), $I(\cdot)$ is the indicator, and it is assumed that T and C are conditionally independent given predictor X. This is the setting considered in the paper, and the problem is to estimate the conditional hazard rate under the mean integrated squared error (MISE) criterion.

A comprehensive discussion of hazard rates and their role in survival analysis can be found in books Cox and Oaks (1984), Klein and Moeschberger (2003), Lee and Wang (2003), Gill (2006), Fleming and Harrington (2011), Kalbfleish and Prentice (2011), Miller (2011), Moore (2016), Efromovich (2018), Ross, Prentice and Zhao (2019), Zhou (2019), Legrand (2021). Let us also present a review of some known approaches and results. Minimax estimation of the hazard rate based on direct observations of the lifetime of interest is considered in Efromovich (2016) where it is shown that the sharp constant is a functional of the survival function. Cox proportional hazards model is a classical approach that performs well in cases where an underlying model fits the Cox's model and may be too restrictive otherwise. Partially linear hazard models with varying coefficients are considered in Cai et al. (2008), additive Cox models in Lu, Lu and Li (2018). Nonparametric estimation of the conditional hazard is more flexible and technically challenging alternative, see a discussion in McKeague and Utikal (1990), Li and Doss (1995), van Keilegom and Veraverbeke (2001) where martingale and counting process techniques are employed. Using splines to estimate the conditional log-hazard rate is discussed in Kooperberg et al. (1995). The intuitively appealing approach, motivated by formula (1.1), is to use ratio of an estimate of the conditional density and an estimate of the conditional survival function. This approach is explored in Spierdijk (2008) and Gneyou (2014) where consistency of the methodology is established. Brunel, Comte and Guilloux (2011, 2013) consider adaptive and rate-optimal estimation in an interesting and more general context of marker-dependent counting processes and the case of missing indicators. Survival analysis often involves a vector of continuous and ordinal/nominal categorical covariates (predictors), see the above-presented literature as well as Kang, Lu and Zhang (2018) where additive models are used, Cui and Hanning (2019) and companion discussions are devoted to nonparametric generalized fiducial inference, the book Ross, Prentice and Zhao (2019) presents a comprehensive overview of multivariate models, interesting discussion of applications can be found in Li et al. (2022), Zhao and Feng (2020) study deep neural networks, Balan and Putter (2020) explores frailty methodology, Hothorn (2020) discusses transformation boosting machines, Huang and Su (2021) use penalized splines in concave extended linear models and study optimal rates, Emura, Sofeu and Rondeau (2022) use copula methods, deep extended hazard (DeepEH) and deep learning for partially linear Cox models are suggested in Zhong, Mueller and Wang (2021, 2022).

Overall, there is a rich literature devoted to conditional hazard estimation, including very interesting ad hoc methods, but neither methodology nor theory of sharp minimax (efficient) conditional hazard estimation are known due to complexity of the problem.

This paper uses an oracle approach to solve the problem of sharp-minimax nonparametric estimation of the conditional hazard of right censored lifetime of interest T. The oracle knows both data and information about the conditional hazard and all nuisance functions, and then develops a sharp lower bound for a minimax MISE. The oracle also develops an oracle-estimator that can be mimicked by a data-driven estimator that attains the lower bound. The oracle approach is of interest because it provides us with: (i) The optimal rate of MISE convergence; (ii) The sharp constant that depends on the so-called coefficient of difficulty; (iii) Natural nuisance functions. Let us comment on these three outcomes. It is shown that the optimal rate is defined by the effective smoothness of $h^{T|X}(t|x)$ in t and x, and this outcome resembles known results for estimation of a bivariate anisotropic regression, see Hoffmann and Lepski (2002). Notions of the coefficient of difficulty and natural nuisance functions may be less known to the reader. Following Efromovich (2018) let us first explain them using a more familiar nonparametric regression model $Y = m(X) + \sigma(X)\xi$ where X is the predictor supported on $[0, 1]$ according to the design density f^X, Y is the response, error ξ is standard normal and independent of X, $m(x)$ is the regression function of interest, and $\sigma(x)$ is the scale function. The coefficient of difficulty d tells us what functional affects the sharp constant of the MISE convergence, and for the regression problem it is $d = \int_0^1 [\sigma^2(x)/f^X(x)]dx$. Importance of knowing d is that the MISE decreases as some function of d/n, that is the larger d is, the larger sample size is needed to get a desired accuracy of estimation. Further, there are two nuisance functions f^X and $\sigma(x)$ in the regression model, but only f^X is the natural one, meaning that the oracle uses only data and f^X to construct a sharp-minimax oracle-estimator. Accordingly, the statistician needs to estimate only f^X to construct an efficient data-driven estimator that mimics the oracle. But in a multivariate additive regression both the design density and scale function are natural nuisance functions, and then both must be estimated for mimicking the oracle, see Efromovich (2013). This is why it is important to know natural nuisance functions. As we will see shortly, for the conditional hazard with univariate predictor the *natural nuisance function* is

$$p(t,x) := f^X(x)S^{C|X}(t|x)S^{T|X}(t|x) = f^X(x)S^{V|X}(t|x). \tag{1.2}$$

Accordingly, now the statistician knows that to mimic the oracle it is sufficient to estimate this single function. As we will see shortly, it is possible to estimate $p(t,x)$ directly without estimating the factors in (1.2). Furthermore, because the natural nuisance function must be estimated with a sufficient accuracy to match performance of the oracle, it is important to untie its assumed smoothness from unknown smoothness of an estimated conditional hazard. As we shall see shortly, the later is possible and this is an important theoretical result. We will continue discussion of natural nuisance functions in the Conclusion.

The content of this paper is as follows. Section 2 presents results for the case of a univariate predictor X. Here a sharp minimax lower bound for the oracle is presented, as well as a data-driven estimator that matches performance of the oracle who knows smoothness of the conditional hazard and all nuisance functions. Moreover, if T and X are independent, then the estimator recognizes this and again matches performance of the oracle. The case of a multivariate mixed (continuous and categorical) vector-predictor is considered in Section 3 and a data-driven estimator is proposed. Here the main emphasis is on the possibility of dimension reduction, and the estimator again matches performance of the oracle who knows the underlying dimensionality. Proof of the lower bound can be found in Section 4, and other proofs in the Supplementary Material. Section 5 presents practical examples devoted to waste treatments that decrease greenhouse gas emissions. Conclusions and topics for future research are presented in Section 6.

In what follows we use notation $q := q_n := \lceil \ln(n+20) \rceil$ for the minimal integer larger than $\ln(n+20)$, and similarly $s := s_n := \lceil \ln(q) \rceil$. These and other specific sequences are chosen based either on asymptotic analysis or numerical studies of proposed estimates. $I(\cdot)$ is the indicator, w denotes generic positive constants, $i \vee j := \max(i,j)$, and $i \wedge j := \min(i,j)$.

2. Conditional hazard estimation for univariate predictor

The section is relatively long, and it is worthwhile to briefly comment on its context. The final result is the data-driven Fourier series estimator (2.20), and as it is stated in Theorem 2, the estimator is sharp-minimax and performs dimension reduction whenever T and X are independent. The dimension reduction property means that the estimator's MISE attains a known in the literature optimal univariate rate of estimating h^T for the case of a directly observed sample from T. The notion of sharp-minimaxity is more involved because we need to establish the best constant and rate for minimax estimation of a bivariate $h^{T|X}$. In classical nonparametric theory this would be done by considering an oracle who may know everything and who develops a sharp-minimax lower bound for oracle-estimators over a global

Sobolev class of conditional hazards $h^{T|X}$. This approach is feasible when an underlying Fisher information does not depend on the estimand, but this is not the case for the problem at hand. Accordingly, we are using a local minimax approach by considering a class of Sobolev conditional hazards near a so-called anchor that is known to the oracle but not to the statistician. Then Theorem 1 presents a lower bound for the local minimax MISE and states that the bound is sharp and attainable by oracle-estimators. Moreover, Theorem 1 presents a blockwise-shrinking oracle estimator (2.12) that inspires construction of the above-mentioned data-driven estimator (2.20), and the estimator is studied under both local and global minimax approaches. These are the main results presented below. Accordingly, in what follows we are introducing the setting, main assumptions, global and local Sobolev classes, blockwise-shrinking estimators, and the main results are presented in Theorems 1 and 2.

Considered setting is as follows. There is a hidden sample $(X_1, T_1), \dots, (X_n, T_n)$ of size n from a pair (X, T). Here T is the lifetime of interest and X is a univariate predictor (covariate) supported on $[0, 1]$. X and T are continuous random variables with a joint density $f^{X,T}(x, t)$ supported on $[0, 1] \times [0, \infty)$. Set $f^X(x) := \int_0^\infty f^{X,T}(x, t)dt$ for the marginal (design) density of X. It is assumed that $f^X(x)$ is positive on $[0, 1]$, and this allows us to introduce the conditional hazard rate $h^{T|X}(t|x)$ defined in (1.1). We do not observe realizations $T_1, \dots, T_n$ directly because they are right censored by independent realizations $C_1, \dots, C_n$ of a censoring lifetime C. Instead we observe a sample $(X_1, V_1, \Delta_1), \dots, (X_n, V_n, \Delta_n)$ from triplet (X, V, Δ) where $V := \min(T, C)$ and $\Delta := I(T \le C)$ is the indicator of censoring.

For given constants $a \ge 0$ and $b > 0$, the aim is to estimate conditional hazard rate $h^{T|X}(t|x)$, defined in (1), over the rectangle $R := [a, a + b] \times [0, 1]$. It will be explained shortly why it is impossible to consider $b = \infty$. The used criterion for an estimator $\check{h}(t|x)$ of $h^{T|X}(t|x)$ is the mean integrated squared error (MISE) $\mathbb{E}\{\int_a^{a+b} \int_0^1 (\check{h}(t|x) - h^{T|X}(t|x))^2 dxdt\} =: \mathbb{E}\{\int_R (\check{h}(t|x) - h^{T|X}(t|x))^2 dxdt\}$. To define a basis on R, set $\psi_0(t) := b^{-1/2}$, $\psi_j(t) := (2/b)^{1/2} \cos(\pi j(t - a)/b)$, $j = 1, 2, \dots$ for elements of the cosine basis on $[a, a + b]$, and $\varphi_0(x) := 1$ and $\varphi_i(x) = 2^{1/2} \cos(\pi i x)$, $i = 1, 2, \dots$ for elements of the cosine basis on $[0, 1]$. Then $\varphi_{ji}(t, x) := \psi_j(t)\varphi_i(x)$, $(j, i) \in \{0, 1, \dots\}^2$ are elements of the tensor-product basis on R.

Now we can proceed to assumptions. The first one is necessary for consistent estimation, see an interesting discussion in Tsiatis (1975).

Assumption 1. *Given predictor X, lifetime of interest T and censoring lifetime C are independent.*

Our next assumption is about smoothness of an estimated conditional hazard rate. Here some explanation is warranted. Conditional hazard rate $h^{T|X}(t|x)$ is a bivariate function. It is a tradition in nonparametric literature to assume that an estimated bivariate function is isotropic meaning that it is as smooth (has the same number of derivatives) in t as in x, see Wasserman (2005). For some settings this assumption is reasonable, but for a conditional hazard $h^{T|X}(t|x)$ there may be a difference between smoothness in t and x. For instance, consider a location model $T = m(X) + \varepsilon$ where the random variable ε is independent of X and its hazard rate is $h^\varepsilon(t)$. Then $h^{T|X}(t|x) = h^\varepsilon(t - m(x))$. Now note that smoothness of $h^{T|X}(t|x)$ in t is defined solely by smoothness of $h^\varepsilon(t)$, while its smoothness in x depends on smoothness of $h^\varepsilon(t)$ and smoothness of $m(x)$. Accordingly, it is prudent to assume that $h^{T|X}(t|x)$ may be an anisotropic bivariate function with different smoothness in t and x.

Another important remark is that while for direct data and classical density or regression estimation problems the asymptotic MISE convergence does not depend on an underlying estimated function, see Efromovich (1999), it will be established shortly that an underlying conditional hazard affects sharp constant of the MISE convergence. Accordingly, in a lower bound for the MISE we are considering a *shrinking local anisotropic* function class of $h^{T|X}$ with an anchor $h_0(t|x)$ which may depend on n. This approach will allow us to understand how an underlying $h^{T|X}$ affects its estimation.

After these comments, let us introduce several functional classes. In what follows α_0 and α_1 are positive integer numbers that define the number of derivatives of the conditional hazard $h^{T|X}(t|x)$ in t

and x, respectively. Introduce two anisotropic Sobolev classes of bivariate functions $g(t,x)$ on R,

$$\mathcal{S}_k := \mathcal{S}_k(\alpha_0,\alpha_1,Q,a,b) := \Big\{g:\ g(t,x) := \sum_{j=k}^{\infty}\sum_{i=0}^{\infty}\theta_{ji}\varphi_{ji}(t,x),\ (t,x)\in R,$$

$$\sum_{j=k}^{\infty}\sum_{i=0}^{\infty}[1+(\pi j/b)^{2\alpha_0}+(\pi i)^{2\alpha_1}]\theta_{ji}^2 \le bQ < \infty\Big\},\quad k\in\{0,1\}. \tag{2.1}$$

The case $k=0$ implies a classical anisotropic Sobolev class $\mathcal{S}_0$, see Hoffmann and Lepski (2002). Now note that if $g\in\mathcal{S}_1$ then $\int_a^{a+b} g(t,x)dt = 0$, and this property is used to define the following shrinking local function class $\mathcal{F}_n$ of conditional hazards. Namely, introduce a continuous on R conditional hazard $h_0(t|x)$, $(t,x)\in[0,\infty)\times[0,1]$ that will be referred to as the anchor, and set

$$\mathcal{F}_n := \mathcal{F}_n(\alpha_0,\alpha_1,Q,a,b,h_0) := \Big\{h^{T|X}:\ h^{T|X}(t|x) = h_0(t|x) + g(t,x)I((t,x)\in R); \tag{2.2}$$

$$g\in\mathcal{S}_1(\alpha_0,\alpha_1,Q,a,b);\ |g(t,x)|<1/q;\ \min_{(t,x)\in R} h_0(t|x) > 1/s,\ \max_{x\in[0,1]}\int_0^{a+b} h_0(u|x)du < \infty; \tag{2.3}$$

$$\sum_{j,i=0}^{\infty}(1+j^{2\alpha_0+\beta}+i^{2\alpha_1+\beta})[\int_R h_0(t|x)\varphi_{ji}(t,x)dxdt]^2 < \infty,\ \beta>0\Big\}. \tag{2.4}$$

In what follows we denote by $\mathcal{F}_n^* := \mathcal{F}_n^*(\alpha_0,\alpha_1,Q,a,b,h_0)$ a class (2.2)-(2.3) without assumption (2.4) about the anchor. This class is used by the oracle, who knows the anchor, to establish the lower bound. To get an upper bound, the oracle may or may not use knowledge of the anchor. The latter case is of the main interest, and then the assumption (2.4) is used. Let us stress that the anchor is not necessarily an underlying conditional hazard, its the only role is to define the local function class.

Remark 1 (Sobolev classes). Using cosine bases allows us to consider aperiodic functions and propose good estimators for small samples, see Efromovich (1999, 2018). This is why Sobolev classes like (2.1), also referred to as Sobolev ellipsoids, are popular in the literature. At the same time, a Sobolev class is a subclass of a corresponding Sobolev function class with integrable squared derivatives. It is apparent that a lower bound should hold for the larger class, but to attain the lower bound a polynomial-cosine basis must be used, see Efromovich (2021). Parameters (α_0,α_1,Q) define the so-called smoothness of Sobolev's functions. The assumed $|g(t,x)|\le 1/q$ and $h_0(t|x)>1/s$ for $(t,x)\in R$ imply that the considered conditional hazards are nonnegative. Now let us comment on the local Sobolev class $\mathcal{F}_n^*$ which is used to derive a lower bound for the corresponding local minimax. It is motivated by the classical local minimax approach for estimation of a parameter θ when the Fisher information $\mathcal{J}(\theta)$ depends on the parameter. Under the parametric approach, it is assumed that θ belongs to a shrinking vicinity of the anchor parameter θ_0. The latter, under a mild assumption, yields a sharp local minimax lower bound for the mean squared error whose sharp constant depends on the Fisher information $\mathcal{J}(\theta_0)$. Then an estimator, typically maximum likelihood or Bayes, attains the sharp lower bound. The anchor h_0, the local Sobolev class $\mathcal{F}_n$ and the coefficient of difficulty $d(h_0)$, defined below in line (2.11), are the corresponding nonparametric analogs of θ_0, the shrinking vicinity of θ_0 and $1/\mathcal{J}(\theta_0)$. A nice discussion of the local minimax approach can be found in Ibragimov and Khasminski (1981) and Golubev(1991).

For the introduced function classes set

$$P := P(\alpha_0,\alpha_1,Q,b) = b\pi^{-4\alpha/(2\alpha+1)}[Q/P_1]^{1/(2\alpha+1)}P_2,\quad \alpha := [\alpha_0^{-1}+\alpha_1^{-1}]^{-1}, \tag{2.5}$$

$$P_1 := \int_{(u,\mathrm{v}):u^{2\alpha_0}+\mathrm{v}^{2\alpha_1}\le 1;u,\mathrm{v}\ge 0} ([u^{2\alpha_0}+\mathrm{v}^{2\alpha_1}]^{1/2} - [u^{2\alpha_0}+\mathrm{v}^{2\alpha_1}])d\mathrm{v}du, \tag{2.6}$$

$$P_2 := \int_{(u,\mathrm{v}):u^{2\alpha_0}+\mathrm{v}^{2\alpha_1}\le 1;u,\mathrm{v}\ge 0} (1 - [u^{2\alpha_0}+\mathrm{v}^{2\alpha_1}]^{1/2})d\mathrm{v}du. \tag{2.7}$$

Assumption 2. *Natural nuisance function $p(t,x)$, defined in (1.2), is continuous and bounded below from zero on R, and its partial derivative in x is bounded and integrable on R.*

Assumption 2 is mild, involves only first-order differentiability in x, and not tied to smoothness of the estimand $h^{T|X}$. The latter is one of the major theoretical achievements.

Now let us introduce several notations that will be used by an estimator. First of all, an estimator must take into account a possibility that T and X are independent, because then a bivariate conditional hazard $h^{T|X}(t|x)$ becomes a univariate hazard $h^T(t)$. To achieve this aim without testing hypotheses, we introduce two tensor-product arrays of indexes (frequencies) (j,i). The first one is $A_1 := \{0,1,2,\ldots\} \times \{0\}$, the second one is $A_2 := \{0,1,2,\ldots\} \times \{1,2,\ldots\}$. Note that $\varphi_{j0}(t,x) = \psi_j(t)$, and then for $(t,x) \in R$ we may write

$$h^{T|X}(t|x) = \sum_{(j,i)\in A_1} \theta_{ji}\varphi_{ji}(t,x) + \sum_{(j,i)\in A_2} \theta_{ji}\varphi_{ji}(t,x) =: h_1(t) + h_2(t,x). \tag{2.8}$$

Here $h_1(t) = h^T(t)$ and $h_2(t,x) = 0$ whenever T and X are independent. Accordingly, (2.8) explains the underlying idea of dimension reduction from a bivariate conditional hazard to a univariate hazard, and it also sheds light on the used arrays of indexes.

To solve the problem of adaptation to unknown smoothness of conditional hazard $h^{T|X}(t|x)$ in t and x, the indexes (or equivalently we may say frequencies) are grouped into special blocks, and then Fourier estimates in a block are shrunk by the same coefficient. Discussion of the introduced below specific sequences is postponed until the end of the section because it is more convenient to explain them after presenting the results. Introduce an increasing sequence of integers $b_1 = 0$, $b_2 = b_1 + 1, \ldots, b_{q+1} = b_q + 1$, and $b_{q+r} = b_{q+r-1} + \lceil (1+s^{-1})^r \rceil$ for $r = 2,3,\ldots$ Introduce blocks of integers $B_k := \{b_k, \ldots, b_{k+1}-1\}$, $k = 1,2,\ldots$, denote their cardinality as L_k, and define K_{0n} and K_{1n} as largest integers such that $b_{K_{0n}+1} \le n^{1/3}s + q + 1$ and $b_{K_{1n}+1} \le n^{1/4}s + q + 1$. Now set $\mathbf{k} := (k_1,k_2)$, recall that arrays A_ν, $\nu = 1,2$ were introduced above line (2.8), and for each array introduce its own array of blocks $B_{\nu\mathbf{k}} := \{B_{k_1} \times B_{k_2}\} \cap A_\nu$. Note that $B_{1(k_1,1)} = B_{k_1} \times \{0\}$, $B_{1(k_1,k_2)}$ is the empty set whenever $k_2 \ge 2$, $B_{2(k_1,1)}$ is always empty, and $B_{2(k_1,k_2)} = B_{k_1} \times B_{k_2}$ whenever $k_2 \ge 2$. We use this particular definition and notation for $B_{\nu\mathbf{k}}$ because they also will be used for a vector-predictor. Using this notation we can approximate (2.8) by a partial sum

$$h_n^{T|X}(t|x) := \sum_{\nu=1}^{2} \sum_{k_1=1}^{K_{0n}} \sum_{k_2=1}^{K_{1n}} \sum_{(j,i)\in B_{\nu\mathbf{k}}} \theta_{ji}\varphi_{ji}(t,x) \tag{2.9}$$

that will be estimated in (2.20). By $L_{\nu\mathbf{k}}$ we denote cardinality of $B_{\nu\mathbf{k}}$, and set $\rho_{\mathbf{k}} := 1/\ln(3 + k_1k_2)$.

The underlying idea of the proposed conditional hazard estimator is to mimic the oracle who knows both data and an underlying model. This approach allows us to develop: (i) A lower bound for the MISE; (ii) An oracle-estimator that attains the lower bound; (iii) A data-driven estimator that mimics the oracle-estimator and matches its properties. We present the corresponding results in turn.

Theorem 1 (Oracle's Sharp Lower Bound). *Let Assumptions 1 and 2 hold, the anchor $h_0(t|x)$ is continuous on R, and a sample of size n from (X,V,Δ) is available. The oracle knows data, the underlying* *function class* $\mathcal{F}_n^*$, *the anchor* h_0, *the design density* f^X, *and the conditional survival* $S^{C|X}$. *Then*

$$\inf_{\check{h}^*} \sup_{h^{T|X} \in \mathcal{F}_n^*} (n/d)^{2\alpha/(2\alpha+1)} \mathbb{E}\{\int_R (\check{h}^*(t|x) - h^{T|X}(t|x))^2 dxdt\} \geq P(1 + o_n(1)), \tag{2.10}$$

where the infimum is taken over all possible oracle-estimators $\check{h}^*$,

$$d := d(h^{T|X}, S^{C|X}, f^X, a, b) := b^{-1} \int_R \frac{h^{T|X}(t|x)}{f^X(x) S^{C|X}(t|x) S^{T|X}(t|x)} dxdt \tag{2.11}$$

is the coefficient of difficulty, and $\alpha := [\alpha_0^{-1} + \alpha_1^{-1}]^{-1}$ *is the effective smoothness of the class* $\mathcal{F}_n^*$. *The lower minimax bound is sharp and attainable by an oracle estimator. Further, if (2.4) holds, then the lower bound is attainable for* $h^{T|X} \in (\mathcal{F}_n \cup \mathcal{S}_0)$ *by oracle-estimators that do not use the anchor, and in particular by the following sharp-minimax oracle-estimator,*

$$\tilde{h}^*(t|x) := \sum_{\nu=1}^{2} \sum_{k_1=1}^{K_{0n}} \sum_{k_2=1}^{K_{1n}} \sum_{(j,i) \in B_{\nu\mathbf{k}}} \Big[I(\theta_{ji}^2 \geq 2qdn^{-1}) I(j \vee i \leq q)$$

$$+ \frac{\Theta_{\nu\mathbf{k}}}{\Theta_{\nu\mathbf{k}} + L_{\nu\mathbf{k}}^{-1} \sum_{(j,i) \in B_{\nu\mathbf{k}}} \mathbb{E}\{(\hat{\theta}_{ji}^* - \theta_{ji})^2\}} I(\Theta_{\nu\mathbf{k}} \geq \rho_{\mathbf{k}} n^{-1}) I(j \vee i > q) \Big] \hat{\theta}_{ji}^* \varphi_{ji}(t, x). \tag{2.12}$$

Here

$$\hat{\theta}_{ji}^* := n^{-1} \sum_{l=1}^{n} \frac{\Delta_l I(V_l \in [a, a+b]) \varphi_{ji}(V_l, X_l)}{p(V_l, X_l)} \tag{2.13}$$

is the oracle-estimator of Fourier coefficients $\theta_{ji} := \int_R h^{T|X}(t|x) \varphi_{ji}(t, x) dxdt$ *and*

$$\Theta_{\nu\mathbf{k}} := L_{\nu\mathbf{k}}^{-1} \sum_{(j,i) \in B_{\nu\mathbf{k}}} \theta_{ji}^2 \tag{2.14}$$

is the Sobolev functional.

Let us comment on results of Theorem 1. The lower bound (2.10) asserts that the MISE decreases with rate $n^{-2\alpha/(2\alpha+1)}$ which is the same as for a bivariate regression of T on a pair of predictors (X_1, X_2) and the case of directly observed data discussed in Hoffmann and Lepski (2002). This is a very positive outcome because it shows that, in a more complicated problem of conditional hazard estimation based on censored data, the rate does not slow down. On the other hand, coefficient of difficulty (2.11) is dramatically more complicated than for the regression where it is a constant that does not depend on an underlying regression function. Formula (2.11) explains how the quartet $(h^{T|X}, S^{C|X}, S^{T|X}, f^X)$, together with the interval $[a, a+b]$ of estimating the conditional hazard, affect the MISE convergence. Further, note that each of the first three functions in the quartet makes the coefficient of difficulty (2.11) larger as b increases because $\int_0^\infty h^{T|X}(t|x)dt = \infty$ and $\max(S^{C|X}(t|x), S^{T|X}(t|x)) \to 0$ as $t \to \infty$. This result also quantifies complexity of the right–tail estimation.

Now let us explain how to construct a data-driven estimator that mimics the oracle-estimator (2.12). We begin with introducing the following estimator of the natural nuisance function,

$$\hat{p}(t, x) := \max(\tilde{p}(t, x), 1/q), \quad \tilde{p}(t, x) := n^{-1} \sum_{l=1}^{n} \sum_{i=0}^{\lceil 2n^{1/3} s \rceil} I(V_l \geq t) \varphi_i(X_l) \varphi_i(x). \tag{2.15}$$

Because $\hat{p}$ is separated from zero, we can plug it in a denominator. Accordingly, we plug $\hat{p}$ in (2.13) and get the Fourier estimator

$$\hat{\theta}_{ji} := n^{-1} \sum_{l=1}^{n} \frac{\Delta_l I(V_l \in [a, a+b])\varphi_{ji}(V_l, X_l)}{\hat{p}(V_l, X_l)}. \tag{2.16}$$

Similarly we get an estimate of the coefficient of difficulty,

$$\hat{d} := \hat{d}(a,b) = n^{-1} \sum_{l=1}^{n} \frac{\Delta_l I(a \le V_l \le a+b)}{b[\hat{p}(V_l, X_l)]^2}. \tag{2.17}$$

To mimic the shrinking weights in (2.12) we calculate two statistics

$$\hat{\Theta}_{\nu\mathbf{k}} := \frac{2}{L_{\nu\mathbf{k}} n(n-1)} \sum_{(j,i)\in B_{\nu\mathbf{k}}} \sum_{1\le l<r\le n} \frac{\Delta_l \Delta_r \varphi_{ji}(V_l, X_l)\varphi_{ji}(V_r, X_r) I((V_l, V_r) \in [a, a+b]^2)}{\hat{p}(V_l, X_l)\hat{p}(V_r, X_r)}, \tag{2.18}$$

and

$$\tilde{\Theta}_{\nu\mathbf{k}} := L_{\nu\mathbf{k}}^{-1} \sum_{(j,i)\in B_{\nu\mathbf{k}}} \hat{\theta}_{ji}^2. \tag{2.19}$$

The proposed conditional hazard estimator, which mimics (2.12), is

$$\hat{h}(t|x) := \sum_{\nu=1}^{2} \sum_{k_1=1}^{K_{0n}} \sum_{k_2=1}^{K_{1n}} \sum_{(j,i)\in B_{\nu\mathbf{k}}} \Big[I(\hat{\theta}_{ji}^2 \ge 2q\hat{d}n^{-1}) I(j \vee i \le q)$$

$$+ \min(1, \hat{\Theta}_{\nu\mathbf{k}}/\tilde{\Theta}_{\nu\mathbf{k}}) I(\hat{\Theta}_{\nu\mathbf{k}} \ge \rho_{\mathbf{k}} n^{-1}) I(j \vee i > q) \Big] \hat{\theta}_{ji}\varphi_{ji}(t,x). \tag{2.20}$$

If necessary, a projection on the class of nonnegative functions can be performed as explained in Efromovich (1999). Now recall that w denotes generic positive constants.

Theorem 2 (Upper bound). *Let Assumptions 1 and 2 hold. Then the data-driven estimator (2.20) is sharp-minimax and*

$$\sup_{h^{T|X} \in (\mathcal{F}_n \cup \mathcal{S}_0)} [n/d]^{2\alpha/(2\alpha+1)}\ \mathbb{E}\{\int_R (\hat{h}^{T|X}(t|x) - h^{T|X}(t|x))^2 dx dt\} \le P(1 + o_n(1)), \tag{2.21}$$

that is the estimator not only matches performance of the oracle that knows the shrinking local class $\mathcal{F}_n$ defined in (2.2), but it also attains the oracle's lower bound over the global Sobolev class $\mathcal{S}_0$ defined in (2.1). If additionally the lifetime of interest T and the predictor X are independent and accordingly $h^{T|X} = h^T$, then the estimator attains the optimal univariate rate and

$$\sup_{h^{T|X} \in \{\mathcal{F}_n \cup \mathcal{S}_0,\ h^{T|X} = h^T\}} \mathbb{E}\{\int_R (\hat{h}^{T|X}(t|x) - h^T(t))^2 dx dt\} \le w n^{-2\alpha_0/(2\alpha_0+1)}. \tag{2.22}$$

We conclude that the proposed data-driven estimator of the conditional hazard rate function: (i) Adapts to an underlying smoothness of the conditional hazard; (ii) Matches performance of the oracle under minimal assumptions on smoothness of nuisance functions, and those assumptions are not tied to smoothness of the conditional hazard; (iii) Takes into account a possible independence between the lifetime of interest T and the predictor X when the bivariate $h^{T|X}(t|x)$ is equal to the hazard rate $h^T(t)$. In that case the estimator attains the optimal univariate rate of MISE convergence; (iv) All these desired properties are achieved without any complementary procedures like a hypothesis testing or an intensive numerical calculation.

As it was mentioned earlier, we finish this section by discussion/explanation of the sequences used in construction of the oracle-estimator (2.12). First of all, the estimator has the low-frequency part, where each Fourier estimate $\hat{\theta}^*_{ji}$ is individually hard thresholded, and the high-frequency part where the same shrinkage is applied to all Fourier estimates from a block. The low-frequency part is motivated by numerical simulations in Efromovich (1999), while the high-frequency part yields the asymptotic sharp minimax. Now let us consider these two parts and explain the involved sequences. We begin with the lower-frequency part. Recall that $q = \lceil \ln(n+20) \rceil$, dn^{-1} is the asymptotic variance of $\hat{\theta}^*_{ji}$, and accordingly the used hard thresholding is classical and it is often referred to as universal. Considering a logarithmic number of individual Fourier coefficients is also well known and supported by numerical studies, see Efromovich (1999). Now consider the high-frequency part. The weakly geometrically increasing blocks are convenient and yield a relatively small, of order $(qs)^2$, number of considered blocks, at the same time slower increasing blocks may be also considered as explained in Efromovich (1985, 1999). The choice of K_{0n} and K_{1n}, and more specifically of the sequences $b_{K_{0n}+1}$ and $b_{K_{1n}+1}$ that do not depend on chosen blocks, is motivated by analysis of the integrated squared bias $\mathcal{B}$ of not estimated Fourier coefficients. Recall that $b_{K_{0n}+1}$ is of order $sn^{1/3}$ and $b_{K_{1n}+1}$ is of order $sn^{1/4}$. We begin with explanation of the choice for K_{0n}. Consider the case when X and T are independent, and set $J_{0n} := b_{K_{0n}+1} - 1$. The proposed blockwise estimator must preseve the optimal rate $n^{-2\alpha_0/(2\alpha_0+1)}$, see (2.22). Because $\theta_{ji} = 0$ whenever $i > 0$, the integrated squared bias of interest is $\sum_{j>J_{0n}} \theta^2_{j0} \leq wJ_{0n}^{-2\alpha_0} \sum_{j>J_{0n}} j^{2\alpha_0}\theta^2_{j0} \leq w(sn^{1/3})^{-2\alpha_0} \leq ws^{-2}n^{-2\alpha_0/(2\alpha_0+1)} = o_n(1)n^{-2\alpha_0/(2\alpha_0+1)}$. This calculation sheds light on the choice of K_{0n}. Now let us explain the choice of K_{1n} which should yield $\mathcal{B} = o_n(1)n^{-2\alpha/(2\alpha+1)}$ if T and X are dependent and correspondingly $h^{T|X}$ is a bivariate function. Set $J_{1n} := b_{K_{1n}+1} - 1$, note that $J_{1n} \leq J_{0n}$, and write $\mathcal{B} \leq \sum_{i\geq 0}\sum_{j>J_{1n}}(\theta^2_{ij} + \theta^2_{ji}) \leq wJ_{1n}^{-2}\sum_{i\geq 0}\sum_{j>J_{1n}}[j^{2\alpha_0}\theta^2_{ji} + j^{2\alpha_1}\theta^2_{ij}] \leq w(sn^{1/4})^{-2} \leq ws^{-2}n^{-2\alpha/(2\alpha+1)} = o_n(1)n^{-2\alpha/(2\alpha+1)}$. What was wished to show. Also note that K_{0n} and K_{1n} are of order qs. Finally, let us comment on the blockwise-shrinking estimation. For the oracle it is ideal to use blocks of unit cardinality and shrinking weights $\theta^2_{ji}/[\theta^2_{ji} + \mathbb{E}\{(\hat{\theta}^*_{ji} - \theta_{ji})^2\}]$, this is the famous Wiener's shrinking that minimizes the MISE. However, the statistician cannot mimic this individual shrinking of each $\hat{\theta}^*_{ji}$ with sufficient accuracy. This is why blocks are used because they yield estimation of averaged shrinking coefficients. It will be explained in the proofs that there are special restrictions on cardinalities of blocks and sequences $\rho_{\mathbf{k}}$ that imply sharp minimaxity. The interested reader can find more discussion of the oracle-estimators in the Supplementary Material.

3. Estimation for a mixed multivariate predictor

We are considering the case of a vector-predictor when the model is extended to a multivariate predictor which may contain mixed (continuous and ordinal/nominal categorical) covariates. Namely, the predictor is $\mathbf{Z} := (\mathbf{X}, \mathbf{U})$ where $\mathbf{X} := (X_1, \ldots, X_\tau) \in [0,1]^\tau$ is the vector of τ continuos covariates and $\mathbf{U} := (U_1, \ldots, U_m) \in \mathcal{M}_m := \prod_{k=1}^m \{0, 1, \ldots, M_k - 1\}$ is the vector of m categorical covariates. The corresponding mixed design density $f^{\mathbf{Z}}(\mathbf{z}) = f^{\mathbf{Z}}(\mathbf{x}, \mathbf{u}) := f^{\mathbf{X}|\mathbf{U}}(\mathbf{x}|\mathbf{u})f^{\mathbf{U}}(\mathbf{u})$ is supported on $R_{\tau m} := [0,1]^\tau \times \mathcal{M}_m$. Also set $M := \prod_{k=1}^m M_k$ and introduce two hyperrectangles $R := [a, a+b] \times [0,1]^\tau$ and $R_{1\tau m} := [a, a+b] \times R_{\tau m}$.

Our main aim is to suggest an estimator whose MISE converges with the rate corresponding to the underlying smoothness and dimensionality of $h^{T|\mathbf{Z}}$ when the lifetime of interest T may depend on a subset of continuous covariates. Similarly to the case of a univariate predictor, we will use a series estimator based on a tensor-product basis defined on the set $R_{1\tau m}$ with the inner product

$$\langle g_1, g_2\rangle := M^{-1} \sum_{\mathbf{u}\in\mathcal{M}_m} \int_R g_1(t,(\mathbf{x},\mathbf{u}))g_2(t,(\mathbf{x},\mathbf{u}))d\mathbf{x}dt. \tag{3.1}$$

The corresponding squared norm is denoted as $\|g\|^2 := \langle g, g\rangle$. A convenient basis is defined as follows. For vector $\mathbf{X}$ of continuous predictors on $[0,1]^\tau$ we use the cosine tensor-product basis $\varphi_{\mathbf{i}}(\mathbf{x}) := \prod_{k=1}^{\tau} \varphi_{i_k}(x_k)$ where $\mathbf{i} := (i_1,\dots,i_\tau) \in \{0,1,\dots\}^\tau$. For a kth categorical covariate $u_k \in \{0,1,\dots,M_k-1\}$ we use a discrete trigonometric basis

$$\eta_0(u_k, M_k) = 1,\ \eta_r(u_k, M_k) := 2^{1/2}\cos(\pi(2u_k+1)r/(2M_k)),\quad r = 1,\dots,M_k-1,$$

which allows us to define the tensor-product basis $\{\eta_{\mathbf{r}}(\mathbf{u}) := \prod_{k=1}^m \eta_{r_k}(u_k, M_k), \mathbf{r}\in\mathcal{M}_m, \mathbf{u}\in\mathcal{M}_m\}$. Finally, for the inner product (3.1) we use the tensor-product basis $\{\psi_{j\mathbf{i}\mathbf{r}}(t,\mathbf{x},\mathbf{u}) := \psi_j(t)\varphi_{\mathbf{i}}(\mathbf{x})\eta_{\mathbf{r}}(\mathbf{u}) =: \psi_j(t)\varphi_{\mathbf{i}\mathbf{r}}(\mathbf{x},\mathbf{u}), (j,\mathbf{i},\mathbf{r})\in\{0,1,\dots\}^{1+\tau}\times\mathcal{M}_m\}$. Using this basis and the Parseval theorem, a conditional hazard with a finite norm may be written as a Fourier series,

$$h^{T|\mathbf{Z}}(t|\mathbf{z}) = \sum_{(j,\mathbf{i},\mathbf{r})\in\mathcal{M}_{(1+\tau)m}} \theta_{j\mathbf{i}\mathbf{r}}\psi_{j\mathbf{i}\mathbf{r}}(t,\mathbf{z}),\quad \text{where}\ \ \theta_{j\mathbf{i}\mathbf{r}} := \langle h^{T|\mathbf{Z}}, \psi_{j\mathbf{i}\mathbf{r}}\rangle,\ (t,\mathbf{z})\in R_{1\tau m}. \tag{3.2}$$

Similarly to the above-discussed univariate predictor case, for dimension reduction and adaptation to smoothness of $h^{T|\mathbf{Z}}(t|\mathbf{x},\mathbf{u})$ in continuous variables $(t,\mathbf{x})$, we define blocks of indexes $(j,\mathbf{i},\mathbf{r})$. Set $A'_0 := \{0,1,\dots\}$, and introduce tensor-product arrays of indexes $A_1 := A'_0\times\{0\}^\tau\times\mathcal{M}_m$, $A_2 := \{A'_0\times A'_0\times\{0\}^{\tau-1}\times\mathcal{M}_m\}\setminus A_1$, $A_3 := \{A'_0\times\{0\}\times A'_0\times\{0\}^{\tau-2}\times\mathcal{M}_m\}\setminus\cup_{k=1}^2 A_k, \dots, A_{1+\tau} := \{A'_0\times\{0\}^{\tau-1}\times A'_0\times\mathcal{M}_m\}\setminus\cup_{k=1}^\tau A_k$, $A_{1+\tau+1} := \{\{A'_0\}^3\times\{0\}^{\tau-2}\times\mathcal{M}_m\}\setminus\cup_{k=1}^{1+\tau}A_k$, $A_{1+\tau+2} := \{\{A'_0\}^2\times\{0\}\times A'_0\times\{0\}^{\tau-3}\times\mathcal{M}_m\}\setminus\cup_{k=1}^{1+\tau+1}A_k, \dots, A_{2^\tau} := \{\{A'_0\}^{1+\tau}\times\mathcal{M}_m\}\}\setminus\cup_{k=1}^{2^\tau-1}A_k$. The arrays are arranged in such a way that the number c_ν, $c_\nu\in\{1,2,\dots,1+\tau\}$ of considered continuous variables in array A_ν is a stepwise function increasing in $\nu = 1,2,\dots,2^\tau$, and note that continuous variables in A_ν include variable t together with $c_\nu - 1$ continuous covariates. Then similarly to (2.8) we can rewrite (3.2) as

$$h^{T|\mathbf{Z}}(t|\mathbf{z}) = \sum_{\nu=1}^{2^\tau} h_\nu(t,\mathbf{z}),\quad \text{where}\quad h_\nu(t,\mathbf{z}) = \sum_{(j,\mathbf{i},\mathbf{r})\in A_\nu}\theta_{j\mathbf{i}\mathbf{r}}\psi_{j\mathbf{i}\mathbf{r}}(t,\mathbf{z}). \tag{3.3}$$

Using blocks B_k introduced in Section 2, for each A_ν we define its own array of tensor-product blocks $B_{\nu\mathbf{k}} := A_s\cap\{\{B_{k_1}, B_{k_2},\dots,B_{k_{1+\tau}}\}^{1+\tau}\times\mathcal{M}_m\}$, $\mathbf{k} := (k_1,\dots,k_{1+\tau})$, and then denote by $L_{\nu\mathbf{k}}$ cardinality of $B_{\nu\mathbf{k}}$. Recall notation K_{0n} and set $K_{\tau n}$ to be the largest integer such that $b_{1+K_{\tau n}} \le n^{1/(2\tau+1)}$. Note that if $\tau = 1$ then K_{1n} is the same as for the univariate predictor. Also set $\rho_{\mathbf{k}} := 1/\ln(3+\prod_{\nu=1}^{1+\tau}k_\nu)$. Then following (2.9) we can approximate conditional hazard (3.3) by

$$h_n^{T|\mathbf{Z}}(t|\mathbf{z}) := \sum_{\nu=1}^{2^\tau}\sum_{k_1=1}^{K_{0n}}\sum_{k_2,\dots,k_{1+\tau}=1}^{K_{\tau n}}\sum_{(j,\mathbf{i},\mathbf{r})\in B_{\nu\mathbf{k}}}\theta_{j\mathbf{i}\mathbf{r}}\psi_{j\mathbf{i}\mathbf{r}}(t,\mathbf{z}). \tag{3.4}$$

This is the partial sum whose Fourier coefficients we intend to estimate.

Now let us introduce a considered function class. For a vector $\vec{\alpha} := (\alpha_0, \alpha_1, \dots, \alpha_\tau)$ of positive integers, whose elements represent number of derivatives of a function in its $1+\tau$ continuous variables, following Hoffmann and Lepski (2002) we define an anisotropic Sobolev function class

$$\mathcal{S}(\vec{\alpha}, Q) := \Big\{ g: \ g(t,\mathbf{z}) \geq 0 \ \text{ for } \ (t,\mathbf{z}) \in [0,\infty) \times R_{\tau m},$$

$$g(t,\mathbf{z}) = \sum_{(j,\mathbf{i},\mathbf{r}) \in \mathcal{M}_{(1+\tau)m}} \theta_{j\mathbf{ir}} \psi_{j\mathbf{ir}}(t,\mathbf{z}) \ \text{ for } \ (t,\mathbf{z}) \in R_{1\tau m},$$

$$\sum_{(j,\mathbf{i},\mathbf{r}) \in \mathcal{M}_{(1+\tau)m}} [1 + (\pi j/b)^{2\alpha_0} + \sum_{k=1}^{\tau} (\pi i_k)^{2\alpha_k}] \theta_{j\mathbf{ir}}^2 \leq bQ < \infty \Big\}, \tag{3.5}$$

and its effective smoothness

$$\alpha := \frac{1}{\sum_{k=0}^{\tau} \alpha_k^{-1}}. \tag{3.6}$$

It is known due to Hoffmann and Lepski (2002) that for this Sobolev class and nonparametric regression with $(1+\tau)$ continuous covariates the optimal rate of the oracle's MISE convergence is $n^{-2\alpha/(2\alpha+1)}$. As we will see shortly, this result can be matched for conditional hazard with $(\tau+m)$ mixed covariates under the following assumption.

Assumption 3. *Given predictor* $\mathbf{Z}$*, lifetime of interest* T *and censoring lifetime* C *are independent. An underlying natural nuisance function* $p(t,\mathbf{z}) := f^{\mathbf{Z}}(\mathbf{z}) S^{T|\mathbf{Z}}(t|\mathbf{z}) S^{C|\mathbf{Z}}(t|\mathbf{z})$ *for all* $t \in [a, a+b]$ *belongs to the Sobolev class*

$$\mathcal{S}_*(t,\tau) := \Big\{ g: \ g(t,\mathbf{z}) = \sum_{(\mathbf{i},\mathbf{r}) \in \mathcal{M}_{\tau m}} \kappa_{\mathbf{ir}}(t) \varphi_{\mathbf{ir}}(\mathbf{z}) \ \textit{for} \ \ (t,\mathbf{z}) \in R_{1\tau m},$$

$$\max_{t \in [a,a+b]} \sum_{(\mathbf{i},\mathbf{r}) \in \mathcal{M}_{\tau m}} [1 + \sum_{k=1}^{\tau} (\pi i_k)^{2\tau}][\kappa_{\mathbf{ir}}(t)]^2 < \infty, \ \ g(t,\mathbf{z}) \geq 0 \ \textit{for} \ \ (t,\mathbf{z}) \in [0,\infty) \times R_{\tau m} \Big\}. \tag{3.7}$$

Remark 2. The oracle knows the natural nuisance function p. Assumption 3 allows us to match the oracle by suggesting an appropriate estimate of p. To overcome the curse of dimensionality, the assumed Sobolev's smoothness in (3.7) depends on τ, but it is not tied to smoothness of $h^{T|\mathbf{Z}}$. Also note that the natural nuisance function is the product of three functions, and accordingly its smoothness in a continuous covariate x_i is defined by a coarsest in x_i function among the triplet $(f^{\mathbf{Z}}, S^{T|\mathbf{Z}}, S^{C|\mathbf{Z}})$. Finally, because $S^{T|\mathbf{Z}}$ is a factor in the natural nuisance function, the estimand $h^{T|\mathbf{Z}}$ is at least as smooth in $\mathbf{x}$ as $p(t,\mathbf{z})$, and accordingly (3.7) implies that in (3.5) we have $\min(\alpha_1, \dots, \alpha_\tau) \geq \tau$.

The introduced notions and notations for the case of a multivariate predictor are similar to the univariate case, and the same similarity will be observed between the proposed estimators.

We begin with introducing a plug-in sample mean Fourier estimator

$$\hat{\theta}_{j\mathbf{ir}} := n^{-1} M^{-1} \sum_{l=1}^{n} \frac{\Delta_l I(V_l \in [a, a+b]) \psi_{j\mathbf{ir}}(V_l, \mathbf{Z}_l)}{\hat{p}(V_l, \mathbf{Z}_l)} \tag{3.8}$$

and an estimator of the coefficient of difficulty,

$$\hat{d} := n^{-1}M^{-1}\sum_{l=1}^{n}\frac{\Delta_l I(V_l \in [a, a+b])}{[\hat{p}(V_l, \mathbf{Z}_l)]^2}. \tag{3.9}$$

Here $\hat{p}$ is an estimator of the natural nuisance function defined as

$$\hat{p}(t, \mathbf{z}) := \max(\tilde{p}(t, \mathbf{z}), q^{-1}), \quad \tilde{p}(t, \mathbf{z}) := n^{-1}M^{-1}\sum_{l=1}^{n}\sum_{(\mathbf{i},\mathbf{r})\in A'} I(V_l \geq t)\varphi_{\mathbf{ir}}(\mathbf{Z}_l)\varphi_{\mathbf{ir}}(\mathbf{z}), \tag{3.10}$$

where $A' := \{0, 1, \dots, n^{1/3\tau}s)\}^{\tau} \times \mathcal{M}_m$. Next, similarly to (2.19)-(2.20) we introduce two statistics

$$\hat{\Theta}_{\nu\mathbf{k}} := \frac{2}{L_{\nu\mathbf{k}}n(n-1)}\sum_{(j,\mathbf{i},\mathbf{r})\in B_{\nu\mathbf{k}}}\sum_{1\leq l_1<l_2\leq n}\prod_{\eta=1}^{2}\frac{\Delta_{l_\eta}\psi_{j\mathbf{ir}}(V_{l_\eta}, \mathbf{Z}_{l_\eta})I((V_{l_\eta} \in [a, a+b])}{\hat{p}(V_{l_\eta}, \mathbf{Z}_{l_\eta})} \tag{3.11}$$

and $\tilde{\Theta}_{\nu\mathbf{k}} := L_{\nu\mathbf{k}}^{-1}\sum_{(j,\mathbf{i},\mathbf{r})\in B_{\nu\mathbf{k}}}\hat{\theta}^2_{j\mathbf{ir}}$. The proposed conditional hazard estimator is

$$\hat{h}^{T|\mathbf{Z}}(t|\mathbf{z}) := \sum_{\nu=1}^{2^\tau}\sum_{k_1=1}^{K_{0n}}\sum_{k_2,\dots,k_{1+\tau}=1}^{K_{\tau n}}\Big[\sum_{(j,\mathbf{i},\mathbf{r})\in B_{\nu\mathbf{k}}} I(\hat{\theta}^2_{j\mathbf{ir}} > 2q\hat{d}n^{-1})I(j \vee \mathbf{i} \leq q)$$

$$+ \min(1, \hat{\Theta}_{\nu\mathbf{k}}/\tilde{\Theta}_{\nu\mathbf{k}})I(\hat{\Theta}_{\nu\mathbf{k}} > \rho_{\mathbf{k}}n^{-1})I((j \vee \mathbf{i} > q)\Big]\hat{\theta}_{j\mathbf{ir}}\psi_{j\mathbf{ir}}(t, \mathbf{z}), \tag{3.12}$$

where $j \vee \mathbf{i} := \max(j, i_1, \dots, i_\tau)$.

Theorem 3 (Multivariate mixed predictor). *Let Assumption 3 hold. Then the conditional hazard estimator (3.12) is rate-optimal over an anisotropic Sobolev class (3.5) and*

$$\sup_{h^{T|\mathbf{Z}}\in\mathcal{S}(\vec{\alpha},Q)} \mathbb{E}\{\|\hat{h}^{T|\mathbf{Z}} - h^{T|\mathbf{Z}}\|^2\} \leq wn^{-2\alpha/(2\alpha+1)}, \tag{3.13}$$

where α is the effective smoothness (3.6) of the Sobolev class. If an estimated conditional hazard $h^{T|\mathbf{Z}}$ depends only on a subvector $(X_{k_1}, \dots, X_{k_\nu})$, $\nu < \tau$ of continuous covariates, then the estimator's MISE satisfies (3.13) with α being replaced by a corresponding effective smoothness $\alpha^* := 1/[\alpha_0^{-1} + \sum_{i=1}^{\nu}\alpha_{k_i}^{-1}]$.

The rates outlined in Theorem 3 are optimal even for the oracle who knows an underlying dimensionality, and this implies that the estimator performs the desired dimension reduction.

4. Proofs

Proof of Theorem 1. We begin with the heuristic that sheds light on steps in the proof and explains new notions. Nonparametric lower bounds for minimax risks over Sobolev classes, that are defined by special restrictions on Fourier coefficients like (2.1), are established using a Bayes approach with least favorable priors for Fourier coefficiens. For the case when Fourier coefficients are observed with additive Gaussian errors, Pinsker (1980) proposed to use normal priors and developed methodology of finding sharp minimax lower bounds. If the additive Gaussian model is not appropriate, as in our problem, the methodology of Efromovich (1989) can be used which is developed for a model where Fisher information, contained in an observation about a Fourier coefficient, is constant or asymptotically approaches a constant. Then, similarly to Pinsker (1980), a normal prior may be used. To make the proof shorter and more concise, it is worthwile to use technical results of that papers. Accordingly, to get a constant Fisher information, we divide the rectangle $R := [a, a+b] \times [0, 1]$ into s^2 subrectangles of sizes $(b/s) \times (1/s)$. Then for each subrectangle its own cosine tensor-product basis and a local Sobolev class are created. Because all involved functions are smooth, for large n the local Fisher information is almost constant, and we can use the above-mentioned methodology of the proof. A pure technical issue to resolve, which complicates presented below notions and notations, is that the local hazard rate functions should be sewed together to get a Sobolev's smooth function over the R. Fortunately, this is a familiar procedure in harmonic analysis and is done using special flattop smoothing kernels. One more comment is due which explains specific notions and sequences used below. Consider a Sobolev ellipsoid with coefficients $a_{sji} := 1 + (\pi js/b)^{2\alpha_0} + (\pi is)^{2\alpha_1}$, $i, j = 0, 1, \dots$ and power bQ_s, and note that a_{1ji} and $Q_1 = Q$ define the ellipsoid (2.1). Further, assume that $\mathcal{I}$ is the constant Fisher information for a problem at hand, for instance $\mathcal{I}$ is reciprocal of the variance of Gaussian white noise in the filtering model of Pinsker (1980). Then it is known that the sharp-minimax estimator over that ellipsoid estimates only Fourier coefficients on frequencies satisfying $a_{sji} \leq a_n^*$ where a_n^* is solution of the equation $\sum_{a_{sji} \leq a_n^*} [(a_{sji}/a_n^*)^{1/2} - a_{sji}] = bQ_s n\mathcal{I}$. Further, a_n^* can be replaced by any other sequence satisfying $a_n = a_n^*(1 + o_n(1))$. Accordingly, we can evaluate a_n by replacing the sum by the integral and solving the equation $\int_{\{(x,t):\, a_{stx} \leq a_n; t,x>0\}} [(a_{stx}/a_n)^{1/2} - a_{stx}] dt dx = bQ_s n\mathcal{I}$. Using the change of variables $v = \pi s x a_n^{-1/(2\alpha_1)}$ and $u = \pi t s b^{-1} a_n^{-1/(2\alpha_0)}$, we conclude that the upper bound on considered a_{sji} may be set to

$$a_n := a_n(s, Q_s, \mathcal{I}) := [s^2 Q_s \mathcal{I} b \pi^2 P_1^{-1} n]^{2\alpha/(2\alpha+1)}, \tag{4.1}$$

where P_1 is defined in (2.6). In its turn, the a_n allows us to calculate the corresponding minimax MISE. Denote this MISE by M_n, and use the Pinsker's formula $M_n = (1/\mathcal{I}n) \sum_{a_{sji} \leq a_n} [1 - (a_{sjn}/a_n)^{1/2}]$. Then repeating approximation of the sum by a corresponding integral and using the same change of variables we get

$$M_n = P(\alpha_0, \alpha_1, s^{-4\alpha} Q_s, b)(n\mathcal{I})^{-2\alpha/(2\alpha+1)}(1 + o_n(1)). \tag{4.2}$$

Here function P is defined in (2.5). Now we can explain the last part of the heuristic. Power Q of the Sobolev class (2.1) over the rectangle R should be spreaded among the above-mentioned Sobolev subclases for s^2 subrectangles. This is done by assigning the power inversely proporional to the local Fisher information, and then the sharp-minimax oracle-estimator verifies that this approach is indeed the least favorable. Of course, to make the above-mentioned step we need to calculate local Fisher informations. The calculation will be done shortly, and now let us explain its heuristic. The mixed density (likelihood) of the triplet (X, V, Δ) is

$$f^{X,V,\Delta}(x, t, \delta) = f^X(x)[S^{C|X}(t|x) h^{T|X}(t|x) e^{-\int_0^t h^{T|X}(z|x)dz}]^\delta$$

$$\times [f^{C|X}(t|x) e^{-\int_0^t h^{T|X}(z|x)dz}]^{1-\delta}. \tag{4.3}$$

Here we used formula $S^{T|X}(t|x) = e^{-\int_0^t h^{T|X}(z|x)dz}$. The Fisher information is calculated for Fourier coefficients of a cosine tensor-product series expansion of $h^{T|X}$. In (4.3) a Fourier coefficient of $h^{T|X}$ is present in the factor $e^{-\int_0^t h^{T|X}(z|x)dz}$ and in the factor $h^{T|X}$ when $\Delta = 1$. In the factor $e^{-\int_0^t h^{T|X}(z|x)dz}$ we get integrals like $\int_0^t \cos(wju)du = (wj)^{-1} \sin(wjt) = o_j(1), j > 0$. In establishing a lower minimax bound the oracle may consider only increasing frequencies $j > q^s$, and then a straightforward calculation shows that the component of Fisher information, created by the factor $e^{-\int_0^t h^{T|X}(z|x)dz}$, vanishes as n increases. For the factor $h^{T|X}$, calculation of the corresponding component of the Fisher information is a standard calculus problem, and as we will see shortly the Fisher information is $[f^X(x)S^{C|X}(t|x)S_0^{T|X}(t|x)/h_0(t|x)](1+o_n(1))$, where (x,t) is a point from the subrectangle and $S_0^{T|X}(t|x) := \int_t^\infty h_0(z|x)dz$ is the conditional survival defined by the anchor. Note that the Fisher information is the natural nuisance function (1.2) divided by the conditional hazard rate. The final step of the proof is to show that the lower bound is attainable by an oracle-estimator. The proposed oracle-estimator (2.12) is convenient for understanding of how to mimic it by a data-driven estimator, but it is a complicated task to directly analyze its MISE. Instead, a ladder of sharp-minimax oracle-estimators is considered that leads us to the (2.12). The ladder will be explained at the beginning of the upper bound proof.

Now we proceed to the proof of the lower bound. As it was explained in the heuristic, the first step is to introduce a new function class that is a subset of $\mathcal{F}_n^*$ and is defined by local Sobolev classes on the above-mentioned subrectangles of R. Set

$$\mathcal{H}_s = \Big\{h:\ h(t|x) = h_0(t|x) + I((t,x)\in R)\sum_{r,k=0}^{s-1}\Big[g_{(rk)}(t,x) - b^{-1}\int_a^{a+b} g_{(rk)}(z,x)dz\Big],$$

$$g_{(rk)}(t,x)\in\mathcal{H}_{srk},\ \ h(t|x)\ge 0\Big\}. \tag{4.4}$$

We will define all the elements of the class shortly, and now note the following. $\mathcal{H}_{srk}$ are local classes on the subrectangles, functions $g_{(rk)}$ are smoothed, at boundaries of the (r,k)th subrectangle, Fourier series so functions $h-h_0$ belong to the Sobolev class $\mathcal{S}_1$. Also note that $\int_a^{a+b}[h(t|x)-h_0(t|x)]dt = 0$, and accordingly $S^{T|X}(t|x)$ is not changed for $t > b$.

Now we define the functions and classes in (4.4). Let $\tilde\phi(x) := \phi(n,x)$ be a sequence of flattop non-negative kernels defined on a real line such that for a given n: the kernel is zero beyond $(0,1)$, it is α_1-fold continuously differentiable on $(-\infty,\infty)$, $0\le\tilde\phi(x)\le 1$, $\tilde\phi(x) = 1$ for $2/q^2\le x\le 1-2/q^2$, and its lth derivative satisfies $\max_x|\tilde\phi^{(l)}(x)|\le Cq^{2l}$, $l=1,\dots,\alpha_1$. For instance, such a kernel may be constructed using so-called mollifiers discussed in Efromovich (1999). Then set $\tilde\phi_{sk}(x) := \tilde\phi(sx-k)$. Absolutely similarly define $\hat\phi_{sr}(t)$ only with α_1 being replaced by α_0 and the interval $[0,1]$ by $[a,a+b]$. Set $\varphi_{ski}(x) := s^{1/2}\varphi_i(sx-k)$ and $\psi_{srj}(t) := s^{1/2}\psi_j(st-br)$. For an (r,k)th subrectangle $[a+br/s, a+b(r+1)/s]\times[k/s,(k+1)/s]$ (as we will see shortly, common boundaries are irrelevant for the proof and simplify formulae), $0\le r,k\le s-1$, set $\phi_{srk}(t,x) := \hat\phi_{sr}(t)\tilde\phi_{sk}(x)$, $\varphi_{srkji}(t,x) := \psi_{srj}(t)\varphi_{ski}(x)$, and

$$g_{[rk]}(t,x) := \sum_{(j,i)\in D(s,r,k)}\eta_{srkji}\varphi_{srkji}(t,x),\ \ g_{(rk)}(t,x) := g_{[rk]}(t,x)]\phi_{srk}(t,x). \tag{4.5}$$

Here $D(s,r,k) := \{(j,i): n^{2\alpha/(2\alpha+1)}s^{-4}\le a_{sji}\le a_n(s,Q_{srk},\mathcal{I}_{srk});\ j,i = q^s+1, q^s+2,\dots\}$, a_n is defined in (4.1), $Q_{srk} := Q(1-1/s)(\overline{I_s^{-1}}I_{srk})^{-1}$, $I_{srk} := f^X(k/s)S^{C|X}(a+br/s|k/s)S_0^{T|X}(a+br/s|k/s)/h_0(a+br/s|k/s)$, $\overline{I_s^{-1}} = \sum_{r,k=0}^{s-1}(1/I_{srk})$, and $S_0^{T|X}(t|x) := e^{-\int_0^t h_0(z|x)dz}$. Let us comment on (4.5). Fourier series $g_{[rk]}$ is the subrectangle's analog of Sobolev's additive perturbations g in (2.2). As it was explained in the heuristic, we need to sew the perturbations from adjoint subrectangles, and this is why the smoothed at the boundaries $g_{(rk)}$ is introduced. Also note that, as was explained in the heuristic, functions $g_{[rk]}$ have only intermediate frequencies. Further, $\mathcal{I}_{srk}$ denotes the local Fisher information, and note that the Sobolev's power is spreaded inversely proportional to local Fisher information.

Now we can define local function classes $\mathcal{H}_{srk}$ used in (4.4). Using defined in (4.5) local Fourier series $g_{[rk]}$ and their smoothed at boundaries versions $g_{(rk)}$, set

$$\mathcal{H}_{srk} := \Big\{ g_{(rk)}(t,x): \; g_{(rk)}(t,x) = [\sum_{(j,i)\in D(s,r,k)} \eta_{srkji}\varphi_{srkji}(t,x)]\phi_{srk}(t,x),$$

$$\sum_{(j,i)\in D(s,r,k)} a_{sji}\eta^2_{srkji} \le bQ_{srk}, \;\; s^{-1} < n\eta^2_{srkji} \le s, \;\; |g_{[rk]}(t,x)|^2 \le s^4 \ln(n) n^{-2\alpha/(2\alpha+1)} \Big\}. \tag{4.6}$$

Note that this is a very special subclass of vanishing, as n increases, and smoothed at boundaries local Sobolev functions with intermediate frequencies and additional restrictions on underlying Fourier coefficients.

Our next step is to verify that $\mathcal{H}_s \subset \mathcal{F}^*_n$ for all large n. Definition of the flattop kernel implies that for $(t,x) \in R$ the difference $h^{T|X}(t|x) - h_0(t|x)$ is α_1-fold differentiable with respect to x and α_0-fold differentiable with respect to t. Second, let us verify that for $h \in \mathcal{H}_s$ this difference belongs to $\mathcal{F}^*_n$. We begin with testing derivatives with respect to x. To simplify formulas, in several next lines we are using $\psi^{(l)}(t,x) := \partial^l \psi(t,x)/\partial x^l$ for the lth derivative. By the Leibniz rule $(g_{[rk]}(t,x)\phi_{srk}(t,x))^{(\alpha_1)} = \sum_{l=0}^{\alpha_1} \mathbf{C}_l^{\alpha_1} g_{[rk]}^{(\alpha_1-l)}(t,x)\phi^{(l)}_{srx}(t,x)$ where $\mathbf{C}_l^{\alpha_1} := \alpha_1!/((\alpha_1-l)!l!)$. For $0 < l \le \alpha_1$ we have $(\phi^{(l)}_{srk}(t,x))^2 \le C(s(\ln(n))^2)^{2l}$, and for $g_{(rk)} \in \mathcal{H}_{srk}$ we can write (recall that w denotes generic positive constants that may be different even in the same line)

$$\int_R [g_{[rk]}^{(\alpha_1-l)}(t,x)\phi^{(l)}_{srk}(t,x)]^2 dxdt \le ws^{2l} \ln^{4l}(n) \int_{a+br/s}^{a+b(r+1)/s} \int_{k/s}^{(k+1)/s} [g_{[rk]}^{(\alpha_1-l)}(t,x)]^2 dxdt$$

$$\le ws^{2l}\ln^{4l}(n) \sum_{(j,i)\in D(s,r,k)} i^{2(\alpha_1-l)}\eta^2_{srkji} \le wq^{4\alpha_1+1} \max_{(j,i)\in D(s,r,k)} \frac{i^{2(\alpha_1-l)}}{a_{ji}} Q_{srk} = o_n(1)q^{-2}Q_{srk}.$$

In the last equality the definition of $\mathcal{H}_{srk}$ and the relation $\min_{(j,i)\in Dsrk} i > q^s$ were used. Absolutely similarly we get

$$\int_R [(\partial^{\alpha_0-1} g_{[rk]}(t,x)\partial t^{\alpha_0-1})(\partial^l \phi_{srk}(t,x)/\partial t^l)]^2 dxdt = o_n(1)q^{-2}Q_{srk}.$$

Next, using the Parseval identity we get for $g_{(rk)} \in \mathcal{H}_{srk}$,

$$\int_R \Big[g^2_{[rk]}(t,x) + (\partial^{\alpha_1} g_{[rk]}(t,x)/\partial x^{\alpha_1})^2 + (\partial^{\alpha_0} g_{[rk]}(t,x)/\partial t^{\alpha_0})^2 \Big] \phi^2_{srk}(t,x) dxdt$$

$$\le \sum_{(j,i)\in D(s,r,k)} a_{ji}\eta^2_{srkji} \le bQ_{srk}.$$

The above-obtained relations, together with Proposition 1 of Efromovich (2001) and $\sum_{r,k=0}^{s-1} Q_{srk} = Q(1-s^{-1})$, yield that for all $g_{(rk)} \in \mathcal{H}_{srk}$ we have $\sum_{r,k=1}^{s-1} g_{(rk)}(t,x) \in \mathcal{F}^*_n(\alpha_0,\alpha_1,Q(1-s^{-1}),a,b,h_0)$. To finish verification of the relation $\mathcal{H}_s \subset \mathcal{F}^*$, we are left with analysis of the function (recall the definition (4.4))

$$g_s(x) := b^{-1} \sum_{r,k=1}^{s-1} \int_a^{a+b} g_{(rk)}(z,x)dz, \;\; x \in [0,1]. \tag{4.7}$$

Using relation $\int_{a+br/s}^{a+b(r+1)/s} g_{[rk]}(t,x)dt = 0$ and definition of the flattop kernel $\phi_{srk}(t,x)$, we conclude that g_s belongs to $\mathcal{F}_n^*(\alpha_0, \alpha_1, o_n(1)s^{-1}, a, b, h_0)$.

This concludes verification of the assertion that for all large n we have $\mathcal{H}_s \subset \mathcal{F}_n^*$. Accordingly, in the proof of the lower minimax bound we can replace $\mathcal{F}_n^*$ by $\mathcal{H}_s$. The latter allows us to consider a parametric estimation problem with conveniently chosen parameters η_{skrji}.

Now we are evaluating from below a component of the studied MISE over a specific subrectangle $R_{srk} := [a + br/s, a + b(r+1)/s] \times [k/s, (k+1)/s]$. Recall that $g_s(x)$ is defined in (4.7), introduce $\hat{h}(t|x) =: h_0(t|x) + \tilde{h}(t|x)$, and write for $h(t|x) \in \mathcal{H}_s$,

$$\int_{R_{srk}} (\hat{h}(t|x) - h(t|x))^2 dxdt = \int_{R_{srk}} (\tilde{h}(t|x) - g_{(rk)}(t,x) + g_s(x))^2 dxdt$$

$$\geq (1 - s^{-1}) \int_{R_{srk}} (\tilde{h}(t|x) - g_{[rk]}(t,x))^2 dxdt - s \int_{R_{srk}} [g_{[rk]}(t,x)(1 - \phi_{srk}(t,x)) + g_s(x)]^2 dxdt$$

$$= (1 - s^{-1}) \int_{R_{srk}} (\tilde{h}(t|x) - g_{[rk]}(t,x))^2 dxdt + o_n(1) s q^{-1/2} n^{-2\alpha/(2\alpha+1)}.$$

To continue, set $\tilde{\eta}_{srkji} := \int_{R_{srk}} \tilde{h}(t|x)\varphi_{srkji}(t,x)dxdt$ for Fourier coefficients of the oracle-estimate $\tilde{h}$. Using definitions (4.4) and (4.6) of function classes $\mathcal{H}_s$ and $\mathcal{H}_{srk}$, and the Bessel inequality we continue

$$\sup_{h \in \mathcal{F}_n^*} \mathbb{E}\left\{\int_R (\hat{h}(t|x) - h(t|x))^2 dxdt\right\} \geq \sup_{h \in \mathcal{H}_s} \mathbb{E}\left\{\int_R (\hat{h}(t|x) - h(t|x))^2 dxdt\right\}$$

$$= \sup_{h \in \mathcal{H}_s} \sum_{r,k=0}^{s-1} \mathbb{E}\left\{\int_{R_{srk}} (\hat{h}(t|x) - h(t|x))^2 dxdt\right\}$$

$$\geq (1 - s^{-1}) \sum_{r,k=0}^{s-1} \sup_{h \in \mathcal{H}_{srk}} \sum_{(j,i) \in T(s,r,k)} \mathbb{E}\left\{(\tilde{\eta}_{srkji} - \eta_{srkji})^2\right\} + o_n(1) n^{-2\alpha/(2\alpha+1)}$$

$$=: (1 - s^{-1}) \sum_{r,k=0}^{s-1} A_{rk} + o_n(1) n^{-2\alpha/(2\alpha+1)}. \tag{4.8}$$

Our next step is to find lower bounds for the parametric risks A_{rk}. To do that, we bound a minimax risk by a Bayes one. Introduce an array of independent normal random variables ζ_{srkji} with zero mean and variance $\tau^2_{srkji} := n^{-1}(1 - 3\gamma_n)\max(\gamma_n, \min(\gamma_n^{-1}, [a_n(s, Q_{srk}, I_{sk})/a_{sji}]^{1/2} - 1))$, here $\gamma_n < 1/3$ tends to zero as slow as desired and the function a_n is defined in (4.1). Using these variables we introduce a stochastic process $\bar{h}^*(t|x)$, defined as the studied $h(t|x) \in \mathcal{H}_s$ but with random ζ_{srkji} used in place of deterministic η_{srkji}. The idea and methodology of considering such a stochastic process goes back to Pinsker (1980) and Efromovich (1989). To use the methodology we need to establish several technical results and then calculate Fisher information. First, following along lines of establishing line (S.18) in Pinsker (1980) we get

$$\mathbb{P}\Big((\bar{h}^*(t|x) - h_0(t|x)) \in \mathcal{H}_s\Big) \geq 1 - |o_n(1)|. \tag{4.9}$$

Second, using $\tau^2_{srkji} \le sn^{-1}$ we get

$$\sum_{(j,i)\in D(s,r,k)} \sup_{t,x}[\tau_{srkji}\varphi_{srkji}(t,x)]^2 \le ws^3 n^{-2\alpha/(2\alpha+1)}.$$

Third, introducing a similarly defined stochastic process $\bar{h}^*_{[rk]}$, and using the above-presented calculations together with Theorem 6.2.3 of Kahane (1985) we get

$$\mathbb{P}\Big(\sup_{(t,x)\in R} |\bar{h}^*_{[rk]}(t|x)|^2 \le s^4 qn^{-2\alpha/(2\alpha+1)}\Big) \ge 1 - |o_n(1)|s^{-2}.$$

Now, following Efromovich (1989), we are calculating Fisher information in the triplet (X, V, Δ) about parameter η_{srkji} of the conditional hazard rate $h^{T|X} \in \mathcal{H}_s$. Recall that formula for the likelihood is presented in (4.1), and it yields that the Fisher information is the sum of two components for $\Delta = 1$ and $\Delta = 0$. We evaluate the components in turn. In what follows we use notation $\mathbb{E}_{srkji,0}$ to stress that the expectation is calculated using $h^{T|X}$ with $\eta_{srkji} = 0$. Write for the case $\Delta = 1$ of the uncensored lifetime of interest T,

$$I_{srkji1} := \mathbb{E}_{srkji,0}\{\Delta[\partial \ln(h^{T|X}(T|X)e^{-\int_0^T h^{T|X}(z|X)dz})/\partial\eta_{srkji}]^2\}. \tag{4.10}$$

Let us consider the derivative on the right side of (4.10),

$$\frac{\partial \ln(h^{T|X}(T|X)e^{-\int_0^T h^{T|X}(z|X)dz})}{\partial\eta_{srkji}} = \frac{\partial \ln(h^{T|X}(T|X))}{\partial\eta_{srkji}} - \frac{\partial \int_0^T h^{T|X}(z|X)dz}{\partial\eta_{srkji}} =: B_1 - B_2. \tag{4.11}$$

Note that

$$h^{T|X}(t,x) = h_0(t|x) + [g_{(rk)}(t,x) - b^{-1}I((t,x)\in R)\int_a^{a+b} g_{(rk)}(z,x)dz], \tag{4.12}$$

where

$$g_{(rk)}(t,x) := g_{[rk]}(t,x)\phi_{srk}(t,x)I(k/s \le x \le (k+1)/s)I(a + br/s \le t \le a + b(r+1)/s), \tag{4.13}$$

and $g_{[rk]}(t,x) := \sum_{(j,i)\in D(s,r,k)} \eta_{srkji}\varphi_{srkji}(t,x)$. Using these formulas we get $\partial g_{[rk]}(t,x)/\partial\eta_{srkji} = \varphi_{srkji}(t,x)$, recall that $\phi_{srk}(t,x)$ is a flattop kernel described in the beginning of the proof, $\varphi_{srkji}(t,x)$ are elements of the tensor-product cosine basis on the (r,k)th subrectangle, and then $\int_{k/s}^{(k+1)/s} \varphi_{srkji}(t,x)dx = 0$ and $\int_{a+br/s}^{a+b(r+1)/s} \varphi_{srkji}(t,x)dt = 0$. Also recall that on the considered rectangle functions $h_0(t|x)$, $f^X(x)$, $S^{C|X}(t|x)$ and $S_0^{T|X}(t|x) = e^{-\int_0^t h_0(z|x)dz}$ are continuous and $|h^{T|X}(t|x) - h_0(t|x)| \le 1/q$. Using these remarks and a straightforward calculation we get

$$\mathbb{E}_{srkji,0}\{B_1^2\} = \frac{f^X(k/s)S^{C|X}(a + br/s|k/s)S_0^{T|X}(a + br/s|k/s)}{h_0(a + br/s|k/s)}(1 + o_n(1)). \tag{4.14}$$

To evaluate the second moment of B_2, we need to make a simple remark. B_2 is the functional of the integral of $h^{T|X}$. For $j > 0$ we have $\int_0^t \cos(\pi jsz/b)dz = (\pi js/b)^{-1}\sin(\pi jst/b) = o_j(1)$, and recall that we are considering only $j > q^s$. This remark and a straightforward calculation yield that

$$\mathbb{E}_{srkji,0}\{B_2^2\} = o_n(1) \tag{4.15}$$

uniformly over considered (r,k,j,i). Using (4.10), (4.14), (4.15) and the Cauchy inequalty we conclude that $I_{srkji1} = \mathbb{E}_{srkji,0}\{B_1^2\}(1+o_n(1)) = \mathcal{I}_{srk}(1+o_n(1))$. Note that the first component of the Fisher information, calculated for the case $\Delta = 1$, is the value used in spreading the Sobolev power among the subrectangles. The reason for this is that the second component, for the case $\Delta = 0$, is

$$I_{srkji0} = \mathbb{E}_{srkji,0}\Big\{(1-\Delta)\Big[\frac{\partial \int_0^T h^{T|X}(z|X)dz}{\partial \eta_{srkji}}\Big]^2\Big\} = o_n(1), \tag{4.16}$$

where the last relation is due to (4.15). We can make an important theoretical conclusion that the oracle can use only uncensored cases, when $\Delta = 1$, for the sharp minimax estimation.

Now we are ready to finish analysis of terms A_{rk} in (4.8). The calculations (4.1) and (4.2) allow us to use results of Efromovich (1989) and get

$$\inf A_{rk} \geq (s^{-4\alpha} Q_{srk})^{1/(2\alpha+1)}(n I_{srk})^{-2\alpha/(2\alpha+1)} P(\alpha_0, \alpha_1, 1, b)(1+o_n(1)). \tag{4.17}$$

Here the infimum is over all possible nonparametric oracle-estimates of $h^{T|X}(t|x)$ considered in Theorem 1. Now we plug in values of Q_{srk} introduced at the beginning of the proof and get

$$\inf \sum_{r,k=0}^{s-1} A_{rk} \geq P n^{-2\alpha/(2\alpha+1)} s^{-4\alpha/(2\alpha+1)} \Big[\sum_{r,k=0}^{s-1} (\overline{I_s^{-1}} I_{srk})^{-1/(2\alpha+1)} I_{srk}^{-2\alpha/(2\alpha+1)}\Big](1+o_n(1)). \tag{4.18}$$

For the sum on the right side of (4.18) we may write,

$$\sum_{r,k=0}^{s-1} (\overline{I_s^{-1}} I_{srk})^{-1/(2\alpha+1)} I_{srk}^{-2\alpha/(2\alpha+1)} = (\overline{I_s^{-1}})^{-1/(2\alpha+1)} \sum_{k,r=0}^{s-1} I_{srk}^{-1} = (\overline{I_s^{-1}})^{2\alpha/(2\alpha+1)}$$

$$= (s^2/b)^{2\alpha/(2\alpha+1)}\Big[(b/s^2)\sum_{r,k=0}^{s-1} \frac{h_0(a+br/s|k/s)}{f^X(k/s) S^{C|X}(a+br/s|k/s) S_0^{T|X}(a+br/s|k/s)}\Big]^{2\alpha/(2\alpha+1)}. \tag{4.19}$$

Because all functions in the last sum are continuous, using (4.19) we continue (4.18) and get

$$\inf \sum_{r,k=0}^{s-1} A_{rk} \geq P\Big[n^{-1}\Big(b^{-1}\int_R \frac{h^{T|X}(t|x)}{f^X(x) S^{C|T}(t|x) S^{T|X}(t|x)} dx dt\Big)\Big]^{2\alpha/(2\alpha+1)}(1+o_n(1)).$$

The oracle's lower bound is proved. Proof of its sharpness can be found in the Supplementary Material.

5. Practical examples

Producing greenhouse gas, also called biogas, greengas or biomethane, occurs in nature when a diverse population of bacteria breaks down organic materials, contained in waste, landfills and livestock manure, into the biogas and a combination of solids and liquid components. Greenhouse gas contains roughly 50-70 percent methane, 30-40 percent carbon dioxide, and trace amounts of other gases. To give a perspective on gravity of the problem, according to EPA in the USA landfills and manure management are responsible for 27% of toxic methane emissions second only to 32% by natural gas and petrolium systems. Instead of escaping into the air, the potent gas can be captured, converted, and used as a clean renewable energy resource. This is what waste treatment plants do, and aerators are the important part of those systems. Moreover, an aerator fouling may force to stop the plant. This is why prediction of the fouling is an important statistical problem.

The section presents four examples of statistical analysis of diffuser fouling in wastewater aerators. We begin with a brief explanation of aeration and diffuser fouling, and then consider experiments conducted by environmental company BIFAR.

Aeration is an essential and most energy demanding process of wastewater treatment plants, and it creates opportunity for wastewater recycling and reducing environmental pollution by cutting greenhouse gas emissions produced by otherwise untreated waste, see Slavov (2017) and Albu et al. (2021). Aeration introduces bubbles of air into a wastewater that supply the required oxygen to the metabolizing microorganisms and provide mixing of waste that allows microorganisms to come into contact with the dissolved and suspended organic matter. While a diffused aeration system (aerator) is a complicated engineering mechanism, we will deal only with two of its parts: (i) Air diffuser which transfers air into wastewater; (ii) Air blower which supplies pressurized air to diffuser. The diffusers are susceptible to chemical and biological fouling that decreases supply of air and causes substantial aeration energy wastage, while the blower may break down due to pollution in poorly filtered air, see a discussion in the above-cited references and Drewnovski et al. (2019), Vinardella et al. (2020), Samuelsson et al. (2021). The environmental company BIFAR has been interested in the effect of several parameters (covariates) on the diffuser fouling time T. Below four controlled BIFAR experiments are explained in turn. In the BIFAR experiments T was right censored by end of the study or breakdown of the blower. The experiments are conducted for different diffusers and wastes, the experiments are labor-intensive and costly, and that explains relatively small sample sizes.

The following aeration terminology is used. CSS, CCR and LAP stand for concentration of suspended solids, concentration of chemical reagents, and level of air polution, respectively.

Experiment 1. BIFAR was interested in fouling time T of a new diffuser given level X of suspended solids in treated wastewater. Perfect air filtration was used to avoid breakdowns of the blower. At the same time, because for some diffusers their lifetime T was too long, BIFAR stopped an experiment when T exceeded a threshold c_0. This created a deterministic right censoring with triplet of observed variables (X, V, Δ) where $V := \min(T, c_0)$ and $\Delta = I(T \leq c_0)$. BIFAR conducted $n = 92$ experiments and among those 20 were censored. The observations, linearly rescaled by BIFAR onto unit square $[0, 1]^2$, are shown in Figure 1. With some obvious but understandable abuse of notations, in the figure and in what follows we denote by X and V the rescaled variables, and as a result if $\Delta = 0$ then $V = 1$.

Before discussion of the data and its statistical analysis, let us make the following important remark. Deterministic censoring, used by BIFAR, is not covered by the paper's theory where C is assumed to be random. On the other hand, because BIFAR is interested only in $T \in [0, 1]$, all assumptions of the theory are fulfilled. Moreover, the interested reader can double-check that the proposed estimator does not use values of V beyond an interval of interest.

Now let us return to the data shown in Figure 1. Let us look at the top diagram where only uncensored observations (pairs $(X, V) = (X, T)$) are shown by the circles. In many instances a visual analysis of a scattergram may reveal a pronounced relationship between predictor X and response T, see examples in Efromovich (2018). This is not the case here, and even our general knowledge that the lifetime should decrease as the concentration of suspended solids increases does not help to see that in the data. Also note that, at least on first glance, Fourier estimator (2.16) uses only uncensored observations, and the oracle does not use censored observations. We return to this issue shortly, and now let us look at the bottom diagram where all available observations are shown. Note how the triangles, exhibiting censored

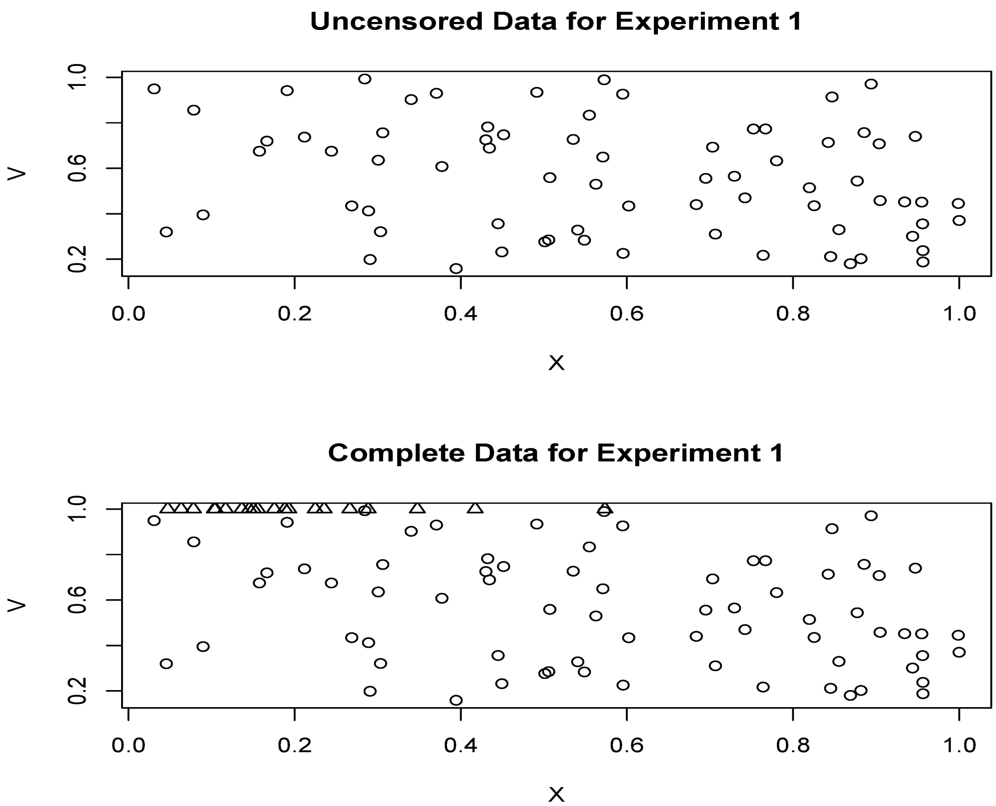

Figure 1. Data for BIFAR experiment 1. X is the CSS (concentration of suspended solids), V is the censored lifetime. Observations are rescaled by BIFAR on $[0,1]^2$. The circles show uncensored observations, that is observations with $\Delta = 1$. The triangles show censored observations whose lifetimes exceeded the BIFAR's threshold. The sample size $n = 92$, and 20 lifetimes are censored by BIFAR.

observations, help us to realize that the diffusers are not everlasting and the fouling increases when X increases. Now we can return to the issue of using only uncensored observations. This is indeed the case for the oracle, but the oracle knows the principle nuisance function p. The Fourier estimate (2.16) utilizes estimate (2.15) of p which uses censored observations.

Now let us present the estimated conditional hazard (recall that $\varphi_1(x) = 2^{1/2}\cos(\pi x)$ and it is a decreasing function on $[0,1]$),

$$\hat{h}(t|x) = 2.1 - \varphi_1(x) - \varphi_1(t)[1.3 - 0.6\varphi_1(x)], \quad (t,x) \in [0,1]^2. \tag{5.1}$$

Recall that t is the time and x is the CSS. As we see, the conditional hazard increases in t and x, and the increase in t accelerates as x increases. This outcome was accepted by BIFAR and supported by the theory of diffuser fouling.

Two comments are due. First, it may be difficult (and even impossible) to “visualize" conditional hazard in data, and accordingly one needs to use an appropriate statistical methodology and software. Second, while practitioners often prefer a polynomial formula in t and x, and this is not difficult to do

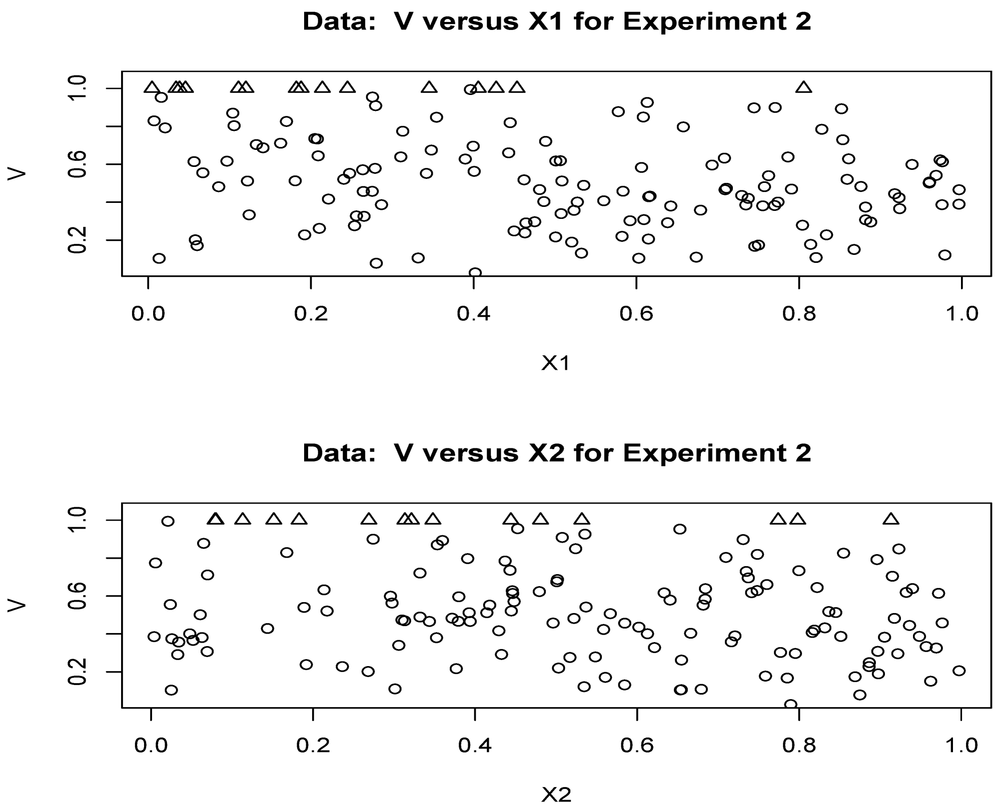

Figure 2. Data for BIFAR experiment 2. $X1$ is the rescaled CSS (concentration of suspended solids), $X2$ is the rescaled CCR (concentration of chemical reagents), V is the rescaled censored time of diffuser fouling. The circles show uncensored lifetimes T, the triangles show censored lifetimes. The sample size $n = 147$, and 15 lifetimes are censored.

by using the Legendre basis discussed in Efromovich (1999), it is the author's recommendation to use cosines because polynomials create a familiar “temptation" to extrapolate a formula beyond a studied set of variables.

Experiment 1 sheds a new light on the problem of conditional hazard estimation and the proposed solution, and it also provides a learning experience that will be useful for understanding next examples.

Experiment 2. This was a more sophisticated experiment when the BIFAR was interested in aerators for dairy wastewater, Slavov (2017). Concentration of chemical reagents in the wastewater from washing milk trucks, tanks, equipment and floor is a familiar source of diffuser fouling. The experiment was similar to the previous one only now, in addition to CSS (concentration of suspended solids) $X1 := X_1 := X$, CCR (concentration of chemical reagents) $X2 := X_2$ was another controlled covariate of interest. Collected data are shown in Figure 2 and explained in the caption. Similarly to data for experiment 1, it is not easy to visualize a relationship between the variables, especially in the bottom diagram

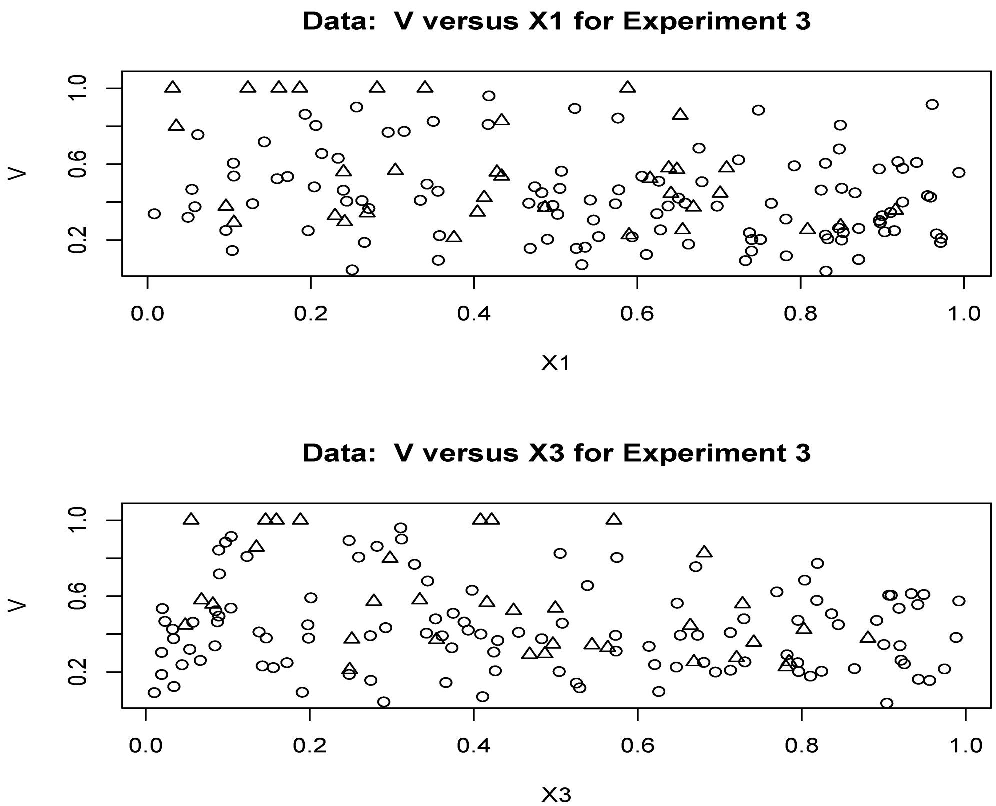

Figure 3. Data for BIFAR experiment 3. $X1 := X_1$ is the CSS (concentration of suspected solids), $X3 := X_3$ is the LAP (level of air pollution), V is the censored time of diffuser fouling. Observations are rescaled by BIFAR onto $[0,1]^3$. The circles and the triangles show uncensored and censored lifetimes, respectively. The sample size $n = 151$ and 35 lifetimes are censored.

where we see the scattergram of censored lifetimes V versus CCR. Estimated conditional hazard is

$$\hat{h}(t|x_1, x_2) = 2.7 - \varphi_1(t)[1.4 - 0.6\varphi_1(x_1) - 0.2\varphi_1(x_2)]$$

$$- 1.1\varphi_1(x_1) - 0.5\varphi_1(x_2), \quad (t, x_1, x_2) \in [0,1]^3. \tag{5.2}$$

As we see, the conditional hazard rate of diffuser fouling increases in t, x_1 and x_2 (that is in time, CSS and CCR), and the increase in t accelerates as CSS and CCR increase. The result is impressive due to the small sample size and estimating the trivariate function.

Experiment 3. Here BIFAR was interested in the effect of CSS X_1 and LAP (level of air pollution) X_3 on diffuser fouling. Specific of the experiment is that air pollution also affects lifetime of the blower, and accordingly diffuser fouling times may be randomly right censored by blower's breakdowns or deterministic censoring by end of the study. Observations are shown in Figure 3 and explained in the caption. The top diagram resembles the bottom diagram in Figure 1, and here again it is difficult to

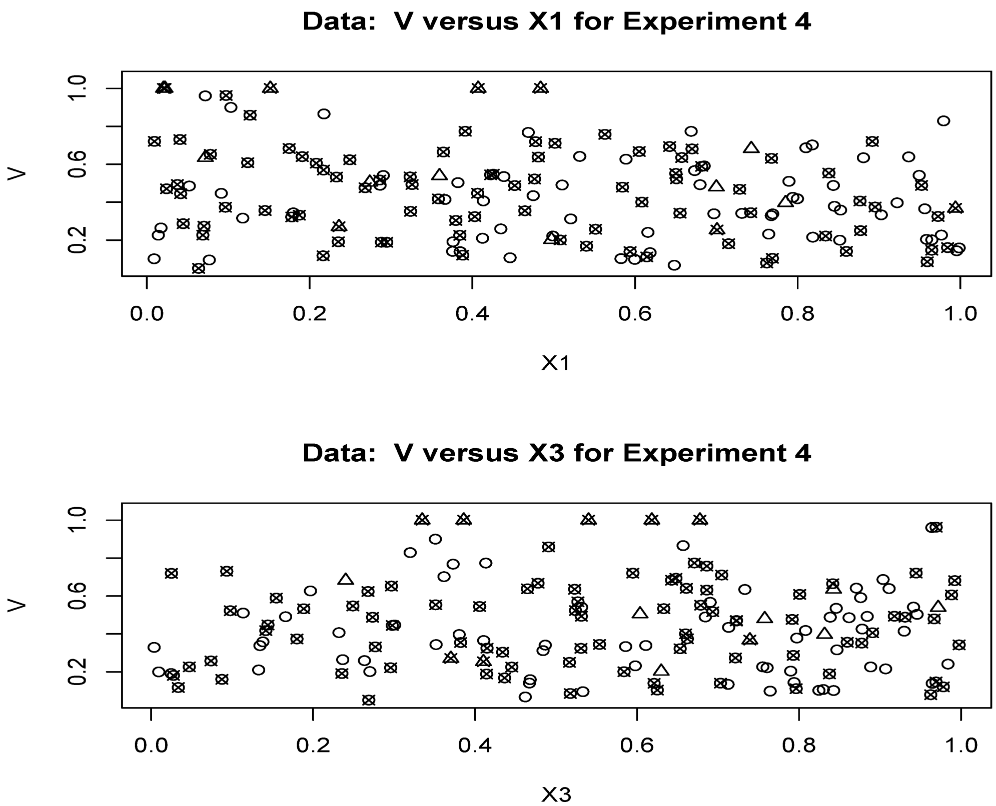

Figure 4. Data for BIFAR experiment 4 where the effect of impulses of increased air flow (cases with $U = 1$) are used to explore feasibility of this method versus controls with constant air flow (controls with $U = 0$). $X1 := X_1$ is the CSS (concentration of suspected solids), $X3 := X_3$ is the LAP (level of air pollution), V is the censored time of diffuser fouling. Observations are rescaled by BIFAR onto $[0, 1]^3$. The circles and the triangles show uncensored and censored lifetimes, respectively. The crosses overlay observations with $U = 1$. The total sample size $n = 163$, the number of controls and cases are 73 and 90, respectively. Among controls and cases 7 and 8 observations are censored, respectively.

visualize a relationship between lifetimes of interest and CSS X_1 which is denoted in Figure 1 by X. The bottom diagram shows us observed pairs (X_3, V) where X_3 is the LAP. Note that the triangles with $V < 1$ show observed lifetimes of the blower, and we can visualize a pronounced decrease in the lifetimes for larger LAP. Further, the uncensored diffuser lifetimes (the circles) indicate a possible minor effect of LAP on diffuser fouling. The obtained formula for the conditional hazard is

$$\hat{h}(t|x_1, x_3) = 2.6 - \varphi_1(t)[0.9 - 0.6\varphi_1(x_1) - 0.2\varphi_1(x_3)] - \varphi_1(x_1) - 0.5\varphi_1(x_3), \quad (t, x_1, x_3) \in [0, 1]^3. \quad (5.3)$$

As we see, the conditional hazard rate increases in time, CSS and LAP, and the increase in time t accelerates as CSS and LAP increase. It is of interest to compare (5.3) with (5.1) where $x = x_1$. We may conclude that despite different studied components of aeration (diffuser, blower, air filtration, waste), shapes of the estimated conditional hazards are similar.

Experiment 4. Methods of combating the diffuser fouling process can be divided into two groups: the ones that require emptying the reactor and the ones that do not. The non-invasive methods include addition of formic acid to the air supplying diffusers and periodic impulses of increased air flow. Adding formic acid only enables to control the pressure losses of diffusers (limited effect on external fouling), whereas increasing the air flow rate results in the detachment of loose external biofilm fragments. BIFAR tested the second method of periodic increases in the air flow rate. As a result, we get a categorical covariate $U \in \{0, 1\}$, and $U = 0$ indicates a control aeration with constant air flow while $U = 1$ indicates a case aeration with periodic impulses of increased air flow. Otherwise the study is similar to Experiment 3 and we are dealing with the triplet (X_1, X_3, U) of covariates and observe a sample from (V, X_1, X_3, U).

Data are shown in Figure 4 and explained in the caption. Here our main attention is to the cases indicated by crosses. Similarly to the previous experiments, without taking into account censored times (the triangles), it is difficult to say something about the benefit of periodic air impulses. Useful information about the experiment is given in the caption. Presented below formula for the conditional hazard rate sheds more light on the experiment. Recall that the two basis functions for the categorical random variable U are $\eta_0(u) = 1$ and $\eta_1(u) = 1 - 2I(u = 1)$, $u \in \{0, 1\}$. The formula is

$$\hat{h}(t|x_1, x_3, u) = 2.8 + .2\eta_1(u) - \varphi_1(t)[0.8 + .2\eta_1(u) - 0.3\varphi_1(x_3) - .2\varphi_1(x_1)\eta_1(u)]$$

$$- .4\varphi_1(x_1) - 0.2\varphi_1(x_3), \quad (t, x_1, x_3) \in [0, 1]^3, u \in \{0, 1\}. \tag{5.4}$$

As we see, periodic impulses of increased air flow slow down diffuser fouling. The formula also sheds light on interaction between the 3 predictors.

6. Conclusion and possible future research

For the first time in the literature, the theory and methodology of efficient nonparametric estimation of conditional hazard rate of a right censored lifetime are developed. The proposed estimator matches performance of the oracle who knows underlying dimensionality and smoothness of the conditional hazard rate and all nuisance functions. It is shown that only a single natural nuisance function is used by the oracle to construct an optimal oracle-estimator. Because the natural nuisance function can be estimated with an accuracy sufficient for matching the oracle, it is possible to match the oracle by a data-driven estimator. Further, the used minimal assumption about smoothness of the natural nuisance function is not tied to smoothness of the conditional hazard, and this is an important theoretical outcome of the paper. It also follows from the developed theory that the oracle uses only uncensored observations for sharp minimax estimation because using censored observations triggers solution of ill-posed problems with slower rates of convergence. Important practical examples, devoted to waste treatment and reducing potent greenhouse gas emissions, show practical feasibility of the estimators for small samples.

Let us comment on the notion of natural nuisance function. The oracle defines a function (or a vector of functions) as the natural nuisance function if using it together with data yields a sharp minimax estimation. Let us present several particular examples that shed light on the notion. For mentioned in the Introduction nonparametric regression model $Y = m(X) + \sigma(X)\xi$ and available sample from pair (X, Y), the natural nuisance function is the design density f^X of the predictor X. The oracle uses the design density f^X and the sample to construct a sharp minimax estimator of the regression. Note that in this example the natural nuisance function is the nuisance function for both the oracle and the statistician because f^X is not related to the estimated regression function. Our second example is estimation of the hazard rate function h^T based on a sample from T considered in Efromovich (2016). For this problem the natural nuisance function is the survival function S^T. This is an interesting example because knowledge of S^T is equivalent to knowing the estimand h^T, and hence S^T is not a nuisance function for the statistician. Further, for this problem the oracle uses data and the natural nuisance function to construct a sharp minimax oracle-estimator of h^T that can be mimicked by a data-driven estimator. Our final example is the setting considered in the paper when the natural nuisance function is $p(t,x) = f^X(x)S^{C|X}(t|x)S^{T|X}(t|x)$. Note that only f^X and $S^{C|T}$ are nuisance functions for the statistician, and the presence of factor $S^{T|X}$ allows the statistician to directly estimate p instead of estimation of the factors. The latter is the interesting specific of the studied problem.

Now let us mention several open problems for future research. An interesting setting, specifically in biostatistical and actuarial applications, is when a categorical variable affects smoothness and dimensionality of conditional hazard in continuous covariates. A specific example is the effect of smoking or zip code on longevity. Missing data are typical in survival data. For a nonparametric regression different estimators are optimal for missing responses and predictors. It is of interest to explore these settings for conditional hazard. Sequential estimation with assigned risk is an interesting and important practical problem due to unknown smoothness of conditional hazard and unknown censoring mechanisms. Measurement errors in covariates is another familiar problem in survival data. It will be of interest to explore optimal nonparametric estimation for this setting.

Acknowledgments

The research was supported by Grant from BIFAR and NSF Grant DMS–1915845. Valuable suggestions and comments of the Editor, Prof. Davy Paindaveine, an Associate Editor and referees are greatly appreciated.

Supplementary Material

Supplement to "Conditional hazard rate estimation for right censored data" (DOI: ....; .pdf). The Supplementary Material includes the remaining proofs.

References

Albu, N., Barani, N. and Constantin, M. (2021). Choosing an conomical solution for water aeration. *Hydraulica* **145** 32–37.

Balan, T. and Putter, H. (2020). Tutorial on frailty models. *Statist. Methods Med. Res.* **29** 3424–3454.

Brunel, E., Comte, F. and Guilloux, A. (2011). Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Ann. de l'I.H.P. Prob. Stat.* **47** 1171–1196.

Brunel, E., Comte, F. and Guilloux, A. (2013). Estimation/Imputation strategies for missing data in survival analysis. In *Statistical Models and Methods for Reliability and Survival Analysis.* Hoboken: John Wiley 229–252.

Butzer, P. and Nessel, R. (1971). *Fourier Analysis and Approximations*. New York: Academic Press.

Cai, J., Fan, J., Jiang, J. and Zhou, H. (2008). Partially linear hazard regression with varying coefficients for multivariate survival data. *J. R. Statist. Soc., Ser. B Statist. Methods* **70** 141–158.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman&Hall.

Cui, Y. and Hanning, L. (2019). Nonparametric generalized fiducial inference for survival functions under censoring (with Discussion). *Biometrika* **106** 501–518.

Drewnowski, J., Remiszewska-Skwarek, A., Duda, S. and Lagod, G. (2019). Aeration Process. In *Bioreactors as the Main Energy Consumer in a Wastewater Treatment Plant. Review of Solutions and Methods of Process Optimization: Review. Processes* **7** 1–21.

Gill, R. (2006). *Lectures on Survival Analysis*, New York: Springer.

Efromovich, S. (1989). On sequential nonparametric estimation of density. *Theory Probab. Appl.* **34** 228–239.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer.

Efromovich, S. (2001). Density estimation under random censorship and order restrictions: from asymptotic to small samples. *J. Amer. Statist. Assoc.* **96** 667–685.

Efromovich, S. (2011). Nonparametric regression with predictors missing at random. *J. Amer. Statist. Assoc.* **106** 306–319.

Efromovich, S. (2013). Nonparametric regression with the scale depending on auxiliary variable. *Ann. Statist.* **41** 1542–1568.

Efromovich, S. (2016). Minimax theory of nonparametric hazard rate estimation: efficiency and adaptation. *Ann. Inst. Math. Statist.* **68** 25–75.

Efromovich, S. (2018). *Missing and Modified Data in Nonparametric Estimation with R Examples.* Boca Raton: Chapman&Hall.

Efromovich, S. (2021). Sharp minimax distribution estimation for current status censoring with or without missing. *Ann. Statist.* **49** 568–589.

Efromovich, S. (2024). Supplement to "Conditional hazard rate estimation for right censored data."

Fleming, T. and Harrington, D. (2011). *Counting Processes and Survival Analysis*. New York: Wiley.

Gneyou, K. (2014). A strong linear representation for the maximum conditional hazard rate estimator in survival analysis. *J. Multivariate Anal.* **128** 10–18.

Golubev, G. (1991). LAN in problems of non-parametric estimation of functions and lower bounds for quadratic risks. *Probl. Inform. Transm.* **36** 152–157.

Hoffmann, M. and Lepski, O. (2002). Random rates in anisotropic regression (with discussion). *Ann. Statist.* **30** 325–396.

Huang, J. and Su, Y. (2021). Asymptotic properties of penalized spline estimators in concave extended linear models: Rates of convergence. *Ann. Statist.* **49** 338–347.

Ibragimov, I. and Khasminskii, R. (1981). *Statistical Estimation: Asymptotic Theory.* New York: Springer.

Jankowski, H. and Wellner, J. (2009). Nonparametric estimation of a convex bathtub-shaped hazard functions. *Bernoulli* **15** 1010–1035

Kahane, J.-P. (1985). *Some Random Series of Functions.* Cambridge: Cambridge Univ. Press.

Kalbfleisch, J. and Prentice, R. (2011). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kang, S., Lu, W. and Zhang, J. (2018). On estimation of the optimal treatment regime with additive hazards model. *Statist. Sinica* **28** 1539–1560.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated data.* New York: Springer.

Kooperberg, C., Stone, C.J. and Truong, Y. (1995). Hazard regression. J. Amer. Statist. Assoc. **90** 78–94.

Lee, E. and Wang, J. (2003). *Statistical Methods for Survival Data Analysis*. New York: Wiley.

Legrand, C. (2021). *Advanced Survival Models*. Boca Raton: Chapman&Hall.

Li, G. and Doss, H. (1995). An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.* **23** 787–823.

Lopuhaa, H. and Musta, E. (2017). Isotonized smooth estimators of a monotone baseline hazard in the Cox model. *J. Statist. Plan. Infer.* **191** 43–67.

Lu, M., Lu, T. and Li, C. (2018). Efficient estimation of partially linear additive Cox model under monotonicity constraint. *J. Statist. Plan. Infer.* **192** 18–34.

McKeague, I. and Utikal, K. (1990). Inference for a nonlinear counting process regression model. *Ann. Statist.* **18** 1172–1185.

Miller, R. (2011). *Survival Analysis*. New York: Wiley.

Moore, D. (2016). *Applied Survival Analysis Using R*. New York: Springer.

Pinsker, M. S. (1980). Optimal filtering a square integrable signal in Gaussian white noise. *Probl. Inform. Transm.* **16** 52–68.

Ross, L., Prentice, R. and Zhao, S. (2019). *The Statistical Analysis of Multivariate Failure Time Data: A Marginal Data Approach.* Boca Raton: Chapman&Hall.

Samuelson, O., Björka, A. and Carlsson, B. (2021). Model-based monitoring of diffuser fouling using standard sensors. *Water Research* **13** 100–118.

Slavov, A. (2017). General characteristics and treatment possibilities of dairy wastewater - a review. *Food Technol. Biotechnol.* **55** 14–28.

Spierdijik, L. (2008). Nonparametric conditional hazard rate estimation: A local linear approach. *Comp. Statist. Data Anal.* **52** 2419–2934.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proc. Nat. Acad. Sciences* **72** 20–22.

Van Keilegom, I. and Veraverbeke, N. (2001). Hazard rate estimation in nonparametric regression with censored data. *Ann. Inst. Statist. Math.* **53** 730–745.

Vinardella, S., Astalsa, S., Pecesb, M, Cardetea, M., Fernandez, I., Mata-Alvareza, J. and Dosta, J. (2020). Advances in anaerobic membrane bioreactor technology for municipal wastewater treatment. *Ren. Sustain. Energy Rev.* **130** 109–136.

Zhao, L. and Feng, D. (2020). Deep neural networks for survival analysis using pseudo values. *IEEE J. Biom. Health Inform.* **24** 3308–3314.

Zhong, Q., Mueller, J. and Wang, J. (2022). Deep learning for the partially linear Cox model. *Ann. Statist.* **50** 1348–1375.

Zhou, M. (2019). *Empirical Likelihood Methods in Survival Analysis*. Boca Raton: Chapman&Hall.

Zygmund, A. (1959). *Trigonometric Series*, 2nd ed. New York: Cambridge Univ. Press.

Wasserman, L. (2005). *All of Nonparametric Statistics*. New York: Springer.