The Annals of Statistics
2021, Vol. 49, No. 1, 568–589
https://doi.org/10.1214/20-AOS1970

SHARP MINIMAX DISTRIBUTION ESTIMATION FOR CURRENT STATUS CENSORING WITH OR WITHOUT MISSING

BY SAM EFROMOVICH

Department of Mathematical Sciences, University of Texas at Dallas, efrom@utdallas.edu

Nonparametric estimation of the cumulative distribution function and the probability density of a lifetime X modified by a current status censoring (CSC), including cases of right and left missing data, is a classical ill-posed problem with biased data. The biased nature of CSC data may preclude us from consistent estimation unless the biasing function is known or may be estimated, and its ill-posed nature slows down rates of convergence. Under a traditionally studied CSC, we observe a sample from (Z, Δ) where a continuous monitoring time Z is independent of X, $\Delta := I(X \leq Z)$ is the status, and the bias of observations is created by the density of Z which is estimable. In presence of right or left missing, we observe corresponding samples from $(\Delta Z, \Delta)$ or $((1 - \Delta)Z, \Delta)$; the data are again biased but now the density of Z cannot be estimated from the data. As a result, to solve the estimation problem, either the density of Z must be known (like in a controlled study) or an extra cross-sectional sampling of Z, which is typically simpler than an underlying CSC study, be conducted. The main aim of the paper is to develop for this biased and ill-posed problem the theory of efficient (sharp-minimax) estimation which is inspired by known results for the case of directly observed X. Among interesting aspects of the developed theory: (i) While sharp-minimax analysis of missing CSC may follow the classical Pinsker's methodology, analysis of CSC requires a more complicated estimation procedure based on a special smoothing in both frequency and time domains; (ii) Efficient estimation requires solving an old-standing problem of approximating aperiodic Sobolev functions; (iii) If smoothness of the cdf of X is known, then its rate-minimax estimation is possible even if the density of Z is rougher. Real and simulated examples, as well as extensions of the core models to dependent X and Z and case-control CSC, are presented.

1. Introduction.

Current status censoring (CSC) is a well-known problem in survival analysis; see a discussion in books Sun (2007), Chen, Sun and Peace (2012), Sun and Zhao (2013), Groeneboom and Jongbloed (2014), Klein et al. (2014), Efromovich (2018), and in more recent papers Becker, Braun and White (2017), Li et al. (2017), van Es and Graafland (2017), Diao and Yuan (2019), Groeneboom and Hendrickx (2018), Li et al. (2019) and Malov (2019) where further references may be found.

We want to estimate the distribution of a time X when an event of interest occurs but cannot constantly monitor the time. Instead, there is a possibility to check status of the event at some random moment of time Z, called the *monitoring* time. Then the available observation is a pair of random variables (Z, Δ) where Z is the monitoring time and $\Delta := I(X \leq Z)$ is the status of the event of interest, namely the status (indicator) is equal to 1 if the event of interest already occurred at moment Z and 0 otherwise. Note that we never observe time X when the event occurred, and only know that the event occurred before or after the monitoring time Z (this is a dramatic difference with a classical censoring). A sample from (Z, Δ)

Received September 2019; revised January 2020.

MSC2020 subject classifications. Primary 62G07, 62G07; secondary 60J10.

Key words and phrases. Adaptation, aperiodic functions, case-control study, cumulative distribution function, density, missing, sharp minimax.

is called current status censored. It is also a typical situation when CSC observations are missing. Two particular missing CSC (MCSC) settings considered in the paper are: (i) Right MCSC (RMCSC) when a realization of (Z, Δ) is observed only when $\Delta = 1$ and otherwise Z is missed, that is, we observe a sample from $(\Delta Z, \Delta)$; (ii) Left MCSC (LMCSC) when a realization of (Z, Δ) is observed only when $\Delta = 0$ and otherwise Z is missed, that is, we observe a sample from $((1 - \Delta)Z, \Delta)$. As we shall see shortly, the missing is destructive and no consistent estimation of the distribution of X is possible unless the distribution of Z is known or an extra sample from Z is collected. In a classical CSC model, the lifetime of interest X and the monitoring time Z are independent, and this is the core model explored in the paper. The model of dependent CSC, when X and Z are dependent, will be considered as an extension of the developed independent CSC theory and then be explored via real and simulated examples. It will be also explained what estimators, suggested for independent CSC, exhibit if X and Z are dependent. A critical aspect of dependent CSC is that even without missing no consistent estimation is possible based solely on dependent CSC data, and a possible remedy will be explored. Moreover, as we will see shortly, consistent estimation based solely on CSC data is the exception rather than the rule.

Another possible extension of the core model is presented in the following remark.

REMARK 1.1 (Case-, control- and case-control CSC). A CSC sampling is binomial and this implies that the number of observed monitoring times with a particular status is random and may be small. To avoid this randomness, another well-known CSC sampling procedure may be used when a fixed number of monitoring times is collected from one or both supplementary (created by the status) subpopulations of monitoring times. To shed light on the sampling and relate it to already discussed CSC models, we are considering three possible scenarios in turn. In a case-CSC, we collect m “cases” from the subpopulation of monitoring times with $\Delta = 1$, that is, the case is a monitoring time with already occurred event of interest. In a control-CSC, we collect cm (here c is a factor) “controls” from the subpopulation of monitoring times with $\Delta = 0$, that is, the control is a monitoring time with not yet occurred event of interest. Finally, in a case-control CSC the both above-defined samples are available. Let us note that we may refer to an observation in the above-introduced RMCSC data as the case and to an observation in LMCSC data as the control; this familiar terminology sheds additional light on the missing CSC. We also may conclude that the difference between RMCSC and case-CSC (or between LMCSC and control-CSC) is in how the corresponding data are collected, namely a binomial sampling with missing not at random (MNAR) versus a deterministic sampling from the subpopulation of monitoring times with status 1 (or correspondingly with status 0). A nice discussion of this setting may be found in Jewell and van der Laan (2004a, 2004b), Vandenbroucke and Pearce (2014), Keogh and Cox (2014), Klein et al. (2014), Hsu, Gorfine and Zucker (2018).

Now let us present a general example, motivated by the above-mentioned literature, that will help us to understand considered models, proposed extra samplings and assumptions.

GENERAL EXAMPLE. Suppose that we are interested in a population of single men ages 25 to 54 and working in the transportation industry. For this population, we would like to estimate the distribution of age at incidence of an occult nonfatal disease for which an accurate diagnostic test is available. A doctor may conduct such a test for a cross-sectional sample of size n and then, if required, submit files with information about sick individuals to agency “S” and about healthy individuals to agency “H.” In the example, X is the (unknown, hidden and of primary interest) age of an individual at the disease incidence, Z is the individual’s age at the time of medical testing, the status $\Delta = 1$ if the disease is present and $\Delta = 0$ otherwise.

If the statistician has access to the doctor's files, then the available data are CSC without missing, if the statistician knows the sample size n and has access only to files of agency "S" then the available data are RMCSC, if the statistician knows n and has access only to files of agency "H," then the available data are LMCSC. Further, if the statistician has access only to m files of agency "S," then the available data are case-CSC, if the statistician has access only to cm files of agency "H," then the available data are control-CSC, and if the statistician can get these files from the both agencies, then we are dealing with the case-control CSC. As we will see shortly, in a number of possible scenarios knowing only CSC data is not enough for consistent estimation. For instance, RMCSC data are biased by the distribution of Z, and hence this distribution either must be known or estimated via an extra cross-sectional study of Z which yields a sample $Z_{E1}, \dots, Z_{Em}$. In our example, this means that either distribution of the age of working men at the time of medical testing is known, or it may be estimated via sampling the age and note that the sampling of Z does not involve a medical testing; it is not related to the disease, not related to the status Δ of monitoring time. As a result, a typical cross-sectional sampling of monitoring times is dramatically simpler than an underlying CSC sampling. The methodology of extra sampling of a monitoring time, which does not involve a lifetime of interest, will be analyzed and then recommended because it is a relatively inexpensive remedy against the complete loss of information about X caused by a missing CSC. Now let us comment on possible assumptions. In many applications, it is reasonable to assume that Z and X are independent, and this is a typical assumption in the literature. Because X is not observed, it is not a trivial (and in many cases impossible) problem to justify this assumption, and nonetheless it is a traditional one. For our example, it is well known that some diseases are age-related (think about cardiovascular diseases or prostate cancer), then Z and X are depended and we are dealing with so-called dependent CSC. As we will see shortly, in this case consistent estimation, based solely on CSC data, is impossible and then a remedy will be suggested and analyzed both theoretically and via examples. Next, it is well known that consistent estimation of the distribution of X is possible only if its support is a subset of the support of Z. In some applications the latter may not be the case, for instance, in our example the doctor may not have access to individuals from specific age groups. It will be explained what can be estimated in this case. Finally, lifetimes are typically bounded (in the example by 54 years), and hence without loss of generality we may assume that the lifetimes take values from [0, 1]. All other used assumptions are technical and will be explained after their introduction.

A large number of interesting CSC examples can be found in the above-mentioned literature, and the online Supplementary Material (Efromovich (2021)) present eight real and simulated examples that shed additional light on the topic.

Now let us formulate core CSC problems and assumptions. There is an underlying (hidden) sample $X_1, \dots, X_n$ from a random variable of interest (lifetime) X. There is also a sample $Z_1, \dots, Z_n$ from an independent monitoring variable Z. Then three possible scenarios for available data are considered: (i) We observe CSC sample $(Z_1, \Delta_1), \dots, (Z_n, \Delta_n)$ where $\Delta_l := I(X_l \le Z_l)$; (ii) We observe RMCSC sample $(\Delta_1 Z_1, \Delta_1), \dots, (\Delta_n Z_n, \Delta_n)$; (iii) We observe LMCSC sample $((1 - \Delta_1)Z_1, \Delta_1), \dots, ((1 - \Delta_n)Z_n, \Delta_n)$. For these three models, we would like to explore sharp minimax estimation of the cumulative distribution function $F^X(x)$ and the density $f^X(x) := dF^X(x)/dx$ of the lifetime of interest X under the mean integrated squared error (MISE) criterion.

ASSUMPTION 1. The lifetime of interest X and the monitoring time Z are independent continuous random variables, $\mathbb{P}(X \in [0, 1]) = 1$, a known density $f^Z(z)$ of the monitoring time is continuous on [0, 1], $\int_0^1 f^Z(z)\,dz = 1$ and $\min_{z \in [0,1]} f^Z(z) \ge c_* > 0$.

The made assumption about independence of X and Z is standard (it will be relaxed in Section 6). The independence implies that the joint (mixed) density of the pair (Z, Δ) is

$$f^{Z,\Delta}(z,\delta) = f^{Z}(z)\big[F^{X}(z)\big]^{\delta}\big[1 - F^{X}(z))\big]^{1-\delta}, \quad \delta \in \{0,1\}. \tag{1.1}$$

This formula shows that consistent estimation of the distribution of X is possible if and only if the support of X is a subset of the support of Z. Further, note that CSC observations are biased by f^Z. Density f^Z can be estimated using CSC data but not missing CSC data. In the latter case, f^Z must be either known or estimated using an extra sample from Z, and the latter approach will be discussed in Section 3 and via examples in the Supplementary Material (Efromovich (2021)).

Our next assumption allows us to develop local minimax lower bounds for the mean integrated squared error of estimation of cdf $F^X(x)$ and density $f^X(x)$ of the lifetime of interest X. Introduce a Sobolev class of α-fold differentiable functions on [0, 1] (here and in what follows $g^{(\alpha)}(x)$ or $(g(x))^{(\alpha)}$ denote the αth derivative)

$$\begin{aligned} \mathcal{F}(\alpha, Q) := \Big\{ & g : g^{(\alpha)}(x) \text{ exists and finite on } [0,1], \\ & \text{and } \int_0^1 \big[g^{(\alpha)}(x)\big]^2\,dx \le Q < \infty \Big\}. \end{aligned} \tag{1.2}$$

Under a local minimax approach, we are assuming that an underlying cdf $F^X(x)$ is close in L_∞-norm to a pivotal cdf $F_0(x)$ and similarly density $f^X(x)$ is close in L_∞-norm to the pivotal density $F_0^{(1)}(x)$.

ASSUMPTION 2. Let $F_0(x)$ be the cdf of a random variable (lifetime) supported on [0, 1]. Introduce a class of cumulative distribution functions supported on [0, 1] and created by a perturbation of F_0,

$$\begin{aligned} \mathcal{F}(F_0, \alpha, Q, c_0, c_1, \rho) := \Big\{ & F : F(x) = F_0(x) + g(x)I(0 \le x \le 1), \\ & F(0) = 0, F(1) = 1, F^{(1)}(x) \ge 0, x \in [0,1], \\ & g \in \mathcal{F}(\alpha, Q), \max_{x\in[0,1]} \max\big(|g(x)|, |g^{(1)}(x)|\big) \le \rho, \\ & F_0(0) = 0, F_0(1) = 1, \min_{x\in[0,1]} F_0^{(1)}(x) \ge c_0, \\ & \max_{x\in[0,1]} F_0^{(1)}(x) \le c_1 \Big\}. \end{aligned} \tag{1.3}$$

It is assumed that an underlying cdf $F^X(x)$ belongs to this class.

Note that the second line in (1.3) implies that the class (1.3) contains only bona fide cumulative distribution functions $F(x)$ supported on [0, 1].

Assumption 2, with the Sobolev class (1.2) being replaced by a Sobolev ellipsoid of Fourier coefficients of $g(x)$, is a classical assumption in sharp-minimax literature whose origin goes back to the pioneering approach of Pinsker (1980); see a discussion in Efromovich (1999, 2018), Tsybakov (2009) and Cai (2012). A Sobolev class is larger than a corresponding Sobolev ellipsoid due to boundary conditions, and this creates new issues that will be resolved via using a special sequence of orthonormal bases. The last two inequalities in (1.3) add a restriction on the pivot $F_0(x)$ which allows us to introduce feasible additive permutations $g(x)$ that preserve bona fide properties of the cdf and the density.

The context of the paper is as follows. We begin with exploring the problem of cdf estimation. Lower minimax bounds for RMCSC, LMCSC and CSC data are presented in Section 2. In Section 3, oracle-estimators (they know more than data) of the cdf are presented whose MISEs attain the corresponding three lower bounds. One outcome to point upon is that for CSC data the oracle-estimator uses a special smoothing in both frequency and time domains (typically smoothing in frequency domain is sufficient for sharp-minimax estimation). Another interesting part of the solution is using a polynomial-cosine basis that depends on the sample size. Adaptive estimators of the cdf are considered in Section 4. Density estimation is considered in Section 5. Section 6 presents several extensions including dependent CSC and case-control CSC. Proofs can be found in the online Supplementary Material (Efromovich (2021)) that also contain real and simulated examples as well as additional references.

Let us finish the Introduction by describing the terminology of a minimax approach used in the paper. The minimax is a game with four participants being the dealer, the nature, the oracle and the statistician. The dealer defines a class of underlying distributions (here (1.3)), a risk (here the MISE), the sample size n (which can be as large as desired because only asymptotic in n is of interest in the game), and then presents this information to the nature and the oracle. The nature chooses the most difficult distribution for estimation, informs the dealer about the choice and then generates a sample of size n which is known to all four participants. Then the dealer proposes a lower bound for the minimax MISE, informs the oracle about it and also provides the oracle with some restricted information about the distribution chosen by the nature. Based on this information, the oracle tries to find an oracle-estimator whose MISE matches the lower bound, and if the latter is possible then the lower bound is announced to be sharp-minimax for the oracle. Otherwise, the game between the dealer and the oracle continues and either the dealer increases the lower bound or the oracle gets more information from the dealer to match the lower bound. When the dealer-oracle game is finished, the dealer informs the statistician about the lower bound and, to benefit the statistician, about the sharp-minimax oracle-estimator. Then the statistician tries to propose a data-driven (adaptive) estimator that matches MISE of the oracle-estimator. If the latter is possible, then the game is over, the lower bound and the adaptive estimator are announced to be sharp-minimax, and the estimator may be referred to as efficient.

In what follows, $o_n(1)$ is a traditional notation for vanishing sequences in n, $s := s_n := 3 + \lceil \ln(\ln(n+3)) \rceil$, and $\lceil a \rceil$ denotes the smallest integer larger or equal to a. To make propositions shorter, we may use notation $(F^X(x))^{(\beta)}$ with $\beta = 0$ and $\beta = 1$ corresponding to the cdf and the density, respectively. In the paper, the parameter β is used solely for this purpose.

2. Dealer's lower minimax bounds for CDF and density.

Given Assumptions 1 and 2, we would like to estimate the cdf $F^X(x) \in \mathcal{F}$ of the lifetime of interest X as well as its density $f^X(x)$. Set

$$\begin{aligned} &\mathcal{J}(\alpha, Q, d, \beta) \\ &\quad := \big[\big(Q(2\alpha+1)\big)^{-\frac{2\beta+1}{2(\alpha-\beta)}} (2\beta+1)^{\frac{2\alpha+1}{2(\alpha-\beta)}} \pi(\alpha+1+\beta)(\alpha-\beta)^{-1}\big]^{\alpha/(\alpha-\beta)}/d. \end{aligned} \tag{2.1}$$

Here, d is a functional of F^X and f^Z which depends on an underlying sampling model. Recall that we simultaneously consider three models of collected data: (i) RMCSC if we observe a sample from $(\Delta Z, \Delta)$; (ii) LMCSC if we observe a sample from $((1-\Delta)Z, \Delta)$; (iii) CSC if no missing occurs and we observe a sample from (Z, Δ). Then for the RMCSC, LMCSC and CSC models the corresponding functionals d are

$$\begin{aligned} &d_{\text{RM}} := \int_0^1 \frac{F^X(x)}{f^Z(x)}\,dx, \qquad d_{\text{LM}} := \int_0^1 \frac{1-F^X(x)}{f^Z(x)}\,dx, \\ &d_{\text{CSC}} := \int_0^1 \frac{F^X(x)(1-F^X(x))}{f^Z(x)}\,dx. \end{aligned} \tag{2.2}$$

It will be explained shortly that $\mathcal{J}$ and d may be referred to as the nonparametric Fisher information and the coefficient of difficulty, respectively. Also recall that $(F^X(x))^{(\beta)}$ with $\beta = 0$ and $\beta = 1$ denotes the cdf and the density, respectively, and that parameter β is used solely to indicate the estimand of interest.

THEOREM 2.1. *RMCSC, LMCSC and CSC models and estimation of the cdf or the density $(F^X(x))^{(\beta)}$, $\beta \in \{0, 1\}$ are considered. Suppose that Assumptions 1 and 2 hold, $\alpha \geq 1+\beta$, and a sample of size n is available. Then the following dealer's lower bound holds*:

$$\inf_{\tilde{\Psi}_\beta} \sup_{F^X} \mathbb{E}_{F^X}\Big\{[n\mathcal{J}(\alpha, Q, d, \beta)]^{2(\alpha-\beta)/(2\alpha+1)} \times \int_0^1 [\tilde{\Psi}_\beta(x) - (F^X(x))^{(\beta)}]^2 \, dx\Big\} \geq (1 + o_n(1)). \tag{2.3}$$

Here, d is defined in (2.2) for an underlying sampling model, the supremum is over $F^X \in \mathcal{F}(F_0, \alpha, Q, 1/s_n, s_n^{1/2}, 1/s_n)$ defined in (1.3), the infimum is taken over all possible dealer-estimators $\tilde{\Psi}_\beta$ knowing the sample, density $f^Z(x)$ of the monitoring time Z and everything about the class (1.3), namely the dealer knows F_0, α, Q and s_n.

Let us comment on the lower bound (2.3). First of all, it will be shown shortly that it is sharp. Second, sequence $s_n \to \infty$ as $n \to \infty$ (it is introduced at the end of the Introduction) and in Theorem 2.1 it is used to consider a local Sobolev function class where all cumulative distribution functions F^X converge in L_∞-norm to the pivot F_0 which, in its turn, is allowed to have a larger derivative as n increases. In other words, in (2.3) we are considering a shrinking minimax with a pivot that may change with n. Third, let us compare the CSC nonparametric lower bound (2.3) with a known sharp lower bound for a classical nonparametric regression $Y = m(V) + \sigma(V)\xi$, where Y is the response, V is the predictor with density f^V supported on [0, 1], ξ is a standard normal error independent of V, $m(v) = \mathbb{E}\{Y|V = v\}$ is the nonparametric regression of interest and $\sigma(v)$ is a nuisance scale function. For this regression and the estimand $m(v) \in \mathcal{F}(\alpha, Q)$, a corresponding regression minimax lower bound is identical to (2.3) where we use $\beta = 0$ and $\mathcal{J}(\alpha, Q, d_R, 0)$ with $d_R := \int_0^1 [\sigma^2(v)/f^V(v)]\, dv$. In nonparametric regression literature, it is a tradition to refer to $\mathcal{J}$ and d_R as the nonparametric Fisher information and the coefficient of difficulty, respectively, and this terminology may be also used for CSC models. Note that in the theory of efficient estimation of a parameter, in a lower bound like (2.3) in place of factor $[n\mathcal{J}(\alpha, Q, d, \beta)]^{2(\alpha-\beta)/(2\alpha+1)}$ we would see $n\mathcal{J}_*$ with $\mathcal{J}_*$ being a classical Fisher information. The latter explains the terminology. Further, the interested reader may compare the regression coefficient of difficulty d_R with the CSC coefficients of difficulty (2.2), and then realize a striking similarity between the coefficients of difficulty. Fourth, in the dealer's lower bound only the coefficient of difficulty d reflects an underlying sampling model. Fifth, because we are assuming that the density $f^Z(z)$ is bounded below from zero on [0, 1], the coefficients of difficulty are finite and also bounded below from zero (see the proof in the Supplementary Material (Efromovich (2021))). Sixth, we can realize via analysis of (2.2) how the missing mechanisms affect the accuracy of estimation of F^X, and how the effect depends on F^X and f^Z. Finally, note that the rate of estimating the cdf is not n^{-1} but $n^{-2\alpha/(2\alpha+1)}$ which is traditional for estimating α-fold differentiable densities based on direct observations of X; see Efromovich (1999). The latter could be expected because when f^Z is known then estimation of F^X is converted into estimation of $f^{Z,\Delta}(x, 1)$ for RMCSC data or $f^{Z,\Delta}(x, 0)$ for LMCSC data. This remark sheds light on why the considered CSC problem is often referred to as ill-posed.

REMARK 2.1. The obtained expressions (2.2) for coefficients of difficulty point upon optimal monitoring times Z that minimize them. It is directly verified by the Cauchy–Schwarz inequality that the corresponding optimal monitoring densities are

$$f_{\mathrm{RM}}^{Z}(z) = \frac{[F^X(z)]^{1/2}}{\int_0^1 [F^X(x)]^{1/2}\,dx} I\bigl(z \in [0,1]\bigr), \tag{2.4}$$

$$f_{\mathrm{LM}}^{Z}(z) = \frac{[1-F^X(z)]^{1/2}}{\int_0^1 [1-F^X(x)]^{1/2}\,dx} I\bigl(z \in [0,1]\bigr), \tag{2.5}$$

and

$$f_{\mathrm{CSC}}^{Z}(z) = \frac{[F^X(z)(1-F^X(z))]^{1/2}}{\int_0^1 [F^X(x)(1-F^X(x))]^{1/2}\,dx} I\bigl(z \in [0,1]\bigr). \tag{2.6}$$

3. Oracle-estimator for CDF. We begin with several new notations. Denote by $\varphi_j(x) := 2^{1/2}\cos(\pi j x)$, $j = 1, 2, \dots$ elements of the cosine basis on $[0,1]$, and by $L_0(x), \dots, L_{s-1}(x)$ denote the first s elements of the so-called "shifted" Legendre orthonormal polynomials on $[0,1]$ (recall that $s := s_n$ is defined at the end of the Introduction). In particular, $L_0(x) = 1$, $L_1(x) = 3^{1/2}(2x-1)$, $L_2(x) = 5^{1/2}(6x^2 - 6x + 1)$, $L_3(x) = 7^{1/2}(20x^3 - 30x^2 + 12x - 1)$, and we also have the recurrence relation

$$\begin{aligned} L_{k+1}(x) = (2k+3)^{1/2}\biggl[&\frac{(2k+1)^{1/2}}{k+1}(2x-1)L_k(x) \\ &- \frac{k}{(k+1)(2k-1)^{1/2}} L_{k-1}(x)\biggr]. \end{aligned} \tag{3.1}$$

Then we apply Gram–Schmidt orthonormalization to $\{L_0(x), \dots, L_{s-1}(x), \varphi_1(x), \varphi_2(x), \dots\}$ and get a new polynomial-cosine basis $\{L_0(x), \dots, L_{s-1}(x), \psi_1(x), \psi_2(x), \dots\}$. Note that the basis depends on n via sequence $s := s_n$. Further, introduce two sequences

$$J(n,d,\alpha,Q) := \bigl\lceil (n/d)^{1/(2\alpha+1)} \bigl[Q\pi^{-2\alpha}(\alpha+1)(2\alpha+1)/\alpha\bigr]^{1/(2\alpha+1)} \bigr\rceil \tag{3.2}$$

and $J'(n,d,\alpha,Q) := \lceil J(n,d,\alpha,Q)/\ln(n+20)\rceil$ where $d > 0$ is a coefficient of difficulty defined in (2.2) for a particular underlying model.

We begin with the case of CSC in presence of missing data. The oracle uses unbiased estimators of Fourier coefficients $\kappa_i := \int_0^1 F^X(x)L_i(x)\,dx$ and $\theta_j := \int_0^1 F^X(x)\psi_j(x)\,dx$. For the RMCSC model, the corresponding Fourier estimators are

$$\tilde{\kappa}_{Ri} := n^{-1}\sum_{l=1}^{n} \frac{\Delta_l L_i(\Delta_l Z_l)}{f^Z(\Delta_l Z_l)}, \qquad \tilde{\theta}_{Rj} := n^{-1}\sum_{l=1}^{n} \frac{\Delta_l \psi_j(\Delta_l Z_l)}{f^Z(\Delta_l Z_l)}. \tag{3.3}$$

REMARK 3.1. The support of Z may be larger than a known support of X. Let the support of X be still $[0,1]$. Then the only required change in the Fourier estimators is to use extra factors $I(\Delta_l Z_l \in [0,1])$. For instance, in (3.3) the modified $\tilde{\kappa}_{Ri}$ will be $n^{-1}\sum_{l=1}^{n}[\Delta_l L_i(\Delta_l Z_l) I(\Delta_l Z_l \in [0,1])/f^Z(\Delta_l Z_l)]$.

Fourier estimators for the LMCSC model are constructed similarly with the underlying idea that in this model the natural estimand is the survival function $1 - F^X(x)$. Keeping in mind that $L_0(x) = 1$ and orthonormality of the polynomial-cosine basis, we write

$$\begin{aligned} \tilde{\kappa}_{Li} &:= I(i=0) - n^{-1}\sum_{l=1}^{n} \frac{(1-\Delta_l)L_i((1-\Delta_l)Z_l)}{f^Z((1-\Delta_l)Z_l)}, \\ \tilde{\theta}_{Lj} &:= -\sum_{l=1}^{n} \frac{(1-\Delta_l)\psi_j((1-\Delta_l)Z_l)}{f^Z((1-\Delta_l)Z_l)}. \end{aligned} \tag{3.4}$$

The oracle's cdf estimator for RMCSC and LMCSC models is

$$\begin{aligned}\tilde{F}^X_*(x) := \sum_{i=0}^{s-1}\tilde{\kappa}_i L_i(x) + \sum_{j=1}^{J'(n,d,\alpha,Q)} \tilde{\theta}_j\psi_j(x)\\ + \sum_{j=J'(n,d,\alpha,Q)+1}^{J(n,d,\alpha,Q)} \big[1-\big(j/J(n,d,\alpha,Q)\big)^{\alpha}\big]\tilde{\theta}_j\psi_j(x).\end{aligned}\tag{3.5}$$

Here, the used $\{\tilde{\kappa}_i, \tilde{\theta}_j, d\}$ are either $\{\tilde{\kappa}_{Ri}, \tilde{\theta}_{Rj}, d_{\mathrm{RM}}\}$ for RMCSC data or $\{\tilde{\kappa}_{Li}, \tilde{\theta}_{Lj}, d_{\mathrm{LM}}\}$ for LMCSC data. The reader familiar with Pinsker (1980) could notice that (3.5) is a modified Pinsker's oracle-estimator which recommends using smoothing coefficients for all Fourier coefficients.

For CSC data with no missing, the oracle proposes a more complicated estimator whose construction involves several steps. First, the support of X is divided into s consecutive subintervals $[b_r, b_{r+1})$, $b_r := (r-1)/s$, $r = 1,\dots,s$. In what follows, to simplify formulas, it is assumed that the last right interval is $[b_s, b_{s+1}]$ and not $[b_s, b_{s+1})$. Second, for an rth subinterval we introduce its own polynomial-cosine basis $\{L_{ri} := s^{1/2}L_i(s(x-b_r))I(x \in [b_r, b_{r+1}]), i = 0,\dots,s-1, \psi_{rj}(x) := s^{1/2}\psi_j(s(x-b_r))I(x \in [b_r, b_{r+1}]), j = 1,2,\dots\}$. Third, the oracle defines a Fourier estimator

$$\begin{aligned}\check{\theta}_{rj} := \lambda_r n^{-1}\sum_{l=1}^{n}\frac{I(Z_l \in [b_r, b_{r+1}))\Delta_l\psi_{rj}(\Delta_l Z_l)}{f^Z(\Delta_l Z_l)}\\ + (1-\lambda_r)(-1)n^{-1}\sum_{l=1}^{n}\frac{I(Z_l \in [b_r, b_{r+1}))(1-\Delta_l)\psi_{rj}((1-\Delta_l)Z_l)}{f^Z((1-\Delta_l)Z_l)},\end{aligned}\tag{3.6}$$

where

$$\lambda_r := \frac{\int_{b_r}^{b_{r+1}}[(1-F^X(x))/f^Z(x)]\,dx}{\int_{b_r}^{b_{r+1}}[1/f^Z(x)]\,dx}.\tag{3.7}$$

Note how RMCSC and LMCSC observations are aggregated in (3.6) and that the aggregation is time-depending.

The final oracle's step for CSC data is to calculate the following nonparametric estimator of the cdf $F^X(x)$:

$$\hat{F}^X_*(x) = \sum_{i=0}^{s-1} n^{-1}\big[N\tilde{\kappa}_{Ri} + (n-N)\tilde{\kappa}_{Li}\big]L_i(x)\tag{3.8}$$

$$+\sum_{r=1}^{s}\Bigg[\sum_{j=1}^{J'(n,d_r,\alpha,Q_r)}\check{\theta}_{rj}\psi_{rj}(x)I\big(x \in [b_r, b_{r+1})\big)\Bigg]\tag{3.9}$$

$$\begin{aligned}+\sum_{r=1}^{s}\Bigg[\sum_{j=J'(n,d_r,\alpha,Q_r)+1}^{J(n,d_r,\alpha,Q_r)}\big[1-\big(j/J(n,d_r,\alpha,Q_r)\big)^{\alpha}\big]\\ \times\check{\theta}_{rj}\psi_{rj}(x)I\big(x \in [b_r, b_{r+1})\big)\Bigg].\end{aligned}\tag{3.10}$$

Here, $N := \sum_{l=1}^{n}\Delta_l$ is the number of "cases" in the CSC sample,

$$d_r := \int_{b_r}^{b_{r+1}}\big[F^X(x)\big(1-F^X(x)\big)/f^Z(x)\big]\,dx,\tag{3.11}$$

and

$$Q_r := \int_{b_r}^{b_{r+1}} \big[\big(F^X(x) - F_0(x)\big)^{(\alpha)}\big]^2\,dx. \tag{3.12}$$

The main component of the estimator, that defines the rate and constant of the MISE convergence, is (3.10). In (3.10), smoothing weights $[1-(j/J(n,d_r,\alpha,Q_r))^\alpha]$ resemble classical Pinsker's weights but the difference is that Pinsker's weights change only in the frequency domain (depend solely on j) while weights in (3.10) change in frequency and time domains. As a result, the oracle uses a rather complicated aggregation of RMCSC and LMCSC data to construct efficient estimator for CSC data. Further, the smoothing weights in (3.10) are based on rather detailed information about $F_0(x)$, $F^X(x)$ and their derivatives. Also, note that the oracle does not aggregate estimators (3.5) for RMCSC and LMCSC data, while the latter looks like a reasonable approach for CSC data and indeed it is often recommended in the literature.

Another important remark is that, while the oracle knows the pivot F_0, the proposed oracle-estimators directly estimate an underlying cdf $F^X(x) = F_0(x) + g(x)$ and not $g(x)$. The reason is that the oracle's aim is to guide the statistician who does not know the pivot $F_0(x)$. This remark explains why in the following theorem it is assumed that $F_0(x)$ is smoother than $g(x)$, and hence does not affect the minimax MISE.

THEOREM 3.1. *RMCSC, LMCSC and CSC models are considered simultaneously. Suppose that Assumptions* 1 *and* 2 *hold*, $\alpha \geq 1$, *and additionally assume that* $F_0 \in \mathcal{F}(\alpha+1, Q')$ *where* $Q' < \infty$. *Then MISEs of the above-defined oracle-estimators attain the lower bound* (2.3), *namely*

$$\begin{aligned}\sup_{F^X\in\mathcal{F}(F_0,\alpha,Q,0,\infty,\infty)} \quad & \mathbb{E}_{F^X}\Big\{\big[n\mathcal{J}(\alpha,Q,d,0)\big]^{2\alpha/(2\alpha+1)} \\ & \times \int_0^1 \big(\bar{F}_*(x) - F^X(x)\big)^2\,dx\Big\} = \big(1+o_n(1)\big),\end{aligned} \tag{3.13}$$

where $\bar{F}_*(x)$ *and* d *are the corresponding oracle-estimator and coefficient of difficulty for an underlying model. In other words, the oracle-estimators are sharp minimax and attain the lower bound of Theorem* 2.1.

Note that in (3.13) the considered class of underlying cdfs is larger than the one used in the lower bound of Theorem 2.1.

Theorem 3.1 ends the first stage of the minimax game when the dealer suggests a lower bound and the oracle finds a corresponding oracle-estimator whose MISE attains this lower bound. In other words, the lower bound (2.3) is sharp for the oracle and the proposed oracle-estimators are efficient.

Before considering a data-driven estimator mimicking the oracle, it is prudent to explore the case when the oracle does not know density f^Z of the monitoring time Z. It can be estimated for a CSC without missing but not for a missing CSC. Indeed, consider RMCSC model when we observe Z only if $\Delta = 1$, then

$$f^{Z,\Delta}(z,1) = f^Z(z)F^X(z). \tag{3.14}$$

We see that RMCSC data are biased by $f^Z(z)$ and we need to either know or estimate $f^Z(z)$ for consistent estimation of F^X. Suppose that we may conduct a cross-sectional study of the monitoring time Z and get an extra sample $Z_{E1},\dots,Z_{Em}$ of size m from Z. Here, the subscript E stresses that this is an extra sample from Z and it is unrelated to the lifetime of interest X. For instance, in the Introduction's example X is the age of a worker at incidence of a nonfatal disease and Z is the age of the worker at the time of study, and hence a cross-sectional study of Z does not require getting information about the disease. Furthermore, information about distribution of the age of workers may be already available. Overall, an extra sampling of Z is less difficult and expansive than an underlying CSC sampling because no information about the lifetime of interest X is collected.

For missing CSC, using an extra sample $Z_{E1}, \dots, Z_{Em}$ from Z we may estimate $f^Z(z)$ by a data-driven (adaptive) estimate $\hat{f}^Z(z)$ (a particular example is the estimate (A.58) in the Supplementary Material (Efromovich (2021))) and then plug it in (3.5). Denote this plug-in oracle-estimator as $\check{F}_*^X(x)$. The same approach may be used for CSC without missing, but instead we may use the available sample $Z_1, \dots, Z_n$ to construct the plug-in estimate $\hat{f}^Z$ and use it in (3.8). The presented below proposition explores the both possibilities in parts A and B, respectively.

THEOREM 3.2. *Suppose that assumptions of Theorem 3.1 hold only now density $f^Z(z)$ of the monitoring time is unknown, and $f^Z \in \mathcal{F}(\alpha_z, Q_z)$, $\alpha_z \geq 1$, $Q_z < \infty$.*

A: *Assume that an extra sample $Z_{E1}, \dots, Z_{Em}$ from Z is available, and an underlying model is either RMCSC or LMCSC. Then*

$$\sup_{F^X \in \mathcal{F}(F_0, \alpha, Q, 0, \infty, \infty), f^Z \in \mathcal{F}(\alpha_z, Q_z)} \mathbb{E}_{F^X}\left\{\int_0^1 \left(\check{F}_*^X(x) - F^X(x)\right)^2 dx\right\} \\ \leq C \max\left(n^{-2\alpha/(2\alpha+1)}, m^{-2\alpha_z/(2\alpha_z+1)}\right), \quad C < \infty. \tag{3.15}$$

If additionally

$$m^{-\alpha_z/(2\alpha_z+1)} = o_n(1) n^{-\alpha/(2\alpha+1)}, \tag{3.16}$$

then

$$\sup_{F^X \in \mathcal{F}(F_0, \alpha, Q, 0, \infty, \infty), f^Z \in \mathcal{F}(\alpha_z, Q_z)} \mathbb{E}_{F^X}\left\{\left[n \mathcal{J}(\alpha, Q, d, 0)\right]^{2\alpha/(2\alpha+1)} \\ \times \int_0^1 \left(\check{F}_*^X(x) - F^X(x)\right)^2 dx\right\} = \left(1 + o_n(1)\right), \tag{3.17}$$

and the plug-in oracle-estimator is sharp minimax and attains the lower bound of Theorem 2.1.

Also, consider the case when $\alpha_z = 1$, α is known, and $m > c_* n$, $c_* > 0$. Then there exists a cdf estimator $\bar{F}(x, \alpha)$ whose MISE decreases with the optimal rate, that is,

$$\sup_{F^X \in \mathcal{F}(F_0, \alpha, Q, 0, \infty, \infty), f^Z \in \mathcal{F}(1, Q_z)} \mathbb{E}_{F^X}\left\{\int_0^1 \left(\bar{F}(x, \alpha) - F^X(x)\right)^2 dx\right\} \\ \leq C n^{-2\alpha/(2\alpha+1)}. \tag{3.18}$$

B: *Consider a setting when the assumption of part* A *about availability of an extra sample from Z is not valid. Then for CSC without missing the above-presented results hold with the estimator $\hat{f}^Z$ being based on the available monitoring times $Z_1, \dots, Z_n$ and with formally setting $m = n$. No consistent estimation of F^X is possible for RMCSC and LMCSC.*

Let us comment on the presented in Theorem 3.2 results. To shed light on (3.15), let us note that the cdf of interest $F^X(z) = f^{Z,\Delta}(z, 1)/f^Z(z)$ is the ratio. As a result, it is natural to expect that the ratio may be estimated with the lower rate among the two. If (3.16) holds, then there is no effect of the nuisance function $f^Z(z)$ on the cdf estimation and even sharp-minimax estimation is possible. $\bar{F}(x,\alpha)$ is correctly referred to as an estimator because it is assumed that we know parameter α but not all other characteristics like Q or F_0. The latter is a traditional assumption when a practitioner is sure that an underlying cdf is, for instance, at least twice differentiable and then $\alpha = 2$ can be used. The upper bound (3.18) shows that even if the nuisance density $f^Z(z)$ is rougher than the cdf, the cdf still can be estimated with the optimal rate corresponding to its known smoothness. Keeping in mind that we essentially estimate the ratio of two unknown densities, this is a nice outcome. Let us also comment on two main assumptions. The first one is that X and Z are independent. This is a traditional assumption in the CSC literature as well as in the survival analysis literature. In some applications, the assumption about independence does not hold, and then the setting is referred to as dependent CSC. We will consider this setting in Section 6 and in Examples 3–6 of the Supplementary Material (Efromovich (2021)). Here, it is worthwhile to explain what happens when the proposed estimators are used for dependent CSC. If X and Z are dependent, then in place of (1.1) we have

$$f^{Z,\Delta}(z,\delta) = f^Z(z)\big[F^{X|Z}(z|z)\big]^{\delta}\big[1 - F^{X|Z}(z|z)\big]^{1-\delta}, \tag{3.19}$$

where $F^{X|Z}(z|z) = \mathbb{P}(X \le z|Z = z)$. Formula (3.19) implies that for dependent CSC the proposed estimators estimate the univariate function $F^{X|Z}(z|z)$ in place of $F^X(z)$. Without additional information, $F^{X|Z}(z|z)$ is the only characteristic of the lifetime of interest that can be estimated, and it may be of interest on its own as shown in Example 3. The second comment is about a possibility that the support of Z is a subset of the support of X. In this case, we are estimating $F^X(x)$ over the support of Z, and the proposed estimators are again minimax if we define the MISE over the support of Z. Example 7 discusses such a setting.

4. Adaptive estimation of CDF. As soon as an oracle-estimator of Pinsker's type is found, there exist data-driven methods; see a discussion in Efromovich (1999). Many prominent methods mimic the oracle directly via using estimates of parameters α and Q that mimimize a penalized empirical risk. A more straightforward approach is to use a blockwise-shrinkage methodology whose attractive theoretical and applied properties are discussed in Efromovich (1999, 2018), Donoho et al. (1996), Hall, Kerkyacharian and Picard (1998), Chicken and Cai (2005) and Zhang (2005). The underlying idea is to avoid estimation of (α, Q) and instead mimic smoothing coefficients $(1 - (j/J)^{\alpha})$ in (3.5) and (3.10) by corresponding statistics.

We begin with the case of RMCSC data when for mimicking oracle (3.5) our tasks are:

(1) Define blocks;
(2) Define sequences J' and J that do not depend on d, α and Q;
(3) Define blockwise-shrinkage statistics that mimic smoothing weights in the third sum of (3.5).

For the first task, following Efromovich (1985) we introduce an increasing to infinity sequence of positive integers $1 = q_1 < q_2 < \cdots$ that create blocks $B_k := \{q_k, q_k + 1, \dots, q_{k+1} - 1\}$ with lengths $T_k := q_{k+1} - q_k$, $k = 1, 2, \dots$, and let us also introduce a sequence of positive and finite thresholds t_k. To be specific, set $T_k = k^2$ and $t_k = 1/\ln(\ln(20 + k))$, and let us also note that there is a wide choice of possible blocks and thresholds discussed in Efromovich (1999). For the second task, J' is defined as the smallest $q_{K'} - 1$ larger than $\lceil n^{1/s} \rceil$, and J as the smallest $q_K - 1$ larger than $\lceil s n^{1/3} \rceil$. For the third task, we estimate the coefficient of difficulty d_{RM} by

$$\tilde{d}_{\mathrm{RM}} := \max\left(1/s, n^{-1}\sum_{l=1}^{n} \Delta_l\big[f^Z(\Delta_l Z_l)\big]^{-2}\right), \tag{4.1}$$

and introduce Sobolev statistics

$$\tilde{\Theta}_k := \sum_{j\in B_k} \sum_{l_1 \neq l_2 = 1}^{n} \frac{\Delta_{l_1}\Delta_{l_2}\psi_j(\Delta_{l_1}Z_{l_1})\psi_j(\Delta_{l_2}Z_{l_2})}{T_k n(n-1) f^Z(\Delta_{l_1}Z_{l_1}) f^Z(\Delta_{l_2}Z_{l_2})}. \tag{4.2}$$

The motivation of (4.2) is that $E\{\tilde{\Theta}_k\} = T_k^{-1}\sum_{j\in B_k}\theta_j^2$, and hence Sobolev statistics are unbiased estimates of the Sobolev functionals; see Giné (1975).

Now we are ready to define an adaptive estimator of the cdf F^X which mimics oracle (3.5) and is based on the same Fourier estimators $\tilde{\kappa}_{Ri}$ and $\tilde{\theta}_{Rj}$ defined in (3.3),

$$\tilde{F}^X(x) := \sum_{i=0}^{s-1} \tilde{\kappa}_{Ri} L_i(x) + \sum_{j=1}^{J'} \tilde{\theta}_{Rj}\psi_j(x) \tag{4.3}$$

$$+ \sum_{k=K'}^{K} \frac{\tilde{\Theta}_k}{\tilde{\Theta}_k + \tilde{d}_{\mathrm{RM}} n^{-1}} I\big(\tilde{\Theta}_k > t_k n^{-1}\big) \sum_{j\in B_k} \tilde{\theta}_{Rj}\psi_j(x). \tag{4.4}$$

If bona fide properties are important, then the corresponding L_2-projection should be added (see Efromovich (1999) and Example 8 in the Supplementary Material (Efromovich (2021))). Estimator for LMCSC data is constructed similarly.

For CSC data, estimator of spatial coefficients of difficulty d_r is defined as

$$\begin{aligned} \hat{d}_r &:= [n(n-1)]^{-1} \\ &\times \sum_{l_1\neq l_2=1}^{n} \frac{\Delta_{l_1}[1-\Delta_{l_2}] I(Z_{l_1} \in [b_r, b_{r+1})) I(Z_{l_2} \in [b_r, b_{r+1}))}{[f^Z(Z_{l_1}) f^Z(Z_{l_2})]^2}, \end{aligned} \tag{4.5}$$

estimator of spatial Sobolev functionals is

$$\begin{aligned} \hat{\Theta}_{rk} &:= T_k^{-1}[n(n-1)]^{-1} \sum_{j\in B_k}\sum_{l_1\neq l_2=1}^{n} \Delta_{l_1}\Delta_{l_2} \\ &\times \frac{\psi_{rj}(Z_{l_1})\psi_{rj}(Z_{l_2}) I(Z_{l_1}\in[b_r,b_{r+1})) I(Z_{l_2}\in[b_r,b_{r+1}))}{f^Z(Z_{l_1}) f^Z(Z_{l_2})}, \end{aligned} \tag{4.6}$$

and the spatial Fourier estimator is

$$\begin{aligned} \hat{\theta}_{rj} &:= \hat{\lambda}_r n^{-1} \sum_{l=1}^{n} \frac{I(Z_l\in[b_r,b_{r+1}))\Delta_l\psi_{rj}(Z_l)}{f^Z(Z_l)} \\ &+ (1-\hat{\lambda}_r)(-1) n^{-1} \sum_{l=1}^{n} \frac{I(Z_l\in[b_r,b_{r+1}))(1-\Delta_l)\psi_{rj}(Z_l)}{f^Z(Z_l)}, \end{aligned} \tag{4.7}$$

where

$$\hat{\lambda}_r := \frac{n^{-1}\sum_{l=1}^{n}(1-\Delta_l) I(Z_l\in[b_r,b_{r+1}))[f^Z(Z_l)]^{-2}}{\int_{b_r}^{b_{r+1}} [1/f^Z(x)]\,dx}. \tag{4.8}$$

The adaptive cdf estimator is (compare with (3.8))

$$\begin{aligned}\hat{F}^X(x) := {} & \sum_{i=0}^{s-1} n^{-1}[N\tilde{\kappa}_{Ri} + (1-N)\tilde{\kappa}_{Li}] \\ & + \sum_{r=1}^{s}\sum_{j=1}^{J'} \hat{\theta}_{rj}\psi_{rj}(x)I\bigl(x \in [b_r, b_{r+1})\bigr) \\ & + \sum_{r=1}^{s}\sum_{k=K'}^{K} \frac{\hat{\Theta}_{rk}}{\hat{\Theta}_{rk} + \hat{d}_r n^{-1}} I\bigl(\hat{\Theta}_{rk} > t_k n^{-1}\bigr) \sum_{j\in B_k} \hat{\theta}_{rj}\psi_{rj}(x)I\bigl(x\in[b_r, b_{r+1})\bigr).\end{aligned} \tag{4.9}$$

THEOREM 4.1. *RMCSC, LMCSC and CSC models are considered simultaneously. Suppose that Assumptions 1 and 2 hold, $\alpha \geq 1$ and additionally assume that $F_0 \in \mathcal{F}(\alpha+1, Q')$ where $Q' < \infty$. Then MISEs of the above-defined estimators attain the lower bound (2.3), namely*

$$\begin{aligned}\sup_{F^X \in \mathcal{F}(F_0,\alpha,Q,0,\infty,\infty)} & \mathbb{E}_{F^X}\Bigl\{[n\mathcal{J}(\alpha, Q, d, 0)]^{2\alpha/(2\alpha+1)} \\ & \times \int_0^1 \bigl(\bar{F}(x) - F^X(x)\bigr)^2\,dx\Bigr\} = \bigl(1 + o_n(1)\bigr),\end{aligned} \tag{4.10}$$

where $\bar{F}(x)$ and d are the estimator and coefficient of difficulty for an underlying model.

Theorem 4.1 ends outlined in the Introduction minimax game for the cdf estimation. We now know that the lower bound of Theorem 2.1 is sharp and the proposed cdf estimators are efficient and adaptive to an underlying smoothness of the cdf.

5. Density estimation.

There are two possible approaches to the problem of estimation of the density. The former is to estimate the cdf and then differentiate the estimate. This approach is absolutely natural for CSC problem and, for instance, it is used in Groeneboom, Jongbloed and Witte (2010) where derivative of a maximum likelihood cdf estimator is used to estimate the density. Further, Efromovich (1999) shows that while derivative of a Pinsker's oracle is not sharp-minimax estimate of the derivative, derivative of a blockwise-shrinkage estimate is.

Another approach, which is traditional in the classical density estimation literature, is to bypass estimation of the cdf and consider density estimation as a self-defined nonparametric problem. In this case, it is assumed that an underlying density $f^X(x)$ belongs to a local Sobolev function class defined by a pivotal density $f_0(x)$ supported on [0, 1] (compare with (1.3))

$$\begin{aligned}\mathcal{F}'(f_0, \nu, Q, c_0, \rho) := \Bigl\{ & f : f(x) = f_0(x) + g(x)I(0 \leq x \leq 1), \\ & \min_{x\in[0,1]} f(x) \geq 0, g \in \mathcal{F}(\nu, Q), \min_{x\in[0,1]} f_0(x) \geq c_0, \\ & \max_{x\in[0,1]}|g(x)| \leq \rho, \int_0^1 g(x)\,dx = 0, \int_0^1 f_0(x)\,dx = 1\Bigr\}.\end{aligned} \tag{5.1}$$

Note that we need to set $\nu = \alpha - 1$ to get a correspondence between function classes (1.3) and (5.1), and the latter explains why we use the new parameter ν and not α in (5.1).

For the class (5.1) and the case of directly observed X the minimax rate of MISE convergence is the familiar $n^{-2\nu/(2\nu+1)}$. For RMCSC, LMCSC and CSC data, the rate slows down and, according to Theorem 2.1, it is $n^{-2\nu/(2\nu+3)}$. Note that the rate is the same as for the case of estimation of a trivariate density (with ν derivatives in each variable) using direct observations. This remark sheds additional light on the "curse of CSC."

Let us explain how to construct a series density estimator for RMCSC model. Suppose that we use a basis $\{\mu_j(x), j = 0, 1, \ldots\}$ on $[0, 1]$ whose elements are differentiable on the interval. We may write down a density of interest as

$$f^X(x) = \sum_{j=0}^{\infty} \theta_j^* \mu_j(x), \quad x \in [0, 1], \tag{5.2}$$

where $\theta_j^* := \int_0^1 f^X(x)\mu_j(x)\,dx$. Furthermore, it is straightforward to write down a Fourier coefficient θ_j^* as the expectation of a function of RMCSC data. Namely, using integration by parts we get

$$\begin{aligned} \theta_j^* &= \int_0^1 f^X(x)\mu_j(x)\,dx \\ &= [\mu_j(1)F^X(1) - \mu_j(0)F^X(0)] - \int_0^1 \mu_j^{(1)}(x)F^X(x)\,dx \qquad (5.3)\\ &= \mu_j(1) - \mathbb{E}\left\{\Delta \frac{\mu_j^{(1)}(\Delta Z)}{f^Z(\Delta Z)}\right\}. \end{aligned}$$

In the last equality, we used $F^X(0) = 0$ and $F^X(1) = 1$.

Using (5.3) and the same sequences J', K', K, B_k, T_k and t_k as in Section 4, we define a data-driven density estimator

$$\begin{aligned} \tilde{f}^X(x) &:= 1 + \sum_{i=1}^{s-1} \tilde{\kappa}_i' L_i(x) + \sum_{j=1}^{J'} \tilde{\theta}_j' \psi_j(x) \qquad (5.4)\\ &\quad + \sum_{k=K'}^{K} \frac{\tilde{\Theta}_k'}{\tilde{\Theta}_k' + \tilde{d}_{\mathrm{RM}} n^{-1}} I(\tilde{\Theta}_k' > t_k n^{-1}) \sum_{j \in B_k} \tilde{\theta}_j' \psi_j(x). \end{aligned}$$

The used in (5.4) statistics are

$$\tilde{\kappa}_i' = L_i(1) - n^{-1} \sum_{l=1}^{n} \Delta_l \frac{L_i^{(1)}(\Delta_l Z_l)}{f^Z(\Delta_l Z_l)}, \tag{5.5}$$

$$\tilde{\theta}_j' = \psi_j(1) - n^{-1} \sum_{l=1}^{n} \Delta_l \frac{\psi_j^{(1)}(\Delta_l Z_l)}{f^Z(\Delta_l Z_l)}, \tag{5.6}$$

$$\begin{aligned} \tilde{\Theta}_k' &:= T_k^{-1} [n(n-1)]^{-1} \sum_{j \in B_k} \sum_{l_1 \neq l_2 = 1}^{n} \left[\psi_j(1) - \frac{\Delta_{l_1} \psi_j^{(1)}(\Delta_{l_1} Z_{l_1})}{f^Z(\Delta_{l_1} Z_{l_1})}\right] \qquad (5.7)\\ &\quad \times \left[\psi_j(1) - \frac{\Delta_{l_2} \psi_j^{(1)}(\Delta_{l_2} Z_{l_2})}{f^Z(\Delta_{l_2} Z_{l_2})}\right], \end{aligned}$$

and $\tilde{d}_{\mathrm{RM}}$ is defined in (4.1). If bona fide properties are important, then the corresponding L_2-projection should be added (see Efromovich (1999) where R-software is also available).

THEOREM 5.1. *Suppose that Assumption 1 holds, the pivotal density $f_0 \in \mathcal{F}(\nu + 1, Q')$, $\nu \geq 1$, $Q' < \infty$ and data are RMCSC. Then the density estimator (5.4) is sharp minimax,* *attains the lower bound of Theorem 2.1 and*

$$\sup_{f^X} \mathbb{E}_{f^X}\biggl\{[n\mathcal{J}(\nu+1, Q, d_{\mathrm{RM}}, 1)]^{2\nu/(2\nu+3)} \times \int_0^1 [\tilde{f}^X(x) - f^X(x))]^2\,dx\biggr\} = (1 + o_n(1)), \tag{5.8}$$

where the supremum is over $f^X \in \mathcal{F}'(f_0, \nu, Q, 0, \infty)$ *defined in* (5.1).

Note that the proposed density estimator (5.4) mimics the corresponding cdf estimator (4.3). Absolutely similarly, following the methodology of Section 4, we define density estimators for LMCSC and CSC cases.

6. Possible extensions.

6.1. *Dependent lifetime of interest and monitoring time* (*dependent CSC*). This setting is referred to as "intriguing" in Jewell and van der Laan (2004a), and see also Wang et al. (2012), Ma, Hu and Sun (2015) and Li et al. (2017) where further references may be found. It was explained at the end of Section 3 that even for a CSC with no missing dependence between X and Z precludes us from consistent estimation of the distribution of X. One of the possibilities to resolve this issue, motivated by Example 4 in the Supplementary Material (Efromovich (2021)), is to find an auxiliary variable V such that X and Z are conditionally independent given V. Then the observed sample of size n is from: $(\Delta V, \Delta Z, \Delta)$ for RMCSC, $((1-\Delta)V, (1-\Delta)Z, \Delta)$ for LMCSC and (V, Z, Δ) for CSC. Here, as usual, $\Delta := I(X \le Z)$.

Suppose that the above-outlined remedy is feasible and there exists an auxiliary continuous variable V such that X and Z are conditionally independent given V, and as before let us assume that each variable is supported on [0, 1]. The proposed solution is as follows. First of all, similar to (1.1) we note that the joint (mixed) density of the triplet (V, Z, Δ) is

$$\begin{aligned} f^{V,Z,\Delta}(v, z, \delta) &= f^V(v) f^{Z|V}(z|v)[F^{X|V}(z|v)]^{\delta} \\ &\quad \times [1 - F^{X|V}(z|v)]^{1-\delta}, \quad (v, z) \in [0, 1]^2, \delta \in \{0, 1\}. \end{aligned} \tag{6.1}$$

Here, $F^{X|V}(x|v) := \mathbb{P}(X \le x|V = v)$, f^V is the density of V, and $f^{Z|V}$ is the conditional density of Z given V. Formula (6.1) allows us to appreciate complexity of the considered dependent CSC. The key difference between (6.1) and formula (1.1) for independent CSC is that for dependent CSC we no longer have a direct access to the cdf of interest $F^X(x)$; instead, the conditional cdf $F^{X|V}(x|v)$ is directly accessible. Nonetheless, let us explain how presented in previous sections methodology may be used for estimating $F^X(x)$. Let $\varphi_0(x) = 1$, $\varphi_j(x)$, $j = 1, 2, \dots$ be elements of a basis on [0, 1]. Using formula $F^X(x) = \mathbb{E}\{F^{X|V}(x|V)\}$, we can write for Fourier coefficients of $F^X(x)$,

$$\begin{aligned} \theta_j &:= \int_0^1 F^X(x)\varphi_j(x)\,dx \\ &= \int_0^1 \mathbb{E}\{F^{X|V}(z|V)\}\varphi_j(z)\,dz = \int_0^1 \mathbb{E}\biggl\{\frac{f^{V,Z,\Delta}(V, z, 1)}{f^V(V) f^{Z|V}(z|V)}\biggr\}\varphi_j(z)\,dz \\ &= \int_{[0,1]^2} \frac{f^{V,Z,\Delta}(v, z, 1)\varphi_j(z)}{f^{Z|V}(z|v)}\,dv\,dz = \mathbb{E}\biggl\{\frac{\Delta\varphi_j(Z)}{f^{Z|V}(Z|V)}\biggr\}. \end{aligned} \tag{6.2}$$

This equation points upon a sample mean Fourier estimator of θ_j, and hence upon a consistent series estimator of $F^X(x)$ for the model of dependent RMCSC and a known conditional density $f^{Z|V}(z|v)$. The conditional density is known in controlled experiments; otherwise an extra cross-sectional sampling of the pair (V, Z) may be required, and let us stress that this sampling does not involve the hidden lifetime of interest X. All other dependent CSC models are considered similarly.

Now let us explain the heuristic of developing a lower minimax bound. Applying the Gram–Schmidt orthonormalization with weight function $f^V(x)$ to the elements $\varphi_0(x), \varphi_j(x)$, $j = 1, 2, \dots$, we create new elements $\mu_j(x)$ such that $\int_0^1 \mu_j(x)\mu_i(x) f^V(x)\,dx = I(i = j)$. Because $\int_0^1 f^V(x)\,dx = 1$, we get $\mu_0(x) = 1$ and $\int_0^1 \mu_j(x) f^V(x)\,dx = 0$ for $j \geq 1$. Set $\nu_{ji} := \int_{[0,1]^2} f^V(v) F^{X|V}(x|v)\varphi_j(x)\mu_i(v)\,dx\,dv$. Then $\theta_j = \int_0^1 F^X(x)\varphi_j(x)\,dx = \nu_{j0}$ and

$$
\begin{aligned}
F^{X|V}(x|v) &= \sum_{j,i=0}^{\infty} \nu_{ji}\varphi_j(x)\mu_i(v) \\
&= \left[\sum_{j=0}^{\infty} \theta_j \varphi_j(x)\right] + \left[\sum_{j=0}^{\infty}\sum_{i=1}^{\infty} \nu_{ji}\varphi_j(x)\mu_i(v)\right] \qquad (6.3)\\
&=: F^X(x) + q(x, v), \quad (v, x) \in [0, 1]^2.
\end{aligned}
$$

Expression (6.3) for $F^{X|V}$ as the sum of the estimand F^X and the nuisance function $q(x, v)$ is the key in suggesting assumptions that will allow us to establish a lower minimax bound.

ASSUMPTION 3. The lifetime of interest X and the auxiliary variable V are continuous random variables and each is supported on $[0, 1]$. The density $f^V(v)$ of V is continuous and positive on $[0, 1]$. The lifetime of interest X and the monitoring time Z are conditionally independent given V. A known conditional density $f^{Z|V}(z|v)$ is continuous on $[0, 1]^2$ and $\min_{(z,v)\in[0,1]^2} f^{Z|V}(z|v) \geq c_* > 0$.

ASSUMPTION 4. Let $F_0^{X|V}(x|v)$, $(x, v) \in [0, 1]^2$ be a pivotal conditional cdf satisfying $F_0^{X|V}(0|v) = 0$, $F_0^{X|V}(1|v) = 1$, $\partial F_0^{X|V}(x|v)/\partial x \geq c_2 > 0$ for $(x, v) \in [0, 1]^2$. Using formula (6.3), we define the corresponding pivotal cdf $F_0^X(x)$ and the pivotal bivariate function $q_0(x, v)$ such that

$$
\begin{aligned}
F_0^X(x) + q_0(x, v) &:= F_0^{X|V}(x|v), \\
F_0^X(x) &:= \int_0^1 f^V(v) F_0^{X|V}(x|v)\,dv, \quad (x, v) \in [0, 1]^2, \qquad (6.4)
\end{aligned}
$$

and it is assumed that F_0^X can be used as a pivotal cdf F_0 for a class $\mathcal{F}(F_0, \alpha, Q, c_0, c_1, \rho)$ defined in (1.3) (in other words, $F_0^{X|V}$ is such that the corresponding F_0^X satisfies the restrictions outlined in (1.3)). Furthermore, it is assumed that an underlying conditional cdf $F^{X|V}(x|v)$ belongs to a class

$$
\begin{aligned}
&\mathcal{D}(\mathcal{F}(F_0^X, \alpha, Q, c_0, c_1, \rho), q_0, f^V, c_2) \\
&\quad := \{F_*^{X|V} : F_*^{X|V}(x|v) = F_*(x) + q_0(x, v), (x, v) \in [0, 1]^2; \qquad (6.5)\\
&\qquad F_* \in \mathcal{F}(F_0^X, \alpha, Q, c_0, c_1, \rho)\},
\end{aligned}
$$

where $F_*^{X|V}$ are bona fide conditional cumulative distribution functions, the class $\mathcal{F}$ is defined in line (1.3) of Assumption 2 and the constant c_2 is the above-introduced lower bound for the partial derivative in x of the pivot $F_0^{X|V}(x|v)$.

For dependent RMCSC, LMCSC and CSC models, introduce the corresponding functionals

$$
\begin{aligned}
b_{\mathrm{RM}} &:= \int_0^1 \biggl[\int_0^1 \frac{f^V(v) f^{Z|V}(x|v)}{F^{X|V}(x|v)}\, dv\biggr]^{-1} dx,\\
b_{\mathrm{LM}} &:= \int_0^1 \biggl[\int_0^1 \frac{f^V(v) f^{Z|V}(x|v)}{(1-F^{X|V}(x|v))}\, dv\biggr]^{-1} dx,\\
b_{\mathrm{CSC}} &:= \int_0^1 \biggl[\int_0^1 \frac{f^V(v) f^{Z|V}(x|v)}{F^{X}(x|v)(1-F^{X|V}(x|v))}\, dv\biggr]^{-1} dx.
\end{aligned}
\tag{6.6}
$$

Recall our notation $(F^X(x))^{(\beta)}$ for the cdf and the density of X when $\beta = 0$ and $\beta = 1$, respectively.

THEOREM 6.1 (Lower minimax bound). *RMCSC, LMCSC and CSC models and estimation of the cdf or the density $(F^X(x))^{(\beta)}$, $\beta \in \{0, 1\}$ are considered. Suppose that Assumptions 3–4 hold, $\alpha \ge 1 + \beta$, and a sample of size n is available. Then the following dealer's lower bound holds*:

$$
\begin{aligned}
&\inf_{\tilde\Psi_\beta} \sup_{F^{X|V}} \mathbb{E}_{F^{X|V}} \Bigl\{ [n\mathcal{J}(\alpha, Q, b, \beta)]^{2(\alpha-\beta)/(2\alpha+1)}\\
&\qquad \times \int_0^1 [\tilde\Psi_\beta(x) - (F^X(x))^{(\beta)}]^2\, dx\Bigr\} \ge (1 + o_n(1)).
\end{aligned}
\tag{6.7}
$$

Here, the coefficient b is defined in (6.6) *for an underlying sampling model, the supremum is over $F^{X|V} \in \mathcal{D}(\mathcal{F}(F_0^X, \alpha, Q, 1/s_n, s_n^{1/2}, 1/s_n), q_0, f^V, 1/s_n)$ defined in* (6.5) *and s_n is defined at the end of the Introduction; the infimum has taken over all possible dealer-estimators $\tilde\Psi_\beta$ that know the sample, densities $f^V(v)$ and $f^{Z|V}(z|v)$, and everything else about the class* (6.5), *namely the dealer also knows $F_0^X(x)$, $q_0(x, v)$, α, Q and s_n.*

Now we can begin to explore upper bounds for oracle-estimators. Correspondingly, for dependent RMCSC, LMCSC and CSC models introduce

$$
\begin{aligned}
d^*_{\mathrm{RM}} &:= \int_{[0,1]^2} \frac{f^V(v) F^{X|V}(x|v)}{f^{Z|V}(x|v)}\, dx\, dv,\\
d^*_{\mathrm{LM}} &:= \int_{[0,1]^2} \frac{f^V(v)(1 - F^{X|V}(x|v))}{f^{Z|V}(x|v)}\, dx\, dv,\\
d^*_{\mathrm{CSC}} &:= \int_{[0,1]^2} \frac{f^V(v) F^{X|V}(x|v)(1 - F^{X|V}(x|v))}{f^{Z|V}(x|v)}\, dx\, dv.
\end{aligned}
\tag{6.8}
$$

Following (3.3), (3.4) and (6.2), define for the considered dependent RMCSC model Fourier estimators

$$
\tilde\kappa^*_{Ri} := n^{-1} \sum_{l=1}^n \frac{\Delta_l L_i(\Delta_l Z_l)}{f^{Z|V}(\Delta_l Z_l | \Delta_l V)}, \qquad \tilde\theta^*_{Rj} := n^{-1} \sum_{l=1}^n \frac{\Delta_l \psi_j(\Delta_l Z_l)}{f^{Z|V}(\Delta_l Z_l | \Delta_l V_l)},
\tag{6.9}
$$

and for dependent LMCSC

$$
\begin{aligned}
\tilde\kappa^*_{Li} &:= I(i = 0) - n^{-1} \sum_{l=1}^n \frac{(1 - \Delta_l) L_i((1 - \Delta_l) Z_l)}{f^{Z|V}((1 - \Delta_l) Z_l | (1 - \Delta_l) V_l)},\\
\tilde\theta^*_{Lj} &:= -n^{-1} \sum_{l=1}^n \frac{(1 - \Delta_l) \psi_j((1 - \Delta_l) Z_l)}{f^{Z|V}((1 - \Delta_l) Z_l | (1 - \Delta) V_l)}.
\end{aligned}
\tag{6.10}
$$

Then the oracle's cdf estimator for RMCSC and LMCSC models is

$$\tag{6.11} \begin{aligned} \tilde{F}^X_*(x) := \sum_{i=0}^{s-1} \tilde{\kappa}^*_i L_i(x) + \sum_{j=1}^{J'(n,d^*,\alpha,Q)} \tilde{\theta}^*_j \psi_j(x) \\ + \sum_{j=J'(n,d^*,\alpha,Q)+1}^{J(n,d^*,\alpha,Q)} \bigl[1 - \bigl(j/J(n,d^*,\alpha,Q)\bigr)^\alpha\bigr] \tilde{\theta}^*_j \psi_j(x). \end{aligned}$$

The used sequences and functions are defined in Section 3, and the used $\{\tilde{\kappa}^*_i, \tilde{\theta}^*_j, d^*\}$ are either $\{\tilde{\kappa}^*_{Ri}, \tilde{\theta}^*_{Rj}, d^*_{\mathrm{RM}}\}$ for RMCSC data or $\{\tilde{\kappa}^*_{Li}, \tilde{\theta}^*_{Lj}, d^*_{\mathrm{LM}}\}$ for LMCSC data. Further, the estimator for the dependent CSC data is again the aggregated one according to (3.8).

THEOREM 6.2 (Upper bound for CDF). *Dependent RMCSC, LMCSC and CSC models are considered simultaneously. Suppose that Assumptions 3–4 hold, $\alpha \geq 1$ and additionally assume that $F_0 \in \mathcal{F}(\alpha+1, Q')$ where $Q' < \infty$. Then MISEs of the above-defined oracle-estimators satisfy the following upper bound:*

$$\tag{6.12} \begin{aligned} \sup_{F^{X|V}} \mathbb{E}_{F^{X|V}} \Bigl\{ \bigl[n\mathcal{J}(\alpha, Q, d^*, 0)\bigr]^{2\alpha/(2\alpha+1)} \\ \times \int_0^1 \bigl(\bar{F}_*(x) - F^X(x)\bigr)^2 dx \Bigr\} \leq \bigl(1 + o_n(1)\bigr), \end{aligned}$$

where $\bar{F}_*(x)$ and d^* are oracle-estimators and coefficients of difficulty for the corresponding models, and the supremum is over $F^{X|V} \in \mathcal{D}(\mathcal{F}(F^X_0, \alpha, Q, 0, \infty, \infty), q_0, f^V, 0)$.

Three comments about the result are due. First, in a traditional minimax setting a class of estimands (here F^X) is considered. This approach cannot be utilized here because F^X does not define the joint distribution of (V, Z, Δ) while the conditional cdf $F^{X|V}$ does. The second comment is that there is a gap between constants b and d^* defined in (6.6) and (6.8). Using the Cauchy–Schwarz inequality, it is straightforward to establish that $b \leq d^*$ for each model. We may conclude that the dependence does not effect rates of the MISE convergence, and we know the range of possible sharp constants. The last comment is that the proposed construction of the estimator does not require knowledge of the distribution of the auxiliary variable V.

We have explained how the more complicated theoretical results may be established for dependent CSC. Further, the presented Fourier estimators allow us to use the R software of Efromovich (2018), and the discussion will be continued in Examples 4–6 of the Supplementary Material (Efromovich (2021)).

6.2. *Case-, control- and case-control CSC.* This sampling model was explained in Remark 1.1. Recall that two samples with deterministic sample sizes may be available. The first one, of size m, is a sample of monitoring times (referred to as "cases") from the subpopulation of monitoring times with already occurred events of interest, that is, with $\Delta = I(X \leq Z) = 1$. To stress that the distribution of a case is different from the underlying distribution of the monitoring time Z, let us denote the observed variable (the case) as T, and note that $f^T(t) = f^{Z|\Delta}(t|1)$. The second sample of size cm is a sample of monitoring times (referred to as "controls") from the subpopulation of monitoring times with not yet occurred events of interest, that is, from the subpopulation with $\Delta = 0$, and we denote the corresponding monitoring time (the control) as Y. The distribution of Y is the conditional distribution of Z given

$\Delta = 0$, that is, $f^Y(y) = f^{Z|\Delta}(y|0)$. Using Assumption 1 and formula (1.1), we conclude that

$$f^T(t) = f^{Z|\Delta}(t|1) = \frac{f^Z(t)F^X(t)}{p},$$
$$f^Y(y) = f^{Z|\Delta}(y|0) = \frac{f^Z(y)[1 - F^X(y)]}{1-p}, \quad p := \mathbb{P}(\Delta = 1). \tag{6.13}$$

If we compare (6.13) with (1.1), then it becomes clear that, in terms of the theory of estimating the distribution of X, the key difference between case-control and traditional CSC samples is that the latter, due to employing a binomial sampling, allows us to estimate probability $p = \mathbb{P}(\Delta = 1)$ by the sample mean $n^{-1}\sum_{l=1}^{n}\Delta_l$, while the former precludes us from estimating p, and hence from consistent estimation of the distribution of X. More discussion of this issue can be found in Jewell and van der Laan (2004b).

Despite the above-made observation, the developed in Sections 2–5 CSC theory allows us to present several interesting theoretical results for the considered models. We begin with a lower minimax bound. Recall that $p := \mathbb{P}(X \le Z) = \mathbb{P}(\Delta = 1)$, constant c is the factor which defines the size cm of the sample of controls, coefficients of difficulty d_{RM} for RMCSC and d_{LM} for LMCSC are defined in line (2.2) of Section 2, and introduce a new coefficient of difficulty

$$d_{\mathrm{CC}} := \int_0^1 \frac{p(1-p)F^X(x)(1-F^X(x))}{f^Z(x)[(1-p)(1-F^X(x)) + cpF^X(x)]}\,dx. \tag{6.14}$$

THEOREM 6.3 (Lower bound). *Models of a case-CSC sampling which collects m cases, a control-CSC sampling which collects cm controls and a case-control CSC sampling which collects m cases and cm controls, are considered simultaneously. Suppose that Assumptions 1 and 2 hold and $\alpha \ge 1 + \beta$. Then the lower bound (2.3) is valid with the following modifications: for the case-CSC replace (n, d) with (m, pd_{RM}); for the control-CSC replace (n, d) with $(cm, (1-p)d_{\mathrm{LM}})$; for the case-control CSC replace (n, d) with (m, d_{CC}).*

As it was explained earlier, the probability $p := \mathbb{P}(X \le Z)$ becomes a main player in constructing a consistent estimator and establishing an upper bound. If p is known (see a discussion in Jewell and van der Laan (2004b)), then a modification of the corresponding efficient estimators of Sections 3–5 is straightforward. For instance, for a case-CSC sample of size m we replace (3.3) by

$$\tilde{\kappa}_i := m^{-1}\sum_{l=1}^{m}\frac{pL_i(T_l)}{f^Z(T_l)}, \qquad \tilde{\theta}_j := m^{-1}\sum_{l=1}^{m}\frac{p\psi_j(T_l)}{f^Z(T_l)}, \tag{6.15}$$

and for a control-CSC sample of size cm we replace (3.4) by

$$\tilde{\kappa}_i := I(i=0) - (cm)^{-1}\sum_{l=1}^{cm}\frac{(1-p)L_i(Y_l)}{f^Z(Y_l)},$$
$$\tilde{\theta}_j := -(cm)^{-1}\sum_{l=1}^{cm}\frac{(1-p)\psi_j(Y_l)}{f^Z(Y_l)}. \tag{6.16}$$

To aggregate these two samples of cases and controls and consider a case-control CSC, we may use the approach of Section 3 with (3.6) being replaced by

$$\check{\theta}_{rj} := \lambda_r m^{-1}\sum_{l=1}^{m}\frac{pI(T_l \in [b_r, b_{r+1}))\psi_{rj}(T_l)}{f^Z(T_l)}$$
$$- (1-\lambda_r)(cm)^{-1}\sum_{l=1}^{cm}\frac{(1-p)I(Y_l \in [b_r, b_{r+1}))\psi_{rj}(Y_l)}{f^Z(Y_l)}, \tag{6.17}$$

and (3.7) being replaced by

$$\lambda_r := \frac{\int_{b_r}^{b_{r+1}}[c(1-p)(1-F_0^X(x))/f^Z(x)]\,dx}{\int_{b_r}^{b_{r+1}}[(pF_0^X(x)+c(1-p)(1-F_0^X(x)))/f^Z(x)]\,dx}. \tag{6.18}$$

THEOREM 6.4. *Let probability $p := \mathbb{P}(\Delta = 1)$ of the "case" be known. Consider cdf estimators of Section 3 with the proposed Fourier estimators (6.15)–(6.17) and weights (6.18). Then the assertions of Theorems 3.1–3.2 hold for case-CSC, control-CSC and case-control CSC samples with corresponding modifications of (n, d) defined in Theorem 6.3.*

In an absolutely similar way, all other previously discussed settings may be considered; in other words, the developed methodology is directly applicable to a case and control CSC whenever p is known.

Probability $p := \mathbb{P}(\Delta = 1)$ is typically unknown, and then in general, additional information is needed for consistent estimation of the distribution of X. Nonetheless, the above-presented estimators shed light on an attractive possibility to estimate the so-called shape of a curve of interest. Indeed, if for instance we check formula (6.15) then we may notice that all Fourier coefficients are proportional to p. As a result, even if p is unknown we can estimate shape of the curve. Of course, similar to missing CSC, we still need to know density f^Z. The interested reader can find continuation of the discussion in Example 7 of the Supplementary Material (Efromovich (2021)).

6.3. *Doubly CSC (DCSC).* Consider a CSC sampling when we collect observations from (Z, Δ), and here as before $\Delta := I(X \le Z)$. If it is additionally known that $X = B + T$, where B is a nuisance random lifetime and T (not X) is the lifetime of interest whose distribution we would like to estimate, then the model is called doubly CSC. Well-understood examples of T are the time to divorce after marriage at age B, or how long it takes to get sick after exposure to a virus. More examples and a discussion of DCSC can be found in Jewell and van der Laan (2004a), Li et al. (2020) and Malov (2019) where further references may be found. To shed light on DCSC, consider as an example an underlying RMCSC model and then write down an analog of (1.1),

$$\begin{aligned} f^{Z,\Delta}(z,1) &= f^Z(z)\mathbb{P}(B+T \le z|Z=z) \\ &= f^Z(z)\int_0^z F^{T|Z,B}(z-b|z,b)\,dF^{B|Z}(b|z). \end{aligned} \tag{6.19}$$

This formula explains complexity of DCSC, and it also sheds light on a possible solution for some settings considered in the DCSC literature. For instance, consider a DCSC model where T and $Z - B$ are mutually independent and B is observed. In this case, the induced monitoring time of T is $Z - B$ and we can get a sample from $(Z - B, I(T \le Z - B))$. As a result, the DCSC problem becomes the already considered CSC problem with $Z - B$ being the induced (and observable) monitoring time. If B is not observed but its distribution is known and (B, T, Z) are mutually independent, then (6.19) implies that we need to solve a deconvolution problem. It is known that the deconvolution problem at hand is ill-posed even if $X = B + T$ is observed directly; see a discussion in Efromovich (1999). The problem becomes even more challenging when the assumption of independence is relaxed and no longer nuisance distributions are known. Then extra information is needed for consistent estimation. Another often considered DCSC model is when a sample from (Δ', Δ, Z) is available where $\Delta' := I(B \le Z)$ and $\Delta := I(X \le Z)$.

It is an open, challenging and practically important problem to develop theory of efficient estimation for DCSC data with unobserved nuisance variables and unknown nuisance distributions.

Acknowledgments. The author would like to thank the Editors, the Associate Editor and the referees for their helpful comments and suggestions which have improved the exposition and presentation of this paper.

The work is supported by NSF Grants DMS-1513461, DMS-1915845 and Grants from BIFAR and the Casualty Actuarial Society.

SUPPLEMENTARY MATERIAL

Supplement to "Sharp minimax distribution estimation for current status censoring with or without missing" (DOI: 10.1214/20-AOS1970SUPP; .pdf). Online supplementary material contains examples, proofs, useful facts about Legendre polynomials and additional references.

REFERENCES

BECKER, D. G., BRAUN, W. J. and WHITE, B. J. G. (2017). Interval-censored unimodal kernel density estimation via data sharpening. *J. Stat. Comput. Simul.* **87** 2023–2037. MR3647585 https://doi.org/10.1080/00949655.2017.1308510

CAI, T. T. (2012). Minimax and adaptive inference in nonparametric function estimation. *Statist. Sci.* **27** 31–50. MR2953494 https://doi.org/10.1214/11-STS355

CHEN, D.-G., SUN, J. and PEACE, K. E., eds. (2012). *Interval-Censored Time-to-Event Data: Methods and Applications. Chapman & Hall/CRC Biostatistics Series*. CRC Press, Boca Raton, FL. MR3053014

CHICKEN, E. and CAI, T. T. (2005). Block thresholding for density estimation: Local and global adaptivity. *J. Multivariate Anal.* **95** 76–106. MR2164124 https://doi.org/10.1016/j.jmva.2004.07.003

DIAO, G. and YUAN, A. (2019). A class of semiparametric cure models with current status data. *Lifetime Data Anal.* **25** 26–51. MR3896658 https://doi.org/10.1007/s10985-018-9420-0

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539. MR1394974 https://doi.org/10.1214/aos/1032894451

EFROMOVICH, S. Y. (1985). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30** 557–568.

EFROMOVICH, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications. Springer Series in Statistics*. Springer, New York. MR1705298

EFROMOVICH, S. (2018). *Missing and Modified Data in Nonparametric Estimation: With R Examples. Monographs on Statistics and Applied Probability* **156**. CRC Press, Boca Raton, FL. MR3752670

EFROMOVICH, S. (2021). Supplement to "Sharp minimax distribution estimation for current status censoring with or without missing." https://doi.org/10.1214/20-AOS1970SUPP

GINÉ, E. (1975). Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev norms. *Ann. Statist.* **3** 1243–1266. MR0388663

GROENEBOOM, P. and HENDRICKX, K. (2018). Current status linear regression. *Ann. Statist.* **46** 1415–1444. MR3819105 https://doi.org/10.1214/17-AOS1589

GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints: Estimators, Algorithms and Asymptotics. Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293 https://doi.org/10.1017/CBO9781139020893

GROENEBOOM, P., JONGBLOED, G. and WITTE, B. I. (2010). Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *Ann. Statist.* **38** 352–387. MR2589325 https://doi.org/10.1214/09-AOS721

HALL, P., KERKYACHARIAN, G. and PICARD, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26** 922–942. MR1635418 https://doi.org/10.1214/aos/1024691082

HSU, L., GORFINE, M. and ZUCKER, D. (2018). On estimation of the hazard function from population-based case-control studies. *J. Amer. Statist. Assoc.* **113** 560–570. MR3832208 https://doi.org/10.1080/01621459.2017.1356315

JEWELL, N. P. and VAN DER LAAN, M. (2004a). Current status data: Review, recent developments and open problems. In *Advances in Survival Analysis. Handbook of Statist.* **23** 625–642. Elsevier, Amsterdam. MR2065792 https://doi.org/10.1016/S0169-7161(03)23035-2

JEWELL, N. P. and VAN DER LAAN, M. (2004b). Case-control current status data. *Biometrika* **91** 529–541. MR2090620 https://doi.org/10.1093/biomet/91.3.529

KEOGH, R. H. and COX, D. R. (2014). *Case-Control Studies. Institute of Mathematical Statistics (IMS) Monographs* **4**. Cambridge Univ. Press, Cambridge. MR3443808 https://doi.org/10.1017/CBO9781139094757

KLEIN, J. P., VAN HOUWELINGEN, H. C., IBRAHIM, J. G. and SCHEIKE, T. H., eds. (2014). *Handbook of Survival Analysis*. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3287588

LI, S., HU, T., WANG, P. and SUN, J. (2017). Regression analysis of current status data in the presence of dependent censoring with applications to tumorigenicity experiments. *Comput*. *Statist*. *Data Anal*. **110** 75–86. MR3612609 https://doi.org/10.1016/j.csda.2016.12.011

LI, S., HU, T., ZHAO, X. and SUN, J. (2019). A class of semiparametric transformation cure models for interval-censored failure time data. *Comput*. *Statist*. *Data Anal*. **133** 153–165. MR3926472 https://doi.org/10.1016/j.csda.2018.09.008

LI, S., SUN, J., TIAN, T. and CUI, X. (2020). Semiparametric regression analysis of doubly censored failure time data from cohort studies. *Lifetime Data Anal*. **26** 315–338. MR4079669 https://doi.org/10.1007/s10985-019-09477-x

MA, L., HU, T. and SUN, J. (2015). Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika* **102** 731–738. MR3394289 https://doi.org/10.1093/biomet/asv020

MALOV, S. V. (2019). Nonparametric estimation for a current status right-censored data model. *Stat*. *Neerl*. **73** 475–495. MR4023706 https://doi.org/10.1111/stan.12180

PINSKER, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl*. *Inf*. *Transm*. **16** 52–68. MR0624591

SUN, J. (2007). *The Statistical Analysis of Interval-Censored Failure Time Data*. *Statistics for Biology and Health*. Springer, New York. MR2287318

SUN, J. and ZHAO, X. (2013). *Statistical Analysis of Panel Count Data*. *Statistics for Biology and Health*. Springer, New York. MR3136574 https://doi.org/10.1007/978-1-4614-8715-9

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

VAN ES, B. and GRAAFLAND, C. (2017). Nonparametric kernel density estimation for univariate current status data. Preprint. Available at arXiv:1707.00544v1.

VANDENBROUCKE, J. and PEARCE, N. (2014). Case-control studies: Basic concepts. *Int*. *J*. *Epidemiol*. **41** 1480–1489.

WANG, C., SUN, J., SUN, L., ZHOU, J. and WANG, D. (2012). Nonparametric estimation of current status data with dependent censoring. *Lifetime Data Anal*. **18** 434–445. MR2973774 https://doi.org/10.1007/s10985-012-9223-7

ZHANG, C.-H. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann*. *Statist*. **33** 54–100. MR2157796 https://doi.org/10.1214/009053604000000995