Study of imputation procedures for nonparametric density estimation based on missing censored lifetimes

Sam Efromovich [a,*], Lirit Fuksman[a]

[a]*Department of Mathematical Sciences, The University of Texas at Dallas, 800 W Campbell Rd, Richardson, 75080, TX, USA*

Abstract

Imputation is a standard procedure in dealing with missing data and there are many competing imputation methods. It is proposed to analyze imputation procedures via comparison with a benchmark developed by the asymptotic theory. Considered model is nonparametric density estimation of the missing right censored lifetime of interest. This model is of a special interest for understanding imputation because each underlying observation is the pair of censored lifetime and indicator of censoring. The latter creates a number of interesting scenarios and challenges for imputation when best methods may or may not be applicable. Further, the theory sheds light on why the effect of imputation depends on an underlying density. The methodology is tested on real life datasets and via intensive simulations. Data and R code are provided.

Keywords: Adaptation, Imputation, Integrated squared error, Missing data, Survival analysis.

1. Introduction

We are interested in nonparametric estimation of the probability density f^T of a lifetime of interest T over an interval $[0, b]$ under the mean integrated squared error (MISE) criterion given that observations of the lifetime are right censored and may be missing. The aim is to develop both theory and methodology of optimal estimation that match known results for models of direct observations of T, missing T, and right censored T. Then we explore and compare popular imputation methods using numerical simulations. Accordingly, in the following first subsection a review of known results for the above-mentioned classical sampling models is presented, and then the second subsection presents the studied problem as well as the terminology and notations used in the paper.

1.1. Review of known results

A standard statistical model is when estimation of density f^T is based on a sample $T_1, \ldots, T_n$ of size n of i.i.d. observations from T. Assuming that the density is α-fold differentiable on $[0, b]$, it is known that the optimal rate of the MISE convergence is $n^{-2\alpha/(2\alpha+1)}$ and there are many different nonparametric estimators, including kernel, spline and series ones, that attain this rate, see Wasserman (2005). Further, for the Sobolev class of α-fold differentiable densities even the smallest constant of the MISE convergence, referred to as the sharp constant, is known. Definition

*Corresponding author

Email address: efrom@utdallas.edu (Sam Efromovich)

of the Sobolev class of densities and formula for the sharp constant will be given shortly in Section 2, and the sharp constant and the optimal rate are attainable by a special adaptive series estimator proposed in Efromovich (1985), see also a discussion in Tsybakov (2009). Two classical modifications of the standard sampling are missing lifetimes and right censored lifetimes that will be reviewed in turn.

The missingness means that we observe a sample from (AT, A) where A is the Bernoulli random variable called the availability and $\mathbb{P}(A = 1|T = t) = w(t) \in (0, 1]$ is called the availability likelihood. Accordingly, some observations of T may be not available and in R they are denoted as "NA" (Not Available) with the corresponding availablity $A = 0$, while available observations of T correspond to $A = 1$. The model with a constant availability likelihood $w(t) = w$ is called Missing Completely At Random (MCAR), and then consistent estimation of f^T is possible. Further, the above-mentioned α-fold differentiable Sobolev densities can be estimated with the optimal rate $n^{-2\alpha/(2\alpha+1)}$ known for direct observations, Efromovich (2018). If the availability likelihood depends on t, that is, the probability of missing T depends on its value, then the missing mechanism is called Missing Not At Random (MNAR) and consistent density estimation is impossible. In this case, extra information would be needed to estimate the availability likelihood. Now let us assume that we observe a sample from a triplet (AT, A, X) such that the availability likelihood $\mathbb{P}(A = 1|T = t, X = x) = w(x)$, that is, the missing mechanism is defined by the always observed variable X but not by the missing T. This missing mechanism is called Missing At Random (MAR), and consistent estimation of the density is possible. Little and Rubin (2002), Molenberghs and Kenward (2007) and Little (2021) give a nice overview of these three missing models. Now, let us briefly explain the main approaches for dealing with missing data. (1) Complete data approach, also referred to as ignore missing approach, discards missing cases and works only with complete ones. This is the default method in many statistical software including R. For the density estimation and MCAR, this approach is optimal according to Efromovich (2013). (2) Weighting procedures, and in particular propensity weighting, is another popular approach. It is a universal method that also can be used for analysis of censored data, see a discussion in Little and Rubin (2002), Satten and Datta (2001), Rotnitzky and Robins (2005). For instance, let the unbiased estimator $n^{-1}\sum_{l=1}^{n} g(T_l)$ of $\mathbb{E}\{g(T)\}$ be recommended for a traditional sample from T. Then for missing data an estimator $[\sum_{l=1}^{n} \pi_{nl}^{-1} g(T_l)]/[\sum_{l=1}^{n} \pi_{nl}^{-1}]$ is used with some specially chosen weights π_{nl}^{-1}. Examples of using this methodology for missing censored data will be presented shortly in Section 2. (3) Imputation–Based procedure is another popular approach for dealing with missing data. The underlying idea is to fill missing values of T using available data and then analyze the resultant completed data using standard methods, see Rubin (1987), Little and Rubin (2002), Huque et al. (2018), van Buuren (2018), He et al. (2021), Heymans and Twisk (2022) and Moghaddam et al. (2022). Description of frequently used imputation procedures can be found in Section 3. (4) Model-Based procedures when an estimator is based on a known underlying model, and then making inferences using the likelihood or Bayesian approaches, see a nice discussion in Schafer (1997), Little and Rubin (2002) and Ungolo et al. (2019). Finally, let us comment about the terminology and notations used in different books. The above-presented terminology and notations follow Efromovich (2018). At the same time the reader may be familiar with other terminology and notations used in other books. For instance, $M := 1 - A$ is called the missing data indicator in Little and Rubin (2002), $R := A$ is called the response indicator and $1 - w$ is called the missing probability in van Buuren (2018), R is called the missing data indicator and w is called the density of the missingness process in Molenberghs and Kenward (2007), $\xi := A$ is called

the missingness indicator and $\pi := w$ is called a specific nuisance parameter in Comte, Guilloux and Brunel (2014). Let us note that notation π for availability likelihood is used by many authors, but in this paper π is the Archimedes constant which is approximately equal to 3.14159.

Next we consider the right censoring when instead of observing a sample from T we have a sample from the pair $(V, \Delta) := (\min(T, C), I(T \leq C))$. Here C is the continuous censoring lifetime independent of T, $I(\cdot)$ is the indicator function, and Δ is the indicator of censoring or simply the indicator. It is fair to say that the modern survival analysis, and nonparametric distribution estimation in particular, are based on the pathbreaking product-limit methodology of Kaplan and Meier (1958) for nonparametric estimation of survival function. Analogous to the classical empirical survival function, which is a step function that jumps down by n^{-1} at each ordered observation, Kaplan-Meier estimator is a step function that jumps down only at uncensored observations and magnitude of each jump is defined by all observations, see a discussion and examples in Chapter 2 of Legrand (2021). Further, similar to the empirical cumulative distribution function, the Kaplan-Meier estimator has optimal asymptotic properties as shown in Wellner (1982). Let us also note that statistical analysis of product-limit estimators is based on the beautiful and mathematically involved theory of counting processes and martingales discussed in Flemming and Harington (1991) and Aalen, Borgan and Gjiessing (2008), and see also interesting results in Tsiatis (2006) and Gijbels, Lin and Ying (2007). There is a vast literature devoted to density estimation reviewed in Efromovich (2018), with more relevant being Efromovich (2001) where, using the Kaplan–Meier estimator, it is shown that for α-fold differentiable Sobolev densities the censoring does not slow down the rate $n^{-2\alpha/(2\alpha+1)}$ of the MISE convergence but affects the sharp constant. Finally, to round up our discussion of censored data and connect it to the above-discussed missing data, let us note that the popular Buckley-James imputation methodology, introduced in Buckley and James (1979), treats censored observations as missing and then replaces them by statistics calculated using uncensored observations, that is, by using the imputation technique.

Now we are in a position to consider more complicated statistical models where both censoring and missing are present. We begin with the model of missing indicators of censoring. In this model, using the above-introduced notations, a sample from the triplet $(V, A\Delta, A)$ is observed. McKeague and Subramanian (1998) developed a modification of the Kaplan-Meier estimator for the MCAR model when the availability likelihood $w(v, \delta) := \mathbb{P}(A = 1|V = v, \Delta = \delta)$ is a constant. Subramanian (2004) and Subramanian (2006) introduced survival function estimators for missing censoring indicators, demonstrating its achievement of the lower bound established in van der Laan and McKeague (1998). Dikta (1998) proposed a convenient framework for estimating the survival function in cases of missing censoring indicators, employing a semiparametric random censorship model that assumes a parametric model for the conditional probability of uncensored observations and estimates the parametric component through maximum likelihood estimation. Subramanian (2003) investigated the semiparametric estimator under MCAR and MAR scenarios, proving that it is more asymptotically efficient than the nonparametric estimator suggested by van der Laan and McKeague (1998). Subramanian (2009) and Subramanian (2011) explored multiple imputation methods where missing censoring indicators were imputed to form several completed datasets, Kaplan-Meier and semiparametric random censorship estimators were then averaged, and their asymptotic properties were derived. Subramanian and Zhang (2013) developed a novel model-based approach for constructing simultaneous confidence bands for the survival function.

Under the MAR, a conditional mean score and mean score density estimators are proposed and their strong consistency is established in Wang et al. (2009). A thorough statistical analysis of hazard rate projection estimators under the MAR is performed in Comte, Guilloux and Brunel (2014) where it is established that the rate $n^{-2\alpha/(2\alpha+1)}$ is still attainable and imputation may be beneficial. Wavelet density estimators are proposed in Zou and Liang (2020), and an adaptive cosine series density estimator in Efromovich (2018) where the corresponding R software can be found. Interesting discussion of imputation versus adaptation for testing hypotheses can be found in Cuparic and Milosevic (2023).

Less is known about a setting when V is missing. Hu, Lawless and Suzuki (1998) consider the case, motivated by analysis of product warranty data, when sampling is from (AV, Δ, A) and the availability likelihood $\mathbb{P}(A = 1|V, \Delta = 0) = 0$ and $\mathbb{P}(A = 1|V, \Delta = 1) = 1$. In other words, all uncensored lifetimes are available but all censored ones are missing. Assuming that the distribution of censoring variable C is known, several estimators of the cumulative distribution function are proposed and analyzed. Further developments can be found in Wang (2008). This is a practically important model of missing lifetimes but it requires knowledge of the distribution of C. In the paper we are considering settings with missing V that allow us to propose data-driven estimates, and accordingly it will be always assumed that the availability likelihood is positive. The next subsection describes studied models.

1.2. Considered models, terminology and notations

Recall that the (right) censoring implies that instead of a sample from the lifetime of interest T there is a sample from the pair $(V, \Delta) = (\min(T, C), I(T \le C))$ where C is the censoring lifetime. Then the observation V may be missed, and the presence of the indicator of censoring Δ creates two interesting and practically important settings. The first one is when the available sample is from (AV, Δ, A), it is referred to as the *partial* missing model, and recall that A is a Bernoulli random variable called the availability. The second setting is when the available sample is from $(AV, A\Delta, A)$ and it is referred to as the *complete* missing model. Examples of these two settings will be presented in Section 3. For the both settings the MCAR (missing completely at random) means that $\mathbb{P}(A = 1|V, \Delta) = w > 0$. The partial missing model is MAR (missing at random) if $\mathbb{P}(A = 1|V, \Delta = \delta) = w(\delta) > 0$ and it is MNAR (missing not at random) otherwise. If the complete missing model is not MCAR, then it is MNAR.

We are interested in estimation of the density $f^T(t), t \in [0, b]$ under the MISE (mean integrated squared error) criterion for the above-defined partial and complete missing models. The aims are: (1) Develop sharp asymptotic theory that allows us to propose a data-driven density estimator that attains both the sharp constant and optimal rate of the MISE convergence; (2) Propose methodology of density estimation for practically important small samples; (3) Explore popular imputation procedures; (4) Test results on real data and simulated examples.

Let us comment on the first three aims in turn. We know from subsection 1.1 that neither censoring nor MAR affects the optimal rate $n^{-2\alpha/(2\alpha+1)}$ of MISE convergence for α-fold differentiable densities f^T but the sharp constants are different and reflect the corresponding loss of information. It is of interest to understand how missing censored lifetimes affect the MISE convergence. Let us also comment on the used MISE criterion. It is a familiar argument in the literature that imputation is important for unbiased estimation, and this is a valid point for estimation of parameters, see Rubin (1987), Efron (1994), van Buuren (2018), He et al. (2021), Little (2021), Heymans and Twisk (2022),

Moghaddam et al. (2022). There are no unbiased nonparametric estimators of the density even for the simplest case of a fully observed sample from T. This is why the MISE criterion, which balances variance and squared bias, is used, see a discussion in Wasserman (2005) and Tsybakov (2009). For the second aim we will use an appropriately modified methodology of E-estimation proposed in Efromovich (2018). For the third aim, recall that even for the simplest parametric problems with classical point estimators, opinion about what method of imputation is the best diverges rather dramatically as shown in the literature reviews presented in Little and Rubin (2002), Molenberghs and Kenward (2007), Bert (2018), Huque et al. (2018), van Buuren (2018), Aust (2021), Little (2021). The approach used in the paper is to use the above-mentioned methodology of E-estimation for creating a benchmark estimator that can be used for both missing censored lifetimes and imputed censored lifetimes. This will allow us to conduct a series of numerical experiments for analysis of different imputation procedures via their comparison with the Benchmark. Just to warm up the reader's interest, it will be shown shortly that an imputation procedure may be good or not so good depending on an underlying density f^T. Why? The theory and simulations show that smoothness of the density plays an important role in performance of imputation, and interestingly enough, the rougher the density the better performance of imputation is. Further, we will be able to compare both theoretical and numerical effects of the availability likelihood on the quality of estimation.

The content of the paper is as follows. Section 2 presents the theory that shed light on the problem and justifies the proposed methodology of numerical analysis of imputation methods. In Section 3 the developed methodology is tested on simulated examples and then on real-life ones. Proofs are in Section 4 and the Conclusion is in Section 5. Colored figures of the paper, more simulation results, datasets, and R code are in the online Supplementary Material.

Finally, let us present several general notations used in the paper. The probability density of a continuous lifetime T is denoted as f^T. The corresponding survival function and the cumulative hazard are $S^T(t) := \int_t^\infty f^T(v)dv$ and $H^T(t) = \int_0^t [f^T(v)/S^T(v)]dv$, respectively. The probability and the expectation are denoted by $\mathbb{P}$ and $\mathbb{E}$, respectively. The density of interest f^T is estimated over a finite interval $[0, b]$. The recommended density estimator is a series estimator that approximates an underlying density via the cosine basis on $[0, b]$ with elements $\varphi_0(t) := b^{-1/2}, \varphi_j(t) := (2/b)^{1/2}\cos(\pi jt/b), j = 1, 2, \ldots$ Namely, the Fourier theorem yields that the density can be written as $f^T(t) = \sum_{j=0}^\infty \theta_j\varphi_j(t)$, $t \in [0, b]$ where $\theta_j := \int_0^b f^T(t)\varphi_j(t)dt$ are called the Fourier coefficients. Now note that the formula $\theta_j = \mathbb{E}\{I(T \le b)\varphi_j(T)\}$ holds. Accordingly, a series density estimator uses a sample mean Fourier estimator of θ_j, and then selects a finite number of significant Fourier coefficients, see Wasserman (2005) and Efromovich (2018). Further, B denotes a generic positive finite constant, $o_j(1)$ denotes a generic vanishing sequence as $j \to \infty$, $q_n := \lceil \ln(n + 3) \rceil$ and $\lceil x \rceil$ is the smallest integer larger or equal to x.

2. Theory

The aim is to estimate density f^T over an interval $[0, b]$ under the MISE criterion and to understand how missingness of censored lifetimes affects the MISE. We begin with the MCAR setting when $\mathbb{P}(A = 1|V, \Delta) = w > 0$. The MCAR allows us to consider both the partial and complete missing models when samples of size n are from (AV, Δ, A) and $(AV, A\Delta, A)$, respectively, and to develop sharp minimax theory of density estimation matching the results for di-

rect observations of the lifetime of interest mentioned in the Introduction. Then we consider the MAR setting where $\mathbb{P}(A = 1|V, \Delta = \delta) = w(\delta) > 0$, and recall that in this case the density is estimable only for the partial missing model.

2.1. MCAR

We begin with assumptions, that are followed by the sharp lower bound for the MISE. Then the proposed methodology of density estimation is presented.

Assumption 1. *The continuous lifetime of interest T and the continuous censoring lifetime C are independent and $S^C(b) > 0$. The availability A is Bernoulli(w) with $w \in (0, 1]$.*

This assumption is necessary for consistent estimation. The next assumption is a classical one about smoothness of underlying density f^T, see Nikolskii (1975), Golubev (1992), Hoffman and Lepski (2002), Section 3.1 in Tsybakov (2009), Efromovich (2022).

Assumption 2. *An underlying density of interest f^T belongs to the shrinking, toward continuous anchor density $f_0(t)$, $t \in [0, \infty)$, local Sobolev class*

$$\mathcal{F}_n(\alpha, Q, f_0) := \{f^T : f^T(t) = f_0(t) + g(t)I(t \in [0, b]), \max_{t\in[0,b]} |g(t)|/f_0(t) \le 1/q_n, \ g \in \mathcal{S}_1(\alpha, Q), t \ge 0\}. \tag{1}$$

Here for $k \in \{0, 1\}$ function classes $\mathcal{S}_k(\alpha, Q)$ are defined via corresponding ellipsoids of Fourier coefficients,

$$\mathcal{S}_k(\alpha, Q) := \{g : g(t) = \sum_{j=k}^{\infty} \mu_j \varphi_j(t), t \in [0, b], \ \sum_{j=k}^{\infty} (1 + \pi j/b)^{2\alpha} \mu_j^2 \le bQ < \infty\}. \tag{2}$$

Let us make several remarks about Assumption 2. A function from the class $\mathcal{S}_1(\alpha, Q)$ is integrated to zero over interval $[0, b]$ because $\int_0^b \varphi_j(t)dt = 0$ for $j \ge 1$. The class $\mathcal{S}_0(\alpha, Q)$ of probability densities is called the global Sobolev class and it will be used shortly. The class $\mathcal{F}_n(\alpha, Q, f_0)$ consists of additive perturbations of the anchor density f_0, and all functions f^T from this class are bona fide densities because we have $\int_0^\infty f^T(t)dt = 1$ and $f^T(t) \ge 0$. The anchor is not necessarily the underlying density, its role is to define a sharp constant of the MISE convergence.

To develop the asymptotic theory, we use a so-called oracle approach, discussed in Efromovich (1999) and Tsybakov (2009). The oracle may know everything about the underlying problem and thus helps us to develop a lower bound for the risk. Namely, the oracle knows all nuisance functions and smoothness of the estimated nonparametric curve. The oracle also proposes an efficient oracle-estimator that can be mimicked by a data-driven estimator.

Our first theoretical result is about oracle-estimators, and we need the following assumption about the oracle.

Assumption 3. *The oracle knows the model of missing, which can be either partial or complete missing. The oracle also knows the sample of size n, distribution of censoring lifetime C, and Assumptions 1 and 2 including knowing all of the involved nuisance parameters and functions (w, α, Q, f_0).*

According to Assumption 3, apart of the underlying density, oracle-estimators know everything about data, even how smooth the underlying density is and what is the censoring mechanism. The following theorem establishes a

lower bound for MISEs of oracle-estimators, that is, we will know what the best oracle-estimator can achieve. Then it will be shown that an estimator can match the best oracle-estimator.

Theorem 1. *Let Assumptions 1-3 hold. Then the following lower bound holds,*

$$\inf_{\tilde{f}^*}\sup_{f^T\in\mathcal{F}_n(\alpha,Q,f_0)}\Big[n^{-1}b^{-1}\int_0^b\frac{f^T(t)}{S^C(t)}dt\Big]^{-2\alpha/(2\alpha+1)}\mathbb{E}_{f^T}\{\int_0^b(\tilde{f}^*(t)-f^T(t))^2dt\}$$

$$\geq w^{-2\alpha/(2\alpha+1)}P(\alpha,Q,b)(1+o_n(1)). \tag{3}$$

Here the infimum is over all oracle-estimators $\tilde{f}^*$ and

$$P(\alpha,Q,b):=Q^{1/(2\alpha+1)}b(2\alpha+1)^{1/(2\alpha+1)}\Big[\frac{\alpha}{\pi(\alpha+1)}\Big]^{2\alpha/(2\alpha+1)}. \tag{4}$$

Our next theorem shows that the oracle's lower bound is sharp, that is, attainable by a special orthonormal series oracle-estimator. Then this oracle-estimator will guide us toward a data-driven estimator that also attains this lower bound. The underlying motivation of the series oracle-estimator is that the Fourier coefficient $\theta_j:=\int_0^b f^T(t)\varphi_j(t)dt$ can be written as the expectation

$$\theta_j=\mathbb{E}\Big\{\frac{A\Delta\varphi_j(AV)I(AV\in[0,b])}{wS^C(AV)}\Big\},$$

and accordingly may be estimated by the sample mean oracle-estimate

$$\check{\theta}_j^*:=n^{-1}\sum_{l=1}^n\frac{A_l\Delta_l\varphi_j(A_lV_l)I(A_lV_l\in[0,b])}{wS^C(A_lV_l)}. \tag{5}$$

Note that if we have a sample $T_1,\dots,T_n$ from T, then the sample mean estimate of θ_j is $n^{-1}\sum_{l=1}^n\varphi_j(T_l)I(T_l\in[0,b])$. Accordingly, (5) can be viewed as a special weighting procedure reviewed in the Introduction.

Now set $d:=b^{-1}\int_0^b\frac{f^T(t)}{wS^C(t)}dt$, $J_n:=\Big\lfloor\Big[\frac{(2\alpha+1)(\alpha+1)}{\alpha}\Big]^{1/(2\alpha+1)}\times\Big[\frac{b}{\pi}\Big]^{\frac{2\alpha}{2\alpha+1}}\Big[\frac{bQn}{d}\Big]^{1/(2\alpha+1)}\Big\rfloor$, $\sigma_j^2:=\mathbb{E}\{(\check{\theta}_j^*-\theta_j)^2\}$, and define the oracle-estimator of the density,

$$\check{f}_*^T(t):=\sum_{j=0}^{q_n}\check{\theta}_j^*I([\check{\theta}_j^*]^2>2q_n\sigma_j^2)\varphi_j(t)+\sum_{j=q_n+1}^{J_n}(1-(j/J_n)^\alpha)\check{\theta}_j^*\varphi_j(t). \tag{6}$$

Note that the oracle does not use the anchor density f_0 in (6) because the oracle assumes that the anchor is smoother than an underlying density. Accordingly, one more assumption is needed.

Assumption 4. *The anchor density $f_0\in\mathcal{S}_0(\beta,Q')$ with $\beta>\alpha$, $Q'<\infty$.*

The following proposition shows that the oracle-estimator (6) is efficient.

Theorem 2. *Suppose that Assumptions 1-4 hold. Then the lower bound (3) is attainable by the oracle-estimator $\check{f}_*^T$, that is, the local lower bound is sharp. Further, this oracle-estimator also attains this lower bound over the global*

Sobolev class $\mathcal{S}_0(\alpha, Q)$, namely

$$\sup_{f^T\in\{\mathcal{F}_n(\alpha,Q,f_0)\cup\mathcal{S}_0(\alpha,Q)\}}\Big[n^{-1}b^{-1}\int_0^b\frac{f^T(t)}{S^C(t)}dt\Big]^{-2\alpha/(2\alpha+1)}\mathbb{E}_{f^T}\{\int_0^b(\check{f}_*^T(t)-f^T(t))^2dt\}\le w^{-2\alpha/(2\alpha+1)}P(\alpha,Q,b)(1+o_n(1)). \quad (7)$$

Let us formulate two important corollaries mentioned in the Introduction. Set $w = 1$, then no missing occurs, and according to Theorems 1 and 2 the oracle-estimator (6) is also efficient (sharp-minimax) for the direct sampling from (V, Δ).

Corollary 1. *Under Assumptions 1-4, the oracle-estimator (6) is universal, meaning that it is efficient regardless of presence of missingness.*

The second corollary is about the effect of availability likelihood w on the MISE.

Corollary 2. *Under Assumptions 1-4, MISE of the efficient oracle-estimator is proportional to $w^{-2\alpha/2(\alpha+1)}$.*

Corollary 2 is an important conclusion for understanding the effect of missingness and a possible improvement in the MISE by imputation. Let us present Table 1 with values of the factor $w^{-2\alpha/(2\alpha+1)}$ which defines the effect of availability likelihood w and smoothness α of f^T on the MISE.

Table 1: Factor $w^{-2\alpha/(2\alpha+1)}$

α	$w = 0.3$	$w = 0.4$	$w = 0.5$	$w = 0.6$	$w = 0.7$	$w = 0.8$	$w = 0.9$	$w = 1$
0.5	1.83	1.58	1.41	1.29	1.20	1.12	1.05	1.00
1	2.23	1.84	1.59	1.41	1.27	1.16	1.07	1.00
2	2.62	2.08	1.74	1.50	1.33	1.20	1.09	1.00
3	2.81	2.19	1.81	1.55	1.36	1.21	1.09	1.00
4	2.92	2.26	1.85	1.57	1.37	1.22	1.10	1.00
10	3.15	2.39	1.94	1.63	1.40	1.24	1.11	1.00

Table 1 clearly indicates that the smoother the estimated density (the larger α) is, the more pronounced the effect of missingness becomes, while rougher densities "mute" the effect of missing data. This is an interesting and reverse effect of missing on nonparametric estimation, because for complete data smoother densities are estimated more accurately, see a nice discussion in Wasserman (2005). Further, the effect of missingness on the MISE accelerates with higher rates of missingness. We will return to these theoretical conclusions in the next section where we will see the same outcomes in simulations.

Now let us explain how the oracle-estimator can be mimicked by a data-driven estimator. The simpler problem is to estimate the unknown availability likelihood w, and this is done by the sample mean estimate $n^{-1}\sum_{l=1}^n A_l$. Next we are estimating S^C. It is possible to use the Kaplan-Meier estimator, but its analysis is relatively complicated as was explained in the Introduction. Instead, a simpler method of moments estimator is employed. Namely, recall notations introduced in the Introduction, and write $S^C(t) = \exp\{-H^C(t)\} = \exp\{-\int_0^t[f^C(v)/S^C(v)]dv\}$. For the likelihood we have

$$f^{AV,A\Delta,A}(v,0,1) = wf^C(v)S^T(v), \quad (8)$$

and also note that $S^V(v) = S^T(v)S^C(v)$ and

$$wS^V(v) = \mathbb{P}(AV \geq v, A = 1) =: S^{AV,A}(v, 1). \tag{9}$$

These relations yield

$$H^C(v) = \int_0^v \frac{f^{AV,A\Delta,A}(u, 0, 1)}{wS^V(u)} du = \mathbb{E}\Big\{\frac{A(1 - A\Delta)I(AV \leq v)}{S^{AV,A}(AV, 1)}\Big\}. \tag{10}$$

In its turn, the last formula yields the following method of moments estimate of S^C,

$$\hat{S}^C(t) = \exp\Big\{ - \sum_{l=1}^{n} \frac{A_l(1 - A_l\Delta_l)I(A_lV_l \leq t)}{\sum_{i=1}^n I(A_iV_i \geq A_lV_l)}\Big\}. \tag{11}$$

Please note that this estimator of the survival function can be used for both partial and complete missing models, and also it is well defined because all denominators in (11) are at least 1.

Now we plug $\hat{w}$ and $\hat{S}^C$ in (5) and get the following data-driven Fourier coefficient estimate

$$\hat{\theta}_j := \sum_{l=1}^{n} \frac{A_l\Delta_l\varphi_j(A_lV_l)I(A_lV_l \in [0, b])}{[\sum_{k=1}^n A_k]\hat{S}^C(A_lV_l)}. \tag{12}$$

Note that this estimate can be used for both partial and complete missing models.

Theorem 3. *The assertion of Theorem 2 holds if the Fourier coefficient estimate (12) is used in place of the oracle-estimate (5).*

We are left with the case of unknown parameters (α, Q), that is, with the case of adaptation to unknown smoothness of f^T. Fortunately, the blockwise-shrinkage adaptation, which mimics the oracle-estimator, is well known in the literature, see Efromovich (1999). Here we present its algorithm for small samples. This algorithm mimics the low-frequency part of the oracle-estimator (6).

Algorithm of density estimation: The estimator depends on three non-negative parameters c_{TH}, c_{J0}, and c_{J1}.
Step 1. Set $J_M := c_{J0} + c_{J1}q_n$. Calculate the sample mean Fourier estimates $\hat{\theta}_j$ for $j = 0, 1, \ldots, J_M$ using (12). Then calculate the empirical variances of $\hat{\theta}_j$ and denote them as $\hat{\sigma}_j^2$.
Step 2. Calculate the empirical cutoff $\hat{J}$

$$\hat{J} := \text{argmin}_{0 \leq J \leq J_M} \sum_{j=1}^{J} (2\hat{\sigma}_j^2 - \hat{\theta}_j^2). \tag{13}$$

Step 3. Calculate the series estimate

$$\hat{f}^T(t) := \sum_{j=0}^{\hat{J}} \hat{\theta}_j I(\hat{\theta}_j^2 > c_{TH}\hat{\sigma}_j^2)\varphi_j(t). \tag{14}$$

Step 4. If the estimate takes on negative values, use the L_2-projection on the class of nonnegative densities. Similarly, if it is known that the density is monotone, use the L_2-projection on monotonic densities.The projections are defined

in Efromovich (1999).

In what follows the above-defined estimator is referred to as the E-estimator. The included R code of the algorithm allows the user to change the default values $c_{TH} = 4$, $c_{J0} = 4$, $c_{J1} = 0.5$, recommended in Efromovich (1999) and used in the numerical studies of Section 3. The interested reader can also find discussion of E-estimator for the case of direct and censored observations in Efromovich (1999) and Efromovich (2001), respectively, and theoretical analysis of hard-threshoding in Johnstone (2023).

2.2. MAR

Suppose that Assumption 1 holds except that now $\mathbb{P}(A = 1|V, \Delta = \delta) = w(\delta) > 0$. The available sample of size n is from (AV, Δ, A), and accordingly we consider the MAR. Our aim is to show that the above-developed methodology of density estimation for MCAR can be extended to MAR.

Let us show how Fourier coefficients θ_j may be estimated under the MAR. Note that

$$f^{AV,\Delta,A}(v, \delta, 1) = \mathbb{P}(A = 1|V = v, \Delta = 1) f^T(v) S^C v = w(1) f^T(v) S^C(v). \tag{15}$$

This formula yields

$$\theta_j := \int_0^b f^T(t)\varphi_j(t)dt = \mathbb{E}\{\frac{A\Delta\varphi_j(AV)I(AV \in [0, b])}{w(1)S^C(AV)}\}. \tag{16}$$

In its turn, (16) yields the sample mean Fourier oracle-estimator

$$\bar{\theta}^*_j := n^{-1} \sum_{l=1}^n \frac{A_l\Delta_l\varphi_j(A_lV_l)I(A_lV_l \in [0, b])}{w(1)S^C(A_lV_l)}. \tag{17}$$

Note how this MAR oracle–estimator mimics the MCAR oracle-estimator (5), and we can also present a proposition that mimics Theorem 2.

Theorem 4. *Consider a sample of size n from (AV, Δ, A). Suppose that Assumptions 1–4 hold only now in Assumption 1 the MCAR is replaced by the MAR, that is, $\mathbb{P}(A = 1|V, \Delta = \delta) = w(\delta) > 0$. Introduce the density oracle–estimate $\bar{f}^T_*$ defined as in (6) only with $\check{\theta}^*_j$ being replaced by $\bar{\theta}^*_j$. Then*

$$\sup_{f^T \in \{\mathcal{F}_n(\alpha,Q,f_0) \cup \mathcal{S}_0(\alpha,Q)\}} \Big[n^{-1}b^{-1} \int_0^b \frac{f^T(t)}{S^C(t)}dt\Big]^{-\frac{2\alpha}{2\alpha+1}} \mathbb{E}_{f^T}\{\int_0^b (\bar{f}^T_*(t) - f^T(t))^2 dt\} \le [w(1)]^{-\frac{2\alpha}{2\alpha+1}} P(\alpha, Q, b)(1 + o_n(1)). \tag{18}$$

Now we are in a position to explain how to construct a data-driven Fourier estimator that mimics $\bar{\theta}^*_j$. The challenging part is to estimate S^C for the MAR. Our first step is to estimate the availability likelihood $w(\delta)$, and we use the sample mean estimates

$$\bar{w}(1) = \max(q_n^{-1}, \frac{\sum_{l=1}^n A_l\Delta_l}{1 + \sum_{l=1}^n \Delta_l}), \quad \bar{w}(0) = \max(q_n^{-1}, \frac{\sum_{l=1}^n A_l(1 - \Delta_l)}{1 + \sum_{l=1}^n (1 - \Delta_l)}). \tag{19}$$

Then the proposed estimate of S^C is

$$\bar{S}^C(t) := \exp\Big(- n^{-1} \sum_{l=1}^{n} \frac{A_l(1-\Delta_l) I(A_l V_l \le t)}{\bar{w}(0) \bar{S}^V(A_l V_l)} \Big), \tag{20}$$

where

$$\bar{S}^V(v) := n^{-1} \frac{\sum_{l=1}^{n} I(A_l \Delta_l V_l \ge v)}{\bar{w}(1)} + n^{-1} \frac{\sum_{l=1}^{n} I(A_l(1-\Delta_l) V_l \ge v)}{\bar{w}(0)}. \tag{21}$$

It will be shown in the Proofs that the assertion of Theorem 4 holds for the density estimator $\bar{f}^T$ defined as the oracle-estimator $\bar{f}^T_*$ only with $w(1)$ and S^C being replaced by $\bar{w}(1)$ and $\bar{S}^C$, respectively. In other words, the classical plug-in methodology yields the desired result. Note that the estimates (19)–(21) are used in denominators, and this explains why they are bounded below from zero.

The next section shows that the proposed methodology is feasible for small samples.

3. Numerical study and real data analysis

The aim of this section is two-fold. First, we would like to understand how the proposed E-estimator for missing censored data, motivated by the asymptotic theory, performs for real–life data. Second, because the E-estimator is also efficient for complete censored data (no missingness), it can be used for analysis of different imputation methods that fill-in the values of V that are missed. The following terminology is used in this section. The E-estimator based on missing censored data (no imputation) is referred to as the Benchmark, and the E-estimator based on imputed data is referred to according to the name of the imputation method. The term Benchmark is used to stress that in all numerical studies imputation-based E-estimators are compared with the Benchmark. Let us also note that while the Benchmark is asymptotically efficient, this does not yield its dominance for small samples, and as we will see shortly, this makes the numerical analysis both interesting and insightful.

We begin with describing the used imputation procedures, then consider a numerical study of MCAR and MAR when the underlying f^T and S^C are known, and finish with analysis of real data.

3.1. Procedures for imputation

The studied imputation procedures, discussed in Rubin (1987), Little and Rubin (2002), van Buuren (2018), He et al. (2021) and Little (2021), are:

(i) *Mean imputation*. The method replaces missing values by the mean of observed values. Mean imputation is the simplest, intuitively appealing, and popular procedure.

(ii) *Resampling imputation*. The method replaces missing lifetimes by randomly chosen lifetimes from available ones with replacement. This is a more sophisticated and often more accurate method than the mean imputation.

(iii) *Hot deck imputation*. This method is similar to the resampling imputation, only now it is done from subsamples with censored and uncensored lifetimes, that is, from subsamples with the same indicator of censoring Δ. According to the literature, hot deck imputation is typically more efficient than the two previous ones. On the other hand, it is not applicable for the complete missing scenario when indicators of censoring are missed.

(iv) *Multiple resampling imputation*. As the name tells us, the resampling imputation is repeated m times, and then the calculated density estimates are averaged. This is a popular technique among practitioners. In all simulations the recommended in van Buuren (2018) choice of $m = 5$ is used, and it will be discussed shortly.
(v) *Multiple hot deck imputation*. The above-defined hot-deck imputation is repeated m times, and then the calculated density estimates are averaged. Following van Buuren (2018), $m = 5$ is used.

3.2. Numerical study

3.2.1. MCAR simulations

We choose the underlying density of interest f^T and the distribution of the censoring lifetime, and then simulate censored data. Accordingly, we know the underlying density and can calculate the ISE of an estimate. To make the subsection shorter, here we are presenting results only for two distinct smooth and rougher densities f^T supported on $[0, 1]$, and more experiments can be found in the Supplementary Material. The former density f^T is the classical Uniform(0,1) and the latter is the Bimodal density of Efromovich (2018). In both cases the censoring lifetime is Uniform(0,2) which yields approximately 25% of censored observations. Results are presented in Tables 2 and 3 and we discuss them in turn. Additionally, the Supplementary Material contains results for exponential C resulting in 30% of observations being censored, and C from Uniform(0,1) resulting in 50% of observations being censored.

Table 2: MCAR simulation results for Uniform $T \sim$ U(0,1) and $C \sim$ U(0,2)

		$\frac{ISEB}{ISEM}$	$\frac{ISEB}{ISER}$	$\frac{ISEB}{ISEHD}$	$\frac{ISEB}{ISEMR}$	$\frac{ISEB}{ISEMHD}$
n=50	$\omega = 0.95$	0.58	0.60	0.51	0.70	0.76
	$\omega = 0.90$	0.41	0.66	0.52	0.92	0.69
	$\omega = 0.85$	0.22	0.50	0.53	0.80	0.73
	$\omega = 0.80$	0.19	0.59	0.61	0.78	0.74
	$\omega = 0.75$	0.12	0.69	0.40	0.81	0.56
	$\omega = 0.70$	0.08	0.44	0.32	0.60	0.48
n=100	$\omega = 0.95$	0.66	0.85	0.84	1.08	1.00
	$\omega = 0.90$	0.36	0.82	0.67	1.08	0.84
	$\omega = 0.85$	0.16	0.64	0.47	0.88	0.67
	$\omega = 0.80$	0.10	0.63	0.46	0.88	0.71
	$\omega = 0.75$	0.07	0.53	0.46	0.89	0.67
	$\omega = 0.70$	0.05	0.61	0.42	0.88	0.63
n=200	$\omega = 0.95$	0.33	0.70	0.63	0.99	0.84
	$\omega = 0.90$	0.14	0.68	0.51	0.89	0.78
	$\omega = 0.85$	0.07	0.62	0.51	0.90	0.74
	$\omega = 0.80$	0.04	0.58	0.42	0.80	0.70
	$\omega = 0.75$	0.03	0.49	0.41	0.78	0.62
	$\omega = 0.70$	0.02	0.44	0.31	0.64	0.49
n=1000	$\omega = 0.95$	0.09	0.93	0.84	1.00	0.87
	$\omega = 0.90$	0.02	0.67	0.50	0.86	0.66
	$\omega = 0.85$	0.01	0.42	0.59	0.70	0.66
	$\omega = 0.80$	0.01	0.46	0.49	0.74	0.63
	$\omega = 0.75$	0.00	0.38	0.31	0.59	0.52
	$\omega = 0.70$	0.00	0.43	0.44	0.67	0.56

We begin with the uniform f^T which is the example of a smooth density. First of all, note that for the smallest $n = 50$ and largest availability likelihood $w = 0.95$ even the mean imputation performs moderately well, but its

performance dwindles as the sample size n and the rate of censoring increases. The resampling dominates the hot deck, and overall the same pattern may be observed for the multiple imputation. In other words, knowing the indicator Δ does not make a difference. Further, recall that the hot deck cannot be used for completely missing sample from $(AV, A\Delta, A)$, and hence we may conclude that among the studied imputation methods the multiple resampling is the best. As for relative performance with respect to the Benchmark, it gradually decreases for larger sample sizes and smaller w. At the same time, multiple resampling performs exceptionally well for the sample size $n = 100$ and $w \geq 0.9$.

Table 3: MCAR simulation results for $T \sim$ Bimodal and $C \sim$ U(0,2)

		$\frac{AISEB}{AISEM}$	$\frac{AISEB}{AISER}$	$\frac{AISEB}{AISEHD}$	$\frac{AISEB}{AISEMR}$	$\frac{AISEB}{AISEMHD}$
n=50	$\omega = 0.95$	0.96	0.98	1.00	1.04	1.07
	$\omega = 0.90$	0.93	0.96	0.96	1.06	1.06
	$\omega = 0.85$	0.84	0.97	0.96	1.08	1.08
	$\omega = 0.80$	0.80	0.96	0.98	1.12	1.14
	$\omega = 0.75$	0.79	0.99	1.03	1.14	1.17
	$\omega = 0.70$	0.67	0.96	0.97	1.16	1.12
n=100	$\omega = 0.95$	0.92	0.99	1.02	1.06	1.06
	$\omega = 0.90$	0.85	1.02	1.02	1.10	1.10
	$\omega = 0.85$	0.75	1.01	1.05	1.14	1.14
	$\omega = 0.80$	0.69	1.01	0.98	1.17	1.15
	$\omega = 0.75$	0.61	1.04	1.06	1.19	1.24
	$\omega = 0.70$	0.49	1.06	1.06	1.25	1.26
n=200	$\omega = 0.95$	0.75	1.01	1.01	1.06	1.08
	$\omega = 0.90$	0.56	0.99	1.00	1.07	1.10
	$\omega = 0.85$	0.50	1.05	1.06	1.13	1.21
	$\omega = 0.80$	0.45	1.03	1.07	1.12	1.24
	$\omega = 0.75$	0.36	1.00	1.10	1.14	1.26
	$\omega = 0.70$	0.26	1.02	1.17	1.17	1.30
n=1000	$\omega = 0.95$	0.52	0.96	0.97	1.01	1.02
	$\omega = 0.90$	0.17	0.95	0.96	1.01	1.02
	$\omega = 0.85$	0.09	0.86	0.90	0.94	1.01
	$\omega = 0.80$	0.07	0.86	1.00	0.90	1.04
	$\omega = 0.75$	0.05	0.81	0.95	0.92	1.13
	$\omega = 0.70$	0.03	0.77	0.95	0.87	1.11

Now let us look at results in Table 3 for the rougher Bimodal density f^T. We observe a remarkably better relative performance for all imputation methods. Even the mean imputation performs very well for the smaller sample sizes and larger w. These outcomes confirm the theory, and the imputation may be a feasible tool for analysis of rough densities based on small samples. Overall the hot deck is a bit better than the resampling, and multiple imputation is definitely a plus. The relative performance of the multiple hot deck is very impressive for $n = 200$ and still very good even for $n = 1000$.

One of the typical questions about multiple imputation is how many repetitions m to perform. According to van Buuren (2018), $m = 5$ repetitions are deemed sufficient, and this number was used in the study. Some authors propose 20-100 repetitions or choosing m equal to the percentage of missing observations. Table 4 sheds some light on the issue. An interesting observation is that the improvement is more pronounced for the smoother Uniform density than for the rougher Bimodal where the effect of m is minor. Overall, the increase in m does not hurt, but the rule of 5-10 repetitions looks reasonable. At the same time, let us stress that no changes in the relative performance of the

Table 4: Effect of the number m of repetitions on $AISB/AISMR$

Model	n	m	$w = 0.95$	$w = 0.90$	$w = 0.85$	$w = 0.80$	$w = 0.75$	$w = 0.70$
$T \sim Unif(0,1)$	50	5	0.70	0.92	0.80	0.78	0.81	0.60
$C \sim Unif(0,2)$	50	10	0.80	0.97	0.85	0.81	0.79	0.69
	50	30	0.88	0.96	0.83	0.87	0.78	0.72
	100	5	1.08	1.08	0.88	0.88	0.89	0.88
	100	10	1.06	1.08	0.97	0.88	0.98	0.95
	100	30	1.08	1.18	1.01	0.96	0.99	0.93
	200	5	0.99	0.89	0.90	0.80	0.78	0.64
	200	10	0.99	0.99	0.95	0.88	0.85	0.71
	200	30	1.03	0.99	0.99	0.92	0.92	0.74
$T \sim Bimodal$	50	5	1.04	1.06	1.08	1.12	1.14	1.16
$C \sim Unif(0,2)$	50	10	1.06	1.09	1.11	1.16	1.17	1.19
	50	30	1.06	1.09	1.13	1.16	1.18	1.20
	100	5	1.06	1.10	1.14	1.17	1.19	1.25
	100	10	1.07	1.12	1.17	1.19	1.22	1.28
	100	30	1.07	1.12	1.17	1.20	1.25	1.31
	200	5	1.06	1.07	1.13	1.12	1.14	1.17
	200	10	1.06	1.09	1.16	1.16	1.19	1.21
	200	30	1.06	1.09	1.18	1.18	1.22	1.24

imputation methods have been noticed for the larger m.

More simulation results for different distributions of the pair (T, C) can be found in the Supplementary Material. Overall they are similar to the above-presented.

3.2.2. MAR simulations

As in the previous subsection, here we are presenting the results for the smoother Uniform(0,1) and the rougher Bimodal underlying densities. In both cases C is generated from Uniform(0,2), leading to approximately 25% of the observations being censored.

Table 5: MAR simulation results for density estimation, $T \sim$ U(0,1)

	n=50		n=100		n=200		n=1000	
$C \sim U(0,2)$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$
$w(\delta) = 0.7(1-\delta) + 0.9\delta$	0.53	0.73	0.68	0.88	0.50	0.70	0.45	0.66
$w(\delta) = 0.9(1-\delta) + 0.7\delta$	0.27	0.39	0.42	0.58	0.28	0.48	0.43	0.58
$w(\delta) = 0.5(1-\delta) + 0.7\delta$	0.30	0.39	0.43	0.62	0.33	0.53	0.41	0.52
$w(\delta) = 0.7(1-\delta) + 0.5\delta$	0.13	0.18	0.23	0.32	0.25	0.39	0.27	0.39

Table 6: MAR simulation results for density estimation, $T \sim$ Bimodal

	n=50		n=100		n=200		n=1000	
$C \sim U(0,2)$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$	$\frac{ISEF}{ISEHD}$	$\frac{ISEF}{ISEMHD}$
$w(\delta) = 0.7(1-\delta) + 0.9\delta$	0.95	1.06	1.01	1.13	1.02	1.12	0.97	1.01
$w(\delta) = 0.9(1-\delta) + 0.7\delta$	0.98	1.11	1.10	1.29	1.16	1.29	0.92	1.10
$w(\delta) = 0.5(1-\delta) + 0.7\delta$	0.97	1.11	1.09	1.29	1.14	1.32	0.91	1.12
$w(\delta) = 0.7(1-\delta) + 0.5\delta$	0.98	1.15	1.05	1.29	1.23	1.52	1.05	1.28

Four missing scenarios are considered: $w(\delta) = 0.7(1-\delta) + 0.9\delta$ leads to approximately 15% of observation being missed, $w(\delta) = 0.9(1-\delta) + 0.7\delta$ leads to 25% missing, $w(\delta) = 0.5(1-\delta) + 0.7\delta$ leads to 35% missing and

$w(\delta) = 0.7(1-\delta) + 0.5\delta$ leads to 45% missing. Each combination is repeated for samples size n equals to 50, 100, 200 and 1000. Since missing is MAR, we only compare hotdeck and multiple hotdeck with the benchmark as those are the only imputation methods that take Δ into account. The multiple hot deck imputation again uses $m = 5$. Results are presented in Tables 5 and 6, and they are similar to the MCAR results. Overall, the multiple imputation helps, and again the imputation methods shine for the rougher Bimodal density. More experiments with different underlying densities and censoring rates can be found in the Supplementary Material.

3.3. Real data examples

Analyzed datasets, their sources, the literature, and colored copies of all figures can be found in the Supplementary Material.

3.3.1. Liver cancer

The dataset, shown in Figure 1, is a sample from (AV, Δ, A). It contains survival status and time to death/censoring for 903 patients with liver cancer. Rate of censoring is 35%, the indicator of censoring is available for all observations, the overall rate of missing is 6.5%. Out of the missing observations, 55 are uncensored and 4 are censored. Recall that the missing is MCAR if $w(1) = w(0)$ and the missing is MAR otherwise. In this example, estimated availability likelihoods are $\bar{w}(1) = 0.905$ and $\bar{w}(0) = 0.988$. Accordingly, we can consider the sample as MAR and the Benchmark is the MAR E-estimate. At the same time, because the availability likelihoods are relatively close to each other, we can use the sample mean estimate $\bar{w} = 0.935$ of $\mathbb{P}(A = 1)$ and compare the MAR Benchmark with MCAR estimate.

Figure 1 allows us to compare the MAR Benchmark (the solid line) with the MCAR E-estimate (the dotted line) and with the naive E-estimate (the dashed line) based only on complete uncensored cases. The MCAR and MAR E-estimates are close due to the above-presented estimate of the availability likelihood. At the same time, the naive estimator yields the dramatically higher mortality during the first 20 days, and this stresses importance of using appropriate statistical tools for survival analysis.

The Benchmark and E-estimates, based on filled-in missing observations by the above-outlined imputation methods, can be found in the colored Figure 1.b presented in the Supplementary Material. The estimates are too close to each other to be recognizable in a black-white figure. The conclusion from that figure is as follows. The multiple hotdeck starts slightly above the Benchmark and becomes undistinguishable after day 20. Multiple resampling and resampling are almost identical until day 20, but start to deviate afterwards. The same outcome is obtained if the number of simulations $m = 5$ is increased to $m = 100$. Overall, in terms of L_2-closeness to the Benchmark, the multiple hotdeck estimator is the closest, followed by the hotdeck, multiple resampling and resampling (performances of the last two are almost identical), and the mean imputation. This conclusion is similar to simulation results where multiple hotdeck also performed the best. However, let us stress that for a real data we do not know the underlying density f^T, and hence can only report on relative performance with respect to the Benchmark.

3.3.2. Breast cancer

The dataset is an example of the complete missing model when sampling is from the triplet $(AV, A\Delta, A)$. Here V is the time in months from surgery to either cancer recurrence or end of study, $n = 212$. As we know from

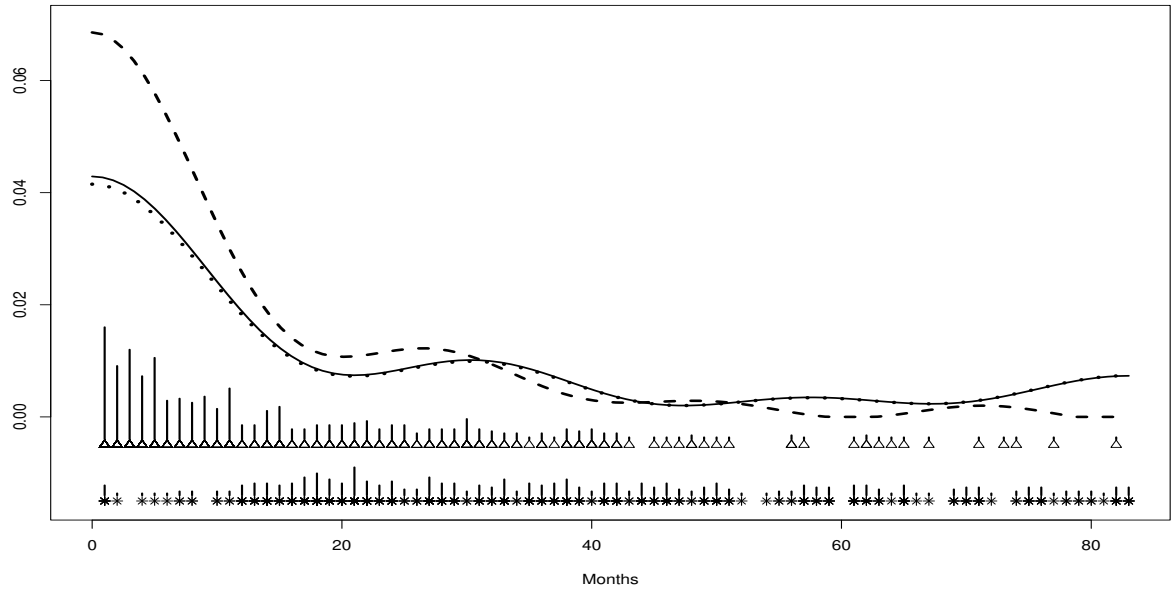

Figure 1: Density estimates for liver cancer data, $n = 903$. The triangles indicate uncensored available lifetimes, the stars indicate censored available lifetimes. The vertical lines are proportional to the number of lifetimes found in the dataset, e.g. at time of 1 month there are 55 uncensored observations and 5 censored observations. The solid line is the MAR E-estimate (the Benchmark), the dotted line is the MCAR E-estimate, the dashed line is the naive estimator which is the E-estimator based only on complete uncensored cases. The density f^T is estimated other the interval $[0, b]$ where $b = 83$ is the largest observed lifetime. The colored figure with added imputation–based estimates can be found in the Supplementary Material.

Section 2, for consistent density estimation of the time to cancer recurrence we need to assume MCAR, meaning that $\mathbb{P}(A = 1) = w > 0$. The missing rate is 34%, the censoring rate is 74.8%.

Figure 2 exhibits data and density estimates. The Benchmark is the MCAR E-estimate. As in the previous example, the naive estimator, which ignores both missingness and censoring, is dramatically more pessimistic about the breast cancer recurrence. The imputation-based estimates are relatively close to the Benchmark. Note that the hotdeck and multiple hotdeck imputations are not applicable in this example because the indicator of censoring is missing.

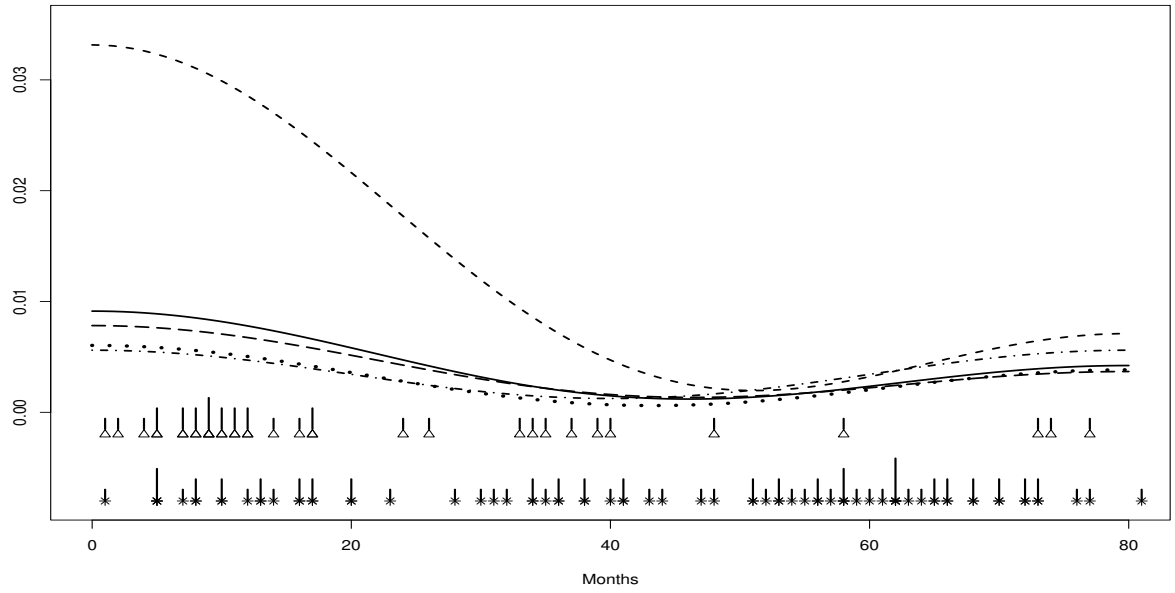

Figure 2: Density estimates for breast cancer data, $n = 212$. The triangles indicate uncensored lifetimes, the stars indicate censored lifetimes. The vertical lines are proportional to the number of lifetimes found in the dataset, e.g. at one month there are 1 uncensored observation and 1 censored observation. There are 35 available uncensored and 105 available censored observations. The solid line is the MCAR E-estimate (the Benchmark), the dashed line is the naive estimator based on uncensored cases, the dotted line is the mean imputation estimator, the long-dashed line is resampling estimator and the dot-dashed line is the multiple resampling estimator.

3.3.3. Mortality of infants

This is the example of a sample from (AV, Δ, A) with MAR missing, $n = 810$. The data is for infant (up to 90 days old) mortality in Ethio-Swedish Children's Hospital in Addis Ababa. For each infant patient, the number of days to death/discharge and the corresponding indicator were collected with some lifetimes being missed. The rate of missing is 66% and the censoring rate is 89.6%. The sample mean availability likelihoods are $\bar{w}(1) = 0.08$ and $\bar{w}(0) = 0.73$. Accordingly, the time to death is missed extremely rare, the time to discharge is missed often, and there is no missingness in the indicator of death/discharge.

The time is in days, in the dataset the time begins with day 0, and the density is estimated over the interval $[0, 10]$ because an overwhelming majority of observations falls within this interval. Accordingly, the lifetime V is discrete, takes on only $M = 11$ possible values, and we are estimating the probability mass function $\mathbb{P}(T = t)$, $t \in \{0, 1, \dots, 10\}$. To apply the developed E-estimation methodology to this case, the discrete cosine basis $\{\phi_0(x) = (1/\sqrt{M}),\ \phi_j(x) = \sqrt{2/M}\cos(\pi(2x+1)j/(2M)),\ j = 1, ..., M-1,\ x \in \{0, 1, \dots, 10\}\}$ is used.

Figure 3 exhibits the data and three probability mass function estimates whose values are connected to simplify visualization. The solid line is the MAR Benchmark, the dashed line is the E-estimate that uses the complete-case approach, and the dotted line is the naive E-estimate that uses only complete and uncensored observations. Comparing the solid line with the dashed and dotted ones, we may conclude that the Benchmark yields the least pessimistic assessment for the survival of infants. Note that because $\mathbb{P}(A = 1|\Delta = 0) < \mathbb{P}(A = 1|\Delta = 1)$, there is more missing

among the infants who were discharged. Accordingly, the complete-case approach leads to ignoring more of the cases where infants survived, that is, for whom $\Delta = 0$. Hence, we are left with more cases of infants who died, and this skews distribution to the left and yields the more pessimistic estimator.

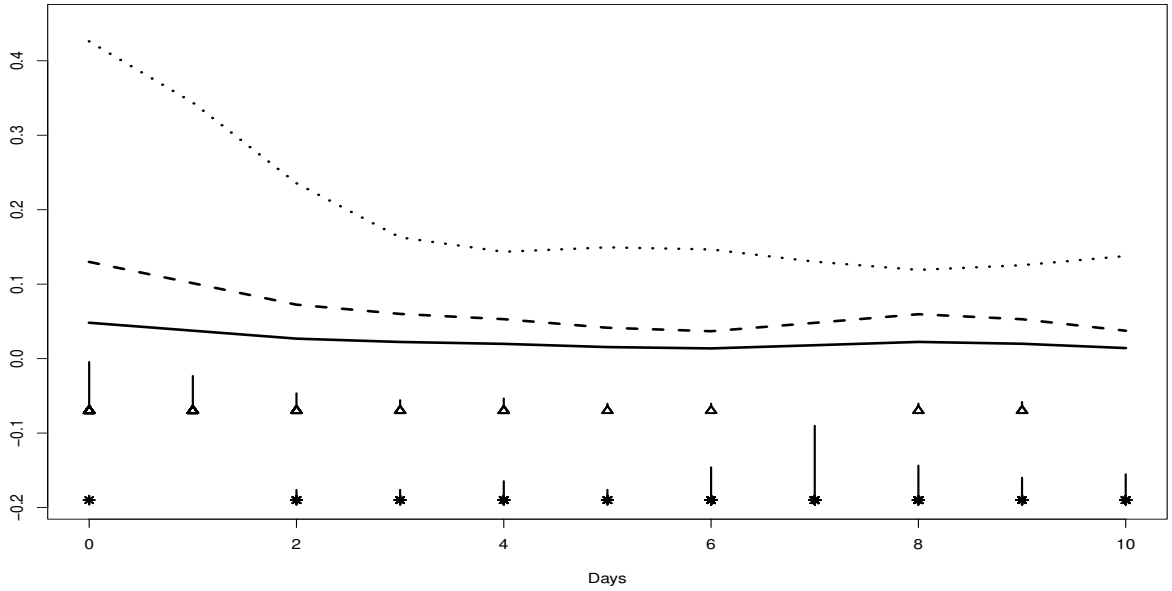

Figure 3: Young infants data. The triangles indicate uncensored available lifetimes, the stars indicate censored available lifetimes in the considered interval. The vertical lines are proportional to the frequency of survival times found in the dataset, e.g. at day 0 there are 26 uncensored observations and 2 censored observations. The curves show connected estimates of the underlying probability mass function $\mathbb{P}(T = t)$, $t \in \{0, 1, \ldots, 10\}$. The solid line is the MAR E-estimate, the dashed line is the naive estimator which ignored censoring and only involves uncensored cases, the dotted line is the naive estimate which ignored both missing and censoring. A color version of the plot is in the Supplementary Material.

3.3.4. Lung cancer data

A clinical study of patients with small cell lung cancer is described in Maksymiuk et al. (1994), the right censored lifetimes are published in Ying et al. (1995) and Bayes density estimates can be found in Poynor and Kottas (2019). Small cell lung cancer is the most aggressive type of lung cancer, and the study was devoted to comparison between two sequences of treatments. Namely, participants were randomly divided into Arm A and Arm B groups. Arm A received cisplatin followed by etoposide while Arm B received the treatment in reverse order. For each patient, either the death time or the time of right censoring due administrative reasons are recorded, the unit for time is a day. According to Ying et al. (1995), the lifetime of interest and the censoring lifetime are independent.

Arm A contains 62 observations with 15 being censored. Arm B contains 59 observations with 8 being censored. Both datasets are complete (no missingness), the data and the density E-estimates are shown in Figure 4. The densities are estimated over interval $[0, b]$, $b = 1400$ because no deaths occurred after that time. The densities are calculated on full data, and in what follows we will treat them as underlying densities. For Arm A the larger left mode is at about 440 days, and there is a 29.2% 2-year survival. For Arm B the larger left mode is at about 297 days, and there is a 16.5% 2-year survival. These characteristics agree with those previously published in the literature. A plausible

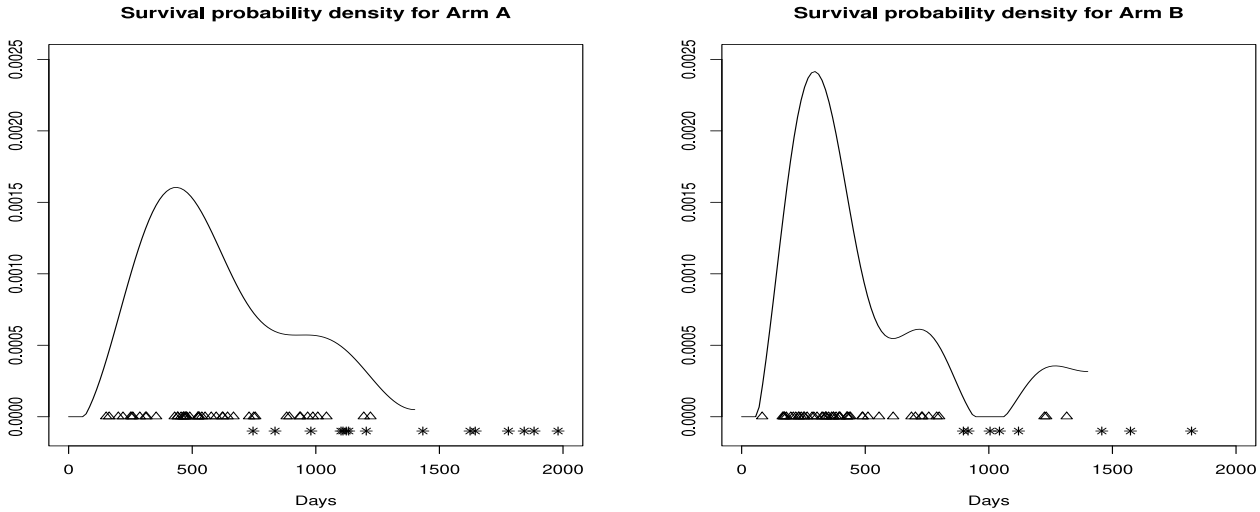

Figure 4: Density estimates for lung cancer data. The triangles indicate uncensored lifetimes, the stars indicate censored lifetimes.

explanation of the multimodality can found in the original publication Maksymiuk et al. (1994). According to the publication, the survival is primarily defined by the stage of cancer (limited or extensive) and by Eastern Cooperative Oncology Group performance status (0, 1, or 2). This is what the multimodality tells us about.

Now the main part of our study begins. We begin with MCAR and simulate samples from (AV, Δ, A), then use the imputation procedures, calculate ISEs of the corresponding E-estimators, and compare them with ISE of the Benchmark, that is the E-estimator (12) for missing data. Let us stress that we are considering partial missing because, as we will see shortly, the better imputation procedures are not applicable for complete missing. Table 7 summarizes results of the numerical study when for each availability likelihood w and the imputation method we repeat a simulated partial missing 1000 times, calculate and then average integrated squared errors of the density estimates with respect to the densities shown in Figure 4. Then we compare the calculated average ISE (AISE) of the benchmark E-estimate for missing data described in Section 2, denoted as AISEB, with corresponding AISEs of the E-estimates based on the imputations. For the above-introduced five imputation procedures we denote the corresponding AISEs as AISEM, AISER, AISEHD, AISEMR, and AISEMHD. Let us note one more time that the same E-estimator methodology, developed for right-censored data with no missingness, is used for all five imputation methods, and then their performances are compared with the Benchmark which is the E-estimator of Section 2 developed for missing censored data.

Table 7: Comparison of MCAR imputation methods for lung cancer data

		$\frac{AISEB}{AISEM}$	$\frac{AISEB}{AISER}$	$\frac{AISEB}{AISEHD}$	$\frac{AISEB}{AISEMR}$	$\frac{AISEB}{AISEMHD}$
Arm A	$\omega = 0.95$	0.17	0.34	0.40	0.64	0.71
	$\omega = 0.9$	0.11	0.36	0.37	0.59	0.60
	$\omega = 0.8$	0.12	0.53	0.50	0.74	0.85
Arm B	$\omega = 0.95$	0.17	0.47	0.52	0.75	0.75
	$\omega = 0.9$	0.12	0.45	0.66	0.85	0.92
	$\omega = 0.8$	0.08	0.57	0.61	0.97	0.98

Let us look at the results. We begin with a general comment and then rank the imputation procedures. What we

clearly observe is that, apart from the mean imputation method, the imputation performs better for Arm B than for Arm A, that is it performs better for the rougher density. We know from the theory why this is the case. Further, apart from the mean imputation, which is a clear outlier for the data at hand, the imputation improves its relative performance with respect to the Benchmark as the rate of missingness increases. The latter is an important outcome because literature, cited in the Introduction, often recommends against using imputation when more than 5-10% of observations are missed. Next, let us look at the individual imputation methods. The mean imputation does not perform well here, and this is due to the complicated shapes of the densities. Resampling and hot deck perform similarly for Arm A, but hot deck is better for Arm B. This is due to the above-explained rougher underlying density for the Arm B. Both multiple imputations dominate all other procedures, and for $w = 0.8$, when 20% of observations are missed, these imputation procedures are comparable with the Benchmark. Multiple hot deck outperforms multiple resampling, but there is a caveat. Indeed, multiple hot deck cannot be used for complete missing while multiple resampling can. Let us also stress that we will see shortly examples where imputation outperforms the Benchmark.

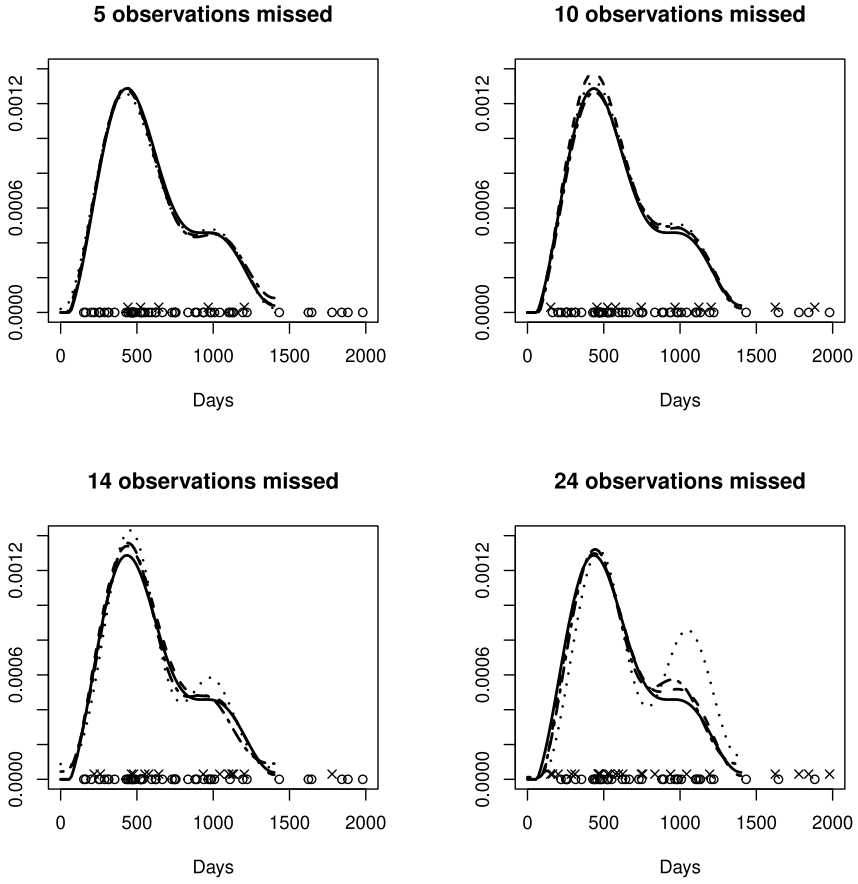

Figure 5: Density estimates for Arm A lung cancer data with various degrees of MCAR missingness. For each missing scenario, the circles and crosses show available and not available censored lifetimes. The solid line is the E-estimate shown in Figure 4. The dashed line is the Benchmark. The dotted line and the dot-dashed line are multiple resampling and multiple hot deck imputation E-estimates, respectively.

Figure 5 complements results of Table 7 by presenting particular simulations and estimates for Arm A data. The four diagrams show outcomes for different rates of missingness. Interestingly, we see how imputation highlights the

second mode, and the resampling does that more aggressively.

Table 8: Comparison of MAR imputation methods for lung cancer data

		$\frac{AISEB}{AISEM}$	$\frac{AISEB}{AISER}$	$\frac{AISEB}{AISEHD}$	$\frac{AISEB}{AISEMR}$	$\frac{AISEB}{AISEMHD}$
Arm A	$w(\delta) = 0.7(1-\delta) + 0.9\delta$	0.06	0.14	0.21	0.21	0.34
	$w(\delta) = 0.9(1-\delta) + 0.7\delta$	0.12	0.31	0.41	0.44	0.62
Arm B	$w(\delta) = 0.7(1-\delta) + 0.9\delta$	0.07	0.25	0.29	0.39	0.44
	$w(\delta) = 0.9(1-\delta) + 0.7\delta$	0.09	0.55	0.50	0.74	0.86

Next we consider simulations of MAR and analysis of the imputation method. Table 8 presents the corresponding results. Here $w(\delta) = 0.7(1-\delta) + 0.9\delta$ and $w(\delta) = 0.9(1-\delta) + 0.7\delta$ yield approximately 15% and %25 of observations being missed. The results show that MAR and MCAR yield a similar pattern in the relative performance of the imputation methods, and the multiple hotdeck is again the better imputation procedure.

4. Proofs

Recall that B's denote generic positive constants that may be different even in the same line, sequences $o_n(1)$ tend to zero as $n \to \infty$ uniformly over all parameters introduced during a proof, $q_n := \lceil \ln(n+3) \rceil$, and set $s := s_n := 1 + \lceil \ln(q_n) \rceil$.

Proof of Theorem 1. Let us begin with heuristic of the proof. Similarly to other known lower bounds for nonparametric curve estimates, the nonparametric problem is converted into a sequence in n of least favorable parametric problems, see an excellent overview in Tsybakov (2009). The traditional approach for Sobolev classes is to use the parametric classes created by the series perturbations $g_n(t) := \sum_{j=1}^{J_n} \theta_j \varphi_j(t)$, $t \in [0, b]$ for some appropriately chosen J_n proportional to $n^{1/(2\alpha+1)}$. Then, following Tsybakov (2009), namely using the Parseval identity, a Bayes approach and a Cramer–Rao type inequality, we can establish the optimal rate $n^{-2\alpha/(2\alpha+1)}$ of the MISE convergence. We also get a constant in the lower bound, but unfortunately it is smaller than the sharp one in Theorem 1. This implies that the classical parametric approach is not least favorable for the studied problem with missing censored lifetimes. To make the parametric approximation of an underlying density f^T more challenging, it is suggested to divide the interval of interest $[0, b]$ into s subintervals, and on each subinterval use the above-described method of finding a lower bound for the local MISE. This is a feasible approach but it requires solving two technical problems. First, the power Q of the Sobolev class (2) should be spread between the subintervals. Second, the underlying density f^T must be α-fold differentiable on $[0, b]$. To solve the first technical problem, let us note that the MISE may only increase if Q increases. Accordingly, for each subinterval the classical Fisher information is calculated, and then the total power Q is spread between the subintervals inversely proportional to the local Fisher informations. To solve the second technical problem, we "sew" local perturbations using special flattop nonnegative kernels that preserve the required α-fold differentiability over the interval $[0, b]$. Now we are ready to proceed to the proof.

To implement the above-introduced idea of converting the problem of estimating density f^T over interval $[0, b]$ into its estimation over a sequence in n of subintervals of $[0, b]$, we introduce the following sequence of function classes

$$\mathcal{H}_s = \{f: \ \ f(t) = f_0(t) + [\sum_{k=0}^{s-2} g_k(t) - \sum_{k=0}^{s-2} \int_0^b g_k(u)du] I(0 \le t \le b), \ \ g_k(t) \in \mathcal{H}_{sk}, f \ge 0\}.$$

Here the function classes $\mathcal{H}_{sk}$ are defined as follows. Denote by $\phi(t) = \phi(n,t)$ a sequence of flattop nonnegative kernels defined on a real line such that for a given n: the kernel is zero beyond $(0,1)$, it is α-fold continuously differentiable on $(-\infty,\infty)$, $0 \le \phi(t) \le 1$, $\phi(t) = 1$ for $2(\ln(n))^{-2} \le t \le 1 - 2(\ln(n))^{-2}$, and $|\phi^{(\alpha)}(t)| \le B(\ln(n))^{2\alpha}$. For instance, such a kernel may be constructed using a mollifier, see Chapter 7 in Efromovich (1999). Then set $\phi_{sk}(t) = \phi(sb^{-1}t - k)$. For a kth subinterval, $0 \le k \le s-2$, define: $\varphi_{skj}(t) = \sqrt{2s/b}\cos(\pi j[sb^{-1}(t-b) - k])$, $g_{[k]}(t) = \sum_{j=\lceil J(k)/\ln(n)\rceil}^{J(k)} \nu_{skj}\varphi_{skj}(t)$, $g_{(k)}(t) = g_{[k]}(t)\phi_{sk}(t)$, $J(k) = \lceil [n(2\alpha+1)(\alpha+1)\alpha^{-1}s^{-2\alpha}Q_{sk}\pi^{-2\alpha}b]^{1/(2\alpha+1)}\rceil + 1$, $Q_{sk} = (Q - 1/s)(\overline{I_s^{-1}}I_{sk})^{-1}$, $I_{sk} = wS^C(bk/s)/f_0(bk/s)$, $\overline{I_s^{-1}} = \sum_{k=0}^{s-2}(1/I_{sk})$. We set

$$\mathcal{H}_{sk} := \{f:\ f(t) = g_{(k)}(t),\ \sum_{j=\lceil J(k)/\ln(n)\rceil}^{J(k)} (\pi j)^{2\alpha}\nu_{skj}^2 \le s^{-2\alpha}Q_{sk},\ |g_{[k]}(t)|^2 \le s^3\ln(n)J(k)n^{-1}\}.$$

We do not use a permutation for the last right subinterval because this allows us to bound from below $S^T(b - 1/s)$ by $\int_{b-1/s}^b f_0(t)dt \ge \min_{t\in[0,b]}(f_0(t))/s$. The latter will be used shortly in evaluation of Fisher information.

Note that we have defined a new function class where additive perturbations for each subinterval are mutually independent, and then the perturbations are sewed together using the flattop kernel. The latter preserves α-fold differentiability of the perturbations over the interval of interest $[0,b]$ as well as smoothness of $f_0(t)$ for $t \ge b$.

Now we are in a position to verify that for sufficiently large n the new set of densities is a subset of the studied class $\mathcal{F}_n$. Definition of the flattop kernel implies that $f - f_0$ is α-fold continuously differentiable on the interval of interest $[0,b]$. Let us show that for $f \in \mathcal{H}_s$ the difference $f - f_0$ belongs to $\mathcal{S}_1(\alpha, Q)$. By Leibniz rule $\int_0^b [(g_{[k]}(t)\phi_{sk}(t))^{(\alpha)}]^2dt = \int_0^b [\sum_{l=0}^{\alpha} \mathbf{C}_l^{\alpha} g_{[k]}^{(\alpha-l)}(t)\phi_{sk}^{(l)}(t)]^2dx$ where $\mathbf{C}_l^{\alpha} = \alpha!/((\alpha-l)!l!)$. Using $\max_{0\le l\le\alpha}\int_0^1(\phi_{sk}^{(l)}(t))^2dt < B(s(\ln(n))^2)^{2\alpha}$ and Cauchy-Schwarz inequality we can write for $0 < l \le \alpha$,

$$|g_{[k]}^{(\alpha-l)}(t)|^2 = |\sum_{j=\lceil J(k)/\ln(n)\rceil}^{J(k)} \nu_{skj}\varphi_{skj}^{(\alpha-l)}(x)|^2 \le Bs^{2(\alpha-l)+1}(\sum_{j=\lceil J(k)/\ln(n)\rceil}^{J(k)} j^{2\alpha}\nu_{skj}^2)(\sum_{j=\lceil J(k)/\ln(n)\rceil}^{J(k)} j^{-2l}) = o_n(1)(J(k))^{-1/2}.$$

Further, we have

$$\int_0^b [g_{[k]}^{(\alpha)}(t)\phi_{sk}(t)]^2dt \le \int_{bk/s}^{b(k+1)/s} [g_{[k]}^{(\alpha)}(t)]^2dx \le Q_{sk}\,,$$

and recall that $\sum_{k=0}^{s-2} Q_{sk} = Q - 1/s$. Combining the results we verify that $f - f_0 \in \mathcal{S}_1(\alpha, Q)$ for $f \in \mathcal{H}_s$ and large n. This is what was wished to show, and from now on we can use the smaller function class in verifying the lower bound.

Set $\hat{f} = f_0 + \tilde{f}$ and $\rho_s := \sum_{k=0}^{s-2}\int_0^b g_{(k)}(u)du$. For any $f \in \mathcal{H}_s$, any positive γ, and a kth subinterval we can write,

$$\int_{bk/s}^{b(k+1)/s} (\hat{f}(t) - f(t))^2dt = \int_{bk/s}^{b(k+1)/s} (\tilde{f}(t) - g_{(k)}(t) + \rho_s)^2dt$$

$$\ge (1-\gamma)\int_{bk/s}^{b(k+1)/s} (\tilde{f}(t) - g_{[k]}(t))^2dt - \gamma^{-1}\int_{bk/s}^{b(k+1)/s} [g_{[k]}(t)(1 - \phi_{sk}(t)) + \rho_s]^2dt$$

$$\ge (1-\gamma)\int_{bk/s}^{b(k+1)/s} (\tilde{f}(t) - g_{[k]}(t))^2dt + o_n(1)\gamma^{-1}(\ln(n))^{-1/2}n^{-2\alpha/(2\alpha+1)}\ .$$

Set $\tilde{\nu}_{skj} := \int_{bk/s}^{b(k+1)/s} \tilde{f}(t)\varphi_{skj}(t)dt$ and choose $\gamma = s^{-1}$. Then we can continue the previous relation and write,

$$\sup_{f\in\mathcal{F}_n(\alpha,Q,f_0)} \mathbb{E}\left\{\int_0^b (\hat{f}(t)-f(t))^2dt\right\} \geq \sup_{f\in\mathcal{H}_s} E\left\{\int_0^b (\hat{f}(t)-f(t))^2dt\right\} = \sup_{f\in\mathcal{H}_s}\sum_{k=0}^{s-2} E\left\{\int_{bk/s}^{b(k+1)/s} (\hat{f}(t)-f(t))^2dt\right\}$$

$$\geq (1-s^{-1})\sum_{k=0}^{s-2}\sup_{f\in\mathcal{H}_{sk}}\sum_{j=\lceil J(k)/\ln(n)\rceil}^{J(k)} E\left\{(\tilde{\nu}_{skj}-\nu_{skj})^2\right\} + o_n(1)n^{-2\alpha/(2\alpha+1)} =: (1-s^{-1})\sum_{k=0}^{s-2} r_k + o_n(1)n^{-2\alpha/(2\alpha+1)}.$$

We converted the verification of the lower bound for the studied minimax MISE into establishing lower bounds for parametric mean squared errors. The latter is a familiar problem in mathematical statistics, and to make the proof shorter we are using the proof of Theorem 1 in Efromovich (1989). To employ it we first need to verify the validity of two relations, and then calculate parametric Fisher informations. Consider independent normal random variables ζ_{skj} with zero mean and variance $(1-\mu_n)\eta_{skj}^2$ where μ_n tends to zero as slowly as desired and $\eta_{ski}^2 := n^{-1}\max(1/B_*, \min(B_*, [(J(k)/j)^\alpha - 1]))$ with $B_* \geq 1$ being a constant that can be as large as desired.

Now introduce a stochastic process $f^*(t)$, $t\in[0,b]$ defined as the studied $f(t)$, $t\in[0,b]$ but with ζ_{skj} being used in place of ν_{skj}. Then Theorem 6.2.2 in Kahane (1985) yields that $\mathbb{P}(f^*\in\mathcal{H}_s) = 1+o_n(1)$. This is the first verified relation. The second relation also follows from that theorem, and for a similarly defined stochastic process $g^*_{[k]}$ we get $\mathbb{P}(\sup_{t\in[0,b]}|g^*_{[k]}(t)|^2 \leq s^3\ln(n)J(k)n^{-1}) = 1+o_n(1)$.

Next we calculate classical parametric Fisher informations. Recall that the underlying likelihood function is

$$f^{AV,\Delta,A}(v,\delta,a) = \Big[w[f^T(v)S^C(v)]^\delta[f^C(v)S^T(v)]^{1-\delta}\Big]^a\Big[(1-w)\mathbb{P}(\Delta=\delta)\Big]^{1-a},$$

and that functions $f_0(t)$, $f^C(t)$ and parameter w are known to the oracle. Set

$$\mathcal{I}_{skj} := \mathbb{E}_{f_0,f^C,w}\Big\{\Big[\frac{\partial}{\partial\nu_{skj}}[A\Delta\ln(wf^T(V)S^C(V)) + A(1-\Delta)\ln(wf^C(V)S^T(V))$$

$$+(1-A)\Delta\ln((1-w)\mathbb{P}(\Delta=1)) + (1-A)(1-\Delta)\ln((1-w)\mathbb{P}(\Delta=0))]\Big]^2\Big\}. \tag{22}$$

Note that the four addends in the expectation are mutually exclusive and this allows us to simplify the formula. Also, let us use notation $\theta := \nu_{skj}$, and we may write f^T_θ, S^T_θ and $\mathbb{P}_\theta$ to stress that these characteristics depend on the underlying parameter θ. Using these remarks we continue formula (22) for the Fisher information,

$$\mathcal{I}_{skj} = \mathbb{E}_{f_0,f^C,w}\{A\Delta[\partial\ln(wf^T_\theta(V)S^C(V))/\partial\theta]^2\} + \mathbb{E}_{f_0,f^C,w}\{A(1-\Delta)[\partial\ln(wf^C(V)S^T_\theta(V))/\partial\theta]^2\}$$

$$+\mathbb{E}_{f_0,f^C,w}\{(1-A)\Delta[\partial\ln((1-w)\mathbb{P}_\theta(\Delta=1))/\partial\theta]^2\} + \mathbb{E}_{f_0,f^C,w}\{(1-A)(1-\Delta)[\partial\ln((1-w)\mathbb{P}_\theta(\Delta=0))/\partial\theta]^2\}$$

$$=: \mathcal{I}_1+\mathcal{I}_2+\mathcal{I}_3+\mathcal{I}_4. \tag{23}$$

As we will see shortly, only the first component $\mathcal{I}_1$ of the Fisher information is not vanishing as $n\to\infty$. The latter explains why the verified lower bound is valid both for all observations and for a subset of observations with

$A\Delta = 1$, that is, for a subset with not missed and uncensored observations. Let us present a heuristic explanation of this phenomenon interesting in its own right. If $A\Delta = 1$, then the likelihood is proportional to an underlying density f_θ^T which depends on the parameter of interest θ. This is a classical statistical model and we will calculate the corresponding Fisher information shortly. If $A(1-\Delta) = 1$, then the likelihood is proportional to the survival function $S_\theta^T(t) = \int_t^\infty f_\theta^T(u)du$. This integration of the very special density, written as a high-frequency Fourier series, yields the vanishing Fisher information. To see that, recall that $\theta = \nu_{skj}$, j is the frequency, $\int_0^z \cos(ujt)dt = (uj)^{-1}\sin(\pi jz) = o_n(1)$ for considered $j > Bn^{1/(2\alpha+1)}/\ln(n)$ and $u > 0$. Overall, in mathematical statistics, it is a familiar fact that it is more difficult to estimate a parameter when the likelihood depends on an integral of the underlying density. If $(1-A)\Delta = 1$, then we get a similar conclusion because the likelihood is proportional to $\mathbb{P}_\theta(\Delta = 1) = \int_0^\infty f_\theta^T(t)S^C(t)dt$, that is, we again integrate the density. Finally, if $(1-A)(1-\Delta) = 1$, then the likelihood is proportional to $\mathbb{P}_\theta(\Delta = 0) = \int_0^\infty f^C(t)S_\theta^T(t)dt$, and again we are dealing with the integral $S_\theta^T(t) = \int_t^\infty f_\theta^T(u)du$.

Next we get several technical formulas for the used cosine elements of the basis and the flattop kernels that are instrumental in evaluating the four components of the Fisher information. Note that $\phi_{sk}(x)$ is zero beyond interval $[bk/s, b(k+1)/s]$, $\varphi_{skj}(t)$ are elements of the cosine basis on $[bk/s, b(k+1)/s]$, and write for $0 \le k \le s-2$,

$$\int_{bk/s}^{b(k+1)/s}[\varphi_{skj}(t)\phi_{sk}(t)]^2dt = 1 + \int_{bk/s}^{b(k+1)/s}\varphi_{skj}^2(t)(\phi_{sk}^2(t)-1)dt.$$

Then, recalling that $\phi_{sk}(t)$ is the special flattop kernel, we get

$$|\int_{bk/s}^{b(k+1)/s}\varphi_{skj}^2(t)(\phi_{sk}^2(t)-1)dt| = o_n(1)(\ln(n))^{-1} \ .$$

Using the same technique we get

$$|\int_{bk/s}^{b(k+1)/s}\varphi_{skj}(t)\phi_{sk}(t)dt| = |\int_{bk/s}^{b(k+1)/s}\varphi_{skj}(t)[\phi_{sk}(t)-1]dt| = o_n(1)(\ln(n))^{-1}.$$

Further, for $bk/s \le z \le b(k+1)/s$ and $j \ge J(k)/\ln(n)$, we get using $\int_0^z \cos(ujt)dt = (uj)^{-1}\sin(ujz)$, $u > 0$ that

$$|\int_{bk/s}^{z}\varphi_{skj}(t)\phi_{sk}(t)dt| = |\int_{bk/s}^{z}\varphi_{skj}(t)dt - \int_{bk/s}^{z}\varphi_{skj}(t)[1-\phi_{sk}(t)]dt| = o_n(1)(\ln(n))^{-1},$$

and similarly for $z > b(k+1)/s$

$$\Big(\int_0^z[\varphi_{skj}(t)\phi_{sk}(t) - \int_{bk/s}^{b(k+1)/s}\varphi_{skj}(u)\phi_{sk}(u)du]I(0 \le t \le b)dt\Big)^2 = o_n(1)(\ln(n))^{-2}.$$

Using these relations we can make the following calculations. For component $\mathcal{I}_1$ of the Fisher information, that considers the case $A\Delta = 1$ of not missed and uncensored observations, we can write using the above-made technical relations,

$$\mathcal{I}_1 = \mathbb{E}_{f_0,f^C,w}\{A\Delta[\partial\ln(wf_\theta^T(V)S^C(V))/\partial\theta]^2\} = \int_{bk/s}^{b(k+1)/s}[wf_0(t)S^C(t)/f_0^2(t)][\varphi_{skj}(t)\phi_{sk}(t)$$

$$-\int_{bk/s}^{b(k+1)/s}\varphi_{skj}(u)\phi_{sk}(u)du]^2dt(1+o_n(1))=[wS^C(bk/s)/f_0(bk/s)](1+o_n(1))=I_{sk}(1+o_n(1)).$$

For component $\mathcal{I}_2$, that considers the case $A(1-\Delta)=1$ of not missed censored observations, we can write using the above-made technical relations,

$$\mathcal{I}_2=\mathbb{E}_{f_0,f^C,w}\{A(1-\Delta)[\partial\ln(wf^C(V)S_\theta^T(V))/\partial\theta]^2\}$$

$$=\mathbb{E}_{f_0,f^C,w}\Big\{(S_\theta^T(V))^{-1}\Big(\int_0^V[\varphi_{skj}(t)\phi_{ks}(t)-\int_{bk/s}^{b(k+1)/s}\varphi_{skj}(u)\phi_{sk}(u)du]1(0\le t\le b)dt\Big)^2\Big\}(1+o_n(1))=o_n(1).$$

In the last equality we used the above-mentioned inequality $F_\theta^T(b(s-1)/s)>Bs^{-1}$ which holds due to the assumed $\min_{t\in[0,b]}f_0(t)>0$.

Now we are considering $\mathcal{I}_3$. We have formula $\mathbb{P}_\theta(\Delta=1)=\int_0^\infty f_\theta^T(t)S^C(t)dt$, which in its turn, with the help of the above-presented technical relations for the cosine basis and flattop kernels, yields

$$\frac{\partial\ln(\mathbb{P}_\theta(\Delta=1))}{\partial\theta}=\frac{\int_{bk/s}^{b(k+1)/s}[\varphi_{skj}(t)\phi_{sk}(t)-\int_{bk/s}^{b(k+1)/s}\varphi_{skj}(u)\phi_{sk}(u)du]S^C(t)dt}{\mathbb{P}_\theta(\Delta=1)}=o_n(1)\ln^{-1}(n).$$

We conclude that $\mathcal{I}_3=o_n(1)$. Absolutely similarly we get $\mathcal{I}_4=o_n(1)$. Combining the obtained results in (22) we get

$$I_{skj}=[wS^C(bk/s)/f_0(bk/s)](1+o_n(1))=I_{sk}(1+o_n(1)). \tag{24}$$

Now we can straightforwardly follow the proof of Theorem 1 in Efromovich (1989). This yields for $k\in\{0,1,\dots,s-2\}$ that

$$\inf r_k\ge(s^{-2\alpha}Q_{sk})^{1/(2\alpha+1)}(nI_{sk})^{-2\alpha/(2\alpha+1)}P(\alpha,1,b)(1+o_n(1)),$$

where the infimum is over all possible nonparametric oracle-estimators of f considered in the theorem. Using this lower bound, definitions of Q_{sk} and I_{sk}, and the assumed $S^C(b)>0$ we continue,

$$\inf\sum_{k=0}^{s-2}r_k\ge P(\alpha,Q,b)\Big[s^{-1}\sum_{k=0}^{s-2}f_0(bk/s)/[wS^C(bk/s)]\Big]^{2\alpha/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}(1+o_n(1))$$

$$=P(\alpha,Q,b)\Big[\sum_{k=0}^{s-2}b^{-1}\int_{bk/s}^{b(k+1)/s}(f_0(t)/[wS^C(t)])dt\Big]^{2\alpha/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}(1+o_n(1))$$

$$=P(\alpha,Q,b)\Big(n/[b^{-1}\int_0^b(f_0(t)/[wS^C(t)])dt]\Big)^{-2\alpha/(2\alpha+1)}(1+o_n(1)).$$

Theorem 1 is proved.

Proof of Theorem 2. First of all, a direct calculation verifies that the Fourier estimator $\check\theta_j^*$ is unbiased, namely

$$\mathbb{E}\{\check\theta_j^*\}=\mathbb{E}\Big\{\frac{A\Delta\varphi_j(V)I(V\in[0,b])}{wS^C(V)}\Big\}=\int_0^b\frac{wf^T(t)S^C(t)\varphi_j(t)}{wS^C(t)}dt=\int_0^bf^T(t)\varphi_j(t)dt=\theta_j.$$

Second, using the trigonometric formula $\varphi_j^2(t) = b^{-1} + (2b)^{-1/2}\varphi_{2j}(t)$ and the fact that Fourier coefficients of a differentiable function vanish as the frequency j increases, see Chapter 2 in Efromovich (1999), we conclude that

$$\mathbb{E}\{(\check{\theta}_j^* - \theta_j)^2\} = n^{-1}[\mathbb{E}\Big\{\frac{A\Delta\varphi_j^2(V)I(V \in [0,b])}{[wS^C(V)]^2}\Big\} - \theta_j^2] = n^{-1}d(1 + o_j(1)).$$

Now we can explore MISE of the density estimator (6) for the considered class $\mathcal{F}_n(\alpha, Q, f_0) \cup \mathcal{S}_0(\alpha, Q)$ of densities. Recall that B denotes generic constants. Write using the Parseval identity,

$$\mathbb{E}\{\int_0^b (\check{f}_*^T(t) - f^T(t))^2 dt\} = \sum_{j=0}^{q_n} \mathbb{E}\{(\check{\theta}_j^* I(\check{\theta}_j^* \geq 2q_n\sigma_j^2) - \theta_j)^2\} + \Big[\sum_{j=q_n+1}^{J_n} [(1 - (j/J_n)^\alpha)\check{\theta}_j^* - \theta_j]^2 + \sum_{j>J_n} \theta_j^2\Big]. \tag{25}$$

We are considering the two terms on the right side of (25) in turn. For the first one we write using the Cauchy inequality,

$$\sum_{j=0}^{q_n} \mathbb{E}\{(\check{\theta}_j^* I(\check{\theta}_j^* \geq 2q_n\sigma_j^2) - \theta_j)^2\} = \sum_{j=0}^{q_n} \mathbb{E}\{[(\check{\theta}_j^* - \theta_j) - \check{\theta}_j I((\check{\theta}_j^*)^2 < 2q_n\sigma_j^2)]^2\}$$

$$\leq 2\sum_{j=0}^{q_n}[\mathbb{E}\{(\check{\theta}_j^* - \theta_j)^2\} + 4q_n^2\sigma_j^4] \leq Bq_n^2 n^{-1}.$$

For the second term on the right side of (25), using notation $d := b^{-1}\int_0^b \frac{f^T(t)}{wS^C(t)}dt$, we can write,

$$\sum_{j=q_n+1}^{J_n} [(1 - (j/J_n)^\alpha)\check{\theta}_j^* - \theta_j]^2 + \sum_{j>J_n} \theta_j^2$$

$$\leq \sum_{j=q_n+1}^{J_n} [(1 - (j/J_n)^\alpha)^2\mathbb{E}\{(\check{\theta}_j^* - \theta_j)^2\} + (j/J_n)^{2\alpha}\theta_j^2] + [\frac{b}{\pi J_n}]^{2\alpha}\sum_{j>J_n}[\pi j/b]^{2\alpha}\theta_j^2$$

$$= dn^{-1}\sum_{j=q_n+1}^{J_n} (1 - (j/J_n)^\alpha)^2(1 + o_j(1)) + [\frac{b}{\pi J_n}]^{2\alpha}\sum_{j>q_n}[\pi j/b]^{2\alpha}\theta_j^2.$$

Using $f^T \in \{\mathcal{F}_n(\alpha, Q, f_0) \cup \mathcal{S}_0(\alpha, Q)\}$, together with elementary calculations and (25), we conclude that

$$\mathbb{E}\{\int_0^b (\check{f}_*^T(t) - f^T(t))^2 dt\} \leq dn^{-1}\sum_{j=0}^{J_n}[1 - (j/J_n)^\alpha] + [\frac{b}{\pi J_n}]^{2\alpha}Q + o_n(1)n^{-2\alpha/(2\alpha+1)}$$

$$= [n/d]^{-2\alpha/(2\alpha+1)}P(\alpha, Q, b)(1 + o_n(1)).$$

Note that $o_n(1)$ is a generic vanishing sequence that does not depend on (α, Q, f_0), and this proves the verified upper bound. Theorem 2 is proved.

Proof of Theorem 3. We need to verify that if the estimates $\hat{S}^C$ and $\bar{w}$ are used in place of unknown S^C and w, then

the upper bound of Theorem 2 is still valid. Consider the data-driven Fourier estimate

$$\hat{\theta}_j := n^{-1} \sum_{l=1}^{n} \frac{A_l \Delta_l \varphi_j(A_l V_l) I(A_l V_l \in [0,b])}{\bar{w} \hat{S}^C(A_l V_l)}. \tag{26}$$

Here $\bar{w} = n^{-1} \sum_{l=1}^{n} A_l$, $\hat{S}^C(v) := e^{-\hat{H}^C(v)}$, $\hat{H}^C(v) := n^{-1} \sum_{l=1}^{n} \frac{A_l(1-\Delta_l) I(A_l V_l \le v)}{\hat{S}^{AV,A}(A_l V_l, 1)}$, and $\hat{S}^{AV,A}(v,1) := n^{-1} \sum_{l=1}^{n} I(A_l V_l \ge v)$. According to the proof of Theorem 2, to verify validity of Theorem 3 it is sufficient to show that

$$|\mathbb{E}\{\hat{\theta}_j - \theta_j\}| \le B n^{-1}, \ \mathbb{E}\{(\hat{\theta}_j - \theta_j)^2\} \le d n^{-1}(1 + o_n(1)). \tag{27}$$

Let us verify (27). Note that $\bar{w}$ is the classical sample mean estimate of the probability of success for Bernoulli random variable A whose statistical properties are well known, see Serfling (1980). Analysis of $\hat{S}^C$ is more involved, and we are presenting properties of the involved statistics one by one. The statistic $\hat{S}^{AV,A}(v,1)$ is the sample mean estimate of the probability $S^{AV,A}(v,1) = \mathbb{P}(V \ge v, A = 1)$ with the following properties highlighted in Serfling (1980). We have $\mathbb{E}\{[\hat{S}^{AV,A}(v,1) - S^{AV,A}(v,1)]^2\} = n^{-1} S^{AV,A}(v,1)[1 - S^{AV,A}(v,1)]$, $\mathbb{E}\{[\hat{S}^{AV,A}(v,1) - S^{AV,A}(v,1)]^{2k}\} \le B_k n^{-k} S^{AV,A}(v,1)[1 - S^{AV,A}(v,1)]$, $\mathbb{P}(\sup_v |\hat{S}^{AV,A}(v,1) - S^{AV,A}(v,1)| > \epsilon) \le B e^{-2n\epsilon^2}$ for $\epsilon > 0$, and recall that B and B_k are generic finite constants. Next, using the Taylor formula we get

$$\hat{S}^C(x) - S^C(x) = e^{-\hat{H}^C(x)} - e^{-H^C(x)} = [H^C(x) - \hat{H}^C(x)] S^C(x) + (1/2)[H^C(x) - \hat{H}^C(x)]^2 S^C(x) + \mu(x), \tag{28}$$

where $|\mu(x)| \le |H^C(x) - \hat{H}^C(x)|^3 S^C(x)$. Analysis of these terms is based on the above-presented properties of $\hat{S}^{AV,A}(v,1)$ and straightforward calculations. We get that $|\mathbb{E}\{[\hat{S}^C(V_1) - S^C(V_1)|V_1 = v]\}| \le B n^{-1}$, $\mathbb{E}\{[\hat{S}^C(V_1) - S^C(V_1)]^{2k} | V_1 = v\} \le B_k n^{-k}$, and $\mathbb{P}(|\hat{S}^C(V_1) - S^C(V_1)| > \epsilon | V_1 = v) \le B n^2 e^{-n\epsilon^2/B}$. Using these relations, (28), the elementary relation

$$\frac{1}{x} = \frac{1}{y} + \frac{y - x}{y^2} + \frac{(y-x)^2}{y^2 x} \quad \text{where} \quad xy \ne 0,$$

and a straightforward calculation we verify (27). Theorem 3 is proved.

Proof of Theorem 4. We begin with several calculations that shed light on the oracle's Fourier estimate $\bar{\theta}_j^*$ defined in (17). First, using (15) and (16), let us show that $\bar{\theta}_j^*$ is unbiased estimate of θ_j. Indeed,

$$\mathbb{E}\{\bar{\theta}_j^*\} = \mathbb{E}\{\frac{A \Delta \varphi_j(AV) I(AV \in [0,b])}{w(1) S^C(AV)}\} = \int_0^b \frac{f^{AV,\Delta,A}(v,1,1) \varphi_j(v)}{w(1) S^C(v)} dv = \int_0^b \frac{w(1) f^T(v) S^C(v) \varphi_j(v)}{w(1) S^C(v)} dv = \theta_j.$$

Next the variance of $\bar{\theta}_j^*$ is evaluated. Set $d_1 := b^{-1} \int_0^b \frac{f^T(t)}{w(1) S^C(t)} dt$, and write using (15),

$$\mathbb{E}\{(\bar{\theta}_j^* - \theta_j)^2\} = n^{-1} \mathbb{E}\{(\frac{A \Delta \varphi_j(AV) I(AV \in [0,b])}{w(1) S^C(AV)} - \theta_j)^2\} = n^{-1} [\int_0^b \frac{f^T(v) \varphi_j^2(v)}{w(1) S^C(v)} dv - \theta_j^2] = n^{-1} d_1 (1 + o_j(1)).$$

Now we can follow the proof of Theorem 2 with the only change in replacing parameter d by d_1. Theorem 4 is verified.

Now let us check the assertion in the paragraph below line (21), which states that the estimate (19) of $w(\delta)$ and the estimate (20) of S^C can be used by the plug-in density estimator. We begin with $\bar{w}(1)$ defined in (19). Because $w(1) > 0$, the estimate $\bar{w}(1)$ dominates $\tilde{w}(1) := \sum_{l=1}^n A_l\Delta_l/[1 + \sum_{i=1}^n \Delta_i] = n^{-1}\sum_{l=1}^n A_l\Delta_l/[n^{-1} + n^{-1}\sum_{i=1}^n \Delta_i]$. Now recall that $w(1) := \mathbb{P}(A = 1|\Delta = 1) = \mathbb{P}(A\Delta = 1|\Delta = 1) = \mathbb{P}(A\Delta = 1)/\mathbb{P}(\Delta = 1)$. Further, a direct calculation allows us to conclude that $\mathbb{E}\{n^{-1} + n^{-1}\sum_{l=1}^n A_l - \mathbb{P}(A = 1)\} = n^{-1}$. Combining the above-presented formulas we get that the moment inequality $\mathbb{E}\{[\tilde{w}(1) - w(1)]^{2k}\} \le B_k n^{-k}$, $B_k < \infty$ holds for any integer k. This is the desired property of the estimate. Due to the symmetry in the right censoring between censored and uncensored observations, the estimate $\bar{w}(0)$ is analyzed similarly and its satisfies the same moment inequality.

Next we are considering the estimator $\bar{S}^V(v)$ of $S^V(v)$ defined in (21). Recall that $V := \min(T, C)$ is a continuous lifetime and write, $S^V(v) = \mathbb{P}(V \ge v) = \mathbb{P}(V \ge v, \Delta = 1) + \mathbb{P}(V \ge v, \Delta = 0)$. Now we need to understand how to express events in the two last probabilities via the observed MAR variables. Write

$$\mathbb{P}(V \ge v, \Delta = 1, A = 1) = \mathbb{P}(A = 1|V \ge v, \Delta = 1)\mathbb{P}(V \ge v, \Delta = 1) = w(1)\mathbb{P}(V \ge v, \Delta = 1).$$

Further, we have $\mathbb{P}(V \ge v, \Delta = 1, A = 1) = \mathbb{P}(AV \ge v, \Delta = 1, A = 1)$. Accordingly, the first sum on the right side of (21) is the plug-in sample mean estimate of $\mathbb{P}(V \ge v, \Delta = 1)$. Absolutely similarly we conclude that the second sum on the right side of (21) is the plug-in sample mean estimate of $\mathbb{P}(V \ge v, \Delta = 0)$. Accordingly, $\bar{S}^V$ is the empirical survival function for the MAR, and note that this estimate is the sum of two independent statistics based on uncensored and censored observations, respectively. Next, to shed light on the estimate $\bar{S}^C$ defined in (20), let us compare it with the MCAR estimate $\hat{S}^C$ defined in (11). Let w be given, then (8) and (9) imply that $(wn)^{-1}\sum_{l=1}^n I(A_lV_l \ge v)$ is the MCAR empirical survival estimate of S^V. This sheds light on the estimate (20) and allows us to analyze the MAR case similarly to the above-presented analysis of the MCAR case.

5. Conclusion

The developed theory of efficient density estimation explains what can and cannot be achieved for missing censored data, and it also quantifies how smoothness of an estimated density, the rate of censoring and the missingness affect the estimation. An interesting corollary of the theory is that the rougher the density, the less sensitive density estimation is to the missingness. The theory also motivated the proposed methodology of E-estimation for small samples. The E-estimator can be used for both missing and complete (no missingness) censored data. The former E-estimator, based on missing censored data and which is of interest on its own, can be also used as the benchmark to study different imputation procedures using the same E-estimator based on imputed (filled in) missing censored observations. The estimators were studied via analysis of real and simulated missing censored datasets. Among tested imputation methods, multiple resampling can be recommended for the complete missing model and multiple hot deck for the partial missing model. Datasets and R code are provided.

Let us mention several open topics for future research: (i) Missing not at random, discussed in the Introduction, prevents us from consistent density estimation. It is of interest to understand what additional information is necessary for consistent estimation, and then develop the corresponding theory and methodology of estimation. (ii) Analysis of

the more complicated missing left truncated and right censored (LTRC) data is another interesting and open problem. Discussion and known results for the case of no missing LTRC can be found in Efromovich (2018). In LTRC without missing we observe (V, Δ) only if V exceeds a truncation lifetime Z and then we observe (Z, V, Δ), otherwise we even do not know that there was a realization of (V, Δ). This model, due to the extra truncating lifetime Z, creates many interesting missing mechanisms. (iii) Another practically important model of missing censored data is when the availability A depends on an auxiliary variable X. For consistent estimation, this variable either must be observed or an extra sample from (X, A) should be available. Then the main issue is how well the availability likelihood can be estimated and what quality of estimation is sufficient for matching the oracle who knows the availability likelihood. (iv) It is of interest to test the Kaplan–Meier estimator and its modifications suggested for small samples in Efron (1967), Lagakos (1979), Dinse (1985) and Meier et al. (2004). (v) Even for the case of no missing, estimation of a bivariate density for censored data is an extremely complicated problem, see Efromovich (2022). It is of interest to understand what may be done for missing censored bivariate data. (vi) Hazard rate estimation is another classical topic to explore.

Acknowledgements

We would like to thank the Co-Editor Prof. Ana Colubi, the Associate Editor and three reviewers for valuable and constructive suggestions which greatly improved the paper. The research was supported by NSF Grant DMS-1915845.

Supplementary Material

Supplementary Material includes color version of the plots in Section 3, more results of the numerical study, data, and R code.

References

Aalen, O., Borgan, O. and Gjessing, H. (2008). *Survival Analysis and Event History Analysis: A Process Point of View.* New York: Springer.

Austin, P., White, I., Lee, D., and van Buuren, S. (2021). Missing data in clinical research: a tutorial on multiple imputation (review). *Canadian Journal of Cardiology* **37**, 1322–1331.

Bertsimas, D., Pawlowski, C. and Zhuo, Y. (2018). From predictive methods to missing data imputation: an optimization approach. *Journal of Machine Learning Reserach* **18**, 1–39.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–436.

Cuparic, M. and Milosevic, B. (2023). To impute or to adapt? Model specification tests' perspective. *Statistical Papers*, 1–19.

Comte, F., Guilloux, A. and Brunel, E. (2014). *Estimation/Imputation Strategies for Missing Data in Survival Analysis*. In *Statistical Models and Methods for Reliability and Survival Analysis* Hoboken: Wiley, 229–252.

Dikta, G. (1998). On semiparametric random censorship models. *Journal of Statistical Planning and Inference* **66**, 253–279.

Dinse, G. (1985). An alternative to Efron's redistribution-of-mass construction of the Kaplan–Meier estimator. *The American Statistician* **39**, 299–300.

Efromovich, S. (1985). Nonparametric estimation of a density with unknown smoothness. *Theory Probability and its Applications* **30**, 557–568.

Efromovich, S. (1989). On sequential nonparametric estimation of a density. *Theory Probability and its Applications* **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. *JASA* **96**, 667–685.

Efromovich,. S. (2013). Adaptive nonparametric density estimation with missing observations. *Journal of Statistical Planning and Inference* **143**, 637–650.

Efromovich, S. (2018). *Missing and Modified Data in Nonparametric Estimation: With R Examples*. Chapman and Hall/CRC.

Efromovich, S. (2022). Nonparametric bivariate density estimation for censored lifetimes. *The Annals of Statistics* **50**, 2767–2792.

Efron , B. (1967). The two sample problem with censored data. *Proceedings of the 5th Berkeley Symposium* **4**, 831–853. Berkeley : University of California Press.

Efron, B. (1994). Missing Data, Imputation, and the Bootstrap. *Journal of the American Statistical Association* **89**, 463–475.

Flemming, T. and Harrington, D (1991). *Counting Processes and Survival Analysis.* Hoboken: Willey.

Gijbels, I., Lin, D. and Ying, Z. (2007). *Non- and Semi-Parametric Analysis of Failure Time Data with Missing Failure Indicators*. Lecture Notes-Monograph Series, **54**, 203–223.

He, Y., Zhang, G. and Hsu, C.-H. (2021). *Multiple Imputation of Missing Data in Practice: Basic Theory and Analysis Strategies*. Boca Raton: CRC Press.

Heymans, M. and Twisk, J. (2022). Handling missing data in clinical research. *Journal of Clinical Epidemiology* **151**, 185–188.

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

Golubev, G. (1992). LAN in problems of non-parametric estimation of functions and lower bounds for quadratic risks. *The Probability Theory and Its Applications* **36**, 152–157.

Hu, X., Lawless, J. and Suzuki, K. (1998). Non parametric estimation of a lifetime distribution when censoring times are missing. *Technometrics* **40**, 3–13.

Huque, M., Carlin, J., Simpson,J. and Lee, K. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Reserach Methodology* **168**, 1–16.

Johnstone (2023). *Gaussian Estimation: Sequence and Wavelet Models.* Manuscript, Stanford: Stanford Univ.

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

Kaplan, E. and Meier, P. (1958). Nonparanmetric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

Lagakos , S. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics* **35**, 139–156.

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

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken: Wiley.

Little, R. (2021). Missing data assumptions. *Annual Review of Statistics and Its Application* **8**, 89–107.

Maksymiuk, A. et al. (1994). Sequencing and schedule effects of cisplatin plus etoposide in small cell lung cancer results of a north central cancer treatment group randomized clinical trial. *Journal of Clinical Oncology* **12**, 70–76.

McKeague, I. and Subramanian, S. (1998). Product-limit estimators and Cox regression with Missing censoring information. *Scandinavian Journal of Statistics* **25**, 589–601.

Meier, P., Karrison, T., Chappell, R. and Xie, H. (2004). The price of Kaplan–Meier. *Journal of the American Statistical Association* **99**, 890–896.

Moghaddam, S., Newell, J. and Hinde, J. (2022). A Bayesian Approach for Imputation of Censored Survival Data. *Stats.* **5**, 89–107.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. Chichester: Wiley.

Nikolskii, S. (1975). *Approximation of Functions of Several Variables and Embedding theorems*. New York: Springer.

Poynor, V. and Kottas, A. (2019). Nonparametric bayesian inference for mean residual life functions in survival analysis. *Biostatistics (Oxford, England)* **20**, 240–255.

Rotnitzky, A. and Robins, J. (2005). Inverse probability weighted estimation in survival analysis. *Encyclopedia of Biostatistics* **4**, 2619–2625.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Satten, G. and Datta, S. (2001) The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average. it The American Statistician **55**, 207–210.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hill.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Subramanian, S. (2003). The Missing Censoring-Indicator Model of Random Censorship. *Handbook of Statistics: Elsevier* **23**, 123–141.

Subramanian, S. (2004). Asymptotically efficient estimation of a survival function in the missing censoring indicator model. *Journal of Nonparametric Statistics* **16**, 797–817.

Subramanian, S. (2006). Survival analysis for the missing censoring indicator model using kernel density estimation techniques. *Stat Methodol.* **3**, 125–136.

Subramanian, S. (2009). The multiple imputations based Kaplan–Meier estimator. *Statistics and Probability Letters* **79**, 1906–1914.

Subramanian, S. (2011). Multiple imputations and the missing censoring indicator model. *Journal of Multivariate Analysis* **102**, 105–117.

Subramanian, S. and Zhang, P. (2013). Model-based confidence bands for survival functions. *Journal of Statistical Planning and Inference* **143**, 1166–1185.

Titterington, D. M. and Mill, G. M. (1983). Kernel-based density estimates from incomplete data. *Journal of the Royal Statistical Society. Series B (Methodological)* **45** , 258–266.

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

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. New York: Springer.

Ungolo, F., Christiansen, M. C., Kleinow, T. and MacDonald, A. S. (2019). Survival analysis of pension scheme mortality when data are missing. *Scandinavian Actuarial Journal* **6**, 523–547.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.

van der Laan, M. and McKeague, I. W. (1998). Efficient estimation from right-censored data when failure indicators are missing at random. *The Annals of Statistics* **26**,164–182.

Wang, Q. (2008). Probability density estimation with data missing at random when covariables are present. *Journal of Statistical Planning and Inference* **138**, 568–587.

Wang J. (2011). Estimation of lifetime distribution with missing censoring. *Journal of Data Science* **9**, 331–343.

Wang, Q., Liu, W. and Liu C. (2009). Probability density estimation for survival data with censoring indicators missing at random. *Journal of Multivariate Analysis* **100**, 835–850.

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

Wellner, J. (1982). Asymptotic optimality of the product limit estimator.*Annals of Statistics* **10**, 595–602.

Ying, Z., Jung, S. H., and Wei, L. J. (1995). Survival analysis with median regression models. *JASA* **90**, 178–184.

Zou, Y.-Y. and Liang, H.-Y. (2020). CLT for integrated square error of density estimators with censoring indicators missing at random. *Statistical Papers* **61**, 2685–2714.

Zou, Y.-Y. , Liang, H.-Y. and Zhang, J.-J. (2015). Nonlinear wavelet density estimation with data missing at random when covariates are present. *Metrika* **78**, 967–995.