UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Bachelor of Science Thesis

Christian Rubjerg Hejstvig-Larsen and Otto Grøn Roepstorff

Summary estimands for survival analysis

Alternatives to the Cox hazard ratio

Supervised by Professor Torben Martinussen
Co-supervised by Professor Susanne Ditlevsen
Department of Mathematical Sciences
University of Copenhagen, Denmark

Handed in October 11, 2024

Abstract

In this thesis, we describe different methods for analyzing survival data. First, we develop the mathematical concepts of survival data and introduce key probabilistic results, and then we discuss the semiparametric Cox model. We show that the Cox model is problematic when misspecified, which motivates the construction of nonparametric estimators. We use functionals and influence functions to derive the one-step estimator for a summary estimand of survival data and display some of the one-step estimator's convenient properties. We conclude the thesis with a discussion of the nonparametric approach.

Contents

1 Introduction

Survival analysis is a statistical field revolving around data where the response of interest is the elapsed time between some baseline event and an event of interest. This time period is called the survival time. Examples of survival times include the time between surgery and death, the interval between the birth of the 1st and 2nd child, and the period between graduation and obtaining a job. Survival analysis is an important component in many research fields such as medical research, where the goal could be to understand the effect of a treatment on the survival times of patients. In the first part of this thesis, we introduce the notation and mathematical framework for survival analysis in order to establish how to analyze survival data.

There are different semiparametric and nonparametric approaches to estimating the effect of covariates on survival times. The Cox model, proposed by Sir David Cox, is one of the most famous semiparametric approaches, and it enjoys widespread use in research and in practice. The extensive use of the model is largely due to its efficient estimation under certain conditions, convenient use and easy interpretation. However, the Cox model is not always a suitable model and should be used with caution. In the second part of the thesis we discuss the strengths and limitations of the Cox model, and we emphasize the need for careful consideration when using the model.

The limitations of the Cox model motivates an alternative estimation approach, which is the focus of the third and final part of the thesis. We ask the research question: *"What is the probability that a treatment prolongs the survival time"* and attempt to answer this question with nonparametric estimation. To do so we introduce the theory of efficient influence functions and their uses in nonparametric estimation. We provide a range of tools to derive efficient influence functions, and we describe how to use them to construct the one-step estimator. Finally, we link the nonparametric framework to survival data and derive the one-step estimator for the estimand that answers the aforementioned research question.

2 Survival data

In survival analysis we aim to conduct inference based on the survival time, T^*, which denotes the time elapsed between a fixed starting time (e.g. the time of surgery) and the time of some event of interest (e.g. death). The variable T^* is a positive, real-valued random variable, which is often recorded with other explanatory variables (e.g. height, age etc.) denoted by W, such that the survival time and covariates can be summarized in the tuple (T^*, W). In practice, we rarely observe the survival time T^* due to *censoring*. Censoring occurs when some real-life circumstance shortens the time interval we observe, e.g research studies have limited lengths or researchers may lose contacts to individuals of interest due to emigration. We will focus on right-censoring, which means that the censoring shortens the observed time interval from the right. Let C be the random variable denoting a (right) *censoring time*. We observe the true survival time, T^*, if $T^* < C$ and C otherwise, which leads to the construction of both the *observed survival time*, T, and the censoring indicator, Δ, defined as

$$T = \min(T^*, C) = T^* \wedge C, \quad \Delta = I(T^* \leq C) = I(\text{True survival time has been observed})$$

It is important to include the information of Δ in observed survival times, since exclusion of censored data points can lead to biased estimates (Skovgaard and Rosthøj, 2019), and survival data is therefore summarized in triples of the form (T, W, Δ).

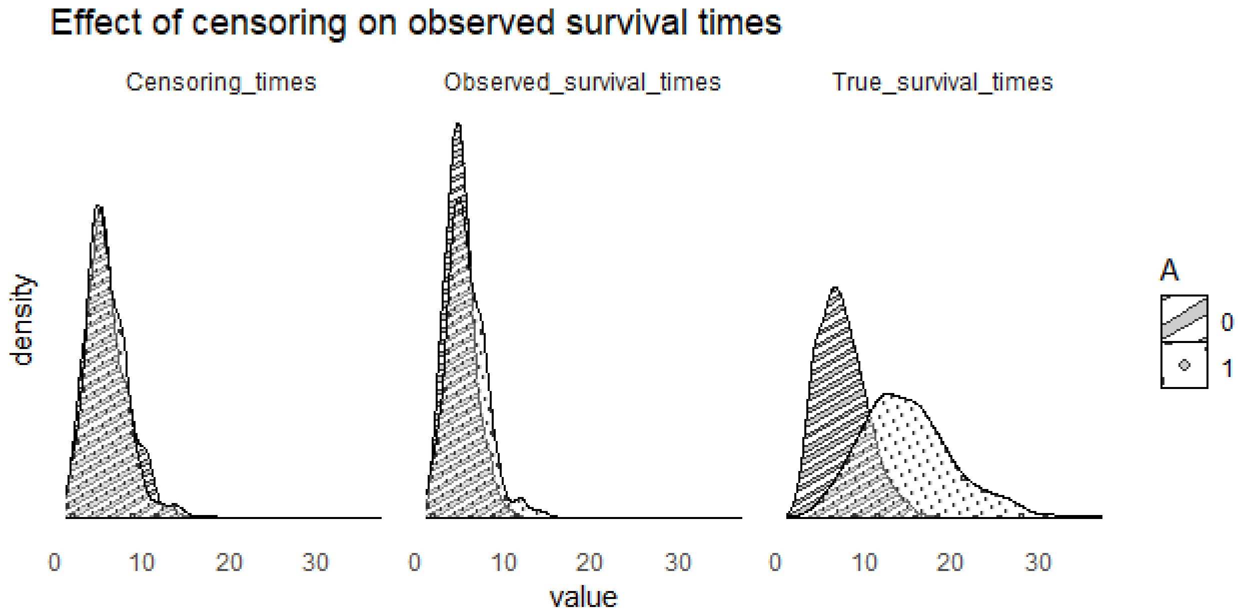

Figure 1: *Density of censoring times C (left), observed survival times T (middle) and true survival times T^* (right) stratified on whether treatment has been given ($A = 1$) or not ($A = 0$)*

Example 2.0.1. A patient with covariates W enters a 4-year study on the survival time after heart surgery. The patient survives until 10 years after the surgery. Since the study has been concluded before the patient dies, the true survival time, $(T^* = 10)$, has been censored by the length of the study, $(C = 4)$, and we only observe the censored survival time $T = 4$. We store the data of the individual as the data triple $(T, W, \Delta) = (4, W, 0)$.

Example 2.0.2 (Importance of censoring indicators)**.** Let T^* be the true survival times that are dependent on the dichotomous variable A, such that the survival times are longer when $A = 1$. This could happen when A is an indicator for a treatment with positive effects on survival times. Let C be the censoring times generated independently of A and T^*. We then generate a data set, where the observed survival times are heavily censored, such that we only observe about 15 % of the true survival times. Figure 1 illustrates how the density of the true survival times (right plot) look entirely different from density of the observed survival times (middle plot), when a heavy censoring scheme is employed (left plot). The density of the observed survival times looks much more like the density of the censoring times. If we didn't keep the Δ-component, it would be difficult to ever realize the possibility that giving the treatment $(A = 1)$ could prolong the survival times.

Censoring poses an issue in inference from survival data, but with the right tools it can be handled relatively easily as long as the assumption of *independent censoring* is satisfied, which is discussed in greater detail in Theorem 3.2.6.

2.1 Survival functions and hazard rates

Definitions and claims in this section are from Aalen, Borgan, and Gjessing (2008) unless noted otherwise. There are two fundamental concepts in survival analysis: *survival functions* and *hazard rates*. A survival function, S, describes the expected proportion of a population that has not yet experienced the event of interest at time t. Formally it can be defined as follows:

Definition 2.1.1. Let $T^* \sim \mathbb{P}$ be a random variable denoting the time elapsed between two events of interest. The survival function $S : \mathbb{R}_+ \to [0, 1]$ is defined as

$$S(t) = \mathbb{P}(T^* > t)$$

If F is the distribution function of T^*, then $S(t) = 1 - F(t)$

The survival function describes the unconditional probability of not having experienced the event of interest at time t, while the hazard rate, $\alpha(t)$, describes the conditional probability of experiencing the event in an infinitesimally small interval $[t, t + dt)$ for an individual that has not yet experienced the event of interest. The formal definition is

Definition 2.1.2. Assume that $T^* \sim \mathbb{P}$ is absolutely continuous with probability density $f(t) = F'(t)$. The *hazard rate* of T^* is given as

$$\alpha(t) = \lim_{h \to 0^+} \frac{1}{h}\mathbb{P}(t \leq T^* < t + h | T^* \geq t) \tag{2.1}$$

The integral

$$H(t) = \int_0^t \alpha(s) ds \tag{2.2}$$

is called the *cumulative hazard rate*.

Equation (2.1) can be interpreted as the instantaneous probability of experiencing the event of interest, given it has not yet occurred. The hazard rate and survival function are closely related. In particular, we can express the hazard rate as

$$\begin{aligned}
\alpha(t) &= \lim_{h \to 0^+} \frac{1}{h}\mathbb{P}(t \leq T^* < t + h \mid T^* \geq t) \\
&= \lim_{h \to 0^+} \frac{1}{h} \frac{\mathbb{P}(t \leq T^* < t + h, T^* \geq t)}{\mathbb{P}(T^* \geq t)} \\
&\overset{(2.1)}{=} \frac{1}{S(t)} \lim_{h \to 0^+} \frac{\mathbb{P}(T^* \leq t + h) - \mathbb{P}(T^* \leq t)}{h} \\
&= \frac{1}{S(t)} \lim_{h \to 0^+} \frac{F(t + h) - F(t)}{h} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}
\end{aligned} \tag{2.3}$$

assuming $\mathbb{P}(T^* \geq t) > 0$. The assumption of absolute continuity ensures that $F'(t) = f(t)$ is well-defined and that $\mathbb{P}(T^* \geq t) = S(t)$. An equally important identity can be achieved by the fundamental theorem of calculus and the fact that $S(0) = 1 - F(0) = 1$, such that

$$S(t) = \exp(-\int_0^t \alpha(s) ds) = \exp(-H(t)) \tag{2.4}$$

since

$$H(t) \overset{(2.2)+(2.3)}{=} -\int_0^t \frac{S'(s)}{S(s)} ds = -\left[\log S(s)\right]_0^t = -\log S(t)$$

The data triple (T, W, Δ), the survival function and the hazard rate cover the basic concepts of survival analysis.

3 Counting processes and martingales

Stochastic processes provide a natural framework for studying properties of estimators and significance tests for data gathered over a period time including survival data. Counting processes and related counting process martingales are frequently encountered in survival analysis, and this section aims to introduce the key results related to these processes. To keep the section brief we assume familiarity with basic measure and probability theory including conditional expectation. We gloss over most technical details, like not writing a.s. after equalities that include conditional expectations, to maintain focus on developing the basic set of tools from stochastic processes required in survival analysis. The invested reader is referred to Fleming and Harrington (1991) for proofs of the theorems and claims, and unless stated otherwise all definitions and theorems are from Fleming and Harrington (1991).

3.1 Preliminary definitions and martingales

The key motivation for stochastic processes is the need to build a mathematical framework to describe how a family of random variables evolves over time. The formal definition of such a stochastic process is

Definition 3.1.1 (Stochastic process)**.** A *stochastic process* is a family of random variables $X = \{X(t) : t \in \Gamma\}$ indexed by a set Γ, and where all $X(t)$ are defined on the same probability space $(\Omega, \mathcal{F}, P)$.

We usually abuse notation and simply write $X(t)$ for the stochastic process $\{X(t) : t \in \Gamma\}$, when Γ is unambiguous from the context. In survival analysis we typically assume Γ to be the set of positive reals $\mathbb{R}_+$ denoting time, and $X(t)$ will hence denote a continuous-time process. We say, that a stochastic process has an attribute if each sample path of the process has this attribute with probability 1. Like random variables, stochastic processes can be continuous, increasing, have left-hand or right-hand limits etc. By common nomenclature we refer to right-continuous processes with left-hand limits as *cadlag* stochastic processes.

When we describe the progress of a stochastic process, it is natural to consider the prior behaviour of the process and keep track of the history of the random variable. This leads to the definition of a filtration.

Definition 3.1.2 (Filtration)**.** A filtration $(\mathcal{F}_t)_{t\geq 0}$ on a probability space $(\Omega, \mathcal{F}, P)$ is an increasing sequence of sub-sigma-algebras of $\mathcal{F}$ such that $\mathcal{F}_t \subset \mathcal{F}_s$ for all $t < s$.

Filtrations can be viewed as a collection of information that increases over time, such that $\mathcal{F}_t$ contains the available information about our background space Ω at time $t \geq 0$. The interplay between the stochastic processes and filtrations can be formalized as follows:

Definition 3.1.3 (Adapted process)**.** A stochastic process $\{X(t) : t \geq 0\}$ on a probability space $(\Omega, \mathcal{F}, P)$ is *adapted* to a filtration $(\mathcal{F}_t)_{t\geq 0}$ if $X(t)$ is $\mathcal{F}_t$-measurable for all $t \geq 0$. If $X(t) \in \mathcal{F}_{t-}$ for all $t \geq 0$, where $\mathcal{F}_{t-} = \sigma(\cup_{s<t}\mathcal{F}_s)$, we say that $X(t)$ is *predictable* with respect to $(\mathcal{F}_t)_{t\geq 0}$.

An adapted process of particular importance is the martingale process:

Definition 3.1.4 (Martingale)**.** A cadlag stochastic process, $M(t)$, $t \geq 0$, on a probability space $(\Omega, \mathcal{F}, P)$ is a *martingale* with respect to $(\mathcal{F}_t)_{t\geq 0}$ if $M(t)$ is adapted to $(\mathcal{F}_t)_{t\geq 0}$, $E|M(t)| < \infty$ for all $t \geq 0$, and if it satisfies the *martingale property*

$$E\{M(t+s)|\mathcal{F}_t\} = M(t) \quad \text{for all } s \geq 0, t \geq 0. \tag{3.1}$$

M is called a *submartingale* if the equality in Eq. (3.1) is replaced by "$\geq$".

Intuitively, a martingale is a process where the expectation to the future value of the process given current knowledge is exactly the present value. An equivalent formulation of Eq. (3.1) is

$$E\{M(t)-M(s)|\mathcal{F}_s\}=0 \quad \text{for all } t\geq s\geq 0 \tag{3.2}$$

Martingales naturally appear in the counting process framework, and martingale theory is useful for developing asymptotic results for estimation in survival analysis. We find it helpful to introduce the following fundamental properties of martingales:

Proposition 3.1.5. *Let $M(t)$ be a martingale with respect to $(\mathcal{F}_t)_{t\geq 0}$ on $(\Omega,\mathcal{F},P)$. The following holds*

(a) $\forall t>0$ *it holds that* $E\{dM(t)|\mathcal{F}_{t_-}\}=0$*, where* $dM(t)=M((t+dt)_-)-M(t_-)$

(b) $EM(t)=E\left\{E(M(t)\mid \mathcal{F}_s)\right\}$ *for* $t\geq 0, s\geq 0$

(c) If $M(t)$ *is a mean-zero martingale such that* $EM(0)=0$*, then for all* t ≥ 0 *it holds that* $EM(t)=0$

(d) $\operatorname{cov}\{M(t)-M(s),M(v)-M(u)\}=0$ *for* $0\leq s\leq t\leq u\leq v$

Proof. The proofs are done one by one.

(a)

$$\begin{aligned}E(dM(t)|\mathcal{F}_{t_-}) &= E\left(M((t+dt)_-)-M(t_-)|\mathcal{F}_{t_-}\right)\\ &= E(M((t+dt)_-)|\mathcal{F}_{t_-})-E(M(t_-)|\mathcal{F}_{t_-})\\ &\overset{\star}{=} M(t_-)-M(t_-)=0\end{aligned}$$

where we use linearity of conditional expectation and the martingale property from Eq. (3.1) at $\star$. This shows that Eq. (3.1) and Eq. (3.2) are equivalent

(b)

$$E\{M(t)\}=\int_\Omega M(t)dP \overset{(1)}{=} \int_\Omega E(M(t)|\mathcal{F}_s)dP=E\left\{E(M(t)|\mathcal{F}_s)\right\} \tag{3.3}$$

where (1) follows by definition of $E(M(t)|\mathcal{F}_t)$ and because $\Omega\in\mathcal{F}_t$. This result is called expectation by conditioning and holds for any conditional expectation (Hansen, 2023).

(c) Follows directly from above

$$EM(t) \overset{Eq.\ (3.3)}{=} E\{E(M(t)|\mathcal{F}_0)\} \overset{Eq.\ (3.1)}{=} EM(0)=0$$

(d)

$$\begin{aligned}&\operatorname{cov}(M(t)-M(s),M(v)-M(u))\\ &=E\{[M(t)-M(s)][M(v)-M(u)]\}-E\{M(t)-M(s)\}E\{M(v)-M(u)\}\\ &\overset{Eq.\ (3.3)}{=} E\{[M(t)-M(s)][M(v)-M(u)]\}-E\{E(M(t)-M(s)\mid \mathcal{F}_s\}E\{M(v)-M(u)\}\\ &\overset{Eq.\ (3.3)+Eq.\ (3.2)}{=} E\{E([M(t)-M(s)][M(v)-M(u)])|\mathcal{F}_t]\}-0\\ &\overset{\star}{=} E\{[M(t)-M(s)]E(M(v)-M(u)|\mathcal{F}_t)\} \overset{Eq.\ (3.1)}{=} E\{(M(t)-M(s))\cdot 0\}=0\end{aligned}$$

where $\star$ uses the predictability of $M(t)$ and $M(s)$ with respect to $\mathcal{F}_t$.

□

3.2 Counting processes

In survival analysis we are concerned with the occurrence of events over time and counting processes arise as a natural framework for analyzing survival data. In this section we describe counting processes and demonstrate how they relate to martingales. We start with the definition of a counting process:

Definition 3.2.1 (Counting process)**.** A *counting process* is a stochastic process $\{N(t) : t \geq 0\}$ adapted to a filtration $\{\mathcal{F}_t : t \geq 0\}$ with $N(0) = 0$ and $N(t) < \infty$ a.s., and whose paths are with probability one right-continuous, piecewise constant and have only jump discontinuities, with jumps of size +1.

The counting processes obtained from right-censored data take the shape $N(t) = I(T \leq t, \Delta = 1)$ and $N^C(t) = I(T \leq t, \Delta = 0)$, where T and Δ follow the notation presented in Section 2. $N(t)$ represents the total number of observed events up to time t and $N^C(t)$ represents the total number of censored events up to time t. Both of these processes are counting processes as $N(0) = N^C(0) = 0$, both are right-continuous and piecewise constant with at most one jump of size +1. Survival data usually consists of several counting processes, one for each observed individual. Grouping the individual counting processes leads to a multivariate counting process:

Definition 3.2.2 (Multivariate counting process)**.** A process $N(t) = (N_1(t), \ldots, N_k(t))$ is called a *multivariate counting process* if

(a) Each $N_j, j = 1, \ldots, k$ is a counting process.

(b) No two component processes, $N_i(t), N_j(t)$ for $i, j = 1, ..., k$, jump at the same time.

To aggregate a multivariate counting process to a one-dimensional counting process, we can simply sum the counting processes

$$N.(t) = \sum_{i=1}^{k} N_i(t)$$

Example 3.2.3. Let $N(t)$ be the counting process that describes the number of individuals suffering from infection t days after surgery. Then $N(5)$ denotes the total number of infections observed after 5 days. Furthermore, the number of individuals that suffered from infection in the time interval u to v days after surgery ($t \in (u, v]$) is given by $N(v) - N(u)$.

It is straight-forward to construct a counting process from survival data, and the goal is to describe the underlying structure of the counting process and go beyond the observed data. This is where the crucial Doob-Meyer decomposition comes into play.

Theorem 3.2.4 (Doob-Meyer decomposition (Fleming and Harrington, 1991))**.** *Let $X(t)$ be a non-negative submartingale with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$ on the measure space $(\Omega, \mathcal{F}, P)$*[1]*. Then there exists an $\mathcal{F}_t$-martingale, $M(t)$, and a right-continuous non-decreasing $\mathcal{F}_t$-predictable process, $\Lambda(t)$, such that $E(\Lambda(t)) < \infty$ and*

$$X(t) = \Lambda(t) + M(t) \tag{3.4}$$

for any $t \geq 0$. If $\Lambda(0) = 0$ a.s. then the decomposition is unique a.s.

[1] Technically $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ is required to be a stochastic basis (see Fleming and Harrington (1991) for additional details).

Since a counting process is a non-negative submartingale, the Doob-Meyer decomposition directly admits the following corollary:

Corollary 3.2.5. *Let $N(t)$ be a counting process adapted to the filtration $(\mathcal{F}_t)_{t\geq 0}$ with $E(N(t)) < \infty$ for any t. Then there exists a unique non-decreasing $\mathcal{F}_t$-predictable process, $\Lambda(t)$, with $\Lambda(0) = 0$ a.s., and $E(\Lambda(t)) < \infty$, such that $M(t) = N(t) - \Lambda(t)$ is a right-continuous $\mathcal{F}_t$-martingale for any $t \geq 0$.*

The Doob-Meyer decomposition and associated corollary states that a counting process, N, can be decomposed into a predictable part, Λ, and a random noise part, M. The predictable part, Λ, is called the *compensator* or *cumulative intensity* of N, and it encaptures the structure of the counting process. It can be shown that $E(N) = E(\Lambda)$ implying that M is a mean-zero martingale, and the martingale therefore has a natural interpretation as an error term with uncorrelated increments and mean zero as established in Proposition 3.1.5 (c) and (d). The Doob-Meyer theorem links counting processes to martingales and we are usually more interested in the behaviour of the martingale $M = N - \Lambda$ than the counting process due to the asymptotic properties of martingales. However, to determine the martingale we need knowledge of Λ, which motivates the following theorem for survival data counting processes:

Theorem 3.2.6. *Let $T^* \sim \mathbb{P}$ be an absolutely continuous survival time random variable with hazard rate $\alpha(t)$. Let C be a censoring time random variable, $T = T^* \wedge C$ the observed survival time random variable and $\Delta = I(T < C)$. Define*

$$N(t) = I(T \leq t, \Delta = 1), \quad N^C(t) \ = I(T \leq t, \Delta = 0), \quad \mathcal{F}_t = \sigma(\{N(u), N^C(u) : 0 \leq u \leq t\})$$

Then the process M given by

$$M(t) = N(t) - \int_0^t I(T \geq u)\alpha(u)du$$

is an $\mathcal{F}_t$ martingale if and only if it holds that

$$\alpha(t) = \lim_{h\to 0^+} \frac{1}{h}\mathbb{P}(t \leq T < t + h \mid T \geq t, C \geq t) \tag{3.5}$$

for $\mathbb{P}(T \geq t, C \geq t) > 0$.

There are three important points to take from Theorem 3.2.6. First, it shows that the process

$$\Lambda(t) = \int_0^t I(T \geq u)\alpha(u)du \tag{3.6}$$

is the compensator for the observed counting process $N(t) = I(T \leq t, \Delta = 1)$. The Doob-Meyer decomposition ensures uniqueness of the compensator since $\Lambda(0) = 0$.

Second, the condition in Eq. (3.5) on the hazard rate of T^* is called the condition of *independent censoring*. Eqs. (2.1) and (3.5) are identical except that $C \geq t$ is added to the conditioning. The right-hand side of Eq. (3.5) can therefore be interpreted as the hazard rate of T^* in the presence of censoring, and it states that censoring does not alter the underlying hazard rate. In the presence of covariates X the condition takes the form

$$\alpha(t \mid X) = \lim_{h\to 0^+} \frac{1}{h}\mathbb{P}(t \leq T^* < t + h \mid T^* \geq t, C \geq t, X) \tag{3.7}$$

Results in survival analysis typically rely on the assumption of independent censoring to hold.

Third, the hazard rate $\alpha(t)$ of T^* appears in $\Lambda(t)$. The cumulative intensity, $\Lambda(t)$, closely resembles the cumulative hazard, $H(t)$, except for the inclusion of the random variable $I(T \geq t)$ in the integral. The indicator is called the *at-risk indicator*, and it keeps track of whether an individual is still susceptible to experiencing the event of interest at time t. The three observations motivate the key concept of *intensity*

Definition 3.2.7 (Intensity process)**.** Let $N(t)$ be a counting process as in Theorem 3.2.6 and $\Lambda(t)$ be the cumulative intensity defined as

$$\Lambda(t) = \int_0^t I(T \geq u)\alpha(u)du = \int_0^t \lambda(u)du.$$

Then $\lambda(t) = I(T \geq t)\alpha(t)$ is called the *intensity process*.

Remark 3.2.8 (Cumulative hazard and cumulative intensity)**.** In later paragraphs it may seem like the cumulative hazard, $H(t)$, and the cumulative intensity, $\Lambda(t)$, are used interchangeably and the same goes for $\alpha(t)$ and $\lambda(t)$. The two concepts are closely related but not the same. The hazard is deterministic whereas the intensity is random due to the random at-risk indicator. Martinussen and Scheike (2006) interpret the hazard rate, $\alpha(t)$, as the deterministic (model) part of the intensity $\lambda(t)$.

The main motivation for the martingale decomposition in Theorem 3.2.6 is that there exists martingale limit theorems that give insight to the asymptotic error in our estimates. In the counting process setting we do not only observe martingale processes, but we also observe stochastic integrals with respect to martingales of the form

$$\sum_{i=1}^{n} \int_0^{\tau} H_i(u)dM_i(u) \tag{3.8}$$

where $M_i(t) = N_i(t) - \Lambda_i(t)$ is a counting process martingale with respect to a filtration $(\mathcal{F}_t)_{t\geq 0}$, and $H_i(t)$ is a $\mathcal{F}_t$-predictable, measurable, and often bounded process. τ represent the final time of interest e.g. the ending time of a clinical trial. As a slight abuse of notation, we concisely write Eq. (3.8) as

$$\sum_{i=1}^{n} \int_0^{\tau} H_i(u)dM_i(u) = \int HdM$$

whenever we only do linear operations. A surprising result is that the stochastic integral with respect to a martingale is itself a martingale under certain regularity conditions.

Theorem 3.2.9. *Let $N(t)$ be a counting process with respect to a right-continuous filtration $(\mathcal{F}_t)_{t\geq 0}$ such that $E(N(t)) < \infty$ for all t ≥ 0. Let*

(a) $M(t) = N(t) - \Lambda(t)$ be an $\mathcal{F}_t$-martingale where $\Lambda(t)$ is the cumulative $\mathcal{F}_t$-intensity with $\Lambda(0) = 0$.

(b) H be a bounded $\mathcal{F}_t$-predictable process.

Then the process

$$\int_0^t H(u)dM(u)$$

is an $\mathcal{F}_t$-martingale.

Theorem 3.2.9 ensures that we can use martingale theorems for stochastic integrals similar to Eq. (3.8). In addition, we can split the integral into the counting process part and intensity process part

$$\int H dM = \int H d(N - \Lambda) = \int H dN - \int H d\Lambda \tag{3.9}$$

by definition of the martingale process M (Aalen, Borgan, and Gjessing, 2008)[2]. The decomposition in Eq. (3.9) and Theorem 3.2.9 are strong tools for analyzing estimators as seen in the following example.

Example 3.2.10 (Nelson-Aalen estimator)**.** We are interested in estimating the cumulative hazard $H(t) = \int_0^t \alpha(u)du$ for some survival time random variable T^* by the *Nelson-Aalen* estimator. Consider a sample of independent observed survival times $(T_i, \Delta_i)_{i=1}^n$ of size n. Let $T_{i_1} < T_{i_2} < \cdots < T_{i_k}$ be the ordered collection of observed event times with $k \leq n$, where T_i is included in the collection if $\Delta_i = 1$. Let $Y(t)$ denote the total number of cases at risk at time t such that $Y(t) = \sum_{i=1}^n I(t \leq T_i)$ and let $N(t) = \sum_{i=1}^n I(T_i \leq t, \Delta = 1)$ be the observed event counting process. The Nelson-Aalen estimator is then defined as

$$\hat{H}(t) = \int_0^t \frac{I(Y(u) > 0)}{Y(u)} dN(u) = \sum_{\substack{T_{i_j} \leq t \\ j=1,\ldots k}} \frac{1}{Y(T_{i_j})} \tag{3.10}$$

where $I(Y(s) > 0)$ coupled with the convention $0/0 = 0$ is a technical condition ensuring that the expression is well defined. The Nelson-Aalen estimator is an estimator of

$$H^*(t) = \int_0^t I(Y(u) > 0)\alpha(u)du$$

which is equivalent to the cumulative hazard if there are individuals at risk for all $u \in [0, t]$ almost surely. The difference between the Nelson-Aalen estimator and $H^*(t)$ can be evaluated to:

$$\begin{aligned}
\hat{H}(t) - H^*(t) &= \int_0^t \frac{I(Y(u) > 0)}{Y(u)} dN(u) - \int_0^t I(Y(u) > 0)\alpha(u)du \\
&= \int_0^t \frac{I(Y(u) > 0)}{Y(u)} dN(u) - \int_0^t \frac{I(Y(u) > 0)}{Y(u)} d\Lambda(u) \\
&= \int_0^t \frac{I(Y(u) > 0)}{Y(u)} d(N(u) - \Lambda(u)) = \int_0^t \frac{I(Y(u) > 0)}{Y(u)} dM(u) \\
&= \sum_{i=1}^n \int_0^t \frac{I(Y(u) > 0)}{Y(u)} dM_i(u)
\end{aligned}$$

which is an integral similar to Eq. (3.8) and it follows from Theorem 3.2.9 that the integral is a martingale. This result can in turn be applied to show, that under regularity conditions $\sqrt{n}(\hat{H} - H)$ converges in distribution to a mean zero Gaussian martingale and the Nelson-Aalen estimator is thus an asymptotically unbiased estimator of the cumulative hazard.

[2] Formally, this requires advanced probabilistic arguments which are beyond the scope of the project

4 Cox's proportional hazards model

In the previous section we established the essential mathematical framework for working with survival data. The framework serves as the basis for understanding some of the models and estimators applied in survival analysis, and in this section we present the proportional hazards model also known as the Cox model. We do not give the probability theoretical proofs and divert the interested reader to Fleming and Harrington (1991) and Martinussen and Scheike (2006) for this.

4.1 Introduction

4.1.1 Notation

In this description of the Cox model, we consider the triple $(N(t), W(t), Y(t))$ adapted to a filtration $(\mathcal{F}_t)_{t\geq 0}$ where

- $N(t) = I(T \leq t, \Delta = 1)$ is the counting process for an individual with survival time T,
- $W(t) = (A, X(t))$ is a $(1 + p)$-dimensional covariate process assumed to be bounded and predictable w.r.t. $\mathcal{F}_t$,
- $A = I($Individual has received treatment$)$ is an indicator for treatment given and $X(t)$ is a p-dimensional row vector that represents all other recorded covariates,
- $Y(t) = I(T \geq t)$ is an at-risk process assumed to be predictable w.r.t. $\mathcal{F}_t$.
- $\mathcal{F}_t$ represents the statistical information accumulated at time t[3],
- $\mathcal{T} = \{t \in \mathbb{R} \cap [0, \tau] \mid \Delta N(t) = 1\}$ is the set of *jump times* for $N(t)$ in the time period $[0, \tau]$, where τ is the maximum length of time we choose to observe individuals.

We let $\Lambda(t)$ be the cumulative intensity process of $N(t)$ and $\lambda(t)$ the corresponding intensity process, such that $d\Lambda(t) = \lambda(t)dt$. The triple $(N(t), W(t), Y(t))$ is the counting process equivalent to the triple (T, W, Δ) introduced in Section 2 and carries the same information.

4.1.2 Cox's proportional hazards model

The proportional hazards model proposed by Sir David Cox in 1972, is a semiparametric model that offers a method for estimating the effect of covariates on survival probabilities. The model is a multiplicative hazards model, implying that the effects of the covariates are modelled on a multiplicative scale. In particular, the Cox model assumes that

(a) the intensity of a right-censored counting process takes the form:

$$\lambda(t \mid W(t)) = Y(t)\alpha(t \mid W(t)) = Y(t)\alpha_0(t)\exp(W(t)\beta) \tag{4.1}$$

where the column vector $\beta \in \mathbb{R}^{p+1}$ is the parameter of interest called the log-*relative* risk parameter, and $\alpha_0(t)$ is a nonparametric integrable baseline hazard (Martinussen and Scheike, 2006).

[3] Formally $\mathcal{F}_t = \sigma\{N(u), Y(u+), X(u+) : 0 \leq u \leq t\}$. A detailed discussion can be found in Fleming and Harrington (1991) chapter 4

(b) the survival times T^* are conditionally independent of the censoring times C given the recorded covariates W as in Eq. (3.7).

Assumption (a) has three important implications. First, it implies the hazard rate has the form $\alpha(t) = \alpha_0(t)\exp(W(t)\beta)$. Second, it assumes that the β-values are time-invariant. This is a strong assumption that is easily violated, for instance in the case of treatments that increase the short term risk of dying, but decrease the long term risk. Third, (a) implies the *proportional hazards* assumption. If we assume $X(t)$ is one-dimensional, the assumption can be shown as the following set of equalities

$$\frac{\alpha(t \mid X(t)+1)}{\alpha(t, X(t))} = \frac{\alpha_0(t)\exp((X(t)+1)\beta)}{\alpha_0(t)\exp(X(t)\beta)} = \exp((X(t)+1-X(t))\beta)) = \exp(\beta) \quad (4.2)$$

which shows that the relative risk is assumed to be constant with time. Since the baseline hazard cancels out in the derivation above, it will often suffice to estimate β only.

The term $\exp(\beta)$ in Eq. (4.2) is called the relative risk or the *Cox hazard ratio*, which is typically the figure of interest. In our thesis, we deal with a binary treatment variable, A, that indicates whether an individual has received treatment $(A = 1)$ or not $(A = 0)$. If we let β_A denote the parameter for the treatment covariate, A, then $\exp(\beta_A)$ is the instantaneous risk of experiencing the event of interest at any given moment in the treated group *relative* to the untreated group. If we estimate β_A to be smaller than zero, then the interpretation is that receiving treatment $(A = 1)$ prolongs the survival time, and opposite if β_A is greater than zero, while $\beta_A = 0$ indicates no effect of the treatment.

Assuming the Cox model structure for the hazard rate, the survival function given $W(t)$ can be expressed as:

$$S(t \mid W(t)) \overset{2.4}{=} \exp\{-H(t \mid W(t))\} = \exp\{-\int_0^t \alpha_0(s)\exp(W(s)\beta)ds\}$$

The Cox model is by far the most used regression model in survival analysis (Martinussen and Scheike, 2006).

4.1.3 Estimation procedure

Assume we have n independent observations of the triple $(N_i(t), W_i(t), Y_i(t))$ adapted to the information accrued over time, $\mathcal{F}_t$, for $t \in [0, \tau]$ similar to the setup outlined in Subsection 4.1.1. Ordinary maximum likelihood methods are insufficient to estimate the regression coefficients, due to the semiparametric nature of the Cox regression model. Instead, we find the parameter, β, as the maximizer of Cox's partial likelihood function:

$$\hat{\beta} = \text{argmax}_{\beta \in \mathbb{R}^{p+1}} L(\beta) = \text{argmax}_{\beta \in \mathbb{R}^{p+1}} \prod_{t=0}^{\tau} \prod_{i=1}^{n} \left(\frac{\exp(W_i(t)\beta)}{S_0(t,\beta)} \right)^{\Delta N_i(t)}$$

where

$$S_0(t,\beta) = \sum_{i=1}^{n} Y_i(t)\exp(W_i(t)\beta),$$

and assuming at least one individual is at risk at all times $t \in [0,\tau]$ such that $S_0(t,\beta) \neq 0$. The product

$$\prod_{t=0}^{\tau}\prod_{i=1}^{n}\left(\frac{\exp(W_i(t)\beta)}{S_0(t,\beta)}\right)^{\Delta N_i(t)}$$

appears to be uncountable, since it is taken over all $t \in [0,\tau]$. However, the $\Delta N_i(t)$ terms are only 1 in the jump times $\mathcal{T}_i$ of $N_i(t)$ and 0 otherwise. Since $N_i(t)$ by definition is almost surely finite, it almost surely jumps a finite number of times. Hence, the cardinality of $\mathcal{T}_i$ is almost surely finite, and the product is an almost surely finite product of terms evaluated at the jump times $\mathcal{T}_i$. To derive the score of the partial likelihood function, we first define $S_1(t,\beta)$ as the derivative of $S_0(t,\beta)$ with respect to β:

$$S_1(t,\beta) = \sum_{i=1}^{n} Y_i(t)\exp(W_i(t)\beta)W_i(t)$$

The partial log-likelihood function can then be determined as

$$\ell(\beta) = \log(L(\beta)) = \sum_{t=0}^{\tau}\sum_{i=1}^{n}[W_i(t)\beta - \log(S_0(t,\beta))]\Delta N_i(t),$$

and finally the score function is derived as

$$\begin{aligned} U(\beta) = \frac{d}{d\beta}\ell(\beta) &= \sum_{t=0}^{\tau}\sum_{i=1}^{n}\left(W_i(t) - \frac{S_1(t,\beta)}{S_0(t,\beta)}\right)\Delta N_i(t) \\ &= \sum_{i=1}^{n}\int_0^{\tau}\left(W_i(t) - \frac{S_1(t,\beta)}{S_0(t,\beta)}\right)dN_i(t) \end{aligned}$$

The estimate of the true β is the vector $\hat{\beta} \in \mathbb{R}^{p+1}$ that solves

$$U(\hat{\beta}) = 0 \tag{4.3}$$

For a fixed $\beta \in \mathbb{R}^{p+1}$ the cumulative baseline hazard, H_0, is estimated with a Nelson-Aalen type estimator:

$$\hat{H}_0(t,\beta) = \sum_{i=1}^{n}\int_0^{t}\frac{1}{S_0(u,\beta)}dN_i(u) = \int_0^{t}\frac{1}{S_0(u,\beta)}dN.(u)$$

However, the β-parameter will often be unknown and it has to be estimated. Using $\hat{\beta}$ obtained from solving Eq. (4.3), we can estimate the cumulative baseline hazard with the *Breslow estimator*:

$$\hat{H}_0(t,\hat{\beta}) = \int_0^{t}\frac{1}{S_0(u,\hat{\beta})}dN.(u)$$

4.1.4 Inference

The Cox model is convenient, because it gives rise to asymptotic distributions for $\hat{\beta}$ and thereby the ability to construct confidence intervals of the point estimates. The key to the asymptotic results, is to realize, that the score of the partial likelihood function evaluated in the true parameter, is a martingale.

Lemma 4.1.1. *Let* $\{N_1(t), W_1(t), Y_1(t)\}, ..., \{N_n(t), W_n(t), Y_n(t)\}$ *be* n *independent data triples obtained from right-censored data, each satisfying the conditions outlined in Subsection 4.1.1 w.r.t. the information accrued over time* $\mathcal{F}_t$, $t \geq 0$. *Let*

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left(W_i(t) - \frac{S_1(t,\beta)}{S_0(t,\beta)} \right) dN_i(t)$$

be the score function of the partial likelihood function and let β_0 *be the true coefficient vector. Then* $U(\beta_0)$ *is a* $\mathcal{F}_t$*-martingale.*

Proof sketch. The Doob-Meyer decomposition (Corollary 3.2.5) is applicable for decomposing N_i, since each $N_i(t)$ is a counting process adapted to the filtration $\mathcal{F}_t$. Using Eq. (3.9) the score function can be rewritten as

$$U(\beta_0) = \sum_{i=1}^{n} \int_0^{\tau} \left(W_i(t) - \frac{S_1(t,\beta_0)}{S_0(t,\beta_0)} \right) d\Lambda_i(t|W_i(t)) + \sum_{i=1}^{n} \int_0^{\tau} \left(W_i(t) - \frac{S_1(t,\beta_0)}{S_0(t,\beta_0)} \right) dM_i(t|W_i(t)) \tag{4.4}$$

where $\Lambda_i(t \mid W_i(t))$ is the conditional cumulative $\mathcal{F}_t$-intensity process of $N_i(t)$ and $M_i(t \mid W_i(t))$ is the associated $\mathcal{F}_t$ counting process martingale. Recalling the intensity of the Cox model in Eq. (4.1) the first term of Eq. (4.4) can be assessed:

$$\begin{aligned}
&\sum_{i=1}^{n} \int_0^{\tau} \left(W_i(t) - \frac{S_1(t,\beta_0)}{S_0(t,\beta_0)} \right) Y_i(t) \exp(W_i(t)\beta_0) dH_0(t) \\
&= \int_0^{\tau} \left(\sum_{i=1}^{n} W_i(t) Y_i(t) \exp(W_i(t)\beta_0) - \frac{S_1(t,\beta_0)}{S_0(t,\beta_0)} \sum_{i=1}^{n} Y_i(t) \exp(W_i(t)\beta_0) \right) dH_0(t) \\
&= \int_0^{\tau} \left(S_1(t,\beta_0) - \frac{S_1(t,\beta_0)}{S_0(t,\beta_0)} S_0(t,\beta_0) \right) dH_0(t) = 0
\end{aligned}$$

Eq. (4.4) simplifies to

$$U(\beta_0) = \sum_{i=1}^{n} \int_0^{\tau} \left(W_i(t) - \frac{S_1(t,\beta_0)}{S_0(t,\beta_0)} \right) dM_i(t \mid W_i(t))$$

Since $W_i(t)$ and $Y_i(t)$ are bounded and predictable with respect to $\mathcal{F}_t$ the integrand is bounded and predictable as long as there are individuals at-risk for $t \in [0, \tau]$. Each $M_i(t \mid W_i(t))$ is a counting process martingale and so by Theorem 3.2.9, $U(\beta_0)$ is a $\mathcal{F}_t$-martingale. □

Lemma 4.1.1 is a key reason for the use of the Cox model, since martingale convergence theorems can be employed to analyze the asymptotic properties of the β-estimates. In particular, it can be shown that under regularity conditions

$$\frac{1}{\sqrt{n}}U(\beta_0) \overset{\mathcal{D}}{\to} N(0,\Sigma), \quad \sqrt{n}(\hat{\beta}-\beta_0) \overset{\mathcal{D}}{\to} N(0,\Sigma^{-1}), \quad \frac{1}{n}I(\hat{\beta}) \overset{\mathcal{P}}{\to} \Sigma$$

Under similar regularity conditions it can be established that the Breslow estimator converges in distribution to a Gaussian process with mean zero.

We refer to Fleming and Harrington (1991) and Martinussen and Scheike (2006) for details on regularity conditions and proofs of the convergence results. The convergence properties show that $\hat{\beta}$ is an unbiased estimator of β_0 and that the estimator asymptotically achieves the Cramér-Rao lower bound for variance, which means $\hat{\beta}$ is an efficient estimator of β_0. In addition, the convergence results enable the construction of confidence intervals for the point estimates. However, all results presented in this section require the assumptions of the Cox model to hold, which is not always the case.

4.2 The misspecified Cox model

We are never guaranteed that the true hazard rate, $\alpha(t)$, is well described by the proportional hazards assumption from Eq. (4.1). When there is a discrepancy between the true hazard rate and the one assumed by the model, we say that the Cox model is *misspecified*. In this subsection we elaborate on some of the ideas presented in Example 6.1.3 in Martinussen and Scheike (2006) and in the paper by Whitney, Shojaie, and Carone (2019). The objective is to show that the parameter estimates, $\hat{\beta}$, may depend on the distribution of the censoring times, C, if the model is misspecified, which is a problem, as the censoring distribution is unrelated to the parameters of interest. We begin with a theoretical argument of the dependency on the censoring distribution and then show a simulation study that supports the theoretical claim.

4.2.1 Theoretical dependency of estimates on censoring distribution

Let $(N_1(t), W_1, Y_1(t)), ..., (N_n(t), W_n, Y_n(t))$ be n i.i.d triples of processes obtained from right-censored survival times. The covariates, $W_i(t) = W_i$ are for notational simplicity assumed constant. The intensity processes of each counting processes with respect to the current information $\mathcal{F}_t$ is given by

$$\lambda(t \mid W_i) = Y_i(t)\alpha(t \mid W_i), \quad i = 1, \dots, n$$

where $Y_i(t)$ is an at-risk indicator and $\alpha(t \mid W_i)$ is the hazard rate given covariates W_i. Under standard regularity assumptions it can be shown that Cox's partial maximum likelihood estimator $\hat{\beta}$ is a consistent estimator of β^*, where β^* is the solution to the equation

$$h(\beta) = \int_0^\tau \left(\frac{s_1(t)}{s_0(t)} - \frac{s_1(t,\beta)}{s_0(t,\beta)}\right) s_0(t)dt = 0$$

and

$$s_j(t) = E[S_j(t)] = E\left[\frac{1}{n}\sum_{i=1}^n W_i^j Y_i(t)\alpha(t \mid W_i)\right]$$

$$s_j(t,\beta) = E[S_j(t,\beta)] = E\left[\frac{1}{n}\sum_{i=1}^n W_i^j Y_i(t)\alpha_0(t)\exp(W_i\beta)\right]$$

for $j = 0, 1$. The function $s_j(t)$ depends on the true data generating process, while $s_j(t, \beta)$ reflects the Cox model's assumption about the data generating process. As a result, $h(\beta)$ quantifies how similar the Cox model is to the true model. If the true model is well described by the Cox model such that

$$s_j(t) = E\left[\frac{1}{n}\sum_{i=1}^{n} W_i^j Y_i(t)\alpha_0(t)\exp(W_i\beta_0)\right]$$

for some true parameter $\beta_0 \in \mathbb{R}^{p+1}$, then β_0 will be the solution to $h(\beta) = 0$. This means $\beta^* = \beta_0$ and consequently $\hat{\beta}$ will converge to the true parameter, β_0. However, if the true conditional hazard rate, $\alpha(t \mid W_i)$, is not of the type given in Eq. (4.1), it can be an arbitrary hazard rate and β^* will depend on the distribution of the censoring times due to the s_0 term because

$$\begin{aligned}
s_0(t) = E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i(t)\alpha(t \mid W_i)\right] &\overset{\dagger}{=} \frac{1}{n}\sum_{i=1}^{n} E[E\{Y_i(t)\alpha(t \mid W_i) \mid W_i\}] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[\alpha(t \mid W_i)E\{I(t \leq T^* \wedge C) \mid W_i\}] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[\alpha(t \mid W_i)E\{I(t \leq T^*, t \leq C) \mid W_i\}] \\
&\overset{\star}{=} \frac{1}{n}\sum_{i=1}^{n} E[\alpha(t \mid W_i)E\{I(t \leq T^*) \mid W_i\}E\{I(t \leq C) \mid W_i\}] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[\alpha(t \mid W_i)\mathbb{P}(T^* > t \mid W_i)\mathbb{P}(C > t \mid W_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[\alpha(t \mid W_i)S(t \mid W_i)K_C(t \mid W_i)].
\end{aligned}$$

We use expectation via conditioning at †, equation $\star$ follows because we assume independent censoring given W_i, and $K_C(t \mid W_i) = \mathbb{P}(C > t \mid W_i)$ is the survival function for the censoring times, C.

4.2.2 Simulation study

To emphasize the result in the previous section, we illustrate a simple case of model misspecification. Let $A = (A_{1,0}, \ldots, A_{n_0,0}, A_{1,1}, \ldots, A_{n_1,1})$ be a vector of covariates with $A_{i,0} = 0$ for $i = 1, \ldots, n_0$ and $A_{i,1} = 1$ for $i = 1, \ldots n_1$. The covariate vector, A, can be interpreted as a treatment covariate where n_0 individuals receive treatment 0, and n_1 individuals receive treatment 1. We assume known treatment-specific hazard rates given by:

$$\alpha_{A_{i,j}}(t) = \begin{cases} \alpha_0(t)\exp(\beta \cdot 0) = \alpha_0(t) & \text{if } j = 0 \\ \theta t^{\theta - 1} & \text{if } j = 1 \text{ for a } \theta \geq 1 \end{cases} \quad \text{for } j \in \{0, 1\}, i \in \{1, \ldots, n_j\}$$

For simplicity we assume that the baseline hazard rate, $\alpha_0(t)$, is equal to 1, and the proportional hazards are thus given by

$$\frac{\alpha_{A_{i,0}}(t)}{\alpha_{A_{i,1}}(t)} = \frac{1}{\theta t^{\theta - 1}}$$

which is constant for $\theta = 1$. When $\theta = 1$, the hazard rate $\alpha_{A_{i,j}}(t)$ is of the type in Eq. (4.1) and the Cox-model assumptions hold. For all $\theta > 1$ the assumption of constant proportional hazards is violated, implying that any Cox model will be misspecified. With known hazard rates we can generate true survival times, $T^*_{i,j}$, for every individual using the quantile transformation method, with

$$F(t) = 1 - S(t) = 1 - \exp(-\int_0^t \alpha(s)ds)$$

The distribution of the true survival times is $T^*_{i,j} \sim Exp(\alpha_{A_{i,j}})$, and we can create $n = n_0 + n_1$ tuples of the form $(A_{i,j}, T^*_{i,j})$. Let the censoring times be distributed as

$$C_{i,j} \sim Exp(\gamma_j)$$

for $\gamma_0, \gamma_1 \in \{0.2, 1.1, 2\}$. Define $T_{i,j} = T^*_{i,j} \wedge C_{i,j}$ and $\Delta_{i,j} = I(T^*_{i,j} < C_{i,j})$, such that we have $n = n_0 + n_1$ data triples $(T_{i,j}, A_{i,j}, \Delta_{i,j})$. To demonstrate how varying θ-values affect estimation of β-values, we run a Cox-regression on the data based on 100 different θ-values in the interval $(1, 2.5)$ and with the 9 different combinations of γ_0 and γ_1. Every simulation consists of $n = 100000$ individuals, where $n_0 = 50000$ and $n_1 = 50000$. A graph containing the point estimate for each simulation across γ_0-, γ_1- and θ-values is displayed in Figure 2. The code for all figures can be found on GitHub [4]

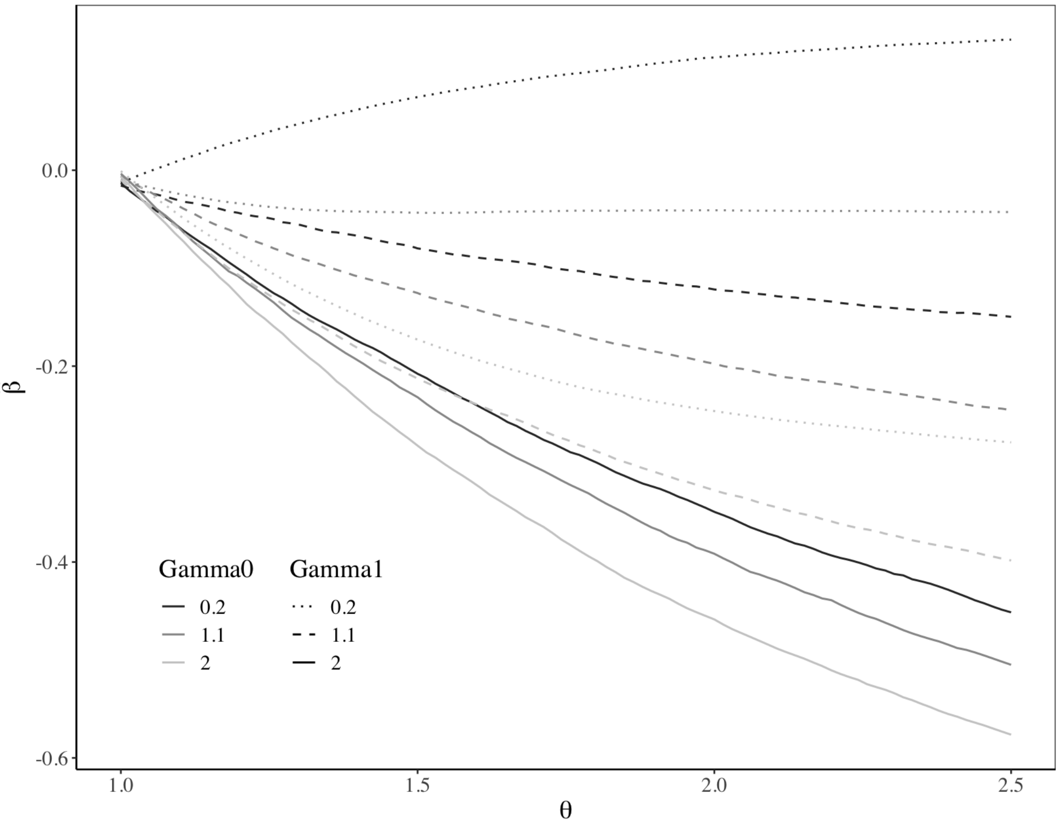

Figure 2: Simulation study of how censoring distribution affects the point estimate, when the Cox model is misspecified. $n = 100000$, $\theta \in [1, 2.5]$

[4] https://github.com/otto-groen-roepstorff/bach_r/tree/master/material_project

When $\theta = 1$, the hazard rates $\alpha_{A_{i,0}}$ and $\alpha_{A_{i,1}}$ are identical, and the Cox-regression correctly estimates that $\hat{\beta} = 0$ for all the combinations of censoring times. This is expected since the true data generating model has been correctly specified when $\theta = 1$. As θ, and thereby the time dependence of the hazard rate $\alpha_{A_{i,1}}$, increases, the estimates of β increasingly depend on the distribution of the censoring times. This happens even when the censoring is independent of covariate A. Otherwise, we would expect that the dotted black line, the dashed dark-grey line and the solid light-grey line would provide exactly the same estimates for β. This illustrates the problematic nature of a misspecified Cox model.

4.3 Cataract surgery data - the Cox model applied

In this subsection we use the Cox proportional hazards model to analyze a data set on cataract surgeries. First, we introduce the data set and the research question, and then we show how to conduct data analysis with the Cox regression. Afterwards we naïvely state the results before we investigate the validity of the model assumptions and discuss the findings.

The data set originates from Rigshospitalets department of eye diseases and was collected between 2010 and 2021 in Copenhagen and Aarhus. The data set consists of 388 observations of children who underwent surgery to prevent cataract - an eye disease leading to low vision or blindness. An observation for each individual eye was recorded if a child was operated in both eyes. After operation the children faced an elevated risk of developing the eye disease glaucoma due to high pressure in the eye, and to reduce the risk of glaucoma the children were prescribed eye drops called steroid responders. For some children the eye drops increased the pressure in the eye, and the eye drops therefore augmented the risk of developing glaucoma. In Copenhagen, a high dosis of the steroid responders was used between 2010 and 2017. Afterwards, between 2017 and 2021, the department in Copenhagen decided to lower the dosis of steroid responders to match the dosis used in Aarhus in the period 2010 to 2021. The research question is: Which factors impact the risk of developing glaucoma after surgery.

The data set consists of 18 covariates including a censoring time variable, `time_until_exam`, that indicates the number of days between surgery and the final examination, and a survival time variable, `time_until_glau`, that indicates the number of days between surgery and the development of glaucoma. We construct the *observed survival time* variable, $T = T^* \wedge C$ and label it `time`. In addition we construct the `status` indicator, $\Delta = I(T^* \leq C)$, that indicates whether the individual developed glaucoma or not. The specialists involved in the study identified the following covariates as the most important factors for developing glaucoma:

- `ster_regi`: Low (0) or high (1) dosis
- `age_at_surg`: Age at surgery in days
- `iol_single`: Artifical lens inserted (1) or not (0)
- `axis_lenght`: Length of the eye axis
- `num_re_op`: Number of reoperations

The proportional hazards model was fitted on these covariates using the function `coxph()` found in the `survival` package in `R`:

```
fit <- coxph(Surv(time,status) ~ ster_regi + iol_single
+ age_at_surg + axis_lenght + num_re_op, data=df)
```

variable	coef	exp(coef)	se(coef)	z	p
ster_regi	-0.209337	0.811121	0.388351	-0.539	0.5899
iol_single	-0.181001	0.834434	0.652909	-0.277	0.7816
age_at_surg	-0.002776	0.997228	0.001200	-2.314	0.0207
axis_lenght	0.039848	1.040652	0.111749	0.357	0.7214
num_re_op	0.976129	2.654162	0.203957	4.786	1.7e-06

Table 1: Summary of coefficients for the covariates of interest, based on Cox regression

Potential ties in survival times are handled by adding a tiny amount of noise sampled from a uniform distribution[5].

The coefficients estimated by the Cox model are displayed in Table 1. According to the model both ster_regi and iol_single have an insignificant effect on the survival times. In fact only age_at_surg and num_re_op are estimated to have a significant effect on the time from surgery until development of glaucoma. The estimate of the instantaneous relative risk of a one day increase in age at surgery is 0.997, and this suggests it is beneficial to operate children when they are older rather than younger. The number of reoperations is estimated to have a negative impact on survival times, with an estimated relative instantaneous risk of 2.654. The naïve conclusion of the results would be to 1) give surgery to children as late as possible, 2) not do any reoperations. However, it is not certain that the point estimates of the relative instantaneous risk have a meaningful interpretation.

An examination of the Kaplan-Meier[6] survival curves for the two distinct levels of steroid responders in Figure 3 reveal that the constant proportional hazards assumption may be violated. This is evident from the crossing survival curves for the covariate ster_regi. The

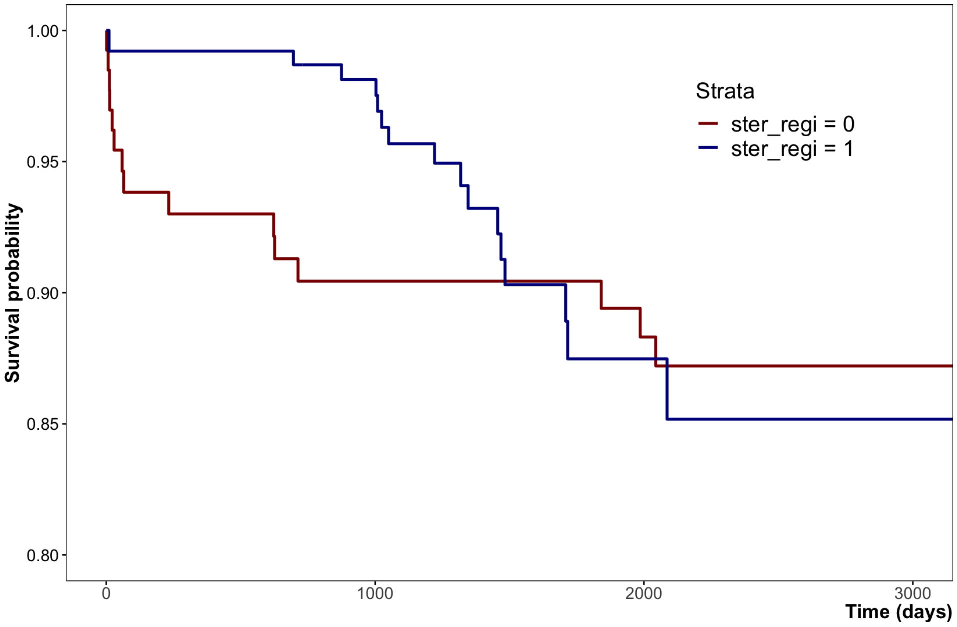

Figure 3: *Kaplan-Meier curves for* ster_regi = 0 *and* ster_regi = 1.

[5] This random error introduced in the survival times is considerably smaller than the amount of error introduced by truncating the survival times into integer values.

[6] The Kaplan-Meier estimator is commonly used for estimating survival curves (see Martinussen and Scheike (2006) for details).

proportional hazards assumption is that the relative instantaneous risk is constant. When the survival curves cross there are some indications that the hazard ratio

$$\frac{\alpha(t \mid X, A = 1)}{\alpha(t \mid X, A = 0)} = \exp(\beta_A)$$

varies with time. The violation of the proportional hazards assumption is further emphasized by simulating cumulative martingale residuals using asymptotic results for the Cox model, when assuming the null-hypothesis of constant proportional hazards holds. The cumulative martingale residuals for each covariate is displayed in Figure 4. For `ster_regi`, `iol_single` and `axis_length_center` the assumption of constant proportional hazards seems to be violated, as the observed cumulative martingale residuals differ notable from the null.

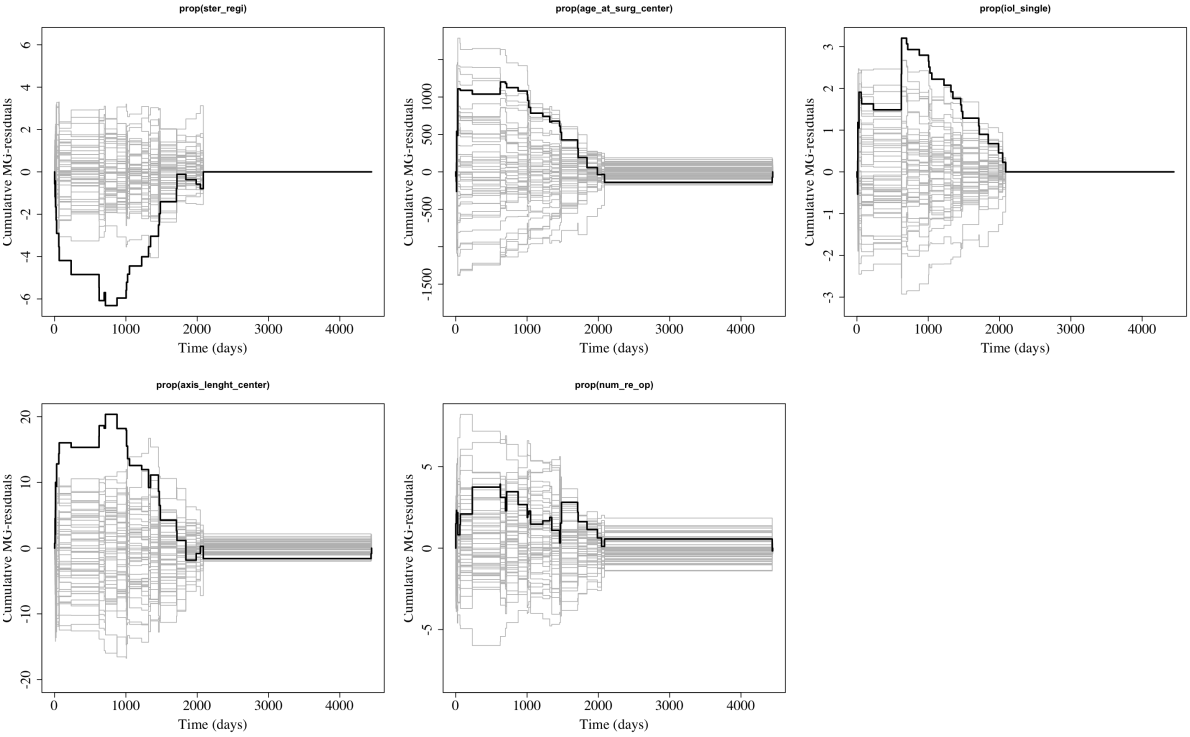

Figure 4: *Cumulative martingale residuals under the assumption of constant proportional hazards assumption against the observed residuals (thick black line).*

The model assessment indicates that the proportional hazards assumption does not hold, and the estimates in Table 1 are therefore subject to the issues presented in Subsection 4.2. They should therefore not be uncritically accepted as facts about the true data generating process of survival times. It could be beneficial to conduct a more thorough exploratory data analysis to find alternative regression models that suit the data better than the naïve approach to fitting a proportional hazards model. [7]

4.4 Discussion of the Cox model

The Cox model is a powerful tool in survival analysis, as it, under the right circumstances, summarizes survival data better than other regression models due to its efficient estimation of

[7] There are many different ways of fitting models including transformation of covariates, adding other covariates, assuming additive hazard or allowing time-dependent covariates

β. In addition, the results obtained from employing a Cox regression model are easy to interpret, when the model is correctly specified. However, there are many problems if the model is misspecified. The $\hat{\beta}$ estimates of a misspecified Cox model will, under regularity conditions, converge to some β^*, but this β^* is a meaningless estimate as it has nothing to do with the log-relative instantaneous risk. Worse, it can be shown, that β^* is prone to depend on the distribution of the censoring times and therefore it has no clear interpretation (Subsection 4.2). This showcases that unreflected use of Cox regression is problematic, since the (possibly wrong) conclusions of the regression model can have great impact on human lives, like in the case study of cataract surgeries (Subsection 4.3).

Ideally, we want estimators of survival data that can compete with the efficiency of the Cox model and is less prone to misspecification errors. This leads us to nonparametric estimation, efficient influence functions and one-step estimators.

5 Nonparametric theory

In the previous section we presented Cox's proportional hazards model - a semiparametric model for analyzing survival data. The model has several convenient properties that makes Cox regression one of the most used tools in survival analysis. However, a misspecified Cox model exhibits some unpleasant characteristic, since the crucial proportional hazards assumption does not always - as demonstrated in Subsection 4.3 - we will often require alternative statistical tools.

In the remainder of this thesis, we focus on estimating summarizing descriptions of a true data generating process, $\mathbb{P}$. The challenge is to identify estimands of interest and construct suitable estimators with desirable properties. The idea is to use nonparametric estimators to estimate only the components of interest of the true distribution, rather than estimating the entire distribution.

The structure of this section is as follows. First, we introduce fundamental concepts of nonparametric estimation and state important smoothness conditions. Then, we introduce influence functions, strategies to derive them, and how influence functions can be used to construct efficient estimators. Last, we connect the nonparametric estimation framework to the survival data framework.

The section is mainly based on Fisher and Kennedy (2019), Kennedy (2022) and Hines et al. (2022). Throughout the section we give examples and illustrations to provide intuition and clarity on the presented material and methods.

5.1 Notation and computational rules

We write $dP(z) = p(z)d\mu(z) = p(z)dz$ whenever P has a density p with respect to some measure μ. For notational convenience, we will suppress the measure μ and for instance write dz instead of $d\mu(z)$. We use the notation $\int HdP - \int Hd\hat{P} = \int Hd(P - \hat{P})$, and we interchange integration and derivative without formally checking conditions from theorem 12.5 in Schilling (2017). Integral signs should be seen a summation signs when "integrating" a countable number terms. $X_n = o_{\mathbb{P}}(r_n)$ means that $X_n/r_n \xrightarrow{p} 0$, and $\mathbb{P}_n$ is the empirical distribution such that $\mathbb{P}_n(f(Z)) = \int f(Z)d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} f(Z_i)$.

In some of the examples we derive influence functions for estimands that answer causal questions, such as "How would an individual have had it, if they had received treatment". Formally this requires a discussion of the identifying assumptions of positivity, exchangeability, non-interference, and consistency of the observed data, and for an introduction to these concepts, the reader is referred to Hernan and Robins (2020). The causal interpretation of our estimands is not the focus of this thesis, so we will assume that the identifying assumptions always hold, when we work with data.

5.2 Functionals and estimands

The initial task is to determine the scientific question of interest, which will usually be raised before data collection takes place. Similarly, the appropriate quantitative target used to answer the question of interest should be determined before the data collection. This approach is more honest compared to fitting a model in the sense that we do not seek to validate a chosen model after the data collection has taken place. We refer to the target of interest as an *estimand*

Definition 5.2.1 (Estimand)**.** An *estimand*, θ, for a random variable $Z \sim \mathbb{P}$ is a summary of the distribution $\mathbb{P}$ constructed to answer a scientific question of interest.

A key task is to identify the estimand of interest. We refrain from going into details on how to identify estimands and focus on the task of estimating estimands. Estimands depend on the underlying distribution they summarize, and they can be expressed as functions of distributions, which leads to the concept of *functionals*

Definition 5.2.2 (Functional)**.** Let $\mathcal{P}$ be a set of distributions. A *functional* is a function

$$\psi : \mathcal{P} \to \mathbb{R}^q \text{ for } q \in \mathbb{N}$$

that maps distributions, $P \in \mathcal{P}$ to a number in $\mathbb{R}^q$.

For simplicity we assume that $q = 1$. The set of distributions $\mathcal{P}$ is usually determined by the assumptions on the data generating process of Z i.e. the assumed model.

Remark 5.2.3 (Functional, estimand and estimator)**.** The concepts "functional" and "estimand" show up as two very similar concepts (Kennedy, 2022; Fisher and Kennedy, 2019; Hines et al., 2022). A *functional*, ψ, is the abstract concept of map from a set of distributions to the reals, and it can take many different shapes, while an estimand for the distribution $\mathbb{P}$ is *a* functional evaluated in $\mathbb{P}$.

$$\theta = \psi(\mathbb{P})$$

Estimands are functionals that are well defined without reference to a (semi)parametric model. The purpose of a functional is to provide tools for summarizing any data distribution, while the purpose of an estimand is to formalize the answer to a scientific question about the data.

The concrete way to determine a value of an estimand is called an *estimator*. For instance, if we estimate $\mathbb{P}$ by some $\hat{\mathbb{P}}$ (i.e. collected by sampling), we can define the *plug-in estimator* for θ as:

$$\hat{\theta}_{PI} = \psi(\hat{\mathbb{P}}) = \hat{\psi}_{PI} \tag{5.1}$$

The relationship between functionals and esimands is shown in the following example:

Example 5.2.4. Let $Z \sim \mathbb{P}$, and let our target of interest be the mean value of Z. We define the *functional*, ψ, for our target as

$$\psi : \mathcal{P} \to \mathbb{R}, \quad Q \mapsto \int z dQ(z) = E_Q(Z),$$

and our *estimand* as

$$\theta = \psi(\mathbb{P}) = E_{\mathbb{P}}(Z)$$

Different estimates of $\mathbb{P}$, yield different estimates of our functional, $\psi(\hat{\mathbb{P}})$, and like any type of estimation it is essential to be able to determine the variance of the estimates. This leads to the theory of influence functions.

5.3 Influence functions

Influence functions are crucial in non- and semiparametric statistics. Influence functions provide several key insights on the behaviour of nonparametric estimators such as how a single observation influences estimators, and they quantify the effect of small changes in the distribution of our observations on the probability limit of estimators (Ichimura and Newey, 2022). Influence functions were originally introduced as a description of estimator stability (Fisher and Kennedy, 2019), but they have proven useful in the construction of estimators. To develop the theory of influence functions we begin with the notion of *parametric submodels*

Definition 5.3.1 (Parametric submodel (Kennedy, 2022))**.** A *parametric submodel* is a smooth parametric model $\mathcal{P}_\epsilon = \{P_\epsilon : \epsilon \in \mathbb{R}\}$ that satisfies

(a) $\mathcal{P}_\epsilon \subseteq \mathcal{P}$

(b) $P_{\epsilon=0} = \mathbb{P}$

The first condition implies that $\mathcal{P}_\epsilon$ has to be a subset of all the distributions of interest such that the functionals $\psi(P_\epsilon)$ are well defined for all $P_\epsilon \in \mathcal{P}_\epsilon$. The second condition ensures that the true distribution, $\mathbb{P}$, is included in $\mathcal{P}_\epsilon$. Parametric submodels are not tools for data analysis as we in practice never know $\mathbb{P}$, but rather a technical device used to study the behaviour of the functional due to changes in the underlying distribution. We limit our attention to parametric submodels $\mathcal{P}_\epsilon = \{P_\epsilon\}_{\epsilon\in[0,1]}$ of the form

$$P_\epsilon = \mathbb{P} + \epsilon(\hat{\mathbb{P}} - \mathbb{P}) \tag{5.2}$$

where P_ϵ is assumed to have density $p_\epsilon(z) = p(z) + \epsilon(\hat{p}(z) - p(z))$ w.r.t. to some measure μ. Here, $p(z)$ is the density of $\mathbb{P}$ and $\hat{p}(z)$ is the density of some estimate, $\hat{\mathbb{P}}$, of $\mathbb{P}$. The results derived in this and the following sections are limited to parametric submodels of this type, although the results hold more generally (Kennedy, 2022).

We refer to the collection $\mathcal{P}_\epsilon = \{P_\epsilon\}_{\epsilon\in[0,1]}$ as a path. The functional value, $\psi(P_\epsilon)$, usually changes as ϵ changes along the path, and Example 5.3.2 shows how a path looks in a hypothetical setting where we have simple, known expressions for $\mathbb{P}$ and $\hat{\mathbb{P}}$.

Example 5.3.2. Let $Z \sim \mathbb{P} = \exp(1)$, and assume that we have estimated $\mathbb{P}$ by $\hat{\mathbb{P}} = \exp(3/2)$. The densities of elements in the collection $\{P_\epsilon\}_{\epsilon\in[0,1]}$ are visualized in panel (A) of Figure 5. Let the estimand of interest be the expected density, such that

$$\theta = \psi(\mathbb{P}) = E_{\mathbb{P}}(f_{\mathbb{P}}(X))$$

where $f_{\mathbb{P}}$ is the density of $\mathbb{P}$. Panel (B) of Figure 5 displays the values attained by the functional, $\psi(P_\epsilon)$ along this specific path for different ϵ values [8].

By looking at panel (B) in Figure 5 it seems like the 'naive' plug-in estimator $\psi(\hat{\mathbb{P}})$ of $\psi(\mathbb{P})$ can be improved by linearly approximating (dashed line) the functional values along the path (solid line). This can be done if the functional is sufficiently smooth. To formalize what we mean by sufficiently smooth we introduce the following definitions:

[8] For more details on calculations see appendix A.

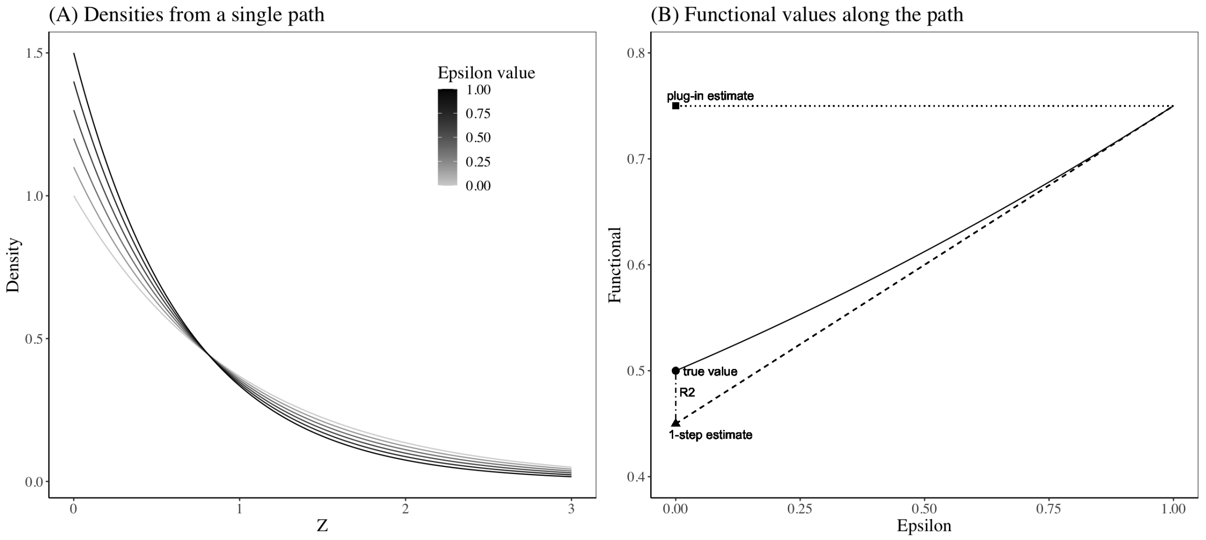

Figure 5: *Panel (A) displays some of the densities of a single path, where* $\mathbb{P} = \exp(1)$ *and* $\hat{\mathbb{P}} = \exp(3/2)$. *As* ϵ *moves from* 1 *to* 0 *the estimates move increasingly closer to the true density. The solid line in panel (B) is the functional values along the path, and the horizontal dotted line is the plug-in estimate. The dashed line is the first-order linear approximation of the solid line at* $\epsilon = 1$. *The slope is derived using the efficient influence function of the functional. "R2" is a visualisation of the second order error - see Subsection 5.5 for additional details.*

Definition 5.3.3. (Von Mises expansion) Let $\psi : \mathcal{P} \to \mathbb{R}$ be a functional, and let $P, \tilde{P} \in \mathcal{P}$ be two distributions. Then ψ is said to admit a von Mises expansion if there exists a function $\phi(Z; P)$ with $\int \phi(z; P)dP(z) = 0$ and $\int \phi(z; P)^2 dP(z) < \infty$ such that

$$\psi(\tilde{P}) - \psi(P) = \int \phi(z; \tilde{P})d(\tilde{P} - P)(z) + R_2(\tilde{P}, P) \tag{5.3}$$

where $R_2(\tilde{P}, P)$ is a second order remainder term, which means the remainder term only depend on products or squares of differences between $\tilde{P}$ and P.

Definition 5.3.4 (Influence function)**.** The function $\phi(Z; P)$ in Eq. (5.3) satisfying $E_P[\phi(Z; P)] = 0$ and $E_P[\phi(Z; P)^2] < \infty$ is the *influence function*. There are potentially many influence functions for semiparametric models. The one that is also a valid score (i.e. part of the submodel tangent space) is called the *efficient* influence function (EIF). In nonparametric models there is only one valid influence function which is also the EIF (Kennedy, 2022).

We only consider nonparametric models, and therefore every influence functions is the *efficient influence function* (EIF). The von Mises expansion can be considered as a distributional Taylor expansion where the influence function $\phi(Z; P)$ serves as the derivative term (Kennedy, 2022). A related smoothness property is pathwise differentiability. To motivate pathwise differentiability we define the Gâteaux derivative.

Definition 5.3.5 (Gâteaux derivative (Hines et al., 2022))**.** Let $\psi : \mathcal{P} \to \mathbb{R}$ be a functional and let $P_\epsilon = \mathbb{P} + \epsilon(\hat{\mathbb{P}} - \mathbb{P})$ be an element of the parametric submodel $\mathcal{P}_\epsilon \subseteq \mathcal{P}$. If the limit

$$\lim_{\epsilon \to 0^+} \frac{\psi(P_\epsilon) - \psi(\mathbb{P})}{\epsilon}$$

exists, it is called the *Gâteaux derivative* of ψ at $\mathbb{P}$ in the direction of $\hat{\mathbb{P}}$ and we denote it by

$$\frac{\partial}{\partial \epsilon}\psi(P_\epsilon)|_{\epsilon=0}$$

As noted in Hines et al. (2022) the Gâteaux derivative can be interpreted as a formalisation of the sensitivity of ψ to changes in the underlying distribution in the direction of $\hat{\mathbb{P}}$. This intuition is useful in understanding the smoothness condition of pathwise differentiability:

Definition 5.3.6. (Pathwise differentiability (Kennedy, 2022)) A functional $\psi : \mathcal{P} \to \mathbb{R}$ is said to be *pathwise differentiable* if there for every smooth submodel $\mathcal{P}_\epsilon \subseteq \mathcal{P}$ exists an influence function $\phi(Z; P)$ such that

$$\frac{\partial}{\partial \epsilon}\psi(P_\epsilon)|_{\epsilon=0} = \int \phi(z; \mathbb{P}) s_\epsilon(z) d\mathbb{P}(z) \tag{5.4}$$

where $s_\epsilon(z)$ is the score function of the submodel P_ϵ.

Definition 5.3.7 (Score function (Kennedy, 2022))**.** The score function $s_\epsilon(z)$ of a μ-dominated submodel $\mathcal{P}_\epsilon$ with associated family of densities $\mathcal{D} = \{p_\epsilon \mid \epsilon \in \mathbb{R}\}$ is a mean-zero function given by

$$s_\epsilon(z) = \frac{\partial}{\partial \epsilon}\log p_\epsilon(z)|_{\epsilon=0}$$

The smoothness conditions are give information about functionals, and the following proposition shows how the von Mises smoothness condition relates to pathwise differentiability:

Proposition 5.3.8. *A functional satisfying the von Mises expansion is pathwise differentiable under regularity conditions.*

Proof. Assume that $\psi : \mathcal{P} \to \mathbb{R}$ is smooth in the sense that it admits a von Mises expansion. Let $P_\epsilon \in \mathcal{P}_\epsilon \subseteq \mathcal{P}$ be given as in Eq. (5.2). Then by definition

$$\psi(\mathbb{P}) - \psi(P_\epsilon) = \int \phi(z; \mathbb{P}) d(\mathbb{P} - P_\epsilon)(z) + R_2(\mathbb{P}, P_\epsilon) \tag{5.5}$$

Taking the derivative of the left hand side of Eq. (5.5) with respect to ϵ yields

$$\frac{\partial}{\partial \epsilon}\left\{\psi(\mathbb{P}) - \psi(P_\epsilon)\right\}|_{\epsilon=0} = -\frac{\partial}{\partial \epsilon}\psi(P_\epsilon)|_{\epsilon=0}. \tag{5.6}$$

By noting that the score of the parametric submodel $\mathcal{P}_\epsilon$ is

$$\begin{aligned} s_\epsilon(z) &= \frac{\partial}{\partial \epsilon}\log(p_\epsilon(z))|_{\epsilon=0} = \frac{\partial}{\partial \epsilon}\log(p(z) + \epsilon(\hat{p}(z) - p(z)))|_{\epsilon=0} \\ &= \frac{1}{p(z) + \epsilon(\hat{p}(z) - p(z))}(\hat{p}(z) - p(z))|_{\epsilon=0} = \frac{\hat{p}(z) - p(z)}{p(z)}, \end{aligned}$$

then the derivative of the right-hand side of Eq. (5.5) with respect to ϵ is

$$
\begin{aligned}
&\frac{\partial}{\partial\epsilon}\left\{\int \phi(z;\mathbb{P})d(\mathbb{P}-P_\epsilon)(z)+R_2(\mathbb{P},P_\epsilon)\right\}|_{\epsilon=0}\\
=&-\frac{\partial}{\partial\epsilon}\left\{\int \phi(z;\mathbb{P})dP_\epsilon(z)\right\}|_{\epsilon=0}+\frac{\partial}{\partial\epsilon}R_2(\mathbb{P},P_\epsilon)|_{\epsilon=0}\\
=&-\int \phi(z;\mathbb{P})\frac{\partial}{\partial\epsilon}p_\epsilon(z)|_{\epsilon=0}dz \qquad (5.7)\\
=&-\int \phi(z;\mathbb{P})\frac{\partial}{\partial\epsilon}((1-\epsilon)p(z)+\epsilon\hat{p}(z))\frac{1}{p(z)}|_{\epsilon=0}d\mathbb{P}(z)\\
=&-\int \phi(z;\mathbb{P})\frac{\hat{p}(z)-p(z)}{p(z)}d\mathbb{P}(z)=-\int \phi(z;\mathbb{P})s_\epsilon(z)d\mathbb{P}(z)
\end{aligned}
$$

using that $\phi(Z;\mathbb{P})$ has mean zero and $\frac{\partial}{\partial\epsilon}R_2(\mathbb{P},P_\epsilon)|_{\epsilon=0}=0$ as it is a second order remainder term. Putting Eqs. (5.6) and (5.7) together we get that

$$\frac{\partial}{\partial\epsilon}\psi(P_\epsilon)|_{\epsilon=0}=\int \phi(z;\mathbb{P})s_\epsilon(z)d\mathbb{P}(z)$$

The above equalities are only valid as long as we can interchange integral and derivative. □

The smoothness conditions stated above are inherently connected to EIFs. In particular, smooth nonparametric functionals possess an EIF. Influence functions play a pivotal role in semi- and nonparametric statistics and come with several valuable properties. To illustrate a few, we return to the issue in Example 5.3.2 that initiated the discussion of smoothness conditions.

Remark 5.3.9 (Slope of functional along a path)**.** Let $\psi:\mathcal{P}\to\mathbb{R}$ be a smooth functional in the sense that it is pathwise differentiable with EIF $\phi(Z;P)$. Using Riez's representation theorem[9], it can be shown (Hines et al., 2022) that

$$\frac{\partial}{\partial\epsilon}\psi(P_\epsilon)|_{\epsilon=1}=-\int \phi(z;\hat{\mathbb{P}})d\mathbb{P}(z) \qquad (5.8)$$

and thus the influence function can be used to estimate distributional derivates corresponding to the slope of the dashed line in panel (B) in Figure 5. We will return to this remark in Subsection 5.5, where we address *one-step* estimators.

EIFs can also be used to determine efficiency bounds for the variance of nonparametric estimators similar to the Cramér-Rao bound for parametric estimators.

Proposition 5.3.10 (Nonparametric efficiency bound (Kennedy, 2022))**.** *Let $\psi:\mathcal{P}\to\mathbb{R}$ be a pathwise differentiable functional with efficient influence function $\phi(Z;\mathbb{P})$. The lowest attainable asymptotic variance for any unbiased nonparametric estimator, $\hat{\psi}$, of ψ is*

$$\mathrm{var}\{\phi(Z;\mathbb{P})\}$$

Proof sketch. For the parametric submodel $\mathcal{P}_\epsilon=\{P_\epsilon\mid\epsilon\in\mathbb{R}\}$ and smooth functional ψ the variance of any unbiased estimator, $\hat{\psi}$, of ψ satisfies

$$\mathrm{var}_{P_\epsilon}(\hat{\psi})\geq\frac{(\frac{\partial}{\partial\epsilon}\psi(P_\epsilon)|_{\epsilon=0})^2}{\mathrm{var}_{P_\epsilon}(s_\epsilon(Z))}$$

[9] See Hines et al. (2022) for details.

where $s_\epsilon(Z)$ is the model score function (Kennedy, 2022). This is the Cramér-Rao lower bound. Since ψ is pathwise differentiable, we have that

$$\left(\frac{\partial}{\partial\epsilon}\psi(P_\epsilon)|_{\epsilon=0}\right)^2 = \left(\int \phi(z;\mathbb{P})s_\epsilon(z)d\mathbb{P}(z)\right)^2 = E_{\mathbb{P}}\left\{\phi(Z;\mathbb{P})s_\epsilon(Z)\right\}^2$$

Hence, over all Cramér-Rao bounds at $\epsilon = 0$ we have

$$\sup_{P_\epsilon\in\mathcal{P}_\epsilon}\frac{(\frac{\partial}{\partial\epsilon}\psi(P_\epsilon)|_{\epsilon=0})^2}{\operatorname{var}(s_\epsilon(Z))} = \sup_{P_\epsilon\in\mathcal{P}_\epsilon}\frac{E_{\mathbb{P}}(\phi(Z;\mathbb{P})s_\epsilon(Z))^2}{E_{\mathbb{P}}(s_\epsilon(Z)^2)} \leq E_{\mathbb{P}}(\phi(Z;\mathbb{P})^2) = \operatorname{var}\{\phi(Z;\mathbb{P})\}$$

where the inequality follows from the Cauchy Schwarz inequality. Since $\phi(Z;\mathbb{P})$ is the efficient influence function it is also a valid score, and the inequality is in fact an equality, which concludes the proof. □

EIFs are useful for determining the lowest possible variance of an estimator for an estimand, and in the next section we establish how influence functions can be derived in practice.

5.4 Deriving influence functions

We present the three approaches to deriving influence functions discussed in Kennedy (2022). Inspired by Kennedy (2022), we introduce the operator, $\mathbb{IF}$, that maps functionals, $\psi : \mathcal{P} \to \mathbb{R}$ to their efficient influence functions, ϕ.

5.4.1 Pathwise differentiability

The first approach is to mathematically manipulate the pathwise derivative $\frac{\partial}{\partial\epsilon}\psi(P_\epsilon)|_{\epsilon=0}$ to show that $\psi(P_\epsilon)$ satisfies the smoothness condition in Definition 5.3.6 and then identify the efficient influence function in the resulting integral.

Example 5.4.1 (Expectation of a random variable)**.** Let $Z \sim \mathbb{P}$ be a finite-variance random variable. Let $\mathcal{P}_\epsilon \subseteq \mathcal{P}$ be a parametric submodel such that $\mathcal{P}_\epsilon$ is smooth and $P_{\epsilon=0} = \mathbb{P}$ for $P_\epsilon \in \mathcal{P}_\epsilon$. Assume $\mathbb{P}$ and P_ϵ have densities p_0, p_ϵ, respectively, w.r.t. μ and let the functional of interest, ψ, be given by

$$\psi : \mathcal{P} \to \mathbb{R}, \quad Q \mapsto E_Q(Z)$$

From Definition 5.3.6 we get that $s_\epsilon(z) = \frac{\partial}{\partial\epsilon}\log p_\epsilon(z)|_{\epsilon=0}$, and the pathwise derivative of ψ is given by

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\psi(P_\epsilon)|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}E_{P_\epsilon}(Z)|_{\epsilon=0} = \frac{\partial}{\partial\epsilon}\int zdP_\epsilon(z)|_{\epsilon=0} \\
&= \frac{\partial}{\partial\epsilon}\int zp_\epsilon(z)dz|_{\epsilon=0} = \int z\frac{\partial}{\partial\epsilon}p_\epsilon(z)|_{\epsilon=0}dz \\
&\overset{\star}{=} \int z\left[\left\{\frac{\partial}{\partial\epsilon}\log p_\epsilon(z)\right\}p_\epsilon(z)\right]|_{\epsilon=0}dz \\
&= \int z\left\{\frac{\partial}{\partial\epsilon}\log p_\epsilon(z)|_{\epsilon=0}\right\}p_0(z)dz \\
&= \int zs_\epsilon(z)d\mathbb{P}(z),
\end{aligned}$$

where $\star$ holds since

$$\frac{\partial}{\partial \epsilon}\log p_\epsilon(z) = \frac{1}{p_\epsilon(z)}\frac{\partial}{\partial \epsilon}p_\epsilon(z).$$

From Eq. (5.4), we recognize $\tilde{\phi}(Z;\mathbb{P}) = Z$ as a potential EIF, and we just need to assure that the potential EIF has mean zero by subtracting the mean $E_{\mathbb{P}}(Z) = \psi(\mathbb{P})$:

$$\begin{aligned}\int zs_\epsilon(z)d\mathbb{P}(z) &\overset{(1)}{=} \int zs_\epsilon(z)d\mathbb{P}(z) - \psi(\mathbb{P})\int s_\epsilon(z)d\mathbb{P}(z)\\ &\overset{(2)}{=} \int \{z-\psi(\mathbb{P})\}s_\epsilon(z)d\mathbb{P}(z)\\ &= \int \phi(z;\mathbb{P})s_\epsilon(z)d\mathbb{P}(z)\end{aligned}$$

where (1) follows because the score has mean zero and (2) because $\psi(\mathbb{P})$ is a constant. This establishes pathwise differentiability of $\psi(\mathbb{P}) = E_{\mathbb{P}}(Z)$ with EIF $\phi(Z;\mathbb{P}) = Z - \psi(\mathbb{P})$. We notate this result as $\mathbb{IF}\{\psi(\mathbb{P})\} = Z - \psi(\mathbb{P})$.

The following example is a special case of Example 5.4.1, which lets us derive the EIF for indicator functions:

Example 5.4.2. Let $Z = I(X = x)$ for some discrete random variable X, and assume everything else is as in Example 5.4.1 with functional

$$\psi_x(\mathbb{P}) = E_{\mathbb{P}}(Z) = \mathbb{P}(X = x).$$

By applying the influence function derived in Example 5.4.1 it follows directly that $\mathbb{IF}\{\mathbb{P}(X = x)\} = I(X = x) - \mathbb{P}(X = x)$

This first approach to deriving EIFs is simple to understand, but it can be rather technical. Although the functional in Example 5.4.1 is simple, we had to solve an integral equation to derive the EIF, which in general may be complicated if not impossible. The advantage of this strategy, however, is that it is guaranteed to yield a valid EIF.

5.4.2 The point mass contamination strategy

The second approach is similar in spirit to the first, except that we assume the data are discrete. We briefly discuss this assumption in Example 5.4.4. In the *point mass contamination strategy*[10], we only look at the smartly chosen subset of submodels, $\mathcal{P}_\epsilon \subseteq \mathcal{P}$, where each distribution $P_\epsilon \in \mathcal{P}_\epsilon$ has probability mass function

$$p_\epsilon(z) = (1-\epsilon)p(z) + \epsilon\delta_{z'}(z), \quad \epsilon \in [0,1], \tag{5.9}$$

where $p_\epsilon(z) = P_\epsilon(Z = z)$, $p(z) = \mathbb{P}(Z = z)$, $\delta_{z'}(z) = I(z' = z)$ is the Dirac measure and z' is a single generic observation. The Gâteaux derivative with respect to this submodel corresponds to pertubating the functional in the direction of a point mass at a single generic observation z'. For this particular submodel the score function evaluates to

$$\begin{aligned}s_\epsilon(z) &= \frac{\partial}{\partial \epsilon}\log p_\epsilon(z)|_{\epsilon=0} = \frac{\partial}{\partial \epsilon}\log\{(1-\epsilon)p(z) + \epsilon\delta_{z'}(z)\}|_{\epsilon=0}\\ &= \frac{\delta_{z'}(z) - p(z)}{(1-\epsilon)p(z) + \epsilon\delta_{z'}(z)}|_{\epsilon=0} = \frac{\delta_{z'}(z) - p(z)}{p(z)}.\end{aligned} \tag{5.10}$$

[10] This is called the Gâteaux derivative approach in Kennedy (2022). We followed nomenclature from Hines et al. (2022) to keep consistent notation

Inserting the score function into Eq. (5.4) we get

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\psi(P_\epsilon)|_{\epsilon=0} &= \int \phi(z;\mathbb{P})s_\epsilon(z)d\mathbb{P} \\
&\overset{Eq.\ (5.10)}{=} \int \phi(z;\mathbb{P})\frac{\delta_{z'}(z)-p(z)}{p(z)}d\mathbb{P}(z) \\
&= \int \phi(z;\mathbb{P})\frac{\delta_{z'}(z)}{p(z)}d\mathbb{P}(z) - \int \phi(z;\mathbb{P})d\mathbb{P}(z) \\
&\overset{\star}{=} \sum_z \phi(z;\mathbb{P})\frac{\delta_{z'}(z)}{p(z)}p(z) = \sum_z \phi(z,\mathbb{P})\delta_{z'}(z) \\
&= \phi(z';\mathbb{P}) = \phi(Z;\mathbb{P}),
\end{aligned}$$

where $\star$ follows because we assume data are discrete and the EIF by definition has mean zero. The last equality follows because z' is any realization of the random variable Z. Hence, the Gâteaux derivative for this type of submodel returns the influence function directly, relieving us from integral equations in contrast to the previous approach.

Example 5.4.3 (Example 5.4.1 continued)**.** Recall that $Z \sim \mathbb{P}$ has finite variance and $\psi(\mathbb{P}) = E_{\mathbb{P}}(Z)$. By this approach we assume Z is discrete, and we let P_ϵ have probability mass function similar to Eq. (5.9). The EIF is

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\psi(P_\epsilon)|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon}\left\{\sum_z zP_\epsilon(Z=z)\right\}|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon}\left\{\sum_z z\left[(1-\epsilon)p(z)+\epsilon\delta_{z'}(z)\right]\right\}|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon}\left\{(1-\epsilon)\sum_z zp(z)+\epsilon\sum_z z\delta_{z'}(z)\right\}|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon}\left\{(1-\epsilon)\psi(\mathbb{P})+\epsilon z'\right\}|_{\epsilon=0} \\
&= z' - \psi(\mathbb{P}) = Z - \psi(\mathbb{P})
\end{aligned}$$

which corresponds to the result derived in Example 5.4.1.

The second approach provides a more direct method of deriving EIFs. Hines et al. (2022) notes that the method also works for finding candidates for EIFs in continuous or mixed cases by exchanging sums with integrals, indicator functions with the Dirac delta measure, and probability mass functions with probability functions. As shown in Example 5.4.4, this does not mean it is always possible to directly compute the EIF by this approach, when the data actually are continuous.

Example 5.4.4 (Discrete data versus continuous data)**.** This example is a continuous version of Example 5.4.2, and it serves as an illustration of why the assumption of discrete data calls for some caution. Let $Y \sim \mathbb{P}$ be a continuous random variable with density $f_{\mathbb{P}}$ w.r.t. the Lebesgue measure, and let

$$\psi(\mathbb{P}) = f_{\mathbb{P}}(y)$$

be the functional of interest. Let $\mathcal{P}_\epsilon$ be a parametric submodel similar to Eq. (5.9), such that each distribution $P_\epsilon \in \mathcal{P}_\epsilon$ has density w.r.t. the Lebesgue measure given by

$$f_{P_\epsilon}(y) = (1-\epsilon)f_{\mathbb{P}}(y) + \epsilon\delta_{y'}(y),$$

where $\delta_{y'}$ is the Dirac delta function. Then

$$\frac{\partial}{\partial \epsilon}\psi(dP_\epsilon) \mid_{\epsilon=0} = \delta_{y'}(y) - f_{\mathbb{P}}(y) = \tilde{\phi}(Z;\mathbb{P}).$$

The variance of the candidate EIF $\tilde{\phi}(Z;\mathbb{P})$ is infinite, because $\delta_{y'}(y)$ is unbounded as Y is continuous. This means $\tilde{\phi}(Z;\mathbb{P})$ violates the requirements of Definition 5.3.4 and is therefore not an EIF in contrast to the discrete case. However, as noted in Hines et al. (2022), the article by Ichimura and Newey (2022) shows that by substituting the Dirac delta function with a probability measure that converges to a point mass measure, the point mass contamination approach is still valid in the continuous case.

Example 5.4.4 illustrates that the point mass contamination strategy is not a foolproof approach to deriving EIFs. To make sure that the identified EIF is valid, we would need to check if the equality in Definition 5.3.6 holds. The strategy, however, has proved to identify valid EIFs and it is considered a reliable strategy (Kennedy, 2022). We conclude this approach with a more complicated example:

Example 5.4.5 (Regression function)**.** Consider a random variable $Z = (X, Y)$, where Y is a response variable and X the explanatory variables. The regression of Y on X is given by the estimand $\psi_x(\mathbb{P}) = E_{\mathbb{P}}(Y \mid X = x)$. Assuming the data are discrete, using the submodel from Eq. (5.9), and using Bayes rule of conditional probability, we get

$$P_\epsilon(Y = y \mid X = x) = \frac{P_\epsilon(Y = y, X = x)}{P_\epsilon(X = x)} = \frac{P_\epsilon(Z = z)}{P_\epsilon(X = x)} = \frac{(1-\epsilon)\mathbb{P}(Z = z) + \epsilon\delta_{z'}(z)}{(1-\epsilon)\mathbb{P}(X = x) + \epsilon\delta_{x'}(x)}, \tag{5.11}$$

assuming $P_\epsilon(X = x) > 0$. The EIF is

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\psi_x(P_\epsilon)|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon}\left[\sum_y y P_\epsilon(Y = y \mid X = x)\right] |_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon}\left[\sum_y y \frac{(1-\epsilon)\mathbb{P}(Z = z) + \epsilon\delta_{z'}(z)}{(1-\epsilon)\mathbb{P}(X = x) + \epsilon\delta_{x'}(x)}\right]\Bigg|_{\epsilon=0} \\
&= \sum_y y \frac{\partial}{\partial \epsilon}\left\{\frac{(1-\epsilon)\mathbb{P}(Z = z) + \epsilon\delta_{z'}(z)}{(1-\epsilon)\mathbb{P}(X = x) + \epsilon\delta_{x'}(x)}\right\}\Bigg|_{\epsilon=0} \\
&= \sum_y y \Bigg\{\frac{[(1-\epsilon)\mathbb{P}(X = x) + \epsilon I(x' = x)]\,[I(z' = z) - \mathbb{P}(Z = z)]}{[(1-\epsilon)\mathbb{P}(X = x) + \epsilon I(x' = x)]^2} \\
&\qquad - \frac{[(1-\epsilon)\mathbb{P}(Z = z) + \epsilon I(z' = z)]\,[I(x' = x) - \mathbb{P}(X = x)]}{[(1-\epsilon)\mathbb{P}(X = x) + \epsilon I(x' = x)]^2}\Bigg\}\Bigg|_{\epsilon=0} \\
&= \sum_y y \Bigg\{\frac{\mathbb{P}(X = x)\,[I(z' = z) - \mathbb{P}(Z = z)] - \mathbb{P}(Z = z)\,[I(x' = x) - \mathbb{P}(X = x)]}{\mathbb{P}(X = x)^2}\Bigg\} \\
&= \sum_y y \left\{\frac{\mathbb{P}(X = x)I(z' = z) - \mathbb{P}(Z = z)I(x' = x)}{\mathbb{P}(X = x)^2}\right\} \\
&= \sum_y y \left\{\frac{I(z' = z)}{\mathbb{P}(X = x)} - \frac{I(x' = x)}{\mathbb{P}(X = x)}\mathbb{P}(Y = y \mid X = x)\right\} \\
&= \frac{y' I(x' = x)}{\mathbb{P}(X = x)} - \frac{I(x' = x)}{\mathbb{P}(X = x)}E_{\mathbb{P}}(Y \mid X = x) = \frac{I(x' = x)}{\mathbb{P}(X = x)}\{y' - E_{\mathbb{P}}(Y \mid X = x)\}
\end{aligned}$$

where the quotient rule is used in the third equality. Thus the EIF for $\psi_x(\mathbb{P}) = E_{\mathbb{P}}(Y \mid X = x)$ is

$$\mathbb{IF}\{\psi_x(\mathbb{P})\} = \frac{I(X = x)}{\mathbb{P}(X = x)}\{Y - \psi_x(\mathbb{P})\}$$

This result holds in general as long as X is discrete and the variance var$(Y \mid X = x)$ is finite (Hines et al., 2022).

The calculations above show that the point mass contamination strategy can still lead to some tedious derivations, but it enables us to derive influence functions for complex functionals using simple calculus rules and basic properties of probability.

5.4.3 Building blocks and derivatives

The final approach introduced in Kennedy (2022) is somewhat similar to the point mass contamination approach, since it relies on derivative calculations to derive influence functions, but there are two key differences. First, we assume that we can treat the $\mathbb{IF}$-operator as a regular derivative and use derivative rules like the chain and product rule. Second, this approach uses already known influence functions as building blocks to skip unnecessary derivative calculations, and we can therefore think of it as a smart notational method for deriving influence functions. We summarize the approach with three tricks:

(T1) Trick 1: Pretend the data are discrete

(T2) Trick 2: Treat influence functions as derivatives

(T3) Trick 3: Use influence functions as building blocks

Just like the point mass contamination approach, there is no guarantee that this approach yields a valid EIF, and we would therefore technically have to check Definition 5.3.6 to ascertain we have found an EIF. In many cases, however, this approach yields valid EIFs. In previous examples we derived the following building blocks:

- Example 5.4.1: If $\psi(\mathbb{P}) = E_{\mathbb{P}}(X)$, then $\mathbb{IF}\{\psi(\mathbb{P})\} = X - \psi(\mathbb{P})$.
- Example 5.4.2: If $\psi_x(\mathbb{P}) = \mathbb{P}(X = x)$, then $\mathbb{IF}\{\psi_x(\mathbb{P})\} = I(X = x) - \mathbb{P}(X = x)$
- Example 5.4.5: If $\psi_x(\mathbb{P}) = E_{\mathbb{P}}(Y \mid X = x)$, then $\mathbb{IF}\{\psi_x(\mathbb{P})\} = \frac{I(X=x)}{\mathbb{P}(X=x)}\{Y - E_{\mathbb{P}}(Y \mid X = x)\}$

Example 5.4.6 (Example 5.4.5 continued)**.** We consider the random variable $Z = (X, Y)$ and the functional $\psi_x(\mathbb{P}) = E_{\mathbb{P}}(Y \mid X = x)$. Using the three tricks we get:

$$
\begin{aligned}
\mathbb{IF}\{\psi_x(\mathbb{P})\} &\overset{T1}{=} \mathbb{IF}\left(\sum_y y\mathbb{P}(Y = y \mid X = x)\right) = \sum_y y\mathbb{IF}\left(\frac{\mathbb{P}(Z = z)}{\mathbb{P}(X = x)}\right) \\
&\overset{T2}{=} \sum_y y\frac{\mathbb{P}(X = x)\mathbb{IF}(\mathbb{P}(Z = z)) - \mathbb{P}(Z = z)\mathbb{IF}(\mathbb{P}(X = x))}{\mathbb{P}(X = x)^2} \\
&\overset{T3}{=} \sum_y y\frac{\mathbb{P}(X = x)(I(Z = z) - \mathbb{P}(Z = z)) - \mathbb{P}(Z = z)(I(X = x) - \mathbb{P}(X = x))}{\mathbb{P}(X = x)^2} \\
&= \sum_y y\frac{\mathbb{P}(X = x)I(Z = z) - \mathbb{P}(Z = z)I(X = x)}{\mathbb{P}(X = x)^2} \\
&= \sum_y y\frac{I(Y = y, X = x)}{\mathbb{P}(X = x)} - y\frac{I(X = x)}{\mathbb{P}(X = x)}\mathbb{P}(Y = y \mid X = x) \\
&= \frac{I(X = x)}{\mathbb{P}(X = x)}\{Y - E_{\mathbb{P}}(Y \mid X = x)\}
\end{aligned}
$$

which is the same conclusion as Example 5.4.5.

Example 5.4.7 (Average treatment effect)**.** Let $Z = (X, A, Y)$ where X are covariates, A is a binary treatment variable taking values 0 and 1, and Y is the response variable. Under suitable identifying assumptions (Kennedy, 2022) the average treatment effect is identified as

$$\psi(\mathbb{P}) = E_{\mathbb{P}}\{E_{\mathbb{P}}(Y \mid X, A = 1)\} = E_{\mathbb{P}}(\mu_{\mathbb{P}}(X)),$$

where $\mu_{\mathbb{P}}(X) = E_{\mathbb{P}}(Y \mid X, A = 1)$. Applying the three tricks we get:

$$
\begin{aligned}
\mathbb{IF}\{\psi(\mathbb{P})\} &= \mathbb{IF}(E_{\mathbb{P}}\{E_{\mathbb{P}}(Y \mid X, A = 1)\}) \\
&\overset{T1}{=} \mathbb{IF}\left(\sum_x \mu_{\mathbb{P}}(x)\mathbb{P}(X = x)\right) \overset{T2}{=} \sum_x \mathbb{IF}(\mu_{\mathbb{P}}(x))\mathbb{P}(X = x) + \mu_{\mathbb{P}}(x)\mathbb{IF}(\mathbb{P}(X = x)) \\
&\overset{T3}{=} \sum_x \left(\frac{I(X = x, A = 1)}{\mathbb{P}(X = x, A = 1)}\{Y - E_{\mathbb{P}}(\mu_{\mathbb{P}}(x)\}\,\mathbb{P}(X = x) + \mu_{\mathbb{P}}(x)\{I(X = x) - \mathbb{P}(X = x)\}\right) \\
&= \frac{I(A = 1)}{\mathbb{P}(A = 1 \mid X)}\{Y - \mu_{\mathbb{P}}(X)\} + \mu_{\mathbb{P}}(X) - \psi(\mathbb{P})
\end{aligned}
$$

Although this final approach relies on more assumptions than the other two, it is also the most direct approach. Each approach has its advantages and disadvantages, so being able to use all three methods is a powerful tool in deriving influence functions in general.

5.5 Influence function based estimators

There are several ways to use EIFs in the construction of estimators, like targeted learning estimators and one-step estimators (Hines et al., 2022). In this thesis, we focus on the *one-step estimator*, since it is often simple to implement and has an intuitive interpretation.

Panel (B) of Figure 5 motivates the construction of the one-step estimator. As noted in Remark 5.3.9, the EIF can be used to find the slope (dashed line) of the functional values along the path (solid line) at $\epsilon = 1$. This provides an estimate (triangle) closer to the true value of the functional (circle), than the plug-in estimate (square). The value of the triangle has been calculated as

$$\psi_{OS}(\hat{\mathbb{P}}) = \psi(\hat{\mathbb{P}}) + \int \phi(z;\hat{\mathbb{P}})d\mathbb{P}(z) \tag{5.12}$$

where $\psi(\hat{\mathbb{P}})$ is the 'naive' plug-in estimator presented in Eq. (5.1). Eq. (5.12) suggests an alternative to the plug-in estimator, and we revisit the von Mises expansion from Eq. (5.3) to justify Eq. (5.12). When a functional $\psi : \mathcal{P} \to \mathbb{R}$ admits a von Mises expansion we can evaluate the error of the naive plug-in estimator:

$$\psi(\hat{\mathbb{P}}) - \psi(\mathbb{P}) = -\int \phi(z;\hat{\mathbb{P}})d\mathbb{P}(z) + R_2(\hat{\mathbb{P}},\mathbb{P}) \tag{5.13}$$

since $\int \phi(z;\hat{\mathbb{P}})d\hat{\mathbb{P}}(z) = 0$ by definition. The term $\int \phi(z;\hat{\mathbb{P}})d\mathbb{P}(z)$ is called a *first order bias term*, and since $\mathbb{P}$ is unknown, we have to estimate it with the naive sample average:

$$-\int \phi(z;\hat{\mathbb{P}})d\mathbb{P}(z) \approx -\int \phi(z;\hat{\mathbb{P}})d\mathbb{P}_n(z) = -\frac{1}{n}\sum_{i=1}^{n}\phi(Z_i,\hat{\mathbb{P}}) = -\mathbb{P}_n\{\phi(Z;\hat{\mathbb{P}})\}$$

The first-order bias-corrected estimator is then given by

$$\hat{\psi}_{OS}(\hat{\mathbb{P}}) = \psi(\hat{\mathbb{P}}) + \mathbb{P}_n\{\phi(Z,\hat{\mathbb{P}})\} \tag{5.14}$$

This estimator is known as the *one-step estimator*, which is a bias corrected version of the plug-in estimator.

Remark 5.5.1 (Oracle vs sample one-step estimator)**.** Eqs. (5.12) and (5.14) are different, because of the first order bias term. Eq. (5.12) is a theoretical expression, since it includes the unknown $\mathbb{P}$, while Eq. (5.14) is an actual estimator, because we estimate $\mathbb{P}$ by $\mathbb{P}_n$. We call Eq. (5.12) the *oracle one-step* estimator, since it is the asymptotic version of the *sample one-step estimator* presented in Eq. (5.14).

Example 5.5.2 (One-step estimator for the average treatment effect)**.** In Example 5.4.7 we derived the EIF for the average treatment effect.

$$\begin{aligned}\psi(\mathbb{P}) &= E_{\mathbb{P}}\{E_{\mathbb{P}}(Y \mid X, A=1)\} = E_{\mathbb{P}}(\mu_{\mathbb{P}}(X))\\ \phi(Z;\mathbb{P}) &= \frac{I(A=1)}{\mathbb{P}(A=1 \mid X)}\{Y - \mu_{\mathbb{P}}(X)\} + \mu_{\mathbb{P}}(X) - \psi(\mathbb{P})\end{aligned}$$

Let $Z_i = (X_i, A_i, Y_i)$ and assume we have n independent observations $Z_1, ..., Z_n$. The one-step estimator for the average treatment effect is:

$$\begin{aligned}\hat{\psi}_{OS}(\hat{\mathbb{P}}) &= \psi(\hat{\mathbb{P}}) + \mathbb{P}_n\{\phi(Z,\hat{\mathbb{P}})\}\\ &\overset{Eq.\ (5.14)}{=} \psi(\hat{\mathbb{P}}) + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(A_i=1)}{\hat{\mathbb{P}}(A=1 \mid X_i)}\{Y_i - \mu_{\hat{\mathbb{P}}}(X_i)\} + \mu_{\hat{\mathbb{P}}}(X_i) - \psi(\hat{\mathbb{P}})\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(A_i=1)}{\hat{\mathbb{P}}(A=1 \mid X_i)}\{Y_i - \mu_{\hat{\mathbb{P}}}(X_i)\} + \mu_{\hat{\mathbb{P}}}(X_i)\right\}\end{aligned} \tag{5.15}$$

Remark 5.5.3 (Nuisance parameters)**.** In Eq. (5.15) the parameters $\hat{\mathbb{P}}(A = 1, \mid X)$ and $\mu_{\hat{\mathbb{P}}}(X)$ are both *estimates* of the true values, $\mathbb{P}(A = 1 \mid X)$ and $\mu_{\mathbb{P}}(X)$. This is an example of *nuisance parameters*. Nuisance parameters are components of the estimator that have to be estimated as well. There are several approaches to estimating $\mathbb{P}(A = 1, \mid X)$ and $\mu_{\mathbb{P}}(X)$, including semiparametric estimators or nonparametric machine learning based estimators. We discuss nuisance parameters in greater detail in Subsection 5.5.2.

To investigate the asymptotic properties of the one-step estimator, we look at the following difference

$$\begin{aligned}
\hat{\psi}_{OS}(\hat{\mathbb{P}}) - \psi(\mathbb{P}) &= \psi(\hat{\mathbb{P}}) + \mathbb{P}_n\{\phi(Z, \hat{\mathbb{P}})\} - \psi(\mathbb{P}) \\
&\overset{Eq.\ (5.13)}{=} - \int \phi(z; \hat{\mathbb{P}}) d\mathbb{P}(z) + R_2(\hat{\mathbb{P}}, \mathbb{P}) + \mathbb{P}_n\{\phi(Z, \hat{\mathbb{P}})\} \\
&= \int \phi(z; \hat{\mathbb{P}}) d(\mathbb{P}_n - \mathbb{P})(z) + R_2(\hat{\mathbb{P}}, \mathbb{P}) \\
&= \int \phi(z; \mathbb{P}) d(\mathbb{P}_n - \mathbb{P})(z) + \int \phi(z; \hat{\mathbb{P}}) - \phi(z; \mathbb{P}) d(\mathbb{P}_n - \mathbb{P})(z) + R_2(\hat{\mathbb{P}}, \mathbb{P}) \\
&= \mathbb{P}_n \left\{\phi(Z; \mathbb{P})\right\} + \int \phi(z; \hat{\mathbb{P}}) - \phi(z; \mathbb{P}) d(\mathbb{P}_n - \mathbb{P})(z) + R_2(\hat{\mathbb{P}}, \mathbb{P}) \\
&= S^* + T_1 + T_2,
\end{aligned} \tag{5.16}$$

where T_1 is called the *empirical process term*, and T_2 is called the *remainder bias term.* The term $S^* = \mathbb{P}_n \left\{\phi(Z; \mathbb{P})\right\}$ is a mean of i.i.d random variables which by the law of large numbers converges in probability to $E_{\mathbb{P}}\{\phi(Z; \mathbb{P})\} = 0$. When $\sqrt{n}T_1$ and $\sqrt{n}T_2$ both converge to 0 in probability, i.e. $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$ and $T_2 = o_{\mathbb{P}}(1/\sqrt{n})$, then

$$\sqrt{n}(\hat{\psi}_{OS}(\hat{\mathbb{P}}) - \psi(\mathbb{P})) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{var}\{\phi(Z; \mathbb{P})\})$$

due to the central limit theorem and Slutsky's theorem (Kennedy, 2022). This implies that under regularity conditions, the one-step estimator is asymptotically efficient by Proposition 5.3.10. We will describe some properties of Eq. (5.16) below.

5.5.1 The empirical process term

The empirical process term is

$$T_1 = \int \phi(z; \hat{\mathbb{P}}) - \phi(z; \mathbb{P}) d(\mathbb{P}_n - \mathbb{P})(z),$$

The challenge with the empirical process term is to estimate both $\mathbb{P}_n(\phi(Z; \hat{\mathbb{P}}))$ and expressions that depend on $\hat{\mathbb{P}}$ with the same finite sample as it may lead to over-fitting issues. One solution is to estimate $\hat{\mathbb{P}}$ and $\mathbb{P}_n$ on independent data sets by cross-fitting procedures as outlined in Hines et al. (2022). If we in addition assume that $\phi(Z; \hat{\mathbb{P}})$ converges to $\phi(Z; \mathbb{P})$ in $\mathcal{L}_2(\mathbb{P})$, then it can be established that $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$ (Kennedy, 2022). Cross-fitting procedures introduce finite-sample bias in the estimator due to estimation on smaller data sets. In general, the empirical process term is negligible and of no concern.

5.5.2 Controlling the remainder bias term

The remainder bias term $T_2 = R_2(\hat{\mathbb{P}}, \mathbb{P})$ will usually have to be studied on a case-specific basis (Kennedy, 2022). Often this term consists of second-order products of differences between functionals (usually nuisance parameters) evaluated in $\mathbb{P}$ and $\hat{\mathbb{P}}$ respectively, and the task is to show that the products are of order $o_{\mathbb{P}}(1/\sqrt{n})$. In some cases, it suffices to show that one of the terms in the product converges fast enough, which allows other terms to converge at a slow rate, or even to be misspecified. Estimators with multiple nuisance parameters that are consistent even if one nuisance parameter is misspecified are called *doubly robust* estimators. The one-step estimator of the average treatment effect is doubly robust.

Example 5.5.4 (Doubly robustness of one-step estimator for average treatment effect)**.** We derive the remainder bias term for the one-step estimator of the average treatment effect from Example 5.5.2 to show that it is doubly robust. Let $\phi(Z; P)$ be the influence function for the functional for the average treatment effect, ψ.

$$\begin{aligned}
E_{\mathbb{P}}\phi(Z;\hat{\mathbb{P}}) &= E_{\mathbb{P}}\left(\frac{I(A=1)}{\hat{\mathbb{P}}(A=1\mid X)}\{Y-\mu_{\hat{\mathbb{P}}}(X)\}+\mu_{\hat{\mathbb{P}}}(X)-\psi(\hat{\mathbb{P}})\right)\\
&= E_{\mathbb{P}}\left(E_{\mathbb{P}}\left[\frac{I(A=1)}{\hat{\mathbb{P}}(A=1\mid X)}\{Y-\mu_{\hat{\mathbb{P}}}(X)\}+\mu_{\hat{\mathbb{P}}}(X)\mid X\right]\right)-\psi(\hat{\mathbb{P}})\\
&= E_{\mathbb{P}}\left(\frac{1}{\hat{\mathbb{P}}(A=1\mid X)}E_{\mathbb{P}}\left[I(A=1)\{Y-\mu_{\hat{\mathbb{P}}}(X)\}\mid X\right]+\mu_{\hat{\mathbb{P}}}(X)\right)-\psi(\hat{\mathbb{P}})\\
&\overset{\star}{=} E_{\mathbb{P}}\left(\frac{\mathbb{P}(A=1\mid X)}{\hat{\mathbb{P}}(A=1\mid X)}\{\mu_{\mathbb{P}}(X)-\mu_{\hat{\mathbb{P}}}(X)\}+\mu_{\hat{\mathbb{P}}}(X)\right)-\psi(\hat{\mathbb{P}})\\
&\overset{\dagger}{=} E_{\mathbb{P}}\left(\frac{\mathbb{P}(A=1\mid X)-\hat{\mathbb{P}}(A=1\mid X)}{\hat{\mathbb{P}}(A=1\mid X)}\{\mu_{\mathbb{P}}(X)-\mu_{\hat{\mathbb{P}}}(X)\}\right)\\
&\quad + E_{\mathbb{P}}\left(\mu_{\mathbb{P}}(X)-\mu_{\hat{\mathbb{P}}}(X)+\mu_{\hat{\mathbb{P}}}(X)\right)-\psi(\hat{\mathbb{P}})\\
&= E_{\mathbb{P}}\left(\frac{\mathbb{P}(A=1\mid X)-\hat{\mathbb{P}}(A=1\mid X)}{\hat{\mathbb{P}}(A=1\mid X)}\{\mu_{\mathbb{P}}(X)-\mu_{\hat{\mathbb{P}}}(X)\}\right)+\psi(\mathbb{P})-\psi(\hat{\mathbb{P}}),
\end{aligned} \tag{5.17}$$

where $\star$ follows because

$$\begin{aligned}
E_{\mathbb{P}}[I(A=1)Y\mid X] &= E_{\mathbb{P}}[E_{\mathbb{P}}\{I(A=1)Y\mid A,X\}\mid X]\\
&= E_{\mathbb{P}}[I(A=1)E_{\mathbb{P}}\{Y\mid A,X\}\mid X]\\
&= E_{\mathbb{P}}[I(A=1)E_{\mathbb{P}}\{Y\mid A=1,X\}\mid X]\\
&= E_{\mathbb{P}}[I(A=1)\mid X]E_{\mathbb{P}}[Y\mid A=1,X]\\
&= \mathbb{P}(A=1\mid X)\mu_{\mathbb{P}}(X),
\end{aligned}$$

which leads to

$$\begin{aligned}
E_{\mathbb{P}}[I(A=1)\{Y-\mu_{\hat{\mathbb{P}}}(X)\}\mid X] &= E_{\mathbb{P}}[I(A=1)Y\mid X]-E_{\mathbb{P}}[I(A=1)\mid X]\mu_{\hat{\mathbb{P}}}(X)\\
&= \mathbb{P}(A=1\mid X)(\mu_{\mathbb{P}}(X)-\mu_{\hat{\mathbb{P}}}(X)),
\end{aligned}$$

and † holds because

$$\mathbb{P}(A=1\mid X)=\mathbb{P}(A=1\mid X)-\hat{\mathbb{P}}(A=1\mid X)+\hat{\mathbb{P}}(A=1\mid X).$$

This implies that

$$\begin{aligned}T_2 &\overset{Eq.~(5.3)}{=} \psi(\hat{\mathbb{P}})-\psi(\mathbb{P})+E_{\mathbb{P}}\phi(Z;\hat{\mathbb{P}})\\ &\overset{Eq.~(5.17)}{=} E_{\mathbb{P}}\left(\frac{\mathbb{P}(A=1\mid X)-\hat{\mathbb{P}}(A=1\mid X)}{\hat{\mathbb{P}}(A=1\mid X)}\{\mu_{\mathbb{P}}(X)-\mu_{\hat{\mathbb{P}}}(X)\}\right)\end{aligned}$$

If $\hat{\mathbb{P}}(A=1\mid X)\to\mathbb{P}(A=1\mid X)$ or $\mu_{\hat{\mathbb{P}}}(X)\to\mu_{\mathbb{P}}(X)$, then $T_2\to 0$, which means one of the estimators $\mu_{\hat{\mathbb{P}}}(X)$ or $\hat{\mathbb{P}}(A=1\mid X)$ can be inconsistent. It is possible to control the propensity score in some cases, which guarantees consistent estimation even when $\mu_{\hat{\mathbb{P}}}(X)$ is misspecified.

5.5.3 Rate of convergence

From the von Mises expansion we noted that the plug-in estimator can be decomposed into

$$\psi(\hat{\mathbb{P}})=\psi(\mathbb{P})+\int\phi(z;\hat{\mathbb{P}})d(\hat{\mathbb{P}}-\mathbb{P})(z)+R_2(\hat{\mathbb{P}},\mathbb{P})$$

Assuming $\hat{\mathbb{P}}$ is a consistent estimator of $\mathbb{P}$, the plug-in estimator has an error of order $O(\|\hat{\mathbb{P}}-\mathbb{P}\|)$ unless the first-order bias term vanishes in the limit (Van Der Vaart, 2014). By removing the first-order bias term, the resulting estimator, i.e. the one-step estimator, achieves an error of order $O(\|\hat{\mathbb{P}}-\mathbb{P}\|^2)$ instead, under reasonable regularity conditions. Hence, the error term need only be of the order $o_{\mathbb{P}}(n^{-1/4})$ for the estimator to be of order $o_{\mathbb{P}}(n^{-1/2})$. The faster convergence of the estimator allows for parametric rates of convergence even when constructed based on flexible nonparametric estimators that themselves converge at slower rates (Fisher and Kennedy, 2019). We illustrate the faster convergence rate of the one-step estimator with the following example.

Example 5.5.5 (Estimated density)**.** Let $\mathbb{P}=\exp(1)$ and our functional of interest be the expected density:

$$\psi(\mathbb{P})=E_{\mathbb{P}}(f(x))=\int f(x)^2dx,\quad \mathbb{IF}(\psi(\mathbb{P}))=2(f(X)-\psi(\mathbb{P}))$$

We let our estimate of $\mathbb{P}$ be $\mathbb{P}_m=\exp(1+\frac{5}{\sqrt{m}})$, and thus our estimate of f is $f_m(x)=(1+\frac{5}{\sqrt{m}})\exp(-(1+\frac{5}{\sqrt{m}})x)$. It is clear that $f_m(x)\to f(x)$ for $m\to\infty$. We have the following three expressions available:

$$\begin{aligned}\hat{\psi}_{PI}&=E_{\mathbb{P}_m}f_m=\int f_m(x)^2dx\\ \psi_{OS}(\mathbb{P}_m)&=\hat{\psi}_{PI}+\int 2(f_m(x)-\psi(\mathbb{P}_m))d\mathbb{P}=\int 2f_m(x)d\mathbb{P}(x)-\psi(\mathbb{P}_m)\\ \hat{\psi}_{OS}(\mathbb{P}_m)&=\hat{\psi}_{PI}+\mathbb{P}_n\left\{2(f_m(x)-\psi(\mathbb{P}_m)\right\}=\frac{2}{n}\sum_{i=1}^n f_m(x_i)-\psi(\mathbb{P}_m)\end{aligned}$$

ψ_{PI} is the plug-in estimator, $\psi_{OS}(\mathbb{P}_m)$ is the *oracle one-step estimator* and $\hat{\psi}_{OS}(\mathbb{P}_m)$ is the *one-step estimator* (see Remark 5.5.1). In reality we would always estimate $\int f_m(x)d\mathbb{P}(x)$

by $\mathbb{P}_n\{f_m(x)\}$, but since we have chosen f_m and f ourselves, we can compute the stochastic integrals analytically to get an idea of how much better the sample one-step estimator can be for big n. The results are shown in Table 2 for different m and with $n = 100$ and visualized in Figure 6, while the theoretical arguments are presented in appendix A. From the top plot of Figure 6 it is clear that the oracle one-step estimator converges much faster than the plug-in estimator, considering the sequence triangles compared to the sequence of squares. Figure 6 bottom plot shows that the one-step estimator is not as good as the oracle one-step, but it is better than the plug-in estimator.

m	ψ_{PI}	ψ_{OS} (oracle)	$\hat{\psi}_{OS}$	ψ_{PI} bias	ψ_{OS} (oracle) bias	$\hat{\psi}_{OS}$ bias
10	1.2915694	0.1509490	-0.08610201	0.7905694	-0.34905097	0.5861020
20	1.0590170	0.2995532	0.11407500	0.5590170	-0.20044682	0.3859250
40	0.8952847	0.3880157	0.24028163	0.3952847	-0.11198431	0.2597184
80	0.7795085	0.4389414	0.31802492	0.2795085	-0.06105860	0.1819751
160	0.6976424	0.4673838	0.36519883	0.1976424	-0.03261616	0.1348012
320	0.6397542	0.4828636	0.39366600	0.1397542	-0.01713637	0.1063340

Table 2: The estimates from the plug-in estimator compared to both the oracle and the sample one-step estimator (from Example 5.5.5) with $n = 100$

So far we have not considered how censoring of the data can affect the nonparametric framework, which leads to the next section on the estimation on censored survival data.

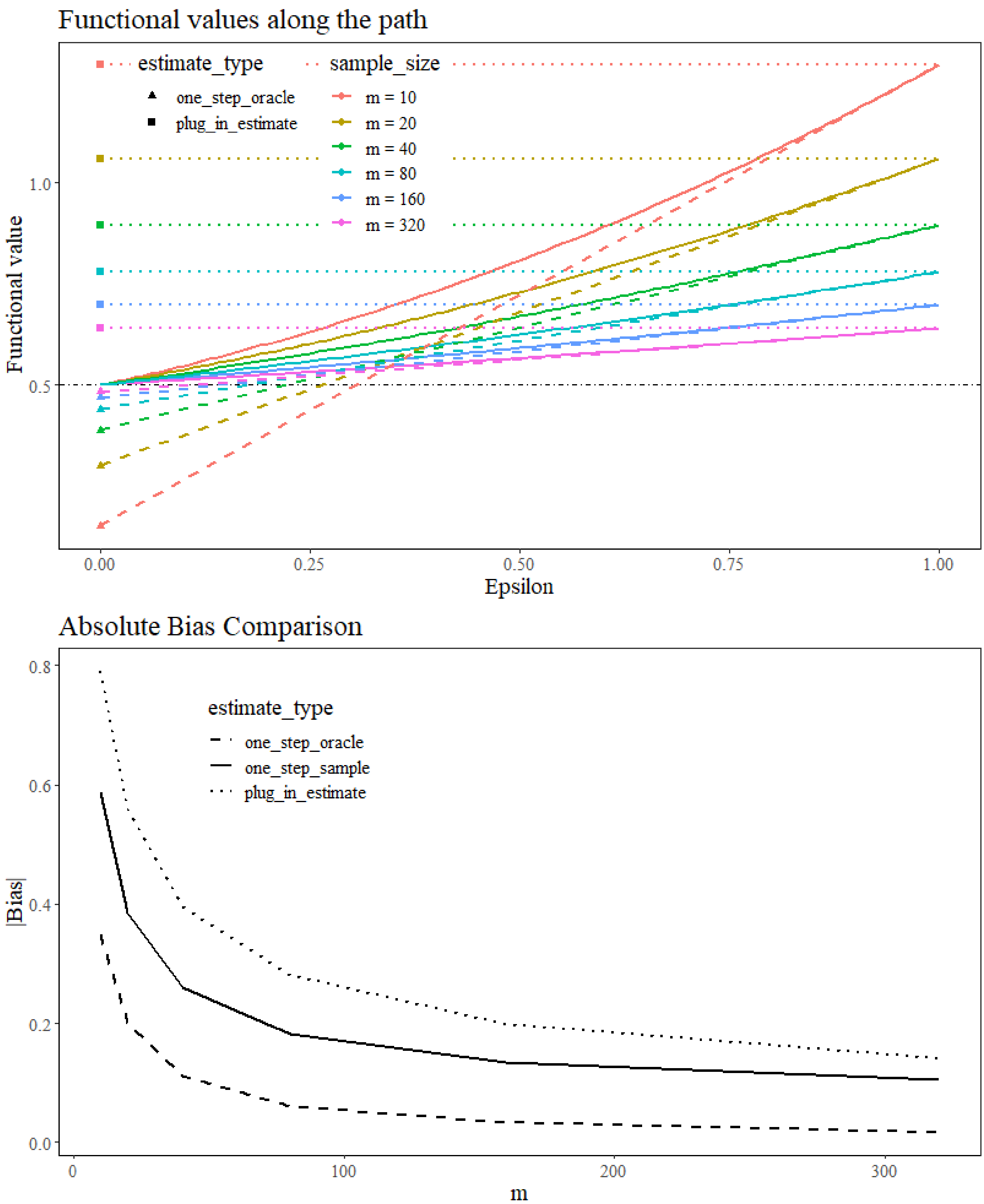

Figure 6: Top: The paths (full line) for the parametric submodels for different m plotted against the corresponding plug-in estimate (squares) and one-step estimates (triangles). The true value is plotted as the dot-dashed horizontal line. Bottom: The absolute bias of the one-step oracle estimator (dashed) vs plug-in estimate (dotted) vs one-step sample esitmator (solid). $n = 100$ for the one-step sample estimator. The figure is explained in Example 5.5.5

5.6 Generalizing to survival data

In this subsection we are interested in constructing estimands that allow us to conduct inference from censored survival data. We therefore introduce propositions for deriving censored data influence functions from full data influence functions.

We start by recalling the relevant notation from Section 2. Let $Z = (T^*, A, X)$ denote the true survival data, such that T^* is the true survival time of interest, A is a binary treatment covariate and X are other confounding variables. We let $W = (A, X)$ for notational convenience. We denote the observed data by $O = (T, \Delta, W)$ where $T = T^* \wedge C$, $\Delta = I(T^* \leq C)$ and C is the censoring time, and we assume conditionally independent censoring given W as in Eq. (3.7). In addition, we let $K_C(t \mid W) = \mathbb{P}(C > t \mid W)$ be the conditional censoring time survival function given W.

Proposition 5.6.1 (From Martinussen (n.d.)). *It suffices to rewrite terms that are not observed when we only have access to the observed data.*

Proposition 5.6.2. *Let $m_t(Z)$ be a full data function given by*

$$m_t(Z) = \int_0^t g(u \mid W) dM_{T^*}(u \mid W) \tag{5.18}$$

where $g(u \mid W)$ is any function such that the integral is well-defined and

$$M_{T^*}(t \mid W) = N_{T^*}(t) - \Lambda_{T^*}(t \mid W) = I(T^* \leq t) - \int_0^t I(T^* \geq u) dH(u \mid W)$$

where $M_{T^*}(t \mid W)$ is the uncensored counting process martingale, $\Lambda_{T^*}(t \mid W)$ is the conditional cumulative intensity process of $N_{T^*}(t)$ given W, and $H(u \mid W)$ is the conditional cumulative hazard rate given W. Assume that the censoring is independent given W. Then the censored data function is given by

$$m_t(O) = \int_0^t \frac{g(u \mid W)}{K_C(u \mid W)} dM_T(u \mid W)$$

where

$$M_T(t \mid W) = N_T(t) - \Lambda_T(t \mid W) = I(T \leq t, \Delta = 1) - \int_0^t I(T \geq u) dH(u \mid W)$$

is the observed counting process martingale.

Proof sketch. It follows from Martinussen (n.d.) that a general censored data function $m_t(O)$ can be obtained from the corresponding full data function $m_t(Z)$ by a mapping G given by

$$G\{m_t(Z)\} = \frac{\Delta \cdot m_t(Z)}{K_C(T \mid W)} + \int \frac{E\{m_t(Z) \mid T^* > u, W\}}{K_C(u \mid W)} dM_C(u \mid W),$$

where Δ is the censoring indicator,

$$M_C(t \mid W) = N_C(t) - \Lambda_C(t \mid W) = I(T \leq t, \Delta = 0) - \int_0^t I(T \geq u) dH_C(u \mid W),$$

and $H_C(u \mid W)$ is the conditional cumulative hazard for C. Let $m_t(Z)$ be given as in Eq. (5.18) such that

$$G\{m_t(Z)\} = \frac{\Delta \cdot \int_0^t g(u \mid W) dM_{T^*}(u \mid W)}{K_C(T \mid W)} + \int \frac{E\{\int_0^t g(u \mid W) dM_{T^*}(u \mid W) \mid T^* > u, W\}}{K_C(u \mid W)} dM_C(u \mid W)$$

It follows from Lemma A.2 in Lu and Tsiatis (2008) that

$$m_t(O) = G(m_t(Z)) = \int_0^t \frac{g(u \mid W)}{K_C(u \mid W)} dM_T(u \mid W)$$

□

Proposition 5.6.3. *Let $S : \mathbb{R}_+ \to (0, 1]$ denote a continuous differentiable survival function such that $S(t \mid W)$ is the probability of being at risk at time t given W. Then*

$$I(T^* > t) - S(t \mid W) = -S(t \mid W) \int_0^t \frac{1}{S(u \mid W)} dM_{T^*}(u \mid W), \tag{5.19}$$

where $M_{T^*}(t \mid W)$ is the full data counting process martingale given W.

Proof. First, we observe the following identities

$$\begin{aligned}
&\int_0^t \frac{1}{S(u \mid W)} I(u \le T^*) \alpha(u \mid W) du \\
&= I(T^* \le t) \int_0^{T^*} \frac{\alpha(u \mid W)}{S(u \mid W)} du + I(T^* > t) \int_0^t \frac{\alpha(u \mid W)}{S(u \mid W)} du \\
&\overset{Eq.\ (2.3)}{=} -I(T^* \le t) \int_0^{T^*} \frac{S'(u \mid W)}{S(u \mid W)^2} du - I(T^* > t) \int_0^t \frac{S'(u \mid W)}{S(u \mid W)^2} du \\
&= I(T^* \le t) \left[\frac{1}{S(u \mid W)} \right]_0^{T^*} + I(T^* > t) \left[\frac{1}{S(u \mid W)} \right]_0^t \qquad (5.20) \\
&= I(T^* \le t) \left(\frac{1}{S(T^* \mid W)} - \frac{1}{S(0 \mid W)} \right) + I(T^* > t) \left(\frac{1}{S(t \mid W)} - \frac{1}{S(0 \mid W)} \right) \\
&= I(T^* \le t) \left(\frac{1}{S(T^* \mid W)} - 1 \right) + I(T^* > t) \left(\frac{1}{S(t \mid W)} - 1 \right) \\
&= \frac{I(T^* \le t)}{S(T^* \mid W)} + \frac{I(T^* > t)}{S(t \mid W)} - 1,
\end{aligned}$$

where we use Riemann-Lebesgue interchangeability and integration by substitution to evaluate the integral in the third equality.

Then it follows that

$$\begin{aligned}
\int_0^t \frac{S(t \mid W)}{S(u \mid W)} dM_{T^*}(u \mid W) &\overset{(3.9)}{=} S(t \mid W) \left(\int_0^t \frac{1}{S(u \mid W)} dN_{T^*}(u) - \int_0^t \frac{1}{S(u \mid W)} d\Lambda_{T^*}(u \mid W) \right) \\
&= S(t \mid W) \left(\frac{I(T^* \le t)}{S(T^* \mid W)} - \int_0^t \frac{1}{S(u \mid W)} I(u \le T^*) \alpha(u \mid W) du \right) \\
&\overset{(5.20)}{=} S(t \mid W) \left(1 - \frac{I(T^* > t)}{S(t \mid W)} \right) \\
&= S(t \mid W) - I(T^* > t)
\end{aligned}$$

where $\alpha(u \mid W)$ is the conditional hazard function of T^* given W.

□

Remark 5.6.4 (On division in Proposition 5.6.3)**.** We divide by $S(T^* \mid W)$ multiple times in the proof of Proposition 5.6.3, which means we have to assume that S is positive for at all times to avoid division by zero. Hence, $S(T^* \mid W) = \mathbb{P}(T^* > T^* \mid W)$ is greater than zero, as $S(T^* \mid W)$ is simply a function evaluated in a random variable, T^*, that takes values in the domain of S. However, division by $S(T^* \mid W)$ may cause problems in practical estimation of the survival function. The estimates of the tail of the survival function may be very small, and consequently the fraction $\frac{1}{S(T^*|W)}$ may be huge which can lead to imprecise estimates.

We illustrate how to use the above propositions to find the average treatment effect for survival data.

Example 5.6.5 (Average treatment effect on survival data)**.** The influence function for the average treatment effect is given by Example 5.4.7

$$\mathbb{IF}\{\psi(\mathbb{P}); Z\} = \frac{I(A=1)}{\mathbb{P}(A=1 \mid X)}\{Y - \mu_{\mathbb{P}}(X)\} + \mu_{\mathbb{P}}(X) - \psi(\mathbb{P})$$

where A is a binary treatment covariate, X denotes all other covariates, and we let $W = (A, X)$ for notational convenience. Let our data be of the type introduced in the beginning of the subsection, such that $Y = I(T^* > t)$ and the estimand and corresponding influence function is

$$\begin{aligned} \psi(\mathbb{P}) &= E_{\mathbb{P}}\{S(t \mid X, A=1)\} \\ \mathbb{IF}\{\psi(\mathbb{P}); Z\} &= \frac{I(A=1)}{\mathbb{P}(A=1 \mid X)}\{I(T^* > t) - S(t \mid A=1, X)\} + S(t \mid A=1, X) - \psi(\mathbb{P}) \end{aligned} \tag{5.21}$$

Assume that T^* is independent of the censoring C given W, and let $\pi_a(X) = \mathbb{P}(A = a \mid X)$. From Proposition 5.6.1 we only have to rewrite the term $I(T^* > t)$ in Eq. (5.21):

$$\begin{aligned} \mathbb{IF}\{\psi(\mathbb{P}); Z\} =& \frac{A}{\pi_1(X)}\{I(T^* > t) - S(t \mid A=1, X)\} + S(t \mid A=1, X) - \psi(\mathbb{P}) \\ \overset{Eq.\ (5.19)}{=}& \frac{A}{\pi_1(X)}\left\{-\int_0^t \frac{S(t \mid A=1, X)}{S(u \mid A=1, X)} dM_{T^*}(u \mid A=1, X)\right\} + S(t \mid A=1, X) - \psi(\mathbb{P}) \\ =& \frac{A}{\pi_1(X)}\left\{-\int_0^t \frac{S(t \mid W)}{S(u \mid W)} dM_{T^*}(u \mid W)\right\} + S(t \mid A=1, X) - \psi(\mathbb{P}) \end{aligned}$$

using that $I(A = 1)S(t \mid A = 1, X) = I(A = 1)S(t \mid W)$. From Proposition 5.6.2 we can go from the full data EIF to the censored data EIF since T^* and C are assumed to be conditionally independent given W:

$$\mathbb{IF}\{\psi(\mathbb{P}); O\} \overset{Proposition\ 5.6.2}{=} S(t \mid A=1, X) - \psi(\mathbb{P}) - \frac{A}{\pi_1(X)}\left\{\int_0^t \frac{S(t \mid W)}{K_C(u \mid W)S(u \mid W)} dM_T(u \mid W)\right\}$$

which yields the censored data EIF.

The three propositions introduced in this section are useful tools for going from full data EIFs to censored data EIFs. In the last section of this thesis we will use the tools presented in this section to derive the censored data EIF for the estimand, $E[\mathbb{P}(T^1 \wedge \tau > T^0 \wedge \tau \mid X)]$.

6 An alternative summary estimand for survival analysis

In the previous sections, we introduced results and examples that allow us to describe new estimands of interest, derive the corresponding efficient influence functions, and construct efficient nonparametric estimators. In this section we define a new estimand for survival data and construct its corresponding one-step estimator.

6.1 Setting

Let $Z = (T^*, A, X) \sim \mathbb{P}$ be a survival time data triple where T^* is the survival time random variable without censoring, A is a binary treatment variable and X denotes all other covariates. Let C be the censoring time variable, $T = T^* \wedge C$ the observed survival time and Δ the censoring indicator. For notational convenience let $W = (A, X)$, and let $O = (T, \Delta, W)$. The estimand of interest is

$$\psi(\mathbb{P}) = E_{\mathbb{P}}\{\mathbb{P}(T^1 \wedge \tau > T^0 \wedge \tau \mid X)\} \tag{6.1}$$

where T^1 is the survival time of an individual that has received treatment $(A = 1)$, T^0 is the survival time of an individual that has not received treatment $(A = 0)$ and τ is the maximum time length of interest (e.g. $\tau = 5$ years in a five year study). The estimand, ψ, quantifies the probability that the survival times in the treated group are longer than the survival times in the untreated group when averaged across all covariates, X. If

$$\frac{E_{\mathbb{P}}\{\mathbb{P}(T^1 \wedge \tau > T^0 \wedge \tau \mid X)\}}{E_{\mathbb{P}}\{\mathbb{P}(T^1 \wedge \tau > T^0 \wedge \tau \mid X)\} + E_{\mathbb{P}}\{\mathbb{P}(T^0 \wedge \tau > T^1 \wedge \tau \mid X)\}} > 0.5 \tag{6.2}$$

it suggests that the treatment $(A = 1)$ prolongs the survival times. We have to construct Eq. (6.2) since τ carries probability mass.

The formal justification for using the potential outcomes, T^1, T^0, in ψ as a tool for making causal claims requires identification assumptions (see Hernan and Robins (2020)), but we only focus on the concrete derivation of an efficient nonparametric estimator.

6.2 Full data efficient influence function for estimand

In order to ease notation, we divide the derivation of the efficient influence function into multiple steps and suppress the dependency on the distribution such that $\psi = \psi(\mathbb{P})$. First, we assume all data are discrete including time. We let $Y = I(T^* > t)$, $U = I(T^* = t)$, and for a given $z = (t, a, x)$ we define the following functionals

$$\begin{aligned}
\psi_{z,0} &= E_{\mathbb{P}}(U \mid A = 0, X = x) = \mathbb{P}(T^* = t \mid A = 0, X = x) \\
\psi_{z,1} &= E_{\mathbb{P}}(Y \mid A = 1, X = x) = \mathbb{P}(T^* > t \mid A = 1, X = x)
\end{aligned}$$

From Example 5.4.6 we get that

$$\begin{aligned}
\mathbb{IF}\{\psi_{z,0}\} &= \frac{I(A = 0, X = x)}{\mathbb{P}(A = 0, X = x)} \{U - E_{\mathbb{P}}(U \mid A = 0, X = x)\} \\
\mathbb{IF}\{\psi_{z,1}\} &= \frac{I(A = 1, X = x)}{\mathbb{P}(A = 1, X = x)} \{Y - E_{\mathbb{P}}(Y \mid A = 1, X = x)\}
\end{aligned}$$

Second, we let

$$\begin{aligned}
\tilde{\psi}_x &= \mathbb{P}(T^1 \wedge \tau > T^0 \wedge \tau \mid X = x) \\
&= \sum_{t=0}^{\tau} \mathbb{P}(T^* > t \mid A = 1, X = x)\mathbb{P}(T^* = t \mid A = 0, X = x) \\
&= \sum_{t=0}^{\tau} \psi_{z,1}\psi_{z,0}
\end{aligned}$$

be another helper-functional. The efficient influence function for $\tilde{\psi}$ is

$$\begin{aligned}
\mathbb{IF}\{\tilde{\psi}_x\} = \mathbb{IF}\left\{\sum_{t=0}^{\tau} \psi_{z,1}\psi_{z,0}\right\} &\overset{T2}{=} \sum_{t=0}^{\tau} \mathbb{IF}\{\psi_{z,1}\}\psi_{z,0} + \psi_{z,1}\mathbb{IF}\{\psi_{z,0}\} \\
\overset{T3}{=} \sum_{t=0}^{\tau} \frac{I(A = 1, X = x)}{\mathbb{P}(A = 1, X = x)} \left\{Y - \psi_{z,1}\right\} \psi_{z,0} &+ \frac{I(A = 0, X = x)}{\mathbb{P}(A = 0, X = x)} \left\{U - \psi_{z,0}\right\} \psi_{z,1}
\end{aligned}$$

In the second equality we use the chain rule, in the third equality we use known EIFs as building blocks and the rest follows from rearranging. Now, let

$$\pi_a(x) = \mathbb{P}(A = a \mid X = x), \quad p_x = \mathbb{P}(X = x)$$

where $p_x = p_x(\mathbb{P})$ is a functional of $\mathbb{P}$. The efficient influence function for the functional, ψ, from Eq. (6.1) is

$$\begin{aligned}
&\mathbb{IF}\{\psi\} \overset{T1}{=} \mathbb{IF}\left\{\sum_x \tilde{\psi}_x p_x\right\} \overset{T2}{=} \sum_x \Big[\mathbb{IF}\{\tilde{\psi}_x\}p_x + \tilde{\psi}_x \mathbb{IF}\{p_x\}\Big] \\
&\overset{T3}{=} \sum_x \Bigg[\sum_{t=0}^{\tau}\Bigg(\frac{I(A = 1, X = x)}{\mathbb{P}(A = 1, X = x)} \left\{Y - \psi_{z,1}\right\} \psi_{z,0} + \frac{I(A = 0, X = x)}{\mathbb{P}(A = 0, X = x)} \left\{U - \psi_{z,0}\right\} \psi_{z,1}\Bigg)p_x \\
&\quad + \tilde{\psi}_x \left(I(X = x) - p_x\right)\Bigg] \\
&= \sum_x \Bigg[\sum_{t=0}^{\tau}\Bigg(\frac{I(A = 1, X = x)}{\pi_1(x)} \left\{Y - \psi_{z,1}\right\} \psi_{z,0} + \frac{I(A = 0, X = x)}{\pi_0(x)} \left\{U - \psi_{z,0}\right\} \psi_{z,1}\Bigg)\Bigg] \\
&\quad + \sum_x \tilde{\psi}_x I(X = x) - \sum_x \tilde{\psi}_x p_x \\
&= \frac{1 - A}{\pi_0(X)} \sum_{t=0}^{\tau}\Bigg(\left\{I(T^* = t) - \mathbb{P}(T^* = t \mid A = 0, X)\right\} \mathbb{P}(T^* > t \mid A = 1, X)\Bigg) \\
&\quad + \frac{A}{\pi_1(X)} \sum_{t=0}^{\tau}\Bigg(\left\{I(T^* > t) - \mathbb{P}(T^* > t \mid A = 1, X)\right\} \mathbb{P}(T^* = t \mid A = 0, X)\Bigg) + \tilde{\psi}_X - \psi
\end{aligned}$$

By swapping sums with integrals and probability mass functions with probability density functions we obtain the continuous time equivalent by:

$$\begin{aligned}\mathbb{IF}(\psi; Z) = &\frac{1-A}{\pi_0(X)} \underbrace{\int_0^\tau I(T^* = t)S(t \mid A=1,X) + S'(t \mid A=0,X)S(t \mid A=1,X)dt}_{(1)} \\ &- \frac{A}{\pi_1(X)} \underbrace{\int_0^\tau \{I(T^* > t) - S(t \mid A=1,X)\}S'(t \mid A=0,X)dt}_{(2)} + \tilde{\psi}_X - \psi\end{aligned} \tag{6.3}$$

where

$$\tilde{\psi}_X = -\int_0^\tau S(t \mid A=1,X)S'(t \mid A=0,X)dt \tag{6.4}$$

This is the full data EIF for the estimand of interest (Eq. (6.1)). However, this EIF involves components like $I(T^* > t)$ and $I(T^* = t)$ that are unknown when we only have censored data available. We therefore rewrite the full data EIF to get the censored data EIF.

Remark 6.2.1 (Discrete time versus continuous time)**.** To use the approach outlined in Subsection 5.4.3, we assume data are discrete, but since time is not discrete, we have to switch back to continuous time to obtain a meaningful EIF. In the derivations above, we do this by swapping sums with integrals and probability mass functions with probability density functions. In Example 5.4.4, we outlined that this is problematic for densities at a single point. We claim this is not an issue in our derivations, as we integrate with respect to t every time we interchange $\mathbb{P}(T^* = t \mid A = 0, X)$ and $-S'(t \mid A = 0, X)$. This means we never consider the density in a point, but rather consider the average value of the density across the interval $[0, \tau]$. Formally we should check Definition 5.3.6 to ensure that the identified EIF in Eq. (6.3) is valid for continuous T^*.

6.3 Censored data efficient influence function for estimand

Subsection 5.6 provides tools to go from the full data EIF to the censored data EIF. To use these tools we assume T^* and C are conditionally independent given $W = (A, X)$. The idea is to rewrite expressions that are unknown when we only have access to the observed data (Proposition 5.6.1), and then apply Propositions 5.6.2 and 5.6.3 where appropriate. We look at the integrals (1) and (2) from Eq. (6.3) one at a time.

First, we split integral (1) into multiple parts and analyze each part:

$$\begin{aligned}&\int_0^\tau I(T^* = t)S(t \mid A=1,X) + S'(t \mid A=0,X)S(t \mid A=1,X)dt \\ &\overset{Remark\ 6.2.1}{=} \underbrace{\int_0^\tau I(T^* = t)S(t \mid A=1,X)dt}_{(\star)} - \tilde{\psi}_X\end{aligned}$$

where $\tilde{\psi}_X$ is defined as in Eq. (6.4). We apply Eq. (3.9) to rewrite $(\star)$:

$$\begin{aligned}\int_0^\tau I(T^* = t)S(t \mid A=1,X)dt &= \int_0^\tau S(t \mid A=1,X)dN_{T^*}(t) \\ &\overset{Eq.\ (3.9)}{=} \int_0^\tau S(t \mid A=1,X)dM_{T^*}(t \mid W) + \int_0^\tau S(t \mid A=1,X)d\Lambda_{T^*}(t \mid W)\end{aligned} \tag{6.5}$$

Using Proposition 5.6.2 and Eq. (3.6) along with the assumption that T^* is independent of C given W, we can rewrite Eq. (6.5) to the corresponding censored data EIF:

$$\int_0^\tau \frac{S(t \mid A=1,X)}{K_C(t \mid W)} dM_T(t \mid W) + \underbrace{\int_0^\tau S(t \mid A=1,X) I(T^* > t)\alpha(t \mid W)dt}_{(\dagger)}$$

Part ($\dagger$) is then rewritten using Proposition 5.6.3 and Eq. (2.3):

$$\begin{aligned}
&\int_0^\tau S(t \mid A=1,X) I(T^* > t)\alpha(t \mid W)dt \\
&= -\int_0^\tau S(t \mid A=1,X)\left\{S(t \mid W) - \int_0^t \frac{S(t \mid W)}{S(u \mid W)} dM_{T^*}(u \mid W)\right\}\frac{S'(t \mid W)}{S(t \mid W)}dt \\
&= -\int_0^\tau S(t \mid A=1,X)\left\{1 - \int_0^t \frac{1}{S(u \mid W)} dM_{T^*}(u \mid W)\right\} S'(t \mid W)dt \\
&= \int_0^\tau \underbrace{\int_0^t \frac{S(t \mid A=1,X)S'(t \mid W)}{S(u \mid W)} dM_{T^*}(u \mid W)\, dt}_{\triangle} - \int_0^\tau S(t \mid A=1,X)S'(t \mid W)dt
\end{aligned} \tag{6.6}$$

Using Proposition 5.6.2 we rewrite the full data function $\triangle$ such that the corresponding censored data version of Eq. (6.6) is

$$\int_0^\tau \int_0^t \frac{S(t \mid A=1,X)S'(t \mid W)}{S(u \mid W)K_C(u \mid W)} dM_T(u \mid W)dt - \int_0^\tau S(t \mid A=1,X)S'(t \mid W)dt$$

Collecting the parts above we get that the integral (1) multiplied by $\frac{1-A}{\pi_0(X)}$ is

$$\begin{aligned}
&\frac{1-A}{\pi_0(X)}\left(\int_0^\tau I(T^* = t)S(t \mid A=1,X)dt - \tilde{\psi}_X\right) \\
&= \frac{1-A}{\pi_0(X)}\Bigg(\int_0^\tau \frac{S(t \mid A=1,X)}{K_C(t \mid W)} dM_T(t \mid W) + \int_0^\tau \int_0^t \frac{S(t \mid A=1,X)S'(t \mid A=0,X)}{S(u \mid W)K_C(u \mid W)} dM_T(u \mid W)dt \\
&\quad - \int_0^\tau S(t \mid A=1,X)S'(t \mid W)dt - \tilde{\psi}_X\Bigg) \\
&= \frac{1-A}{\pi_0(X)}\left(\int_0^\tau \frac{S(t \mid A=1,X)}{K_C(t \mid W)} dM_T(t \mid W) + \int_0^\tau \int_0^t \frac{S(t \mid A=1,X)S'(t \mid A=0,X)}{S(u \mid W)K_C(u \mid W)} dM_T(u \mid W)dt\right)
\end{aligned} \tag{6.7}$$

since

$$\begin{aligned}
\frac{1-A}{\pi_0(X)}\int_0^\tau S(t \mid A=1,X)S'(t \mid W)dt &= \frac{1-A}{\pi_0(X)}\int_0^\tau S(t \mid A=1,X)S'(t \mid A=0,X)dt \\
&= -\frac{1-A}{\pi_0(X)}\tilde{\psi}_X
\end{aligned} \tag{6.8}$$

The integral (2) in Eq. (6.3) can be rewritten as:

$$
\begin{aligned}
&\int_0^\tau \{I(T^* > t) - S(t \mid A = 1, X)\} S'(t \mid A = 0, X)dt \\
&\overset{\mathit{Proposition}\ 5.6.3}{=} -\int_0^\tau \int_0^t \frac{S(t \mid A = 1, X)}{S(u \mid A = 1, X)} dM_{T^*}(u \mid A = 1, X) S'(t \mid A = 0, X)dt
\end{aligned}
$$

Using Proposition 5.6.2 and multiplying by $\frac{A}{\pi_1(X)}$ we get the censored data version:

$$
\begin{aligned}
&-\frac{A}{\pi_1(X)} \int_0^\tau \left\{ \int_0^t \frac{S(t \mid A = 1, X)}{S(u \mid A = 1, X) K_C(u \mid A = 1, X)} dM_T(u \mid A = 1, X) \right\} S'(t \mid A = 0, X)dt \\
&\overset{\star}{=} -\frac{A}{\pi_1(X)} \int_0^\tau \left\{ \int_0^t \frac{S(t \mid A = 1, X)}{S(u \mid W) K_C(u \mid W)} dM_T(u \mid W) \right\} S'(t \mid A = 0, X)dt
\end{aligned}
\tag{6.9}
$$

where $\star$ uses the same logic as in Eq. (6.8). Collecting the expressions in Eqs. (6.3), (6.7) and (6.9) we get the censored data EIF:

$$
\begin{aligned}
\mathbb{IF}(\psi; O) = &\frac{1 - A}{\pi_0(X)} \int_0^\tau \frac{S(t \mid A = 1, X)}{K_C(t \mid W)} dM_T(t \mid W) \\
&+ \frac{1 - A}{\pi_0(X)} \int_0^\tau \left\{ \int_0^t \frac{S(t \mid A = 1, X)}{S(u \mid W) K_C(u \mid W)} dM_T(u \mid W) S'(t \mid A = 0, X) \right\} dt \\
&+ \frac{A}{\pi_1(X)} \int_0^\tau \left\{ \int_0^t \frac{S(t \mid A = 1, X)}{S(u \mid W) K_C(u \mid W)} dM_T(u \mid W) S'(t \mid A = 0, X) \right\} dt \\
&+ \tilde{\psi}_X - \psi \\
&\overset{\mathit{Eq.}\ (2.3)}{=} \frac{1 - A}{\pi_0(X)} \int_0^\tau \frac{S(t \mid A = 1, X)}{K_C(t \mid W)} dM_T(t \mid W) + \tilde{\psi}_X - \psi \\
&- \left(\frac{A}{\pi_1(X)} + \frac{1 - A}{\pi_0(X)} \right) \int_0^\tau S(t \mid A = 1, X) S(t \mid A = 0, X) \tilde{M}(t; W) dH(t \mid A = 0, X)
\end{aligned}
\tag{6.10}
$$

where $\tilde{M}(t; W) = \int_0^t \frac{dM_T(u|W)}{S(u|W)K_C(u|W)}$ and H is the cumulative hazard. Thus, Eq. (6.10) is the censored data EIF. Formally we would have to check Definition 5.3.6 to make sure that this EIF is valid, but we refrain from doing so, and assume that the approach presented in Kennedy (2022) is valid.

6.4 One-step estimator

The derived EIF from Eq. (6.10) can now be used to construct the one-step estimator of the estimand in Eq. (6.1) by Eq. (5.14):

$$
\begin{aligned}
\hat{\psi}_{OS} &= \hat{\psi}_{PI} + \mathbb{P}_n\{\phi(O;\hat{\mathbb{P}})\} \\
&= \psi(\hat{\mathbb{P}}) + \frac{1}{n}\sum_{i=1}^{n}\Bigg[\frac{1-A_i}{\hat{\pi}_0(X_i)}\int_0^{\tau}\frac{\hat{S}(t \mid A_i=1,X_i)}{\hat{K}_C(t \mid W_i)}d\hat{M}_{T_i}(t \mid W_i) + \tilde{\psi}_{X_i}(\hat{\mathbb{P}}) - \psi(\hat{\mathbb{P}}) \\
&\quad - \left(\frac{A_i}{\hat{\pi}_1(X_i)} + \frac{1-A_i}{\hat{\pi}_0(X_i)}\right)\int_0^{\tau}\left\{\hat{S}(t \mid A_i=1,X_i)\hat{S}(t \mid A_i=0,X_i)\hat{\tilde{M}}(t;W_i)\right\}d\hat{H}(t \mid A_i=0,X_i)\Bigg] \\
&= \frac{1}{n}\sum_{i=1}^{n}\Bigg[\frac{1-A_i}{\hat{\pi}_0(X_i)}\int_0^{\tau}\frac{\hat{S}(t \mid A_i=1,X_i)}{\hat{K}_C(t \mid W_i)}d\hat{M}_{T_i}(t \mid W_i) + \tilde{\psi}_{X_i}(\hat{\mathbb{P}}) \\
&\quad - \left(\frac{A_i}{\hat{\pi}_1(X_i)} + \frac{1-A_i}{\hat{\pi}_0(X_i)}\right)\int_0^{\tau}\left\{\hat{S}(t \mid A_i=1,X_i)\hat{S}(t \mid A_i=0,X_i)\hat{\tilde{M}}(t;W_i)\right\}d\hat{H}(t \mid A_i=0,X_i)\Bigg]
\end{aligned}
\tag{6.11}
$$

The remaining challenge is to estimate all nuisance parameters i.e. the survival function (S), the survival function for censoring times (K_C) and the propensity score (π_a).

6.5 Discussion and perspectives

In theory, the one-step estimator from Eq. (6.11) is a nonparametric estimator of the average probability of prolonging survival times up to some time τ given a treatment. If we by divine insight or study design knew the true value of the nuisance parameters, the survival function (S), the censoring distribution (K_C) and the propensity score (π_a), we could obtain an estimate of the estimand in Eq. (6.1). Moreover, we could construct confidence intervals for the estimate using the asymptotic variance, $\text{var}\{\mathbb{IF}(\psi(\mathbb{P}))\}$ and finally conclude whether or not the treatment, A, has a positive effect. However, reality dictates that we have to construct working models to estimate each nuisance parameter[11]. There are different procedures for estimating the nuisance parameters: Semiparametric, nonparametric and machine learning methods. The Cox model from Section 4 is an example of a semiparametric model that efficiently estimates S and K_C under certain conditions. Introducing semiparametric models however, leads to the issues discussed in Subsection 4.2.

In general, it is difficult to construct entirely nonparametric estimators as they will typically involve nuisance parameters which have to be estimated with parametric or semiparamtric methods (Hines et al., 2022). However, as Hines et al. (2022) writes: "*The crucial difference is that compared with the parametric modelling approach, estimators based on the nonparametric model do not 'extract efficiency' from highly parametric modelling assumptions*". Nonparametric estimators may exhibit higher variance compared to their semiparametric counterparts, but they are less exposed to misspecification bias.

The nonparametric EIF-based estimators presented in this thesis exhibit some useful properties which justify their use. First, one-step estimators, like the one in Example 5.5.2, may be doubly robust under regularity conditions allowing for non-consistent estimation of at least one nuisance parameter.

In addition, EIF-based estimators enable the use of consistent estimators that converge at slow rates, which is impossible with the plug-in estimator. In later development of nonparametric estimation, the faster convergence rate of EIF-based estimators has proven useful as

[11] In some cases it is possible to control π_a by study design (like randomized control trials).

they can attain sufficiently fast convergence even when they are based on data-adaptive estimators, i.e. nonparametric and machine learning based estimators (Van Der Vaart, 2014; Fisher and Kennedy, 2019; David Benkeser and Van Der Laan, 2016; Benkeser et al., 2017). Ultimately this makes fully nonparametric estimators a feasible alternative to semiparametric or fully parametric estimation methods.

Even with data-adaptive estimation of nuisance parameters, we still rely on identifying assumptions and the assumption of conditional independence of censoring and survival times to get interpretable results, while consistency of our estimators requires regularity conditions on our functionals and error terms. Every assumption and condition is at the risk of being violated just like the model assumptions of parametric and semiparametric models. In general, one-step estimators and nonparametric estimation should not be seen as a substitute to semiparametric estimation like Cox regression, but rather as a supplement to the toolbox of survival analysis.

Future work could include development and analysis of nonparametric estimation of nuisance parameters such as highly adaptive lasso (HAL) estimation (David Benkeser and Van Der Laan, 2016), and analysis of other EIF-based estimators such as targeted minimum-loss estimators (TMLE's) (Benkeser et al., 2017). It could also be interesting to generalize the study of remainder terms of one-step estimators and find conditions for double robustness, in addition to looking into higher order influence functions to remove higher-order bias terms for even faster convergence rates.

References

Aalen, Odd O., Ørnulf Borgan, and Håkon K. Gjessing (2008). Survival and Event History Analysis. Red. by M. Gail et al. Statistics for Biology and Health. Springer New York: New York, NY. ISBN: 978-0-387-20287-7. DOI: 10.1007/978-0-387-68560-1. URL: http://link.springer.com/10.1007/978-0-387-68560-1 (visited on 10/08/2023).

Benkeser, D et al. (Dec. 1, 2017). "Doubly robust nonparametric inference on the average treatment effect". In: Biometrika 104(4), pp. 863–880. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/asx053. URL: https://academic.oup.com/biomet/article/104/4/863/4554445 (visited on 10/13/2023).

Benkeser, David and Mark Van Der Laan (Oct. 2016). "The Highly Adaptive Lasso Estimator". In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE: Montreal, QC, Canada, pp. 689–696. ISBN: 978-1-5090-5206-6. DOI: 10.1109/DSAA.2016.93. URL: http://ieeexplore.ieee.org/document/7796956/ (visited on 12/13/2023).

Fisher, Aaron and Edward H. Kennedy (Oct. 27, 2019). Visually Communicating and Teaching Intuition for In arXiv: 1810.03260[math, stat]. URL: http://arxiv.org/abs/1810.03260 (visited on 10/12/2023).

Fleming, Thomas R. and David P. Harrington (1991). Counting processes and survival analysis. Wiley series in probability and mathematical statistics. Wiley: New York. 429 pp. ISBN: 978-0-471-52218-8.

Hansen, Ernst (2023). Stochastic Processes. 4th ed. Københavns Universitet: Insitute for matematiske fag. 666 pp. ISBN: 978-87-7125-259-0.

Hernan, Miguel A. and James M. Robins (2020). Causal Inference: What If. Chapman & Hall/CRC.: Boca Raton. ISBN: 9781420076165. URL: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Hines, Oliver et al. (July 3, 2022). "Demystifying Statistical Learning Based on Efficient Influence Functions". In: The American Statistician 76(3), pp. 292–304. ISSN: 0003-1305, 1537-2731. DOI: 10.1080/00031305.2021.2021984. URL: https://www.tandfonline.com/doi/full/10.1080/00031305.2021.2021984 (visited on 10/12/2023).

Ichimura, Hidehiko and Whitney K. Newey (2022). "The influence function of semiparametric estimators". In: Quantitative Economics 13(1), pp. 29–61. ISSN: 1759-7323. DOI: 10.3982/QE826. URL: http://qeconomics.org/ojs/index.php/qe/article/view/QE826 (visited on 12/26/2023).

Kennedy, Edward H. (2022). "Semiparametric doubly robust targeted double machine learning: a review". In: Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.2203.06469. URL: https://arxiv.org/abs/2203.06469 (visited on 10/08/2023).

Lu, X. and A. A. Tsiatis (Sept. 1, 2008). "Improving the efficiency of the log-rank test using auxiliary covariates". In: Biometrika 95(3), pp. 679–694. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/asn003. URL: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asn003 (visited on 11/10/2023).

Martinussen, Torben (n.d.). "A note on survival analysis and martingales and calculation of efficient influence function based right-censored data". In: ().

Martinussen, Torben and Thomas H. Scheike (2006). Dynamic regression models for survival data. Statistics for biology and health. Springer: New York, NY. 470 pp. ISBN: 978-0-387-20274-7.

Schilling, René L. (2017). Measures, integrals and martingales. Second edition. Cambridge University Press: Cambridge, United Kingdom ; New York, NY. 476 pp. ISBN: 978-1-316-62024-3.

Skovgaard, Lene Theil and Susanne Rosthøj (2019). "Basal statistik - Overlevelsesanalyse".

Van Der Vaart, Aad (Nov. 1, 2014). "Higher Order Tangent Spaces and Influence Functions". In: Statistical Science 29(4). ISSN: 0883-4237. DOI: 10.1214/14-STS478. URL: https://projecteuclid.org/journals/statistical-science/volume-29/issue-4/Higher-Order-Tangent-Spaces-and-Influence-Functions/10.1214/14-STS478.full (visited on 12/26/2023).

Whitney, David, Ali Shojaie, and Marco Carone (Nov. 1, 2019). "Comment: Models as (Deliberate) Approximations". In: Statistical Science 34(4). ISSN: 0883-4237. DOI: 10.1214/19-STS747. URL: https://projecteuclid.org/journals/statistical-science/volume-34/issue-4/Comment-Models-as-Deliberate-Approximations/10.1214/19-STS747.full (visited on 10/08/2023).

A Visualizing influence functions

Here, we provide the mathematics behind Examples 5.3.2 and 5.5.5. Let $\mathbb{P} = \exp(1)$ and our estimate $\mathbb{P}_m = \exp\left(1 + \frac{5}{\sqrt{m}}\right)$ be two distributions with the densities:

$$
\begin{aligned}
f(x) &= \exp(-x) \text{ for } x \geq 0 \\
f_m(x) &= \left(1 + \frac{5}{\sqrt{m}}\right) \exp\left(-\left(1 + \frac{5}{\sqrt{m}}\right) x\right) \text{ for } x \geq 0, m \in \mathbb{N}
\end{aligned}
$$

respectively. In Example 5.3.2 we use $m = 100$, while m varies in Example 5.5.5. The functional of interest is the expected density

$$\psi(\mathbb{P}) = \int_0^\infty f^2(x)dx = E_{\mathbb{P}}(f(X))$$

The influence function can be derived using approach 3 as in Kennedy (2022):

$$
\begin{aligned}
\mathbb{IF}(\psi(P)) &= \sum_x \mathbb{IF}(f_P^2(x)) = \sum_x 2f_P(x)\mathbb{IF}(f_P(x)) \\
&= \sum_x 2f_P(x)(I(X = x) - f_P(x))) \\
&= 2(f_P(X) - \psi(P))
\end{aligned}
$$

For $\epsilon \in [0, 1]$ we define the submodel P_ϵ such that

$$f_\epsilon(x) = \epsilon f_m(x) + (1 - \epsilon)f(x) = \epsilon \left\{(1 + \frac{5}{\sqrt{m}}) \exp\left(-(1 + \frac{5}{\sqrt{m}})x\right)\right\} + (1 - \epsilon)\exp(-x)$$

is the density of P_ϵ with respect to the Lebesgue measure. In addition,

$$\psi(\mathbb{P}_m) = E_{\mathbb{P}_m} f_m(X) = \frac{1 + \frac{5}{\sqrt{m}}}{2}, \quad \psi(\mathbb{P}) = E_{\mathbb{P}} f(X) = \frac{1}{2}, \quad E_{\mathbb{P}} f_m(X) = \frac{1 + \frac{5}{\sqrt{m}}}{2 + \frac{5}{\sqrt{m}}}$$

Since

$$
\begin{aligned}
\int f(x)^2 dx &= \int \exp(-2x)dx = \frac{1}{2} \\
\int f_m^2(x)dx &= \int (1+\frac{5}{\sqrt{m}})^2 \exp(-2(1+\frac{5}{\sqrt{m}})x)dx = \frac{1+\frac{5}{\sqrt{m}}}{2} \\
\int f_m f dx &= \int (1+\frac{5}{\sqrt{m}}) \exp(-(2+\frac{5}{\sqrt{m}})x)dx = \frac{1+\frac{5}{\sqrt{m}}}{2+\frac{5}{\sqrt{m}}}
\end{aligned}
$$

Our functionals are

$$\psi_{PI} = \psi(\mathbb{P}_m) = E_{\mathbb{P}_m} f_m = \frac{1+\frac{5}{\sqrt{m}}}{2}$$

$$\psi_{OS}(\mathbb{P}_m) = \psi_{PI} + \int 2(f_m(x) - \psi(\mathbb{P}_m))d\mathbb{P} = 2E_{\mathbb{P}} f_m(X) - E_{\mathbb{P}_m} f_m(X) = \frac{2(1+\frac{5}{\sqrt{m}})}{2+\frac{5}{\sqrt{m}}} - \frac{1+\frac{5}{\sqrt{m}}}{2}$$

$$\hat{\psi}_{OS}(\mathbb{P}_m) = \psi_{PI} + \mathbb{P}_n \left\{2(f_m(x) - \psi(\mathbb{P}_m)\right\} = \frac{2}{n}\sum_{i=1}^{n} f_m(X_i) - \psi(\mathbb{P}_m)$$

where $X_i \overset{i.i.d}{\sim} \exp(1)$ for $i = 1..n$. Thus we have that

$$
\begin{aligned}
\psi(P_\epsilon) &= \int_0^\infty (f_\epsilon(x))^2 dx = \int_0^\infty (\epsilon f_m(x) + (1-\epsilon) f(x))^2 dx \\
&= \epsilon^2 \int_0^\infty f_m(x)^2 dx + (1-\epsilon)^2 \int_0^\infty f(x)^2 dx + 2\epsilon(1-\epsilon) \int_0^\infty f(x) f_m(x) dx \\
&= \epsilon^2 E_{\mathbb{P}_m} f_m(x) + (1-\epsilon)^2 E_{\mathbb{P}} f(x) + 2\epsilon(1-\epsilon) E_{\mathbb{P}} f_m(x) \\
&= \epsilon^2 \frac{1+\frac{5}{\sqrt{m}}}{2} + (1-\epsilon)^2 \frac{1}{2} + \epsilon(1-\epsilon)\frac{1+\frac{5}{\sqrt{m}}}{2+\frac{5}{\sqrt{m}}}
\end{aligned}
$$

From Riesz Representation Theorem (Hines et al., 2022) we obtain that the slope for the functional, $\psi(P_\epsilon)$ at $\epsilon = 1$ is

$$
\begin{aligned}
\frac{\partial}{\partial \epsilon}\psi(P_\epsilon)|_{\epsilon=1} &= -E_{\mathbb{P}}(\mathbb{IF}(\psi(\mathbb{P}_m))) \\
&= -\int_0^\infty 2(f_m(x) - \psi(\mathbb{P}_m)) f(x) dx \\
&= -\int_0^\infty 2 f_m(x) f(x) dx - 2E_{\mathbb{P}_m} f_m \int_0^\infty f(x) dx \\
&= 2(E_{\mathbb{P}_m} f_m - E_{\mathbb{P}} f_m) \\
&= 2\left(\frac{1+\frac{5}{\sqrt{m}}}{2} - 2\frac{1+\frac{5}{\sqrt{m}}}{2+\frac{5}{\sqrt{m}}}\right) = \frac{25+5\sqrt{m}}{2n+5\sqrt{m}}
\end{aligned}
$$

Furthermore, the error term (Hines et al., 2022) can be evaluated as

$$
\begin{aligned}
R_2 &= -\int (f_m(x) - f(x))^2 dx \\
&= 2E_{\mathbb{P}} f_m(x) - E_{\mathbb{P}} f(X) - E_{\mathbb{P}_m} f_m(x) \\
&= \frac{2(1+\frac{5}{\sqrt{m}})}{2+\frac{5}{\sqrt{m}}} - \frac{1}{2} - \frac{1+\frac{5}{\sqrt{m}}}{2}
\end{aligned}
$$