Received: 25 February 2020 | Revised: 13 August 2021 | Accepted: 8 November 2021

DOI: 10.1111/sjos.12570

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

# Robust inference with censored survival data

**Pierre-Yves Deléamont**[1] | **Elvezio Ronchetti**[2]

[1]Institute of Statistics, Faculty of Science, University of Neuchâtel, Neuchâtel, Switzerland

[2]Research Center for Statistics, Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland

**Correspondence**
Pierre-Yves Deléamont, Institute of Statistics, Faculty of Science, University of Neuchâtel, 2000 Neuchâtel, Switzerland.
Email: pierre-yves.deleamont@unine.ch

[Correction added on 19 April 2022, after first online publication: CSAL funding statement has been added.]

**Abstract**

Randomly censored survival data appear in a wide variety of applications in which the time until the occurrence of a certain event is not completely observable. In this paper, we assume that the statistician observes a possibly censored survival time along with a censoring indicator. In this setting, we study a class of M-estimators with a bounded influence function, in the spirit of the infinitesimal approach to robustness. We outline the main asymptotic properties of the robust M-estimators and characterize the optimal B-robust estimator according to two possible measures of sensitivity. Building on these results, we define robust testing procedures which are natural counterparts to the classical Wald, score, and likelihood ratio tests. The empirical performance of our robust estimators and tests is assessed in two extensive simulation studies. An application to data from a well-known medical study on head and neck cancer is also presented.

**KEYWORDS**

censoring, influence function, multiplicative intensity model, robustness, survival analysis

## 1 | INTRODUCTION

Inference based on censored survival data has been a central issue in statistics over the last decades. The seminal work of Kaplan and Meier (1958), for instance, is one of the most

cited papers in the statistical literature, with over 50,000 references. The reason for such an overwhelming interest in this type of data is certainly due to the great variety of fields in which these are encountered, ranging from medicine to engineering or finance. It seems fair to argue that nonparametric and semiparametric techniques have enjoyed a dominant position in the literature. Obviously, finding an appropriate parametric model, which is already far from trivial when the data are completely observed, becomes even more delicate in the presence of censoring. Nevertheless, nonparametric techniques are not the panacea. Aside from usual difficulties such as bandwidth selection, specific issues related to the nature of censored survival data come into play. For instance, Miller (1983) advocated against the blind use of the Kaplan–Meier estimator with no consideration for parametric alternatives. In particular, he showed that the Kaplan–Meier estimator featured a very large efficiency loss relative to a well-specified parametric model, especially for the estimation of tail probabilities. His findings were corroborated by Efron (1988), and largely supported by subsequent simulation studies led by Aranda-Ordaz (1987) and Klein and Moeschberger (1989). These authors noted, however, that the parametric maximum likelihood estimator (MLE) could be severely affected by outliers, reducing the efficiency gain of using a parametric model.

In numerous applications, estimating survival or hazard rates in the tail of the distribution is of utmost interest to the researcher. This is true, for instance, in many actuarial studies; see, for example, Gavrilov and Gavrilova (2011). When the risk set is small, the potential instability of the Kaplan–Meier and Ramlau–Hansen estimators (arguably the most widely used estimators of survival and hazard functions, respectively) is a cause for concern. As an example, Gámiz et al. (2016) recently analyzed mortality data for women in four countries (United States, United Kingdom, Denmark, and Iceland) in 2006, with the aim of estimating the hazard function. They reported the following issue with the data for Iceland. At very high ages, the risk set can be so small that the ratio of occurrences to exposures used to estimate the hazard can behave very wildly. Figure 1, reproduced from Gámiz et al. (2016), shows the effect of a small change in the exposure at age 106 on a Ramlau–Hansen-type estimator of the hazard. This change in the exposure results from assuming that the individual who died in March 2006 actually died in January 2006, at the same age.

The estimator is obviously highly unstable. The hazard estimates on the corrupted data are approximately multiplied by 10 at all high ages, and the shape of the estimated hazard function is now clearly decreasing in the tail of the distribution. Gámiz et al. (2016) proposed an estimator which appears to be able to cope with such issues satisfactorily, by introducing nonclassical weighting in local linear hazard estimation. Nevertheless, these authors did suggest to use a parametric model to transform the data in a first stage before using their nonparametric approach on the transformed data. In that case, as noted by the authors, the quality of the estimates is likely to depend crucially on the adequacy of the parametric model used in the first stage.

In this paper, we consider robust inference with censored survival data. Robust statistics can loosely be described as the statistics of approximate parametric models; see Huber (1981) (second edition by Huber & Ronchetti, 2009) and Hampel et al. (1986) for general references, and Heritier et al. (2009) for a more introductory treatment with applications in biostatistics. Based on the discussion above, robust statistics appears to be well suited to the context at hand. Indeed, it preserves the main advantages related to parametric inference while explicitly taking into account the fact that, in nearly any situation of interest, a parametric model is not flexible enough to capture all of the information contained in the data. We shall focus on infinitesimal robustness in the sense of Hampel et al. (1986). In other words, we examine the impact of small deviations from the model. The issue of global robustness, when larger deviations are considered, is left for

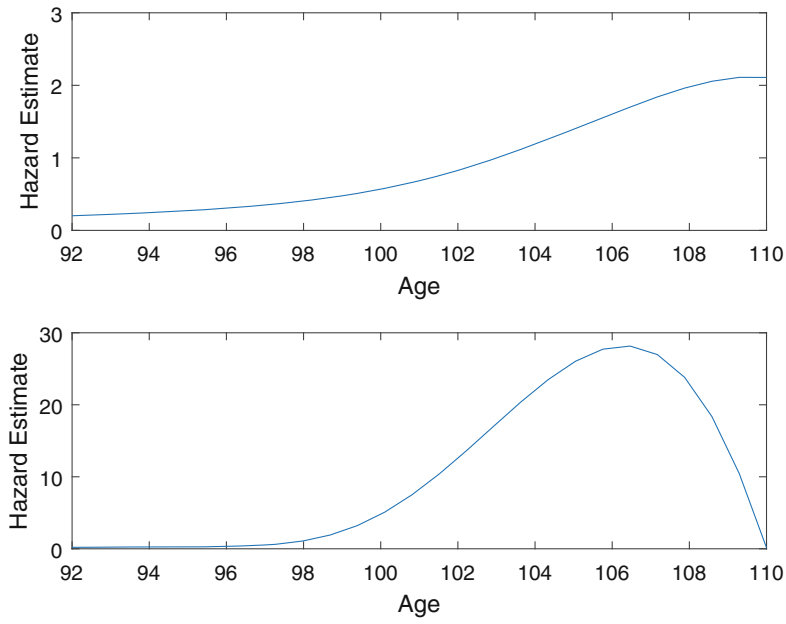


**FIGURE 1** Estimates of the hazard function with a Ramlau–Hansen-type estimator on mortality data for women in Iceland in 2006 (top) and on the same data with one corrupted observation (bottom)

future research. The presence of censoring, which we assume to be random, is challenging from a robustness standpoint. It creates a gap between the distribution of the actual survival times, which we want to model, and the distribution of the data that we actually observe. This point was already stressed by Samuels (1978) in an early attempt to adapt the infinitesimal robustness framework to the censored case. One important issue is related to Fisher consistency. It is well known from the robustness literature with completely observed data that it is in general necessary to recenter the estimating equation to preserve the Fisher consistency of the underlying M-functional. In the presence of censored data, we might expect that this recentering constant would depend on the distribution of the censoring times, which we arguably do not want to have to specify parametrically. We show that, for a large class of M-estimators first considered by Hjort (1985, 1992), there is no need for a recentering constant vector in order to have Fisher consistency. Building on this fact, we follow Hampel et al. (1986) and characterize the optimal B-robust estimator in this class of M-estimators. In other words, we look for the most efficient estimator among those which possess a bounded influence function. This estimator happens to have a very natural interpretation. We then use this optimal B-robust estimator to construct robust tests as in Heritier and Ronchetti (1994). These tests are designed to ensure the stability of the level under contamination.

The paper is organized as follows. Section 2 gives an overview of parametric inference with censored survival data. We focus on maximum likelihood estimation and on the three classical tests (Wald, score, and likelihood ratio), and present alternatives which have been designed to overcome the lack of robustness of the classical methods. In Section 3, we focus on the class of M-estimators considered by Hjort (1985, 1992) and outline the main asymptotic properties of these estimators. Section 4 contains the main results of this paper. We characterize the subclass of M-estimators with a bounded sensitivity to deviations from the model. Then, we present robust tests, which can be seen as natural extensions of the classical ones. In Section 5, we begin to

examine the finite-sample performance of the proposed estimator and tests with two simulation studies, considering the Weibull and log-normal models. In Section 6, we present an application of our methods to the well-known Head and Neck Cancer Study dataset introduced in the statistical literature by Efron (1988). Section 7 summarizes the main findings and presents opportunities for future research. The proofs of the main results can be found in the Appendix.

### 1.1 | Setup and notation

We close this introduction with a brief presentation of the setting considered in this paper and the related notation. Throughout we assume that our data consist of $n$ independent and identically distributed (i.i.d.) pairs $(X_1, \Delta_1), \dots, (X_n, \Delta_n)$, with $X_i := \min(T_i, C_i)$ and $\Delta_i := \mathbb{1}_{\{T_i \leq C_i\}}$, where $T_i$ and $C_i$ are the actual survival time and the censoring time for the $i$th observation, respectively. By the assumption of random censoring, $T_i$ is independent of $C_i$ for every $i = 1, \dots, n$. We denote the cumulative distribution function of any given set of variables by $F$ along with the appropriate subscript(s). The corresponding density, whenever it exists, is denoted by a lowercase $f$ with the same subscript(s). We consider a parametric model for $F_T$, so that $F_T(\cdot) \equiv F_T(\cdot; \boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. We use $\boldsymbol{\theta}$ as the general parameter argument, while $\boldsymbol{\theta}^*$ is used to indicate the true vector of parameters. The hazard function for the survival times is denoted by $\alpha(\cdot; \boldsymbol{\theta}) := f_T(\cdot; \boldsymbol{\theta})/(1 - F_T(\cdot; \boldsymbol{\theta}))$. The likelihood score and the logarithmic derivative of the hazard with respect to the parameter are denoted by $\boldsymbol{s}(\cdot; \boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}^\top} \log f_T(\cdot; \boldsymbol{\theta})$ and $\boldsymbol{h}(\cdot; \boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}^\top} \log \alpha(\cdot; \boldsymbol{\theta})$, respectively. We sometimes also make use of counting processes, in the spirit of most of the survival analysis literature since the seminal work of Aalen (1978); see the book by Andersen et al. (1993) for a detailed account. Specifically, let $(N_1, Y_1), \dots, (N_n, Y_n)$ be i.i.d. pairs of counting processes with respect to an increasing, right-continuous, complete filtration $\{\mathcal{F}_t, t \in \mathbb{R}_+\}$ satisfying the usual regularity conditions as in p. 60 of Andersen et al. (1993), with $N_i(t) := \mathbb{1}_{\{X_i \leq t, \Delta_i = 1\}}$ and $Y_i(t) := \mathbb{1}_{\{X_i \geq t\}}$. A process $\{N_i(t), t \in \mathbb{R}_+\}$ thus records whether an individual who is known to fail during the observation period has failed at or before a given time point, while a process $\{Y_i(t), t \in \mathbb{R}_+\}$ indicates whether an individual is still at risk of failure at a given time point. It is clear that the pair $(N_i, Y_i)$ contains as much information as the pair $(X_i, \Delta_i)$. We will follow Hjort (1992) and often work with the sum of these processes. Consequently, we define $N(t) := \sum_{i=1}^n N_i(t)$ and $Y(t) := \sum_{i=1}^n Y_i(t)$. We also assume that $\frac{1}{n} Y(t)$ converges to a limit denoted by $y(t)$ for every $t$ as $n \to \infty$. With this notation, our problem can be studied as a parametric version of Aalen (1978)'s multiplicative intensity model for the sum processes. In that case, the random intensity of the counting process $N$ is given by $\alpha(t; \boldsymbol{\theta})Y(t)$ at time $t$. It is important to realize that the process $\{M(t) := N(t) - \int_0^t \alpha(s; \boldsymbol{\theta})Y(s)ds, t \in \mathbb{R}_+\}$ is an $\{\mathcal{F}_t\}-$martingale at the model. The rest of the notation used in this paper is quite standard and will be clarified below as needed.

## 2 | OVERVIEW OF INFERENCE WITH CENSORED SURVIVAL DATA

In this section, we begin by reviewing maximum likelihood estimation and other M-estimation procedures proposed in the survival analysis literature. Then, we briefly address the issue of testing, by presenting the classical tests derived from the likelihood theory in the presence of censoring. We emphasize that our survey is by no means a thorough account of all of the literature on censored survival data. We deliberately focus on parametric techniques and only present the key results specialized to our needs.

## 2.1 | Estimation

Given a noninformative censoring scheme such as random censoring, the (partial) likelihood for $\theta$ given the observed data $(x_1, \delta_1), \ldots, (x_n, \delta_n)$ is proportional to $\prod_{i=1}^{n} \alpha(x_i; \theta)^{\delta_i} \exp\left(-\int_0^{x_i} \alpha(t; \theta) dt\right)$. The estimating equation for the MLE, $\hat{\theta}_{\mathbf{MLE}}$, can be expressed in counting process notation as $\int_0^{\infty} \boldsymbol{h}(t; \theta) \{dN(t) - \alpha(t; \theta) Y(t) dt\} = \mathbf{0}$. Maximum likelihood estimation in a multiplicative intensity model with application to censored survival data was studied by Borgan (1984). He showed that, under mild regularity conditions, the MLE is consistent and asymptotically normal, with asymptotic variance

$$\boldsymbol{V}_{\mathbf{MLE}}(\theta^*) = \boldsymbol{Q}_{\mathbf{MLE}}(\theta^*)^{-1} = \left[\int_0^{\infty} \boldsymbol{h}(t; \theta^*) \boldsymbol{h}(t; \theta^*)^{\top} \alpha(t; \theta^*) y(t) dt\right]^{-1}. \tag{1}$$

By analogy with M-estimation in the uncensored case, Hjort (1985) suggested to consider the solution to

$$\boldsymbol{U}(\theta) := \int_0^{\infty} \boldsymbol{\Psi}(t; \theta) \{dN(t) - \alpha(t; \theta) Y(t) dt\} = \mathbf{0}, \tag{2}$$

where $\boldsymbol{\Psi}(\cdot; \theta)$ is a vector of deterministic functions or predictable locally bounded processes chosen appropriately. In a subsequent contribution, Hjort (1992) proposed a specific "M-type estimator" defined through the equation

$$\int_0^{\infty} W(t) \boldsymbol{h}(t; \theta) \{dN(t) - \alpha(t; \theta) Y(t) dt\} = \mathbf{0},$$

where $W$ is a possibly random weighting function. This estimator can thus be seen as a weighted likelihood estimator. We note that although the author introduced this class of estimators with issues related to model misspecification in mind, he did not perform a thorough analysis of their robustness in the sense of the present paper.

A different approach to M-estimation for censored survival data was considered by Wang (1999). Following Reid (1981), she suggested an alternative extension of the classical M-estimators, with an estimating equation of the form

$$\int_0^{\infty} \boldsymbol{\Psi}(t; \theta) d\hat{F}_n(t) = \mathbf{0}, \tag{3}$$

where $\hat{F}_n$ is the Kaplan–Meier estimator of the survival time distribution. She established sufficient conditions under which such M-estimators are consistent and asymptotically normal. In that paper as well, issues related to robustness were mentioned, and some examples were presented, but a theoretical treatment was left for future research. In our view, the M-estimator proposed by Hjort (1985) appears to be more general, in the sense that it is formulated in the

context of a multiplicative intensity model and can thus be extended to other settings more easily. Also, as acknowledged by Wang (1999), it may have better properties in small samples because it is based on an unbiased estimating equation, unlike the Kaplan–Meier integrals as in (3) which can be severely biased. On the other hand, the type of M-estimator introduced by Hjort (1985) relies on a parametrically specified hazard and is thus subject to misspecification. Therefore, it seems to be a good compromise to consider a robust version of Hjort (1985)'s M-estimator, and this is what we shall focus on in this work.

A different strand of the literature has focused on adapting divergence-based approaches to the presence of censoring. Yang (1991) considered minimum Hellinger distance estimation in a random censorship model, extending previous work by Beran (1977). The author showed that the estimator features desirable asymptotic properties, in the sense that it is asymptotically efficient among the class of regular estimators when the parametric model is correctly specified, and also minimax robust in Hellinger neighbourhoods of the parametric family. One drawback of this type of approach is that it requires a smooth nonparametric estimate of the density of the survival time distribution. Hence, even though the approach is parametric in spirit, issues such as bandwidth selection still need to be addressed. Additionally, it is well known that the Hellinger distance leads to the consideration of only very small deviations from the model.

Another divergence-based approach, the so-called minimum density power divergence approach, was proposed by Basu et al. (2006), extending previous work by Basu et al. (1998). This method is based on a family of divergences, indexed by a tuning parameter $\alpha \in [0, 1]$. For a given value of $\alpha$, an estimator of the parameter $\boldsymbol{\theta}$ is obtained by minimizing the sample version of the divergence, given (in our notation) by $\int_0^\infty f_T(x;\boldsymbol{\theta})^{1+\alpha}dx - (1+1/\alpha)\int_0^\infty f_T(x;\boldsymbol{\theta})^{\alpha}d\hat{F}_n(x)$, where $\hat{F}_n$ once again denotes the Kaplan–Meier estimator. The authors showed that a small value of $\alpha$ results in a robust and reasonably efficient estimator. In fact, this estimator can be viewed as a particular case of the M-estimator proposed by Wang (1999), and thus potentially suffers from the drawbacks mentioned above. On the other hand, unlike the method proposed by Yang (1991), there is no need for a nonparametric estimate of a density.

## 2.2 | Testing

We now turn to the issue of testing in the presence of censored survival data. Given a parametric model, it is possible to define analogues of the three classical tests (Wald, score, and likelihood ratio) in this context in a fairly straightforward fashion. We only provide a short review; see, for example, chapter VI of Andersen et al. (1993) for more details. Suppose that we want to test the null hypothesis that $q < p$ linearly estimable functions of $\boldsymbol{\theta}$ are equal to zero. We partition the vector of parameters $\boldsymbol{\theta}^\top = \left(\boldsymbol{\theta}_{(1)}^\top, \boldsymbol{\theta}_{(2)}^\top\right)$, and denote the submatrices of any given matrix $\boldsymbol{A}$ corresponding to such a partition by $\boldsymbol{A}_{(mn)}$, $m, n \in \{1, 2\}$. By using a linear transformation, the hypothesis testing problem can be formulated as the test of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta^0}$, with $\boldsymbol{\theta}_{(1)}^{\boldsymbol{0}}$ left unspecified and $\boldsymbol{\theta}_{(2)}^{\boldsymbol{0}} = \mathbf{0}$, against the alternative $H_1 : \boldsymbol{\theta}_{(1)}$ unspecified, $\boldsymbol{\theta}_{(2)}^{\boldsymbol{0}} \neq \mathbf{0}$. Recall that the MLE in the full model, which we denote by $\hat{\boldsymbol{\theta}}_{\mathbf{MLE}}$ like in the previous subsection, is the minimizer of the log-likelihood $C_{\mathrm{MLE}}(\boldsymbol{\theta}) := \int_0^\infty \log\alpha(t;\boldsymbol{\theta})dN(t) - \int_0^\infty \alpha(t;\boldsymbol{\theta})Y(t)dt$, and thus solves $\boldsymbol{U}_{\mathbf{MLE}}(\boldsymbol{\theta}) := \int_0^\infty \boldsymbol{h}(t;\boldsymbol{\theta})\{dN(t) - \alpha(t;\boldsymbol{\theta})Y(t)dt\} = \mathbf{0}$. Its asymptotic variance, evaluated at the true parameter, is given by (1). Additionally, we denote by $\tilde{\boldsymbol{\theta}}_{\mathbf{MLE}}$ the MLE in the reduced model, which satisfies $\boldsymbol{U}_{\boldsymbol{MLE},(1)}(\tilde{\boldsymbol{\theta}}_{\mathbf{MLE}}) = \mathbf{0}$ with $\tilde{\boldsymbol{\theta}}_{\mathbf{MLE},(2)} = \mathbf{0}$. We can then define:

- The Wald test statistic, $W_n^2 := n\hat{\boldsymbol{\theta}}_{\mathbf{MLE},(2)}^{\top}\left(\boldsymbol{V}_{\mathbf{MLE}}\left(\hat{\boldsymbol{\theta}}_{\mathbf{MLE}}\right)_{(22)}\right)^{-1}\hat{\boldsymbol{\theta}}_{\mathbf{MLE},(2)}$;
- The score test statistic, $S_n^2 := n\boldsymbol{Z}_{\boldsymbol{n}}^{\top}\boldsymbol{V}_{\mathbf{MLE}}\left(\tilde{\boldsymbol{\theta}}_{\mathbf{MLE}}\right)_{(22)}\boldsymbol{Z}_{\boldsymbol{n}}$, with $\boldsymbol{Z}_{\boldsymbol{n}} := \boldsymbol{U}_{\mathbf{MLE},(2)}\left(\tilde{\boldsymbol{\theta}}_{\mathbf{MLE}}\right)$;
- The likelihood ratio test statistic, $L_n^2 := 2n\left[C_{\mathrm{MLE}}\left(\hat{\boldsymbol{\theta}}_{\mathbf{MLE}}\right) - C_{\mathrm{MLE}}\left(\tilde{\boldsymbol{\theta}}_{\mathbf{MLE}}\right)\right]$.

Using standard arguments, it follows from the asymptotic distribution of the MLE that these three tests are asymptotically $\chi^2$-distributed under the null and the alternative hypotheses—with a nonzero noncentrality parameter in the latter case. As is unfortunately common in the robustness literature, the issue of robust testing with randomly censored survival data has received much less attention than robust estimation, in spite of the fact that it is a crucial aspect of statistical inference. A recent exception is the work of Ghosh et al. (2017), where a robust version of the Wald test based on the minimum density power divergence approach of Basu et al. (2006) was studied. To the best of our knowledge, other contributions to the literature were mainly designed for a specific model; see, for example, Denecke and Müller (2014) for the Weibull case.

## 3 | PROPERTIES OF HJORT'S M-ESTIMATORS

In this section, we describe the main asymptotic properties of the class of M-estimators proposed by Hjort (1985), defined as the solution to (2).

### 3.1 | Fisher consistency

We start by addressing Fisher consistency. For this part of the analysis, we must work with the M-functional corresponding to the M-estimator when evaluated at the empirical distribution. Following Reid (1981) and Hjort (1992), we introduce the subdistribution functions of $X$ associated to $\Delta = 0$ and $\Delta = 1$, denoted by $F_X^0$ and $F_X^1$, respectively. These are simply defined by $F_X^{\delta}(x) = \mathbb{P}(X \leq x, \Delta = \delta)$, for $\delta \in \{0, 1\}$. We remark that (2) can be rewritten in terms of the empirical subdistribution functions, and that the M-estimator can be obtained as a solution to $\int_0^{\infty}\boldsymbol{\Psi}(t;\boldsymbol{\theta})\left\{d\hat{F}_X^1(t) - \alpha(t;\boldsymbol{\theta})\left[1 - \hat{F}_X^0(t) - \hat{F}_X^1(t)\right]dt\right\} = \mathbf{0}$. Thus, we can view the M-estimator as a functional, denoted by $\boldsymbol{S}$, evaluated at the empirical subdistribution functions, that is, $\hat{\boldsymbol{\theta}} = \boldsymbol{S}\left(\hat{F}_X^0, \hat{F}_X^1\right)$. Asymptotically, $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{S}\left(F_X^0, F_X^1\right)$, which solves

$$\int_0^{\infty}\boldsymbol{\Psi}(t;\boldsymbol{\theta})\left\{dF_X^1(t) - \alpha(t;\boldsymbol{\theta})\left[1 - F_X^0(t) - F_X^1(t)\right]dt\right\} = \mathbf{0}, \tag{4}$$

for $\boldsymbol{\theta}$. With this formulation, the M-functional is Fisher consistent in the following sense. Let $\left(F_X^{0,\boldsymbol{\theta}}, F_X^{1,\boldsymbol{\theta}}\right)$ denote any pair of subdistribution functions which are "consistent with the model" for the distribution of the actual survival times. We remark that we use the notation with $\boldsymbol{\theta}$ as a superscript—rather than as an argument as for $F_T$—to make it clear that these subdistribution functions are not fully parametrized by $\boldsymbol{\theta}$. Under the assumption of random censoring, such subdistribution functions are characterized by $F_X^{0,\boldsymbol{\theta}}(t) = \int_0^t$

$(1 - F_T(u;\boldsymbol{\theta}))dF_C(u)$ and $F_X^{1,\boldsymbol{\theta}}(t) = \int_0^t (1 - F_C(u))dF_T(u;\boldsymbol{\theta})$, for some censoring distribution $F_C$. We define Fisher consistency in this setting by the requirement that $\boldsymbol{\theta} = \boldsymbol{S}\left(F_X^{0,\boldsymbol{\theta}}, F_X^{1,\boldsymbol{\theta}}\right)\ \forall \boldsymbol{\theta} \in \Theta$. It is then trivial to notice that our M-functional satisfies this definition for any choice of $\boldsymbol{\Psi}$. Indeed,

$$\begin{aligned}\mathbf{0} &= \int_0^\infty \boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^{0,\boldsymbol{\theta}}, F_X^{1,\boldsymbol{\theta}}\right)\right)\left\{dF_X^{1,\boldsymbol{\theta}}(t) - \alpha\left(t; \boldsymbol{S}\left(F_X^{0,\boldsymbol{\theta}}, F_X^{1,\boldsymbol{\theta}}\right)\right)\left[1 - F_X^{0,\boldsymbol{\theta}}(t) - F_X^{1,\boldsymbol{\theta}}(t)\right]dt\right\} \\ &= \int_0^\infty \boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^{0,\boldsymbol{\theta}}, F_X^{1,\boldsymbol{\theta}}\right)\right)\left[1 - F_X^{0,\boldsymbol{\theta}}(t) - F_X^{1,\boldsymbol{\theta}}(t)\right]\left\{\frac{dF_T(t;\boldsymbol{\theta})}{1 - F_T(t;\boldsymbol{\theta})} - \frac{dF_T(t; \boldsymbol{S}(F_X^{0,\boldsymbol{\theta}}, F_X^{1,\boldsymbol{\theta}}))}{1 - F_T(t; \boldsymbol{S}(F_X^{0,\boldsymbol{\theta}}, F_X^{1,\boldsymbol{\theta}}))}\right\},\end{aligned}$$

where we used the fact that $dF_X^{1,\boldsymbol{\theta}}(t) = \left(1 - F_X^{0,\boldsymbol{\theta}}(t) - F_X^{1,\boldsymbol{\theta}}(t)\right)(1 - F_T(t;\boldsymbol{\theta}))^{-1}dF_T(t;\boldsymbol{\theta})$ for any censoring distribution (see e.g., van der Vaart (1998), p. 408), as well as the fact that $\alpha(t;\boldsymbol{\theta})dt = (1 - F_T(t;\boldsymbol{\theta}))^{-1}dF_T(t;\boldsymbol{\theta})$ by definition of the hazard function. It is obvious that $\boldsymbol{\theta} = \boldsymbol{S}\left(F_X^{0,\boldsymbol{\theta}}, F_X^{1,\boldsymbol{\theta}}\right)$ is a solution to this functional equation.

*Remark* 1. A similar argument for Fisher consistency, though in a quite different setting, was given in Assunção and Guttorp (1999). The key point is that the estimating equation is unbiased because it is a martingale at the model. The unbiasedness of the estimating equation leads to a Fisher consistent estimator (see e.g., Welsh, 1996, p. 191).

As a consequence, we do not need to do any recentering to achieve Fisher consistency. As argued above, this may be of importance in the presence of censoring.

## 3.2 | Consistency and asymptotic normality

The consistency of solutions to (2) in the context of censored survival data follows from the Fisher consistency property established above by assuming the continuity of the M-functional. Alternatively, it can be shown by adapting the arguments from the seminal paper by Huber (1967), to which we refer the reader for the appropriate results and proofs.

*Remark* 2. For a proof more in the spirit of the counting process approach followed by Hjort (1985), it would also be possible to adapt the proof of theorem 1 of Assunção and Guttorp (1999) to the present context. However, the conditions that they impose are much more stringent than what is required with the type of approach followed by Huber (1967). For instance, the matrix $\boldsymbol{P}(\boldsymbol{\theta}^*)$, defined below in Assumption 1, is required to be positive definite. As discussed in Andersen et al. (1993, pp. 442–443), consistency of M-estimators based on counting processes is difficult to prove without this kind of assumption.

Asymptotic normality is also easily obtained from the general theory of M-estimators. Since the matrices serving as building blocks for the standard sandwich formula of asymptotic variance will play an important role throughout this chapter, we believe that it is important to state the result formally. In doing so, we shall assume that the following conditions hold (see condition VI.2.1 in Andersen et al. (1993) for a similar set of conditions).

**Assumption 1.**

(a) There exists a neighborhood $\Theta^*$ of $\boldsymbol{\theta}^*$ such that, for all $\boldsymbol{\theta} \in \Theta^*$ and almost all $t \in \mathbb{R}_+$, the partial derivatives with respect to $\boldsymbol{\theta}$ of $\alpha(t;\boldsymbol{\theta})$ and $\Psi_j(t;\boldsymbol{\theta})$, $j \in \{1,\dots,p\}$, exist up to second order and are continuous in $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \Theta^*$.

(b) $\boldsymbol{U}(\boldsymbol{\theta})$ may be differentiated twice with respect to $\boldsymbol{\theta} \in \Theta^*$ by interchanging the order of integration and differentiation.

(c) There exist bounded functions $P_{jk}(\boldsymbol{\theta})$ and $Q_{jk}(\boldsymbol{\theta})$ defined on $\Theta^*$ such that for all $j,k \in \{1,\dots,p\}$,

$$\frac{1}{n}\int_0^\infty \Psi_j(t;\boldsymbol{\theta}^*)h_k(t;\boldsymbol{\theta}^*)\alpha(t;\boldsymbol{\theta}^*)Y(t)dt \xrightarrow{p} P_{jk}(\boldsymbol{\theta}^*),$$

and

$$\frac{1}{n}\int_0^\infty \Psi_j(t;\boldsymbol{\theta}^*)\Psi_k(t;\boldsymbol{\theta}^*)\alpha(t;\boldsymbol{\theta}^*)Y(t)dt \xrightarrow{p} Q_{jk}(\boldsymbol{\theta}^*),$$

as $n \to \infty$. Moreover, $\boldsymbol{P}(\boldsymbol{\theta}^*) = \left[P_{jk}(\boldsymbol{\theta}^*)\right]$ is nonsingular and $\boldsymbol{Q}(\boldsymbol{\theta}^*) = \left[Q_{jk}(\boldsymbol{\theta}^*)\right]$ is positive definite.

(d) For any $j \in \{1,\dots,p\}$ and $\epsilon > 0$, we have

$$\frac{1}{n}\int_0^\infty \Psi_j(t;\boldsymbol{\theta}^*)^2 \mathbb{1}_{\left\{\left|\frac{1}{\sqrt{n}}\Psi_j(t;\boldsymbol{\theta}^*)\right|>\epsilon\right\}}\alpha(t;\boldsymbol{\theta}^*)Y(t)dt \xrightarrow{p} 0,$$

as $n \to \infty$.

(e) There exist predictable processes $\{G_n(t), t \in \mathbb{R}_+\}$ and $\{H_n(t), t \in \mathbb{R}_+\}$, not depending on $\boldsymbol{\theta}$, such that for all $t \in \mathbb{R}_+$

$$\sup_{\boldsymbol{\theta}\in\Theta^*}\left|\frac{\partial^2}{\partial\theta_k\partial\theta_l}\Psi_j(t;\boldsymbol{\theta})\right| \le G_n(t),$$

and

$$\sup_{\boldsymbol{\theta}\in\Theta^*}\left|\frac{\partial^2}{\partial\theta_k\partial\theta_l}\left[\Psi_j(t;\boldsymbol{\theta})\alpha(t;\boldsymbol{\theta})\right]\right| \le H_n(t),$$

for all $j,k,l \in \{1,\dots,p\}$. Moreover,

$$\frac{1}{n}\int_0^\infty G_n(t)\alpha(t;\boldsymbol{\theta}^*)Y(t)dt \xrightarrow{p} C_1 < \infty,$$

$$\frac{1}{n}\int_0^\infty H_n(t)Y(t)dt \xrightarrow{p} C_2 < \infty,$$

for some constants $C_1$ and $C_2$, and for any $\epsilon > 0$,

$$\frac{1}{n}\int_0^\infty G_n(t)\mathbb{1}_{\left\{\sqrt{\frac{G_n(t)}{n}}>\epsilon\right\}}\alpha(t;\boldsymbol{\theta}^*)Y(t)dt \xrightarrow{p} 0,$$

as $n \to \infty$.

We briefly comment on these conditions. Conditions (a) and (b) ensure the validity of the Taylor expansions used in the proof of asymptotic normality. Condition (c) essentially guarantees the existence and natural properties of the matrices which will appear in the expression for the asymptotic variance. Condition (d) is a (Lindeberg-type) condition required for the use of the martingale central limit theorem. Finally, condition (e) guarantees that the remainder term in the Taylor expansion remains under control. We note that these conditions are sufficient but need not be necessary. Under these conditions, asymptotic normality of the M-estimator is established in the following proposition.

**Proposition 1.** *Suppose that Assumption 1 holds, and let* $\{\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}, n \in \mathbb{N}\}$ *be a consistent sequence of estimators of* $\boldsymbol{\theta}^*$. *Then*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{P}(\boldsymbol{\theta}^*)^{-1}\boldsymbol{Q}(\boldsymbol{\theta}^*)\boldsymbol{P}(\boldsymbol{\theta}^*)^{-\top}\right),$$

*with* $\boldsymbol{P}(\boldsymbol{\theta}^*) = \int_0^\infty \boldsymbol{\Psi}(t;\boldsymbol{\theta}^*)\boldsymbol{h}(t;\boldsymbol{\theta}^*)^\top \alpha(t;\boldsymbol{\theta}^*)y(t)dt$ *and* $\boldsymbol{Q}(\boldsymbol{\theta}^*) = \int_0^\infty \boldsymbol{\Psi}(t;\boldsymbol{\theta}^*)\boldsymbol{\Psi}(t;\boldsymbol{\theta}^*)^\top \alpha(t;\boldsymbol{\theta}^*)y(t)dt$.

Proposition 1 shows that the asymptotic variance of the M-estimator has the usual sandwich form, and reduces to the variance of the MLE if the logarithmic derivative of the hazard with respect to the parameter is chosen as the function $\boldsymbol{\Psi}$.

## 3.3 | Influence function

We now investigate the form of the influence function, which will play a central role in the remainder of this paper. The influence function can be interpreted as the Gâteaux derivative of the M-functional $\boldsymbol{S}(F_X^0, F_X^1)$ solving (4) in the direction of a point mass distribution at some point $(\tilde{x}, \tilde{\delta})$; we refer the reader to the seminal contributions of Hampel (1968, 1974) for more details. We consider a contamination of the joint distribution of $(X, \Delta)$ of the type $F_{X,\Delta}^{\epsilon}(x,\delta) = (1-\epsilon)F_{X,\Delta}(x,\delta) + \epsilon \mathbb{1}_{\{x\geq\tilde{x},\delta\geq\tilde{\delta}\}}(x,\delta)$, where $0 < \epsilon < 0.5$ is a fixed number and $\mathbb{1}_{\{x\geq\tilde{x},\delta\geq\tilde{\delta}\}}$ denotes a point mass distribution at the point $(\tilde{x},\tilde{\delta})$. Under such a contamination scheme, the implied contaminated subdistributions are of the form

$$\begin{aligned} F_X^{0,\epsilon}(x) &= (1-\epsilon)F_X^0(x) + \epsilon(1-\tilde{\delta})\mathbb{1}_{\{x\geq\tilde{x}\}}(x) \\ F_X^{1,\epsilon}(x) &= (1-\epsilon)F_X^1(x) + \epsilon\tilde{\delta}\mathbb{1}_{\{x\geq\tilde{x}\}}(x). \end{aligned} \tag{5}$$

*Remark* 3. We note that it is always possible to relate these contaminated subdistributions to the corresponding contaminated distribution for the actual survival times. Indeed, following van der Vaart (1998, p. 407), we have

$$F_T^{\epsilon}(x) = 1 - \prod_{0\leq t\leq x}\left[1-\Lambda\{t\}\right]\exp\left(-\Lambda^c(x)\right),$$

where $\prod_{0\leq t\leq x}$ denotes a product-integral (Gill & Johansen, 1990) on $[0,x]$ and

$$\Lambda(x) = \int_0^x \frac{dF_X^{1,\epsilon}(t)}{1 - F_X^{0,\epsilon}(t) - F_X^{1,\epsilon}(t)},$$

with $\Lambda\{t\}$ denoting the jump of $\Lambda$ at t and $\Lambda^c$ denoting the continuous part of $\Lambda$.

We have the following generalization of the result given in Hjort (1992, pp. 372–373).

**Proposition 2.** *Consider the contamination scheme (5) at some given point* $(\tilde{x}, \tilde{\delta})$*. The influence function of the functional* $\boldsymbol{S}$ *at a pair of subdistribution functions* $(F_X^0, F_X^1)$ *is given by*

$$\boldsymbol{IF}\left(\boldsymbol{S}, \left(F_X^0, F_X^1\right), (\tilde{x}, \tilde{\delta})\right) = \boldsymbol{P}\left(F_X^0, F_X^1\right)^{-1} \boldsymbol{\Xi}\left(\tilde{x}, \tilde{\delta}; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right), \tag{6}$$

*with*

$$\boldsymbol{\Xi}\left(\tilde{x}, \tilde{\delta}; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right) = \tilde{\delta}\boldsymbol{\Psi}\left(\tilde{x}; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right) - \int_0^{\tilde{x}} \boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\alpha\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)dt, \tag{7}$$

*and*

$$\begin{aligned}\boldsymbol{P}\left(F_X^0, F_X^1\right) = &\int_0^{\infty} \boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\boldsymbol{h}\left(t; \boldsymbol{S}(F_X^0, F_X^1)\right)^{\top}\alpha\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\left[1 - F_X^0(t) - F_X^1(t)\right]dt \\ &- \int_0^{\infty} \nabla_{\theta^{\top}}\boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\left\{dF_X^1(t) - \alpha\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\left[1 - F_X^0(t) - F_X^1(t)\right]dt\right\},\end{aligned}$$

*where* $\nabla_{\theta^{\top}}\boldsymbol{\Psi}(t; \cdot)$ *denotes the gradient of* $\boldsymbol{\Psi}(t; \cdot)$*. In particular, by the Fisher consistency of* $\boldsymbol{S}$*, we have* $\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{P}\left(F_X^{0,\boldsymbol{\theta}}, F_X^{1,\boldsymbol{\theta}}\right) = \int_0^{\infty} \boldsymbol{\Psi}(t; \boldsymbol{\theta})\boldsymbol{h}(t; \boldsymbol{\theta})^{\top} dF_X^{1,\boldsymbol{\theta}}(t)$.

Some comments are in order. First of all, we note that the influence function is proportional to $\boldsymbol{\Xi}$ as given in (7). This is the quantity that should be bounded in order to have a robust estimator in the infinitesimal sense. Clearly, the influence function associated to the MLE is unbounded in $(\tilde{x}, \tilde{\delta})$ for most parametric models of interest, as can be verified by setting $\boldsymbol{\Psi}(\cdot; \boldsymbol{\theta}) = \boldsymbol{h}(\cdot; \boldsymbol{\theta})$. Second, we remark that the matrix $\boldsymbol{P}$ was obviously already encountered in Proposition 1 as part of the expression of the asymptotic variance. This is a general property of M-estimators, which is also seen to hold in this setting. Finally, it is interesting to notice that the last expression for $\boldsymbol{P}$ given in Proposition 2 is an integral with respect to the subdistribution function for uncensored observations only. A similar expression can be derived in the same manner for $\boldsymbol{Q}$, defined in Proposition 1. This can be viewed as an application of the isometry with randomly censored data studied by Sasieni (1992) and Janssen (1994).

## 4 | ROBUST APPROACH

Now that the main properties of the class of M-estimators under consideration are well understood, we can address the robustness issue. We begin by considering robust estimation, before using our results in the testing context.

### 4.1 | Robust estimation

As noted above, a robust estimator in the sense of Hampel et al. (1986) is generally characterized by a bounded influence function. This implies that (6) should be bounded, in some metric, for

any $(\tilde{x}, \tilde{\delta})$. In problems with multidimensional parameters, various measures of the magnitude of the influence of a point can be considered. We shall consider two such measures in what follows, namely, the unstandardized and self-standardized sensitivities. In each case, we will determine the optimal B-robust estimator (OBRE), that is, the best estimator (in some sense to be clarified below) given the bound on the sensitivity.

### 4.1.1 | Unstandardized sensitivity

We start with the case of unstandardized sensitivity, which is probably the most straightforward extension of the one-dimensional parameter case. It is defined as $\gamma_U := \sup_{(x,\delta)\in\mathbb{R}_+\times\{0,1\}} \left\| \boldsymbol{IF}\left(\boldsymbol{S}, \left(F_X^0, F_X^1\right), (x,\delta)\right)\right\|$. From the inspection of (6) and (7), it is clear that $\gamma_U$ is bounded if and only if

$$\sup_{(x,\delta)\in\mathbb{R}_+\times\{0,1\}} \left\| \delta\boldsymbol{\Psi}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}(t;\boldsymbol{\theta})\alpha(t;\boldsymbol{\theta})dt \right\| \leq c, \tag{8}$$

for a suitably chosen constant $c$. Specifically, a smaller value of $c$ will lead to a more robust estimator, but below a certain point, there may be no desirable $\boldsymbol{\Psi}$ function satisfying (8). The issue of adequately selecting the constant bounding the sensitivity will be addressed below for the case of self-standardized sensitivity, where more results are available; see Hampel et al. (1986, p. 252). Our first task in this robustness analysis is to characterize the class of $\boldsymbol{\Psi}$ functions satisfying (8). To do so, we will need to introduce two operators, $R$ and $L$, allowing us to relate functions acting as substitutes for the logarithmic derivative of the hazard with respect to the parameter to score-type functions of the parametric model. The following results were given by Ritov and Wellner (1988) and Efron and Johnstone (1990); see also Bickel et al. (1993, pp. 420–424) for a summary. We omit $\boldsymbol{\theta}$, which remains fixed, in the statement of these results.

**Lemma 1.** *Let* $\boldsymbol{u}, \boldsymbol{v} \in L_2(F_T) := \{\boldsymbol{w} \mid \int_{-\infty}^{\infty} \|\boldsymbol{w}(s)\|^2 dF_T(s) < \infty\}$. *Consider the linear operators* $R : L_2(F_T) \to L_2(F_T)$ *and* $L : L_2(F_T) \to L_2(F_T)$ *acting as* $R\boldsymbol{u}(t) := \boldsymbol{u}(t) - (1 - F_T(t))^{-1} \int_t^{\infty} \boldsymbol{u}(s) f_T(s) ds$ *and* $L\boldsymbol{v}(t) := \boldsymbol{v}(t) - \int_0^t \boldsymbol{v}(s)\alpha(s)ds$, *respectively. Then the following properties hold:*

*(a)* $(R \circ L)\boldsymbol{v} = \boldsymbol{v}$;
*(b)* $(L \circ R)\boldsymbol{u} = \boldsymbol{u} - \int_{-\infty}^{\infty} \boldsymbol{u}(s) f_T(s) ds$;
*(c)* $\mathrm{im}(R) = L_2(F_T)$;
*(d)* $\mathrm{im}(L) = L_2^0(F_T)$,

*where* $\mathrm{im}(\cdot)$ *denotes the image of a linear operator and* $L_2^0(F_T) := \{\boldsymbol{w} \in L_2(F_T) \mid \int_{-\infty}^{\infty} \boldsymbol{w}(s) dF_T(s) = \boldsymbol{0}\}$.

By Lemma 1, any function $\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta}) \in L_2(F_T)$ is the image of a function $\boldsymbol{\Phi}(\cdot;\boldsymbol{\theta}) \in L_2^0(F_T)$ through the operator $R$. This implies that we can rewrite the norm in (8) as

$$\left\| \delta\boldsymbol{\Psi}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}(t;\boldsymbol{\theta})\alpha(t;\boldsymbol{\theta})dt \right\| = \left\| \delta\boldsymbol{\Phi}(x;\boldsymbol{\theta}) + \frac{1-\delta}{1 - F_T(x;\boldsymbol{\theta})} \int_x^{\infty} \boldsymbol{\Phi}(t;\boldsymbol{\theta}) f_T(t;\boldsymbol{\theta}) dt \right\|. \tag{9}$$

*Remark* 4. We note that the expression within the norm in (9) can be viewed as a James-type estimating function, where the second term is essentially a conditional expectation; see James (1986), in particular equation (3), p. 36. However, the approach followed by James (1986) is different, since he then uses a nonparametric estimator of the conditional expectation based on the product-limit estimator. This leads to an estimator which is equivalent to the one studied by Wang (1999); see our comments in Section 2.1.

From (9), it is straightforward to obtain a simple robustness criterion, as can be seen from the following lemma.

**Lemma 2.** *Let* $\boldsymbol{\Phi}(\cdot;\boldsymbol{\theta}) \in L_2^0(F_T)$. *Then the following statements are equivalent:*

$$(a) \quad \left\| \delta \boldsymbol{\Phi}(x;\boldsymbol{\theta}) + \frac{1-\delta}{1-F_T(x;\boldsymbol{\theta})} \int_x^\infty \boldsymbol{\Phi}(t;\boldsymbol{\theta}) f_T(t;\boldsymbol{\theta}) dt \right\| \leq c \ \forall (x,\delta);$$

$$(b) \quad \|\boldsymbol{\Phi}(x;\boldsymbol{\theta})\| \leq c \ \forall x.$$

Lemma 2 leads to the following observation: any $\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta}) \in L_2(F_T)$ which is the image, through the operator $R$, of some $\boldsymbol{\Phi}(\cdot;\boldsymbol{\theta}) \in L_2^0(F_T)$ satisfying $\sup_x \|\boldsymbol{\Phi}(x;\boldsymbol{\theta})\| \leq c$ leads to an estimator with a bounded unstandardized sensitivity. This result has a very natural interpretation, and shows that the relevant quantity to bound is still the score function of the parametric model, even in the presence of censoring. Making use of this result, we now address the issue of finding the optimal B-robust estimator with unstandardized sensitivity. Following Hampel et al. (1986, p. 238) we consider the trace of the asymptotic covariance matrix as the measure of efficiency. Therefore, the problem that we wish to solve can be formulated as follows.

**Problem 1.**

$$\min_{\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta})} \operatorname{tr} \left( \int \left[ \delta \boldsymbol{\Psi}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}(t;\boldsymbol{\theta}) \alpha(t;\boldsymbol{\theta}) dt \right] \left[ \delta \boldsymbol{\Psi}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}(t;\boldsymbol{\theta}) \alpha(t;\boldsymbol{\theta}) dt \right]^\top dF_{X,\Delta}^{\boldsymbol{\theta}}(x,\delta) \right)$$

$$s.t. \quad \sup_{(x,\delta) \in \mathbb{R}_+ \times \{0,1\}} \left\| \delta \boldsymbol{\Psi}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}(t;\boldsymbol{\theta}) \alpha(t;\boldsymbol{\theta}) dt \right\| \leq c$$

$$\int \left[ \delta \boldsymbol{\Psi}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}(t;\boldsymbol{\theta}) \alpha(t;\boldsymbol{\theta}) dt \right] \left[ \delta \boldsymbol{h}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{h}(t;\boldsymbol{\theta}) \alpha(t;\boldsymbol{\theta}) dt \right]^\top dF_{X,\Delta}^{\boldsymbol{\theta}}(x,\delta) = \boldsymbol{I_p}.$$

Our objective is thus to minimize the trace of the asymptotic variance at the model, subject to a bound on the Euclidean norm of the influence function. Since $\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta})$ is only determined up to multiplication by a matrix, without loss of generality, we set the matrix $\boldsymbol{P}(\boldsymbol{\theta})$ to be the identity matrix $\boldsymbol{I_p}$. Taken together, the two constraints then imply that the unstandardized sensitivity is bounded by $c$. The following proposition states the solution to Problem 1.

**Proposition 3.** *The OBRE in the unstandardized case is obtained by setting*

$$\boldsymbol{\Psi}_U(x;\boldsymbol{\theta}) = \boldsymbol{H_c}\left(\boldsymbol{A}(\boldsymbol{\theta})\left[\boldsymbol{s}(x;\boldsymbol{\theta}) - \boldsymbol{a}(\boldsymbol{\theta})\right]\right) - \frac{1}{1-F_T(x;\boldsymbol{\theta})} \int_x^\infty \boldsymbol{H_c}\left(\boldsymbol{A}(\boldsymbol{\theta})\left[\boldsymbol{s}(t;\boldsymbol{\theta}) - \boldsymbol{a}(\boldsymbol{\theta})\right]\right) f_T(t;\boldsymbol{\theta}) dt,$$

*where* $\boldsymbol{H_c}(\boldsymbol{r}) := \boldsymbol{r}\min\{1, c/\|\boldsymbol{r}\|\}$ *is the multivariate Huber function with tuning parameter* $c$*, and* $\boldsymbol{A}(\boldsymbol{\theta})$ *and* $\boldsymbol{a}(\boldsymbol{\theta})$ *are, respectively, a* $p \times p$ *matrix and a* $p \times 1$ *vector solving*

$$\int_0^\infty \boldsymbol{\Psi_U}(t;\boldsymbol{\theta})\boldsymbol{h}(t;\boldsymbol{\theta})^\top dF_X^{1,\boldsymbol{\theta}}(t) = \boldsymbol{I_p},$$

*and*

$$\int_0^\infty \boldsymbol{H_c}\left(\boldsymbol{A}(\boldsymbol{\theta})\left[\boldsymbol{s}(t;\boldsymbol{\theta}) - \boldsymbol{a}(\boldsymbol{\theta})\right]\right) f_T(t;\boldsymbol{\theta})dt = \boldsymbol{0}.$$

The result of Proposition 3 appears as an expected consequence of the result of Lemma 2. The function $\boldsymbol{\Psi_U}$ leading to the unstandardized OBRE is obtained by applying the operator $R$ to the Huberized scaled and centered likelihood score function. We note that the recentering constant vector $\boldsymbol{a}(\boldsymbol{\theta})$ can be seen as dictated by a Fisher consistency requirement *in the space of score functions* of the parametric model. Importantly, this does not contradict our claim for the absence of a recentering constant vector when viewing the estimating function in terms of logarithmic derivatives of hazard functions. The parallel between the optimality result of Proposition 3 and the corresponding optimality result in the absence of censoring is clear. In addition, as in the uncensored case, the unstandardized OBRE has the natural property that when $c$ is sufficiently large, the OBRE is equivalent to the MLE. Indeed, as $c$ tends to infinity, $\boldsymbol{a}(\boldsymbol{\theta})$ and $\boldsymbol{A}(\boldsymbol{\theta})$ become the null vector and the identity matrix, respectively, and $\boldsymbol{\Psi_U}$ reduces to the logarithmic derivative of the hazard. On the other hand, for a small enough value of $c$, the constraint on the influence function becomes binding and the two estimators differ.

### 4.1.2 | Self-standardized sensitivity

So far, we have focused on the case of unstandardized sensitivity. However, it is well known that this measure is not invariant to scale transformations of individual parameters. To overcome that issue, we now consider the self-standardized sensitivity, defined as

$$\gamma_{SS} := \sup_{(x,\delta)\in\mathbb{R}_+\times\{0,1\}} \left[\boldsymbol{IF}\left(\boldsymbol{S}, \left(F_X^0, F_X^1\right), (x,\delta)\right)^\top \boldsymbol{V}\left(\boldsymbol{S}, \left(F_X^0, F_X^1\right)\right)^{-1} \boldsymbol{IF}\left(\boldsymbol{S}, \left(F_X^0, F_X^1\right), (x,\delta)\right)\right]^{\frac{1}{2}}, \quad (10)$$

where $\boldsymbol{V}\left(\boldsymbol{S}, \left(F_X^0, F_X^1\right)\right)$ is the asymptotic covariance matrix, assumed to be nonsingular. The influence function is thus measured in the metric given by the asymptotic covariance matrix of the estimator; see Hampel et al. (1986, p. 228). Building on our results in the unstandardized case, we now look for the OBRE in the self-standardized case. In what follows, it will be important to make the dependence of all quantities on the underlying $\boldsymbol{\Psi}$ function explicit. To achieve this, we will complement our usual notation with a subscript $\boldsymbol{\Psi}$; in particular, $\boldsymbol{\Psi_{SS}}$ refers to the $\boldsymbol{\Psi}$ function leading to the OBRE in the self-standardized case. As argued by Hampel et al. (1986, p. 243), the asymptotic mean squared error that we wish to minimize should be measured in the same metric as the sensitivity. Our problem can thus be formulated as follows.

**Problem 2.**

$$\min_{\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta})} \operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\Psi}}\boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}^{-1}\right)$$

$$s.t. \sup_{(x,\delta)\in\mathbb{R}_+\times\{0,1\}} \left(\delta\boldsymbol{\Psi}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}(t;\boldsymbol{\theta})\alpha(t;\boldsymbol{\theta})dt\right)^\top \boldsymbol{P}_{\boldsymbol{\Psi}}^{-\top}\boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}^{-1}\boldsymbol{P}_{\boldsymbol{\Psi}}^{-1}\left(\delta\boldsymbol{\Psi}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}(t;\boldsymbol{\theta})\alpha(t;\boldsymbol{\theta})dt\right) \leq c^2.$$

Even though the objective was written in a compact way, it is obviously equivalent to minimizing the trace of

$$\int \boldsymbol{P}_{\boldsymbol{\Psi}}^{-1}\left(\delta\boldsymbol{\Psi}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}(t;\boldsymbol{\theta})\alpha(t;\boldsymbol{\theta})dt\right)\boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}^{-1}\left(\delta\boldsymbol{\Psi}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}(t;\boldsymbol{\theta})\alpha(t;\boldsymbol{\theta})dt\right)^\top \boldsymbol{P}_{\boldsymbol{\Psi}}^{-\top} dF_{X,\Delta}^{\boldsymbol{\theta}}(x,\delta),$$

which, upon setting $\boldsymbol{P}_{\boldsymbol{\Psi}}$ to the identity matrix, is seen to coincide with the objective of Problem 1 up to the standardization by $\boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}$. The formulation of Problem 2 may look somewhat peculiar at first, since the conditions for the optimality of $\boldsymbol{\Psi}_{SS}$ depend on $\boldsymbol{\Psi}_{SS}$ itself through $\boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}$ (Stefanski et al., 1986). However, the introduction of this standardization leads to the desirable invariance property mentioned above. The following proposition states our optimality result in the self-standardized case.

**Proposition 4.** *The OBRE in the self-standardized case is obtained by setting*

$$\boldsymbol{\Psi}_{SS}(x;\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{\Psi}_{\boldsymbol{B}}(x;\boldsymbol{\theta}), \tag{11}$$

*with*

$$\boldsymbol{\Psi}_{\boldsymbol{B}}(x;\boldsymbol{\theta}) = [\boldsymbol{s}(x;\boldsymbol{\theta}) - \boldsymbol{a}(\boldsymbol{\theta})]\, w_c(x;\boldsymbol{\theta}) - \frac{1}{1-F_T(x;\boldsymbol{\theta})}\int_x^\infty [\boldsymbol{s}(t;\boldsymbol{\theta}) - \boldsymbol{a}(\boldsymbol{\theta})]\, w_c(t;\boldsymbol{\theta}) f_T(t;\boldsymbol{\theta})dt,$$

*where $w_c(\cdot;\boldsymbol{\theta})$ is a weighting function defined by*

$$w_c(t;\boldsymbol{\theta}) := \min\left\{1, \frac{c}{\|\boldsymbol{A}(\boldsymbol{\theta})\,[\boldsymbol{s}(t;\boldsymbol{\theta}) - \boldsymbol{a}(\boldsymbol{\theta})]\|}\right\}, \tag{12}$$

*at time t, and the $p \times p$ matrix $\boldsymbol{A}(\boldsymbol{\theta})$ and $p \times 1$ vector $\boldsymbol{a}(\boldsymbol{\theta})$ are implicit solutions to*

$$\int_0^\infty \boldsymbol{\Psi}_{SS}(t;\boldsymbol{\theta})\boldsymbol{\Psi}_{SS}(t;\boldsymbol{\theta})^\top dF_X^{1,\boldsymbol{\theta}}(t) = \boldsymbol{I}_{\boldsymbol{p}}, \tag{13}$$

*and*

$$\int_0^\infty [\boldsymbol{s}(t;\boldsymbol{\theta}) - \boldsymbol{a}(\boldsymbol{\theta})]\, w_c(t;\boldsymbol{\theta}) f_T(t;\boldsymbol{\theta})dt = \boldsymbol{0}. \tag{14}$$

Unsurprisingly, the general structure of the OBRE in the self-standardized case is the same as in the unstandardized case. For the reason mentioned above (invariance to scale transformations), we prefer to use this estimator rather than the unstandardized version.

We note that one potential issue in the computation of the OBRE is that the left-hand side of (13), which is used to determine the matrix $\boldsymbol{A}(\boldsymbol{\theta})$ (for a given $\boldsymbol{a}(\boldsymbol{\theta})$), does involve the censoring distribution. We propose to overcome this problem by replacing the left-hand side of (13) by its empirical counterpart. Provided that the sample size is not too small, this should not have a large quantitative impact on the estimates. All the results presented in this paper have been obtained in this way, without encountering any convergence problems. We remark that a similar strategy has been used several times in the robustness literature; see, for example, Mancini et al. (2005) and La Vecchia and Trojani (2010). A full description of the algorithm used to compute the OBRE in the self-standardized case is provided in the Supporting Information.

We close this subsection by giving some general guidelines for picking the tuning constant $c$, even though we emphasize that the specific choice of $c$ should always be made by the researcher after careful consideration of the problem at hand. As a starting point, we note that $c$ cannot be chosen smaller than $\sqrt{p}$, as the following proposition shows.

**Proposition 5.** *Let $\gamma_{SS}$ denote the self-standardized sensitivity as defined in (10). Then we have $\gamma_{SS}^2 \geq p$.*

Beyond that, a natural idea is to select a value for $c$ so as to limit the influence of outliers while maintaining a sufficiently high level of efficiency at the model. Let us assume for concreteness that our objective is a 90% asymptotic relative efficiency (ARE) at the model. In that case, we can pick $c$ in such a way that $0.90 = \mathrm{tr}\left(\boldsymbol{V}_{\mathbf{MLE}}\right) / \mathrm{tr}\left(\boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}\right)$. Figure 2 illustrates this approach for the case of the Weibull(2,5) distribution which will be studied in Section 5.1, for two different levels of censoring. The covariance matrices of the OBRE and the MLE were estimated empirically based on a simulated sample of 10,000 observations.

According to this criterion, we should select a value of $c$ which is around 3.5. However, in our experience, the use of such a criterion leads to estimators which are not very robust. This point has already been made in various contexts; see, for example, Hampel et al. (1986, p. 252), for the

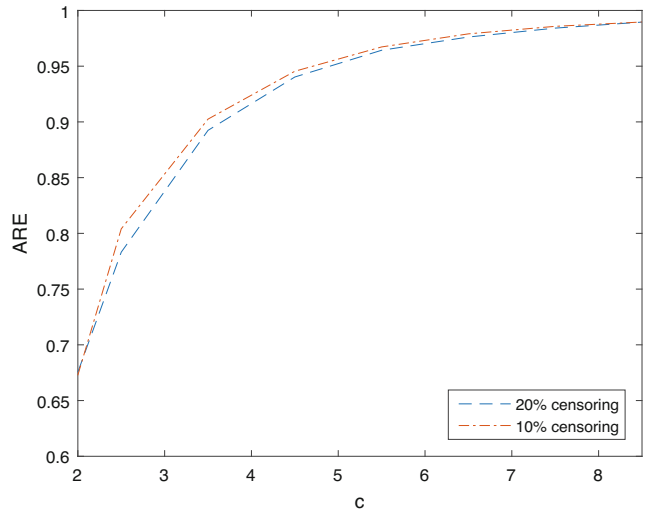


**FIGURE 2** Asymptotic efficiency of the optimal B-robust estimator relative to the maximum likelihood estimator for the Weibull(2,5) distribution at different levels of the tuning parameter

regression case. In fact, these authors suggested that a value of $c$ near its lower bound should be chosen in many applications, and we also tend to adopt this point of view. We note that a different type of approach to the selection of the tuning parameter, motivated by testing, has been proposed by Ronchetti and Trojani (2001); see also Mancini et al. (2005) for more details.

## 4.2 | Robust testing

Building on the results obtained for robust estimation, we now address robust testing. Consider the same setting as in Section 2.2. Let $\hat{\boldsymbol{\theta}}$ denote a robust M-estimator in the full model and $\tilde{\boldsymbol{\theta}}$ the corresponding M-estimator in the reduced model. In the spirit of Heritier and Ronchetti (1994), we introduce the following robust versions of the classical tests presented in Section 2.2:

- A Wald-type test statistic, $W^2_{R,n} := n\hat{\boldsymbol{\theta}}_{(2)}^\top \left( \boldsymbol{V}\left(\hat{\theta}\right)_{(22)} \right)^{-1} \hat{\boldsymbol{\theta}}_{(2)}$, where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{P}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{Q}(\hat{\boldsymbol{\theta}})\boldsymbol{P}(\hat{\boldsymbol{\theta}})^{-\top}$ is the estimated asymptotic variance of the M-estimator;
- A score-type test statistic, $S^2_{R,n} := n\boldsymbol{Z}^\top_{\boldsymbol{R},\boldsymbol{n}} \left( \boldsymbol{P}\left(\tilde{\theta}\right)_{(22.1)} \boldsymbol{V}\left(\tilde{\theta}\right)_{(22)} \boldsymbol{P}\left(\tilde{\theta}\right)^\top_{(22.1)} \right)^{-1} \boldsymbol{Z}_{\boldsymbol{R},\boldsymbol{n}}$, with $\boldsymbol{Z}_{\boldsymbol{R},\boldsymbol{n}} := \boldsymbol{U}_{(2)}\left(\tilde{\boldsymbol{\theta}}\right)$ and $\boldsymbol{P}\left(\tilde{\theta}\right)_{(22.1)} := \boldsymbol{P}\left(\tilde{\theta}\right)_{(22)} - \boldsymbol{P}\left(\tilde{\theta}\right)_{(21)} \boldsymbol{P}\left(\tilde{\theta}\right)^{-1}_{(11)} \boldsymbol{P}\left(\tilde{\theta}\right)_{(12)}$;
- A likelihood ratio-type test statistic, $L^2_{R,n} := 2n\left[\rho(\hat{\boldsymbol{\theta}}) - \rho(\tilde{\boldsymbol{\theta}})\right]$, with $\rho$ such that $\rho(\mathbf{0}) = 0$ and $\nabla_{\boldsymbol{\theta}^\top}\rho(\boldsymbol{\theta}) = \boldsymbol{U}(\boldsymbol{\theta})$.

Heritier and Ronchetti (1994) showed that the Wald-type and score-type test statistics follow a noncentral chi-square distribution, with noncentrality parameter equal to zero under the null hypothesis. The properties of the likelihood ratio-type test, on the other hand, are more difficult to characterize. First, as shown in Heritier and Ronchetti (1994), the asymptotic distribution of the likelihood ratio-type test statistic is different from the one of the other two test statistics, unlike in the classical case. Second, this test requires defining an appropriate $\rho$ function, which is not a trivial task given that it requires integrating $\boldsymbol{U}(\cdot)$ over the parameter space $\Theta$. We leave the study of the likelihood ratio-type test for future research. An important result given by Heritier and Ronchetti (1994) is that optimal bounded-influence tests are obtained from optimal self-standardized B-robust estimators in the context of M-estimators in parametric models with complete information. We conjecture that the same conclusion applies in our framework in the presence of censoring, under similar conditions to the ones given in Appendix A.2 of the cited article.

# 5 | SIMULATION STUDY

In order to assess the practical relevance of our robust approach, we first perform a simulation study. We shall consider two families of distributions, namely the Weibull and the log-normal models. In each case, we construct $N = 500$ samples of size $n$ in the following way. We simulate the actual survival times $t_i$ from a particular member of the family and, independently, censoring times $c_i$ from a given censoring distribution. The distribution of the censoring times is chosen in such a way that a predetermined expected censoring proportion, denoted by $p_c$, is achieved. The observed simulated data are then given by $x_i = \min(t_i, c_i)$ and $\delta_i = \mathbb{1}_{\{t_i = x_i\}}$, for $i = 1, \ldots, n$.

In terms of estimation, we mainly examine the performance of the OBRE relative to the MLE, with and without contamination. For the Weibull model, we additionally compare the OBRE to the $L_2E$ proposed by Yang and Scott (2013), which is obtained by minimizing the $L_2$ distance to the unknown density of survival times. The latter estimator is a special case of the minimum density power divergence approach of Basu et al. (2006), with $\alpha = 1$ in the notation of Section 2.1. For our purposes, the $L_2E$ method was implemented in Matlab. The function fmincon was used, with settings matching those of R's nlminb function used by Yang and Scott (2013) as closely as possible.

Regarding the contamination, we assume that a fraction $\epsilon$ of our observed simulated data come from some contaminating distribution. In general, this means that we consider a neighborhood of the joint distribution of $(X, \Delta)$. Actually, our preliminary analyses revealed that contaminating the censoring indicator (e.g., by random 0/1 switching) had little effect on both the OBRE and the MLE. This seems to be in line with the result given in Lemma 2, which shows that selecting a bounded function of the censored survival times only is sufficient to yield a robust estimator. Therefore, for the results that we present below, we focus on a contamination of $X$ of the form $F_X^{\epsilon}(x) = (1-\epsilon)F_X(x) + \epsilon G_X(x)$, where $G_X$ is an arbitrary contaminating distribution. We look at the impact of various aspects of the problem at hand on our results. For both models, we consider two sample sizes ($n \in \{50,200\}$), two expected censoring proportions ($p_c \in \{0.1, 0.2\}$), and two contamination levels ($\epsilon \in \{0, 0.05\}$). The type of censoring distribution and the type of contaminating distribution used in each simulation are also different; this will be explained in more detail below. We are interested in the general performance of the estimators in terms of bias, variance, and mean squared error, as well as in the construction of confidence intervals. For the sake of readability, the presentation of the latter results is deferred to Data S1.

In terms of testing, we are mainly interested in the level under contamination of the robust and classical Wald and score tests, but we also explore the power of the score tests in the simulations for the log-normal model.

## 5.1 | Weibull model

In the first example, we consider the Weibull distribution with scale and shape parameters $\lambda$ and $\kappa$, meaning that the density of the actual survival times is $f_T(t; \lambda, \kappa) = \kappa t^{\kappa-1}/\lambda^{\kappa} \exp\left[-(t/\lambda)^{\kappa}\right]$, $t \geq 0$. We set $\lambda = 2$ and $\kappa = 5$. The censoring times are generated from an exponential distribution with parameter $\theta$, that is, $f_C(t; \theta) = \theta \exp[-\theta t]$, $t \geq 0$. Given the values of $\lambda$ and $\kappa$, the value of $\theta$ is uniquely determined by the expected censoring proportion $p_c$. For instance, when $p_c = 0.1$, we have $\theta = 0.0575$. Regarding contamination, we choose $G_X$ to be a point mass distribution at the point 3, which is above the 99th percentile of the actual survival time distribution. We set the tuning parameter to $c = 1.5$. In this two-parameter model, such a value of $c$ essentially corresponds to the lower bound of $\sqrt{2}$ derived in Proposition 5 and is consistent with our comments at the end of Section 4.1.2. The results depicted in Figures 3 and 4 thus illustrate a situation in which the researcher is primarily interested in controlling the bias under contamination, at the expense of increased variance at the model.

In the absence of contamination (Figure 3), the MLE outperforms the OBRE. Indeed, as expected, the latter features a greater variability. In addition, the OBRE is slightly biased. We note that, from a theoretical standpoint, the OBRE is only guaranteed to be consistent. In our experience, the convergence is often slower for estimators of scale or shape than for estimators of the mean, for instance. In finite samples, the bias is assured to be bounded in a full neighborhood of the model and may exceed the bias of the MLE at the model, especially for a small value of

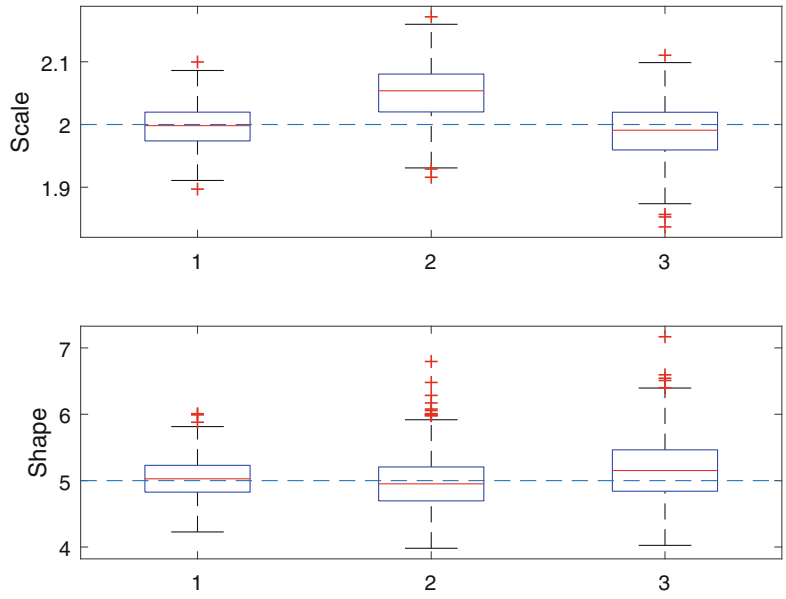


**FIGURE 3** Boxplots of maximum likelihood estimator (left), $L_2E$ (middle) and optimal B-robust estimator (right) for the scale (top) and shape (bottom) parameters with $n = 200$, $p_c = 0.2$, $\epsilon = 0$ for the Weibull model

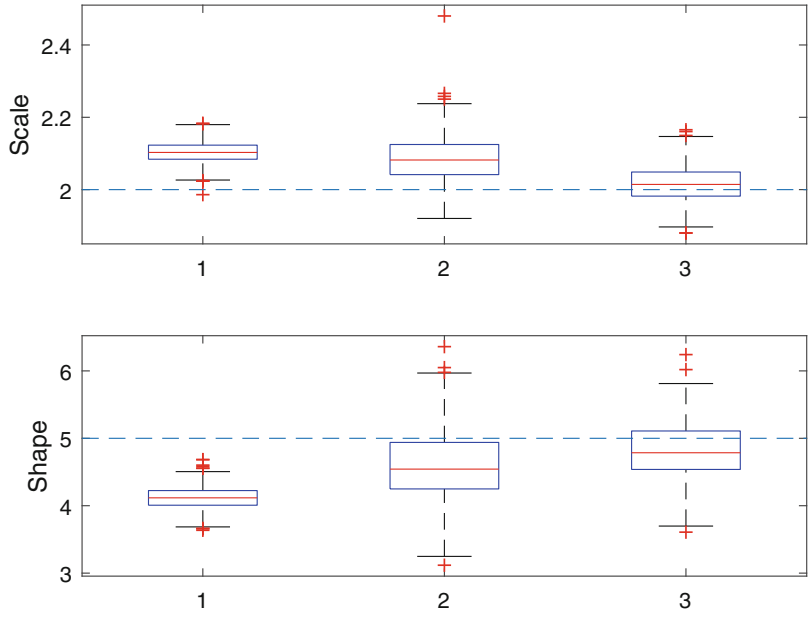


**FIGURE 4** Boxplots of maximum likelihood estimator (left), $L_2E$ (middle) and optimal B-robust estimator (right) for the scale (top) and shape (bottom) parameters with $n = 200$, $p_c = 0.2$, $\epsilon = 0.05$ for the Weibull model

*c*. In some sense, this is a price to pay for a broader bias control in the vicinity of the model. In addition, the bias may be exacerbated by numerical issues in the computation of the vector $\boldsymbol{a}(\boldsymbol{\theta})$ used to recenter the score function; see (14) in the statement of Proposition 4. The $L_2E$ is even more biased, while its variability is comparable to the one of the OBRE. On the other hand, in Figure 4, we can observe that the maximum likelihood estimates for both parameters are considerably affected by the contamination, while the optimal B-robust estimates remain well behaved. We note that, in this case, the variability of the OBRE is still greater than the variability of the MLE. As mentioned above, this is due to the very small value of $c$ that we consider in this example. The researcher may be willing to accept a larger bias in exchange for a decreased variance, and this could be achieved by modifying the tuning parameter. The $L_2E$ controls the bias better than the MLE, but has a much greater variance. Overall, it performs considerably worse than the OBRE.

Table 1 presents our full results for $n = 50$, while the corresponding results in the case $n = 200$ are given in Table 2.

**TABLE 1** Bias, variance and mean squared error (MSE) of the maximum likelihood estimator (MLE), the $L_2E$ and the optimal B-robust estimator (OBRE) of scale ($\lambda = 2$) and shape ($\kappa = 5$) for an expected censoring proportion $p_c \in \{0.1, 0.2\}$ and a contamination level $\epsilon \in \{0, 0.05\}$ based on 500 samples of size $n = 50$ for the Weibull model

| | | | $p_c = 0.1$ | | $p_c = 0.2$ | |
|---|---|---|---|---|---|---|
| | | | $\epsilon = 0$ | $\epsilon = 0.05$ | $\epsilon = 0$ | $\epsilon = 0.05$ |
| Scale ($\lambda = 2$) | Bias | MLE | −0.0025 | 0.0723 | 0.0026 | 0.0861 |
| | | $L_2E$ | 0.0195 | 0.0365 | 0.0538 | 0.0812 |
| | | OBRE | −0.0147 | 0.0068 | −0.0081 | 0.0101 |
| | Variance | MLE | 0.0039 | 0.0037 | 0.0041 | 0.0040 |
| | | $L_2E$ | 0.0056 | 0.0072 | 0.0078 | 0.0117 |
| | | OBRE | 0.0066 | 0.0073 | 0.0077 | 0.0084 |
| | MSE | MLE | 0.0039 | 0.0089 | 0.0041 | 0.0114 |
| | | $L_2E$ | 0.0060 | 0.0086 | 0.0107 | 0.0183 |
| | | OBRE | 0.0068 | 0.0073 | 0.0077 | 0.0085 |
| Shape ($\kappa = 5$) | Bias | MLE | 0.1610 | −0.6771 | 0.2250 | −0.7269 |
| | | $L_2E$ | 0.2297 | 0.0161 | 0.3084 | 0.0087 |
| | | OBRE | 0.3199 | 0.0393 | 0.4571 | 0.0908 |
| | Variance | MLE | 0.3712 | 0.1282 | 0.4081 | 0.1446 |
| | | $L_2E$ | 0.7314 | 0.8188 | 1.0967 | 1.2087 |
| | | OBRE | 0.7156 | 0.6358 | 0.9369 | 0.7355 |
| | MSE | MLE | 0.3971 | 0.5867 | 0.4587 | 0.6730 |
| | | $L_2E$ | 0.7842 | 0.8191 | 1.1918 | 1.2087 |
| | | OBRE | 0.8180 | 0.6373 | 1.1458 | 0.7437 |

**TABLE 2** Bias, variance and mean squared error (MSE) of the maximum likelihood estimator (MLE), the $L_2E$ and the optimal B-robust estimator (OBRE) of scale ($\lambda = 2$) and shape ($\kappa = 5$) for an expected censoring proportion $p_c \in \{0.1, 0.2\}$ and a contamination level $\epsilon \in \{0, 0.05\}$ based on 500 samples of size $n = 200$ for the Weibull model

| | | | $p_c = 0.1$ | | $p_c = 0.2$ | |
|---|---|---|---|---|---|---|
| | | | $\epsilon = 0$ | $\epsilon = 0.05$ | $\epsilon = 0$ | $\epsilon = 0.05$ |
| Scale ($\lambda = 2$) | Bias | MLE | 0.0005 | 0.0879 | −0.0028 | 0.1036 |
| | | $L_2E$ | 0.0248 | 0.0339 | 0.0503 | 0.0847 |
| | | OBRE | −0.0066 | 0.0077 | −0.0111 | 0.0150 |
| | Variance | MLE | 0.0009 | 0.0008 | 0.0011 | 0.0009 |
| | | $L_2E$ | 0.0015 | 0.0019 | 0.0019 | 0.0039 |
| | | OBRE | 0.0018 | 0.0019 | 0.0021 | 0.0024 |
| | MSE | MLE | 0.0009 | 0.0086 | 0.0011 | 0.0117 |
| | | $L_2E$ | 0.0021 | 0.0030 | 0.0044 | 0.0111 |
| | | OBRE | 0.0018 | 0.0020 | 0.0022 | 0.0026 |
| Shape ($\kappa = 5$) | Bias | MLE | 0.0506 | −0.8026 | 0.0388 | −0.8833 |
| | | $L_2E$ | 0.0553 | −0.2300 | −0.0281 | −0.4100 |
| | | OBRE | 0.1772 | −0.1267 | 0.1801 | −0.1814 |
| | Variance | MLE | 0.0947 | 0.0309 | 0.0914 | 0.0298 |
| | | $L_2E$ | 0.1859 | 0.1762 | 0.1716 | 0.2617 |
| | | OBRE | 0.2179 | 0.1508 | 0.2332 | 0.1731 |
| | MSE | MLE | 0.0972 | 0.6751 | 0.0929 | 0.8100 |
| | | $L_2E$ | 0.1890 | 0.2291 | 0.1724 | 0.4297 |
| | | OBRE | 0.2493 | 0.1668 | 0.2656 | 0.2060 |

In agreement with our earlier comments, we can see that the OBRE really outperforms the MLE and the $L_2E$ in controlling the bias under contamination and, despite its greater variability than the MLE, achieves the lowest mean squared error by a fairly large margin when $n = 200$. For $n = 50$, the greater variance of the OBRE relative to the MLE is sometimes relatively more important than the smaller bias. As a result, the performance of the two estimators in terms of mean squared error is comparable for this sample size. We note that the level of censoring does not appear to affect the results qualitatively.

We also examine the performance of the Wald and score tests in this setting. Specifically, we test, at the nominal level of 5%, the hypothesis that the value of the shape parameter is 5, while leaving the value of the scale parameter unrestricted. Results are presented in Table 3.

It is clear that the performance of the classical tests is disastrous in this case, especially with the larger sample size. On the other hand, the robust tests are able to maintain a sensible level under contamination, the score test being slightly more conservative than the Wald test.

**TABLE 3** Empirical level of classical and robust Wald and score tests of $\kappa = 5$ for an expected censoring proportion $p_c \in \{0.1, 0.2\}$ and a contamination level $\epsilon = 0.05$ based on 500 samples of size $n \in \{50,200\}$ for the Weibull model

| | | $n = 50$ | | $n = 200$ | |
|---|---|---|---|---|---|
| | | $p_c = 0.1$ | $p_c = 0.2$ | $p_c = 0.1$ | $p_c = 0.2$ |
| Wald | Classical | 0.2000 | 0.2040 | 0.9220 | 0.9440 |
| | Robust | 0.0820 | 0.0500 | 0.0680 | 0.0980 |
| Score | Classical | 0.2880 | 0.3260 | 0.9320 | 0.9560 |
| | Robust | 0.0220 | 0.0280 | 0.0520 | 0.0660 |

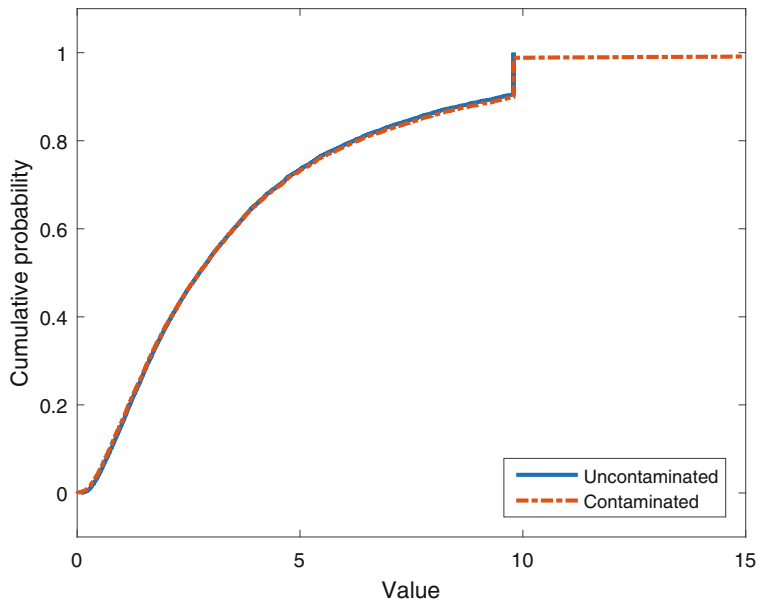


**FIGURE 5** Estimates of the cumulative distribution functions $F_X$ and $F_X^\epsilon$ ($\epsilon = 0.05$) for $p_c = 0.1$ based on a sample of size 10,000 for the log-normal model

## 5.2 | Log-normal model

As a second example, we focus on the log-normal distribution with log-mean and log-scale parameters $\mu$ and $\sigma$, that is, $f_T(t;\mu,\sigma) = 1/(\sqrt{2\pi}\sigma t)\exp\left[-\left(\log(t)-\mu\right)^2/2\sigma^2\right]$. We set $\mu = 1$ and $\sigma = 1$. The distribution of the censoring times is a point mass distribution at a point $\rho$ selected according to the expected censoring proportion. In other words, $\rho$ is simply the $(1 - p_c)$-quantile of $F_T(\cdot;1,1)$. The contaminating distribution $G_X$ is a log-normal distribution with $\mu = 1$ and $\sigma = 2$. Figure 5 illustrates the impact of contamination on the distribution of $X$. In order to have a more readable graph, a few values larger than 15 occurring in the contaminated case are not displayed.

Figures 6 and 7 give an example of the performance of the OBRE relative to the MLE under contamination, for a tuning parameter value of $c = 2$. Compared to the previous simulation study,

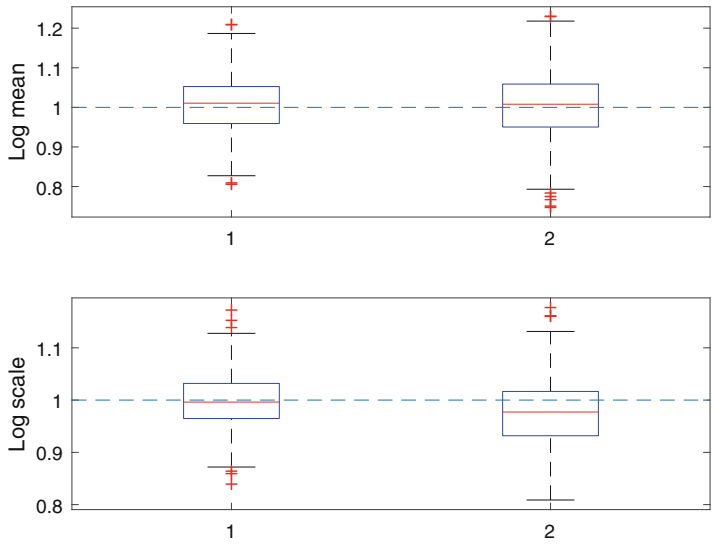


**FIGURE 6** Boxplots of maximum likelihood estimator (left) and optimal B-robust estimator (right) for the log-mean (top) and log-scale (bottom) parameters with $n = 200$, $p_c = 0.1$, $\epsilon = 0$ for the log-normal model

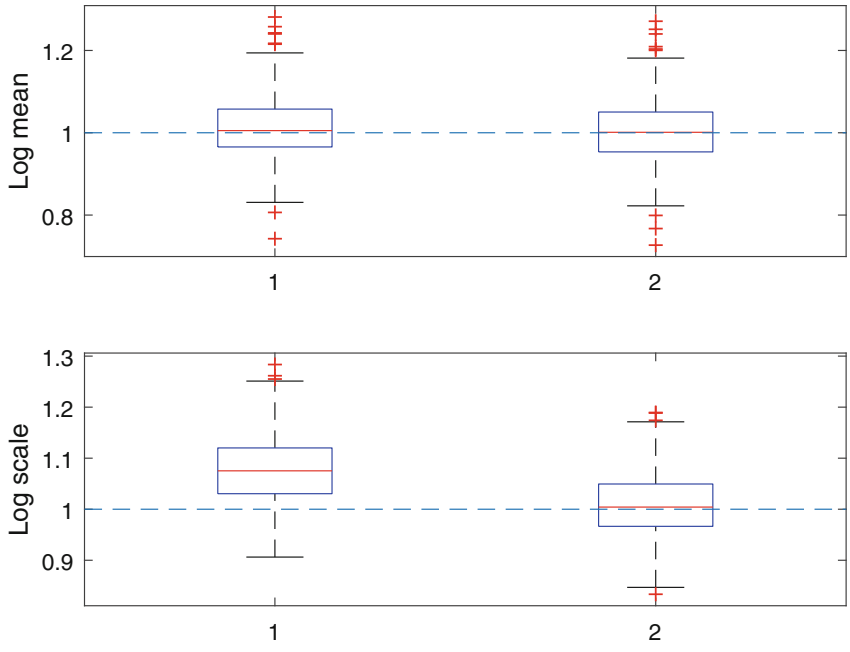


**FIGURE 7** Boxplots of maximum likelihood estimator (left) and optimal B-robust estimator (right) for the log-mean (top) and log-scale (bottom) parameters with $n = 200$, $p_c = 0.1$, $\epsilon = 0.05$ for the log-normal model

**TABLE 4** Bias, variance and mean squared error (MSE) of the maximum likelihood estimator (MLE) and the optimal B-robust estimator (OBRE) of log mean ($\mu = 1$) and log scale ($\sigma = 1$) for an expected censoring proportion $p_c \in \{0.1, 0.2\}$ and a contamination level $\epsilon \in \{0, 0.05\}$ based on 500 samples of size $n = 50$ for the log-normal model

| | | | $p_c = 0.1$ | | $p_c = 0.2$ | |
|---|---|---|---|---|---|---|
| | | | $\epsilon = 0$ | $\epsilon = 0.05$ | $\epsilon = 0$ | $\epsilon = 0.05$ |
| Log mean ($\mu = 1$) | Bias | MLE | 0.0150 | 0.0048 | 0.0039 | 0.0262 |
| | | OBRE | 0.0049 | -0.0124 | -0.0028 | 0.0040 |
| | Variance | MLE | 0.0212 | 0.0226 | 0.0224 | 0.0269 |
| | | OBRE | 0.0262 | 0.0240 | 0.0253 | 0.0265 |
| | MSE | MLE | 0.0214 | 0.0227 | 0.0224 | 0.0275 |
| | | OBRE | 0.0262 | 0.0242 | 0.0253 | 0.0266 |
| Log scale ($\sigma = 1$) | Bias | MLE | -0.0058 | 0.0599 | -0.0084 | 0.0592 |
| | | OBRE | -0.0261 | 0.0062 | -0.0234 | -0.0032 |
| | Variance | MLE | 0.0112 | 0.0183 | 0.0143 | 0.0199 |
| | | OBRE | 0.0140 | 0.0175 | 0.0192 | 0.0193 |
| | MSE | MLE | 0.0113 | 0.0219 | 0.0144 | 0.0234 |
| | | OBRE | 0.0147 | 0.0176 | 0.0197 | 0.0193 |

we thus select a slightly higher value of $c$ which should lead to a reasonable balance between bias under contamination and variance in the absence thereof.

The MLE for the log-scale parameter is affected by contamination while the OBRE is not. On the other hand, the performance of the two estimators is comparable for the log-mean parameter, which is not surprising given that our contamination scheme targets the log-scale parameter. Table 4 summarizes all of our results for the case $n = 50$. The corresponding results for the case $n = 200$ are given in Table 5.

We observe that the use of the OBRE results in an often much lower level of the mean squared error under contamination for the log-scale parameter, irrespective of its the sample size or the level of censoring. This is due to the fact that the variance of both estimators is comparable in all cases, which implies that the difference in mean squared error is mainly driven by the difference in bias.

We evaluate the performance of the robust score-type test in this example, by testing the hypothesis that the log-scale parameter is 1, while the log-mean parameter remains unspecified. The top of Table 6 shows that the level of the robust score-type test remains very close to the nominal level under contamination, being slightly too conservative. On the other hand, the performance of the classical score test is poor, though somewhat less than in the previous simulation study. This motivates us to conduct a comparison of the power of the two testing procedures. We thus simulate 500 samples using the same setting as above, except that we set the true log-scale parameter to 0.9 (respectively, 0.8) instead of 1. We then once again test the hypothesis that the value of the log-scale parameter is 1, leaving the value of the log-mean parameter unrestricted. Results are displayed in Table 6 as well.

We can observe that the robust score-type test exhibits a satisfactory power, especially at the sample size $n = 200$.

**TABLE 5** Bias, variance and mean squared error (MSE) of the maximum likelihood estimator (MLE) and the optimal B-robust estimator (OBRE) of log mean ($\mu = 1$) and log scale ($\sigma = 1$) for an expected censoring proportion $p_c \in \{0.1, 0.2\}$ and a contamination level $\epsilon \in \{0, 0.05\}$ based on 500 samples of size $n = 200$ for the log-normal model

| | | | $p_c = 0.1$ | | $p_c = 0.2$ | |
|---|---|---|---|---|---|---|
| | | | $\epsilon = 0$ | $\epsilon = 0.05$ | $\epsilon = 0$ | $\epsilon = 0.05$ |
| Log mean ($\mu = 1$) | Bias | MLE | 0.0060 | 0.0118 | 0.0037 | 0.0295 |
| | | OBRE | 0.0042 | 0.0005 | 0.0047 | 0.0079 |
| | Variance | MLE | 0.0052 | 0.0056 | 0.0052 | 0.0059 |
| | | OBRE | 0.0063 | 0.0057 | 0.0058 | 0.0059 |
| | MSE | MLE | 0.0052 | 0.0058 | 0.0052 | 0.0068 |
| | | OBRE | 0.0064 | 0.0057 | 0.0058 | 0.0060 |
| Log scale ($\sigma = 1$) | Bias | MLE | −0.0016 | 0.0771 | −0.0044 | 0.0944 |
| | | OBRE | −0.0233 | 0.0073 | −0.0257 | 0.0158 |
| | Variance | MLE | 0.0027 | 0.0044 | 0.0034 | 0.0063 |
| | | OBRE | 0.0038 | 0.0041 | 0.0043 | 0.0055 |
| | MSE | MLE | 0.0027 | 0.0103 | 0.0034 | 0.0152 |
| | | OBRE | 0.0043 | 0.0041 | 0.0049 | 0.0058 |

**TABLE 6** Empirical rejection frequency of classical and robust score tests of $\sigma = 1$ for a true value of the log-scale parameter $\sigma \in \{1, 0.9, 0.8\}$, an expected censoring proportion $p_c \in \{0.1, 0.2\}$ and a contamination level $\epsilon = 0.05$ based on 500 samples of size $n \in \{50,200\}$ for the log-normal model

| | | $n = 50$ | | $n = 200$ | |
|---|---|---|---|---|---|
| | | $p_c = 0.1$ | $p_c = 0.2$ | $p_c = 0.1$ | $p_c = 0.2$ |
| Log scale = 1 | Classical | 0.1300 | 0.1220 | 0.2380 | 0.2920 |
| | Robust | 0.0380 | 0.0360 | 0.0460 | 0.0500 |
| Log scale = 0.9 | Classical | 0.1660 | 0.1640 | 0.1240 | 0.1300 |
| | Robust | 0.1520 | 0.1260 | 0.4620 | 0.4040 |
| Log scale = 0.8 | Classical | 0.4160 | 0.3400 | 0.4740 | 0.3860 |
| | Robust | 0.4000 | 0.3100 | 0.9300 | 0.8600 |

## 6 | HEAD AND NECK CANCER STUDY DATA

In this section, we consider data from the Head and Neck Cancer Study conducted by the Northern California Oncology Group as reported by Efron (1988). This is a well-known dataset in the survival analysis literature which has been analyzed by several authors since Efron's original analysis; see, for example, Meier et al. (2004) and Basu et al. (2006). More details on the data including some characteristics of the patients can be found in Fu et al. (1987). The data are quite challenging

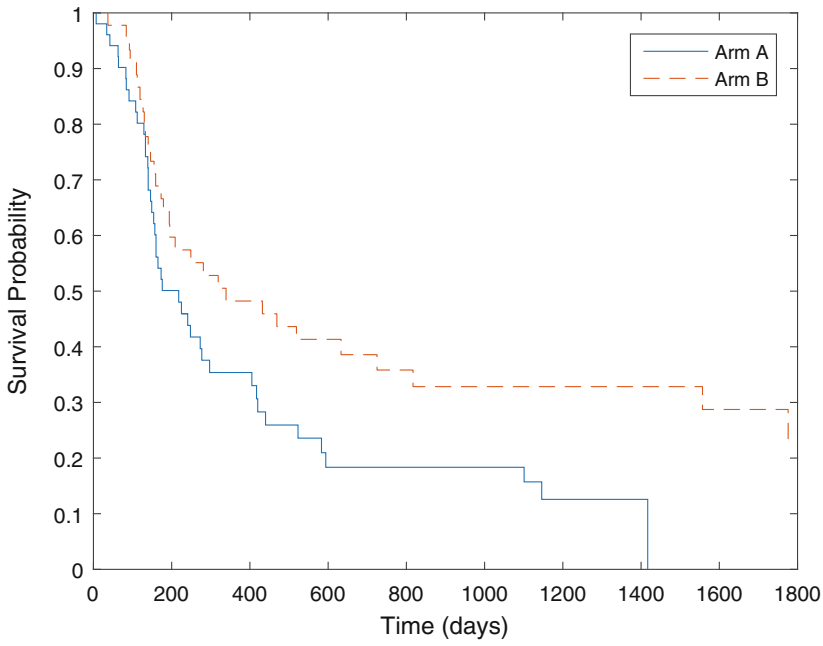


**FIGURE 8** Kaplan–Meier estimates of the survival probability for Arms A and B of the Head and Neck Cancer Study. Risk set sizes at times 0, 400, 800, 1200 and 1600 are 51, 15, 7, 4 and 0 in Arm A and 45, 21, 12, 10 and 7 in Arm B

to analyze because of the relatively small sample sizes and high prevalence of censoring. The data consist of two groups of patients, Arm A and Arm B. The 51 patients in Arm A undergo radiation therapy, while the 45 patients in Arm B are treated with both radiation therapy and chemotherapy. The proportion of censored ("lost to follow-up") observations is 0.1765 in Arm A and 0.3111 in Arm B. Figure 8 presents the Kaplan–Meier estimated survival functions for the two groups. Upon visual inspection of Figure 8, it appears that the group of patients with chemotherapy have a considerably higher survival probability, with an increasing difference over time. We note, however, that the estimated survival probabilities in roughly the second half of the range of survival times are based on risk sets which are quite small. For instance, in Arm A, only 7 out of 51 patients have a survival time exceeding 594 days.

We try to fit a Gamma model to these data. The results displayed below are obtained with the tuning parameter for the OBRE set to $c = 2.5$, although we provide evidence that other values yield qualitatively similar results in Data S1. The estimated survival functions obtained by the MLE and the OBRE are plotted in Figure 9, along with the Kaplan–Meier estimates.

Figure 9 is quite useful in assessing the potential appeal of the OBRE. It appears that, by assigning less weight to (very) long-term survivors, the OBRE is able to keep the effect of such observations under control. Correspondingly, the agreement of the OBRE with the Kaplan–Meier estimate at times when most patients are still at risk is considerably higher than for the MLE. To gain further insights, we now focus on Arm A. Figure 10 displays the weights—as given by (12)—that the OBRE assigns to observations, ranked from lowest to highest observed survival time.

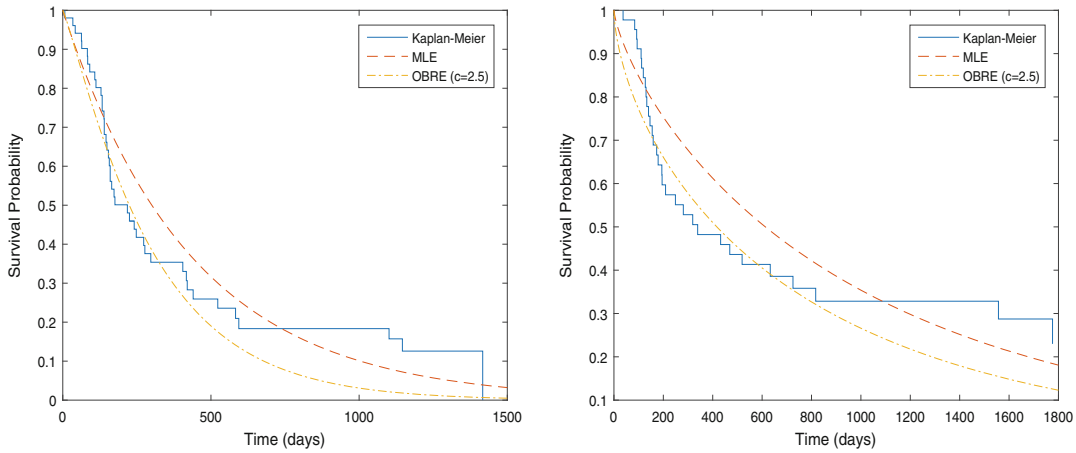


**FIGURE 9** Comparison of estimates of the survival function for Arms A (left) and B (right) of the Head and Neck Cancer Study

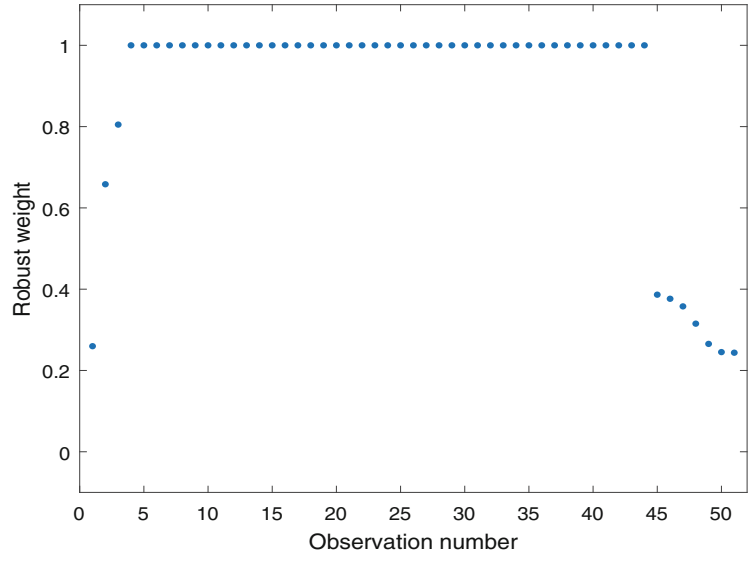


**FIGURE 10** Weights assigned by the optimal B-robust estimator to the observations in Arm A of the Head and Neck Cancer Study

As expected, the OBRE assigns a rather small weight to some long-term survivors. Perhaps more interestingly, the three patients with the smallest survival times also receive lower weights than with the MLE, showing that influential observations may be of various types. We note that these patients do exhibit small survival times (7, 34, and 42 days) compared to the median (176 days) or mean (357.84 days) survival times in Arm A. Even so, these survival times are not sufficiently far away from subsequent observations (e.g., 63, 64, and 74 days for the next three) that it would be trivial to detect them through *ad hoc* methods. As an additional experiment, we remove the $m \in \{1, 2, 3, 4\}$ observation(s) with the largest survival time(s) from this group of patients and recompute the estimates. Figure 11 shows the corresponding estimated survival functions. We

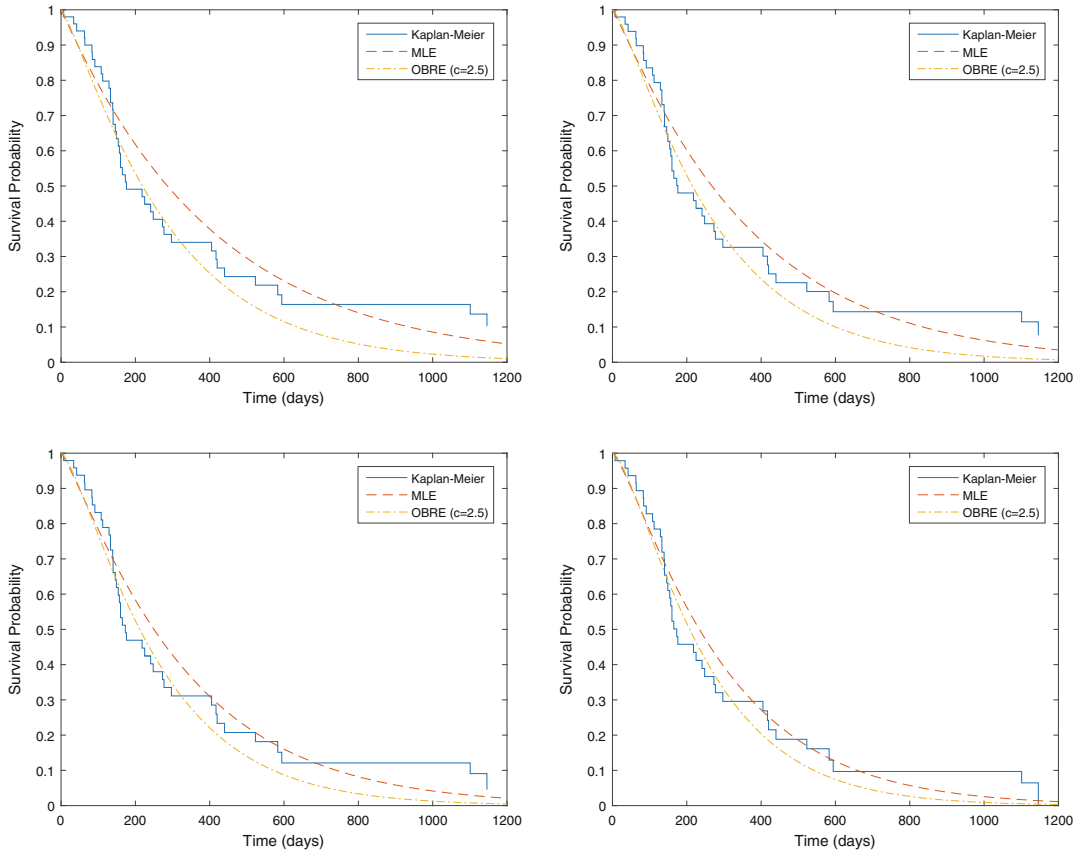


**FIGURE 11** Sensitivity analysis for Arm A of the Head and Neck Cancer Study. From left to right, top to bottom: estimated survival functions with $m \in \{1, 2, 3, 4\}$ largest survival time(s) removed

can clearly see the effect of the largest observed survival times on the curve corresponding to the MLE, which is shifted downwards as the "contamination" decreases. The curve corresponding to the OBRE, on the other hand, is largely unaffected by these changes.

An alternative graphical representation of this sensitivity analysis, in which the behavior of each estimator is depicted separately, is available in the Supporting Information.

## 7 | CONCLUDING REMARKS

In this paper, we studied robust inference with censored survival data. We defined a class of robust M-estimators based on the work of Hjort (1985, 1992) and characterized the optimal estimators with a bounded influence function in that class. We also addressed the issue of testing and showed that simple robust extensions of the three main classical tests can be defined in the presence of censoring. We illustrated the finite-sample performance of our self-standardized OBRE and some related tests in two simulation studies related to the Weibull and log-normal models, in which the impact of various aspects of the inference problem (such as the type of contamination, the proportion of censoring, and the sample size) was investigated. We also proposed an application

to a real dataset related to head and neck cancer, which is well known in the field of survival analysis.

Our analysis showed that our robust estimator is appealing at least in some situations. This does not mean that it is always absolutely necessary to rely on a robust estimator; in some situations, the MLE works sufficiently well. Nevertheless, we would advise to at least use a robust estimator as a safety check. In case the results differ greatly from those obtained with classical methods, an analysis of the weights assigned to the observations by the robust estimator may be useful in detecting highly influential observations. We emphasize, as seen in Section 7, that identifying influential data points by heuristic methods is generally difficult and can even be misleading; see also section 1.3 of Heritier et al. (2009).

Obviously, our work could be extended in different directions. First of all, it is clear that, in most applications, the researcher is faced with more than one type of incomplete information. It would thus be of interest to see how our results could be extended to account for multiple types of censoring and/or truncation. An analysis of robustness with truncated data, with applications to income data, was performed by Victoria-Feser and Ronchetti (1997); see also Victoria-Feser (1993). It would therefore be interesting to see how these results could be combined. A general formulation in terms of counting processes would certainly be required to address such a problem. One issue is then to determine a proper way of defining a contamination neighborhood in terms of these counting processes, which is certainly not trivial. The analysis of Assunção and Guttorp (1999) could serve as a useful starting point.

Second, our framework was kept as simple as possible under random censoring, in the sense that we did not assume to have any information aside from the censored survival times and the censoring indicators. In most studies, information on some patient characteristics is also recorded. In that case, our setup should be expanded to account for the presence of covariates. Robust estimation in the standard (semiparametric) Cox model was already considered by Bednarski (1993) and Sasieni (1993). An alternative and probably more direct extension of our work would be to address robust estimation and testing in a parametric version of the Cox model, as suggested in section 6A of Hjort (1992). A brief description of that setting is as follows. Let us assume that we observe $n$ triples $(X_1, \Delta_1, \boldsymbol{Z}_1), \ldots, (X_n, \Delta_n, \boldsymbol{Z}_n)$, where the first two components of each triple are defined as above and the last component is a vector of (time-invariant) covariates. The hazard rate for individual $i$ is then specified as $\alpha_i(t|\boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) = \alpha(t; \boldsymbol{\theta}) \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_i\right]$, where $\alpha(\cdot; \boldsymbol{\theta})$ is a parametrically specified baseline hazard and $\boldsymbol{\beta}$ is a vector of regression coefficients. The maximum likelihood estimator of $\left(\boldsymbol{\theta}^\top, \boldsymbol{\beta}^\top\right)^\top$ is then given by the solution to

$$\boldsymbol{U}^\dagger(\boldsymbol{\theta}, \boldsymbol{\beta}) := \begin{pmatrix} \sum_{i=1}^{n} \int_0^\infty \boldsymbol{h}(t; \boldsymbol{\theta}) \left\{ dN_i(t) - \alpha(t; \boldsymbol{\theta}) \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_i\right] Y_i(t) dt \right\} \\ \sum_{i=1}^{n} \int_0^\infty \boldsymbol{Z}_i \left\{ dN_i(t) - \alpha(t; \boldsymbol{\theta}) \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_i\right] Y_i(t) dt \right\} \end{pmatrix} = \boldsymbol{0}, \tag{15}$$

where $\boldsymbol{h}$ denotes the logarithmic derivative of the baseline hazard in this setting. Upon inspection of (15), it seems legitimate to assume that a robust estimator could be obtained by replacing $\boldsymbol{h}$ by a bounded $\boldsymbol{\psi}$ function and appropriately downweighing "large" covariates. We note, in agreement with Hjort (1992, p. 375), that even though much of the success of the standard Cox model stems from the fact that the baseline hazard need not be specified, some efficiency in the estimation of survival probabilities may be gained from an approximate knowledge of the baseline

hazard. Further investigations of robust estimation and testing in the setting just described would therefore be of importance in our view.

Finally, it would be of interest to explore potential improvements to the finite-sample properties of our robust estimators and tests. In particular, we believe that the performance of our robust methods for small sample sizes could be improved with the help of saddlepoint approximations. This type of technique is well developed for M-estimators in the case of complete information; see Robinson et al. (2003). Chapter 11 of O'Quigley (2008) contains some useful information in the context of proportional hazards regression.

## ACKNOWLEDGMENTS

Authors thank Eva Cantoni, Davide La Vecchia, Loriano Mancini, an Associate Editor and a referee for valuable comments which greatly improved the paper. A part of this work was conducted while the first author was at the Research Center for Statistics of the University of Geneva, whose support is gratefully acknowledged. Open Access Funding provided by Universite de Neuchatel.

## ORCID

*Pierre-Yves Deléamont* https://orcid.org/0000-0003-1729-7253

## REFERENCES

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, *6*, 701–726.

Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. Springer Science & Business Media.

Aranda-Ordaz, F. J. (1987). Relative efficiency of the Kaplan-Meier estimator under contamination. *Communications in Statistics-Simulation and Computation*, *16*, 987–997.

Assunção, R., & Guttorp, P. (1999). Robustness for inhomogeneous Poisson point processes. *Annals of the Institute of Statistical Mathematics*, *51*, 657–678.

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*, 549–559.

Basu, S., Basu, A., & Jones, M. (2006). Robust and efficient parametric estimation for censored survival data. *Annals of the Institute of Statistical Mathematics*, *58*, 341–355.

Bednarski, T. (1993). Robust estimation in Cox's regression model. *Scandinavian Journal of Statistics*, *20*, 213–225.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, *5*, 445–463.

Bickel, P. J., Klaassen, C., Ritov, Y., & Wellner, J. (1993). *Efficient and adaptive inference in semiparametric models*. Johns Hopkins University Press.

Borgan, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics*, *11*, 1–16.

Denecke, L., & Müller, C. H. (2014). New robust tests for the parameters of the Weibull distribution for complete and censored data. *Metrika*, *77*, 585–607.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, *83*, 414–425.

Efron, B., & Johnstone, I. M. (1990). Fisher's information in terms of the hazard rate. *The Annals of Statistics*, *18*, 38–62.

Fu, K. K., Phillips, T. L., Silverberg, I. J., Jacobs, C., Goffinet, D. R., Chun, C., Friedman, M. A., Kohler, M., McWhirter, K., & Carter, S. K. (1987). Combined radiotherapy and chemotherapy with bleomycin and methotrexate for advanced inoperable head and neck cancer: Update of a Northern California Oncology Group randomized trial. *Journal of Clinical Oncology*, *5*, 1410–1418.

Gámiz, M. L., Mammen, E., Martínez Miranda, M. D., & Nielsen, J. P. (2016). Double one-sided cross-validation of local linear hazards. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*, 755–779.

Gavrilov, L. A., & Gavrilova, N. S. (2011). Mortality measurement at advanced ages: A study of the social security administration death master file. *North American Actuarial Journal*, *15*, 432–447.

Ghosh, A., Basu, A. & Pardo, L. (2017). Robust Wald-type tests under random censoring with applications to clinical trial analyses. *arXiv preprint arXiv:1708.09695*.

Gill, R. D., & Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, *18*, 1501–1555.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation* [Ph.D. thesis]. University of California Berkeley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. John Wiley & Sons.

Heritier, S., Cantoni, E., Copt, S., & Victoria-Feser, M.-P. (2009). *Robust methods in biostatistics*. John Wiley & Sons.

Heritier, S., & Ronchetti, E. M. (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association*, *89*, 897–904.

Hjort, N. L. (1985). Discussion of 'Counting process models for life history data: A review' by Andersen, P. K. and Borgan, Ø. *Scandinavian Journal of Statistics*, *12*, 141–150.

Hjort, N. L. (1992). On inference in parametric survival data models. *International Statistical Review*, *60*, 355–387.

Huber, P. J. (1967). *The behavior of maximum likelihood estimates under nonstandard conditions*. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 221–233). Berkeley.

Huber, P. J. (1981). *Robust statistics*. Wiley.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). Wiley.

James, I. R. (1986). On estimating equations with censored data. *Biometrika*, *73*, 35–42.

Janssen, A. (1994). On local odds and hazard rate models in survival analysis. *Statistics & Probability Letters*, *20*, 355–365.

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

Klein, J. P., & Moeschberger, M. L. (1989). The robustness of several estimators of the survivorship function with randomly censored data. *Communications in Statistics - Simulation and Computation*, *18*, 1087–1112.

La Vecchia, D., & Trojani, F. (2010). Infinitesimal robustness for diffusions. *Journal of the American Statistical Association*, *105*, 703–712.

Mancini, L., Ronchetti, E. M., & Trojani, F. (2005). Optimal conditionally unbiased bounded-influence inference in dynamic location and scale models. *Journal of the American Statistical Association*, *100*, 628–641.

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

Miller, R. G. (1983). What price Kaplan-Meier? *Biometrics*, *39*, 1077–1081.

O'Quigley, J. (2008). *Proportional hazards regression*. Springer.

Rebolledo, R. (1980). Central limit theorems for local martingales. *Probability Theory and Related Fields*, *51*, 269–286.

Reid, N. (1981). Influence functions for censored data. *The Annals of Statistics*, *9*, 78–92.

Ritov, Y., & Wellner, J. A. (1988). Censoring, martingales, and the Cox model. *Contemporary Mathematics*, *80*, 191–219.

Robinson, J., Ronchetti, E. M., & Young, G. A. (2003). Saddlepoint approximations and tests based on multivariate M-estimates. *The Annals of Statistics*, *31*, 1154–1169.

Ronchetti, E. M., & Trojani, F. (2001). Robust inference with GMM estimators. *Journal of Econometrics*, *101*, 37–69.

Samuels, S. J. (1978). *Robustness of survival estimators* [Ph.D. thesis]. University of Washington.

Sasieni, P. (1992). Non-orthogonal projections and their application to calculating the information in a partly linear Cox model. *Scandinavian Journal of Statistics*, *19*, 215–233.

Sasieni, P. (1993). Maximum weighted partial likelihood estimators for the Cox model. *Journal of the American Statistical Association*, *88*, 144–152.

Stefanski, L. A., Carroll, R. J., & Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika*, *73*, 413–424.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Victoria-Feser, M.-P. (1993). *Robust methods for personal income distribution models* [Ph.D. thesis]. University of Geneva.

Victoria-Feser, M.-P., & Ronchetti, E. M. (1997). Robust estimation for grouped data. *Journal of the American Statistical Association*, *92*, 333–340.

Wang, J.-L. (1999). Asymptotic properties of M-estimators based on estimating equations and censored data. *Scandinavian Journal of Statistics*, *26*, 297–318.

Welsh, A. H. (1996). *Aspects of statistical inference*. John Wiley & Sons.

Yang, J., & Scott, D. W. (2013). Robust fitting of a Weibull model with optional censoring. *Computational Statistics & Data Analysis*, *67*, 149–161.

Yang, S. (1991). Minimum Hellinger distance estimation of parameter in the random censorship model. *The Annals of Statistics*, *19*, 579–602.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Deléamont, P. -Y., & Ronchetti, E. (2022). Robust inference with censored survival data. *Scandinavian Journal of Statistics*, *49*(4), 1496–1533. https://doi.org/10.1111/sjos.12570

## APPENDIX

### Proof of Proposition 1

The proof relies on the arguments given in chapter VI of Andersen et al. (1993). By a Taylor expansion, for $\hat{\boldsymbol{\theta}} \in \Theta^*$,

$$0 = \frac{1}{\sqrt{n}} U_j(\hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}} U_j(\boldsymbol{\theta}^*) + \sum_{k=1}^{p} \sqrt{n} \left(\hat{\theta}_k - \theta_k^*\right) \frac{1}{n} \frac{\partial}{\partial \theta_k} U_j(\tilde{\boldsymbol{\theta}}),$$

for any $j \in \{1, \dots, p\}$, where $\tilde{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$. We now proceed in two steps.

First, we show that $n^{-1/2}\boldsymbol{U}(\boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{Q}(\boldsymbol{\theta}^*)\right)$. Let $U_j^t(\boldsymbol{\theta}) := \int_0^t \Psi_j(u;\boldsymbol{\theta})\{dN(u) - \alpha(u;\boldsymbol{\theta})Y(u)du\}$. Since $\boldsymbol{\theta}^*$ is the true parameter, we can consider $\{n^{-1/2}U_j^t(\boldsymbol{\theta}^*), t \in \mathbb{R}_+\}$, $j \in \{1, \dots, p\}$, as local square integrable martingales (this is true with probability one asymptotically). Then, as $t \to \infty$, the predictable covariation process for two such local square integrable martingales converges to

$$\lim_{t\to\infty} < \frac{1}{\sqrt{n}} U_j^t(\boldsymbol{\theta}^*), \frac{1}{\sqrt{n}} U_k^t(\boldsymbol{\theta}^*) > (t) = \frac{1}{n} \int_0^\infty \Psi_j(u;\boldsymbol{\theta}^*)\Psi_k(u;\boldsymbol{\theta}^*)\alpha(u;\boldsymbol{\theta}^*)Y(u)du,$$

which converges in probability to $Q_{jk}(\boldsymbol{\theta}^*)$ by condition 1 of Assumption 1. The result now follows from a martingale central limit theorem as in Rebolledo (1980) and the Cramér–Wold device.

Second, we show that, for any $j, k \in \{1, \dots, p\}$, $n^{-1}\frac{\partial}{\partial\theta_k}U_j(\tilde{\boldsymbol{\theta}}) \xrightarrow{p} -P_{jk}(\boldsymbol{\theta}^*)$. By a Taylor expansion, for $\tilde{\boldsymbol{\theta}} \in \Theta^*$, we have

$$\frac{1}{n}\frac{\partial}{\partial\theta_k}U_j(\tilde{\boldsymbol{\theta}}) = \frac{1}{n}\frac{\partial}{\partial\theta_k}U_j(\boldsymbol{\theta}^*) + \frac{1}{n}\sum_{l=1}^{p}\left(\tilde{\theta}_l - \theta_l^*\right)\frac{\partial^2}{\partial\theta_k\partial\theta_l}U_j(\boldsymbol{\theta}^\dagger),$$

where $\boldsymbol{\theta}^\dagger$ lies between $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$. The first term on the right-hand side converges in probability to $-P_{jk}(\boldsymbol{\theta}^*)$. Indeed, it can be expressed as the difference of two terms in the following way:

$$\frac{1}{n}\frac{\partial}{\partial\theta_k}U_j(\boldsymbol{\theta}^*) = \frac{1}{n}\int_0^\infty \frac{\partial}{\partial\theta_k}\Psi_j(t;\boldsymbol{\theta}^*)\{dN(t) - \alpha(t;\boldsymbol{\theta}^*)Y(t)dt\} - \frac{1}{n}\int_0^\infty \Psi_j(t;\boldsymbol{\theta}^*)h_k(t;\boldsymbol{\theta}^*)\alpha(t;\boldsymbol{\theta}^*)Y(t)dt.$$

Using once more the fact that $\boldsymbol{\theta}^*$ is the true parameter, the first term converges to zero in probability, whereas the second term converges to $P_{jk}(\boldsymbol{\theta}^*)$ by condition 1 of Assumption 1.

The proof will be complete if we can show that $n^{-1}\sum_{l=1}^{p}\left(\tilde{\theta}_l - \theta_l^*\right)\frac{\partial^2}{\partial\theta_k\partial\theta_l}U_j(\boldsymbol{\theta}^\dagger)$ is bounded in probability by $K\left\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|$ for some constant $K$ not depending on $\boldsymbol{\theta}$. By the Cauchy–Schwarz inequality, we have

$$\frac{1}{n}\left|\sum_{l=1}^{p}\left(\tilde{\theta}_l - \theta_l^*\right)\frac{\partial^2}{\partial\theta_k\partial\theta_l}U_j(\boldsymbol{\theta}^\dagger)\right| \le \left\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|\sqrt{\sum_{l=1}^{p}\left(\frac{1}{n}\frac{\partial^2}{\partial\theta_k\partial\theta_l}U_j(\boldsymbol{\theta}^\dagger)\right)^2}.$$

Moreover, by condition 1 of Assumption 1, we get

$$\begin{aligned}\left|\frac{1}{n}\frac{\partial^2}{\partial\theta_k\partial\theta_l}U_j(\boldsymbol{\theta}^\dagger)\right| &= \left|\frac{1}{n}\int_0^\infty \frac{\partial^2}{\partial\theta_k\partial\theta_l}\Psi_j(t;\boldsymbol{\theta}^\dagger)dN(t) - \frac{1}{n}\int_0^\infty \frac{\partial^2}{\partial\theta_k\partial\theta_l}\left[\Psi_j(t;\boldsymbol{\theta}^\dagger)\alpha(t;\boldsymbol{\theta}^\dagger)\right]Y(t)dt\right| \\ &\le \left|\frac{1}{n}\int_0^\infty G_n(t)\{dN(t) - \alpha(t;\boldsymbol{\theta}^*)Y(t)dt\}\right| \\ &\quad + \left|\frac{1}{n}\int_0^\infty G_n(t)\alpha(t;\boldsymbol{\theta}^*)Y(t)dt\right| + \left|\frac{1}{n}\int_0^\infty H_n(t)Y(t)dt\right|,\end{aligned}$$

where the first term converges in probability to zero while the last two terms converge in probability to a finite constant. This ends the proof of the proposition.

**Proof of Proposition 2**

Given the contamination scheme under consideration, the influence function can be regarded as the derivative of $\boldsymbol{S}\left(F_X^{0,\epsilon}, F_X^{1,\epsilon}\right)$ with respect to $\epsilon$, evaluated at $\epsilon = 0$. We have

$$\mathbf{0} = \int_0^\infty \boldsymbol{\Psi}\left(t;\boldsymbol{S}\left(F_X^{0,\epsilon}, F_X^{1,\epsilon}\right)\right)\left\{dF_X^{1,\epsilon}(t) - \alpha\left(t;\boldsymbol{S}\left(F_X^{0,\epsilon}, F_X^{1,\epsilon}\right)\right)\left[1 - F_X^{0,\epsilon}(t) - F_X^{1,\epsilon}(t)\right]dt\right\}$$

$$= (1-\epsilon)\int_0^\infty \boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^{0,\epsilon}, F_X^{0,\epsilon}\right)\right) dF_X^1(t) + \epsilon\tilde{\delta}\boldsymbol{\Psi}\left(\tilde{x}; \boldsymbol{S}\left(F_X^{0,\epsilon}, F_X^{0,\epsilon}\right)\right)$$
$$- \int_0^\infty \boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^{0,\epsilon}, F_X^{1,\epsilon}\right)\right) \alpha\left(t; \boldsymbol{S}\left(F_X^{0,\epsilon}, F_X^{1,\epsilon}\right)\right) \left[1 - F_X(t) + \epsilon\left(F_X(t) - \mathbb{1}_{\{x\geq\tilde{x}\}}(t)\right)\right] dt,$$

with $F_X = F_X^0 + F_X^1$ by definition. Taking derivatives with respect to $\epsilon$ on both sides, evaluating them at $\epsilon = 0$, and rearranging terms yields

$$\frac{\partial}{\partial\epsilon}\boldsymbol{S}\left(F_X^{0,\epsilon}, F_X^{1,\epsilon}\right)|_{\epsilon=0} = \boldsymbol{P}\left(F_X^0, F_X^1\right)^{-1}\boldsymbol{\Xi}(\tilde{x}, \tilde{\delta}; \boldsymbol{S}(F_X^0, F_X^1)),$$

with

$$\boldsymbol{P}\left(F_X^0, F_X^1\right) = \int_0^\infty \boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\boldsymbol{h}\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)^\top \alpha\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\left[1 - F_X^0(t) - F_X^1(t)\right] dt$$
$$- \int_0^\infty \nabla_{\theta^\top}\boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\left\{dF_X^1(t) - \alpha\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\left[1 - F_X^0(t) - F_X^1(t)\right] dt\right\}.$$

and

$$\boldsymbol{\Xi}\left(\tilde{x}, \tilde{\delta}; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right) = \left[\tilde{\delta}\boldsymbol{\Psi}\left(\tilde{x}; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right) - \int_0^{\tilde{x}} \boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\alpha\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)dt\right]$$
$$- \left[\int_0^\infty \boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\left\{dF_X^1(t) - \alpha\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)[1 - F_X(t)]\, dt\right\}\right]$$
$$= \tilde{\delta}\boldsymbol{\Psi}\left(\tilde{x}; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right) - \int_0^{\tilde{x}} \boldsymbol{\Psi}\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)\alpha\left(t; \boldsymbol{S}\left(F_X^0, F_X^1\right)\right)dt,$$

where the last equality follows from the definition of the M-functional. The last statement in the proposition is a direct consequence of Fisher consistency, using the fact that $dF_X^{1,\boldsymbol{\theta}}(t) = \alpha(t;\boldsymbol{\theta})\left[1 - F_X^{0,\boldsymbol{\theta}}(t) - F_X^{1,\boldsymbol{\theta}}(t)\right] dt$.

**Proof of Lemma 1**

The results of Lemma 1 were given in the thorough treatments by Ritov and Wellner (1988) and Efron and Johnstone (1990), to which we refer the reader for many interesting insights on the role played by the operators $R$ and $L$. We briefly sketch their arguments for completeness. Part 1 can easily be shown with the help of Fubini's theorem. Indeed,

$$(R\circ L)\boldsymbol{v}(t) = \boldsymbol{v}(t) - \int_0^t \boldsymbol{v}(s)\alpha(s)ds - \frac{1}{1-F_T(t)}\int_t^\infty \boldsymbol{v}(s)f_T(s)ds + \frac{1}{1-F_T(t)}\int_t^\infty\int_0^s \boldsymbol{v}(u)\alpha(u)duf_T(s)ds$$

$$
\begin{aligned}
&= \boldsymbol{\nu}(t) - \frac{1}{1-F_T(t)} \int_0^t (1-F_T(t))\, \boldsymbol{\nu}(s)\alpha(s)ds - \frac{1}{1-F_T(t)} \int_t^\infty (1-F_T(s))\, \boldsymbol{\nu}(s)\alpha(s)ds \\
&\quad + \frac{1}{1-F_T(t)} \int_0^\infty \min\{1-F_T(u), 1-F_T(t)\} \boldsymbol{\nu}(u)\alpha(u)du \\
&= \boldsymbol{\nu}(t).
\end{aligned}
$$

Part 1 can be proved analogously. Parts (c) and (d) follow directly from these two properties, using the fact that $R$ and $L$ are isometries of $L_2(F)$.

**Proof of Lemma 2**

Define $\boldsymbol{\mu}(x, \delta; \theta) := \delta\boldsymbol{\Phi}(x;\theta) + (1-\delta)(1-F_T(x;\theta))^{-1} \int_x^\infty \boldsymbol{\Phi}(t;\theta) f_T(t;\theta)dt$. Assume first that $\|\boldsymbol{\Phi}(x;\theta)\| \leq c\ \forall x$. Then, for any $x$, we trivially have $\|\boldsymbol{\mu}(x,1;\theta)\| = \|\boldsymbol{\Phi}(x;\theta)\| \leq c$ and

$$
\begin{aligned}
\|\boldsymbol{\mu}(x,0;\theta)\| &= \left\| \frac{1}{1-F_T(x;\theta)} \int_x^\infty \boldsymbol{\Phi}(t;\theta) f_T(t;\theta)dt \right\| \\
&\leq \frac{1}{1-F_T(x;\theta)} \int_x^\infty \|\boldsymbol{\Phi}(t;\theta)\| f_T(t;\theta)dt \\
&\leq c.
\end{aligned}
$$

Conversely, suppose that $\|\boldsymbol{\mu}(x,\delta;\theta)\| \leq c\ \forall (x,\delta)$. Then, in particular, $\|\boldsymbol{\mu}(x,1;\theta)\| = \|\boldsymbol{\Phi}(x;\theta)\| \leq c$ for any $x$, which completes the proof.

**Proof of Proposition 3**

We start by rewriting the problem in terms of a function $\boldsymbol{\Phi} \in L_2^0(F_T)$, satisfying $\boldsymbol{\Psi} = R\boldsymbol{\Phi}$. In doing so, we will suppress $\theta$, which is fixed throughout, from the notation for simplicity. By (9), the problem consists of minimizing

$$
\operatorname{tr}\left( \int \left[ \delta\boldsymbol{\Phi}(x) + \frac{1-\delta}{1-F_T(x)} \int_x^\infty \boldsymbol{\Phi}(t) f_T(t)dt \right] \left[ \delta\boldsymbol{\Phi}(x) + \frac{1-\delta}{1-F_T(x)} \int_x^\infty \boldsymbol{\Phi}(t) f_T(t)dt \right]^\top dF_{X,\Delta}(x,\delta) \right), \quad \text{(A1)}
$$

subject to

$$
\sup_{(x,\delta)\in\mathbb{R}_+\times\{0,1\}} \left\| \delta\boldsymbol{\Phi}(x) + \frac{1-\delta}{1-F_T(x)} \int_x^\infty \boldsymbol{\Phi}(t) f_T(t)dt \right\| \leq c, \quad \text{(A2)}
$$

and

$$\int \left[\delta\boldsymbol{\Phi}(x) + \frac{1-\delta}{1-F_T(x)}\int_x^\infty \boldsymbol{\Phi}(t)f_T(t)dt\right]\left[\delta\boldsymbol{s}(x) + \frac{1-\delta}{1-F_T(x)}\int_x^\infty \boldsymbol{s}(t)f_T(t)dt\right]^\top dF_{X,\Delta}(x,\delta) = \boldsymbol{I_p}. \quad \text{(A3)}$$

Our aim is now to rewrite the problem in a simpler way. First, we note that, by Lemma 2, the constraint (A2) is equivalent to

$$\sup_{x\in\mathbb{R}_+} \|\boldsymbol{\Phi}(x)\| \leq c. \quad \text{(A4)}$$

Second, we remark that the minimization of (A1) subject to (A3) is equivalent to the minimization of

$$\begin{aligned}&\int \delta\|\boldsymbol{\Phi}(x) - \boldsymbol{A}\left[\boldsymbol{s}(x) - \boldsymbol{a}\right]\|^2 dF_{X,\Delta}(x,\delta)\\ &+ \int (1-\delta)\left\|\frac{1}{1-F_T(x)}\int_x^\infty \left[\boldsymbol{\Phi}(t) - \boldsymbol{A}\left[\boldsymbol{s}(t) - \boldsymbol{a}\right]\right]f_T(t)dt\right\|^2 dF_{X,\Delta}(x,\delta),\end{aligned} \quad \text{(A5)}$$

where the matrix $\boldsymbol{A}$ can be chosen in such a way that the constraint (A3) holds. This can be readily verified by exploiting (A3) and the property that

$$\begin{aligned}&\int \left[\delta\boldsymbol{\Phi}(x) + \frac{1-\delta}{1-F_T(x)}\int_x^\infty \boldsymbol{\Phi}(t)f_T(t)dt\right]dF_{X,\Delta}(x,\delta)\\ &= \int_0^\infty \boldsymbol{\Phi}(x)\left[1-F_C(x)\right]dF_T(x) + \int_0^\infty \left[\frac{1}{1-F_T(x)}\int_x^\infty \boldsymbol{\Phi}(t)f_T(t)dt\right]\left[1-F_T(x)\right]dF_C(x)\\ &= \int \boldsymbol{\Phi}(x)dF_T(x)\\ &= \mathbf{0}.\end{aligned}$$

for $\boldsymbol{\Phi} \in L_2^0(F_T)$, where we made use of Fubini's theorem in the second-to-last line. By analogy with the case of complete information, it is clear that the minimizer of (A5) subject to (A4) is

$$\boldsymbol{\Phi_U}(x) = \boldsymbol{H_c}\left(\boldsymbol{A}\left[\boldsymbol{s}(x) - \boldsymbol{a}\right]\right).$$

The result given in Proposition 3 now follows directly by applying $R$ to $\boldsymbol{\Phi_U}$.

**Proof of Proposition 4**

The proof builds on Proposition 3 as well as on theorem 1 in Stefanski et al. (1986). As in the proof of Proposition 3, we omit $\boldsymbol{\theta}$ from the notation. We want to show that any "competitor" $\boldsymbol{\Psi}$ of $\boldsymbol{\Psi_B}$ is essentially the same as $\boldsymbol{\Psi_B}$, up to multiplication by a constant matrix. Without loss of generality,

we assume that $\boldsymbol{P}_{\boldsymbol{\Psi}} = \boldsymbol{I}_{\boldsymbol{p}}$, which implies that $\boldsymbol{V}_{\boldsymbol{\Psi}} = \boldsymbol{Q}_{\boldsymbol{\Psi}}$. Note that

$$\begin{aligned}\boldsymbol{Q}_{\boldsymbol{\Psi}} &= \int_0^\infty \boldsymbol{\Psi}(t)\boldsymbol{\Psi}(t)^\top dF_X^1(t) \\ &= \int_0^\infty \left[\boldsymbol{\Psi}(t) - \boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}\boldsymbol{h}(t)\right]\left[\boldsymbol{\Psi}(t) - \boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}\boldsymbol{h}(t)\right]^\top dF_X^1(t) + \boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}\int_0^\infty \boldsymbol{h}(t)\boldsymbol{\Psi}(t)^\top dF_X^1(t) \\ &\quad + \int_0^\infty \boldsymbol{\Psi}(t)\boldsymbol{h}(t)^\top dF_X^1(t)\boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-\top} - \boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}\int_0^\infty \boldsymbol{h}(t)\boldsymbol{h}(t)^\top dF_X^1(t)\boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-\top}.\end{aligned}$$

Since $\boldsymbol{P}_{\boldsymbol{\Psi}} = \boldsymbol{I}_{\boldsymbol{p}}$, the second and third terms on the right-hand side are equal to $\boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}$ and $\boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-\top}$, respectively, whereas the fourth term does not depend on $\boldsymbol{\Psi}$. Therefore, in our optimality problem, we can equivalently minimize

$$\int_0^\infty \left[\boldsymbol{\Psi}(t) - \boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}\boldsymbol{h}(t)\right]^\top \boldsymbol{V}_{\boldsymbol{\Psi}_B}^{-1}\left[\boldsymbol{\Psi}(t) - \boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}\boldsymbol{h}(t)\right] dF_X^1(t),$$

with $\boldsymbol{V}_{\boldsymbol{\Psi}_B} = \boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}\boldsymbol{Q}_{\boldsymbol{\Psi}_B}\boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-\top}$. In fact, defining $\boldsymbol{\Omega}(t) := \boldsymbol{V}_{\boldsymbol{\Psi}_B}^{-\frac{1}{2}}\boldsymbol{\Psi}(t)$, our objective in terms of $\boldsymbol{\Omega}$ is to minimize

$$\int_0^\infty \left\|\boldsymbol{\Omega}(t) - \boldsymbol{V}_{\boldsymbol{\Psi}_B}^{-\frac{1}{2}}\boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}\boldsymbol{h}(t)\right\|^2 dF_X^1(t),$$

subject to

$$\sup_{(x,\delta)\in\mathbb{R}_+\times\{0,1\}} \left\|\delta\boldsymbol{\Omega}(x) - \int_0^x \boldsymbol{\Omega}(t)\alpha(t)dt\right\|^2 \leq c^2.$$

This problem, stated in terms of $\boldsymbol{\Omega}$, can be seen to be equivalent to the unstandardized problem stated in terms of $\boldsymbol{\Psi}$, with $\boldsymbol{A} = \boldsymbol{V}_{\boldsymbol{\Psi}_B}^{-\frac{1}{2}}\boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}$. Hence, it follows from Proposition 3 that the solution for $\boldsymbol{\Omega}$ is

$$\begin{aligned}\boldsymbol{\Omega}(x) &= \boldsymbol{H}_{\boldsymbol{c}}\left(\boldsymbol{A}\left[\boldsymbol{s}(x) - \boldsymbol{a}\right]\right) - \frac{1}{1 - F_T(x)}\int_x^\infty \boldsymbol{H}_{\boldsymbol{c}}\left(\boldsymbol{A}\left[\boldsymbol{s}(t) - \boldsymbol{a}\right]\right) f_T(t)dt \\ &= \boldsymbol{A}\left[\left[\boldsymbol{s}(x) - \boldsymbol{a}\right] w_c(x) - \frac{1}{1 - F_T(x)}\int_x^\infty \left[\boldsymbol{s}(t) - \boldsymbol{a}\right] w_c(t) f_T(t)dt\right],\end{aligned}$$

with $w_c(t) = \min\left\{1, c\left|[\boldsymbol{s}(t)-\boldsymbol{a}]^\top \boldsymbol{A}^\top \boldsymbol{A}\,[\boldsymbol{s}(t)-\boldsymbol{a}]\right|^{-\frac{1}{2}}\right\}$, where $\boldsymbol{A}$ satisfies $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{Q}_{\boldsymbol{\Psi}_B}^{-1}$. This implies that, up to multiplication by a constant matrix, the solution in terms of $\boldsymbol{\Psi}$ is given by

$$\boldsymbol{\Psi}(x) = \boldsymbol{P}_{\boldsymbol{\Psi}_B}^{-1}\left[[\boldsymbol{s}(x)-\boldsymbol{a}]\,w_c(x) - \frac{1}{1-F_T(x)}\int_x^\infty [\boldsymbol{s}(t)-\boldsymbol{a}]\,w_c(t) f_T(t)dt\right],$$

which is what we wanted to show.

**Proof of Proposition 5**

Letting $\boldsymbol{\Xi}_{\boldsymbol{\Psi}_{SS}}(x,\delta;\boldsymbol{\theta}) := \delta\boldsymbol{\Psi}_{\boldsymbol{SS}}(x;\boldsymbol{\theta}) - \int_0^x \boldsymbol{\Psi}_{\boldsymbol{SS}}(t;\boldsymbol{\theta})\alpha(t;\boldsymbol{\theta})dt$, we have

$$\begin{aligned}
\gamma_{SS}^2 &\geq \operatorname{tr}\left(\int_{(x,\delta)\in\mathbb{R}_+\times\{0,1\}} \left[\boldsymbol{P}_{\boldsymbol{\Psi}_{SS}}^{-1}\boldsymbol{\Xi}_{\boldsymbol{\Psi}_{SS}}(x,\delta;\boldsymbol{\theta})\right]^\top \boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}^{-1}\left[\boldsymbol{P}_{\boldsymbol{\Psi}_{SS}}^{-1}\boldsymbol{\Xi}_{\boldsymbol{\Psi}_{SS}}(x,\delta;\boldsymbol{\theta})\right] dF_{X,\Delta}^{\boldsymbol{\theta}}(x,\delta)\right)\\
&= \operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}^{-1}\boldsymbol{P}_{\boldsymbol{\Psi}_{SS}}^{-1}\int_{(x,\delta)\in\mathbb{R}_+\times\{0,1\}} \boldsymbol{\Xi}_{\boldsymbol{\Psi}_{SS}}(x,\delta;\boldsymbol{\theta})\boldsymbol{\Xi}_{\boldsymbol{\Psi}_{SS}}(x,\delta;\boldsymbol{\theta})^\top dF_{X,\Delta}^{\boldsymbol{\theta}}(x,\delta)\boldsymbol{P}_{\boldsymbol{\Psi}_{SS}}^{-\top}\right)\\
&= \operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}^{-1}\boldsymbol{P}_{\boldsymbol{\Psi}_{SS}}^{-1}\int_0^\infty \boldsymbol{\Psi}_{\boldsymbol{SS}}(x;\boldsymbol{\theta})\boldsymbol{\Psi}_{\boldsymbol{SS}}(x;\boldsymbol{\theta})^\top dF_X^{1,\boldsymbol{\theta}}(x)\boldsymbol{P}_{\boldsymbol{\Psi}_{SS}}^{-\top}\right)\\
&= \operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}^{-1}\boldsymbol{V}_{\boldsymbol{\Psi}_{SS}}\right)\\
&= p.
\end{aligned}$$