A two-step estimation procedure for semiparametric mixture cure models

Musta, E.; Patilea, Valentin; Van Keilegom, Ingrid

Publication date
2024

Document Version
Final published version

Published in
Scandinavian Journal of Statistics

License
CC BY

Citation for published version (APA):
A two-step estimation procedure for semiparametric mixture cure models

Eni Musta¹ | Valentin Patilea² | Ingrid Van Keilegom³

¹Korteweg de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands
²CREST, Ensaï, Bruz, France
³ORSTAT, KU Leuven, Leuven, Belgium

Abstract
In survival analysis, cure models have been developed to account for the presence of cured subjects that will never experience the event of interest. Mixture cure models with a parametric model for the incidence and a semiparametric model for the survival of the susceptibles are particularly common in practice. Because of the latent cure status, maximum likelihood estimation is performed via the iterative EM algorithm. Here, we focus on the cure probabilities and propose a two-step procedure to improve upon the maximum likelihood estimator when the sample size is not large. The new method is based on presmoothing by first constructing a nonparametric estimator and then projecting it on the desired parametric class. We investigate the theoretical properties of the resulting estimator and show through an extensive simulation study for the logistic-Cox model that it outperforms the existing method. Practical use of the method is illustrated through two melanoma datasets.

Keywords
cure model, logistic model, presmoothing, survival analysis

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

1 | INTRODUCTION

Cure models are used to analyze time until occurrence of an event of interest when a proportion of the study population is immune to that event (cured). They are recently becoming increasingly popular in oncology as curative treatments are now a possibility, meaning that some patients will not experience cancer relapse/death (Legrand & Bertrand, 2019; Othus et al., 2012). More broadly cure models find applications in studies of fertility (Van Geloven et al., 2013), hospitalization of COVID-19 patients (Pedrosa-Laza et al., 2022), equipment failure in engineering (Meeker, 1987), credit scoring in economics (Dirick et al., 2017, 2019), etc. What makes statistical modeling and estimation challenging when not all subjects are susceptible to the event of interest is the unobserved cure status. As a consequence of limited follow-up period, all cured subject are observed as censored, hence mixed with the uncured ones.

There are two main families of cure models: promotion time models and mixture cure models, see Amico and Van Keilegom (2018) and Peng and Yu (2021) for an overview. The later ones are particularly attractive in practice because, by modeling separately the uncure probability (incidence) and the survival of the susceptibles (latency) given possibly different sets of covariates, they are able to distinguish a curative from a life-prolonging effect. Early works on mixture cure models were fully parametric approaches (Farewell, 1982; Kuk & Chen, 1992; Yamaguchi, 1992), while more recently semiparametric (Li & Taylor, 2002; Peng & Dear, 2000; Sy & Taylor, 2000; Zhang & Peng, 2007) and nonparametric (Amico et al., 2019; López-Cheda, Cao, et al., 2017; López-Cheda, Jácome, & Cao, 2017; Patilea & Van Keilegom, 2020; Xu & Peng, 2014) models have been proposed. Among them, the semiparametric models are often used in practice as a reasonable compromise between flexibility and simplicity. These models assume a parametric form of the incidence and a semiparametric form for the latency, with the most common choice being the mixture of the logistic with the Cox proportional hazards model (Stringer et al., 2016; Wycinka & Jurkiewicz, 2017; Yilmaz et al., 2013).

Estimation in the logistic-Cox or in general semi-parametric mixture cure models is mostly carried out via the Expectation-Maximization algorithm because of the latent cure status. Such estimators were proposed in Peng and Dear (2000) and Sy and Taylor (2000) for the logistic-Cox model; in Li and Taylor (2002), Zhang and Peng (2007), and Lu (2010) for the logistic-accelerated failure time (AFT) model. The procedure is implemented in the R-package smcure (Cai et al., 2012). However, for limited sample sizes which are common in practice, such iterative procedures are characterized by large mean-squared-error (MSE), convergence problems and instability of the estimators for the incidence component depending on which variables are included in the latency model (Musta et al., 2022). This might lead to incorrect conclusions regarding significant effects.

Here we propose a new two-step estimator based on presmoothing with the aim to improve upon an initially available estimator, that can for example be the smcure estimator, for small and moderate sample sizes. Presmoothing refers to the introduction of a preliminary smoothing step before the construction of the final estimates. It has been previously proposed in the context of linear regression, variable selection, functional linear regression, instrumental regression among others (Aerts et al., 2010; Cristobal et al., 1987; Ferraty et al., 2012; Tedesco et al., 2023). The idea is to replace the response variable by a smooth nonparametric estimate of the regression function and then compute the parameter estimates via least squares or maximum likelihood. In the linear regression model presmoothing has been shown to reduce the second-order term of the asymptotic variance of the least squares estimators, leading to improvement in terms of mean squared error (MSE) for finite samples (Cristobal et al., 1987). The intuition is that replacing the response
variable by a smooth nonparametric estimator filters out some of the error and gets closer to the parametric regression function. Motivated by this, we propose the following estimation procedure in the cure model setting. An initial estimator is used to construct an one-dimensional covariate, conditional on which we compute a nonparametric estimator of the cure probabilities. Afterwards, the nonparametric estimator is projected on the desired parametric class (e.g., logistic). The novelty of our method lies in using presmoothing as a second stage estimator where a preliminary available estimator is used to reduce the covariate dimension to one. In this way we only need to choose one bandwidth independently of the number of covariates and still profit from the advantages of presmoothing: lower MSE and more stable estimators. Presmoothing allows direct estimation of the parametric incidence component despite the latent cure status. We focus on the cure fraction, but once that is estimated, one can also fit a semiparametric model to the latency component. Compared to the method proposed in Musta et al. (2022), this approach does not require multidimensional smoothing, which is essential for practical purposes. Moreover, if there is more than one continuous covariate, consistency of the estimators in Musta et al. (2022) is not guaranteed without further assumptions.

The paper is organized as follows. In Sections 2 and 3 we describe the model and the estimation procedure. In Section 4 we show that the resulting estimator is consistent and square-root-n convergent with a Gaussian limit distribution, provided that the initial estimator is consistent. As a particular case, we focus on the logistic-Cox mixture cure model in Section 5 and compare the proposed estimator with the \texttt{smcure} estimator and the method of Musta et al. (2022) through an extensive simulation study. In particular, we observe that the two-step approach has almost always lower variance than \texttt{smcure} and often also lower bias. In addition, it is also more stable toward misspecifications in the latency model. In Section 6, we apply the method to two medical datasets and show that it can in practice lead to different conclusions compared to the \texttt{smcure} estimator. We conclude with a discussion in Section 7. Proofs and the extensive simulation results are provided in the Data S1. The R code is available open access on Github and can be reached via the following link: https://github.com/eni-musta/2step_logCox_MCM.git.

2 | THE SEMIPARAMETRIC MIXTURE CURE MODEL

Suppose we are interested in the time $T$ until a certain event happens for a mixed population of cured ($T = \infty$) and uncured ($T = T_u < \infty$) subjects. The subscript “u” will be used to indicate quantities that correspond to the uncured subjects. Let $B$ be a 0-1 random variable indicating the uncured status: $B = 1$ for susceptible individuals and $B = 0$ otherwise. Due to the limited follow-up period we cannot actually observe $T$ and $B$. Instead we observe a finite survival time $Y = \min(T, C)$ and a censoring indicator $\Delta = 1_{\{T \leq C\}}$, where $C$ denotes the censoring time. As a result, for all the censored observations the cure status is unknown. In the mixture cure model, the survival function of $T$ given two covariate vectors $X \in \mathbb{R}^p$, $Z \in \mathbb{R}^q$, is given by

$$S(t|x, z) = \mathbb{P}(T > t|X = x, Z = z) = \pi_0(x) + (1 - \pi_0(x))S_u(t|z),$$

where $\pi_0(x) = \mathbb{P}(B = 0|X = x)$ denotes the cure probability and $S_u(t|z) = \mathbb{P}(T > t|Z = z, B = 1) = \mathbb{P}(T_u > t|Z = z)$ is the survival function of the susceptibles. Using two covariate vectors $X$ and $Z$ for modeling the incidence and the latency allows the cure probability and the survival of the uncured to be affected by different variables. However it does not exclude situations in which the two vectors $X$ and $Z$ are exactly the same or share some components.
In the context of mixture cure models, the classical survival analysis assumption of independent censoring, means that \( T_u \perp (C, X | Z) \) and \( B \perp (C, T_u, Z | X) \), which imply that

\[
T \perp C | (X, Z), \tag{1}
\]

see Lemma 1 in the supplementary material of Musta et al. (2022). As a result, we also have

\[
\mathbb{P}(T = \infty | X, Z) = \mathbb{P}(T = \infty | X) \quad \text{and} \quad \mathbb{P}(T_u \leq t | X, Z) = \mathbb{P}(T_u \leq t | Z). \tag{2}
\]

Among various modeling approaches for the incidence and the latency, the most common choice is a parametric (logistic) model for the incidence and a semiparametric (Cox or AFT) model for the latency (Burke & Patilea, 2020; Legrand & Bertrand, 2019; Patilea & Van Keilegom, 2020; Stringer et al., 2016; Wycinka & Jurkiewicz, 2017; Yilmaz et al., 2013). The popularity of such choice is due to the simplicity and ease of interpretation. We focus on this type of models and assume that

\[
\pi_0(x) = 1 - \phi(\gamma_0^T x),
\]

where \( \phi : \mathbb{R} \rightarrow [0, 1] \) is a known function, \( \gamma_0 \in \mathbb{R}^{p+1} \) and \( \gamma_0^T \) denotes the transpose of the vector \( \gamma_0 \). Here the first component of \( x \) is taken to be equal to one and the first component of \( \gamma \) corresponds to the intercept. In particular, for the logistic model we have

\[
\phi(u) = \frac{e^u}{1 + e^u}. \tag{3}
\]

To check the fit of this model in practice, one can compare the prediction error with that of a more flexible single-index model as done in Amico et al. (2019) or use the test proposed in Müller and Van Keilegom (2019) which is currently developed only for an one-dimensional covariate.

For the latency, we assume a semiparametric model \( S_u(t | z) = S_u(t | z; \beta_0, \Lambda_0) \) depending on a finite-dimensional parameter \( \beta_0 \in \mathbb{R}^q \), and an infinite-dimensional parameter \( \Lambda_0 \). The main examples we keep in mind are the Cox proportional hazards (PH) model

\[
S_u(t | z) = \exp\{-\Lambda_0(t) \exp(\beta_0^T z)\}, \tag{4}
\]

and the AFT model

\[
S_u(t | z) = \exp\{-\Lambda_0(\exp(\beta_0^T z) t)\},
\]

where \( \Lambda_0 \) is the baseline cumulative hazard.

The goal is to estimate the true parameters \( \gamma_0, \beta_0 \) and \( \Lambda_0 \) on the basis of \( n \) i.i.d. observations \((Y_1, \Delta_1, X_1, Z_1), \ldots, (Y_n, \Delta_n, X_n, Z_n)\). The general conditions under which the semiparametric mixture cure model is identifiable, meaning that different parameter values lead to different distributions of the observed variables \((Y, \Delta, X, Z)\), were derived by Parsa and Van Keilegom (2022) and are the following:

\begin{itemize}
  \item[(I1)] if \( \phi(\gamma^T X) = \phi(\tilde{\gamma}^T X) \) almost surely, then \( \gamma = \tilde{\gamma} \),
  \item[(I2)] the function \( S_u(\cdot | z) \) has support \([0, \tau(z)]\),
  \item[(I3)] \( \mathbb{P}(C > \tau(Z) | X, Z) > 0 \) for almost all \( X \) and \( Z \),
  \item[(I4)] if, for all \( t \geq 0 \), we have \( S_u(t | Z; \Lambda, \beta) = S_u(t | Z; \tilde{\Lambda}, \tilde{\beta}) \) almost surely, then \( \Lambda = \tilde{\Lambda} \) and \( \beta = \tilde{\beta} \).
\end{itemize}
In the particular case of the logistic-Cox model the conditions become:

(I1') for all \( x, 0 < \phi(\gamma_0^T x) < 1 \),

(I2') the function \( S_u \) has support \([0; \tau_0]\) for some \( \tau_0 < \infty \),

(I3') \( P(C > \tau_0 | X; Z) > 0 \) for almost all \( X \) and \( Z \),

(I4') the matrices \( \text{Var}(X) \) and \( \text{Var}(Z) \) are positive definite,

see Prop. 1 and 2 in Parsa and Van Keilegom (2022). Conditions I3 and I3' are of particular importance in the context of mixture cure models and essentially tell us that, in order to correctly identify the cure proportion, we need sufficiently long follow-up beyond the time when the events occur. In practice, this can be evaluated based on the plateau of the Kaplan–Meier estimator and the expert (medical) knowledge. Throughout the paper we assume that conditions (I1)-(I4), or the corresponding ones in the case of the logistic-Cox model are satisfied. In absence of model identification, one cannot expect to have a consistent preliminary estimator required for our proposed two-step procedure.

### 3 | THE TWO-STEP ESTIMATION PROCEDURE

Estimation in semiparametric mixture cure models is usually performed via the expectation maximization algorithm because of the latent cure status. Such method has been proposed by Peng and Dear (2000) and Sy and Taylor (2000) for the logistic-Cox mixture cure model, and by Li and Taylor (2002) and Zhang and Peng (2007) for the logistic-AFT model. The procedure is implemented in the R package smcure (Cai et al., 2012). Despite the simplicity of the method, simultaneous computation of \( \gamma_0, \beta_0, \) and \( \Lambda_0 \) through an iterative procedure leads to several problems for finite, not large sample sizes which are commonly encountered in practice. This has been previously reported and illustrated in Musta et al. (2022), Burke and Patilea (2020), and Han (2017). The main concerns are the large MSE, convergence problems and instability of the estimator for the incidence component depending on which variables are included in the latency model. In particular, if the latency model is misspecified, even the estimators of the incidence parameters suffer from induced bias. To alleviate these problems, we propose the following two-step estimation procedure that makes use of presmoothing.

We start with some preliminary estimator \( \tilde{\gamma}_n \) of \( \gamma_0 \). This can be any estimator that satisfies the conditions described in Section 4 and in particular for the logistic-Cox or logistic-AFT model we can use the smcure estimator. We use this preliminary estimator to construct the one-dimensional index \( \tilde{V} = \tilde{\gamma}_n^T X \) estimating \( V = \gamma_0^T X \). Based on this new one-dimensional covariate we compute a nonparametric estimator of the cure probability for each subject defined as follows

\[
\hat{\pi}_n(x) = \prod_{t \in \mathbb{R}} \left( 1 - \frac{\hat{H}_{1,\tilde{\gamma}_n}(dt|\tilde{\gamma}_n^T X)}{\hat{H}_{1,\tilde{\gamma}_n}(\{t, \infty\}|\tilde{\gamma}_n^T X)} \right),
\]

where

\[
\hat{H}_{1,\tilde{\gamma}_n}([t, \infty]|u) = \hat{H}_{1,\tilde{\gamma}_n}([t, \infty]|u) + \hat{H}_{0,\tilde{\gamma}_n}([t, \infty]|u), \quad \hat{H}_{1,\tilde{\gamma}_n}(dt|u) = \hat{H}_{1,\tilde{\gamma}_n}(\{t - dt, t\}|u) \quad \text{for small} \quad dt
\]

and

\[
\hat{H}_{1,\tilde{\gamma}_n}([t, \infty]|u) = \sum_{i=1}^{n} \frac{k_b(\tilde{\gamma}_n^T X_i - u)}{\sum_{j=1}^{n} k_b(\tilde{\gamma}_n^T X_j - u)} \mathbb{I}_{(Y_i \geq t, \Delta_i = l)} \quad l = 0, 1,
\]
are estimators of
\[ H_l((t, \infty)|u) = \mathbb{P}(Y \geq t, \Delta = l|\gamma_0^TX = u), \quad l = 0, 1, \tag{6} \]
and \( H((t, \infty)|u) = H_1((t, \infty)|u) + H_0((t, \infty)|u) \). Here \( k \) is a one-dimensional kernel function, \( b = b_n \) is a bandwidth sequence and \( k_b(\cdot) = k(-\cdot/b)/b \). \( H((t, \infty)|u) \) represents the conditional survival function of the observed survival times given that \( \gamma_0^TX = u \), while \( H_0 \) and \( H_1 \) represent the conditional subsurvival functions of the censored and uncensored observed survival times, respectively. Note also that \( \prod_{t \in \mathbb{R}} \) denotes the product integral but it actually a product of a finite number of terms. Moreover the nonparametric estimator \( \hat{\gamma}_n \) depends on \( \hat{\gamma}_n \) but to simplify the notation we omit the subscript \( \hat{\gamma}_n \).

The estimator \( \hat{\gamma}_n(x) \) is a Beran-type estimator of the conditional survival function \( S \), given the estimated index \( \hat{\gamma}_n^T X \) instead of the covariate \( X \), computed at the largest observed event time \( Y_{(m)} \) and does not require any specification of \( \tau_0 \). Since \( \hat{H}_1,\hat{\gamma}_n \left( \frac{\text{d}t}{\hat{\gamma}_n^T X} \right) \) is different from zero only at the observed event times, computation of \( \hat{\gamma}_n(x) \) requires only a product over \( t \) in the set of the observed event times. Afterwards, we consider the logistic-type likelihood

\[ \hat{L}_{n,1}(\gamma) = \prod_{i=1}^{n} \phi(\gamma^T X_i) \left( 1 - \hat{\gamma}_n^T X_i \right) \left( 1 - \phi(\gamma^T X_i) \right)^{\hat{\gamma}_n^T X_i}, \]

where the binary latent variable \( B_i \) has been replaced by the nonparametric estimator \( \hat{\gamma}_n(X_i) \) of its success probability. The estimator \( \hat{\gamma}_n \) is then defined as the maximizer of

\[ \log \hat{L}_{n,1}(\gamma) = \sum_{i=1}^{n} \left\{ [1 - \hat{\gamma}_n(X_i)] \log \phi(\gamma^T X_i) + \hat{\gamma}_n(X_i) \log \left[ 1 - \phi(\gamma^T X_i) \right] \right\} \tau_i(X_i). \tag{7} \]

We introduce a trimming function \( \tau(\cdot) \geq 0 \) to avoid regions where the density function of the index \( \gamma^T X \), for \( \gamma \) in a neighborhood of \( \gamma_0 \), approaches zero (as done e.g., in Lopez et al., 2013). We discuss possible choices of \( \tau(\cdot) \) in Section 4. However, trimming is introduced mainly for the asymptotic study. In practice we observe that it does not affect the results and can be avoided (see discussion in Section 5), so the practitioner does not have to worry about the choice of the trimming function. Existence and uniqueness of \( \hat{\gamma}_n \) hold under the same conditions as for the maximum likelihood estimator in the binary outcome regression model where \( 1 - \hat{\gamma}_n(X_i) \) is replaced by the outcome \( B_i \). For example, in the logistic model, it is required that \( p < n \) and the matrix of the variables \( X \) has full rank. Estimation of the latency component can then be performed by maximizing the likelihood of the mixture model

\[ \prod_{i=1}^{n} \left\{ \phi(\gamma^T X_i) f_u(Y_i|Z_i; \beta, \Lambda) \right\}^{\Delta_i} \left\{ 1 - \phi(\gamma^T X_i) + \phi(\gamma^T X_i) S_u(Y_i|Z_i; \beta, \Lambda) \right\}^{1-\Delta_i}, \]

with respect to \( \beta \) and \( \Lambda \) for \( \gamma = \hat{\gamma}_n \). Here \( f_u(t|Z; \beta, \Lambda) = -dS_u(t|Z; \beta, \Lambda)/dt \). This maximization can in practice be done via an ‘EM-type’ algorithm but now we keep \( \gamma = \hat{\gamma}_n \) fixed and only update \( \beta \) and \( \Lambda \) in each iteration. In particular, at the \( (m+1) \)th iteration, the E-step consists in computing \( W_i = \mathbb{E}[B_i|Y_i, \Delta_i, X_i, Z_i; \hat{\gamma}_n, \hat{\beta}_n^{(m)}, \hat{\Lambda}_n^{(m)}] \) and the M-step consists in updating \( \hat{\beta}_n^{(m)}, \hat{\Lambda}_n^{(m)} \) by maximizing

\[ \sum_{i=1}^{n} W_i \{ \Delta f_u(Y_i|Z_i; \beta, \Lambda) + (1 - \Delta_i) S_u(Y_i|Z_i; \beta, \Lambda) \}. \]
The iterates \( \hat{\beta}_n^{(m+1)} \) and \( \hat{\Lambda}_n^{(m+1)} \) are then computed as in the smcure package but the difference is that the weights \( W_i \) are always computed using the fixed estimator \( \hat{g}_n \).

We call this a two-step estimator because it relies on a preliminary estimator \( \hat{g}_n \), which is used to construct the 1-dimensional covariate \( \hat{g}_n^T X \). In this way, independently of the dimension of \( X \), the kernel estimator requires only one bandwidth parameter. The idea of a single-index structure is used in several papers to avoid multidimensional regression (Lopez et al., 2013; Strzalkowska-Kominiak & Cao, 2014), but has previously not been exploited in the context of cure models and presmoothing. A nonparametric estimator for the cure probability in (5) could be obtained using any nonparametric estimator of the conditional survival function as done for example in Peláez Suárez, Cao, and Vilar (2021); Peláez Suárez, Cao Abad, and Vilar Fernández (2021) for estimation of the default probability. Here we use the Beran-type estimator since it is easy to compute and exhibits a good behavior.

### 3.1 Rationale behind the new approach

By definition we have

\[
\pi_0(x) = \mathbb{P}(T = \infty | X = x) = \mathbb{P}(T = \infty | \gamma_0^T X = \gamma_0^T x).
\]

Moreover, since from our model it follows that \( T \perp C | \gamma_0^T X, \beta_0^T Z \), we have

\[
H_1(\mathbb{d}t | \gamma_0^T x, \beta_0^T z) = F_C(t, \infty) \gamma_0^T x, \beta_0^T z) F_T(\mathbb{d}t | \gamma_0^T x, \beta_0^T z),
\]

\[
H(t, \infty) | \gamma_0^T x, \beta_0^T z) = F_T(t, \infty) | \gamma_0^T x, \beta_0^T z).
\]

where \( H_1, H \) are defined as in (6) and \( F_T, F_C \) denote the distribution functions of \( T \) and \( C \) respectively. This yields

\[
\frac{F_T(\mathbb{d}t | \gamma_0^T x, \beta_0^T z)}{F_T(t, \infty) | \gamma_0^T x, \beta_0^T z)} = \frac{H_1(\mathbb{d}t | \gamma_0^T x, \beta_0^T z)}{H(t, \infty) | \gamma_0^T x, \beta_0^T z)}.
\]

As in Patiilea and Van Keilegom (2020), we obtain that

\[
\mathbb{P}(T = \infty | \gamma_0^T X = \gamma_0^T x, \beta_0^T Z = \beta_0^T z) = \prod_{t \in \mathbb{R}} \left\{ 1 - \frac{H_1(\mathbb{d}t | \gamma_0^T x, \beta_0^T z)}{H(t, \infty) | \gamma_0^T x, \beta_0^T z)} \right\},
\]

where \( \prod_{t \in \mathbb{R}} \) denotes the product integral. By the first part of (2), the product integral is also equal to \( P(T = \infty | \gamma_0^T X = \gamma_0^T x) \). Similarly, if \( T \perp C | \gamma_0^T X \), we obtain

\[
\mathbb{P}(T = \infty | \gamma_0^T X = \gamma_0^T x) = \prod_{t \in \mathbb{R}} \left\{ 1 - \frac{H_1(\mathbb{d}t | \gamma_0^T x)}{H(t, \infty) | \gamma_0^T x)} \right\},
\]

which justifies the definition of our estimator in (5). This assumption is satisfied if \( C \perp (X,Z) | \gamma_0^T X \). We restrict to this case for simplicity in order to have conditioning on only one index. However, the method can be used in general conditioning on both \( \gamma_0^T X \) and \( \beta_0^T Z \). We illustrate this through one of the simulation settings in Section 5.
4 \hspace{1em} \textbf{ASYMPTOTIC RESULTS}

In this section we focus on models where the survival function of the susceptible \( S_u \) has fixed support \([0, \tau_0]\) and there exists a constant \( c > 0 \) such that

\[
\inf_{\gamma \in G} \mathbb{P}(C > \tau_0 | \gamma^T X) > c \text{ almost surely,} \tag{8}
\]

for some compact set \( G \subset \mathbb{R}^p \) containing \( \gamma_0 \) (see Assumption C2 below). When the latency follows a Cox regression model, this is related to the identifiability assumption (I3'). Moreover, we assume that \( T \perp C | \gamma_0^T X \) but the results can be generalized as mentioned in the previous section.

Let us sketch the arguments we will use to obtain asymptotic properties of \( \hat{\gamma}_n \). Note that \( \hat{\gamma}_n \) is the maximizer of the Bernoulli-type log-likelihood in (7). Hence, the main issue is dealing with the nonparametric estimator \( \hat{\gamma}_n \), which replaces the latent binary outcome. By definition we have

\[
\hat{\gamma}_n(x) = \hat{g}_{\gamma_n}(\tilde{\gamma}_n T x),
\]

where

\[
\hat{g}_{\gamma}(u) = 1 - \hat{F}_n(\tau_0 | \gamma^T X = u) = \prod_{t \in \mathbb{R}} \left( 1 - \frac{\hat{H}_{1,t}(dt|u)}{\hat{H}_t([t, \infty)|u)} \right),
\] \tag{9}

and \( \hat{F}_n \) is a nonparametric estimator of \( F_T \). Note that the product integral actually over \( t \in \mathbb{R} \) is the same as over \( t \leq \tau_0 \) because \( \hat{H}_{1,t}(dt|u) = 0 \) for \( t > Y_{(m)} \), where \( Y_{(m)} \leq \tau_0 \) is the last observed event time. We can also write

\[
\pi_0(x) = g_{\gamma_0}(\gamma_0^T x),
\]

with

\[
g_{\gamma}(u) = 1 - F_T(\tau_0 | \gamma^T X = u) = \prod_{t \in \mathbb{R}} \left( 1 - \frac{H(t|u)}{H([t, \infty)|u)} \right). \tag{10}
\]

Hence

\[
\hat{\gamma}_n(x) - \pi_0(x) = \hat{g}_{\gamma_n}(\tilde{\gamma}_n T x) - g_{\gamma_0}(\gamma_0^T x)
= \left\{ \hat{g}_{\gamma_0}(\gamma_0^T x) - g_{\gamma_0}(\gamma_0^T x) \right\} + \left\{ \hat{g}_{\gamma_n}(\tilde{\gamma}_n T x) - \hat{g}_{\gamma_0}(\gamma_0^T x) \right\}
= \left\{ \hat{F}_n(\tau_0 | \gamma_0^T X = \gamma_0^T x) - F_T(\tau_0 | \gamma_0^T X = \gamma_0^T x) \right\}
+ \left\{ \hat{g}_{\gamma_n}(\tilde{\gamma}_n T x) - \hat{g}_{\gamma_0}(\gamma_0^T x) \right\}. \tag{11}
\]

The first term on the right-hand side of the equation, can be dealt as usual being the difference between \( \hat{F}_n \) and \( F_T \) conditionally on a one-dimensional covariate \( \gamma_0^T X \). The second term results from using \( \tilde{\gamma}_n \) instead of \( \gamma_0 \) when constructing the one-dimensional covariate \( \tilde{\gamma}_n T x \). The behavior of this term depends on the properties of the preliminary estimator \( \tilde{\gamma}_n \). We first formulate the results for a general preliminary estimator \( \tilde{\gamma}_n \) and a general parametric function \( \phi \). Then we show that, for the logistic-Cox model, the maximum likelihood estimator satisfies the required conditions.
The following assumptions are needed for consistency of $\hat{\gamma}_n$.

(C1) The preliminary estimator is consistent, that is, $\hat{\gamma}_n - \gamma_0 = o_P(1)$.
(C2) The parameter $\gamma_0$ lies in the interior of a compact set $G \subset \mathbb{R}^p$.
(C3) There exist some constants $a > 0$, $c > 0$ such that
\[
|\phi(\gamma_1^T x) - \phi(\gamma_2^T x)| \leq c|\gamma_1 - \gamma_2|^a, \quad \forall \gamma_1, \gamma_2 \in G, \forall x \in \mathcal{X},
\]
where $\| \cdot \|$ denotes the Euclidean norm and $\mathcal{X} \subset \mathbb{R}^p$ is the support of $X$.
(C4) $\inf_{\gamma \in G} \inf_{x \in \mathcal{X}} \phi(\gamma^T x) > 0$ and $\sup_{\gamma \in G} \sup_{x \in \mathcal{X}} \phi(\gamma^T x) < 1$.
(C5) For any $\gamma \in G$, the support $\mathcal{X}_\gamma$ of $X_\gamma = \gamma^T X$ is a bounded convex subset of $\mathbb{R}$. The density $f_{X_\gamma}(\cdot)$ of $X_\gamma$ is twice differentiable with bounded second derivative.
(C6) The bandwidth $b$ is such that $nb^4 \to 0$ and $nb^{3+\xi}/(\log b^{-1}) \to \infty$ for some $\xi > 0$.
(C7) The kernel $k$ is a twice continuously differentiable, symmetric probability density function with compact support.
(C8) (i) The functions $H_0([0, t]) u$, $H_1([0, t]) u$ defined in (6) are twice differentiable with respect to $u$, with uniformly bounded derivatives for all $t \leq \tau_0$, $u \in \mathcal{X}_{\gamma_0}$. Moreover, there exist continuous nondecreasing functions $L_1, L_2, L_3$ such that $L_0(0) = 0$, $L_i(\tau_0) < \infty$ and for all $t, s \in [0, \tau_0], u \in \mathcal{X}_{\gamma_0}$,
\[
|H_c(t) u - H_c(s) u| \leq |L_1(t) - L_1(s)|,
\]
\[
|H_{1c}(t) u - H_{1c}(s) u| \leq |L_1(t) - L_1(s)|,
\]
\[
\left| \frac{\partial H_c(t) u}{\partial u} - \frac{\partial H_c(s) u}{\partial u} \right| \leq |L_2(t) - L_2(s)|,
\]
\[
\left| \frac{\partial H_{1c}(t) u}{\partial u} - \frac{\partial H_{1c}(s) u}{\partial u} \right| \leq |L_3(t) - L_3(s)|,
\]
where the subscript $c$ denotes the continuous part of a function.
(ii) The number of jump points for the distribution function $F_C(t|u)$ of the censoring times given the index $\gamma_0^T X = u$, are finite and the same for all $u$. The partial derivative of $F_C(t|u)$ with respect to $u$ exists and is uniformly bounded for all $t \leq \tau_0$, $u \in \mathcal{X}_{\gamma_0}$. Moreover, the partial derivative with respect to $u$ of $F_R(t|u)$ (distribution function of the survival times $T$ given $\gamma_0^T X = u$) exists and is uniformly bounded for all $t \leq \tau_0$, $u \in \mathcal{X}_{\gamma_0}$.
(C9) The function $(x, \gamma) \mapsto g_\gamma(\gamma^T x)$ is continuously differentiable with respect to $\gamma$ and the vector $\nabla_\gamma g_\gamma(\gamma^T x)$ is continuous with respect to $(x, \gamma)$.

Assumptions (C2)–(C4), (C6)–(C8) are standard assumptions (Musta et al., 2022; Patilea & Van Keilegom, 2020; Van Keilegom & Akritas, 1999). Assumptions (C5) and (C9) are needed because we compute the nonparametric estimator using the index $\hat{\gamma}_n^T X$ instead of $\gamma_0^T X$. Such assumptions appear for example in Lopez et al. (2013).

A possible choice of the trimming function $\tau(\cdot)$ in (7) could be $\tau(\alpha) = 1_{\hat{\mathcal{X}}}(\alpha)$ if we know a set $\hat{\mathcal{X}}$ such as
\[
\inf_{\gamma \in G} \inf_{x \in \mathcal{X}} f_{X_\gamma}(\gamma^T x) = c > 0,
\]
where $f_{X_{\hat{\gamma}}}$ is defined in Assumption (C5). Otherwise, as shown in Lopez et al. (2013) one can take $\tau(\alpha) = 1_{[\hat{f}_{X_{\hat{\gamma}}}(\hat{\gamma}^T x) > c]}$ for some $c > 0$, which is asymptotically equivalent to the previous proposal. In practice, we can use $\tau(\alpha) = 1_{[\hat{f}_{X_{\hat{\gamma}}}(\hat{\gamma}^T x) > c]}$ based on the preliminary estimator $\hat{\gamma}_n$. 

Theorem 1. Assume that conditions (C1)–(C9) are satisfied. Then

$$\hat{\gamma}_n - \gamma_0 = o_P(1).$$

In order to obtain asymptotic normality of $\hat{\gamma}_n$ at rate $\sqrt{n}$ we need the following additional assumptions.

(N1) For each $x \in \mathcal{X}$, the function $\gamma \mapsto \phi(\gamma^T x)$ is twice continuously differentiable with uniformly bounded derivatives on $G \times \mathcal{X}$.

(N2) The matrix $E\left[\phi'(\gamma_0^T X) XX^T\right]$ is positive definite.

(N3) The preliminary estimator $\tilde{\gamma}_n$ is $\sqrt{n}$ consistent and such that there exists a function $\zeta$ such that

$$\tilde{\gamma}_n - \gamma_0 = 1_n \sum_{i=1}^n \zeta(Y_i, \Delta_i, X_i, Z_i) + R_n,$$

with $\|R_n\| = o_P(n^{-1/2})$ and $E[\zeta(Y, \Delta, X, Z)] = 0$.

Again (N1)-(N2) are standard assumptions, while (N3) arises from the use of the index $\tilde{\gamma}_n^T X$ instead of $\gamma_0^T X$. As a result the asymptotic variance of $\hat{\gamma}_n$ will also depend on the asymptotic variance of the preliminary estimator $\tilde{\gamma}_n$.

Theorem 2. Assume that conditions (C1)–(C9), (N1)–(N3) are satisfied. Then

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \xrightarrow{d} N(0, \Sigma_\gamma),$$

with covariance matrix $\Sigma_\gamma = \Gamma_1^{-1} V T_1^{-1}$, where $V = \text{Var}(\Psi(Y, \Delta, X, Z) + Q \zeta(Y, \Delta, X, Z))$,

$$\Psi(Y, \Delta, X, Z) = - \left\{ \frac{\Delta T \{ Y \leq s_0 \}}{H([Y, \infty] \gamma_0^T X)} - \int_0^{Y \wedge s_0} \frac{H_1(ds|\gamma_0^T X)}{H^2([s, \infty] \gamma_0^T X)} \frac{\phi'(\gamma_0^T X)}{\phi(\gamma_0^T X)} X \tau(X) \right\}. \quad (13)$$

$$Q = E \left[ \frac{\phi'(\gamma_0^T X)^2 \{ E[X^T Y_0^T X] E[X \tau(X) | Y_0^T X] - E[X^T X \tau(X)] Y_0^T X \} - E[X^T X \tau(X)] \gamma_0^T X}{\phi(\gamma_0^T X) \{ 1 - \phi(\gamma_0^T X) \}} \right],$$

and

$$\Gamma_1 = -E \left[ \left( \frac{1}{\phi(\gamma_0^T X)} + \frac{1}{1 - \phi(\gamma_0^T X)} \right) \phi'(\gamma_0^T X)^2 XX^T \tau(X) \right]. \quad (14)$$

Given the complicated form of the covariance matrix $\Sigma_\gamma$, we suggest using a bootstrap procedure for estimation of the SEs as done also for the maximum likelihood estimator of a semiparametric mixture cure model.

If we consider the particular case of a logistic-Cox mixture cure model and take the maximum likelihood estimator as a preliminary estimator $\tilde{\gamma}_n$, then assumptions (C2)–(C4), (C9), (N1), and (N2) are obviously satisfied for the logistic model. Moreover (C1) and (N2) are satisfied if the
cumulative baseline function $\Lambda_0$ is strictly increasing and continuously differentiable under the condition

$$\inf_z \mathbb{P}(T_u \geq \tau_0|Z = z) > 0,$$

see Thms. 2 and 3 in Lu (2008). Then, from Theorems 1 and 2 it follows that the two-step estimator is also consistent and $\sqrt{n}$-convergent. If we continue estimation of the latency submodel using this estimator of $\gamma_0$, then the resulting estimator of $\beta_0$ and $\Lambda_0$ have the desired asymptotic behavior as in Thms. 2 and 4 in Musta et al. (2022). The proof remains the same given that they only use consistency and the asymptotic i.i.d. expression of the estimator as in assumption (N3). As in Musta et al. (2022), we expect to lose in terms of efficiency of the estimators since estimation of the incidence parameters is performed without using the information on the latency model. However, efficiency is an asymptotic concept, hence less relevant in this context since our purpose is to improve estimation for small and moderate sample sizes. For large $n$, the MLE performs well and we do not expect to do better than that.

5 | SIMULATION STUDY

In this section we investigate the finite-sample behavior of the proposed two-step approach in the logistic-Cox mixture cure model and compare it with the maximum likelihood estimator implemented in the R package smcure and the method of Musta et al. (2022). We use the smcure estimator as preliminary estimator $\tilde{\gamma}_n$ for the new method.

We make some standard and common choices when computing the nonparametric estimator in (5). The kernel function $k$ is taken to be the Epanechnikov kernel $k(u) = (3/4)(1 - u^2)I_{|u| \leq 1}$. Using the preliminary estimator $\tilde{\gamma}_n$, we compute the smoothing bandwidth by cross-validation as implemented in the R package np for kernel estimators of conditional distribution functions, in our case for estimation of $H = H_0 + H_1$ given $\tilde{\gamma}_TX$. In addition, we restrict to the interval $[0, Y(m)]$, where $Y(m)$ is the last observed event time since the estimator of the cure probability $\hat{\pi}$ in (5) is essentially a product over values of $t$ that are equal to the observed event times. This means that we use the cross-validation bandwidth for estimation of the conditional distribution $H(t|\tilde{\gamma}_TX)$ for $t \leq Y(m)$. We could use a trimming function

$$\tau(x) = 1 \{\tilde{\gamma}_X(x) \geq c\},$$

for some small value of $c$ as proposed in Section 4. However, we observe that in practice this does not affect the results since $c$ can be chosen arbitrarily small. Hence, we do not do any trimming so there is no need to choose a trimming constant.

We consider four different models and for each of them three scenarios, covering a wide range of settings with different number and choice of covariates (continuous and discrete), different cure and censoring rate, and different censoring mechanisms (independent of covariates, depending on the same index as the incidence model, depending on both indexes of the incidence and latency). The models are as follows.

**Model 1.** Both incidence and latency depend on two independent continuous covariates $X_1 = Z_1 \sim N(0, 1)$ and $X_2 = Z_2 \sim \text{Unif}(-1, 1)$. We generate the cure status $B$ as a Bernoulli random variable with success probability $\phi(\gamma^TX)$ where $\phi$ is the logistic function in (3) and $\gamma = (\gamma_0, 1.5, 1.5)$. 
The survival times for the uncured observations are generated according to a Weibull proportional hazards model

\[ S_u(t \mid z) = \exp \left( -\mu t^\rho \exp \left( \beta^T z \right) \right), \]

and are truncated at \( \tau_0 = 15 \) for \( \rho = 0.75, \mu = 1.5 \) and \( \beta = (0.5, 0.3) \). The censoring times are independent from \( X \) and \( T \). They are generated from the exponential distribution with parameter \( \lambda_C \) and are truncated at \( \tau = 17 \).

**Model 2.** Both incidence and latency depend on three independent covariates \( X_1 = Z_1 \sim N(0,1) \), \( X_2 = Z_2 \sim \text{Bernoulli}(0.3) \) and \( X_3 = Z_3 \sim \text{Bernoulli}(0.7) \). The cure status and the survival times for the uncured observations are generated as in Model 1 for \( \gamma = (\gamma_0, -1, 1, -0.3) \), \( \beta = (-0.8, 1.5, -0.5) \), \( \rho = 0.75 \), \( \mu = 1.5 \), and \( \tau_0 = 7 \). The censoring times are generated according to a Weibull proportional hazards model

\[ S_C(t \mid x) = \exp \left( -\lambda_C \mu t^\rho \exp \left( \gamma^T x \right) \right), \]

for various choices of \( \lambda_C \) and are truncated at \( \tau = 9 \).

**Model 3.** For the incidence we consider four independent covariates: \( X_1 \sim N(0,1) \), \( X_2 \sim \text{Unif}(-1,1) \), \( X_3 \) and \( X_4 \) are Bernoulli random variables with parameters 0.4 and 0.6, respectively. The latency depends on three covariates: \( Z_1 \sim N(0,1) \), \( Z_2 = X_2 \) and \( Z_3 = X_4 \). The cure status and the survival times for the uncured observations are generated as in Model 1 for \( \gamma = (\gamma_0, -0.3, 0.8, 0.5, -1) \), \( \rho = 0.75 \), \( \mu = 1.5 \), \( \beta = (0.1, 0.4, -0.2) \) and \( \tau_0 = 10 \). The censoring times are generated according to a Weibull proportional hazards model

\[ S_C(t \mid x) = \exp \left( -\lambda_C \mu t^\rho \exp \left( 0.4\gamma^T x + 0.5\beta^T z \right) \right), \]

for various choices of \( \lambda_C \) and are truncated at \( \tau = 12 \).

**Model 4.** For the incidence we consider five independent covariates: \( X_1 \sim N(0,1) \), \( X_2 \sim \text{Unif}(-1,1) \), \( X_3 \) is Binomial with parameters 2 and 0.5, \( X_4 \) and \( X_5 \) are Bernoulli random variables with parameters 0.4 and 0.6, respectively. The latency depends on three covariates: \( Z_1 \sim N(0,1) \), \( Z_2 = X_3 \) and \( Z_3 = X_4 \). The cure status and the survival times for the uncured observations are generated as in Model 1 for \( \gamma = (\gamma_0, -0.8, 0.3, -0.4, 0.5, 0.6) \), \( \rho = 0.75 \), \( \mu = 1.5 \), \( \beta = (0.2, -0.5, 0.3) \) and \( \tau_0 = 7 \). The censoring times are independent from \( X \), \( Z \), and \( T \). They are generated from the exponential distribution with parameter \( \lambda_C \) and are truncated at \( \tau = 9 \).

To summarize, the main differences between the four models are the following:

- Model 1 contains two continuous covariates, same for both incidence and latency and censoring is independent of the covariates.
- Model 2 contains one continuous covariate and two discrete, same for both incidence and latency, and censoring depends on the same index \( \gamma^T X \) as the incidence.
- Model 3 contains two continuous covariates and two discrete for the incidence and only some of them affect the latency. Censoring depends on both indices \( \gamma^T X \) and \( \beta^T Z \).
- Model 4 contains two continuous covariates and three discrete for the incidence and there is an additional independent covariate that only affects the latency. Censoring is independent of the covariates.
**Table 1** Parameter values and model characteristics for each scenario.

<table>
<thead>
<tr>
<th>Model</th>
<th>Scenario</th>
<th>$\gamma_0$</th>
<th>$\lambda_C$</th>
<th>Cens. rate</th>
<th>Cure rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.4</td>
<td>36%</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.6</td>
<td>0.4</td>
<td>50%</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-0.5</td>
<td>0.3</td>
<td>63%</td>
<td>58%</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.6</td>
<td>1/35</td>
<td>30%</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.4</td>
<td>1/20</td>
<td>45%</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-0.6</td>
<td>1</td>
<td>75%</td>
<td>60%</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1/9</td>
<td>35%</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.9</td>
<td>1/7</td>
<td>50%</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-0.1</td>
<td>1/7</td>
<td>65%</td>
<td>60%</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1.5</td>
<td>0.6</td>
<td>25%</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.3</td>
<td>0.3</td>
<td>55%</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-0.6</td>
<td>0.4</td>
<td>70%</td>
<td>60%</td>
</tr>
</tbody>
</table>

For the four models we choose the values of the unspecified parameters $\gamma_0$ and $\lambda_C$ in such a way that the cure rate is around 20%, 40%, or 60% and the difference between the cure and the censoring rate is around 5%, 10%, or 15%. The specification of the parameters and the corresponding censoring and cure rates are given Table 1. The truncation of the survival and censoring times on $[0, \tau_0]$ and $[0, \tau]$ is made in such a way that $\tau_0 < \tau$ and condition (15) is satisfied but in practice it is unlikely to observe event times at $\tau_0$. In this way, we try to find a compromise between theoretical assumptions and real-life scenarios. Model 3 illustrates the behavior of the method when the censoring times depends on both the indexes of the incidence and latency models (which was assumed for simplicity in the theoretical study).

We consider samples of size $n = 200$ and $n = 400$ since we aim to provide a method that improves upon the maximum likelihood estimator for small and moderate sample size. For each configuration 1000 datasets were generated and the estimators of $\beta$ and $\gamma$ were computed through smcure and the proposed two-step approach. We report complete results in terms of bias, variance and MSE of the estimators, computed over the iterations for which the smcure procedure converges, in Tables S1–S5. For models 1 and 4 we also report the results for the presmoothing estimator proposed in Musta et al. (2022), which uses multivariate kernel smoothing. To simplify the exposition, here we just present a summary of the results and illustrate the behavior in some of the settings, specifically for sample size 200 and Models 1, 2, and 4, shown, respectively, in scenarios 1, 3, and 2. These settings correspond to three different cure rates and high censoring. Boxplots of the $\gamma$ estimates using the smcure package, the presmoothing method of Musta et al. (2022) and the proposed two-step approach are shown in Figures 1–3. $\beta$ estimates are almost the same for all methods, see Tables S1–S5. In some scenarios, mainly corresponding to the ones with 15% additional censoring compared to the cure rate and smaller sample size, the iterative procedure of the EM algorithm in smcure does not converge. The most problematic setting in this regard is Model 2 scenario 3, for which 73/1000 iterations do not converge for $n = 200$ and 36/1000 for $n = 400$. The boxplots of the estimators for both methods in these non-convergent iterations are shown in Figure 4. We have limited the range of the values to improve readability, as a result of which in each of the boxplots for smcure and $n = 200$ two extreme points have been left out. In the other settings, only around 2% or less of the iterations do not converge.
From the simulation results we first notice that the multivariate presmoothing approach of Musta et al. (2022) often leads to biased estimators and the bias does not go away as the sample size increases (see for example Model 1 in Table S1 and $\gamma_2$ for Model 4 in Tables S4 and S5). In addition, this procedure is computationally expensive since it requires computation of multiple bandwidths via cross-validation. We hence do not recommend its use for multivariate covariates. Regarding \texttt{smcure} and the proposed two-step approach we make the following observations:

- The two-step approach improves considerably upon \texttt{smcure} for estimation of $\gamma$ when $n = 200$ and the censoring rate among the uncured observations is higher.
- In almost all scenarios the two-step approach has a smaller variance, which is expected due to presmoothing, but it often exhibits also lower bias.
- As the sample size increases or the censoring rate decreases, we see less difference between the two methods.
- In terms of $\beta$ estimators, both approaches give very similar results.
- Even when \texttt{smcure} does not converge, the two-step approach usually gives more reasonable estimates than \texttt{smcure}.

One might think that the proposed procedure might be iterated further beyond two steps, where the $\hat{\gamma}_n$ estimator from maximizing (7) is used to reconstruct the one-dimensional covariate.
used in (5), after which (7) is remaximized and so on. We conducted a small simulation study to investigate this and observe that the results remain quite stable and in any case there is no gain in repeating the procedure (see Table S6). Moreover, since the second step of the new method does not depend on the latency model, we expect it to be more stable than smcure with respect to misspecifications of the latency model. We investigate this issue by considering two additional settings: one corresponding to a non-Cox latency model (Model 5 below) for which we still apply the two methods as if the Cox model was true and one corresponding to a logistic-Cox model but in which we do not use the correct covariates. For the latter, we use Model 4, scenario 2 described above but fit a latency model with covariates \( X_1, \ldots, X_5 \) instead of \( Z_1, Z_2, Z_3 \). In particular, this means that we are including covariates \( X_1, X_2, X_3 \) that actually do not have any effect and excluding \( Z_1 \) which effects the survival of the uncured.

**Model 5.** Both incidence and latency depend on three independent covariates \( X_1 = Z_1 \sim \text{Unif}(-1, 1), X_2 = Z_2 \sim \text{Bernoulli}(0.4) \) and \( X_3 = Z_3 \sim \text{Bernoulli}(0.6) \). The cure status and the survival times for the uncured are generated as in Model 1 for \( \gamma = (1.4, 2.1, -1) \), \( \rho \) depending on \( z \), \( \rho(z) = 0.75 + \exp(\beta^T z) \), \( \mu = 1.5 \), \( \beta = (1, 0.4, -0.6) \). In particular this means that the latency model does not satisfy the proportional hazards assumption. For an observation with covariate

![Figure 2](https://onlinelibrary.wiley.com/doi/10.1111/sjos.12713)
FIGURE 3  Boxplots of estimates of the six $\gamma$ parameters for Model 4, scenario 2 and $n = 200$. The horizontal lines correspond to the true values of the parameters. Top line: $\gamma_1, \gamma_2, \gamma_3$. Bottom line: $\gamma_4, \gamma_5, \gamma_6$.

$z$, the event time is truncated at $\tau_0(z)$ equal to the 97% quantile of the Weibull distribution with parameters $\rho(z)$ and $\mu^{-1/\rho(z)}$. The censoring times are generated according to a Weibull proportional hazards model

$$S_C(t|x) = \exp\left(-\frac{1}{22} \mu^\hat{\rho} \exp(\gamma^Tx)\right),$$

with $\mu = 1.5$ and $\hat{\rho} = 2.5$, truncated at $\tau = \max_z \tau_0(z) + 2$. This corresponds to a cure rate of 30% and a censoring rate of 45%.

Results for sample size 200 and 400, reported in Table S7, show that when the true latency model is not a Cox proportional hazards model, even $\gamma$ estimates are biased. However, the two-step approach has lower bias and MSE, hence suffers less from the misspecification of the latency. On the other hand, misspecification of the latency covariates when the model is still Cox, seems to be less critical. It leads to a slight increase in bias and variance compared to the results in Tables S4 and S5 but again the two-step approach performs better.

6  |  APPLICATION

In this section we illustrate the practical use of the method through two medical datasets for melanoma cancer patients and compare the results with those provided by the smcure package. Melanoma is a common skin cancer type for which nowadays it is expected that a considerable
fraction of the patients get cured as a consequence of medical advances in diagnostics and treatment. Therefore, it is important to account for the presence of cured patients in the statistical analysis of melanoma survival data and to evaluate new treatments focusing on cure and not only survival prolongation.

6.1 Eastern Cooperative Oncology Group data

The Eastern Cooperative Oncology Group (ECOG) phase III clinical trial c1684 aimed at evaluating the effect of treatment (high dose interferon alpha-2b regimen) as the postoperative adjuvant therapy for melanoma patients. The corresponding dataset, consisting of 284 observations (after deleting missing data), is available in the \texttt{smcure} package (Cai et al., 2012). The event time is the time from initial treatment to recurrence of melanoma and three covariates have been considered: age (continuous variable centered to the mean), gender ($0 = \text{male}$ and $1 = \text{female}$) and treatment ($0 = \text{control}$ and $1 = \text{treatment}$). Around 30\% of the observations are censored. The Kaplan–Meier curve is shown in Figure 5.

We fit a logistic-Cox mixture cure model by using the maximum likelihood principle (\texttt{smcure} package) and the proposed two-step approach. For our method we use the \texttt{smcure} estimator as a preliminary estimator. In both cases, SEs are computed through 500 naive bootstrap samples. The resulting parameter estimates, SEs and corresponding $p$-values for the Wald test are reported in Table 2.
FIGURE 5 Left panel: Kaplan–Meier survival curve for Eastern Cooperative Oncology Group (ECOG) data. Right panel: Kaplan–Meier survival curves for the treatment group (solid) and control group (dotted) in the ECOG data.

TABLE 2 Results for the incidence (logistic component) and the latency (Cox PH component) from the Eastern Cooperative Oncology Group data.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>smcure package</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimates</td>
<td>SE</td>
</tr>
<tr>
<td>Incidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1.3649</td>
<td>0.3457</td>
</tr>
<tr>
<td>Age</td>
<td>0.0203</td>
<td>0.0159</td>
</tr>
<tr>
<td>Gender</td>
<td>−0.0869</td>
<td>0.3347</td>
</tr>
<tr>
<td>Treatment</td>
<td>−0.5884</td>
<td>0.3706</td>
</tr>
<tr>
<td>Latency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>−0.0077</td>
<td>0.0069</td>
</tr>
<tr>
<td>Gender</td>
<td>0.0994</td>
<td>0.1932</td>
</tr>
<tr>
<td>Treatment</td>
<td>−0.1535</td>
<td>0.1715</td>
</tr>
</tbody>
</table>

We observe that, despite exhibiting the same effect directions for all covariates, the two approaches give quite different results in terms of treatment effect. Age and treatment are both found to have a significant effect on the cure fraction when using the two-step method, while smcure does not detect any significant effect. The same data were also analyzed by Musta et al. (2022) and we observe that the two-step approach agrees with the multivariate presmoothing in terms of sign of the estimates and significance of the covariates. Since the two-step approach focuses on estimation of the incidence component, we also compare the two methods in terms of prediction accuracy for the cure probabilities. We estimate the expected prediction error (EPECP) as proposed in Jiang et al. (2017) via 10-fold cross-validation and 100 replications. The EPECP is defined as $E[(B - \phi(\hat{\gamma}_T X)]^2$ and is estimated by

$$\frac{1}{n} \sum_{k=1}^{10} \sum_{j \in \text{fold } k} \frac{1 - \mathbb{1}_{[Y_j < Y_{(m)}, \Delta = 0]}}{\hat{G}_n(Y_j \land Y_{(m)})} \left( B_j - \phi(\hat{\gamma}_{n-j} X_j) \right)^2,$$

where $\hat{G}_n$ is the Kaplan–Meier estimate of the survival function for censoring times, $Y_{(m)}$ is the last observed event time and $\hat{\gamma}_{n-j}$ denotes the estimator of $\gamma$ based on the observed data with the jth
fold removed. Note that for this we only need to know $B_j$ for the uncensored observations (which are uncured) and for the censored observations in the plateau (which are assumed to be cured).

The boxplot of the difference between the log of EPECP of the new method and the log of EPECP of `smcure` is given in the left panel of Figure 6. The log-scale is chosen to improve readability.

We also evaluate the performance based on the Bernoulli likelihood as follows. We split the data into a training and a test set (at a 2:1 ratio), fit the model in the training set and then compute the prediction error for the test set according to the formula

$$PE = - \sum_{j \in \text{test set}} \log \left[ \phi(\hat{\gamma}_n^T X_j) \hat{w}_j \{1 - \phi(\hat{\gamma}_n^T X_j)\}^{1 - \hat{w}_j} \right],$$

where $\hat{\gamma}_n$ are the parameter estimates from the training set and $\hat{w}_j$ are the predicted uncure probabilities given the observations, that is,

$$\hat{w}_j = \Delta_j + (1 - \Delta_j) \frac{\phi(\hat{\gamma}_n^T X_j) \hat{S}_u(Y_j | Z_j)}{1 - \phi(\hat{\gamma}_n^T X_j) + \phi(\hat{\gamma}_n^T X_j) \hat{S}_u(Y_j | Z_j)}.$$

We repeat this procedure 1000 times, for random selection of the train and test set. The boxplot of the difference between the log of PE of the new method and the log of PE of `smcure`, over these 1000 iteration, is given in the right panel of Figure 6. We observe that the two-step approach leads to lower PE and EPECP (negative difference) in more than 50% of the cases and the improvement in PE for the new method is usually larger compared to the cases in which `smcure` does better.

In addition, we expect the new approach to be more stable with respect to the latency model since that does not influence the second step of the estimation. To illustrate this point, we also fit a cure model with only gender as covariate for the survival of uncured patients (see Table 3) and see that in that case `smcure` also detects the effect of the treatment to be significant.

6.2 Surveillance, epidemiology and end results database

Here we consider melanoma data extracted from the Surveillance, Epidemiology and End Results (SEER) database to illustrate the performance of the method for more than one continuous
TABLE 3 Results for the incidence (logistic component) and the latency (Cox PH component) from the Eastern Cooperative Oncology Group (ECOG) data.

| Covariates | smcure package | | | Proposed method | | |
|------------|----------------|----------------|----------------|----------------|----------------|
|            | Estimates      | SE             | p-Value        | Estimates      | SE             | p-Value        |
| Incidence  | Intercept      | 1.4000         | 0.2791         | 5 · 10^{-7}    | 1.9073         | 0.5225         | 0.0002         |
|            | Age            | 0.0165         | 0.0121         | 0.1709         | 0.0357         | 0.0174         | 0.0399         |
|            | Gender         | -0.0538        | 0.3101         | 0.8623         | -0.0429        | 0.3979         | 0.9141         |
|            | Treatment      | -0.6765        | 0.3118         | 0.0300         | -1.2590        | 0.5809         | 0.0302         |
| Latency    | Gender         | 0.0637         | 0.1935         | 0.7421         | 0.0235         | 0.1924         | 0.9028         |

FIGURE 7 Kaplan–Meier survival curve for the surveillance, epidemiology and end results data.

covariates. The SEER database collects cancer incidence data from population-based cancer registries in United States. We select the database “Incidence – SEER 18 Regs Research Data” and, in order to have a reasonable sample size, we extract the melanoma cancer data for the county of San Francisco in California during the period 2005–2010. We consider only patients with known follow-up time and tumor size (in the range 1–90 mm) and restrict the study to white people because of the very small number of cases from other races. The event of interest is death because of melanoma. This cohort consists of 384 observations out of which 228 are male. The age ranges from 23 to 101 years, the follow-up from 1 to 143 months with no events observed after 108 months. Because of the high expected cure rate, 89% of the observations are censored. We consider as covariates in the model: gender (0 = male, 1 = female), age and tumor size (continuous). The use of cure models is justified from the presence of a long plateau containing around 25% of the observations (see the Kaplan–Meier curve in Figure 7).

We compute parameter estimates, standard errors and corresponding p-values for the two-step approach, smcure and the multivariate presmoothing proposed in Musta et al. (2022) (see Table 4). We observe that all three methods agree on the directions of the effects and give similar parameter estimates. Note that the sample size in this case is larger than in the previous data example. However, smcure only finds age to be significant while the other two methods also detect the tumor size. In addition, presmoothing finds age to be significant at 0.05 level. For the latency, none of the covariates are found significant with all methods. In this case, the new
TABLE 4 Results for the incidence (logistic component) and the latency (Cox PH component) from the surveillance, epidemiology and end results data.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>smcure package</th>
<th>Proposed method</th>
<th>Musta et al. (2022)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>p-Value</td>
</tr>
<tr>
<td>Incidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−4.4952</td>
<td>0.9670</td>
<td>3 \cdot 10^{-6}</td>
</tr>
<tr>
<td>Age</td>
<td>0.0411</td>
<td>0.0142</td>
<td>0.0037</td>
</tr>
<tr>
<td>Gender</td>
<td>−0.6805</td>
<td>0.4551</td>
<td>0.1349</td>
</tr>
<tr>
<td>Tumor size</td>
<td>0.0214</td>
<td>0.0136</td>
<td>0.1147</td>
</tr>
<tr>
<td>Latency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>−0.0161</td>
<td>0.0152</td>
<td>0.2874</td>
</tr>
<tr>
<td>Gender</td>
<td>−0.3599</td>
<td>0.4217</td>
<td>0.3934</td>
</tr>
<tr>
<td>Tumor size</td>
<td>0.0308</td>
<td>0.0224</td>
<td>0.1695</td>
</tr>
</tbody>
</table>

FIGURE 8 Left: Boxplot of the difference between the log of expected prediction error (EPECP) of the proposed method and the log of EPECP of smcure, over these 100 replications for the surveillance, epidemiology and end results (SEER) data. Right: Boxplot of the difference between the log of PE of the proposed method and the log of PE of smcure, over these 1000 iterations for the SEER data.

FIGURE 9 Left: Boxplot of the difference between the log of expected prediction error (EPECP) of the proposed method and the log of EPECP of multivariate presmoothing Musta et al. (2022), over these 100 replications for the surveillance, epidemiology and end results (SEER) data. Right: Boxplot of the difference between the log of PE of the proposed method and the log of PE of multivariate presmoothing Musta et al. (2022), over these 1000 iterations for the SEER data.
### TABLE 5
Results for the incidence (logistic component) and the latency (Cox PH component) from the surveillance, epidemiology and end results (SEER) data.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>smcure package</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimates</td>
<td>SE</td>
</tr>
<tr>
<td>Incidence</td>
<td>Intercept</td>
<td>-4.7297</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>0.0423</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>-0.6393</td>
</tr>
<tr>
<td></td>
<td>Tumor size</td>
<td>0.0288</td>
</tr>
<tr>
<td>Latency</td>
<td>Age</td>
<td>-0.0165</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>-0.3208</td>
</tr>
</tbody>
</table>

method leads to a slight improvement compared to *smcure* in terms of prediction errors, computed as described in the previous subsection (see Figure 8). On the other hand, multivariate presmoothing behaves worse than the proposed method in terms of EPECP but slightly better in terms of PE, see Figure 9. In addition, we also observe that if we remove the covariate tumor size from the latency model, the proposed two-step approach gives similar results while this time *smcure* detects also tumor size as significant for the incidence component (see Table 5). Once more, this behavior reflects the strong dependence of the incidence estimates on the latency model for *smcure*.

### 7 | DISCUSSION

In this article we considered estimation of the mixture cure model with a parametric model for the incidence and a semiparametric model for the latency, in presence of multivariate covariates. Our goal is to improve upon the standard maximum likelihood estimator, computed via the iterated EM algorithm, for small and moderate sample sizes that are usually encountered in practice. We proposed a new estimation method relying on the idea of presmoothing, which is expected to reduce variance of the estimates with finite samples. The novelty consists in using presmoothing as a second-stage estimator, while in the first step a preliminary estimator is used to project the covariates into an one-dimensional index. Based on this pseudo-covariate, we first compute a nonparametric incidence estimator and then maximize a logistic-type likelihood with the unobserved non-cure indicators replaced by the nonparametric estimates. In this way, the new method does not require multidimensional smoothing, which is a great advantage compared to the method proposed in Musta et al. (2022) because it guarantees consistency even for dimensions larger than one and reduces drastically the computational time. We use the MLE computed via the *smcure* package as a preliminary estimator, but the method would work in general for any available estimator. We studied the asymptotic properties of the estimators and showed through simulation studies that it has better performance than the MLE for small and moderate sample sizes. We also observed that the multivariate presmoothing, despite having smaller variance, often leads to biased estimators for dimension larger than one. This confirms the advantage of the two-step procedure. We also note that kernel smoothing is performed using a standard choice of the bandwidth computed via cross-validation as implemented in the np-package. This is observed to work well and, since smoothing is only an intermediate step, the choice of the bandwidth is not crucial. Finally, despite focusing mainly on the incidence component, the latency is also
estimated consequently by maximizing the corresponding component of the likelihood. However, simulation results show that the estimators remain quite stable independently of the method used for incidence estimation.

**ACKNOWLEDGMENTS**

Van Keilegom acknowledges financial support from the European Research Council (2016–2021, Horizon 2020 and grant agreement 694409). For the simulations we used the Lisa cluster of the Dutch national Supercomputer. The data that were analyzed in Section 5 have been provided by the SEER programme (https://www.seer.cancer.gov) research data (1973–2014), National Cancer Institute, Division of Cancer Control and Population Sciences, Surveillance Research Program.

**ORCID**

Ingrid Van Keilegom https://orcid.org/0000-0001-8827-7642

**REFERENCES**


SUPPORTING INFORMATION

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