Single-index mixture cure model under monotonicity constraints

Musta, E.; Yuen, T.P.

DOI
10.1214/24-EJS2273

Publication date
2024

Document Version
Final published version

Published in
Electronic Journal of Statistics

Electronic Journal of Statistics
Vol. 18 (2024) 3376–3436
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2273

Single-index mixture cure model under monotonicity constraints

Eni Musta and Tsz Pang Yuen

Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Netherlands,
e-mail: e.musta@uva.nl; t.p.yuen@uva.nl

Abstract: We consider survival data in the presence of a cure fraction, meaning that some subjects will never experience the event of interest. We assume a mixture cure model consisting of two sub-models: one for the probability of being uncured (incidence) and one for the survival of the uncured subjects (latency). Various approaches, ranging from parametric to nonparametric, have been used to model the effect of covariates on the incidence, with the logistic model being the most common one. We propose a monotone single-index model for the incidence and introduce a new estimation method that is based on the profile maximum likelihood approach and techniques from isotonic regression. The monotone single-index structure relaxes the parametric logistic assumption while maintaining interpretability of the regression coefficients. We investigate the consistency of the proposed estimator and show through a simulation study that, when the monotonicity assumption is satisfied, it performs better compared to the non-constrained single-index/Cox mixture cure model. To illustrate its practical use, we use the new method to study melanoma cancer survival data.

MSC2020 subject classifications: Primary 62N02.
Keywords and phrases: Survival analysis, mixture cure model, single-index model, isotonic estimation, kernel smoothing.

Received March 2023.

Contents

arXiv: 2211.09464

1. Introduction

Modelling time-to-event data in the presence of subjects that will never experience the event of interest has gained popularity over the recent decades. For instance, the advancement of cancer treatments has led to a larger fraction of patients being cured of their diseases [20]. Another example can be found in fertility studies [39], where one is interested in the time to pregnancy while there are infertile couples for whom natural conception is impossible. Cure models have also been utilized in credit scoring to model the time to default of a loan applicant and default does not occur for the majority of debtors [9]. In all these scenarios, the subjects that are immune to the event of interest are referred to as 'cured' (non susceptible). There are two types of cure models: mixture cure models and promotion time models. We refer the reader to [1] and [32] for a comprehensive review of these models.

Mixture cure models assume that the population is a mixture of cured and susceptible subjects and consist of two sub-models: one for the uncured probability (incidence) and one for the conditional survival function of the susceptibles (latency). Initially, fully parametric models with logistic regression assumption for the incidence and different parametric distributions for the latency were proposed [10, 11]. Later on, extensions to semi-parametric models for the latency, such as the Cox proportional hazard (PH) model [31, 34] and the accelerated failure time model [21, 41], were introduced. For the incidence, a nonparametric estimator was developed in [40] based on the Beran estimator for the conditional survival function. However, such method is problematic for multivariate covariates since it requires multi-dimensional smoothing. To circumvent the curse-of-dimensionality, a single-index model for the incidence was introduced in [2], linking a linear predictor (index) to the incidence probability via an unspecified link function. The index achieves dimension reduction and alleviates the dimensionality issues when estimating the link function nonparametrically. In certain contexts, the link function is expected to be monotone, meaning that the cure probability increases/decreases as the risk score of an individual, given by the index, increases. The widely used logistic model for the incidence is in particular a monotone single-index model. In such cases, it is more appropriate to estimate the link function under monotonicity constraints, while the estimate proposed in [2] is not guaranteed to be monotone. Another advantage of a monotone link function is interpretability since the sign of the coefficients of the index tells us whether a given covariate increases or decreases the cure chances. This motivates us to investigate estimation of a single-index model for the incidence

under monotonicity constraints. We focus mainly on the incidence component and assume a Cox PH model for the latency.

Single-index models have been thriving because of their flexibility over linear and parametric models, while avoiding dimensionality problems of general nonparametric models. The monotonicity of the link function appears in numerous applications leading to the popularity of generalized linear models. Therefore, it is appealing to impose monotonicity constraint on the link function, which leads to monotone single-index models and has recently become an active research area [5, 12, 3, 4]. It is also worth mentioning that the binary choice model in econometrics and the current status linear regression model [12] are special cases of the monotone single-index model. More in general, there has been a growing interest in statistical inference under shape-constraints, such as monotonicity, convexity, log-concavity, etc., which arise naturally in a wide range of applications [14]. One advantage of such methods is that they allow for nonparametric estimation without using tuning parameters. However, a combination of smoothing and shape-constrained estimation often leads to better finite sample performance [23, 24].

In this paper we introduce an estimation method for the monotone single-index mixture cure model that is based on the profile maximum likelihood principle and techniques from isotonic regression in combination with kernel smoothing. Despite the fact that the monotone single-index model has already been studied in the literature, its use within the mixture cure model has some unique features that make the problem more challenging (see Section 3 for a more detailed discussion). First, in contrast to the standard monotone single-index models, including the current status linear regression model, where the response is directly observed, the cure status in mixture cure model is unknown for the censored subjects. As a result, the likelihood has a more complicated expression and iterative procedures such as the EM algorithm are required to solve the optimization problem. Secondly, apart from the coefficients of the index and the link function, our model contains additional parameters (finite and infinite dimensional) because of the extra latency component. Challenges also arise when studying the theoretical properties of the estimators, which are established in a less straightforward manner as compared to the monotone single-index model or the current status model, see Section 4 for a more detailed discussion. We study consistency of the proposed estimator, which to our best knowledge has not been investigated even for the single-index mixture cure model without monotonicity constraints [2]. We illustrate through simulations that imposing monotonicity improves the behavior of the estimator and makes it more stable with respect to the choice of the bandwidth compared to the smooth non-monotone estimator. In particular, one does not need to use time consuming bandwidth selection procedures since a simple bandwidth choice performs reasonably well. In addition, we observe that our decision to incorporate a smoothing step to the isotonic estimation of the link function is motivated by its improved practical performance for finite sample sizes.

To summarize, the main contributions and innovative aspects of the present paper are the following:

- We introduce a novel model for the incidence of mixture cure models using a monotone single-index. This differs from existing literature, which relies on parametric approaches or nonparametric methods without imposing monotonicity constraints. Our model relaxes the parametric assumption, while maintaining interpretability of the regression coefficients.
- This paper extends the applicability of monotone single-index models to a more complicated setting, where the response variable is not observed and there are additional finite and infinite dimensional model parameters.
- The proposed estimation method is a novel combination of several techniques including the profile maximum likelihood approach, expectation maximization, shape-constrained estimation methods and kernel smoothing. The combination of monotonicity and smoothing simplifies the bandwidth selection procedure eliminating the need for computationally intensive methods like cross-validation.
- We provide consistency results of the proposed estimators, which are not available in the existing literature even for single-index mixture cure models without monotonicity constraint.
- The improved finite sample behavior of the proposed method compared to methods relying on just smoothing or monotonicity alone is illustrated via a simulation study.

The paper is organized as follows. Section 2 describes the monotone single-index mixture cure model and the conditions for model identification. Section 3 introduces the estimation procedure, while Section 4 focuses on establishing consistency of the estimator. The finite sample properties of the proposed method are investigated through a simulation study and the results are reported in Section 5. Finally, an illustration of the practical use through a study of a medical dataset of melanoma cancer patients is provided in Section 6. The proofs and additional simulation results can be found in the Appendix. Software in the form of R code is available on the GitHub repository `https://github.com/tp-yuen/msic`.

2. Model description

Let T be a nonnegative random variable denoting the survival time, i.e., time until occurrence of an event of interest, which can be equal to infinity indicating the possibility of cure. Let $B = \mathbb{1}\{T < \infty\}$ be the uncured status: $B = 1$ if the subject is uncured, i.e. susceptible to the event of interest, and $B = 0$ if the subject is cured, i.e. not susceptible. Under the assumption that the survival time is subject to random right censoring, we observe the follow-up time $Y = \min(T, C)$ and the censoring indicator $\Delta = \mathbb{1}\{T \leq C\}$, where C is the censoring time. Since the duration of the studies is in practice limited, we assume that C has bounded support. As a result of censoring, the uncured status B is a latent variable and the cured subjects cannot be distinguished from the censored uncured ones. Since the cure probability and the survival time of the uncured do not necessarily depend on the same predictors, we use two sets of

covariates $\mathbf{X} \in \mathbb{R}^d$ for the incidence and $\mathbf{Z} \in \mathbb{R}^q$ for the latency. $\mathbf{X}$ and $\mathbf{Z}$ can possibly be the same or partially/completely different. We assume that C and T are conditionally independent given the covariates $(\mathbf{X}, \mathbf{Z})$, which is a rather standard assumption in survival analysis.

In the mixture cure model the survival function is given by

$$\begin{aligned} S(t|\mathbf{x},\mathbf{z}) &= \mathbb{P}(T > t|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}) \\ &= \mathbb{P}(B=0|\mathbf{X}=\mathbf{x}) + \mathbb{P}(B=1|\mathbf{X}=\mathbf{x})\,\mathbb{P}(T>t|B=1,\mathbf{Z}=\mathbf{z}) \\ &= 1 - p(\mathbf{x}) + p(\mathbf{x})\,S_u(t|\mathbf{z}), \end{aligned} \tag{1}$$

where $p(\mathbf{x}) = \mathbb{P}(B=1|\mathbf{X}=\mathbf{x})$ is the conditional uncure probability (incidence) and $S_u(t|\mathbf{z})$ is the conditional survival function for the uncured (latency). Note that S_u is a proper survival function, while $\lim_{t\to\infty} S(t|\mathbf{x},\mathbf{z}) = 1 - p(\mathbf{x})$. In terms of distribution functions we have

$$F(t|\mathbf{x},\mathbf{z}) = 1 - S(t|\mathbf{x},\mathbf{z}) = p(\mathbf{x})\,F_u(t|\mathbf{z}),$$

where $F_u(t|\mathbf{z}) = 1 - S_u(t|\mathbf{z})$ denotes the conditional distribution function of the uncured. We consider a monotone single-index model for the incidence component, that is

$$p(\mathbf{x}) = \varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{x}\right), \tag{2}$$

for some unknown regression coefficient $\boldsymbol{\gamma}_0 \in \mathbb{R}^d$ and an unknown link function φ_0 belonging to the set $\mathcal{M} := \{\varphi : \mathbb{R} \to (0,1) :\ \varphi \text{ is monotone non-decreasing}\}$. For the latency we assume a Cox proportional hazard model, i.e.

$$S_u(t|\mathbf{z}) = \mathbb{P}(T > t|B=1,\mathbf{Z}=\mathbf{z}) = \exp\left[-\Lambda_0(t)\exp\left(\boldsymbol{\beta}_0^T\mathbf{z}\right)\right], \tag{3}$$

where Λ_0 denotes the baseline cumulative hazard function and $\boldsymbol{\beta}_0 \in \mathbb{R}^q$ is a vector of regression parameters. Both Λ_0 and β_0 are left unspecified.

A crucial issue for both the single-index and the mixture cure model is identifiability meaning that

$$\begin{aligned} &l(Y,\Delta,\mathbf{X},\mathbf{Z};\boldsymbol{\gamma},\boldsymbol{\beta},\Lambda,\varphi) = l(Y,\Delta,\mathbf{X},\mathbf{Z};\tilde{\boldsymbol{\gamma}},\tilde{\boldsymbol{\beta}},\tilde{\Lambda},\tilde{\varphi}) \text{ a.s.} \\ &\implies \boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}, \quad \boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}, \quad \Lambda = \tilde{\Lambda}, \quad \varphi = \tilde{\varphi}, \end{aligned}$$

where $l(Y,\Delta,\mathbf{X},\mathbf{Z};\boldsymbol{\gamma},\boldsymbol{\beta},\Lambda,\varphi)$ denotes the log-likelihood of the model given the parameters. For the single-index model, the parameters $(\varphi_0,\boldsymbol{\gamma}_0)$ are not identifiable without further restrictions since, for example, for any $a \in \mathbb{R}$ we can define $\tilde{\varphi}(t) = \varphi_0(t/a)$, $\tilde{\boldsymbol{\gamma}} = a\boldsymbol{\gamma}_0$ and have $\tilde{\varphi}(\tilde{\boldsymbol{\gamma}}^T\mathbf{x}) = \varphi_0(\boldsymbol{\gamma}_0^T\mathbf{x})$. The identifiability of single-index models has been studied in [18] and in [22] under weaker conditions. Section 2.3 of [16] provides detailed discussions of the conditions for identification of single-index models. Assumption (A1) below states the required conditions for the identification of the monotone single-index model for the incidence component. For the mixture cure model, the parameters are not identifiable if the follow-up of the study does not contain the support of the event times since it is not possible to distinguish the event of being cured from the one of being

uncured with survival time larger than the follow-up of the study. To guarantee identifiability of the model we require the following set of assumptions. Here, $\mathcal{X}$ denotes the support of the covariate $\mathbf{X}$ and $\mathcal{I}_{\boldsymbol{\gamma}} := \left\{\boldsymbol{\gamma}^T\mathbf{x} : \ \mathbf{x} \in \mathcal{X}\right\}$ denotes the support of the index $\boldsymbol{\gamma}^T\mathbf{X}$.

(A1) (i) φ_0 is differentiable and not constant on $\mathcal{I}_{\boldsymbol{\gamma}_0}$.

(ii) The parameter $\boldsymbol{\gamma}_0$ belongs to the $d-1$ dimensional unit sphere $\mathcal{S}_{d-1} := \{\boldsymbol{\gamma} \in \mathbb{R}^d \colon \|\boldsymbol{\gamma}\|_2 = 1\}$ with respect to the Euclidean norm $\|\cdot\|_2$ and the model does not have an intercept term, i.e. $\mathbf{X}$ does not contain constant component.

(iii) The covariate $\mathbf{X}$ contains at least one continuous variable and the continuous components have a joint probability density function.

(iv) $\mathcal{X}$ is not contained in a proper linear subspace of $\mathbb{R}^d$.

(v) Varying the values of the discrete components of $\mathbf{X}$ does not divide the support of $\boldsymbol{\gamma}_0^T\mathbf{X}$ into disjoint subsets.

(A2) (i) $\boldsymbol{\beta}_0 \in \mathbb{R}^q$ and $\mathbf{Z}$ does not contain constant (intercept) component.

(ii) The covariance matrix of $\mathbf{Z}$ has full rank.

(A3) (i) There exists a cure threshold $\tau_0 < \infty$ such that $T > \tau_0 \iff T = \infty$. Moreover $\mathbb{P}\left(C > \tau_0 | \mathbf{X}, \mathbf{Z}\right) > 0$ for almost all $\mathbf{X}$ and $\mathbf{Z}$.

(ii) The incidence $p\left(\mathbf{x}\right)$ in (2) satisfies $0 < p\left(\mathbf{x}\right) < 1$ for all $\mathbf{x} \in \mathcal{X}$.

Assumptions (A1)–(A3) are almost identical to the ones in [2]. Note that, in (A1)(ii), we do not need the assumption that the sign of the first entry of $\boldsymbol{\gamma}_0$ is fixed because we are fixing the direction of monotonicity for the link function. Assumption (A1)(v) is required since we allow the covariate $\mathbf{X}$ to contain both continuous and discrete variables. It essentially means that we want to avoid situations such as the following. Suppose $\mathbf{X}$ has one continuous components X_1 with support $[0, 1]$ and one Bernoulli component X_2, which are independent. Let $\gamma_2 > \gamma_1 > 0$. When $X_2 = 0$, the support of $\boldsymbol{\gamma}^T\mathbf{X}$ is $[0, \gamma_1]$, while when $X_2 = 1$ the support is $[\gamma_2, \gamma_2 + \gamma_1]$, which are disjoint. Example 2.4 of [16, p .14] illustrates the non-identifiability of $\boldsymbol{\gamma}_0$ in such scenario that does not satisfy Assumption (A1)(v). To check this assumption in practice, one can observe the histogram of $\hat{\boldsymbol{\gamma}}_n^T\mathbf{x}_i$, where $\hat{\boldsymbol{\gamma}}_n$ is an estimator of $\boldsymbol{\gamma}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the given sample. If such histogram has disconnected components, it is an indication that the assumption is not satisfied. The existence of a cure threshold as in Assumption (A3) depends on the knowledge about the problem (e.g. medical knowledge) and is usually considered valid if the Kaplan–Meier estimator of the survival function exhibits a long plateau in the right tail with many censored observations and no events. We also note that an additional assumption of φ_0 being non-periodic is not required, since φ_0 is monotonically non-decreasing. Assumption (A1) entails the identifiability of the monotone single-index model for the incidence in (2) (Theorem 2.1 in [16]). Given that the incidence in (2) is identifiable, using the same argument as in the proof of Proposition 1 of [2], it follows that the monotone-single-index/Cox mixture cure model is identifiable.

3. Estimation method

Assume that we have i.i.d. realizations $(y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \cdots, n$ of $(Y, \Delta, \mathbf{X}, \mathbf{Z})$. The observed likelihood function of the mixture cure model is given by

$$L_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda, \varphi) = \prod_{i=1}^{n} \left[\varphi\left(\boldsymbol{\gamma}^T \mathbf{x}_i\right) \lambda(y_i) e^{\boldsymbol{\beta}^T \mathbf{z}_i} e^{-\Lambda(y_i)\exp\left(\boldsymbol{\beta}^T \mathbf{z}_i\right)}\right]^{\delta_i} \times \left[1 - \varphi\left(\boldsymbol{\gamma}^T \mathbf{x}_i\right) + \varphi\left(\boldsymbol{\gamma}^T \mathbf{x}_i\right) e^{-\Lambda(y_i)\exp\left(\boldsymbol{\beta}^T \mathbf{z}_i\right)}\right]^{1-\delta_i}. \tag{4}$$

When the link function is assumed to be known, the parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda)$ are estimated via the maximum likelihood principle. Here, we treat φ as a nuisance (infinite dimensional) parameter and, for any fixed $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda)$, we construct a smooth monotone estimator of φ. Finally, we consider a new likelihood criterion with the plug-in estimator of φ and apply the maximum likelihood method. Hence, the estimation procedure consists of the following three steps:

1. For fixed $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda)$, we estimate the link by

$$\hat{\varphi}_{n,\boldsymbol{\theta}} = \underset{\varphi \in \mathcal{M}_{\epsilon'}}{\arg\max}\, L_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda, \varphi), \tag{5}$$

   where $\mathcal{M}_{\epsilon'} = \{\varphi : \mathbb{R} \to [\epsilon', 1-\epsilon'] : \varphi \text{ is monotone non-decreasing}\}$ and $\epsilon' > 0$ is a fixed small constant for a truncation on the uncured probability. See Remark 3.2 below for a discussion on this truncation and the choice of ϵ'. The estimator $\hat{\varphi}_{n,\boldsymbol{\theta}}$ is computed using the EM algorithm and techniques from isotonic estimation as explained in Subsection 3.2 below. The maximizer is not unique but it is uniquely defined at the points $\boldsymbol{\gamma}^T \mathbf{x}_i$, $i = 1, \ldots, n$. We consider $\hat{\varphi}_{n,\boldsymbol{\theta}}$ to be a left-continuous step function that extends constantly to the entire real line.
2. A kernel smoothed version of $\hat{\varphi}_{n,\boldsymbol{\theta}}$ is defined by

$$\hat{\varphi}^s_{n,\boldsymbol{\theta}}(u) = \int_{u-h}^{u+h} \frac{1}{h} k\left(\frac{u-t}{h}\right) \hat{\varphi}_{n,\boldsymbol{\theta}}(t)\, \mathrm{d}t, \tag{6}$$

   where k is a symmetric kernel with bounded support $[-1, 1]$ and h is a chosen bandwidth. By definition and the monotonicity of $\hat{\varphi}_{n,\boldsymbol{\theta}}$ it follows that $\hat{\varphi}^s_{n,\boldsymbol{\theta}}$ is a smooth non-decreasing function. We illustrate in Appendix B.4 that this smoothing step indeed improves the behavior of the estimator. For the bandwidth h we follow a common choice in the literature of smooth isotonic estimators by taking $h = rn^{-1/5}$, where r is the range of the observed index $\boldsymbol{\gamma}^T \mathbf{X}$. In Appendix B.3 we investigate the sensitivity of the estimators with respect to the choice of the bandwidth and conclude that, despite not being the optimal bandwidth, this is a satisfactory and quick solution. The fact that a combination of smoothing and shape constraints makes the estimators more robust to the choice of tuning parameters, such as kernel bandwidth or number of knots for splines, has been previously

reported in related literature, see for example [6, 28]. We do not apply any boundary correction for the kernel estimator but instead extend the isotonic estimator $\hat{\varphi}_{n,\boldsymbol{\theta}}$ to be constant outside of the range of the observed data.

3. Using the plug-in approach and the maximum likelihood principle, $\boldsymbol{\theta}_0$ is estimated by

$$\hat{\boldsymbol{\theta}}_n = \left(\hat{\boldsymbol{\gamma}}_n, \hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n\right) = \arg\max_{\boldsymbol{\theta}} L_n\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda, \hat{\varphi}^s_{n,\boldsymbol{\theta}}\right), \tag{7}$$

where the maximization is done over $\boldsymbol{\gamma} \in \mathcal{S}_{d-1}$, $\boldsymbol{\beta} \in \mathbb{R}^q$ and non-decreasing positive functions Λ. The estimator is computed iteratively using the EM algorithm as explained in Subsection 3.2 below. The whole estimation procedure is described in Algorithm 2 in Appendix B.7. As in the standard logistic/Cox mixture cure model, we impose the zero tail constraint meaning that the observations in the plateau are assumed to be cured. This corresponds to setting $S_u(t|\mathbf{z}; \Lambda, \boldsymbol{\beta}) = \exp(-\Lambda(t)\exp(\boldsymbol{\beta}^T\mathbf{z})) = 0$ for $t > y_{(r)}$ where $y_{(r)}$ denotes the largest observed event time.

3.1. Comparison with similar problems in the literature

Before explaining the computation of the proposed estimators, we comment on how our problem and method relate to the existing literature on the standard monotone single-index model [3, 5, 4, 13], which assumes that $\mathbb{E}[Y \mid \mathbf{X}] = \psi_0(\boldsymbol{\alpha}_0^T\mathbf{X})$ for some unknown $\boldsymbol{\alpha}_0$ and monotone link function ψ_0. Different methods for estimation of $(\boldsymbol{\alpha}_0, \psi_0)$ have been proposed based on the least-squares principle and adaptations of it without imposing any smoothness assumptions. The main idea is the following. For fixed $\boldsymbol{\alpha}$, one can minimize the least squares criterion $h_n(\psi, \boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^n\{Y_i - \psi(\boldsymbol{\alpha}^T\mathbf{X}_i)\}^2$ with respect to ψ on the class of monotone functions, which gives a $\boldsymbol{\alpha}$-dependent function $\psi_{n,\boldsymbol{\alpha}}$. In a second step, the function $h_n(\psi_{n,\boldsymbol{\alpha}}, \boldsymbol{\alpha})$ is then minimized over $\boldsymbol{\alpha}$. This would be the standard profile least squares estimator. Note that since this criterion function for $\boldsymbol{\alpha}$ is not smooth but piecewise constant, the estimator of $\boldsymbol{\alpha}$ is not unique. Moreover, alternative ways to estimate $\boldsymbol{\alpha}$ in the second step have been proposed by using the score approach and computing the zero-crossings of $\frac{1}{n}\sum_{i=1}^n\{Y_i - \psi_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\mathbf{X}_i)\}\mathbf{X}_i$ or minimizing its squared norm. If the criterion function was continuous in $\boldsymbol{\alpha}$, these alternative approaches would result in the same least squares estimator.

The current status linear regression problem can also be seen as a monotone single-index model where the link is actually a distribution function [13]. In that setting estimation can be performed via the maximum likelihood principle, again by first maximizing the likelihood for a fixed index on the class of distribution functions and then maximizing with respect to $\boldsymbol{\alpha}$ or solving score equations [12].

In our setting, the response variable that corresponds to the single-index model for the incidence is the latent cure status B. The fact that B is not always observed makes the use of the least-squares approach not suitable. Hence,

our method is based on the maximum likelihood principle similar to the one for the current status linear regression problem. However, in the current status model, the response (the current status Δ) is observed and the only unknown parameters are the index and the link function. In our model, the presence of additional unknown finite and infinite dimensional parameters $\boldsymbol{\beta}, \Lambda$ makes the estimation problem much more challenging. In particular, both optimization problems in (5) and (7) cannot be solved directly but only through iterative procedures such as the EM algorithm. To the best of our knowledge, this is the first case for which a maximum likelihood estimator under monotonicity constraints, as in (5), cannot be characterized explicitly. In the proof of Proposition 3.1 below we comment that, even if we try to use the standard techniques from isotonic estimation to characterize the maximizer as the left derivative of a greatest convex minorant, we would end up with an iterative procedure that is the same as the EM algorithm. In addition, we include a smoothing step which leads to a smooth monotone estimator of the link function and a continuous criterion for estimation of $\boldsymbol{\theta}$ in the next step. In Appendix B.4 we illustrate that smoothing indeed improves the performance of the estimator compared to the monotone (non-smooth) estimator. As in our case, a truncation is also needed in the current status regression problem in order to avoid the link function from being close to 0 and 1. However, since for that setting the link is a distribution function which necessarily attains the values 0 and 1, the truncation is imposed for the likelihood criterion excluding the extreme observations. For our model, given the assumption (A3)(ii), it seems easier and more reasonable to restrict to link functions that are bounded away from 0 and 1.

3.2. Computation of the estimators

Unlike the Cox proportional hazard model [7], for which the regression coefficients $\boldsymbol{\beta}$ can be estimated using a profile likelihood approach independently of the baseline cumulative hazard function Λ, the mixture cure model does not possess a likelihood function that can take advantage of such an approach due to the latent uncure status B. The maximization problems in (5) and (7) are solved via the expectation-maximization (EM) algorithm as in the standard logistic/Cox mixture cure model [34].

EM algorithm

The uncure status B_i of the i-th subject is $B_i = 1$ (uncured) if $\delta_i = 1$ and is unknown otherwise. Given the observed data $\mathbf{v}_i = (y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \cdots, n$, the complete-data likelihood function is

$$L_{nc}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda, \varphi) = \prod_{i=1}^{n} \varphi(\boldsymbol{\gamma}^T \mathbf{x}_i)^{B_i} [1 - \varphi(\boldsymbol{\gamma}^T \mathbf{x}_i)]^{1-B_i} \times \prod_{i=1}^{n} \left[\lambda(y_i)\, e^{\boldsymbol{\beta}^T \mathbf{z}_i} e^{-\Lambda(y_i) \exp\left(\boldsymbol{\beta}^T \mathbf{z}_i\right)} \right]^{\delta_i B_i} \tag{8}$$

$$\times \prod_{i=1}^{n} \left[e^{-\Lambda(y_i)\exp\left(\boldsymbol{\beta}^T \mathbf{z}_i\right)} \right]^{(1-\delta_i)B_i}.$$

In the $(k+1)$-th iteration of the EM algorithm the parameters are updated as follows. The E-step of the EM algorithm computes the conditional expectation of the complete-data log-likelihood $\log L_{nc}$ with respect to the uncured status B given the parameters of the previous iteration and the observed data. Specifically, by virtue of the partially observed nature of B and its linearity in the complete-data log-likelihood function, the E-step is equivalent to computing

$$\begin{aligned} w_i^{(k)} &= \mathbb{E}^{(k)}[B_i \mid \mathbf{v}_i] \\ &= \delta_i + (1-\delta_i)\frac{\varphi^{(k)}(\boldsymbol{\gamma}^{(k)T}\mathbf{x}_i) S_u(y_i|\mathbf{z}_i;\boldsymbol{\beta}^{(k)},\Lambda^{(k)})}{1-\varphi^{(k)}(\boldsymbol{\gamma}^{(k)T}\mathbf{x}_i)+\varphi^{(k)}(\boldsymbol{\gamma}^{(k)T}\mathbf{x}_i) S_u(y_i|\mathbf{z}_i;\boldsymbol{\beta}^{(k)},\Lambda^{(k)})}, \end{aligned} \tag{9}$$

where the expectation $\mathbb{E}^{(k)}$ is computed using the parameters $\boldsymbol{\theta}^{(k)}, \varphi^{(k)}$ and $S_u(\cdot|\cdot;\boldsymbol{\beta},\Lambda)$ is obtained according to (3). Substituting B_i with w_i in (8), we obtain the expected complete-data likelihood

$$\begin{aligned} &\prod_{i=1}^{n} \varphi(\boldsymbol{\gamma}^T\mathbf{x}_i)^{w_i}[1-\varphi(\boldsymbol{\gamma}^T\mathbf{x}_i)]^{1-w_i} \prod_{i=1}^{n} f_u(y_i|\mathbf{z}_i)^{\delta_i w_i} S_u(y_i|\mathbf{z}_i)^{(1-\delta_i)w_i} \\ &= \tilde{L}_{nc}^1(\boldsymbol{\gamma},\varphi)\,\tilde{L}_{nc}^2(\boldsymbol{\beta},\Lambda), \end{aligned} \tag{10}$$

where, to simplify the notation, we have denoted $w_i^{(k)}$ by just w_i. The M-step of the algorithm consists in maximizing the expected complete-data likelihood with respect to the parameters of interest. Specifically, in (5), we maximize with respect to φ over $\mathcal{M}_{\epsilon'}$ while keeping $(\boldsymbol{\gamma},\boldsymbol{\beta},\Lambda)$ fixed in all iterations, while in (7) we maximize over $(\boldsymbol{\gamma},\boldsymbol{\beta},\Lambda)$ for a given $\varphi = \hat{\varphi}^s_{n,\boldsymbol{\theta}}$. To simplify the maximization problem in (7), the link function could be kept fixed, equal to the one obtained in the previous iteration of the EM algorithm with $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m-1)}$. However, we observe that in some cases such procedure does not behave well computationally since the link estimator is very sensitive to the estimator of $\boldsymbol{\gamma}$. On the other hand, the link estimate seems to be stable with respect to small changes of the latency parameters $\boldsymbol{\beta}$ and Λ from one iteration to the other. Hence, we only fix $\boldsymbol{\beta}$ and Λ as the estimates of the previous iteration and allow $\varphi = \hat{\varphi}^s_{n,\boldsymbol{\theta}}$ with $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}^{(m-1)}, \Lambda^{(m-1)})$ to still depend on $\boldsymbol{\gamma}$. From (10), we can see that the expected complete-data likelihood can be factorized into two parts. The first part only consists of the parameters of the incidence part, while the second part contains the parameters of the latency part only. Therefore we can maximize the likelihood for the two parts separately.

Monotone link estimator

Algorithm 1 in Appendix B.7 describes the procedure of the EM algorithm for estimation of the monotone link function in (5). The M-step of the EM algorithm

is equivalent to the following maximization problem for a fixed $\boldsymbol{\gamma}$

$$\underset{\varphi \in \mathcal{M}_{\epsilon'}}{\text{maximize}} \sum_{i=1}^{n} \left\{ w_i \log \varphi \left(\boldsymbol{\gamma}^T \mathbf{x}_i \right) + (1 - w_i) \log \left[1 - \varphi \left(\boldsymbol{\gamma}^T \mathbf{x}_i \right) \right] \right\}, \tag{11}$$

which belongs to the class of order-restricted maximum likelihood estimation problems [33]. In addition to the order restriction, a uniform bound restriction is imposed to the maximum likelihood estimation (MLE) of φ for fixed $\boldsymbol{\gamma}$. Such class of order and uniform bound restricted problems is studied by [17]. Using results from [33] and [17] we obtain the following characterization, a proof of which can be found in Appendix A.

Proposition 3.1. *The maximizer in (11) exists, it is not unique but it is uniquely defined at the ordered points* $\boldsymbol{\gamma}^T \mathbf{x}_{(1)} < \cdots < \boldsymbol{\gamma}^T \mathbf{x}_{(n)}$ *with corresponding values* $\hat{\varphi}_i = \max(\epsilon', \min(\tilde{\varphi}_i, 1 - \epsilon'))$*, where* $(\tilde{\varphi}_1, \ldots, \tilde{\varphi}_n)$ *are the left derivatives of the greatest convex minorant of the cumulative sum diagram*

$$\left\{ (0,0), \left(i, \sum_{j=1}^{i} w_{(j)} \right), i = 1, \cdots, n \right\}.$$

Here $w_{(j)}$ *corresponds to the* j*-th order statistic of* $\boldsymbol{\gamma}^T \mathbf{x}_i$*,* $i = 1, \ldots, n$*.*

Remark 3.2. The truncation of the link function is introduced in order to avoid that it takes the extreme values 0 and 1 which would create theoretical problems with terms that explode to infinity. When ϵ' is chosen to be very small (for example of the order 10^{-6}), in practice there would be almost no difference between the truncated and not truncated version of the link function since by construction $\tilde{\varphi}_i \in [0, 1]$. Truncation is also compatible with Assumption (A3)(ii) above. When the support of $\mathbf{x}$ is assumed to be bounded, there exists ϵ' such that the true link belongs to $\mathcal{M}_{\epsilon'}$.

Incidence regression parameter estimator

For the estimation of the incidence regression parameter $\boldsymbol{\gamma}$ in (7), the M-step of the EM algorithm is equivalent to the following maximization problem

$$\underset{\boldsymbol{\gamma} \in \mathcal{S}_{d-1}}{\text{maximize}} \sum_{i=1}^{n} \left\{ w_i \log \varphi_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^T \mathbf{x}_i) + (1 - w_i) \log \left[1 - \varphi_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^T \mathbf{x}_i) \right] \right\}, \tag{12}$$

where $\varphi_{\boldsymbol{\gamma}}$ denotes the smooth monotone link estimate $\hat{\varphi}^s_{n,\boldsymbol{\theta}}$ for $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}^{(m-1)}, \Lambda^{(m-1)})$, i.e. the parameters $\boldsymbol{\beta}$, Λ are fixed to the estimates of the previous iteration. Here, we impose the identifiability constraint, $\|\boldsymbol{\gamma}\|_2 = 1$, leading to a maximization problem with a nonlinear constraint. This problem can be solved by the augmented Lagrangian method. Such method first reformulates the problem as an unconstrained problem by introducing penalty terms for the equality constraints and solving this unconstrained problem by some interior-point algorithms. The penalty terms are then updated. These two steps are repeated until convergence. See Chapter 17 of [30] for more details.

Latency estimation

For the latency component, the M-step of the EM algorithm can be performed as in [34]. Specifically, the estimator $\hat{\boldsymbol{\beta}}_n$ for $\boldsymbol{\beta}$ is computed using the profile likelihood approach as the maximizer of

$$\prod_{i=1}^{n}\left(\frac{\exp(\boldsymbol{\beta}^T\mathbf{z}_i)}{\sum_{j\in\mathcal{R}_i} w_j\exp(\boldsymbol{\beta}^T\mathbf{z}_j)}\right)^{\delta_i},$$

where $\mathcal{R}_i$ denotes the risk set just before time y_i. The nonparametric estimator of Λ is given by

$$\hat{\Lambda}_n(t)=\sum_{i:y_i\le t}\frac{d_i}{\sum_{j\in\mathcal{R}_i} w_j\exp(\hat{\boldsymbol{\beta}}_n^T\mathbf{z}_j)},$$

where d_i denotes the number of events at time y_i. As suggested in [35], the conditional survival function $\hat{S}_u(t|\mathbf{z})$ is set to zero when $t > y_{(r)}$.

4. Asymptotic properties

We start by providing some technical intuition on the approach and illustrate the additional challenges that we face compared to the existing literature on the monotone single-index model. We will derive consistency of our estimators using results from the theory of semiparametric M-estimation. Indeed, our estimator $\hat{\boldsymbol{\theta}}_n$ corresponds to the maximizer of an empirical criterion function that depends on an infinite dimensional nuisance parameter φ. The unknown φ is replaced by a nonparametric estimator $\hat{\varphi}^s_{n,\boldsymbol{\theta}}$ depending on $\boldsymbol{\theta}=(\boldsymbol{\gamma},\boldsymbol{\beta},\Lambda)$. Denote the log-likelihood for a single observation by

$$l(y,\delta,\mathbf{x},\mathbf{z};\boldsymbol{\theta},\varphi)=\delta\log f_u(y|\mathbf{z})+\delta\log\varphi(\boldsymbol{\gamma}^T\mathbf{x})+(1-\delta)\log[1-\varphi(\boldsymbol{\gamma}^T\mathbf{x})F_u(y|\mathbf{z})].$$

Then we can write

$$\hat{\boldsymbol{\theta}}_n=\arg\max_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta},\hat{\varphi}^s_{n,\boldsymbol{\theta}}),\qquad M_n(\boldsymbol{\theta},\hat{\varphi}^s_{n,\boldsymbol{\theta}})=\frac{1}{n}\sum_{i=1}^{n} l\left(y_i,\delta_i,\mathbf{x}_i,\mathbf{z}_i;\boldsymbol{\theta},\hat{\varphi}^s_{n,\boldsymbol{\theta}}\right).$$

We will show in Proposition 4.3 below that such maximizer exists and is finite. The asymptotic version of M_n is given by $M(\boldsymbol{\theta},\varphi_{\boldsymbol{\theta}})=\mathbb{E}\left[l\left(Y,\Delta,\mathbf{X},\mathbf{Z};\boldsymbol{\theta},\varphi_{\boldsymbol{\theta}}\right)\right]$, where $\varphi_{\boldsymbol{\theta}}$ denotes any infinite dimensional nuisance parameter φ that might depend on $\boldsymbol{\theta}$. On the other hand, for fixed $\boldsymbol{\theta}=(\boldsymbol{\gamma},\boldsymbol{\beta},\Lambda)$, we define $l_{\boldsymbol{\theta}}(\varphi):=\mathbb{E}\left[l\left(Y,\Delta,\mathbf{X},\mathbf{Z};\boldsymbol{\theta},\varphi\right)\right]$ and

$$\varphi_{0,\boldsymbol{\theta}}:=\arg\max_{\varphi\in\mathcal{M}_{\epsilon'}} l_{\boldsymbol{\theta}}(\varphi),\tag{13}$$

where $\mathcal{M}_{\epsilon'}:=\{\varphi:\mathbb{R}\to[\epsilon',1-\epsilon']:\ \varphi \text{ is monotone non-decreasing}\}$. Note that $M(\boldsymbol{\theta},\varphi)=l_{\boldsymbol{\theta}}(\varphi)$, but $l_{\boldsymbol{\theta}}$ is viewed as a function of φ for a fixed $\boldsymbol{\theta}$, and $\varphi_{0,\boldsymbol{\theta}}$ corresponds to the asymptotic version of the estimator $\hat{\varphi}_{n,\boldsymbol{\theta}}$ in (5). Proposition 4.1

below guarantees that the maximizer $\varphi_{0,\boldsymbol{\theta}}$ exists and is unique. Note that from Assumptions (A1)(i) and (A4) below, it follows that there exists $\epsilon > 0$ such that $\epsilon \leq \varphi_0(\boldsymbol{\gamma}_0^T \mathbf{x}) \leq 1 - \epsilon$ for all $\mathbf{x} \in \mathcal{X}$. We assume that ϵ' in the definition of $\hat{\varphi}_{n,\boldsymbol{\theta}}$ is chosen such that $\epsilon' < \epsilon$. In this way, $\varphi_0 \in \mathcal{M}_{\epsilon'}$ and in particular we have that, if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then $\varphi_{0,\boldsymbol{\theta}} = \varphi_0$ (Proposition 4.2). We will show that $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} M(\boldsymbol{\theta}, \varphi_{0,\boldsymbol{\theta}})$, which is the foundation behind the estimation strategy. Then, to obtain consistency of of the estimator for $\boldsymbol{\theta}_0$ we will check the conditions of Theorem 1 in [8]. Specifically we need that $\hat{\varphi}^s_{n,\boldsymbol{\theta}}$ is a consistent estimator for $\varphi_{0,\boldsymbol{\theta}}$ uniformly over $\boldsymbol{\theta}$, the empirical criterion function $M_n(\boldsymbol{\theta}, \varphi)$ is a good approximation of the asymptotic criterion $M(\boldsymbol{\theta}, \varphi)$ uniformly over $\boldsymbol{\theta}$ and φ, the function $M(\boldsymbol{\theta}, \varphi)$ is continuous with respect to φ at $\varphi_{0,\boldsymbol{\theta}}$ uniformly over $\boldsymbol{\theta}$. Such conditions will be proved in Theorem 4.7.

The function $\varphi_{0,\boldsymbol{\theta}}$ defined in (13) plays a fundamental role in the theoretical analysis of our estimators and the main challenges we face arise from the fact that we do not have an explicit characterization of this function. The counterpart of this function in the monotone single-index model is $\psi_{\boldsymbol{\alpha}}(u) = \mathbb{E}[\psi_0(\boldsymbol{\alpha}_0^T \mathbf{X}) \mid \boldsymbol{\alpha}^T \mathbf{X} = u]$ (see for example equation (5) in [4]), while in the current status model is $F_{\boldsymbol{\beta}}(u) = \mathbb{E}[F_0(T - \boldsymbol{\beta}_0^T \mathbf{X}) \mid T - \boldsymbol{\beta}^T \mathbf{X} = u]$ (see equation 3.2 in [12]). In both cases, this function can be seen as the expected value of the true single-index model when we fix the index to a given value. Having this explicit characterization makes it easier to deal with this function and in particular, properties such as continuity, differentiability of $\psi_{\boldsymbol{\alpha}}$ or $F_{\boldsymbol{\beta}}$ can be derived from assumptions on the true link function. In our case, we can characterize $\varphi_{0,\boldsymbol{\theta}}$ as

$$\varphi_{0,\boldsymbol{\theta}}(u) = \mathbb{E}\left[\varphi_0(\boldsymbol{\gamma}_0^T \mathbf{X}) \mid \boldsymbol{\gamma}^T \mathbf{X} = u\right] \tag{14}$$

only if the covariates $\mathbf{X}$ and $\mathbf{Z}$ are independent, censoring is independent of all the other variables and only for $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)$ (see details in Appendix A). However, such assumptions are too strong for practical purposes since it is quite common in particular to have $\mathbf{X} = \mathbf{Z}$. Hence, we prefer not to restrict ourselves to such scenario. Without such characterization, even just arguing continuity of $\varphi_{0,\boldsymbol{\theta}}$ is quite challenging and technical (see Proposition 4.4) and requires assumptions that are more difficult to interpret such as assumption (A8) below. For a fixed $\boldsymbol{\theta}$, the smooth kernel estimator $\hat{\varphi}^s_{n,\boldsymbol{\theta}}$ is an estimator of $\varphi_{0,\boldsymbol{\theta}}$, hence in order to obtain the rate of convergence of the estimators, one would need also (twice) differentiability of $\varphi_{0,\boldsymbol{\theta}}$. Given the technicalities of proving by contradiction that $\varphi_{0,\boldsymbol{\theta}}$ is continuous, we do not explore this direction further.

The second challenge arises from the fact that our parameter $\boldsymbol{\theta}$ contains not only the index $\boldsymbol{\gamma}$ but also the latency parameters $\boldsymbol{\beta}$, Λ. In particular, $\boldsymbol{\theta}$ is not finite dimensional. The theory of semiparametric M-estimators in [8] allows $\boldsymbol{\theta}$ to be infinite dimensional only for the consistency part but obtaining the asymptotic normality requires an Euclidean parameter $\boldsymbol{\theta}$. Hence, one would first need to extend the standard results of semiparametric M-estimation to this scenario. Note that obtaining the limit distribution of the latency parameters $\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n$ is equally challenging since they are dependent on $\hat{\varphi}^s_{n,\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}_n$ (see Definition (7)) and cannot be dealt with separately. Because of these two main

issues, within this paper, we focus on the consistency property of the estimators and leave the rate of convergence and asymptotic normality to be subject of future research.

We expect that the rate of convergence of the finite dimensional parameters is square-root n, as for the smooth single-index model [15] and the standard logistic-Cox mixture cure model [26]. In the setting of the monotone (non smooth) single-index model, the rate of convergence of the least-squares index estimator is only known to be faster than $n^{1/3}$ and slower than $n^{1/2}$, mainly as a result of a non-smooth criterion function when using a piecewise constant link estimate [3]. This problem is avoided by means of a score estimator which achieves the parametric convergence rate $n^{1/2}$ [5]. However, such issues are not encountered in our setting since we are using a smooth estimate of the link function, resulting in a smooth criterion with respect to the index parameter. We also expect the link estimator to converge at rate $(nh)^{1/2}$, i.e. $n^{2/5}$ for a bandwidth of order $n^{-1/5}$, since it has been observed in the literature that asymptotically the kernel smoothed monotone estimator behaves the same as the kernel (unconstrained) estimator [37, 25]. A possible approach to obtain the rate of convergence would be to use Theorem 2 in [8], which requires checking conditions B1–B4 of that paper. Condition B1 regarding the rate of convergence of the smooth monotone estimator is expected to hold with $v_n = n^{2/5}$ if $\varphi_{n,\boldsymbol{\theta}}$ is twice continuously differentiable. Condition B4 is automatically fulfilled by definition of the estimator. For condition B2, one could use the strategy described in Remark 2(ii) of [8], requiring stronger results on entropy numbers of certain classes of functions similar to those in our Lemmas A.1–A.3. It is however not clear how to verify condition B3, which relies on some type of Taylor expansion for the criterion function, since in our setting the parameter $\boldsymbol{\theta}$ is not finite dimensional. In addition, our criterion function M depends not only on $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda)$ but also on λ defined as the derivative of Λ or the jump size at the points of jump of Λ. We investigated the rate of convergence of the estimators via a small simulation study, the results of which are described in the next section.

The results that we previously described intuitively are formulated rigorously below. First we list the required assumptions. The identifiability assumptions (A1)–(A3) are assumed throughout this section. In what follows $B_{\mathbf{c}}(r)$ will denote a closed ball of center $\mathbf{c}$ and radius r in a given metric space, which is for some integer d with the Euclidean norm if not specified otherwise.

(A4) The covariates $\mathbf{X}$ and $\mathbf{Z}$ have bounded supports $\mathcal{X}$ and $\mathcal{Z}$ respectively. That is, $\mathcal{X} \subset B_{\mathbf{0}}(r_1)$ for some $r_1 > 0$ and $\mathcal{Z} \subset B_{\mathbf{0}}(r_2)$ for some $r_2 > 0$.

(A5) There exists $\delta_0 > 0$ such that, for all $\boldsymbol{\gamma} \in B_{\boldsymbol{\gamma}_0}(\delta_0)$, $\boldsymbol{\gamma}^T\mathbf{X}$ has a density with respect to the Lebesgue measure. We denote such density by $g_{\boldsymbol{\gamma}^T\mathbf{X}}(\cdot)$.

(A6) $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_0$ lie in the interior of compact sets $\Gamma = \mathcal{S}_{d-1} \cap B_{\boldsymbol{\gamma}_0}(\delta_0)$ and $\mathcal{B}$ respectively.

(A7) The function $\Lambda_0(t)$ defined on $[0, \tau_0)$ is continuous, non-decreasing and $\Lambda_0(\tau_0-) < \infty$. We define $\Lambda_0(\tau_0) = \Lambda_0(\tau_0-)$, i.e. extend Λ at τ_0 by left-continuity. In addition, its derivative λ_0, which is defined a.e. on $(0, \tau_0)$

since Λ_0 is monotone, has bounded variation.

Consider $\boldsymbol{\theta} \in \Theta = \Gamma \times \mathcal{B} \times \mathcal{D}$, where $\mathcal{D}$ is the space of nondecreasing functions Λ on $[0, \tau_0]$ such that $\Lambda(0) = 0$ and $\Lambda(\tau_0) < \infty$.

(A8) For any $\boldsymbol{\theta} \in \Theta$, the function

$$u \longmapsto \mathbb{E}\left[\frac{1 - \varphi_0\left(\boldsymbol{\gamma}_0^T \mathbf{X}\right) F_u\left(Y|\mathbf{Z}; \boldsymbol{\beta}_0, \Lambda_0\right)}{1 - \varphi_{0,\boldsymbol{\theta}}\left(\tilde{u}\right) F_u\left(Y|\mathbf{Z}; \boldsymbol{\beta}, \Lambda\right)} \middle| \boldsymbol{\gamma}^T \mathbf{X} = u \right]$$

is continuous for all $\tilde{u} \in \mathcal{I}_{\boldsymbol{\gamma}}$, where $F_u(Y|\mathbf{Z}; \boldsymbol{\beta}, \Lambda) = 1 - S_u(Y|\mathbf{Z}; \boldsymbol{\beta}, \Lambda)$ is defined as in (3) but with general parameters $\boldsymbol{\beta}$ and Λ.

(A9) The family of density functions $\{g_{\boldsymbol{\gamma}^T \mathbf{X}} : \boldsymbol{\gamma} \in \Gamma\}$ is uniformly equicontinuous.

(A10) The density functions $g_{\boldsymbol{\gamma}^T \mathbf{X}}$ for $\boldsymbol{\gamma} \in \Gamma$ and $g_{\boldsymbol{\beta}^T \mathbf{Z}}$ for $\boldsymbol{\beta} \in \mathcal{B}$ are uniformly bounded from above by some positive constants $\bar{q}_1$, $\bar{q}_2$ respectively.

Assumptions (A4)–(A6) are standard assumptions made also in the standard single index model and the current status model. Assumption (A7) is a standard assumption of the mixture cure model, see [26, 29]. In general λ_0 is required to be continuous. Here, since in the proof we use an argument based on entropy numbers, continuity of λ_0 is not sufficient. Bounded variation can be implied for example by Lipschitz continuity. Together with assumption (A3), it essentially means that the distribution of the survival times for the uncured subjects has a jump at τ_0, i.e. there is positive probability for the event to happen at τ_0. This is mainly a technical condition for consistency of the estimators (see discussion in [29]) and the probability mass at τ_0 can be arbitrarily small, hence reasonable in practice. Assumption (A8) is needed to guarantee the continuity of $\varphi_{0,\boldsymbol{\theta}}$ since we do not have an explicit expression for such function. If there was no latency component F_u then this assumption reduces to continuity of $\mathbb{E}[\varphi_0\left(\boldsymbol{\gamma}_0^T \mathbf{X}\right) \mid \boldsymbol{\gamma}^T \mathbf{X} = u]$, which is standard in the single-index model. Assumption (A8') below is a sufficient condition for (A8). This can be shown by first writing the expectation in (A8) as an integral with respect to the conditional distribution of the covariates $\mathbf{X}$ and $\mathbf{Z}$ given the index $\boldsymbol{\gamma}^T \mathbf{X} = u$. Then we can conclude the sufficiency by using the generalized dominated convergence theorem together with the continuity and boundedness of φ_0, the continuity of $F_u(Y|\mathbf{Z}; \boldsymbol{\beta}, \Lambda)$ with respect to $\mathbf{Z}$, and the boundedness of $\varphi_{0,\boldsymbol{\theta}}$.

(A8') The joint probability density function of the continuous components of $\mathbf{X}$ is continuous with respect to at least one of the components and the corresponding components of $\boldsymbol{\gamma}$ is non-zero.

Assumptions (A9) and (A10) are required in order to get uniform consistency of the link estimate. (A9) is for example satisfied if the density functions $g_{\boldsymbol{\gamma}^T \mathbf{X}}(\cdot)$ are continuously differentiable on their support $\mathcal{I}_{\boldsymbol{\gamma}}$ for all $\boldsymbol{\gamma} \in \Gamma$ with uniformly bounded derivative.

Proposition 4.1. *Suppose that Assumptions (A4) and (A5) hold. Then, for any $\boldsymbol{\theta} \in \Theta$, the maximizer $\varphi_{0,\boldsymbol{\theta}}$ in (13) exists and is unique.*

Proposition 4.2. *For $\boldsymbol{\theta} = (\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0, \Lambda_0)$, we have $\varphi_{0,\boldsymbol{\theta}} = \varphi_0$.*

Proposition 4.3. *Suppose that Assumptions (A4) and (A6) hold. Then the maximum likelihood estimator $(\hat{\boldsymbol{\gamma}}_n, \hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n)$ defined as in (7) exists and is finite.*

Proposition 4.4. *Suppose that Assumptions (A4), (A5) and (A8) hold. Then, for any $\boldsymbol{\theta} \in \Theta$, the function $u \longmapsto \varphi_{0,\boldsymbol{\theta}}(u)$ is continuous.*

As in Lemma 1 of [26], using Assumption (A4), it can be shown that there exists $M > 0$ such that $\sup_n \hat{\Lambda}_n(\tau_0) \leq M$ a.s.. Hence, we can restrict to $\boldsymbol{\theta} \in \tilde{\Theta} = \Gamma \times \mathcal{B} \times \tilde{\mathcal{D}}$, where $\tilde{\mathcal{D}}$ is the subset of $\mathcal{D}$ consisting of functions Λ that are uniformly bounded by M. In addition, we can also assume that functions $\Lambda \in \tilde{\mathcal{D}}$ have λ of bounded variation. Here we will use the notation: $\lambda(t) = \Lambda'(t)$ at the points where Λ is differentiable and $\lambda(t)$ is equal to the jump size at the points of jump of Λ. Both Λ_0 and $\hat{\Lambda}_n$ belong to this class, so the maximization can be restricted to such functions.

Proposition 4.5. *Suppose that Assumptions (A4)–(A7) and (A10) hold, then*

$$\mathbb{P}\left[\lim_{n\to\infty} \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\mathcal{X}} \left\{\hat{\varphi}_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) = 0\right] = 1,$$

where $Q_{\mathbf{X}}(\cdot)$ denotes the distribution function of $\mathbf{X}$.

Let k be a symmetric kernel density function with support $[-1, 1]$ that satisfies $k(x) \leq K < \infty$ for all $x \in [-1, 1]$. To simplify the notation, we define $k_h(u) = h^{-1}k(u/h)$. Here $h = h_n > 0$ is a bandwidth that depends on the sample size n and satisfies $h_n \to 0$ as $n \to \infty$. Since within this paper we do not investigate further the rate of convergence of the estimators, no additional restrictions are imposed on the bandwidth. However, it is known that the optimal order bandwidth for estimation of a twice differentiable function is $n^{-1/5}$ and, if the bandwidth is of order $n^{-1/3}$ or smaller, the smooth estimator is not much different from the non-smooth one. Hence, in practice we take $h = rn^{-1/5}$, where r is the range of the index $\boldsymbol{\gamma}^T\mathbf{X}$. This is a common choice in the literature of smooth isotonic estimators that is simple and behaves well. For a more detailed investigation of the role of the bandwidth see Appendix B.3.

Proposition 4.6. *Suppose that Assumptions (A4)–(A10) hold, then*

$$\mathbb{P}\left[\lim_{n\to\infty} \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\mathcal{X}} \left\{\hat{\varphi}^s_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) = 0\right] = 1.$$

Next we show that the estimator $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\gamma}}_n, \hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n)$ of $\boldsymbol{\theta}_0$ is weakly consistent.

Theorem 4.7. *Suppose that Assumptions (A4)–(A10) hold, then*

$$\|\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\|_2, \qquad \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \qquad \text{and} \qquad \sup_{t\in[0,\tau_0]} |\hat{\Lambda}_n(t) - \Lambda_0(t)|$$

converge to zero in probability as $n \to \infty$.

The following corollary shows the consistency of the estimated cure probabilities and the estimated survival function for the uncured.

Corollary 4.8. *Suppose that Assumptions (A4)–(A10) hold. We suppose further that φ_0 has bounded derivative on $\mathcal{I}_0$. Then, for any $\mathbf{z} \in \mathcal{Z}$,*

$$\int_{\mathcal{X}} \left\{ \hat{\varphi}^s_{n,\hat{\boldsymbol{\theta}}_n}(\hat{\boldsymbol{\gamma}}_n^T \mathbf{x}) - \varphi_0\left(\boldsymbol{\gamma}_0^T \mathbf{x}\right) \right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) \qquad \text{and} \qquad \sup_{t \in [0,\tau_0]} |\hat{S}_u(t|\mathbf{z}) - S_u(t|\mathbf{z})|$$

converge to zero in probability as $n \to \infty$, where $S_u(t|\mathbf{z}) = S_u(t|\mathbf{z}; \boldsymbol{\beta}_0, \Lambda_0)$ and $\hat{S}_u(t|\mathbf{z}) = S_u(t|\mathbf{z}; \hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n)$ are defined as in (3) using $\boldsymbol{\beta}_0$, Λ_0 and the estimated parameters $\hat{\boldsymbol{\beta}}_n$, $\hat{\Lambda}_n$ respectively.

5. Simulation study

In the simulation study, we consider different settings to evaluate the finite sample behavior of the estimator proposed in Section 3 and compare it with the SIC method proposed in [2]. We simulate 500 datasets from the mixture cure model introduced in Section 2, where $\mathbf{X} = (X_1, \cdots, X_4)^T$ contains four independent covariates: $X_1 \sim U[0,1]$, $X_2 \sim N(0,1)$, X_3 and X_4 are Bernoulli variables with parameters 0.3 and 0.6 respectively, and $\mathbf{Z} = (Z_1, Z_2)^T$ with $Z_1 = X_1$ and $Z_2 = X_4$. We consider a Weibull model with parameters $\lambda = 1.5$ and $k = 2.2$ for the baseline distribution of the uncured subjects. The random right censoring time C follows the exponential distribution with rate λ_C. We consider three experiments A, B and C, with three different non-decreasing link functions.

$$\varphi_A(u) = \frac{\exp(c+u)}{1+\exp(c+u)}, \; \varphi_B(u) = \frac{\exp[\psi(c,u)]}{1+\exp[\psi(c,u)]}, \; \varphi_C(u) = \frac{1+\tanh(c+u^3)}{2},$$

where c is an intercept term, $\psi(c,u) = 0.75\Phi\{(c+u)+0.5\}+0.25\Phi\{0.5\,(c+u)^3\}$ and Φ is the cdf of the standard normal distribution. The first two links correspond to Scenarios 1 and 2 of the simulation study investigated by [2], while φ_C is a scaled hyperbolic tangent function. These three link functions are considered for exploring the influence of the shape and the steepness of the true link function to the model estimation performance. Figure 1 shows the plots of the three link functions over $[-4,4]$ when the intercept term $c = 0$. Among these three link functions, φ_C is the steepest and φ_B is the flattest.

The choices for the parameters $\boldsymbol{\gamma}_0$, $\boldsymbol{\beta}_0$, and λ_C are given in Table 1 as well as the averages, over the 500 simulated datasets for each setting, of the cure proportion, the censoring rate and the proportion of observations in the plateau. $\boldsymbol{\gamma}_0$ is chosen such that it has unit Euclidean norm to ensure model identifiability and it leads to different cure proportions. For each experiment we also consider two different censoring scenarios while maintaining a reasonable percentage of observations in the plateau.

The simulation experiments are carried out with sample size n of 250 and 500. The SIC method proposed by [2] and our estimation method (mSIC) are applied

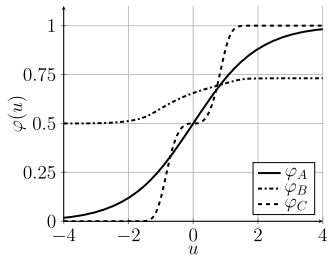


FIG 1. *Link functions when the intercept term* $c = 0$.

TABLE 1
Simulation settings

Expt.	c	γ_{01}	γ_{02}	γ_{03}	γ_{04}	β_{01}	β_{02}	λ_C	Cure prop.	Cens. rate	Plateau
A	1.2	-0.2383	0.7423	0.3156	0.5409	-0.8	0.5	0.1	0.2090	0.2674	0.1675
								0.3		0.3691	0.1108
B	0.5	-0.7826	0.4368	-0.2599	0.3594	-0.6	0.8	0.1	0.3352	0.3792	0.2734
								0.4		0.4902	0.1554
C	0.2	0.1057	0.7899	-0.4883	0.3556	0.6	0.4	0.15	0.3912	0.4415	0.3092
								0.5		0.5390	0.1867

to each dataset. For comparison we compute the mean squared error (MSE) of the estimate of the cure probability, bias and variance of the estimates of both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. The MSE for the cure probability is defined as

$$\text{MSE}(\hat{p}, p_0) = \frac{1}{K}\sum_{k=1}^{K}\{\hat{\varphi}(\hat{\boldsymbol{\gamma}}^T\mathbf{x}_k) - \varphi_0(\boldsymbol{\gamma}_0^T\mathbf{x}_k)\}^2, \tag{15}$$

where $\hat{\varphi}$ is the smooth monotone link estimate $\hat{\varphi}^s_{\hat{\boldsymbol{\theta}}}$ for mSIC (smooth non-monotone link for SIC) and the summation is over a grid of points. For x_{k1} and x_{k2} we take a grid of size 0.01 on $[0, 1]$ and $[-3, 3]$ respectively, while $x_{k3}, x_{k4} \in \{0, 1\}$. The bias of the coefficient estimates is the mean of the Euclidean norms of the differences between the coefficient estimate and the true parameter over 500 replications. The variance of the coefficient estimates is the sample variance of the Euclidean norms of the coefficient estimates.

For both methods, we initialize the algorithms as follows. The initial link estimate is the logistic function, $\hat{\boldsymbol{\gamma}}_{(0)}$ is the estimate from fitting a logistic regression model to the censoring indicator against the covariates $\mathbf{X}$, initial estimates for the latency are obtained from fitting the standard Cox model to the uncensored observations. Both algorithms terminate when the difference of the estimators from one iteration to the other, in Eucledian norm, is smaller then 10^{-5}. Other configurations for the SIC method are set as stated in Section 3 of [2]. Since the SIC method sets $\hat{\gamma}_1 = \pm 1$ for the model identifiability condition, at the end we normalize the obtained estimate. For our method we used the triweight kernel function for smoothing the monotone link estimate and set the bandwidth pa-

rameter at the k-th iteration to be $h_k = r_k n^{-1/5}$, where r_k is the range of the index $\boldsymbol{\gamma}^T\mathbf{X}$ computed at the k-th iteration. This is a common choice for smooth isotonic estimators and behaves well in practice. We note that the bandwidth choice is less problematic in the case of smooth monotone estimators than for the standard kernel estimator. A more detailed investigation of the role of the bandwidth is provided in Appendix B.3. For the SIC method, cross-validation is used to select the bandwidth but in our experience that is not stable in practice and the search interval needs to be chosen depending on the range of the index (see discussion in Appendix B.6). Regarding the truncation parameter, we set $\epsilon' = 10^{-6}$ for both experiments A and C. For experiment B, given the small range of values for the incidence, the estimation is very challenging. Since the isotonic estimator is known to have problems at the boundary, in these type of situations it is better to determine the upper and lower truncation in a data-driven way instead of taking a very small ϵ'. We use the method proposed by [27] for the range-regularized isotonic regression problem. Details are given in Appendix B.1.

TABLE 2
Simulation results

Expt.	Size	λ_C	Method	MSE $(\hat{p}, p_0)$		$\hat{\boldsymbol{\gamma}}$		$\hat{\boldsymbol{\beta}}$	
				Mean	Variance	Bias	Variance	Bias	Variance
A	250	0.1	mSIC	0.00939	4.39E-05	0.57368	0.05038	0.28080	0.02709
			SIC	0.01688	1.41E-04	0.71096	0.05077	0.28151	0.02708
		0.3	mSIC	0.01141	5.92E-05	0.61280	0.05145	0.30371	0.03408
			SIC	0.01940	1.55E-04	0.76052	0.06128	0.30605	0.03442
	500	0.1	mSIC	0.00562	1.50E-05	0.43937	0.03810	0.19793	0.01321
			SIC	0.00967	4.57E-05	0.58563	0.04825	0.19799	0.01331
		0.3	mSIC	0.00668	2.01E-05	0.48043	0.03989	0.22269	0.01667
			SIC	0.01183	7.44E-05	0.61136	0.04907	0.22371	0.01716
B	250	0.1	mSIC	0.00697	6.06E-05	0.99875	0.20072	0.30717	0.03557
			SIC	0.00831	9.70E-05	1.02873	0.26451	0.30798	0.03525
		0.4	mSIC	0.00912	9.79E-05	0.94890	0.18028	0.35858	0.04353
			SIC	0.01041	1.34E-04	1.04764	0.25388	0.35971	0.04412
	500	0.1	mSIC	0.00433	2.25E-05	0.88094	0.17064	0.21738	0.01595
			SIC	0.00504	4.17E-05	0.91734	0.26178	0.21715	0.01596
		0.4	mSIC	0.00591	3.50E-05	0.86886	0.17217	0.24466	0.02036
			SIC	0.00578	4.44E-05	0.97191	0.25303	0.24623	0.02122
C	250	0.15	mSIC	0.00583	3.49E-06	0.29442	0.01686	0.32893	0.03600
			SIC	0.01474	1.79E-04	0.40554	0.01987	0.32966	0.03599
		0.5	mSIC	0.00666	6.10E-06	0.33582	0.02433	0.37579	0.05071
			SIC	0.01847	2.82E-04	0.45743	0.02754	0.38216	0.05140
	500	0.15	mSIC	0.00446	4.52E-05	0.20682	0.01075	0.22675	0.01747
			SIC	0.00811	5.61E-05	0.32336	0.01965	0.22658	0.01761
		0.5	mSIC	0.00459	1.41E-06	0.23880	0.01169	0.26455	0.02287
			SIC	0.01038	8.51E-05	0.35384	0.02201	0.26582	0.02345

Table 2 summarizes the simulation results, including the MSE of the link estimates, bias and variance of the coefficient estimates, for both SIC and mSIC

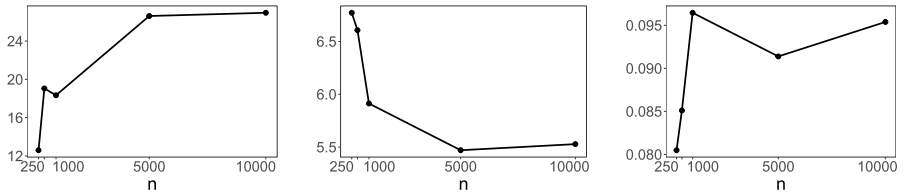


FIG 2. *Empirical variances of* $\sqrt{n}\,\|\hat{\gamma}_n - \gamma_0\|_2$ *(left),* $\sqrt{n}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2$ *(center) and* $n^{2/5}\sqrt{MSE(\hat{p}_n, p_0)}$ *(right) for different sample sizes in Experiment A with* $\lambda_C = 0.1$.

methods. In terms of the MSE for the cure probability, the mSIC method has lower mean and variance compared to the SIC method among all simulation settings, except for Experiment B with sample size of 500 and $\lambda_C = 0.4$. This indicates that mSIC performs better in inferring the incidence and gives less dispersed estimates. The mSIC method behaves better in estimating $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, in terms of bias and variance among all simulation configurations. As expected, as the sample size increases the performance of both methods improves, while it deteriorates when the censoring rate increases.

For experiment B, we find that range-regularization method shrinks the range of the upper and lower truncation to zero for about one-fifth of the replications for each simulation settings. This leads to a constant estimate of the link function which poses a practical identification issue for $\boldsymbol{\gamma}$ and causes the high variability of $\hat{\boldsymbol{\gamma}}$. We look separately at the cases with a constant (non-constant) estimate for the link function and report the correspondent MSE, bias and variance in Tables 4–5 in Appendix B.1. Results indicate that, when our method estimates a constant link function, the SIC method behaves no better than a constant link estimate. Table 5 also suggests mSIC performs better in estimating the incidence for almost all cases with a non-constant link estimate. As expected, the variance of $\hat{\boldsymbol{\gamma}}$ for the cases with a constant estimated link function is higher than that for a non-constant link. This practical identifiability issue may be a result of the flatness of the true link function φ_B, although it is not a constant function. Note that the true incidence probability ranges from around 0.52 to around 0.73. We recommend that the range-regularized method should be used in practice to determine the upper and lower truncation in order to identify situations for which there is risk of practical identifiability as in experiment B. In such cases, one should be careful in interpreting the results.

In Section 4, we discussed that the expected rates of convergence for $\hat{\boldsymbol{\gamma}}_n$ and $\hat{\boldsymbol{\beta}}_n$ are $\sqrt{n}$, while for $\hat{\varphi}^s_{n,\hat{\boldsymbol{\theta}}_n}$ it is $n^{2/5}$. To investigate this, we conducted further simulations for experiment A with $\lambda_C = 0.1$ and $n \in \{250, 500, 1000, 5000, 10000\}$. For each sample size, we calculated the variances of $\sqrt{n}\,\|\hat{\gamma}_n - \gamma_0\|_2$, $\sqrt{n}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2$, and of $n^{2/5}\sqrt{\mathrm{MSE}(\hat{p}_n, p_0)}$ (see (15)) over 500 replications. In theory, if the rate of convergence is correctly specified, we would expect these variances to stabilize and reach a constant value as n increases. Figure 2 illustrates the variances of $\sqrt{n}\,\|\hat{\gamma}_n - \gamma_0\|_2$ and $\sqrt{n}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2$ plotted against n. We observed

that the values become stable for large n (specifically, $n \geq 5000$), suggesting that $\hat{\boldsymbol{\gamma}}_n$ and $\hat{\boldsymbol{\beta}}_n$ likely exhibit $\sqrt{n}$ rate of convergence. To confirm this, larger sample sizes need to be considered but the simulation become too time consuming. We also notice that the variance of $n^{2/5}\sqrt{\text{MSE}(\hat{p}_n, p_0)}$ in Figure 2 stabilizes as n becomes large ($n \geq 5000$) indicating that the rate of convergence for the smooth monotone link estimator $\hat{\varphi}^s_{n,\hat{\boldsymbol{\theta}}_n}$ is likely $n^{2/5}$.

We note that if the true link function is not monotone, using mSIC would lead to model mispecification and SIC would be preferred over mSIC. We illustrate this through an additional simulation setting in Appendix B.2. However, we expect that for small deviations from monotonicity, mSIC would still provide reasonable estimates. Finally, we also investigated the role of the additional smoothing step 2 in the estimation procedure described in Section 3. Results in Appendix B.4 illustrate that the smooth monotone estimator behaves better than the piecewise constant isotonic estimator.

For the computational aspect, the proposed method consumes slightly more time to complete the model estimation comparing to the SIC method. For instance, the average elapsed time, over the 500 replications, on the model estimation for experiment C with $\lambda_C = 0.15$ and a sample size of 250 (500) are 12.28 (14.20) seconds for mSIC and 11.64 (13.67) seconds for SIC, with a Core i7-1165G7 CPU laptop.

6. Real data application

We apply the proposed method to a dataset of melanoma patients extracted from the Surveillance, Epidemiology and End Results (SEER) database. Such dataset has also been studied by [29] and it consists of 1445 melanoma patients diagnosed between 2004 – 2015. The event time of interest is the time to death because of melanoma and the follow-up time ranges from 1 to 155 months. The age of the patients varies between 11 to 104 years old. Among the 1445 patients, 596 are females and 849 are males. The cancer stage at diagnosis (localized: 1302 cases, regional: 101 cases, distant: 42 cases) is also recorded. The Kaplan–Meier estimate of the survival function in Figure 3 has a long plateau, which contains around 20% of the observations. Combined with medical evidence of possibility of cure for melanoma, this suggests that the cure model is appropriate for this dataset.

To compare the estimates of the logistic/Cox (LC), the single-index/Cox (SIC) and the monotone single-index/Cox (mSIC) models, we split the data into a training set (with size of 964) and a testing set (with size of 481) and estimate the models using the training set. The standardized age is considered in the incidence in order to select a reasonable bandwidth for the SIC model. We note that the SIC method selected an optimal bandwidth of 1, which is the upper bound of the search interval, $[0.4, 1]$, considered in the SIC algorithm. However, extending further the search interval for the bandwidth gave similar results. Table 3 shows the parameter estimates, and the percentile-based bootstrap confidence intervals with approximate 95% confidence level. The con-

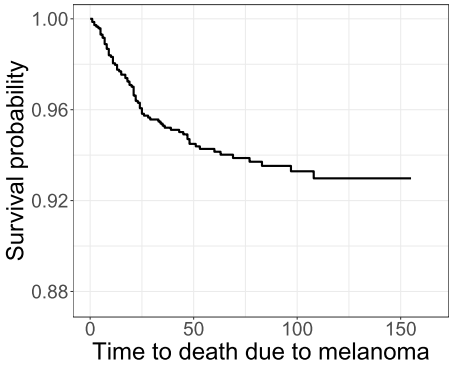


FIG 3. *Kaplan–Meier estimate of the survival function for the SEER data.*

fidence intervals are computed using 500 naive bootstrap samples. The effects of the covariates have the same direction for all methods. Note that the $\boldsymbol{\gamma}$ estimate of the LC model are normalized, so that the coefficients corresponding to the non-intercept terms possess a norm of one for better comparison with the other two methods. To compare the performance of the three models in predicting the uncure probability, we compute the prediction error (PE) using the testing data

$$\mathrm{PE} = -\sum_{i=1}^{481} \hat{w}_i \log\bigl(1 - \hat{p}(\mathbf{x}_i^{\mathrm{test}})\bigr) + (1 - \hat{w}_i) \log \hat{p}(\mathbf{x}_i^{\mathrm{test}}),$$

where $\hat{w}_i$ is computed using (9) with the estimated parameters (and the estimated link for SIC and mSIC.) The prediction errors are 95.47 for the LC model, 101.14 for the SIC model and 81.46 for the mSIC model. This indicates that mSIC performs better in predicting the uncure status among the three approaches. As mentioned previously, the SIC algorithm selected an optimal bandwidth of 1. The PE for SIC reduced slightly (to 97.15) when we allowed the SIC algorithm to search for some larger bandwidth.

TABLE 3
Parameter estimates for the SEER data (LC's $\hat{\gamma}_n$ is normalized)

	Covariates	LC			SIC			mSIC		
		Est.	Lower CI	Upper CI	Est.	Lower CI	Upper CI	Est.	Lower CI	Upper CI
	Intercept	−4.2129	−4.8391	−3.3821	-	-	-	-	-	-
$\hat{\boldsymbol{\gamma}}_n$	Age	0.0620	−0.0189	0.1463	0.1221	0.0586	0.3640	0.0487	−0.0633	0.0843
	Gender	0.1298	−0.0475	0.3019	0.2167	−0.0132	0.7204	0.1713	0.0246	0.3119
	Regional	0.5842	0.4812	0.7208	0.5217	0.1992	0.6674	0.5658	0.4843	0.7121
	Distant	0.7987	0.7266	0.8935	0.8161	0.7393	1.1943	0.8051	0.7270	0.8794
	Age	−0.0079	−0.0277	0.0163	−0.0086	−0.0283	0.0121	−0.0038	−0.0196	0.0188
	Gender	−0.1809	−1.0741	0.9811	−0.201	−1.0711	0.7866	−0.1609	−0.8706	0.7711
$\hat{\boldsymbol{\beta}}_n$	Regional	0.3908	−0.5542	1.4638	0.479	−0.3423	1.5802	0.2916	−0.5600	1.0817
	Distant	1.4290	0.1771	2.5478	1.5064	0.2783	2.6326	1.1356	−0.0604	2.0322

Figure 4 shows the link estimates of the three methods, plotted over the same range of the index. The LC estimate is rescaled according to the normalization

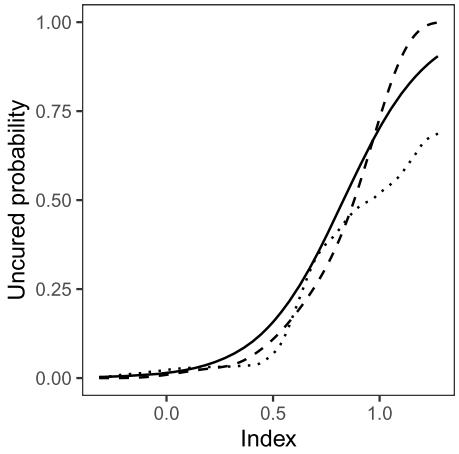


FIG 4. *Link estimates (solid: LC, dotted: SIC, dashed: mSIC.)*

of $\hat{\gamma}_n$. For this particular split of the data, all three methods give a monotone link function, and the links of LC and mSIC have similar shape. The estimated link function from SIC differs considerably from the other two methods but it has the largest prediction error. However, as illustrated in Appendix B.6, the SIC estimate is in general non-monotone and unstable. Part of this behavior seems to be due to its sensitivity to the bandwidth choice and the behavior of the cross-validation bandwidth selection method. mSIC on the other hand, apart from guaranteeing a monotone link estimate, is also more stable, less sensitive to the bandwidth and behaves well even for a simple (non-optimal) bandwidth choice. To achieve a more robust comparison among the three models using the prediction error, we also generate 10 random splits of the data (training: 964, testing: 481) and compute the prediction error (PE) using the testing data for each split. The averages (standard deviations) of the prediction errors over the 10 random splits are 89.64 (5.73) for LC; 91.84 (9.51) for SIC; and 71.84 (6.66) for mSIC. This suggests the same conclusion as before that mSIC behaves better in predicting the uncure probability among the three methods.

Appendix A: Technical lemmas and proofs

Proof of Proposition 3.1. We start by making an analogy with the standard isotonic regression problem, which corresponds to solving the following minimization problem

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}\in\mathcal{C}} \sum_{i=1}^{n} (x_i - y_i)^2 ,$$

where $\mathcal{C} = \{\mathbf{y} \in \mathbb{R}^n : \ y_1 \le y_2 \le \cdots \le y_n\}$, for given observations $x_1, \ldots, x_n$. From the theory of isotonic regression [33, Section 1.5], it is known that the

isotonic least squares estimate $\hat{\mathbf{y}}$ is also the solution of the following optimization problem

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}\in\mathcal{C}} \sum_{i=1}^{n} \Delta_{\Psi}\left(x_i, y_i\right), \tag{16}$$

where

$$\Delta_{\Psi}\left(x, y\right) = \Psi\left(x\right) - \Psi\left(y\right) - \left(x - y\right)\psi\left(y\right), \tag{17}$$

and Ψ is a convex function with derivative ψ. This equivalence is usually used to transform maximum likelihood isotonic estimation into an isotonic least-squares problem. Moreover, $\hat{\mathbf{y}}$ can be characterized as the left derivative of the greatest convex minorant (GCM) of the cumulative sum diagram (CSD)

$$\left\{\left(0,0\right), \left(i, \sum_{j=1}^{i} x_{(j)}\right), i = 1, \cdots, n\right\},$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ correspond to the ordered observations. If in addition to the order restriction, a uniform bound restriction is imposed on the solution, the problem becomes

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}\in\mathcal{C}'} \sum_{i=1}^{n} \left(x_i - y_i\right)^2,$$

with $\mathcal{C}' = \{\mathbf{y} \in \mathbb{R}^n : \ a \leq y_1 \leq y_2 \leq \cdots \leq y_n \leq b\}$ and is studied by [17]. Specifically, it has shown that the minimizer under the order and uniform bound restrictions is $\tilde{\mathbf{y}} = (\tilde{y}_1, \cdots, \tilde{y}_n)$ given by

$$\tilde{y}_i = \max\left(a, \min\left(\hat{y}_i, b\right)\right), \qquad i = 1, \cdots, n,$$

where $\hat{\mathbf{y}} = (\hat{y}_1, \cdots, \hat{y}_n)$ is the solution of (16).

Going back to our maximization problem in (11), first note that since the criterion depends only on the values of φ at the observed points $\boldsymbol{\gamma}^T\mathbf{x}_{(1)} < \cdots < \boldsymbol{\gamma}^T\mathbf{x}_{(n)}$ we can only identify the corresponding values $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$. If we choose the convex function $\Psi\left(u\right) = u\log u + (1-u)\log\left(1-u\right)$ in (17), we obtain

$$\begin{aligned} &\sum_{i=1}^{n} \Delta_{\Psi}\left(w_i, \varphi\left(\boldsymbol{\gamma}^T\mathbf{x}_i\right)\right) \\ &= \sum_{i=1}^{n} \left\{-w_i \log\varphi\left(\boldsymbol{\gamma}^T\mathbf{x}_i\right) - \left(1 - w_i\right)\log\left[1 - \varphi\left(\boldsymbol{\gamma}^T\mathbf{x}_i\right)\right]\right\} + C, \end{aligned} \tag{18}$$

for some C that depends on the observations and the fixed parameters but not on φ. Therefore, the maximization problem in (11) is equivalent to solving

$$\widehat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}\in\mathcal{C}_{\epsilon'}} \sum_{i=1}^{n} \Delta_{\Psi}\left(w_{(i)}, \phi_i\right),$$

where $\mathcal{C}_{\epsilon'} = \{\boldsymbol{\phi} \in \mathbb{R}^n : \ \epsilon' \leq \phi_1 \leq \phi_2 \leq \cdots \leq \phi_n \leq 1 - \epsilon'\}$. Consequently, using the previous results for the isotonic regression problem, $\widehat{\boldsymbol{\phi}}$ can be characterized as

$$\hat{\phi}_i = \max\left(\epsilon', \min\left(\tilde{\phi}_i, 1 - \epsilon'\right)\right), \qquad i = 1, \cdots, n,$$

where $\left(\tilde{\phi}_1, \cdots, \tilde{\phi}_n\right)$ are the left derivatives of the GCM of the CSD

$$\left\{ (0,0), \left(i, \sum_{j=1}^{i} w_{(j)} \right), i = 1, \cdots, n \right\}$$

with $w_{(j)}$ corresponding to the j-th order statistic of $(\boldsymbol{\gamma}^T \mathbf{x}_i)_i$. Note also that the procedure can easily accomodate ties in the observations of the index by summing the corresponding w_j.

Finally we note that, if we try to directly characterize the maximizer of the likelihood criterion in (5) over the non-decreasing function φ, we would still end up with an iterative procedure that is the same as the EM algorithm. Indeed, the problem is equivalent to maximizing

$$\mathcal{L}_n(\boldsymbol{\phi}) = \sum_{i=1}^{n} \Delta_{(i)} \log \phi_i + (1 - \Delta_{(i)}) \log\left\{1 - \phi_i + \phi_i S_u(Y_{(i)} | \mathbf{Z}_{(i)})\right\}$$

over $\phi_1 \leq \cdots \leq \phi_n$ denoting again the values of the link function on the ordered points $\boldsymbol{\gamma}^T \mathbf{x}_{(1)} < \cdots < \boldsymbol{\gamma}^T \mathbf{x}_{(n)}$. If we assume to know

$$W_{(i)} = \Delta_{(i)} + (1 - \Delta_{(i)}) \frac{\phi_i S_u(Y_{(i)} | \mathbf{Z}_{(i)})}{1 - \phi_i + \phi_i S_u(Y_{(i)} | \mathbf{Z}_{(i)})}$$

and choose $\Psi(u) = u \log u + (1 - u) \log(1 - u)$ as before, we obtain

$$\sum_{i=1}^{n} \Delta_{\Psi}\left(W_{(i)}, \phi_i\right) = -\mathcal{L}_n(\boldsymbol{\phi}) - \sum_{i=1}^{n} (1 - \Delta_{(i)}) W_{(i)} \log S_u(Y_{(i)} | \mathbf{Z}_{(i)}).$$

As a result, maximizing $\mathcal{L}_n(\boldsymbol{\phi})$ is equivalent to minimizing the expression on the left hand side of the equation. The solution can be characterized as the slope of the GCM of the cumulative sum diagram $\{(0,0), (i, \sum_{i=1}^{n} W_{(i)}), i = 1, \cdots, n\}$. However, the $W_{(i)}$ are actually not known and depend on $\boldsymbol{\phi}$. One could construct an iterative algorithm in which $\boldsymbol{\phi}$ from the previous step is used to compute $W_{(i)}$ and then update $\boldsymbol{\phi}$ again. This is the same as what EM algorithm does since the $W_{(i)}$ coincide with the ones defined in (9). □

Deriavation of the characterization in (14). By definition we have

$$\varphi_{0,\boldsymbol{\theta}} = \underset{\varphi \in \mathcal{M}_{\epsilon'}}{\arg\max}\, \mathbb{E}\left[\Delta \log \varphi(\boldsymbol{\gamma}^T \mathbf{X}) + (1 - \Delta) \log\left\{1 - \varphi(\boldsymbol{\gamma}^T \mathbf{X}) + \varphi(\boldsymbol{\gamma}^T \mathbf{X}) S_u(Y | \mathbf{Z}; \boldsymbol{\beta}, \Lambda)\right\}\right].$$

Let us assume that the covariates $\mathbf{X}$ and $\mathbf{Z}$ are independent and the censoring is independent of the other variables. Define $U = \boldsymbol{\gamma}^T\mathbf{X}$. Then the true likelihood for the observations $(Y, \Delta, U, \mathbf{Z})$ is

$$\begin{aligned}
&\Delta\mathbb{P}\left(C > Y, B = 1, T = Y \mid U, \mathbf{Z}\right) + (1-\Delta)\mathbb{P}\left(C = Y, T > Y \mid U, \mathbf{Z}\right) \\
&= \Delta\mathbb{P}\left(C > Y\right)\mathbb{P}\left(B = 1 \mid U\right)\mathbb{P}\left(T = Y \mid B = 1, \mathbf{Z}\right) \\
&\quad + (1-\Delta)\mathbb{P}\left(C = Y\right)\left\{\mathbb{P}\left(B = 0 \mid U\right) + \mathbb{P}\left(B = 1 \mid U\right)\mathbb{P}\left(T > Y \mid B = 1, \mathbf{Z}\right)\right\},
\end{aligned}$$

where with a slight abuse of notation we have used $\mathbb{P}\left(\cdot\right)$ instead of the density function for the continuous variables. By the law of conditional expectations we have

$$\mathbb{P}\left(B = 1 \mid U\right) = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{B = 1\right\} \mid \mathbf{X}\right] \mid U\right] = \mathbb{E}\left[\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X}) \mid U\right] =: \varphi^*_{\boldsymbol{\gamma}}(U).$$

Then the true log-likelihood for the observations $(Y, \Delta, U, \mathbf{Z})$ can be written as

$$\begin{aligned}
&= \Delta\log\varphi^*_{\boldsymbol{\gamma}}(U) + \Delta\log f_u(Y|\mathbf{Z}; \boldsymbol{\beta}_0, \Lambda_0) \\
&\quad + (1-\Delta)\log\left\{1 - \varphi^*_{\boldsymbol{\gamma}}(U) + \varphi^*_{\boldsymbol{\gamma}}(U)S_u(Y|\mathbf{Z}; \boldsymbol{\beta}_0, \Lambda_0)\right\} + c,
\end{aligned}$$

where c denotes other terms that do not depend on the model parameters (related to the censoring distribution). If we fix $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}_0, \Lambda_0)$, it follows from the Kullback–Leibler inequality that

$$\arg\max_{\varphi\in\mathcal{M}} \mathbb{E}\left[\Delta\log\varphi(U) + (1-\Delta)\log\left\{1 - \varphi(U) + \varphi(U)S_u(Y|\mathbf{Z}; \boldsymbol{\beta}_0, \Lambda_0)\right\}\right]$$

is obtained for $\varphi = \varphi^*_{\boldsymbol{\gamma}}$. On the other hand, Assumptions (A4), (A1)(i-ii), (A3)(ii) imply that there exists $\epsilon > 0$ such that $\epsilon \leq \varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X}) \leq 1 - \epsilon$. As a result, if we choose $\epsilon' < \epsilon$, $\varphi^*_{\boldsymbol{\gamma}} \in \mathcal{M}_{\epsilon'}$. Hence we conclude that under these more restrictive assumptions, $\varphi_{0,\boldsymbol{\theta}}$ is given by the expression in (14).

Next we comment on the necessity of such assumptions. To apply the previous argument we needed that $C \perp (B, T) \mid U, \mathbf{Z}$. Hence just conditional independence $C \perp (B, T) \mid \mathbf{X}, \mathbf{Z}$ would not be sufficient to split the probability into a product of probabilities. In addition, if $\mathbf{X}$ and $\mathbf{Z}$ would be dependent then we also would not have $\mathbb{P}\left(B = 1 \mid U, \mathbf{Z}\right) = \mathbb{P}\left(B = 1 \mid U\right)$. For example, if $\mathbf{X} = \mathbf{Z}$, we would have $\mathbb{P}\left(B = 1 \mid U, \mathbf{Z}\right) = \varphi_0(\boldsymbol{\gamma}_0^T\mathbf{Z})$ and there would be no connection between the function $\varphi^*_{\boldsymbol{\gamma}}$ and the true likelihood. Finally, if the latency parameters were not fixed to their true values $\boldsymbol{\beta}_0, \Lambda_0$, then the Kullback–Leibler inequality could not be used to conclude that the maximizer is $\varphi^*_{\boldsymbol{\gamma}}$. □

Proof of Proposition 4.1. Let $Q_{\mathbf{X}}$ denote the distribution of $\mathbf{X}$. For any $\varphi \in \mathcal{M}_{\epsilon'}$ and $\boldsymbol{\gamma} \in \Gamma$ define the L_2-norm

$$\|\varphi\|_{Q_{\mathbf{X}},\boldsymbol{\gamma}} = \left(\int_{\mathcal{X}} \varphi(\boldsymbol{\gamma}^T\mathbf{x})^2 \mathrm{d}Q_{\mathbf{X}}\left(\mathbf{x}\right)\right)^{\frac{1}{2}}.$$

To show the existence of a maximizer of $l_{\boldsymbol{\theta}}\left(\varphi\right)$ over $\mathcal{M}_{\epsilon'}$, it suffices to show that $l_{\boldsymbol{\theta}}$ is continuous on $(\mathcal{M}_{\epsilon'}, \|\cdot\|_{Q_{\mathbf{X}},\boldsymbol{\gamma}})$ and $(\mathcal{M}_{\epsilon'}, \|\cdot\|_{Q_{\mathbf{X}},\boldsymbol{\gamma}})$ is compact. To obtain

uniqueness of the maximizer, we show that $\varphi \longmapsto l_{\boldsymbol{\theta}}(\varphi)$ is strictly concave using the Gateaux derivative and $\mathcal{M}_{\epsilon'}$ is a convex set.

For simplicity we consider separately the two terms $l_{1,\boldsymbol{\theta}}(\varphi)=\mathbb{E}\left[\Delta \log \varphi\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right]$ and $l_{2,\boldsymbol{\theta}}(\varphi) = \mathbb{E}[(1-\Delta)\log\{1-\varphi(\boldsymbol{\gamma}^T\mathbf{X})F_u(Y|\mathbf{Z})\}]$. Let $(\varphi_m)_m \subset \mathcal{M}_{\epsilon'}$ be any sequence on $\mathcal{M}_{\epsilon'}$ that converges to $\tilde{\varphi} \in \mathcal{M}_{\epsilon'}$. For any $\eta > 0$, we can find an $N(\eta) \in \mathbb{N}$ such that $\|\varphi_m - \tilde{\varphi}\|_{Q_{\mathbf{X},\boldsymbol{\gamma}}} < \eta$, whenever $m \geq N(\eta)$. For $m \geq N(\eta)$, we have

$$\begin{aligned}
|l_{1,\boldsymbol{\theta}}(\varphi_m) - l_{1,\boldsymbol{\theta}}(\tilde{\varphi})| &= \left|\mathbb{E}\left[\Delta \log \varphi_m(\boldsymbol{\gamma}^T\mathbf{X})\right] - \mathbb{E}\left[\Delta \log \tilde{\varphi}(\boldsymbol{\gamma}^T\mathbf{X})\right]\right| \\
&\leq \frac{1}{\epsilon'}\left(\mathbb{E}\left[\Delta^2\right]\mathbb{E}\left[\left\{\varphi_m\left(\boldsymbol{\gamma}^T\mathbf{X}\right) - \tilde{\varphi}\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right\}^2\right]\right)^{\frac{1}{2}} \\
&\leq \frac{1}{\epsilon'}\|\varphi_m - \tilde{\varphi}\|_{Q_{\mathbf{X},\boldsymbol{\gamma}}} < \frac{\eta}{\epsilon'}.
\end{aligned}$$

Here, the second inequality follows from the Cauchy–Schwarz inequality, the mean value theorem and the bound on $\epsilon' \leq \tilde{\varphi}, \varphi_m \leq 1-\epsilon'$. In a similar way, for $m \geq N(\eta)$,

$$\begin{aligned}
|l_{2,\boldsymbol{\theta}}(\varphi_m) - l_{2,\boldsymbol{\theta}}(\tilde{\varphi})| &\leq \frac{1}{\epsilon'}\left(\mathbb{E}\left[\left\{(1-\Delta)F_u(Y|\mathbf{Z})\right\}^2\right]\right. \\
&\qquad \left.\times\mathbb{E}\left[\left\{\varphi_m\left(\boldsymbol{\gamma}^T\mathbf{X}\right) - \tilde{\varphi}\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right\}^2\right]\right)^{\frac{1}{2}} \\
&\leq \frac{1}{\epsilon'}\|\varphi_m - \tilde{\varphi}\|_{Q_{\mathbf{X},\boldsymbol{\gamma}}} < \frac{\eta}{\epsilon'}.
\end{aligned}$$

Therefore, take $\xi = 2\eta/\epsilon' > 0$ and $N = N(\xi\epsilon'/2)$, we have $|l_{\boldsymbol{\theta}}(\varphi_m) - l_{\boldsymbol{\theta}}(\tilde{\varphi})| < \xi$, whenever $m \geq N$, and hence $l_{\boldsymbol{\theta}}$ is continuous on $(\mathcal{M}_{\epsilon'}, \|\cdot\|_{Q_{\mathbf{X},\boldsymbol{\gamma}}})$.

Next, we show the compactness of $(\mathcal{M}_{\epsilon'}, \|\cdot\|_{Q_{\mathbf{X},\boldsymbol{\gamma}}})$. By Helly's selection theorem, since $(\varphi_m)_m \subset \mathcal{M}_{\epsilon'}$ is a sequence of uniformly bounded monotonically increasing functions, there exists a convergent subsequence $(\varphi_{m_k})_{k\in\mathbb{N}}$ of $(\varphi_m)_{m\in\mathbb{N}}$ such that $\varphi_{m_k}(u) \to \varphi^*(u)$ for almost all $u \in \mathbb{R}$ with respect to the Lebesgue measure. φ_{m_k} are all uniformly bounded monotonically increasing functions, so as its limit φ^*. By Assumption (A5), the distribution of $\boldsymbol{\gamma}^T\mathbf{X}$ has a density with respect to the Lebesgue measure, we have $\|\varphi_{m_k} - \varphi^*\|_{Q_{\mathbf{X},\boldsymbol{\gamma}}} \to 0$ almost everywhere by the dominated convergence theorem.

We show the convexity of $\mathcal{M}_{\epsilon'}$ by showing that the convex combination $t\varphi_1 + (1-t)\varphi_2$ belongs to $\mathcal{M}_{\epsilon'}$, for any $\varphi_1, \varphi_2 \in \mathcal{M}_{\epsilon'}$, and $t \in [0,1]$. Let $-\infty < x \leq y < \infty$. We have $\epsilon' \leq \varphi_1(x) \leq \varphi_1(y) \leq 1-\epsilon'$ and $\epsilon' \leq \varphi_2(x) \leq \varphi_2(y) \leq 1-\epsilon'$. These inequalities imply that $\epsilon' \leq t\varphi_1(x) + (1-t)\varphi_2(x) \leq t\varphi_1(y) + (1-t)\varphi_2(y) \leq 1-\epsilon'$. Hence, $t\varphi_1 + (1-t)\varphi_2 \in \mathcal{M}_{\epsilon'}$.

Define $H(t) = l_{\boldsymbol{\theta}}(t\varphi_1 + (1-t)\varphi_2)$, where $0 < t < 1$ and $\varphi_1, \varphi_2 \in \mathcal{M}_{\epsilon'}$ with $\varphi_1 \neq \varphi_2$. We show that the second derivative of $H(t)$ is negative on $(0,1)$ using the Gateaux derivative and hence $\varphi \longmapsto l_{\boldsymbol{\theta}}(\varphi)$ is strictly concave. For simplicity, denote $F_{u,0} = 1 - S_u(Y|\mathbf{Z})$ with the true parameters $\boldsymbol{\beta}_0$ and Λ_0, as in (3), and

$F_u = 1 - S_u(Y|\mathbf{Z})$ with parameters $\boldsymbol{\beta}$ and Λ. One can show that

$$H''(t) = \mathbb{E}\left[\frac{-\Delta\left[\varphi_1\left(\boldsymbol{\gamma}^T\mathbf{X}\right) - \varphi_2\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right]^2}{\left[t\varphi_1\left(\boldsymbol{\gamma}^T\mathbf{X}\right) + (1-t)\varphi_2\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right]^2}\right] + \mathbb{E}\left[\frac{-(1-\Delta)F_u^2\left[\varphi_1\left(\boldsymbol{\gamma}^T\mathbf{X}\right) - \varphi_2\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right]^2}{\left[1 - \left\{t\varphi_1\left(\boldsymbol{\gamma}^T\mathbf{X}\right) + (1-t)\varphi_2\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right\}F_u\right]^2}\right].$$

Bounding the denominators by one and replacing Δ by

$$\mathbb{E}\left[\Delta|\mathbf{X}, Y, \mathbf{Z}\right] = \varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}\left(Y|\mathbf{Z}\right)$$

using the law of iterated expectations, we obtain

$$H''(t) \leq -\mathbb{E}\left[\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{x})F_{u,0}\left\{1 - \varphi_0(\boldsymbol{\gamma}_0^T\mathbf{x})F_{u,0}\right\}\left\{\varphi_1(\boldsymbol{\gamma}^T\mathbf{x}) - \varphi_2(\boldsymbol{\gamma}^T\mathbf{x})\right\}^2\right]. \tag{19}$$

By Assumption (A3)(i), we have $\inf_{\mathbf{x}\in\mathcal{X}}\mathbb{P}\left[C > \tau_0|\mathbf{X} = \mathbf{x}\right] = c > 0$. Therefore, since $\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{x}) \geq \epsilon$,

$$\begin{aligned}\mathbb{E}\left[\Delta|\mathbf{X}\right] &= \mathbb{P}\left[B = 1, T \leq C|\mathbf{X}\right] \geq \mathbb{P}\left[B = 1, C \geq \tau_0|\mathbf{X}\right] = \mathbb{P}\left[B = 1|\mathbf{X}\right]\mathbb{P}\left[C \geq \tau_0|\mathbf{X}\right] \\ &\geq \epsilon \cdot \inf_{\mathbf{x}\in\mathcal{X}}\mathbb{P}\left[C > \tau_0|\mathbf{X} = \mathbf{x}\right] = c\epsilon > 0.\end{aligned} \tag{20}$$

Thus, using the law of iterated expectations and $\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0} \leq 1 - \epsilon$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}\left[1 - \varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}\right]\left[\varphi_1\left(\boldsymbol{\gamma}^T\mathbf{X}\right) - \varphi_2\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right]^2\right] \\ &\geq \epsilon \cdot \mathbb{E}\left[\mathbb{E}\left[\Delta|\mathbf{X}, Y, \mathbf{Z}\right]\left[\varphi_1\left(\boldsymbol{\gamma}^T\mathbf{X}\right) - \varphi_2\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right]^2\right] \\ &= \epsilon \cdot \mathbb{E}\left[\mathbb{E}\left[\Delta|\mathbf{X}\right]\left[\varphi_1\left(\boldsymbol{\gamma}^T\mathbf{X}\right) - \varphi_2\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right]^2\right] \\ &\geq c\epsilon^2 \cdot \mathbb{E}\left[\left\{\varphi_1\left(\boldsymbol{\gamma}^T\mathbf{X}\right) - \varphi_2\left(\boldsymbol{\gamma}^T\mathbf{X}\right)\right\}^2\right] > 0.\end{aligned} \tag{21}$$

Hence $H''(t) < 0$, for $0 < t < 1$. □

Proof of Proposition 4.2. If $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0, \Lambda_0)$, by the definition of $\varphi_{0,\boldsymbol{\theta}}$ in Proposition 4.1, we have

$$\mathbb{E}\left[l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}_0, \varphi\right)\right] \leq \mathbb{E}\left[l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}_0, \varphi_{0,\boldsymbol{\theta}_0}\right)\right] \quad \text{for all } \varphi \in \mathcal{M}_{\epsilon'},$$

with equality if $\varphi = \varphi_{0,\boldsymbol{\theta}}$. From the Kullback–Leibler inequality, we also have

$$\mathbb{E}\left[l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}_0, \varphi\right)\right] \leq \mathbb{E}\left[l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}_0, \varphi_0\right)\right] \quad \text{for all } \varphi \in \mathcal{M}_{\epsilon'},$$

with equality if $\varphi = \varphi_0$. Since, as shown in Proposition 4.1, $\varphi_{0,\boldsymbol{\theta}_0}$ is the unique maximizer of $\varphi \mapsto \mathbb{E}\left[l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}_0, \varphi\right)\right]$, it follows that $\varphi_{0,\boldsymbol{\theta}_0} = \varphi_0$. □

Proof of Proposition 4.3. We observe that, as in the standard logistic-Cox cure model, the maximizer $\hat{\Lambda}_n$ must be a step function with jumps at the observed times. Let $\lambda(t)$ be the jump size of Λ at t. Also, $u \longmapsto \hat{\varphi}^s_{n,\boldsymbol{\theta}}(u)$ is continuous on $\mathbb{R}$, therefore the likelihood function L_n is continuous with respect to $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and the jump sizes of Λ. Hence the existence and finiteness of the maximum likelihood estimator $(\hat{\boldsymbol{\gamma}}_n, \hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n)$ follows as in the proof of Theorem 1 in [26]. □

Proof of Proposition 4.4. Recall that $\varphi_{0,\boldsymbol{\theta}} = \arg\max_{\varphi \in \mathcal{M}_{\epsilon'}} l_{\boldsymbol{\theta}}(\varphi)$ for fixed $\boldsymbol{\theta}$. Let $\tilde{\varphi} \in \mathcal{M}_{\epsilon'}$ and define $H(t) = l_{\boldsymbol{\theta}}(t\tilde{\varphi} + (1-t)\varphi_{0,\boldsymbol{\theta}})$, where $0 \le t \le 1$. By the definition of $\varphi_{0,\boldsymbol{\theta}}$ and the convexity of $\mathcal{M}_{\epsilon'}$, $H'(0) \le 0$ for all $\tilde{\varphi} \in \mathcal{M}_{\epsilon'}$. We show the continuity of $\varphi_{0,\boldsymbol{\theta}}$ by contradiction.

For simplifying the notation, we denote $\tilde{\varphi} = \tilde{\varphi}(\boldsymbol{\gamma}^T\mathbf{X})$, $\varphi_{0,\boldsymbol{\theta}} = \varphi_{0,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T\mathbf{X})$, $F_{u,0} = F_u(Y|\mathbf{Z};\boldsymbol{\beta}_0,\Lambda_0)$, and $F_u = F_u(Y|\mathbf{Z};\boldsymbol{\beta},\Lambda)$. Then, for all $\tilde{\varphi} \in \mathcal{M}_{\epsilon'}$,

$$
\begin{aligned}
H'(0) &= \mathbb{E}\left[\frac{\Delta(\tilde{\varphi}-\varphi_{0,\boldsymbol{\theta}})}{\varphi_{0,\boldsymbol{\theta}}} - \frac{(1-\Delta)(\tilde{\varphi}-\varphi_{0,\boldsymbol{\theta}})F_u}{1-\varphi_{0,\boldsymbol{\theta}}F_u}\right] \\
&= \mathbb{E}\left[\frac{\tilde{\varphi}-\varphi_{0,\boldsymbol{\theta}}}{\varphi_{0,\boldsymbol{\theta}}}\frac{\Delta-\varphi_{0,\boldsymbol{\theta}}F_u}{1-\varphi_{0,\boldsymbol{\theta}}F_u}\right] \\
&= \mathbb{E}\left[\frac{\tilde{\varphi}-\varphi_{0,\boldsymbol{\theta}}}{\varphi_{0,\boldsymbol{\theta}}}\frac{\mathbb{E}[\Delta|\mathbf{X},Y,\mathbf{Z}]-\varphi_{0,\boldsymbol{\theta}}F_u}{1-\varphi_{0,\boldsymbol{\theta}}F_u}\right] \\
&= \mathbb{E}\left[\frac{\tilde{\varphi}-\varphi_{0,\boldsymbol{\theta}}}{\varphi_{0,\boldsymbol{\theta}}}\frac{\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}-\varphi_{0,\boldsymbol{\theta}}F_u}{1-\varphi_{0,\boldsymbol{\theta}}F_u}\right] \\
&= \mathbb{E}\left[\frac{\tilde{\varphi}-\varphi_{0,\boldsymbol{\theta}}}{\varphi_{0,\boldsymbol{\theta}}}\left\{1-\mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}\right]\right\}\right] \le 0.
\end{aligned}
\tag{22}
$$

Next we argue that, if $\varphi_{0,\boldsymbol{\theta}}$ was not continuous, we can construct $\tilde{\varphi} \in \mathcal{M}_{\epsilon'}$ for which the inequality (22) is not satisfied. Hence, by contradiction we can conclude that $\varphi_{0,\boldsymbol{\theta}}$ is continuous.

Since $\varphi_{0,\boldsymbol{\theta}}$ is monotone non-decreasing, discontinuity points would be points of jump. Assume that $\varphi_{0,\boldsymbol{\theta}}$ has a jump at $\boldsymbol{\gamma}^T\tilde{\mathbf{x}}$ with a size of κ. Let

$$
A(u) = \mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T\tilde{\mathbf{x}})F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right].
$$

We consider three cases based on whether $A(\boldsymbol{\gamma}^T\tilde{\mathbf{x}})$ is smaller than one, larger than one or equal to one.

Case 1. Suppose that $A(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}) < 1$. By Assumption (A8), there exists $\eta > 0$, such that $A(u) < 1$ for $u \in (\boldsymbol{\gamma}^T\tilde{\mathbf{x}}-\eta, \boldsymbol{\gamma}^T\tilde{\mathbf{x}}+\eta)$. Since $\varphi_{0,\boldsymbol{\theta}}$ is monotone non-decreasing and $\varphi_{0,\boldsymbol{\theta}}$ has a jump at $\boldsymbol{\gamma}^T\tilde{\mathbf{x}}$, we consider that $\varphi_{0,\boldsymbol{\theta}}$ is either right- or left-continuous at this point and construct $\tilde{\varphi}$ which coincides with $\varphi_{0,\boldsymbol{\theta}}$ apart from in a small neighborhood of $\boldsymbol{\gamma}^T\tilde{\mathbf{x}}$.

(i) If $\varphi_{0,\boldsymbol{\theta}}$ is right-continuous at $\boldsymbol{\gamma}^T\tilde{\mathbf{x}}$, $\varphi_{0,\boldsymbol{\theta}}(u) < \varphi_{0,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T\tilde{\mathbf{x}})$ for $u \in [\boldsymbol{\gamma}^T\tilde{\mathbf{x}} - \frac{\eta}{2}, \boldsymbol{\gamma}^T\tilde{\mathbf{x}})$. We define $\tilde{\varphi}(u) = \varphi_{0,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}) > \varphi_{0,\boldsymbol{\theta}}(u)$ for $u \in [\boldsymbol{\gamma}^T\tilde{\mathbf{x}} - \frac{\eta}{2}, \boldsymbol{\gamma}^T\tilde{\mathbf{x}})$ and $\tilde{\varphi}(u) = \varphi_{0,\boldsymbol{\theta}}(u)$ otherwise.

(ii) If $\varphi_{0,\boldsymbol{\theta}}$ is left-continuous at $\boldsymbol{\gamma}^T\tilde{\mathbf{x}}$, $\varphi_{0,\boldsymbol{\theta}}(u) \leq \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right) < \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right) + \kappa$ for $u \in [\boldsymbol{\gamma}^T\tilde{\mathbf{x}} - \frac{\eta}{2}, \boldsymbol{\gamma}^T\tilde{\mathbf{x}})$. We define $\tilde{\varphi}(u) = \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right) + \kappa > \varphi_{0,\boldsymbol{\theta}}(u)$ for $u \in [\boldsymbol{\gamma}^T\tilde{\mathbf{x}} - \frac{\eta}{2}, \boldsymbol{\gamma}^T\tilde{\mathbf{x}})$ and $\tilde{\varphi}(u) = \varphi_{0,\boldsymbol{\theta}}(u)$ otherwise.

In addition, monotonicity of $\varphi_{0,\boldsymbol{\theta}}$ implies that

$$\mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}(u)F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] \leq \mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T\tilde{\mathbf{x}})F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] = A(u) < 1,$$

for $u \in [\boldsymbol{\gamma}^T\tilde{\mathbf{x}} - \frac{\eta}{2}, \boldsymbol{\gamma}^T\tilde{\mathbf{x}})$. This inequality together with the construction of $\tilde{\varphi}$ in either case give

$$\begin{aligned} H'(0) &= \mathbb{E}\left[\frac{\tilde{\varphi}-\varphi_{0,\boldsymbol{\theta}}}{\varphi_{0,\boldsymbol{\theta}}}\left\{1-\mathbb{E}\left[\frac{1-\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}\right]\right\}\right] \\ &= \mathbb{E}\left[\mathbb{1}\left\{\boldsymbol{\gamma}^T\tilde{\mathbf{x}} - \frac{\eta}{2} \leq \boldsymbol{\gamma}^T\mathbf{X} < \boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right\}\frac{\tilde{\varphi}-\varphi_{0,\boldsymbol{\theta}}}{\varphi_{0,\boldsymbol{\theta}}}\right. \\ &\qquad \left.\left\{1-\mathbb{E}\left[\frac{1-\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}\right]\right\}\right], \end{aligned}$$

is strictly larger than zero, which contradicts the fact that $H'(0) \leq 0$ for all $\tilde{\varphi} \in \mathcal{M}_{\epsilon'}$.

Case 2. Suppose that $A(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}) > 1$. By Assumption (A8), there exists $\eta > 0$, such that $A(u) > 1$ for $u \in (\boldsymbol{\gamma}^T\tilde{\mathbf{x}} - \eta, \boldsymbol{\gamma}^T\tilde{\mathbf{x}} + \eta)$. Similar to the argument in the first case, we can show that

$$\mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}(u)F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] \geq \mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T\tilde{\mathbf{x}})F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] = A(u) > 1$$

for $u \in (\boldsymbol{\gamma}^T\tilde{\mathbf{x}}, \boldsymbol{\gamma}^T\tilde{\mathbf{x}} + \frac{\eta}{2}]$. Similarly to the first case, we can construct $\tilde{\varphi}$ which coincides with $\varphi_{0,\boldsymbol{\theta}}$ apart from in a small neighbourhood of $\boldsymbol{\gamma}^T\tilde{\mathbf{x}}$ where it is constant and strictly smaller than $\varphi_{0,\boldsymbol{\theta}}$. We then have $H'(0) > 0$ and leads to a contradiction.

Case 3. Suppose that $A(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}) = 1$.

(i) If $\varphi_{0,\boldsymbol{\theta}}$ is right-continuous at $\boldsymbol{\gamma}^T\tilde{\mathbf{x}}$, we have $\varphi_{0,\boldsymbol{\theta}}(u) \leq \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right) - \kappa < \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)$ for $u < \boldsymbol{\gamma}^T\tilde{\mathbf{x}}$. This implies that for $u < \boldsymbol{\gamma}^T\tilde{\mathbf{x}}$,

$$\begin{aligned} \mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}(u)F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] &\leq \mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)-\kappa\right]F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] \\ &< \mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] = A(u). \end{aligned} \tag{23}$$

We want to show that the left hand side of the previous inequality is smaller than one in a left neighborhood of $\boldsymbol{\gamma}^T\tilde{\mathbf{x}}$ and then proceed with the construction of $\tilde{\varphi}$ as in Case 1. By the mean value theorem, we have

$$\frac{1}{1-\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)F_u\left(Y|\mathbf{Z}\right)}-\frac{1}{1-\left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)-\kappa\right]F_u\left(Y|\mathbf{Z}\right)}=\frac{\kappa F_u\left(Y|\mathbf{Z}\right)}{\left(1-\xi\right)^2}$$

for some $\left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)-\kappa\right]F_u\left(Y|\mathbf{Z}\right)<\xi<\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)F_u\left(Y|\mathbf{Z}\right)$. Since $\varphi_{0,\boldsymbol{\theta}}\in\mathcal{M}_{\epsilon'}$, we have $\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)-\kappa\geq\epsilon'$ and $1>\xi>\epsilon' F_u\left(Y|\mathbf{Z}\right)\geq 0$. It follows that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1-\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)F_u\left(Y|\mathbf{Z}\right)}-\frac{1-\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}}{1-\left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)-\kappa\right]F_u\left(Y|\mathbf{Z}\right)}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right]\\
&\geq\mathbb{E}\left[\left\{1-\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}\right\}\kappa F_u\left(Y|\mathbf{Z}\right)\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right]\\
&=\kappa\cdot\mathbb{E}\left[\mathbb{E}\left[1-\Delta|\mathbf{X},Y,\mathbf{Z}\right]F_u\left(Y|\mathbf{Z}\right)|\boldsymbol{\gamma}^T\mathbf{X}=u\right]\\
&\geq\kappa\cdot\mathbb{E}\left[\left(1-\Delta\right)F_u\left(Y|\mathbf{Z}\right)\mathbb{1}\left\{Y>\tau_0\right\}|\boldsymbol{\gamma}^T\mathbf{X}=u\right]\\
&=\kappa\cdot\mathbb{P}\left[Y>\tau_0|\boldsymbol{\gamma}^T\mathbf{X}=u\right]\\
&=\kappa\cdot\mathbb{P}\left[B=0,C>\tau_0|\boldsymbol{\gamma}^T\mathbf{X}=u\right]\\
&=\kappa\cdot\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{B=0\right\}\mathbb{1}\left\{C>\tau_0\right\}|\mathbf{X}\right]|\boldsymbol{\gamma}^T\mathbf{X}=u\right]\\
&\geq c\epsilon\kappa>0.
\end{aligned}\tag{24}$$

Here the second equality follows from the definition of τ_0 in Assumption (A3)(i) and $F_u\left(y|\mathbf{z}\right)=1$ if $y>\tau_0$. The last inequality follows from $\mathbb{P}\left[B=1,C\geq\tau_0|\mathbf{X}\right]\geq c\epsilon$ as shown in (20). By Assumption (A8), there exists $\eta>0$, such that $A(u)<1+c\epsilon\kappa$ for $u\in(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}-\eta,\boldsymbol{\gamma}^T\tilde{\mathbf{x}}+\eta)$. Equation (24) implies that, for $u\in(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}-\eta,\boldsymbol{\gamma}^T\tilde{\mathbf{x}})$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1-\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}}{1-\left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)-\kappa\right]F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right]+c\epsilon\kappa\\
&\leq\mathbb{E}\left[\frac{1-\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right]=A(u)<1+c\epsilon\kappa.
\end{aligned}$$

From (23), it follows that, for $u\in(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}-\eta,\boldsymbol{\gamma}^T\tilde{\mathbf{x}})$,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}\left(u\right)F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right]\\
&\leq\mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)-\kappa\right]F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right]<1.
\end{aligned}$$

By a similar construction of $\tilde{\varphi}$ as in the first case, we get $H'(0) > 0$ which leads to a contradiction.
(ii) If $\varphi_{0,\boldsymbol{\theta}}$ is left-continuous at $\boldsymbol{\gamma}^T\tilde{\mathbf{x}}$, we have $\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right) < \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right) + \kappa \leq \varphi_{0,\boldsymbol{\theta}}(u)$ for $u > \boldsymbol{\gamma}^T\tilde{\mathbf{x}}$. This implies that

$$\mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] \leq \mathbb{E}\left[\frac{1-\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}}{1-\left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)+\kappa\right]F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right]$$
$$< \mathbb{E}\left[\frac{1-\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}(u)F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right].$$

As in (i),

$$\frac{1}{1-\left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)+\kappa\right]F_u(Y|\mathbf{Z})} - \frac{1}{1-\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)F_u} = \frac{\kappa F_u}{(1-\xi)^2}$$

for some $\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)F_u < \xi < \left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)+\kappa\right]F_u$. Using a similar argument in (i), one can show that

$$\mathbb{E}\left[\frac{1-\varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{X}\right)F_{u,0}}{1-\left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\tilde{\mathbf{x}}\right)+\kappa\right]F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] > A(u) + c\epsilon\kappa.$$

By Assumption (A8), there exists $\eta > 0$, such that $A(u) > 1 - c\epsilon\kappa$ for $u \in (\boldsymbol{\gamma}^T\tilde{\mathbf{x}} - \eta, \boldsymbol{\gamma}^T\tilde{\mathbf{x}} + \eta)$. Hence, we have for $u \in (\boldsymbol{\gamma}^T\tilde{\mathbf{x}}, \boldsymbol{\gamma}^T\tilde{\mathbf{x}} + \eta)$,

$$\mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-\varphi_{0,\boldsymbol{\theta}}(u)F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] \geq \mathbb{E}\left[\frac{1-\varphi_0(\boldsymbol{\gamma}_0^T\mathbf{X})F_{u,0}}{1-[\varphi_{0,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T\tilde{\mathbf{x}})+\kappa]F_u}\middle|\boldsymbol{\gamma}^T\mathbf{X}=u\right] > 1.$$

By a similar construction of $\tilde{\varphi}$ as in the second case, we get $H'(0) > 0$ and leads to a contradiction. □

For the proof of Proposition 4.5 we need results on the entropy numbers of certain classes of functions, which we derive below. First we introduce some notation that will be used in the following series of Lemmas. Consider a class of functions $\mathcal{F}$ equipped with a norm $\|\cdot\|$. For $\zeta > 0$, the bracketing number $N_B(\zeta, \mathcal{F}, \|\cdot\|) = N$ is the minimal number of pairs of functions $\{[f_j^L, f_j^U]\colon j = 1, \cdots, N\}$ such that $\|f_j^U - f_j^L\| \leq \zeta$, for all $j = 1, \cdots, N$ and for each $f \in \mathcal{F}$, there is a $j \in \{1, \cdots, N\}$ such that $f_j^L \leq f \leq f_j^U$. The ζ–entropy with bracketing of $\mathcal{F}$ is the logarithm of the bracketing number, $H_B(\zeta, \mathcal{F}, \|\cdot\|) = \log N_B(\zeta, \mathcal{F}, \|\cdot\|)$. $\mathbb{P}_0$ denotes the distribution of $(Y, \Delta, \mathbf{X}, \mathbf{Z})$. We will consider the following classes of functions.

- $\mathcal{M}$ is the class of all non-decreasing bounded functions on $\mathbb{R}$.
- $\mathcal{F}$ is the class of functions $F(y, \mathbf{z}) = 1 - \exp\left[-\Lambda(y)\exp\left(\boldsymbol{\beta}^T\mathbf{z}\right)\right]$, $y \in [0, \tau_0]$ and $\mathbf{z} \in \mathcal{Z}$, where $\Lambda \in \tilde{\mathcal{D}}$ and $\boldsymbol{\beta} \in \mathcal{B}$.

- $\mathcal{G}_{\epsilon'}$ is the class of functions $g\left(y,\delta,\mathbf{x},\mathbf{z}\right)=\left(1-\delta\right)\log\left[1-\varphi\left(\boldsymbol{\gamma}^{T}\mathbf{x}\right)F\left(y,\mathbf{z}\right)\right]$, $y\in[0,\tau_0]$, $\delta\in\{0,1\}$, $\mathbf{x}\in\mathcal{X}$ and $\mathbf{z}\in\mathcal{Z}$, where $\boldsymbol{\gamma}\in\mathcal{S}_{d-1}$, $\varphi\in\mathcal{M}_{\epsilon'}$ and $F\in\mathcal{F}$,
- $\mathcal{H}_{\epsilon'}$ is the class of functions $h\left(\delta,\mathbf{x}\right)=\delta\log\varphi\left(\boldsymbol{\gamma}^{T}\mathbf{x}\right)$, $\delta\in\{0,1\}$ and $\mathbf{x}\in\mathcal{X}$, where $\boldsymbol{\gamma}\in\mathcal{S}_{d-1}$ and $\varphi\in\mathcal{M}_{\epsilon'}$.
- $\mathcal{L}_{\epsilon'}$ is the class of functions

$$l\left(y,\delta,\mathbf{x},\mathbf{z}\right)=\delta\log\varphi\left(\boldsymbol{\gamma}^{T}\mathbf{x}\right)+\left(1-\delta\right)\log\left[1-\varphi\left(\boldsymbol{\gamma}^{T}\mathbf{x}\right)F\left(y,\mathbf{z}\right)\right],$$

  $y\in[0,\tau_0]$, $\delta\in\{0,1\}$, $\mathbf{x}\in\mathcal{X}$ and $\mathbf{z}\in\mathcal{Z}$, where $\boldsymbol{\gamma}\in\mathcal{S}_{d-1}$, $\varphi\in\mathcal{M}_{\epsilon'}$ and $F\in\mathcal{F}$.
- $\tilde{\mathcal{L}}_{\epsilon'}$ is the class of functions

$$\begin{aligned}\tilde{l}\left(y,\delta,\mathbf{x},\mathbf{z}\right)=&\,\delta\left[\log\varphi_1(\boldsymbol{\gamma}^{T}\mathbf{x})-\log\varphi_2(\boldsymbol{\gamma}^{T}\mathbf{x})\right]\\&+\left(1-\delta\right)\{\log[1-\varphi_1(\boldsymbol{\gamma}^{T}\mathbf{x})F(y,\mathbf{z})]-\log[1-\varphi_1(\boldsymbol{\gamma}^{T}\mathbf{x})F(y,\mathbf{z})]\},\end{aligned}$$

  for $y\in[0,\tau_0]$, $\delta\in\{0,1\}$, $\mathbf{x}\in\mathcal{X}$, $\mathbf{z}\in\mathcal{Z}$, $\boldsymbol{\gamma}\in\mathcal{S}_{d-1}$, $\varphi_1,\varphi_2\in\mathcal{M}_{\epsilon'}$ and $F\in\mathcal{F}$.

We begin with a result from [38] and we use it to construct ζ–brackets for the other classes.

Lemma A.1 (Theorem 2.7.5 of [38])**.** *There exists a constant $A>0$ such that*

$$H_B\left(\zeta,\mathcal{M},\|\cdot\|_{r,Q}\right)\leq\frac{A}{\zeta},$$

for all $\zeta>0$, $r\geq1$, and all probability measures Q on $\mathbb{R}$, where $\|\cdot\|_{r,Q}$ is the L_r–norm corresponding to Q.

Lemma A.2. *Suppose that Assumptions (A4) and (A10) are satisfied. Let $\zeta>0$. There exists a constant $C_1>0$ depending on $\bar{q}_2$, and r_2 only such that*

$$H_B\left(\zeta,\mathcal{F},\|\cdot\|_{1,\mathbb{P}_0}\right)\leq\frac{C_1\left(q+1\right)}{\zeta}.$$

Furthermore, there exists a constant $C_2>0$ depending on $\bar{q}_1$, and r_1 and a constant $C_3>0$ depending on $\epsilon',\bar{q}_1,\bar{q}_2,r_1$, and r_2 such that

$$H_B\left(\zeta,\mathcal{G}_{\epsilon'},\|\cdot\|_{1,\mathbb{P}_0}\right)\leq\frac{C_2d+C_3\left(q+1\right)}{\zeta}.$$

Moreover, there exists a constant $C_4>0$ depending on $\epsilon',\bar{q}_1$, and r_1 such that

$$H_B\left(\zeta,\mathcal{H}_{\epsilon'},\|\cdot\|_{1,\mathbb{P}_0}\right)\leq\frac{C_4\left(d+1\right)}{\zeta}.$$

Proof. Let $\zeta_{\boldsymbol{\beta}}\in(0,1)$. Since $\mathcal{B}$ is a compact subset of $\mathbb{R}^q$, it can be covered by N_1 balls with radius $\zeta_{\boldsymbol{\beta}}$, where $N_1\leq(A_1/\zeta_{\boldsymbol{\beta}})^q$ for some constant $A_1>0$. Let $\left\{\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_{N_1}\right\}$ be the centers of such balls. Consider $f\left(y,\mathbf{z}\right)=1-$

$\exp\left[-\Lambda(y)\exp\left(\boldsymbol{\beta}^T\mathbf{z}\right)\right] \in \mathcal{F}$, $y \in [0,\tau_0]$ and $\mathbf{z} \in \mathcal{Z}$, for some $\Lambda \in \tilde{\mathcal{D}}$ and $\boldsymbol{\beta} \in \mathcal{B}$. We can find $j \in \{1,\cdots,N_1\}$ such that $\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_j\right\|_2 \le \zeta_{\boldsymbol{\beta}}$. By the monotonicity of the exponential function and the Cauchy–Schwarz inequality, we have

$$\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}-\zeta_{\boldsymbol{\beta}}r_2\right) \le \exp\left(\boldsymbol{\beta}^T\mathbf{z}\right) \le \exp\left(\boldsymbol{\beta}_j^T\mathbf{z}+\zeta_{\boldsymbol{\beta}}r_2\right),$$

for all $\mathbf{z} \in \mathcal{Z}$. Let $\zeta_\Lambda > 0$. By Lemma A.1, the class $\tilde{\mathcal{D}}$ can be covered by ζ_Λ–brackets $[\Lambda_l^L, \Lambda_l^U]$, $l = 1,\cdots,N_2$, such that

$$\int_{[0,\tau_0]} \left[\Lambda_l^U(y)-\Lambda_l^L(y)\right]^2 \mathrm{d}Q_Y(y) \le \zeta_\Lambda^2,$$

where Q_Y denotes the distribution of Y and $N_2 \le \exp(A_2/\zeta_\Lambda)$ for some constant $A_2 > 0$. We note that Λ_l^L and Λ_l^U can always be taken to be bounded below by 0 and bounded above by the uniform upper bound M of the class $\tilde{\mathcal{D}}$. Otherwise, we can take $\Lambda_l^L \vee 0$ and $\Lambda_l^U \wedge M$. Then, we have

$$1-\exp\left[-\Lambda_l^L(y)\exp(\boldsymbol{\beta}_j^T\mathbf{z}-\zeta_{\boldsymbol{\beta}}r_2)\right] \le f(y,\mathbf{z}) \le 1-\exp\left[-\Lambda_l^U(y)\exp(\boldsymbol{\beta}_j^T\mathbf{z}+\zeta_{\boldsymbol{\beta}}r_2)\right],$$

for some $l = 1,\cdots,N_2$, and for all $y \in [0,\tau_0]$ and $\mathbf{z} \in \mathcal{Z}$. We then show that, for certain choices of ζ_Λ and $\zeta_{\boldsymbol{\beta}}$, the brackets

$$[f_l^L, f_l^U] = \Big[1-\exp\left\{-\Lambda_l^L(y)\exp(\boldsymbol{\beta}_j^T\mathbf{z}-\zeta_{\boldsymbol{\beta}}r_2)\right\}, 1-\exp\left\{-\Lambda_l^U(y)\exp(\boldsymbol{\beta}_j^T\mathbf{z}+\zeta_{\boldsymbol{\beta}}r_2)\right\}\Big]$$

are ζ–brackets with respect to $\|\cdot\|_{1,\mathbb{P}_0}$ that cover $\mathcal{F}$. We have

$$\begin{aligned}\|f_l^U-f_l^L\|_{1,\mathbb{P}_0} = \int_{[0,\tau_0]\times\mathcal{Z}} \Big|&\exp\left[-\Lambda_l^U(y)\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}+\zeta_{\boldsymbol{\beta}}r_2\right)\right] \\ &-\exp\left[-\Lambda_l^L(y)\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}-\zeta_{\boldsymbol{\beta}}r_2\right)\right]\Big| \mathrm{d}Q_{Y,\mathbf{Z}}(y,\mathbf{z}).\end{aligned} \tag{25}$$

Considering the integrand of (25), by the mean value theorem we obtain

$$\begin{aligned}&\left|\exp\left[-\Lambda_l^U(y)\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}+\zeta_{\boldsymbol{\beta}}r_2\right)\right]-\exp\left[-\Lambda_l^L(y)\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}-\zeta_{\boldsymbol{\beta}}r_2\right)\right]\right| \\ &= \left|\exp(-\xi)\left[\Lambda_l^L(y)\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}-\zeta_{\boldsymbol{\beta}}r_2\right)-\Lambda_l^U(y)\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}+\zeta_{\boldsymbol{\beta}}r_2\right)\right]\right| \\ &\le \left|\Lambda_l^L(y)\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}-\zeta_{\boldsymbol{\beta}}r_2\right)-\Lambda_l^U(y)\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}+\zeta_{\boldsymbol{\beta}}r_2\right)\right|,\end{aligned}$$

for some $\xi \in (\Lambda_l^L(y)\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}-\zeta_{\boldsymbol{\beta}}r_2\right), \Lambda_l^U(y)\exp\left(\boldsymbol{\beta}_j^T\mathbf{z}+\zeta_{\boldsymbol{\beta}}r_2\right))$. As a result,

from Minkowski inequality it follows that

$$\begin{aligned}\|f_l^U - f_l^L\|_{1,\mathbb{P}_0} \leq & \int_{[0,\tau_0]\times\mathcal{Z}} \left|\Lambda_l^L(y) - \Lambda(y)\right| \exp(\boldsymbol{\beta}_j^T\mathbf{z} - \zeta_{\boldsymbol{\beta}} r_2) \mathrm{d}Q_{Y,\mathbf{Z}}(y,\mathbf{z}) \\ & + \int_{[0,\tau_0]\times\mathcal{Z}} \Lambda(y) \left|\exp(\boldsymbol{\beta}_j^T\mathbf{z} - \zeta_{\boldsymbol{\beta}} r_2) - \exp(\boldsymbol{\beta}_j^T\mathbf{z} + \zeta_{\boldsymbol{\beta}} r_2)\right| \mathrm{d}Q_{Y,\mathbf{Z}}(y,\mathbf{z}) \\ & + \int_{[0,\tau_0]\times\mathcal{Z}} \left|\Lambda(y) - \Lambda_l^U(y)\right| \exp(\boldsymbol{\beta}_j^T\mathbf{z} + \zeta_{\boldsymbol{\beta}} r_2) \mathrm{d}Q_{Y,\mathbf{Z}}(y,\mathbf{z}). \end{aligned} \tag{26}$$

Consider the first integral on the right hand side of the last inequality in (26). Using the Cauchy–Schwarz inequality, we have

$$\begin{aligned} & \int_{[0,\tau_0]\times\mathcal{Z}} \left|\Lambda_l^L(y) - \Lambda(y)\right| \exp\left(\boldsymbol{\beta}_j^T\mathbf{z} - \zeta_{\boldsymbol{\beta}} r_2\right) \mathrm{d}Q_{Y,\mathbf{Z}}(y,\mathbf{z}) \\ & \leq \left\{ \int_{[0,\tau_0]} \left[\Lambda_l^L(y) - \Lambda(y)\right]^2 \mathrm{d}Q_Y(y) \cdot \int_{\mathcal{Z}} \left[\exp\left(\boldsymbol{\beta}_j^T\mathbf{z} - \zeta_{\boldsymbol{\beta}} r_2\right)\right]^2 \mathrm{d}Q_{\mathbf{Z}}(\mathbf{z}) \right\}^{\frac{1}{2}} \\ & \leq \exp(b)\,\zeta_\Lambda \leq \exp(b + r_2)\,\zeta_\Lambda, \end{aligned}$$

where b is such that the support of $\boldsymbol{\beta}_j^T\mathbf{Z}$ is included in an interval $[a,b]$ for all j. Similarly, we also have

$$\int_{[0,\tau_0]\times\mathcal{Z}} \left|\Lambda(y) - \Lambda_l^U(y)\right| \exp\left(\boldsymbol{\beta}_j^T\mathbf{z} + \zeta_{\boldsymbol{\beta}} r_2\right) \mathrm{d}Q_{Y,\mathbf{Z}}(y,\mathbf{z}) \leq \exp(b + r_2)\,\zeta_\Lambda.$$

For the second integral on the right hand side of (26), we can show that

$$\begin{aligned} & \int_{[0,\tau_0]\times\mathcal{Z}} \Lambda(y) \left|\exp\left(\boldsymbol{\beta}_j^T\mathbf{z} - \zeta_{\boldsymbol{\beta}} r_2\right) - \exp\left(\boldsymbol{\beta}_j^T\mathbf{z} + \zeta_{\boldsymbol{\beta}} r_2\right)\right| \mathrm{d}Q_{Y,\mathbf{Z}}(y,\mathbf{z}) \\ & \leq \bar{q}_2 M \int_a^b \exp(t + \zeta_{\boldsymbol{\beta}} r_2) - \exp(t - \zeta_{\boldsymbol{\beta}} r_2)\,\mathrm{d}t \\ & = \bar{q}_2 M \left(\int_{b-\zeta_{\boldsymbol{\beta}} r_2}^{b+\zeta_{\boldsymbol{\beta}} r_2} \exp(u)\,\mathrm{d}u - \int_{a-\zeta_{\boldsymbol{\beta}} r_2}^{a+\zeta_{\boldsymbol{\beta}} r_2} \exp(u)\,\mathrm{d}u \right) \\ & \leq 2M\bar{q}_2 \left[\exp(b + r_2) - \exp(a - r_2)\right] r_2 \zeta_{\boldsymbol{\beta}} \\ & \leq 2M\bar{q}_2 r_2 \exp(b + r_2)\,\zeta_{\boldsymbol{\beta}}, \end{aligned}$$

where M denotes the upper bound of the class of functions $\tilde{\mathcal{D}}$. If we take $\zeta_\Lambda = \frac{\zeta}{3\exp(b+r_2)}$ and $\zeta_{\boldsymbol{\beta}} = \frac{\zeta}{6M\bar{q}_2 r_2 \exp(b+r_2)}$, we obtain

$$\|f_l^U - f_l^L\|_{1,\mathbb{P}_0} \leq \zeta.$$

Hence, using $\log x \leq x-1$, for $x>0$, we have

$$\begin{aligned} H_B\left(\zeta, \mathcal{F}, \|\cdot\|_{1,\mathbb{P}_0}\right) &\leq \log N_1 + \log N_2 \\ &\leq q \log\left(\frac{6A_1 M \bar{q}_2 r_2 \exp\left(b+r_2\right)}{\zeta}\right) + \frac{3A_2 \exp\left(b+r_2\right)}{\zeta} \\ &\leq \frac{C_1\left(q+1\right)}{\zeta}. \end{aligned}$$

Next we construct brackets for the class of functions $\mathcal{G}_{\epsilon'}$ to show the second assertion of the Lemma. Let $\zeta_{\boldsymbol{\gamma}}>0$. By Lemma 7.5 of [3], $\mathcal{S}_{d-1}$ can be covered by N_3 neighborhoods with diameter at most $\zeta_{\boldsymbol{\gamma}}$, where $N_3 \leq \left(A_3/\zeta_{\boldsymbol{\gamma}}\right)^d$ with a constant $A_3>0$. Let $\left\{\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{N_3}\right\}$ be elements of each these neighborhoods. Let $\zeta_\varphi>0$ and consider ζ_φ–brackets $[\varphi_k^L, \varphi_k^U]$, $k=1, \cdots, N_4$, covering the class $\mathcal{M}_{\epsilon'}$ such that

$$\int_{\mathbb{R}} \left|\varphi_k^U(t) - \varphi_k^L(t)\right| \mathrm{d}Q_i^{\pm}(t) \leq \zeta_\varphi, \quad k=1, \cdots, N_4,$$

where $Q_i^{\pm}(t)$ denotes the distribution of $\boldsymbol{\gamma}_i^T\mathbf{X} \pm \zeta_{\boldsymbol{\gamma}} r_1$, $i=1, \cdots, N_3$, and $N_4 \leq \exp\left(A_4/\zeta_\varphi\right)$ for some constant $A_4>0$, by Lemma A.1. We note that φ_k^L and φ_k^U can always be taken to be bounded below by ϵ' and bounded above by $1-\epsilon'$, respectively. Otherwise, we can take $\varphi_k^L \vee \epsilon'$ and $\varphi_k^U \wedge 1-\epsilon'$. Using the monotonicity of φ and the Cauchy–Schwarz inequality, we have, for all $\mathbf{x} \in \mathcal{X}$,

$$\varphi_k^L\left(\boldsymbol{\gamma}_i^T\mathbf{x} - \zeta_{\boldsymbol{\gamma}} r_1\right) \leq \varphi\left(\boldsymbol{\gamma}^T\mathbf{x}\right) \leq \varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x} + \zeta_{\boldsymbol{\gamma}} r_1\right),$$

for some $i=1, \cdots, N_3$ and $k=1, \cdots, N_4$. Using the result from the first statement of the Lemma, we can consider a ζ_F–bracket $[F_j^L, F_j^U]$, $j=1, \cdots, N$, covering the class $\mathcal{F}$ such that

$$\int \left|F_j^U(y|\mathbf{z}) - F_j^L(y|\mathbf{z})\right| \mathrm{d}\mathbb{P}_0(y, \delta, \mathbf{x}, \mathbf{z}) \leq \zeta_F, \quad j=1, \cdots, N,$$

where $N \leq \exp[C_1(q+1)/\zeta_F]$. Then, we have

$$\begin{aligned} \log\left[1-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x} + \zeta_{\boldsymbol{\gamma}} r_1\right) F_j^U(y|\mathbf{z})\right] &\leq \log\left[1-\varphi\left(\boldsymbol{\gamma}^T\mathbf{x}\right) F_u(y|\mathbf{z})\right] \\ &\leq \log\left[1-\varphi_k^L\left(\boldsymbol{\gamma}_i^T\mathbf{x} - \zeta_{\boldsymbol{\gamma}} r_1\right) F_j^L(y|\mathbf{z})\right], \end{aligned}$$

for some $j=1, \cdots, N$. We then show that, for a certain choice of ζ_F, the brackets

$$\begin{aligned} [g_{i,j,k}^L, g_{i,j,k}^U] = \big[&(1-\delta)\log\left\{1-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x} + \zeta_{\boldsymbol{\gamma}} r_1\right) F_j^U(y|\mathbf{z})\right\}, \\ &(1-\delta)\log\left\{1-\varphi_k^L\left(\boldsymbol{\gamma}_i^T\mathbf{x} - \zeta_{\boldsymbol{\gamma}} r_1\right) F_j^L(y|\mathbf{z})\right\}\big] \end{aligned}$$

are ζ–brackets with respect to $\|\cdot\|_{1,\mathbb{P}_0}$ for the class $\mathcal{G}_{\epsilon'}$. Using the Minkowski inequality, we have

$$\|g_{i,j,k}^U - g_{i,j,k}^L\|_{1,\mathbb{P}_0} \leq \int_{[0,\tau_0]\times\mathcal{X}\times\mathcal{Z}} \left|\log\left[1-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x} - \zeta_{\boldsymbol{\gamma}} r_1\right) F_j^L(y|\mathbf{z})\right]\right.$$

$$
\begin{aligned}
&\quad - \log\left[1-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}+\zeta_{\boldsymbol{\gamma}}r_1\right)F_j^U\left(y|\mathbf{z}\right)\right]\Big|\mathrm{d}Q_{Y,\mathbf{X},\mathbf{Z}}\left(y,\mathbf{x},\mathbf{z}\right)\\
&+\int_{[0,\tau_0]\times\mathcal{X}\times\mathcal{Z}}\Big|\log\left[1-\varphi_k^L\left(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1\right)F_j^L\left(y|\mathbf{z}\right)\right]\\
&\quad - \log\left[1-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1\right)F_j^L\left(y|\mathbf{z}\right)\right]\Big|\mathrm{d}Q_{Y,\mathbf{X},\mathbf{Z}}\left(y,\mathbf{x},\mathbf{z}\right).
\end{aligned}
\tag{27}
$$

For the second integral in (27) we have

$$
\begin{aligned}
&\int_{[0,\tau_0]\times\mathcal{X}\times\mathcal{Z}}\Big|\log\left[1-\varphi_k^L(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1)F_j^L(y|\mathbf{z})\right]\\
&\qquad\qquad - \log\left[1-\varphi_k^U(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1)F_j^L(y|\mathbf{z})\right]\Big|\mathrm{d}Q_{Y,\mathbf{X},\mathbf{Z}}\left(y,\mathbf{x},\mathbf{z}\right)\\
&\leq\frac{1}{\epsilon'}\int_{[0,\tau_0]\times\mathcal{X}\times\mathcal{Z}}\left|\varphi_k^U(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1)-\varphi_k^L(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1)\right|F_j^L(y|\mathbf{z})\mathrm{d}Q_{Y,\mathbf{X},\mathbf{Z}}\left(y,\mathbf{x},\mathbf{z}\right)\\
&\leq\frac{1}{\epsilon'}\int_{\mathbb{R}}\left|\varphi_k^U\left(t\right)-\varphi_k^L\left(t\right)\right|\mathrm{d}Q_i^-\left(t\right)\leq\frac{\zeta_\varphi}{\epsilon'},
\end{aligned}
$$

where we have used the mean value theorem, $0\leq F_j^L\leq 1$ and $\epsilon'\leq\varphi_k^L,\varphi_k^U\leq 1-\epsilon'$. For the integrand of the first integral in (27), we have

$$
\begin{aligned}
&\left|\log\left[1-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1\right)F_j^L\left(y|\mathbf{z}\right)\right]-\log\left[1-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}+\zeta_{\boldsymbol{\gamma}}r_1\right)F_j^U\left(y|\mathbf{z}\right)\right]\right|\\
&\leq\frac{1}{\epsilon'}\left|\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}+\zeta_{\boldsymbol{\gamma}}r_1\right)F_j^U\left(y|\mathbf{z}\right)-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1\right)F_j^L\left(y|\mathbf{z}\right)\right|\\
&\leq\frac{1}{\epsilon'}\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}+\zeta_{\boldsymbol{\gamma}}r_1\right)\left|F_j^U\left(y|\mathbf{z}\right)-F_j^L\left(y|\mathbf{z}\right)\right|\\
&\quad+\frac{1}{\epsilon'}F_j^L\left(y|\mathbf{z}\right)\left|\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}+\zeta_{\boldsymbol{\gamma}}r_1\right)-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1\right)\right|\\
&\leq\frac{1}{\epsilon'}\left|F_j^U\left(y|\mathbf{z}\right)-F_j^L\left(y|\mathbf{z}\right)\right|+\frac{1}{\epsilon'}\left|\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}+\zeta_{\boldsymbol{\gamma}}r_1\right)-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1\right)\right|.
\end{aligned}
$$

Note that, by the monotonicity of φ_k^U, we have

$$
\begin{aligned}
&\int_{\mathcal{X}}\left|\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}+\zeta_{\boldsymbol{\gamma}}r_1\right)-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1\right)\right|\mathrm{d}Q_{\mathbf{X}}\left(\mathbf{x}\right)\\
&=\int_{\mathcal{X}}\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}+\zeta_{\boldsymbol{\gamma}}r_1\right)-\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x}-\zeta_{\boldsymbol{\gamma}}r_1\right)\mathrm{d}Q_{\mathbf{X}}\left(\mathbf{x}\right)\\
&\leq\bar{q}_1\int_{-r_1}^{r_1}\left[\varphi_k^U\left(t+\zeta_{\boldsymbol{\gamma}}r_1\right)-\varphi_k^U\left(t-\zeta_{\boldsymbol{\gamma}}r_1\right)\right]\mathrm{d}t\\
&=\bar{q}_1\left(\int_{r_1-\zeta_{\boldsymbol{\gamma}}r_1}^{r_1+\zeta_{\boldsymbol{\gamma}}r_1}\varphi_k^U\left(u\right)\mathrm{d}u-\int_{-r_1-\zeta_{\boldsymbol{\gamma}}r_1}^{-r_1+\zeta_{\boldsymbol{\gamma}}r_1}\varphi_k^U\left(u\right)\mathrm{d}u\right)\\
&=2r_1\bar{q}_1\left(1-2\epsilon'\right)\zeta_{\boldsymbol{\gamma}}.
\end{aligned}
\tag{28}
$$

Hence, the first integral in (27) can be bounded by

$$\frac{1}{\epsilon'}\left[\int \left|F_j^U(y|\mathbf{z}) - F_j^L(y|\mathbf{z})\right| \mathrm{d}\mathbb{P}_0\left(y,\delta,\mathbf{x},\mathbf{z}\right) + 2r_1\bar{q}_1\left(1-2\epsilon'\right)\zeta_{\boldsymbol{\gamma}}\right] \leq \frac{\zeta_F + 2r_1\bar{q}_1\zeta_{\boldsymbol{\gamma}}}{\epsilon'}.$$

Choosing $\zeta_\varphi = \epsilon'\zeta/3$, $\zeta_F = \epsilon'\zeta/3$, $\zeta_{\boldsymbol{\gamma}} = \epsilon'\zeta/6r_1\bar{q}_1$, we get $\|g^U_{i,j,k} - g^L_{i,j,k}\|_{1,\mathbb{P}_0} \leq \zeta$. Hence, using $\log x \leq x$, we have

$$\begin{aligned} H_B\left(\zeta, \mathcal{G}_{\epsilon'}, \|\cdot\|_{1,\mathbb{P}_0}\right) &\leq \log N + \log N_3 + \log N_4 \\ &\leq \frac{C_1\left(q+1\right)}{\zeta_F} + d\log\left(\frac{A_3}{\zeta_{\boldsymbol{\gamma}}}\right) + \frac{A_4}{\zeta_\varphi} \\ &\leq \frac{C_2 d + C_3(q+1)}{\zeta}, \end{aligned}$$

for some positive constants C_2, C_3.

To prove the third statement of the Lemma, we can use a similar argument as for the second statement and show that the brackets $[\log\varphi_k^L\left(\boldsymbol{\gamma}_i^T\mathbf{x} - \zeta_{\boldsymbol{\gamma}} r_1\right), \log\varphi_k^U\left(\boldsymbol{\gamma}_i^T\mathbf{x} + \zeta_{\boldsymbol{\gamma}} r_1\right)]$ for $k \in \{1,\ldots,N_4\}$, $i \in \{1,\ldots,N_3\}$ are ζ–brackets with respect to $\|\cdot\|_{1,\mathbb{P}_0}$ if we choose $\zeta_\varphi = \epsilon'\zeta/2$ and $\zeta_{\boldsymbol{\gamma}} = \epsilon'\zeta/4r_1\bar{q}_1$. Hence

$$H_B\left(\zeta, \mathcal{H}_{\epsilon'}, \|\cdot\|_{1,\mathbb{P}_0}\right) \leq \log N_3 + \log N_4 \leq d\log\left(\frac{A_3}{\zeta_{\boldsymbol{\gamma}}}\right) + \frac{A_4}{\zeta_\varphi} \leq \frac{C_4\left(d+1\right)}{\zeta}.$$

□

Lemma A.3. *Suppose that Assumptions (A4) and (A10) are satisfied. Let* $\zeta > 0$. *There exists a constant* $A_1 > 0$ *depending on* $\epsilon', \bar{q}_1$, *and* r_1 *and a constant* $A_2 > 0$ *depending on* $\epsilon', \bar{q}_1, \bar{q}_2, r_1$, *and* r_2, *such that*

$$H_B\left(\zeta, \mathcal{L}_{\epsilon'}, \|\cdot\|_{1,\mathbb{P}_0}\right) \leq \frac{A_1\left(d+1\right) + A_2\left(q+1\right)}{\zeta}.$$

Moreover, there exists a constant $A_3 > 0$ *depending on* $\epsilon', \bar{q}_1$, *and* r_1 *and a constant* $A_4 > 0$ *depending on* $\epsilon', \bar{q}_1, \bar{q}_2, r_1$, *and* r_2, *such that*

$$H_B\left(\zeta, \tilde{\mathcal{L}}_{\epsilon'}, \|\cdot\|_{1,\mathbb{P}_0}\right) \leq \frac{A_3\left(d+1\right) + A_4\left(q+1\right)}{\zeta}.$$

Proof. Let $\zeta > 0$. Using the last two statements of Lemma A.2, we can consider $\zeta/2$–brackets $[h_i^L, h_i^U]$, $i = 1,\cdots,N_1$, covering the class $\mathcal{H}_{\epsilon'}$ where $N_1 \leq \exp\left(\frac{2C_4(d+1)}{\zeta}\right)$. Also, consider $\zeta/2$–brackets $[g_j^L, g_j^U]$, $j = 1,\cdots,N_2$, covering the class $\mathcal{G}_{\epsilon'}$, where $N_2 \leq \exp\left(\frac{2C_2 d + 2C_3(q+1)}{\zeta}\right)$. Then, for any $l \in \mathcal{L}_{\epsilon'}$, we have

$$h_i^L\left(\delta,\mathbf{x}\right) + g_j^L\left(y,\delta,\mathbf{x},\mathbf{z}\right) \leq l\left(y,\delta,\mathbf{x},\mathbf{z}\right) \leq h_i^U\left(\delta,\mathbf{x}\right) + g_j^U\left(y,\delta,\mathbf{x},\mathbf{z}\right),$$

for some $i = 1, \cdots N_1$ and $j = 1, \cdots, N_2$. Using Minkowski inequality, we obtain

$$\begin{aligned}
&\int \left|h_i^U(\delta, \mathbf{x}) + g_j^U(y, \delta, \mathbf{x}, \mathbf{z}) - h_i^L(\delta, \mathbf{x}) - g_j^L(y, \delta, \mathbf{x}, \mathbf{z})\right| \mathrm{d}\mathbb{P}_0(y, \delta, \mathbf{x}, \mathbf{z}) \\
&\leq \int \left|h_i^U(\delta, \mathbf{x}) - h_i^L(\delta, \mathbf{x})\right| \mathrm{d}\mathbb{P}_0(y, \delta, \mathbf{x}, \mathbf{z}) \\
&\quad + \int \left|g_j^U(y, \delta, \mathbf{x}, \mathbf{z}) - g_j^L(y, \delta, \mathbf{x}, \mathbf{z})\right| \mathrm{d}\mathbb{P}_0(y, \delta, \mathbf{x}, \mathbf{z}) \\
&\leq \zeta/2 + \zeta/2 = \zeta.
\end{aligned}$$

Hence, the ζ–brackets $[h_i^U + g_j^U, h_i^L + g_j^L]$, $i = 1, \ldots, N_1$, $j = 1, \ldots, , N_2$, cover $\mathcal{L}_{\epsilon'}$, and

$$H_B\left(\zeta, \mathcal{L}_{\epsilon'}, \|\cdot\|_{1,\mathbb{P}_0}\right) \leq \log N_1 + \log N_2 \leq \frac{A_1(d+1) + A_2(q+1)}{\zeta}.$$

The second statement of the Lemma can be shown with the same type of argument. □

Lemma A.4. *Let Q be any probability measures on $\mathbb{R}$ and denote by $\|\cdot\|_{r,Q}$ the L_r–norm corresponding to Q. Let $\tilde{\mathcal{M}}_{\epsilon'}$ be the class of continuous monotone functions on $\mathbb{R}$ with values in $[\epsilon', 1-\epsilon']$. For any $\zeta > 0$, there exist $M \leq \exp\left(\frac{A}{\zeta}\right)$ balls of radius of ζ with respect to the L_r–norm and centered in $\tilde{\mathcal{M}}_{\epsilon'}$ such that their union covers $\tilde{\mathcal{M}}_{\epsilon'}$. The constant $A > 0$ does not depend on ζ.*

Proof of Lemma A.4. By Lemma A.1 and Lemma 2.1 in [36], $\mathcal{M}_{\epsilon'}$ can be covered by N balls of radius of $\zeta/2$, where $N \leq \exp(A/\zeta)$ for some constant $A > 0$. Let $\{\varphi_1, \cdots, \varphi_N\}$ be the centers of each of these balls which are not necessarily elements of $\tilde{\mathcal{M}}_{\epsilon'}$. Next we construct new balls whose union covers $\tilde{\mathcal{M}}_{\epsilon'}$. For each $i \in \{1, \cdots, N\}$, if there is a continuous monotone function $\tilde{\varphi}_i$ in the ball $B_{\varphi_i}(\zeta/2)$, we define a new ball, $B_{\tilde{\varphi}_i}(\zeta)$, centered at $\tilde{\varphi}_i$ with a radius of ζ. Otherwise, if the ball $B_{\varphi_i}(\zeta/2)$ does not contain any continuous monotone function we eliminate it. Thus we have at most $M \leq N \leq \exp(A/\zeta)$ of such balls with radius ζ. Next we show that they cover $\tilde{\mathcal{M}}_{\epsilon'}$.

Consider $\varphi \in \tilde{\mathcal{M}}_{\epsilon'} \subset \mathcal{M}_{\epsilon'}$. We can find $i \in \{1, \cdots, N\}$ such that $\|\varphi - \varphi_i\|_{r,Q} \leq \zeta/2$. Let $\tilde{\varphi}_j$ be center of the new ball constructed using the preceding argument that corresponds to $B_{\varphi_i}(\zeta/2)$. We have $\|\varphi - \tilde{\varphi}_j\|_{r,Q} \leq \|\varphi - \varphi_i\|_{r,Q} + \|\varphi_i - \tilde{\varphi}_j\|_{r,Q} \leq \zeta$. Hence $\varphi \in B_{\tilde{\varphi}_j}(\zeta)$ for some $j \in \{1, \cdots M\}$. We conclude that $\tilde{\mathcal{M}}_{\epsilon'} \subset \bigcup_{j=1}^{M} B_{\tilde{\varphi}_j}(\zeta)$. □

Proof of Proposition 4.5. Recall that for fixed $\boldsymbol{\theta}$, $\varphi_{0,\boldsymbol{\theta}} = \arg\max_{\varphi \in \mathcal{M}_{\epsilon'}} l_{\boldsymbol{\theta}}(\varphi)$, where

$$l_{\boldsymbol{\theta}}(\varphi) = \mathbb{E}\left[l(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}, \varphi)\right] = \int l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi) \, \mathrm{d}\mathbb{P}_0(y, \delta, \mathbf{x}, \mathbf{z}),$$

and $\mathbb{P}_0$ denotes the distribution of $(Y, \Delta, \mathbf{X}, \mathbf{Z})$. On the other hand, for fixed $\boldsymbol{\theta}$, we have defined $\hat{\varphi}_{n,\boldsymbol{\theta}} = \arg\max_{\varphi \in \mathcal{M}_{\epsilon'}} l_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda, \varphi)$, where

$$l_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda, \varphi) = \int l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi) \, \mathrm{d}\mathbb{P}_n(y, \delta, \mathbf{x}, \mathbf{z}),$$

and $\mathbb{P}_n$ denotes the empirical distribution of the $(y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$. We first show that

$$\begin{aligned} &\int l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \hat{\varphi}_{n,\boldsymbol{\theta}}) - l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi_{0,\boldsymbol{\theta}}) \, \mathrm{d}(\mathbb{P}_n - \mathbb{P}_0)(y, \delta, \mathbf{x}, \mathbf{z}) \\ &\quad \geq \tilde{c} \int_{\mathcal{X}} \left\{ \hat{\varphi}_{n,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T \mathbf{x}) - \varphi_{0,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T \mathbf{x}) \right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) = \tilde{c} \| \hat{\varphi}_{n,\boldsymbol{\theta}} - \varphi_{0,\boldsymbol{\theta}} \|_{Q_{\mathbf{X}},\boldsymbol{\gamma}}, \end{aligned} \tag{29}$$

for some constant $\tilde{c} > 0$. Then it suffices to show that

$$\sup_{\boldsymbol{\theta} \in \tilde{\Theta}} \int l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \hat{\varphi}_{n,\boldsymbol{\theta}}) - l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi_{0,\boldsymbol{\theta}}) \, \mathrm{d}(\mathbb{P}_n - \mathbb{P}_0)(y, \delta, \mathbf{x}, \mathbf{z}) \tag{30}$$

converges almost surely to zero.

For fixed $\boldsymbol{\theta} \in \tilde{\Theta}$ and $\varphi \in \mathcal{M}_{\epsilon'}$, define $H(t) = \mathbb{E}\left[l_{\boldsymbol{\theta}}(t\varphi + (1-t)\varphi_{0,\boldsymbol{\theta}}) \right]$, where $0 < t < 1$. We have

$$\begin{aligned} &\int l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi) - l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi_{0,\boldsymbol{\theta}}) \, \mathrm{d}\mathbb{P}_0(y, \delta, \mathbf{x}, \mathbf{z}) \\ &= H(1) - H(0) = H'(0) + \frac{1}{2} H''(t^*) \end{aligned}$$

for some $t^* \in (0, 1)$. By the definition of $\varphi_{0,\boldsymbol{\theta}}$ and the convexity of $\mathcal{M}_{\epsilon'}$, $H'(0) \leq 0$ for all $\varphi \in \mathcal{M}_{\epsilon'}$. From (19) and (21), we have

$$\begin{aligned} H''(t) &\leq -\mathbb{E}\left[\varphi_0(\boldsymbol{\gamma}_0^T \mathbf{X}) F_{u,0} \left\{ 1 - \varphi_0(\boldsymbol{\gamma}_0^T \mathbf{X}) F_{u,0} \right\} \left\{ \varphi(\boldsymbol{\gamma}^T \mathbf{X}) - \varphi_{0,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T \mathbf{X}) \right\}^2 \right] \\ &\leq -c\epsilon^2 \int_{\mathcal{X}} \left\{ \varphi(\boldsymbol{\gamma}^T \mathbf{x}) - \varphi_{0,\boldsymbol{\theta}}(\boldsymbol{\gamma}^T \mathbf{x}) \right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}). \end{aligned}$$

Then,

$$\int l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi) - l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi_{0,\boldsymbol{\theta}}) \, \mathrm{d}\mathbb{P}_0(y, \delta, \mathbf{x}, \mathbf{z}) \leq -\tilde{c} \| \hat{\varphi}_{n,\boldsymbol{\theta}} - \varphi_{0,\boldsymbol{\theta}} \|_{Q_{\mathbf{X}},\boldsymbol{\gamma}},$$

where $\tilde{c} = \frac{1}{2} c\epsilon^2 > 0$. By the definition of $\hat{\varphi}_{n,\boldsymbol{\theta}}$, we have

$$\int l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \hat{\varphi}_{n,\boldsymbol{\theta}}) - l(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi_{0,\boldsymbol{\theta}}) \, \mathrm{d}\mathbb{P}_n(y, \delta, \mathbf{x}, \mathbf{z}) \geq 0.$$

Combining the two inequalities above and the fact that $\hat{\varphi}_{n,\boldsymbol{\theta}} \in \mathcal{M}_{\epsilon'}$ gives (29).

To obtain the almost sure convergence of the expression in (30), we consider the class $\tilde{\mathcal{L}}_{\epsilon'}$ of functions of the form

$$\begin{aligned} \tilde{l}(y, \delta, \mathbf{x}, \mathbf{z}) = \delta &\left[\log \varphi_1(\boldsymbol{\gamma}^T \mathbf{x}) - \log \varphi_2(\boldsymbol{\gamma}^T \mathbf{x}) \right] \\ &+ (1 - \delta) \left\{ \log \left[1 - \varphi_1(\boldsymbol{\gamma}^T \mathbf{x}) F_u(y|\mathbf{z}) \right] - \log \left[1 - \varphi_2(\boldsymbol{\gamma}^T \mathbf{x}) F_u(y|\mathbf{z}) \right] \right\}, \end{aligned}$$

for some $\varphi_1, \varphi_2 \in \mathcal{M}_{\epsilon'}$, $F_u(y|\mathbf{z}) = 1 - \exp\{-\Lambda(y)\exp(\boldsymbol{\beta}^T\mathbf{z})\}$ and $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda) \in \tilde{\Theta}$. With a series of Lemmas in Appendix we show that $\tilde{\mathcal{L}}_{\epsilon'}$ is a Glivenko–Cantelli class of functions (see Lemma A.3). Note also that the class is uniformly bounded since for any $\tilde{l} \in \tilde{\mathcal{L}}_{\epsilon'}$, $|\tilde{l}| \leq -\log \epsilon' < \infty$. As a result,

$$\sup_{\tilde{l} \in \tilde{\mathcal{L}}_{\epsilon'}} \left| \int \tilde{l}\,(y, \delta, \mathbf{x}, \mathbf{z})\, \mathrm{d}\,(\mathbb{P}_n - \mathbb{P}_0)\,(y, \delta, \mathbf{x}, \mathbf{z}) \right| \to 0 \quad a.s. \text{ as } n \to \infty.$$

Since $\hat{\varphi}_{n,\boldsymbol{\theta}}, \varphi_{0,\boldsymbol{\theta}} \in \mathcal{M}_{\epsilon'}$, we complete the proof by the following inequality,

$$\begin{aligned} &\sup_{\boldsymbol{\theta} \in \tilde{\Theta}} \int l\,(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \hat{\varphi}_{n,\boldsymbol{\theta}}) - l\,(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi_{0,\boldsymbol{\theta}})\, \mathrm{d}\,(\mathbb{P}_n - \mathbb{P}_0)\,(y, \delta, \mathbf{x}, \mathbf{z}) \\ &\quad \leq \sup_{\tilde{l} \in \tilde{\mathcal{L}}_{\epsilon'}} \left| \int \tilde{l}\,(y, \delta, \mathbf{x}, \mathbf{z})\, \mathrm{d}\,(\mathbb{P}_n - \mathbb{P}_0)\,(y, \delta, \mathbf{x}, \mathbf{z}) \right| \to 0 \quad a.s. \text{ as } n \to \infty. \end{aligned}$$

□

Proof of Proposition 4.6. For fixed $\boldsymbol{\theta} \in \tilde{\Theta}$ and $\mathbf{x} \in \mathcal{X}$, using the properties of the kernel density k and a change of variable, we have

$$\begin{aligned} \hat{\varphi}^s_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right) &= \int_{-1}^{1} k\,(u)\left[\hat{\varphi}_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hu\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hu\right)\right] \mathrm{d}u \\ &\quad + \int_{-1}^{1} k\,(v)\left[\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right] \mathrm{d}v. \end{aligned}$$

Then, we have

$$\begin{aligned} &\left\{\int_{\mathcal{X}} \left[\hat{\varphi}^s_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right]^2 \mathrm{d}Q_{\mathbf{X}}\,(\mathbf{x})\right\}^{\frac{1}{2}} \\ &\leq \left\{\int_{\mathcal{X}} \left[\int_{-1}^{1} k\,(u)\left\{\hat{\varphi}_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hu\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hu\right)\right\} \mathrm{d}u\right]^2 \mathrm{d}Q_{\mathbf{X}}\,(\mathbf{x})\right\}^{\frac{1}{2}} \\ &+ \left\{\int_{\mathcal{X}} \left[\int_{-1}^{1} k\,(v)\left\{\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\} \mathrm{d}v\right]^2 \mathrm{d}Q_{\mathbf{X}}\,(\mathbf{x})\right\}^{\frac{1}{2}} = (I) + (II). \end{aligned} \tag{31}$$

We proceed by showing that (I) and (II) converge to zero uniformly on $\tilde{\Theta}$ with probability one.

Consider the inner integral of (I) in (31) and using Jensen's inequality gives

$$\begin{aligned} &\left[\int_{-1}^{1} k\,(u)\left\{\hat{\varphi}_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hu\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hu\right)\right\} \mathrm{d}u\right]^2 \\ &\leq K^2 \int_{-1}^{1} \left\{\hat{\varphi}_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hu\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hu\right)\right\}^2 \mathrm{d}u, \end{aligned}$$

where K is the upper bound of the kernel function k. By Assumption (A1)(v) and for $\boldsymbol{\gamma}$ in a neighborhood of $\boldsymbol{\gamma}_0$ we have $\mathcal{I}_{\boldsymbol{\gamma}} = [\underline{I}_{\boldsymbol{\gamma}}, \bar{I}_{\boldsymbol{\gamma}}]$. Then, we can write

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\tilde{\Theta}} (I)^2 &\le K^2 \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\mathcal{X}} \int_{-1}^{1} \left\{\hat{\varphi}_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hu\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hu\right)\right\}^2 \mathrm{d}u \, \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) \\
&= K^2 \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\mathcal{I}_{\boldsymbol{\gamma}}} \int_{-1}^{1} \left\{\hat{\varphi}_{n,\boldsymbol{\theta}}(v - hu) - \varphi_{0,\boldsymbol{\theta}}(v - hu)\right\}^2 \mathrm{d}u \, \mathrm{d}G_{\boldsymbol{\gamma}^T\mathbf{X}}(v) \\
&= K^2 \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\mathcal{I}_{\boldsymbol{\gamma}}} \frac{1}{h} \int_{v-h}^{v+h} \left\{\hat{\varphi}_{n,\boldsymbol{\theta}}(s) - \varphi_{0,\boldsymbol{\theta}}(s)\right\}^2 \mathrm{d}s \, \mathrm{d}G_{\boldsymbol{\gamma}^T\mathbf{X}}(v) \\
&= K^2 \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\underline{I}_{\boldsymbol{\gamma}}-h}^{\bar{I}_{\boldsymbol{\gamma}}+h} \left\{\hat{\varphi}_{n,\boldsymbol{\theta}}(s) - \varphi_{0,\boldsymbol{\theta}}(s)\right\}^2 \frac{1}{h} \int_{s-h}^{s+h} \mathrm{d}G_{\boldsymbol{\gamma}^T\mathbf{X}}(v) \, \mathrm{d}s,
\end{aligned} \tag{32}$$

where $G_{\boldsymbol{\gamma}^T\mathbf{X}}$ denotes the distribution function of $\boldsymbol{\gamma}^T\mathbf{X}$. Recall that $h = h_n \to 0$ as $n \to \infty$. For $s \in [\underline{I}_{\boldsymbol{\gamma}} + h, \bar{I}_{\boldsymbol{\gamma}} - h]$, since the density function $g_{\boldsymbol{\gamma}^T\mathbf{X}}$ is continuous, by the mean value theorem, there exists $\xi_n \in (s - h_n, s + h_n)$ such that

$$\frac{1}{h} \int_{s-h}^{s+h} \mathrm{d}G_{\boldsymbol{\gamma}^T\mathbf{X}}(v) = \frac{1}{h}\left\{G_{\boldsymbol{\gamma}^T\mathbf{X}}(s+h) - G_{\boldsymbol{\gamma}^T\mathbf{X}}(s-h)\right\} = 2g_{\boldsymbol{\gamma}^T\mathbf{X}}(\xi_n). \tag{33}$$

By Assumption (A9) it follows that

$$\sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \sup_{s\in[\underline{I}_{\boldsymbol{\gamma}}+h, \bar{I}_{\boldsymbol{\gamma}}-h]} \left| \frac{1}{h} \int_{s-h}^{s+h} \mathrm{d}G_{\boldsymbol{\gamma}^T\mathbf{X}}(v) - 2g_{\boldsymbol{\gamma}^T\mathbf{X}}(s) \right| \to 0 \quad \text{as } n \to \infty. \tag{34}$$

Since the length $\mathcal{I}_{\boldsymbol{\gamma}}$ is uniformly bounded, $\{\hat{\varphi}_{n,\boldsymbol{\theta}}(s) - \varphi_{0,\boldsymbol{\theta}}(s)\}^2 \le (1 - 2\epsilon')^2$, from (32), (34) and Proposition 4.5 we obtain

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\tilde{\Theta}} (I)^2 \le \tilde{K}h + 2K^2 \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\mathcal{I}_{\boldsymbol{\gamma}}} \left\{\hat{\varphi}_{n,\boldsymbol{\theta}}(s) - \varphi_{0,\boldsymbol{\theta}}(s)\right\}^2 \mathrm{d}G_{\boldsymbol{\gamma}^T\mathbf{X}}(s) \\
&+ K^2 \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\underline{I}_{\boldsymbol{\gamma}}+h}^{\bar{I}_{\boldsymbol{\gamma}}-h} \{\hat{\varphi}_{n,\boldsymbol{\theta}}(s) - \varphi_{0,\boldsymbol{\theta}}(s)\}^2 \left| \int_{s-h}^{s+h} \frac{1}{h} \mathrm{d}G_{\boldsymbol{\gamma}^T\mathbf{X}}(v) - 2g_{\boldsymbol{\gamma}^T\mathbf{X}}(s) \right| \mathrm{d}s \xrightarrow{a.s.} 0,
\end{aligned} \tag{35}$$

where the term $\tilde{K}h$ comes from dealing with the boundary regions of the integrals and the constant $\tilde{K}$ depends on the uniform bound of $g_{\boldsymbol{\gamma}^T\mathbf{X}}(\cdot)$. This means that the first term (I) in (31) converges to zero uniformly on $\tilde{\Theta}$ with probability one.

Next, we deal with (II) in (31), which is a deterministic term. With a similar argument as above, one can show that

$$\sup_{\boldsymbol{\theta}\in\tilde{\Theta}} (II)^2 \le K^2 \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\mathcal{X}} \int_{-1}^{1} \left\{\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}v \, \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}). \tag{36}$$

For a fixed $\boldsymbol{\theta} \in \tilde{\Theta}$, the inner integral in the previous display converges to zero since $\varphi_{0,\boldsymbol{\theta}}$ is continuous and $h \to 0$. However, this does not hold uniformly on $\tilde{\Theta}$. To circumvent the non-uniformity issue, we construct a covering of the space of bounded continuous monotone functions by a finite number of balls and approximate $\varphi_{0,\boldsymbol{\theta}}$ by one of the centers of such balls. Let $\zeta > 0$ and let $[\tilde{A}, \tilde{B}]$ be an interval that contains all the intervals $[\underline{I}_{\boldsymbol{\gamma}} - h, \bar{I}_{\gamma} + h]$. Denote by Q the uniform distribution on $[\tilde{A}, \tilde{B}]$. By Lemma A.4, where $M \leq \exp[A/r]$ with a constant $A > 0$. For a fixed $\boldsymbol{\theta} \in \tilde{\Theta}$, using the Minkowski inequality, we have

$$
\begin{aligned}
&\left\{\int_{\mathcal{X}} \int_{-1}^{1} \left\{\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}v \, \mathrm{d}Q_{\mathbf{X}}(\mathbf{x})\right\}^{\frac{1}{2}} \\
&\leq \left\{\int_{\mathcal{X}} \int_{-1}^{1} \left\{\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right) - \tilde{\varphi}_{i(\boldsymbol{\theta})}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right)\right\}^2 \mathrm{d}v \, \mathrm{d}Q_{\mathbf{X}}(\mathbf{x})\right\}^{\frac{1}{2}} \\
&\quad + \left\{\int_{\mathcal{X}} \int_{-1}^{1} \left\{\tilde{\varphi}_{i(\boldsymbol{\theta})}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right) - \tilde{\varphi}_{i(\boldsymbol{\theta})}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}v \, \mathrm{d}Q_{\mathbf{X}}(\mathbf{x})\right\}^{\frac{1}{2}} \\
&\quad + \left\{2\int_{\mathcal{X}} \left\{\tilde{\varphi}_{i(\boldsymbol{\theta})}\left(\boldsymbol{\gamma}^T\mathbf{x}\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x})\right\}^{\frac{1}{2}} = (III) + (IV) + (V).
\end{aligned}
\tag{37}
$$

We will show that there exists N independent of $\boldsymbol{\theta}$ such that each of these three terms are smaller than $\zeta/3$ for $n > N$ and any $\boldsymbol{\theta} \in \tilde{\Theta}$. Since this holds for any $\zeta > 0$, we can then conclude that (II) converges to zero uniformly in $\boldsymbol{\theta}$. Consider the square of (III) in (37). As in (32)–(35), we obtain

$$
\begin{aligned}
&\int_{\mathcal{X}} \int_{-1}^{1} \left\{\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right) - \tilde{\varphi}_{i(\boldsymbol{\theta})}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right)\right\}^2 \mathrm{d}v \, \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) \\
&\leq c\left\{h + \int_{I_{\boldsymbol{\gamma}}} \left\{\varphi_{0,\boldsymbol{\theta}}(s) - \tilde{\varphi}_{i(\boldsymbol{\theta})}(s)\right\}^2 \mathrm{d}G_{\boldsymbol{\gamma}^T\mathbf{X}}(s)\right. \\
&\quad \left. + \sup_{s \in [\underline{I}_{\boldsymbol{\gamma}} + h, \bar{I}_{\boldsymbol{\gamma}} - h]} \left| \frac{1}{h} \int_{s-h}^{s+h} \mathrm{d}G_{\boldsymbol{\gamma}^T\mathbf{X}}(v) - 2g_{\boldsymbol{\gamma}^T\mathbf{X}}(s) \right| \right\} \\
&\leq c\left(h + \bar{q}_1(\tilde{B} - \tilde{A})r^2 + 2 \sup_{s \in [\underline{I}_{\boldsymbol{\gamma}} + h, \bar{I}_{\boldsymbol{\gamma}} - h]} \left| g_{\boldsymbol{\gamma}^T\mathbf{X}}(\xi_{n,s}) - g_{\boldsymbol{\gamma}^T\mathbf{X}}(s) \right| \right),
\end{aligned}
$$

where $s - h \leq \xi_{n,s} \leq s + h$, $\bar{q}_1$ denotes the uniform upper bound on the density functions $g_{\boldsymbol{\gamma}^T\mathbf{X}}(\cdot)$, and c is a positive constant. By uniform equicontinuity of the family of functions $\{g_{\boldsymbol{\gamma}^T\mathbf{X}}\}$ it follows that there exists N_1 such that,

$$
\sup_{s \in [\underline{I}_{\boldsymbol{\gamma}} + h, \bar{I}_{\boldsymbol{\gamma}} - h]} \left| g_{\boldsymbol{\gamma}^T\mathbf{X}}(\xi_{n,s}) - g_{\boldsymbol{\gamma}^T\mathbf{X}}(s) \right| \leq \frac{\zeta^2}{54c}
$$

for all $\boldsymbol{\theta} \in \tilde{\Theta}$, $h \leq \zeta^2/27c$ whenever $n \geq N_1$. If we choose $r = \zeta/\sqrt{27c\bar{q}_1(\tilde{B} - \tilde{A})}$ we obtain $(III)^2 \leq \zeta^2/9$ for all $\boldsymbol{\theta} \in \tilde{\Theta}$ and $n > N_1$. In the same way, for the

square of (V) in (37) we obtain, for $n > N_2$,

$$(V)^2 \leq 2 \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\mathcal{X}} \left\{\tilde{\varphi}_{i(\boldsymbol{\theta})}\left(\boldsymbol{\gamma}^T\mathbf{x}\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}Q_{\mathbf{X}}\left(\mathbf{x}\right) \leq \frac{\zeta^2}{9}. \tag{38}$$

For the square of (IV) in (37), we have

$$\begin{aligned} &\int_{\mathcal{X}} \int_{-1}^{1} \left\{\tilde{\varphi}_{i(\boldsymbol{\theta})}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right) - \tilde{\varphi}_{i(\boldsymbol{\theta})}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}v\, \mathrm{d}Q_{\mathbf{X}}\left(\mathbf{x}\right) \\ &\leq \bar{q}_1 \int_{\tilde{A}}^{\tilde{B}} \int_{-1}^{1} \left\{\tilde{\varphi}_{i(\boldsymbol{\theta})}\left(u - hv\right) - \tilde{\varphi}_{i(\boldsymbol{\theta})}\left(u\right)\right\}^2 \mathrm{d}v\, \mathrm{d}u. \end{aligned} \tag{39}$$

Since the functions $\tilde{\varphi}_{i(\boldsymbol{\theta})}$ are continuous on a compact, they are uniformly continuous and since we are dealing with a finite number M of such functions the family $\{\tilde{\varphi}_{i(\boldsymbol{\theta})} : i = 1, \ldots, M\}$ is uniformly equicontinuous. Therefore, since $h = h_n \to 0$ as $n \to \infty$, we can find an integer N_3 such that

$$|\tilde{\varphi}_{i(\boldsymbol{\theta})}\left(u - hv\right) - \tilde{\varphi}_{i(\boldsymbol{\theta})}\left(u\right)| \leq \zeta / \sqrt{18\bar{q}_1(\tilde{B} - \tilde{A})}$$

for all $u \in [\tilde{A}, \tilde{B}]$, $v \in [-1, 1]$ and $n \geq N_3$. Consequently, $(IV)^2 \leq \zeta^2/9$ for $n \geq N_3$ and any $\boldsymbol{\theta} \in \tilde{\Theta}$. Combining these results, we have for $n \geq \max\{N_1, N_2, N_3\}$,

$$\sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \int_{\mathcal{X}} \int_{-1}^{1} \left\{\varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x} - hv\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}v\, \mathrm{d}Q_{\mathbf{X}}\left(\mathbf{x}\right) \leq \zeta^2,$$

from which the uniform almost sure convergence of (II) follows. □

Proof of Theorem 4.7. This can be proved by applying Theorem 1 of [8], which boils down to verifying the required conditions. First we define some notation. We equip the space $\tilde{\Theta} = \Gamma \times \mathcal{B} \times \tilde{\mathcal{D}}$ with the metric d which is defined by

$$d\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) = \max\left\{\left\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\right\|_2, \left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right\|_2, \sup_{t\in[0,\tau_0]} \left|\Lambda_1\left(t\right) - \Lambda_2\left(t\right)\right|\right\},$$

where $\boldsymbol{\theta}_j = \left(\boldsymbol{\gamma}_j, \boldsymbol{\beta}_j, \Lambda_j\right) \in \tilde{\Theta}$, $j = 1, 2$. Let $\mathcal{H}$ denote the infinite-dimensional parameter space for the link, defined by $\mathcal{H} = \{h : \mathcal{X} \times \tilde{\Theta} \to [\epsilon', 1 - \epsilon'],\ h\left(\mathbf{x}, \boldsymbol{\theta}\right) = \varphi_{\boldsymbol{\theta}}(\boldsymbol{\gamma}^T\mathbf{x})$ for some $\varphi_{\boldsymbol{\theta}} \in \mathcal{M}_{\epsilon'}\}$. We equip $\mathcal{H}$ with the metric

$$d_{\mathcal{H}}\left(h_1, h_2\right) = \sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \left(\int_{\mathcal{X}} \left\{\varphi_{1,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right) - \varphi_{2,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}Q_{\mathbf{X}}\left(\mathbf{x}\right)\right)^{\frac{1}{2}},$$

where $h_j\left(\mathbf{x}, \boldsymbol{\theta}\right) = \varphi_{j,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)$ for some $\varphi_{j,\boldsymbol{\theta}} \in \mathcal{M}_{\epsilon'}, j = 1, 2$. Let $h_0\left(\mathbf{x}, \boldsymbol{\theta}\right) = \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)$, where $\varphi_{0,\boldsymbol{\theta}}$ is defined in (13). Note that, because of the Kullback–Leibler inequality,

$$\mathbb{E}\left[l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}, \varphi\right)\right] \leq \mathbb{E}\left[l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}_0, \varphi_0\right)\right]$$

for all $\boldsymbol{\theta} \in \tilde{\Theta}$, $\varphi \in \mathcal{M}_{\epsilon'}$ and, under the identifiability conditions, the equality holds only for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\varphi = \varphi_0$. By Proposition 4.2 it follows that $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\tilde{\Theta}} \mathbb{E}\left[l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}, \varphi_{0,\boldsymbol{\theta}}\right)\right]$. Then, the non-parametric estimator of h_0 is $\hat{h}(\mathbf{x}, \boldsymbol{\theta}) = \hat{\varphi}^s_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)$, where $\hat{\varphi}^s_{n,\boldsymbol{\theta}}$ is our smoothed monotone link estimator defined in (6). Consider i.i.d. realization $(y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \cdots, n$ of $(Y, \Delta, \mathbf{X}, \mathbf{Z})$. Let $h \in \mathcal{H}$ such that $h\left(\mathbf{x}, \boldsymbol{\theta}\right) = \varphi_{\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)$ for some $\varphi_{\boldsymbol{\theta}} \in \mathcal{M}_{\epsilon'}$. We define

$$M\left(\boldsymbol{\theta}, h\right) = \mathbb{E}\left[l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}, \varphi_{\boldsymbol{\theta}}\right)\right] \quad \text{and} \quad M_n\left(\boldsymbol{\theta}, h\right) = \frac{1}{n}\sum_{i=1}^{n} l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}, \varphi_{\boldsymbol{\theta}}\right).$$

Next we verify the required conditions (A1)–(A5) of Theorem 1 of [8].

Condition (A1) is satisfied by definition of $\hat{\boldsymbol{\theta}}_n$. For condition (A2), note that $M\left(\boldsymbol{\theta}, h_0\right) = \mathbb{E}\left[l\left(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}, \varphi_{0,\boldsymbol{\theta}}\right)\right]$. Under the model identifiability assumptions, the negative Kullback–Leibler divergence $M\left(\boldsymbol{\theta}_0, h_0\right) - M\left(\boldsymbol{\theta}, h_0\right)$ attains its maximum uniquely at $\boldsymbol{\theta}_0$. Since $M\left(\boldsymbol{\theta}, h_0\right)$ is continuous with respect to $\boldsymbol{\theta}$, on each compact set $\tilde{\Theta} \cap \{\boldsymbol{\theta} : d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq \delta\}$, it will obtain a maximum strictly smaller than $M\left(\boldsymbol{\theta}_0, h_0\right)$, which indicates that (A2) is fulfilled. Condition (A3) follows immediately from Proposition 4.6 and the fact that $\hat{\varphi}^s_{n,\boldsymbol{\theta}} \in \mathcal{M}_{\epsilon'}$ for all $\boldsymbol{\theta} \in \tilde{\Theta}$. Using Remark (ii) of [8], Condition (A4) follows from the fact that the family of uniformly bounded functions

$$\left\{ l\left(y, \delta, \mathbf{x}, \mathbf{z}\right) = (1-\delta)\log\left[1-\varphi(\boldsymbol{\gamma}^T\mathbf{x})\left\{1-\exp\left(-\Lambda\left(y\right)e^{\boldsymbol{\beta}^T\mathbf{z}}\right)\right\}\right] + \delta\log\varphi\left(\boldsymbol{\gamma}^T\mathbf{x}\right) \right.$$
$$\left. + \delta\left[\log\lambda(y) + \boldsymbol{\beta}^T\mathbf{z} - \Lambda\left(y\right)e^{\boldsymbol{\beta}^T\mathbf{z}}\right]; (\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda) \in \tilde{\Theta},\ \varphi \in \mathcal{M}_{\epsilon'} \right\}$$

is Glivenko–Cantelli. We have dealt with most terms in Lemma A.3. The remaining terms can be handled similarly. Dealing with the term involving λ is even simpler since that belongs to the class of functions of bounded variation, which has bracketing entropy of the same order as the class of monotone bounded functions. For condition (A5), note that for fixed $\boldsymbol{\theta} \in \tilde{\Theta}$, we have, from the proof of Proposition 4.1, that

$$\left|M\left(\boldsymbol{\theta}, h\right) - M\left(\boldsymbol{\theta}, h_0\right)\right| \leq \frac{2}{\epsilon'}\left(\int_{\mathcal{X}}\left\{\varphi_{\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)\right\}^2 \mathrm{d}Q_{\mathbf{X}}\left(\mathbf{x}\right)\right)^{\frac{1}{2}},$$

where $h\left(\mathbf{x}, \boldsymbol{\theta}\right) = \varphi_{\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T\mathbf{x}\right)$ for some $\varphi_{\boldsymbol{\theta}} \in \mathcal{M}_{\epsilon'}$. Then it follows immediately from the definition of $d_{\mathcal{H}}$ that $\sup_{\boldsymbol{\theta}\in\Theta}\left|M\left(\boldsymbol{\theta}, h\right) - M\left(\boldsymbol{\theta}, h_0\right)\right|$ converges to zero as $d_{\mathcal{H}}\left(h, h_0\right) \to 0$. □

Proof of Corollary 4.8. We first show the convergence of link estimator $\hat{p}_n$. Using Minkowski inequality and $\varphi_{0,\boldsymbol{\theta}_0} = \varphi_0$, we have

$$\left\{\int_{\mathcal{X}}\left\{\hat{\varphi}^s_{n,\hat{\boldsymbol{\theta}}_n}(\hat{\boldsymbol{\gamma}}_n^T\mathbf{x}) - \varphi_0\left(\boldsymbol{\gamma}_0^T\mathbf{x}\right)\right\}^2 \mathrm{d}Q_{\mathbf{X}}\left(\mathbf{x}\right)\right\}^{\frac{1}{2}}$$

$$
\begin{aligned}
&\leq \left\{ \int_{\mathcal{X}} \left\{ \hat{\varphi}^{s}_{n,\hat{\boldsymbol{\theta}}_n}(\hat{\boldsymbol{\gamma}}_n^T \mathbf{x}) - \varphi_{0,\hat{\boldsymbol{\theta}}_n}(\hat{\boldsymbol{\gamma}}_n^T \mathbf{x}) \right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) \right\}^{\frac{1}{2}} \\
&\quad + \left\{ \int_{\mathcal{X}} \left\{ \varphi_{0,\hat{\boldsymbol{\theta}}_n}(\hat{\boldsymbol{\gamma}}_n^T \mathbf{x}) - \varphi_{0,\boldsymbol{\theta}_0}(\hat{\boldsymbol{\gamma}}_n^T \mathbf{x}) \right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) \right\}^{\frac{1}{2}} \\
&\quad + \left\{ \int_{\mathcal{X}} \left\{ \varphi_0(\hat{\boldsymbol{\gamma}}_n^T \mathbf{x}) - \varphi_0\left(\boldsymbol{\gamma}_0^T \mathbf{x}\right) \right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) \right\}^{\frac{1}{2}} \\
&= (I) + (II) + (III).
\end{aligned}
$$

Consider the square of (I), using the result of Proposition 4.6, we have

$$
(I)^2 \leq \sup_{\boldsymbol{\theta} \in \tilde{\Theta}} \int_{\mathcal{X}} \left\{ \hat{\varphi}^{s}_{n,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T \mathbf{x}\right) - \varphi_{0,\boldsymbol{\theta}}\left(\boldsymbol{\gamma}^T \mathbf{x}\right) \right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) \xrightarrow{P} 0 \quad \text{as } n \to \infty.
$$

For the square of (III), by the assumption that φ_0' is uniformly bounded on $\mathcal{I}_0$ and using the mean value theorem followed by the Cauchy–Schwarz inequality, we have, for every $\mathbf{x} \in \mathcal{X}$,

$$
|\varphi_0(\hat{\boldsymbol{\gamma}}_n^T \mathbf{x}) - \varphi_0(\boldsymbol{\gamma}_0^T \mathbf{x})| \leq c_1 r_1 \left\| \hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0 \right\|_2,
$$

where c_1 a positive constant that does not depend on $\mathbf{x}$. Therefore, using the result of Theorem 4.7, we have

$$
(III)^2 = \int_{\mathcal{X}} \left\{ \varphi_0(\hat{\boldsymbol{\gamma}}_n^T \mathbf{x}) - \varphi_0\left(\boldsymbol{\gamma}_0^T \mathbf{x}\right) \right\}^2 \mathrm{d}Q_{\mathbf{X}}(\mathbf{x}) \leq c_1^2 r_1^2 \left\| \hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0 \right\|_2^2 \xrightarrow{P} 0.
$$

For the square of (II), we apply the argmax continuous mapping theorem [36, Theorem 3.2.2] to show $\sup_{x \in \mathbb{R}} |\varphi_{0,\hat{\boldsymbol{\theta}}_n}(x) - \varphi_{0,\boldsymbol{\theta}_0}(x)| \xrightarrow{P} 0$ if $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$. Given this uniform convergence, the convergence of (II) to zero in probability follows immediately. Let

$$
m(y, \delta, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \varphi) = \delta \log \varphi\left(\boldsymbol{\gamma}^T \mathbf{x}\right) + (1 - \delta) \log\left[1 - \varphi\left(\boldsymbol{\gamma}^T \mathbf{x}\right) F_u(y|\mathbf{z})\right].
$$

Define $\mathbb{M}_n(\varphi) = \mathbb{E}[m(Y, \Delta, \mathbf{X}, \mathbf{Z}; \hat{\boldsymbol{\theta}}_n, \varphi)]$ and $\mathbb{M}(\varphi) = \mathbb{E}[m(Y, \Delta, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}_0, \varphi)]$, where φ belongs to the metric space $H = (\mathcal{M}_{\epsilon'}, \|\cdot\|_\infty)$ with $\|\varphi\|_\infty = \sup_{x \in \mathbb{R}} |\varphi(x)|$. By the definition of $\varphi_{0,\boldsymbol{\theta}}$ in (13), we have $\varphi_{0,\hat{\boldsymbol{\theta}}_n} = \arg\max_{\varphi \in \mathcal{M}_{\epsilon'}} \mathbb{M}_n(\varphi)$ and $\varphi_{0,\boldsymbol{\theta}_0} = \arg\max_{\varphi \in \mathcal{M}_{\epsilon'}} \mathbb{M}(\varphi)$. We note that the continuity of the sample paths $\varphi \mapsto \mathbb{M}(\varphi)$ on H follows from a similar argument in the proof of Proposition 4.1. $\varphi_{0,\boldsymbol{\theta}_0}$ being tight and being the unique maximizer of $\mathbb{M}$, $\varphi_{0,\hat{\boldsymbol{\theta}}_n}$ being uniformly tight and maximizing $\mathbb{M}_n$ follow from their definitions. Therefore, it remains to show $\mathbb{M}_n$ converges weakly to $\mathbb{M}$ in $\ell^\infty(K)$ for every compact $K \subset H$. By Theorem 4.7 and since every $\varphi \in \mathcal{M}_{\epsilon'}$ are uniformly bounded, we can show a stronger statement that $\sup_{\varphi \in \mathcal{M}_{\epsilon'}} |\mathbb{M}_n(\varphi) - \mathbb{M}(\varphi)| \xrightarrow{P} 0$ and hence the weak convergence condition holds. For a fixed $\varphi \in \mathcal{M}_{\epsilon'}$, we have

$$
\begin{aligned}
&|\mathbb{M}_n(\varphi) - \mathbb{M}(\varphi)| \\
&\leq \left| \mathbb{E}\left[\Delta \log \frac{\varphi(\hat{\boldsymbol{\gamma}}_n^T \mathbf{X})}{\varphi(\boldsymbol{\gamma}_0^T \mathbf{X})} \right] \right| + \left| \mathbb{E}\left[(1 - \Delta) \log \frac{1 - \varphi(\hat{\boldsymbol{\gamma}}_n^T \mathbf{X}) F_u(Y|\mathbf{Z}; \hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n)}{1 - \varphi(\boldsymbol{\gamma}_0^T \mathbf{X}) F_u(Y|\mathbf{Z}; \boldsymbol{\beta}_0, \Lambda_0)} \right] \right| \\
&= (IV) + (V),
\end{aligned}
$$

where $F_u(t|\mathbf{z};\boldsymbol{\beta},\Lambda) = 1-\exp[-\Lambda(t)\exp(\boldsymbol{\beta}^T\mathbf{z})]$. For (IV), using a similar argument as in the proof of Proposition 4.1, we have

$$(IV) \leq \frac{1}{\epsilon'}\left(\mathbb{E}\left[\left\{\varphi(\hat{\boldsymbol{\gamma}}_n^T\mathbf{X})-\varphi(\boldsymbol{\gamma}_0^T\mathbf{X})\right\}^2\right]\right)^{\frac{1}{2}}.$$

To show that the term on the right-hand side of the above inequality converges to zero in probability, let $\zeta>0$ and consider ζ–brackets $[\varphi_k^L,\varphi_k^U]$, $k=1,\cdots,N$, covering the class $\mathcal{M}_{\epsilon'}$ such that

$$\int_{\mathcal{I}_0}\left\{\varphi_k^U(t)-\varphi_k^L(t)\right\}^2 \mathrm{d}G_{\boldsymbol{\gamma}_0^T\mathbf{X}}(t)\leq\zeta,\quad k=1,\cdots,N,$$

where $N\leq\exp(A/\zeta)$ for some constant $A>0$, by Lemma A.1. Using a similar argument in the proof of Lemma A.2, we have,

$$\begin{aligned}
&\left(\int_{\mathcal{X}}\left\{\varphi(\hat{\boldsymbol{\gamma}}_n^T\mathbf{x})-\varphi(\boldsymbol{\gamma}_0^T\mathbf{x})\right\}^2\mathrm{d}G_{\mathbf{X}}(\mathbf{x})\right)^{\frac{1}{2}}\\
&\leq\left(\int_{\mathcal{X}}\left\{\varphi_k^U(\boldsymbol{\gamma}_0^T\mathbf{x}+r_1\|\mathbf{e}_n\|_2)-\varphi_k^U(\boldsymbol{\gamma}_0^T\mathbf{x})\right\}^2\mathrm{d}G_{\mathbf{X}}(\mathbf{x})\right)^{\frac{1}{2}}\\
&\quad+\left(\int_{\mathcal{X}}\left\{\varphi_k^U(\boldsymbol{\gamma}_0^T\mathbf{x})-\varphi_k^L(\boldsymbol{\gamma}_0^T\mathbf{x})\right\}^2\mathrm{d}G_{\mathbf{X}}(\mathbf{x})\right)^{\frac{1}{2}}=(A)+(B),
\end{aligned}$$

for some $k=1,\cdots,N$, where $\mathbf{e}_n=\hat{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma}_0$. For the square of (A), using a similar argument in (28) in the proof of Lemma A.2, we have

$$\begin{aligned}
(A)^2&\leq\bar{q}_1(1-2\epsilon')\left(\int_{-r_1+r_1\|\mathbf{e}_n\|_2}^{r_1+r_1\|\mathbf{e}_n\|_2}\varphi_k^U(t)\,\mathrm{d}t-\int_{-r_1}^{r_1}\varphi_k^U(t)\,\mathrm{d}t\right)\\
&=\bar{q}_1(1-2\epsilon')\left(\int_{r_1}^{r_1+r_1\|\mathbf{e}_n\|_2}\varphi_k^U(t)\,\mathrm{d}t-\int_{-r_1}^{-r_1+r_1\|\mathbf{e}_n\|_2}\varphi_k^U(t)\,\mathrm{d}t\right)\\
&\leq\bar{q}_1 r_1(1-2\epsilon')^2\|\mathbf{e}_n\|_2.
\end{aligned}$$

For the square of (B), since $[\varphi_k^L,\varphi_k^U]$, $k=1,\cdots,N$, are ζ–brackets covering $\mathcal{M}_{\epsilon'}$, choosing $\zeta=\|\mathbf{e}_n\|_2$ we have $(B)^2\leq\|\mathbf{e}_n\|_2$. Note that ζ depends on n and on the realization of the data ω but the argument can be followed for any fixed n and ω. Therefore, by Theorem 4.7 together with the results on (A) and (B),

we obtain

$$\begin{aligned}
&\sup_{\varphi\in\mathcal{M}_{\epsilon'}}\left|\mathbb{E}\left[\Delta\log\frac{\varphi(\hat{\boldsymbol{\gamma}}_n^T\mathbf{X})}{\varphi(\boldsymbol{\gamma}_0^T\mathbf{X})}\right]\right| \\
&\le \frac{1}{\epsilon'}\left\{\sup_k\left(\int_{\mathcal{X}}\left\{\varphi_k^U(\boldsymbol{\gamma}_0^T\mathbf{x}+r_1\|\mathbf{e}_n\|_2)-\varphi_k^U(\boldsymbol{\gamma}_0^T\mathbf{x})\right\}^2\mathrm{d}G_{\mathbf{X}}(\mathbf{x})\right)^{\frac{1}{2}}\right. \\
&\qquad\left.+\sup_k\left(\int_{\mathcal{X}}\left\{\varphi_k^U(\boldsymbol{\gamma}_0^T\mathbf{x})-\varphi_k^L(\boldsymbol{\gamma}_0^T\mathbf{x})\right\}^2\mathrm{d}G_{\mathbf{X}}(\mathbf{x})\right)^{\frac{1}{2}}\right\} \\
&\le \frac{1}{\epsilon'}\left\{\left(1+(1-2\epsilon')\sqrt{\bar{q}_1 r_1}\right)\sqrt{\|\mathbf{e}_n\|_2}\right\}\xrightarrow{P}0\quad\text{as } n\to\infty.
\end{aligned}$$

Consider (V), using the mean value theorem, the Cauchy–Schwarz inequality, $\epsilon'\le\varphi\le 1-\epsilon'$, $F_u\le 1$, and the Minkowski inequality, we have

$$\begin{aligned}
(V)&\le\frac{1}{\epsilon'}\left\{\left(\mathbb{E}\left[\left\{\varphi(\hat{\boldsymbol{\gamma}}_n^T\mathbf{X})-\varphi(\boldsymbol{\gamma}_0^T\mathbf{X})\right\}^2F_u^2(Y|\mathbf{Z};\boldsymbol{\beta}_0,\Lambda_0)\right]\right)^{\frac{1}{2}}\right. \\
&\qquad\left.+\left(\mathbb{E}\left[\left\{F_u(Y|\mathbf{Z};\hat{\boldsymbol{\beta}}_n,\hat{\Lambda}_n)-F_u(Y|\mathbf{Z};\boldsymbol{\beta}_0,\Lambda_0)\right\}^2\varphi^2(\hat{\boldsymbol{\gamma}}_n^T\mathbf{X})\right]\right)^{\frac{1}{2}}\right\} \\
&\le\frac{1}{\epsilon'}\left(\mathbb{E}\left[\left\{\varphi(\hat{\boldsymbol{\gamma}}_n^T\mathbf{X})-\varphi(\boldsymbol{\gamma}_0^T\mathbf{X})\right\}^2\right]\right)^{\frac{1}{2}} \\
&\qquad+\frac{1-\epsilon'}{\epsilon'}\left(\mathbb{E}\left[\left\{F_u(Y|\mathbf{Z};\hat{\boldsymbol{\beta}}_n,\hat{\Lambda}_n)-F_u(Y|\mathbf{Z};\boldsymbol{\beta}_0,\Lambda_0)\right\}^2\right]\right)^{\frac{1}{2}} \\
&=(C)+(D).
\end{aligned}$$

The convergence of (C) to zero in probability follows from the previous results on (IV). For the square of (D), using the mean value theorem, $F_u\le 1$, and the Minkowski inequality, we have

$$\begin{aligned}
\left[\frac{\epsilon'}{1-\epsilon'}(D)\right]^2&\le\left(\mathbb{E}\left[\left\{\hat{\Lambda}_n(Y)-\Lambda_0(Y)\right\}^2\exp(2\hat{\boldsymbol{\beta}}_n^T\mathbf{Z})\right]\right)^{\frac{1}{2}} \\
&\qquad+\left(\mathbb{E}\left[\left\{\exp(\hat{\boldsymbol{\beta}}_n^T\mathbf{Z})-\exp(\boldsymbol{\beta}_0^T\mathbf{Z})\right\}^2\Lambda_0^2(Y)\right]\right)^{\frac{1}{2}} \\
&=(E)+(F).
\end{aligned}$$

For (E), we note that, for every $\mathbf{z}\in\mathcal{Z}$,

$$\exp(2\hat{\boldsymbol{\beta}}_n^T\mathbf{z})\le\exp(2\boldsymbol{\beta}_0^T\mathbf{z}+2r_2\,\|\mathbf{e}_n'\|_2)\le c_2\exp(\|\mathbf{e}_n'\|_2),$$

where $\mathbf{e}_n'=\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0$ and c_2 is a positive constant. The existence of the constant c_2 follows from assumptions (A4) and (A6). By Theorem 4.7 that $\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2$ and $\sup_{t\in[0,\tau_0]}|\hat{\Lambda}_n(t)-\Lambda_0(t)|$ converge to zero in probability, (E) converges to

zero in probability. For (F), we note that Λ_0 is strictly increasing and $\Lambda_0(\tau_0) < \infty$ by Assumption (A7). Using the mean value theorem, the Cauchy–Schwarz inequality, assumptions (A4) and (A6), and Theorem 4.7, we have

$$(F)^2 \leq c_3 \exp(\|\mathbf{e}_n'\|_2) \|\mathbf{e}_n'\|_2 \xrightarrow{P} 0 \quad \text{as } n \to \infty,$$

where c_3 is a positive constant. Thus, combining the above results on (IV) and (V), we have $\sup_{\varphi \in \mathcal{M}_{\epsilon'}} |\mathbb{M}_n(\varphi) - \mathbb{M}(\varphi)| \xrightarrow{P} 0$ as $n \to \infty$.

To show the convergence of the latency estimator $\hat{S}_u$, using the mean value theorem, $F_u \leq 1$, and the triangle inequality, we have, for every $t \in [0, \tau_0]$ and $\mathbf{z} \in \mathcal{Z}$,

$$\left|\hat{S}_u(t|\mathbf{z}) - S_u(t|\mathbf{z})\right| \leq \left|[\hat{\Lambda}_n(t) - \Lambda_0(t)] \exp(\hat{\boldsymbol{\beta}}_n^T \mathbf{z})\right| + \left|[\exp(\hat{\boldsymbol{\beta}}_n^T \mathbf{z}) - \exp(\boldsymbol{\beta}_0^T \mathbf{z})]\Lambda_0(t)\right|.$$

Using similar arguments as in (E) and (F), we have

$$\sup_{t \in [0,\tau_0]} \left|\hat{S}_u(t|\mathbf{z}) - S_u(t|\mathbf{z})\right| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

□

Appendix B: Additional simulation results

B.1. Selection of truncation parameters

When the link function has a small range as in experiment B of the simulation study, it is better to determine the truncation parameters for the link function in a data driven way instead of taking fixed bounds $\epsilon' = 10^{-6}$ and $1 - \epsilon'$. In practice, determining the truncation parameters from the data is challenging. We use ideas form the range-regularized isotonic regression problem studied in [27]. Specifically, we replace the uniform bound restricted MLE in step 7 of Algorithm 1 by considering the following range-regularized isotonic regression problem

$$\begin{aligned} \underset{\mathbf{y},a,b}{\text{minimize}} \quad & \sum_{i=1}^{n} (w_i - y_i)^2 + \mu\,(b - a) \\ \text{subject to} \quad & a \leq y_1 \leq y_2 \leq \cdots \leq y_n \leq b, \end{aligned} \tag{40}$$

where $a = \inf_{i=1,\cdots,n} y_i$, $b = \sup_{i=1,\cdots,n} y_i$, and μ is the regularization parameter for shrinking the range $b - a$. To reduce computational cost, the data driven truncation parameters are computed only at the first iteration of the EM algorithm. We applied the bounded isotonic regression algorithm proposed by [27] to obtain solutions over the regularization path. The optimal regularization parameter and hence the lower and upper truncation, are determined by a K-fold cross-validation with the expected prediction error for cure probability (EPECP) [19] as the performance metric. EPECP is a weight adjusted Brier score for assessing

the prediction accuracy of a mixture cure model in predicting cure probability. [19] proposed an estimator for EPECP and studied the statistical properties of this estimator.

In Tables 4–5 below we provide extra simulation results for experiment B using this choice of data-dependent truncation parameters.

TABLE 4
Simulation results for experiment B when the truncation parameters are equal

Sample Size	λ_C	Number of Cases	Method	MSE $(\hat{p}, p_0)$		$\hat{\gamma}$	
				Mean	Variance	Bias	Variance
250	0.1	129	mSIC	0.00350	1.77E-06	1.03314	0.23414
			SIC	0.00666	8.47E-05	1.10141	0.30039
	0.4	131	mSIC	0.00515	9.97E-06	0.95056	0.21264
			SIC	0.00936	1.51E-04	1.17386	0.29076
500	0.1	101	mSIC	0.00295	3.26E-07	0.87385	0.16862
			SIC	0.00400	2.94E-05	0.98576	0.30896
	0.4	116	mSIC	0.00489	4.57E-06	0.84465	0.18364
			SIC	0.00504	3.19E-05	1.11335	0.29095

TABLE 5
Simulation results for experiment B when the truncation parameters are different

Sample Size	λ_C	Number of Cases	Method	MSE $(\hat{p}, p_0)$		$\hat{\gamma}$	
				Mean	Variance	Bias	Variance
250	0.1	371	mSIC	0.00818	7.55E-05	0.98945	0.18702
			SIC	0.00888	1.00E-04	1.00346	0.25033
	0.4	369	mSIC	0.01052	1.22E-04	0.94831	0.16934
			SIC	0.01079	1.28E-04	1.00270	0.23379
500	0.1	399	mSIC	0.00468	2.75E-05	0.88273	0.17155
			SIC	0.00531	4.46E-05	0.90003	0.24909
	0.4	384	mSIC	0.00622	4.38E-05	0.87617	0.16895
			SIC	0.00601	4.80E-05	0.92918	0.23441

B.2. Non-monotone true link function

To investigate how the method performs when the true link function is not monotone, we consider an additional simulation experiment with the following non-monotone link function,

$$\varphi_D(u) = \frac{\exp[\psi(u-c)]}{1+\exp[\psi(u-c)]},$$

where c is an intercept term and $\psi(x) = 0.5x^3 - 0.1x^2 - 0.8x + 1$. Figure 5 depicts such function over $[-4, 4]$ when the intercept term $c = 0$. Table 6 shows the parameters $\boldsymbol{\gamma}_0$, $\boldsymbol{\beta}_0$, and λ_C of the additional simulation. The averages of the cure proportion, the censoring rate and the proportion of observations in the plateau are also reported in the same table. The truncation parameter ϵ'

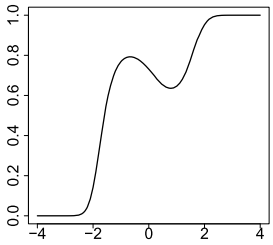


FIG 5. *The non-monotone link when the intercept term* $c = 0$.

TABLE 6
Additional simulation settings

Expt.	c	γ_{01}	γ_{02}	γ_{03}	γ_{04}	β_{01}	β_{02}	λ_C	Cure prop.	Cens. rate	Plateau
D	-0.5	0.6718	0.2896	-0.1547	0.6640	-0.4	-0.6	0.1	0.2679	0.3347	0.2079
								0.25		0.4215	0.1463

is set to 10^{-6}. Other settings, such as covariates, baseline distribution of the uncured subjects, censoring time distribution, remain the same as described in Section 5. Table 7 summarizes the simulation results, including the MSE of the link estimates, bias and variance of the coefficient estimates, for both SIC and mSIC methods. As expected, when the true link function is non-monotone, the SIC method performs better in estimating the link and $\boldsymbol{\gamma}$. In general, the effect would depend on the amount of deviation from the monotonicity assumption and mSIC would still perform well for small deviations. On the other hand, we observe that both methods behave similarly in estimating $\boldsymbol{\beta}$ meaning that the latency component is not very sensitive to non-monotonicity of the link function.

TABLE 7
Additional simulation results

Expt.	Size	λ_C	Method	MSE $(\hat{p}, p_0)$		$\hat{\boldsymbol{\gamma}}$		$\hat{\boldsymbol{\beta}}$	
				Mean	Variance	Bias	Variance	Bias	Variance
D	250	0.1	mSIC	0.01921	1.24E-04	0.6041	0.1425	0.2896	0.0301
			SIC	0.01160	9.05E-05	0.5012	0.1283	0.2894	0.0300
		0.25	mSIC	0.02312	1.45E-04	0.7073	0.1959	0.3241	0.0377
			SIC	0.01457	1.46E-04	0.6338	0.1816	0.3190	0.0375
	500	0.1	mSIC	0.01122	5.06E-05	0.4313	0.0813	0.2121	0.0147
			SIC	0.00601	2.69E-05	0.3250	0.0551	0.2127	0.0150
		0.25	mSIC	0.01324	6.41E-05	0.5126	0.1344	0.2349	0.0169
			SIC	0.00818	6.73E-05	0.4236	0.0973	0.2348	0.0171

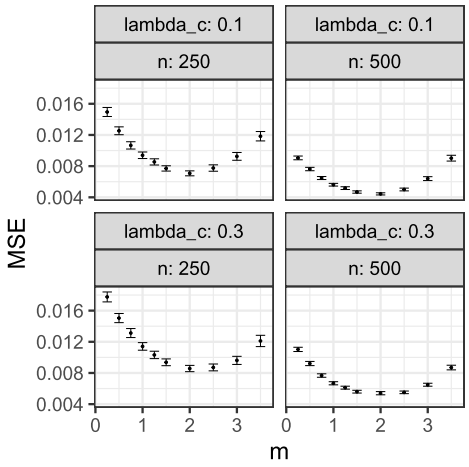


FIG 6. *$MSE(\hat{p}, p_0)$ against different bandwidths $h = mrn^{-1/5}$, where r is the range of the index $\boldsymbol{\gamma}^T\mathbf{X}$ and $m \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 2.5, 3\}$.*

B.3. Sensitivity to the choice of bandwidth

We re-consider Experiment A and investigate the influence of bandwidth on the mean squared error (MSE) of the estimated cure probability as mentioned in Section 5. In particular, we set the bandwidth parameter at the k-th iteration as $h_k = mr_k n^{-1/5}$, where r_k is the range of the index $\boldsymbol{\gamma}^T\mathbf{X}$ computed at the k-th iteration and $m \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 2.5, 3\}$. The averages MSE over the 500 simulated datasets for each setting are computed and depicted in Figure 6. Recall that in Section 5 the bandwidth is set to $r_k n^{-1/5}$, i.e., $m = 1$, and the mSIC method performs better than SIC. This figure shows that the MSE reaches minimum when $m = 2$ indicating that there is still room for improvement on estimating the link for mSIC. However, $m = 1$ is a satisfactory choice that does not the increase computational cost of the method. Even without an optimal bandwidth mSIC outperforms the SIC method. Actually for almost all the considered values of m the average MSE of mSIC remains below the one of SIC (given in Table 2). In general we observe that mSIC is more stable with respect to the choice of the bandwidth than the unconstrained estimator. Furthermore, the influence of the choice of bandwidth reduces when the sample size increases to 500.

B.4. Estimation without smoothing the monotone link estimator

We re-consider Experiment A and study the estimation of $\boldsymbol{\gamma}$ and the link p when the non-smoothed monotone link estimator is used, i.e. without applying step 2 in the estimation method described in Section 3. Without smoothing the monotone link estimator, the M-step of the EM algorithm in (12) to estimate

$\boldsymbol{\gamma}$ can be replaced by the score approach. The estimate of $\boldsymbol{\gamma}$ is obtained by computing the zero-crossings of

$$\sum_{i=1}^{n} \left[\frac{w_i}{\varphi_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^T \mathbf{x}_i)} - \frac{1 - w_i}{1 - \varphi_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^T \mathbf{x}_i)} \right] \mathbf{x}_i, \tag{41}$$

where $\varphi_{\boldsymbol{\gamma}}$ denotes the monotone link estimate $\hat{\varphi}_{n,\boldsymbol{\theta}}$ for $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}^{(m-1)}, \Lambda^{(m-1)})$, or minimizing the squared norm of (41) over $\mathcal{S}_{d-1}$. Table 8 shows MSE of the estimated cure probability, bias and variance of the estimated $\boldsymbol{\gamma}$. In summary, the mSIC method with the smoothed monotone link estimates performs better in estimating $\boldsymbol{\gamma}$ and the link p, among all simulation settings, when comparing with the score approach where the non-smoothed monotone link estimator is used.

TABLE 8
Simulation results of the score approach

Size	λ_C	Method	MSE $(\hat{p}, p_0)$		$\hat{\gamma}$	
			Mean	Variance	Bias	Variance
250	0.1	mSIC	0.00939	4.39E-05	0.57368	0.05038
		mSIC (score)	0.01645	7.95E-05	0.59970	0.05342
	0.3	mSIC	0.01141	5.92E-05	0.61280	0.05145
		mSIC (score)	0.01983	1.21E-04	0.64170	0.05902
500	0.1	mSIC	0.00562	1.50E-05	0.43937	0.03810
		mSIC (score)	0.01043	3.25E-05	0.49248	0.04135
	0.3	mSIC	0.00668	2.01E-05	0.48043	0.03989
		mSIC (score)	0.01198	4.08E-05	0.49994	0.04055

B.5. Coverage and length of confidence intervals

To compare the performance of the mSIC and SIC methods in terms of coverage rate and length of confidence intervals, we re-consider Experiment A and compute the percentile-based bootstrap 95% confidence intervals. The confidence intervals are computed using 500 naive bootstrap samples. It is worth noting that the SIC method requires fixing the first component of $\boldsymbol{\gamma}$ to be 1 or -1 for model identifiability. The sign is determined from the corresponding $\boldsymbol{\gamma}$ estimate of the LC model. In each bootstrap iteration, the sign is determined from the corresponding $\boldsymbol{\gamma}$ estimate of the LC model for the bootstrap sample. The estimator is then scaled to have norm one, in order to make it comparable to our estimator.

Table 9 summarizes the coverage rate and average length of confidence intervals for $\boldsymbol{\gamma}$ over the 500 simulated datasets. Note that the $\boldsymbol{\gamma}$ estimate of the LC method are not normalized, which leads to a larger average length of confidence intervals when comparing to the other two methods. In summary, the average confidence interval length for $\boldsymbol{\gamma}$ using the mSIC method is smaller than that of the SIC method. The coverage rate for $\boldsymbol{\gamma}$ using mSIC is also lower when comparing with the SIC method. Hence SIC intervals perform better in this setting.

TABLE 9
Coverage rate and average length of confidence intervals for $\boldsymbol{\gamma}$

Expt.	λ_C	Size	Method	CI length					Coverage				
				γ_0	γ_1	γ_2	γ_3	γ_4	γ_0	γ_1	γ_2	γ_3	γ_4
A	0.1	250	mSIC	–	1.106	0.558	0.934	0.845	–	0.934	0.892	0.926	0.888
			SIC	–	1.333	0.629	1.069	0.970	–	0.982	0.922	0.969	0.941
			LC	1.656	2.565	0.823	1.791	1.485	0.918	0.96	0.914	0.948	0.930
		500	mSIC	–	0.958	0.435	0.766	0.665	–	0.93	0.92	0.916	0.92
			SIC	–	1.239	0.536	0.921	0.839	–	0.983	0.973	0.960	0.943
			LC	1.108	1.718	0.544	1.138	1.001	0.916	0.946	0.934	0.938	0.94
	0.3	250	mSIC	–	1.037	0.538	0.878	0.791	–	0.898	0.8	0.856	0.882
			SIC	–	1.342	0.690	1.111	1.012	–	0.984	0.905	0.960	0.960
			LC	2.005	3.019	0.983	2.506	1.767	0.914	0.950	0.920	0.936	0.932
		500	mSIC	–	0.890	0.432	0.715	0.616	–	0.858	0.824	0.876	0.894
			SIC	–	1.208	0.579	0.957	0.859	–	0.977	0.961	0.955	0.959
			LC	1.288	1.967	0.622	1.332	1.142	0.928	0.928	0.922	0.932	0.936

However, as sample size increases, it seems like the SIC intervals remain more conservative (wider and with coverage larger than 95%), while mSIC interval would be shorter and with coverage closer to 95%. We did not perform complete simulation studies for larger sample sizes because bootstrap then becomes computationally intensive. It remains interesting to further investigate whether the bootstrap confidence intervals can be computed in a different way (e.g using smooth bootstrap) that might lead to better coverage for mSIC, since in terms of bias and variance such estimator performs better. We tried also basic (reverse percentile) confidence intervals but the performance for sample sizes 250 and 500 was considerably worse for both methods. Table 10 provides the coverage rate and average length of confidence intervals for $\boldsymbol{\beta}$ over the 500 simulated datasets. All methods performs similarly in inferring $\boldsymbol{\beta}$ in terms of the coverage rate and average confidence interval length. This is in line with our previous results which show that the estimation of $\boldsymbol{\beta}$ is not affected much by the estimation of the incidence component.

B.6. Real data application revisited

In Section 6, we compared the performance of the three models (LC, SIC, and mSIC) in predicting the uncure probability using the prediction error (PE) and observed that mSIC behaved the best. However, even the link estimate of the SIC method in Figure 4 is monotone, which makes it intriguing to understand the reason behind the better performance of mSIC. We looked at the link estimates over the 10 random splits for each method, which are shown in Figure 7. We observe that the link estimates of the SIC method are in general not monotone and it was a coincidence that the particular random split considered in

TABLE 10
Coverage rate and average length of confidence intervals for $\boldsymbol{\beta}$

Expt.	λ_C	Size	Method	CI length		Coverage	
				β_1	β_2	β_1	β_2
A	0.1	250	mSIC	1.083	0.653	0.930	0.938
			SIC	1.084	0.653	0.930	0.934
			LC	1.085	0.654	0.928	0.934
		500	mSIC	0.746	0.450	0.936	0.950
			SIC	0.747	0.450	0.934	0.948
			LC	0.747	0.450	0.938	0.950
	0.3	250	mSIC	1.204	0.728	0.936	0.942
			SIC	1.209	0.730	0.932	0.946
			LC	1.216	0.736	0.938	0.940
		500	mSIC	0.828	0.501	0.928	0.946
			SIC	0.831	0.501	0.934	0.948
			LC	0.835	0.505	0.926	0.950

Section 6 gave a monotone estimator. This shows the need to impose the monotonicity assumption in the estimation procedure. Out of the 10 random splits, one in particular leads to a non-monotone link estimate that differs considerably from the others and also reaches value 1 for index in the range $(0.81, 0.87)$. For this case, we observe that the support of the fitted index $\hat{\boldsymbol{\gamma}}_n^T \mathbf{x}_i^{\text{train}}$ for SIC, a histogram of which is shown in Figure 8a, is divided into two disjoint intervals. The lack of observations in the region around $(0.81, 0.87)$ leads to the strange behavior of link estimate. Recall that the SIC method employs a leave-one-out cross-validation approach to search the optimal bandwidth over the interval $[0.4, 1]$. For this particular case, the selected bandwidth was the upper bound 1. We increased the upper bound of the search interval from 1 to 5 and fitted again the SIC model for the same 10 random splits of the data. The average PE (standard deviation) for SIC with bandwidth adjustment is 90.54 (5.82), which improves slightly as compared to the SIC models without adjusting the bandwidth, with average PE (standard deviation) 91.84 (9.51). Figure 8b shows that the fitted index after adjusting the search of bandwidth is not divided into two disjoint intervals. This suggests that the cross-validation applied to choose the bandwidth of the SIC method is not stable in practice and the search interval should depend on the range of the index, hence should be data dependent. In contrast, we showed that the mSIC method performs well even with a simple bandwidth choice. Figure 9b depicts the link estimates of the SIC method with the bandwidth adjustment. Again, the link estimates are non-monotone in general but possess less variability as compared to the un-adjusted ones. The estimated links using SIC are more volatile in the region where the fitted index is sparse (see the + sign in Figures 7 and 9) compared to mSIC. Imposing the monotonicity constraint in mSIC increases the stability of the link estimate particularly in the region of index that possesses less available observations.

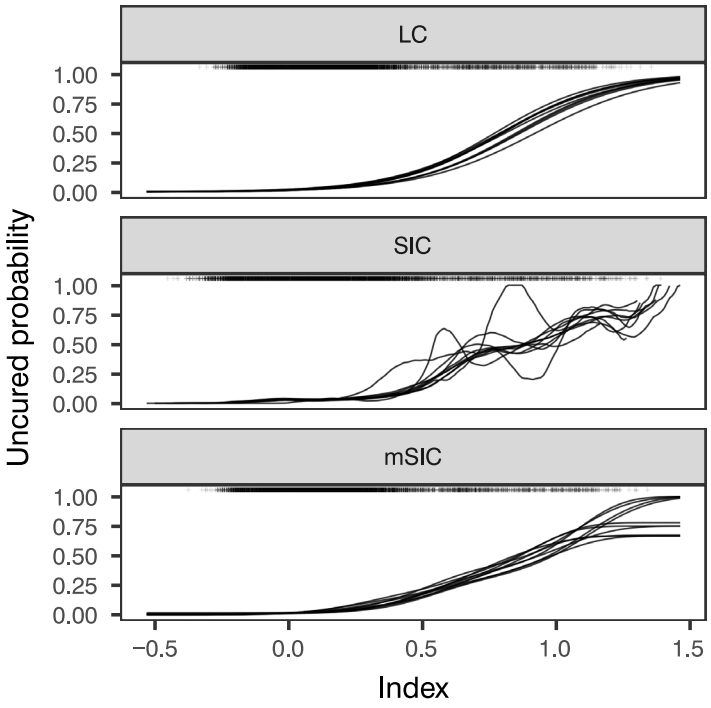
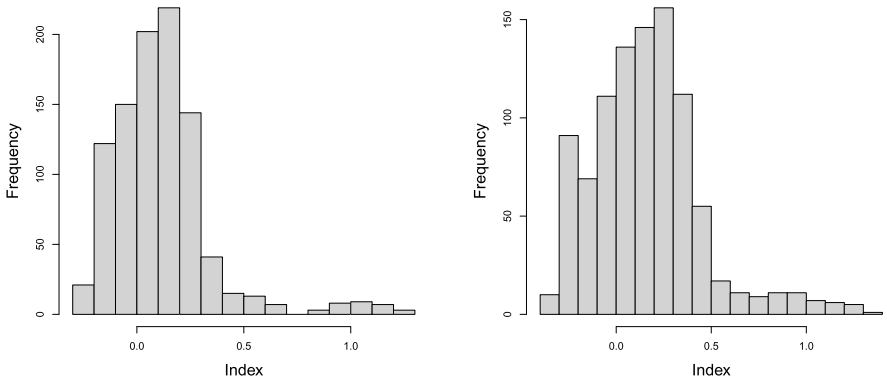
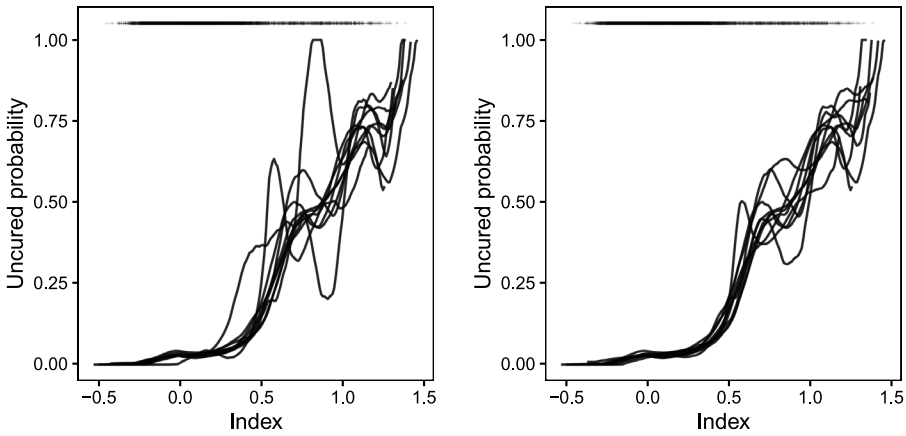


FIG 7. *Link estimates over the 10 random splits (+ indicates the density of the fitted index.)*



(a) Without bandwidth adjustment (b) With bandwidth adjustment

FIG 8. *Histogram of the fitted index of the selected SIC model.*



(a) Without bandwidth adjustment (b) With bandwidth adjustment

FIG 9. *SIC link estimates over the 10 random splits (+ indicates the density of the fitted index.)*

B.7. Algorithms

Algorithm 1 Monotone Link Estimator $\hat{\varphi}_{n,\boldsymbol{\theta}}$ for fixed $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda)$

Require:

Observed data $\{(y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \cdots, n\}$

Fixed parameter values $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \Lambda)$

Initial link $\hat{\varphi}^{(0)}$

Trunction parameter $\epsilon' > 0$

Let $S_u(y_i|\mathbf{z}_i) = \exp\left[-\Lambda(y_i)\exp\left(\boldsymbol{\beta}^T\mathbf{z}_i\right)\right]$ and $F_u = 1 - S_u$

$\hat{\varphi}^{(0)}_{n,\boldsymbol{\theta}} \leftarrow \hat{\varphi}^{(0)}$ and $k \leftarrow 0$

repeat

  $k \leftarrow k + 1$

  $\hat{p}(\mathbf{x}_i) \leftarrow \hat{\varphi}^{(k-1)}_{n,\boldsymbol{\theta}}(\boldsymbol{\gamma}'\mathbf{x}_i)$ for all $i = 1, \cdots, n$

  E-step

  $w_i \leftarrow \delta_i + (1 - \delta_i)\frac{\hat{p}(\mathbf{x}_i)S_u(y_i|\mathbf{z}_i)}{1-\hat{p}(\mathbf{x}_i)F_u(y_i|\mathbf{z}_i)}$ for all $i = 1, \cdots, n$

  M-step

  $\hat{\varphi}^{(k)}_{n,\boldsymbol{\theta}} \leftarrow \arg\max_{\varphi \in \mathcal{M}_{\epsilon'}} \sum_{i=1}^{n}\left[w_i \log\varphi\left(\boldsymbol{\gamma}^T\mathbf{x}_i\right) + (1 - w_i)\log\left\{1 - \varphi\left(\boldsymbol{\gamma}^T\mathbf{x}_i\right)\right\}\right]$ ▷ Using uniform bound restricted MLE

until Termination criterion is satisfied

return $\hat{\varphi}^{(k)}_{n,\boldsymbol{\theta}}$

Algorithm 2 Model Estimation

Require:
Observed data $\{(y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \cdots, n\}$
Trunction parameter $\epsilon' > 0$
Kernel function k
Bandwidth $h > 0$

Initialize $\hat{\boldsymbol{\gamma}}^{(0)}$, $\hat{\boldsymbol{\beta}}^{(0)}$, $\hat{\Lambda}^{(0)}$ and $\hat{\varphi}^{(0)}$; m $\leftarrow 0$
$\hat{\varphi} \leftarrow \arg\max_{\varphi \in \mathcal{M}_{\epsilon'}} L_n\left(\hat{\boldsymbol{\gamma}}^{(0)}, \hat{\boldsymbol{\beta}}^{(0)}, \hat{\Lambda}^{(0)}, \varphi\right)$
▷ Using Algorithm 1 with $\hat{\boldsymbol{\theta}}^{(0)}$ and $\hat{\varphi}^{(0)}$
$\hat{\varphi}^{(0)}(u) \leftarrow \int \frac{1}{h} k\left(\frac{u-t}{h}\right) \hat{\varphi}(t)\, dt$
▷ Smoothing as in (6)
repeat
  $m \leftarrow m + 1$
  E-step
  $\hat{p}(\mathbf{x}_i) \leftarrow \hat{\varphi}^{(m-1)}\left(\hat{\boldsymbol{\gamma}}^{(m-1)^T} \mathbf{x}_i\right)$ for all $i = 1, \cdots, n$
  $\hat{S}_u(y_i|\mathbf{z}_i) \leftarrow \exp\left[-\hat{\Lambda}^{(m-1)}(y_i) \exp\left(\hat{\boldsymbol{\beta}}^{(m-1)^T} \mathbf{z}_i\right)\right]$ and $\hat{F}_u = 1 - \hat{S}_u$
  $\hat{w}_i \leftarrow \delta_i + (1 - \delta_i) \frac{\hat{p}(\mathbf{x}_i) \hat{S}_u(y_i|\mathbf{z}_i)}{1 - \hat{p}(\mathbf{x}_i) \hat{F}_u(y_i|\mathbf{z}_i)}$ for all $i = 1, \cdots, n$
  M-step
  $\hat{\boldsymbol{\gamma}}^{(m)} \leftarrow \arg\max_{\boldsymbol{\gamma} \in \mathcal{S}_{d-1}} \tilde{L}_{nc,1}\left(\boldsymbol{\gamma}, \hat{\varphi}^s_{n,\boldsymbol{\gamma},\boldsymbol{\beta}^{(m-1)},\Lambda^{(m-1)}}\right)$
  ▷ Using the augmented Lagrangian method
  $\left(\hat{\boldsymbol{\beta}}^{(m)}, \hat{\Lambda}^{(m)}\right) \leftarrow \arg\max_{\boldsymbol{\beta},\Lambda} \tilde{L}_{nc,2}(\boldsymbol{\beta}, \Lambda)$ ▷ Using the profile likelihood approach
  Update the link
  $\hat{\varphi} \leftarrow \arg\max_{\varphi \in \mathcal{M}_{\epsilon'}} L_n\left(\hat{\boldsymbol{\gamma}}^{(m)}, \hat{\boldsymbol{\beta}}^{(m)}, \hat{\Lambda}^{(m)}, \varphi\right)$
  ▷ Using Algorithm 1 with $\hat{\boldsymbol{\theta}}^{(m)}$ and $\hat{\varphi}^{(m-1)}$
  $\hat{\varphi}^{(m)}(u) \leftarrow \int \frac{1}{h} k\left(\frac{u-t}{h}\right) \hat{\varphi}(t)\, dt$ ▷ Smoothing as in (6)
until Termination criterion is satisfied:
return $\hat{\boldsymbol{\gamma}}^{(m)}$, $\hat{\boldsymbol{\beta}}^{(m)}$, $\hat{\Lambda}^{(m)}$, $\hat{\varphi}^{(m)}$

Acknowledgments

The authors would like to thank the referees and the Associate Editor for their constructive suggestions and comments that improved the quality of this paper.

References

[1] Amico, M. and Van Keilegom, I. (2018). Cure Models in Survival Analysis. *Annual Review of Statistics and Its Application* **5** 311–342. MR3774750
[2] Amico, M., Van Keilegom, I. and Legrand, C. (2019). The single-index/Cox mixture cure model. *Biometrics* **75** 452–462. MR3999169
[3] Balabdaoui, F., Durot, C. and Jankowski, H. (2019). Least squares estimation in the monotone single index model. *Bernoulli* **25** 3276–3310. MR4010955
[4] Balabdaoui, F. and Groeneboom, P. (2021). Profile least squares estimators in the monotone single index model. In *Advances in Contemporary Statistics and Econometrics* 3–22. Springer. MR4299271

[5] Balabdaoui, F., Groeneboom, P. and Hendrickx, K. (2019). Score estimation in the monotone single-index model. *Scandinavian Journal of Statistics* **46** 517–544. MR3948566
[6] Cheng, M.-Y., Gasser, T. and Hall, P. (1999). Nonparametric Density Estimation under Unimodality and Monotonicity Constraints. *Journal of Computational and Graphical Statistics* **8** 1–21. MR1705906
[7] Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B (Methodological)* **34** 187–220. MR0341758
[8] Delsol, L. and Van Keilegom, I. (2020). Semiparametric M-estimation with non-smooth criterion functions. *Annals of the Institute of Statistical Mathematics* **72** 577–605. MR4067237
[9] Dirick, L., Claeskens, G. and Baesens, B. (2017). Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society* **68** 652–665.
[10] Farewell, V. T. (1977). A Model for a Binary Variable with Time-Censored Observations. *Biometrika* **64** 43–46.
[11] Farewell, V. T. (1982). The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors. *Biometrics* **38** 1041–1046.
[12] Groeneboom, P. and Hendrickx, K. (2018). Current status linear regression. *The Annals of Statistics* **46** 1415–1444. MR3819105
[13] Groeneboom, P. and Hendrickx, K. (2019). Estimation in monotone single-index models. *Statistica Neerlandica* **73** 78–99. MR3905873
[14] Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. MR3445293
[15] Hardle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *The annals of Statistics* **21** 157–178. MR1212171
[16] Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*, 1 ed. Springer New York, New York, NY. MR2535631
[17] Hu, X. (1997). Maximum-likelihood estimation under bound restriction and order and uniform bound restrictions. *Statistics & Probability Letters* **35** 165–171. MR1483270
[18] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58** 71–120. MR1230981
[19] Jiang, W., Sun, H. and Peng, Y. (2017). Prediction accuracy for the cure probabilities in mixture cure models. *Stat Methods Med Res* **26** 2029–2041. MR3712218
[20] Legrand, C. and Bertrand, A. (2019). *Cure Models in Cancer Clinical Trials*, 1st ed. *Textbook of Clinical Trials in Oncology: A Statistical Perspective* 465–492. Chapman and Hall/CRC, New York, NY. MR3738479
[21] Li, C.-S. and Taylor, J. M. G. (2002). A semi-parametric accelerated failure time cure model. *Statistics in medicine* **21** 3235–3247.
[22] Lin, W. and Kulasekera, K. B. (2007). Identifiability of Single-Index

Models and Additive-Index Models. *Biometrika* **94** 496–501. MR2380574
[23] Lopuhaä, H. P. and Musta, E. (2017). Smooth estimation of a monotone hazard and a monotone density under random censoring. *Statistica Neerlandica* **71** 58–82. MR3605364
[24] Lopuhaä, H. P. and Musta, E. (2018). Smoothed isotonic estimators of a monotone baseline hazard in the Cox model. *Scandinavian Journal of Statistics* **45** 753–791. MR3858954
[25] Lopuhaä, H. P. and Musta, E. (2019). Central limit theorems for the L_p-error of smooth isotonic estimators. *Electronic Journal of Statistics* **13** 1031–1098. MR3935844
[26] Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Annals of the Institute of Statistical Mathematics* **60** 545–574. MR2434411
[27] Luss, R. and Rosset, S. (2017). Bounded isotonic regression. *Electronic Journal of Statistics* **11** 4488–4514. MR3724487
[28] Meyer, M. C. (2008). Inference Using Shape-Restricted Regression Splines. *The Annals of Applied Statistics* **2** 1013–1033. MR2516802
[29] Musta, E., Patilea, V. and Van Keilegom, I. (2022). A presmoothing approach for estimation in the semiparametric Cox mixture cure model. *Bernoulli* **28** 2689–2715. MR4474559
[30] Nocedal, J. and Wright, S. (2006). *Numerical Optimization*, 2 ed. Springer, New York, NY. MR2244940
[31] Peng, Y. and Dear, K. B. G. (2000). A Nonparametric Mixture Model for Cure Rate Estimation. *Biometrics* **56** 237–243.
[32] Peng, Y. and Yu, B. (2021). *Cure Models: Methods, Applications, and Implementation*, 1st ed. *Chapman & Hall/CRC Biostatistics Series*. Chapman and Hall/CRC, New York, NY. MR4596124
[33] Robertson, T., Wright, F. T. and Dykstra, R. (1988). *Order Restricted Statistical Inference*. Wiley, New York. MR0961262
[34] Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox Proportional Hazards Cure Model. *Biometrics* **56** 227–236. MR1767631
[35] Taylor, J. M. G. (1995). Semi-Parametric Estimation in Failure Time Mixture Models. *Biometrics* **51** 899–907.
[36] Van de Geer, S. A. (2000). *Empirical Processes in M-estimation* **6**. Cambridge university press. MR0937256
[37] Van Der Vaart, A. and Van Der Laan, M. (2003). Smooth estimation of a monotone density. *Statistics: A Journal of Theoretical and Applied Statistics* **37** 189–203. MR1986176
[38] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York, N.Y. MR1385671
[39] Van Geloven, N., Van der Veen, F., Bossuyt, P. M. M., Hompes, P. G., Zwinderman, A. H. and Mol, B. W. (2012). Can we distinguish between infertility and subfertility when predicting natural conception in couples with an unfulfilled child wish? *Human Reproduction* **28** 658–665.

[40] Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics* **42** 1–17. MR3181580

[41] Zhang, J. and Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in medicine* **26** 3157–3171. MR2380509