engineering proceedings

MDPI

Proceeding Paper

Nonparametric Inference for Mixture Cure Model When Cure Information Is Partially Available †

Wende Clarence Safari [1,2,*], Ignacio López-de-Ullibarri [2] and María Amalia Jácome [1,2]

[1] Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), Universidade da Coruña, 15071 A Coruña, Spain
[2] MODES Group, Department of Mathematics, Universidade da Coruña, 15071 A Coruña, Spain; ilu@udc.es (I.L.-d.-U.); majacome@udc.es (M.A.J.)
* Correspondence: wende.safari@udc.es
† Presented at the 4th XoveTIC Conference, A Coruña, Spain, 7–8 October 2021.

Abstract: We introduce nonparametric estimators to estimate the conditional survival function, cure probability and latency function in the setting of a mixture cure model when the cure status is partially known. For the sake of illustration, we present an application concerning patients hospitalized with COVID-19 in Galicia (Spain) during the first outbreak of the epidemic.

Keywords: COVID-19; ICU; kernel estimators; mixture cure model; survival analysis

Citation: Safari, WC.; López-de-Ullibarri, I.; Jácome, M.A. Nonparametric Inference for Mixture Cure Model When the Cure Information Is Partially Available. *Eng. Proc.* **2021**, *7*, 17. https://doi.org/10.3390/engproc2021007017

Academic Editors: Joaquim de Moura, Marco A. González, Javier Pereira and Manuel G. Penedo

Published: 9 October 2021

1. Introduction

Survival analysis arises in many applications where we want to reason about the amount of time until the considered event happens. A common assumption in standard survival modeling is that all individuals can experience the event if observed for a sufficient amount of time. Cure models [1] have been developed because there might be situations where the standard survival model is not true, for example, in the event of a recurrence in some diseases or death from some types of cancer. One challenge with time-to-event data is that the event is not always observed (censored observations). Standard cure models typically make inferences based on the assumption that the cure status information is an unobserved (latent) variable as the event is only known for the uncensored (uncured) subjects, but it is unknown for the censored observations whether it is cured or not. There are situations where cure status information is known for some of the censored individuals as they can be identified to be insusceptible to the considered event, that is, known to be cured. For example, when a medical test ascertains that a disease has entirely disappeared after treatment.

In this paper, we present kernel methods to estimate the conditional survival function, cure probability and latency function in the presence of cure status information. The proposed approach contributes to state-of-the-art in time-to-event data, as it extends previous works in the mixture cure model.

2. Estimation When the Cure Status Is Partially Available

Let Y be the time until the event of interest, X is a vector of covariates and $F(t \mid \mathbf{x}) = P(Y \leq t \mid \mathbf{X} = \mathbf{x})$ is the distribution function of Y conditional on $\mathbf{X} = \mathbf{x}$. In follow-up studies, the event of interest may not be observed due to, for example, the end of the study or loss to follow up, which occurs at censoring time C^* with conditional distribution function $G(t \mid \mathbf{x}) = P(C^* \leq t \mid \mathbf{X} = \mathbf{x})$. As a consequence, instead of observing Y, only the possibly censored survival time $T^* = \min(Y, C^*)$ and the indicator of the event $\delta = \mathbf{1}(Y < C^*)$ can be observed. The random variables Y and C^* are assumed to be conditionally independent given $\mathbf{X} = \mathbf{x}$, which is a widely used assumption in most studies. We set $Y = \infty$ if the subject will not experience the event and so is cured. Let $\nu = \mathbf{1}(Y = \infty)$

be the indicator of being cured. Note that ν is partially observed because the individual is known not to be cured ($\nu = 0$) when the event is observed ($\delta = 1$), but in the general situation, ν is unknown when $\delta = 0$. When the cure status is partially known, some censored individuals are identified to be cured, so $\nu = 1$ is observed.

To accommodate the cure status information, we include an additional random variable ξ, which indicates whether the cure status ν is known ($\xi = 1$) or not ($\xi = 0$). Furthermore, let the censoring distribution be an improper distribution function $G(t \mid \mathbf{x}) = (1 - \pi(\mathbf{x}))G_0(t \mid \mathbf{x})$. Thus, with probability $\pi(\mathbf{x})$, the censoring variable is $C^* = \infty$, and with probability $1 - \pi(\mathbf{x})$ the value of the censoring variable C^* corresponds to the value of a random variable C with proper continuous distribution function $G_0(t \mid \mathbf{x})$. A cured individual is identified with probability $P(\xi = 1 \mid \nu = 1, \mathbf{X} = \mathbf{x}) = P(C^* = \infty \mid \mathbf{X} = \mathbf{x}) = \pi(\mathbf{x})$. In this setup, the data actually observed are $\{(\mathbf{X}_i, T_i, \delta_i, \xi_i, \xi_i\nu_i) : i = 1, \ldots, n\}$, where the observed time is $T_i = \min(Y_i, C_i^*) = T_i^*$, except for those identified as cured which is $T_i = C_i$. Hence, the observations $\{(\mathbf{X}_i, T_i, \delta_i, \xi_i, \xi_i\nu_i) : i = 1, \ldots, n\}$ can be classified into three groups: (a) the individual is observed to have experienced the event and, therefore, is known to be uncured ($\mathbf{X}_i, T_i = Y_i, \delta_i = 1, \xi_i = 1, \xi_i\nu_i = 0$); (b) the lifetime is censored and the cure status is unknown ($\mathbf{X}_i, T_i = C_i, \delta_i = 0, \xi_i = 0, \xi_i\nu_i = 0$); and (c) the lifetime is censored and the individual is known to be cured ($\mathbf{X}_i, T_i = C_i, \delta_i = 0, \xi_i = 1, \xi_i\nu_i = 1$). In standard cure models where the cure status is unknown for all the censored observations, only groups (a) and (b) are considered.

The probability of cure is $1 - p(\mathbf{x}) = P(Y = \infty \mid \mathbf{X} = \mathbf{x})$, and the conditional survival function of the uncured individuals, also known as latency, is $S_0(t \mid \mathbf{x}) = P(Y > t \mid Y < \infty, \mathbf{X} = \mathbf{x})$. The mixture cure model specifies the survival function $S(t \mid \mathbf{x}) = P(Y > t \mid \mathbf{X} = \mathbf{x})$ as the following.

$$S(t \mid \mathbf{x}) = 1 - p(\mathbf{x}) + p(\mathbf{x})S_0(t \mid \mathbf{x}). \tag{1}$$

Assuming model (1) and the availability of a suitable estimator of the $S(t \mid x)$, estimators of the cure probability and the latency can be derived by considering the following relationships.

$$1 - p(x) = \lim_{t \to \infty} S(t \mid x) > 0, \; S_0(t \mid x) = \frac{S(t \mid x) - \{1 - p(x)\}}{p(x)}. \tag{2}$$

Safari et al. [2] proposed the generalized product-limit estimator of the conditional survival function $S(t \mid x)$ when the cure status is partially known, which is the following:

$$\widehat{S}_h^c(t \mid x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]} B_{h[i]}(x) \mathbf{1}\left(T_{(i)} \leq t\right)}{\sum_{j=i}^{n} B_{h[j]}(x) + \sum_{j=1}^{i-1} B_{h[j]}(x) \mathbf{1}\left(\xi_{[j]}\nu_{[j]} = 1\right)} \right), \tag{3}$$

where $X_{[i]}$, $\delta_{[i]}$, $\xi_{[i]}$, and $\nu_{[i]}$ are the concomitants of the ordered observed times $T_{(1)} \leq \ldots \leq T_{(n)}$, and $B_{h[i]}(x)$ is the Nadaraya–Watson (NW) weight of the following:

$$B_{h[i]}(x) = \frac{K_h\left(x - X_{[i]}\right)}{\sum_{j=1}^{n} K_h(x - X_j)},$$

$K_h(\cdot) = K(\cdot/h)/h$ is a kernel function $K(\cdot)$ rescaled with bandwidth h. The corresponding estimator of the cure rate $1 - p(x)$ [3] is the following:

$$1 - \widehat{p}_h^c(x) = \widehat{S}_h^c\left(T_{(n)}^1 \mid x\right), \tag{4}$$

where $T_{(n)}^1$ is the largest uncensored observed time. Here, in light of (3), (4), and the relation in (2), a nonparametric estimator of the latency function is given by the following.

$$\widehat{S}^c_{0,h_1,h_2}(t \mid x) = \begin{cases} \dfrac{\widehat{S}^c_{h_2}(t \mid x) - (1 - \widehat{p}^c_{h_1}(x))}{\widehat{p}^c_{h_1}(x)} & \text{if } 0 \le t \le T^1_{(n)} \text{ and } \widehat{S}^c_{h_2}(t \mid x) > 1 - \widehat{p}^c_{h_1}(x) \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

The optimal bandwidth for $\widehat{S}^c_h(t \mid x)$ in (3) is not necessarily the optimal bandwidth for $1 - \widehat{p}^c_h(x)$ in (4); therefore, the estimator in (5) is a more general estimator that uses two different bandwidths for estimating $S(t \mid x)$ and $1 - p(x)$. Note that if $h = h_1 = h_2$, then the estimator in (5) reduces to the following estimator.

$$\widehat{S}^c_{0,h}(t \mid x) = \frac{\widehat{S}^c_h(t \mid x) - (1 - \widehat{p}^c_h(x))}{\widehat{p}^c_h(x)}.$$

3. Application to COVID-19 Data

For illustration of the nonparametric estimators stated in Section 2, we present an application concerning patients hospitalized with COVID-19 in Galicia (Spain) during the first outbreak of the epidemic. We have a medical database of 10,454 COVID-19 patients reported by the Galician Healthcare Service between 6 March and 7 May 2020. This database contains some information on sex, age, and the dates of different medical outcomes such as admission to the intensive care unit (ICU), discharge, or death. The aim was to estimate the time from hospital ward until admission to ICU while adjusting for age and sex. In our analysis we included only 2380 patients who had been hospitalized for at least a day. Among them, 8.3% were admitted to ICU and 91.7% were censored. In the censored group, 68.8% patients were discharged from the hospital alive and without the need for ICU, and 13.8% died without entering the ICU. Therefore, these patients were identified to be "cured" from the event of interest, which is admission to ICU. Note that in this example, "being cured" means being free of experiencing admission to ICU and not being cured in medical terms.

Acknowledgments: This work has been supported by MINECO grant MTM2017-82724-R and the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-014) and we wish to acknowledge the support received from the Centro de Investigación de Galicia "CITIC" funded by Xunta de Galicia and the European Union (European Regional Development Fund Galicia 2014–2020 Program) by grant ED431G 2019/01. The authors are grateful to Andrés Paz-Ares Rodríguez (General Director of Public Health), Xurxo Hervada Vidal (General Deputy Director of Information on Health and Epidemiology), and Benigno Rosón Calvo (general deputy director of the SERGAS information system) for providing the COVID-19 data.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

References

1. Peng, Y.; Yu, B. *Cure Models: Methods, Applications, and Implementation*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2021.
2. Safari, W.C.; López-de-Ullibarri, I.; Jácome, M.A. A product-limit estimator of the conditional survival function when cure status is partially known. *Biometr. J.* **2021**, *63*, 984–1005. [CrossRef] [PubMed]
3. Safari, W.C.; López-de-Ullibarri, I.; Jácome, M.A. Nonparametric Kernel Estimation of the Probability of Cure in a Mixture Cure Model When the Cure Status Is Partially Observed. Submitted. 2021. Available online: https://dm.udc.es/preprint/main_paper_cure_rate_Safari_et_al.pdf (accessed on 29 September 2021).