proceedings

MDPI

Extended Abstract

Nonparametric Inference in Mixture Cure Models [†]

Ana López-Cheda [1,*], Ricardo Cao [1], Mª Amalia Jácome [1] and Ingrid Van Keilegom [2]

[1] Department of Mathematics, University of A Coruña, 15071 A Coruña, Spain; rcao@udc.es (R.C.); majacome@udc.es (M.A.J.)

[2] ORSTAT, KU Leuven, 3000 Leuven, Belgium; ingrid.vankeilegom@kuleuven.be

[*] Correspondence: ana.lopez.cheda@udc.es; Tel.: +34-981-167-000 (ext. 1301)

[†] Presented at the XoveTIC Congress, A Coruña, Spain, 27–28 September 2018.

Published: 17 September 2018

Abstract: A completely nonparametric method for the estimation of mixture cure models is proposed. Nonparametric estimators for the cure probability (incidence) and for the survival function of the uncured population (latency) are introduced. In addition, a bootstrap bandwidth selection method for each nonparametric estimator is considered. The methodology is applied to a dataset of colorectal cancer patients from the University Hospital of A Coruña (CHUAC). Furthermore, a nonparametric covariate significance test for the incidence is proposed. The test is extended to non-continuous covariates: binary, discrete and qualitative, and also to contexts with a large number of covariates. The method is applied to a sarcomas dataset from the University Hospital of Santiago (CHUS).

Keywords: bandwidth selection; bootstrap; censored data; kernel estimation; survival analysis

1. Introduction

In the last two decades there has been a remarkable progress in cancer treatments, which led to longer patient survival and improved their quality of life. Consequently, a spate of statistical research to develop cure models arose. These models are useful tools to analyze and describe survival data with long-term survivors, since they express and predict the prognosis of a patient considering, as a novelty, the real possibility that the subject may never experience the event of interest. Cure models allow to estimate the cured proportion, $1 - p(x)$, and also the probability of survival of the uncured patients up to a given time point, or latency, $S_0(t|x)$. In the literature, ref. [1] proposed the nonparametric incidence estimator: $1 - \hat{p}_h(x) = \hat{S}_h(T^1_{\max}|x)$, where $\hat{S}_h()$ is the conditional Kaplan-Meier estimator with bandwidth h, and $T^1_{\max}$ is the largest uncensored failure time. The first completely nonparametric approach in mixture cure models was proposed by [2], who introduced the nonparametric latency estimator: $\hat{S}_{0,b}(t|x) = \frac{\hat{S}_b(t|x)-(1-\hat{p}_b(x))}{\hat{p}_b(x)}$, studied in detail by [3]. Furthermore, in cancer studies it is interesting to test if a covariate has some influence on the cure rate or on the survival time of the susceptible patients. Since no significance testing has been proposed yet for nonparametric cure models, this important gap is filled with the proposal of a covariate significance test for the incidence. This test allows to identify which covariates must be included in the incidence in a mixture cure model. Following [4], the proposed statistics is based on the process:

$$T_n(z) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\eta}_i - \left(\frac{1}{n}\sum_{j=1}^{n}\hat{\eta}_j\right)\right) I\left(Z_i \le z\right),$$

where n is the sample size, $\hat{\eta}_i$ is an estimator of the cure indicator for each individual, and Z is the covariate. Possible test statistics are the Cramér-von Mises (CvM) or the Kolmogorov-Smirnov (KS) tests. Moreover, the test statistic null distribution is approximated by bootstrap, using an independent naive resampling. For the case with an m-dimensional covariate, $\mathbf{Z}$, the method consists of considering

m hypotheses in H_0 to be tested independently. In order to control the false discovery rate, the approach by [5] to problems of multiple significance testing is studied. In addition, to achieve the family wise error rate control, the conservative method by [6] is considered.

Application to Medical Data

The proposed methodology is applied to a dataset of 414 colorectal cancer patients from CHUAC. The goal is to estimate the cure rate as a function of the stage (from 1 to 4) and the age. The event of interest is the death due to colorectal cancer, and the censoring percentage is between 30.77% (Stage 4) and 70.97% (Stage 1). Figure S1 in the Supplementary Materials shows that the effect of the age on the cure rate changes with the stage. For example, in Stage 1, patients have a probability of survival between 0.25 and 0.65, depending on the age; whereas in Stage 3, for patients above 60, in a 10 years gap that probability decreases considerably from 0.4 to almost 0. The latency estimation for three specific ages is shown in Figure S2 in the Supplementary Materials. For Stages 1–2, the age does not seem to be determining for the survival of the uncured patients. On the contrary, for Stages 3–4, the latency estimation varies considerably depending on the age. For example, the probability that the follow-up time since the diagnostic until death is larger than 4.5 years is around 0.2 for patients with ages 35 and 50, whereas for 80 year old patients, that probability is larger than 0.4.

Moreover, a dataset related to patients with sarcomas, provided by CHUS, is studied. It consists of 261 observations with 372,420 covariates with information about DNA methylations and 32 covariates with clinical data. The event of interest is the death due to sarcomas, and a total of 195 observations are censored. Regarding the conservative method, the results show that only one covariate is significant for the cure rate: "Year of initial pathologic diagnosis". With respect to the non-conservative alternative, the results for $B = 10^5$ bootstrap resamples show that for the CvM statistic, there are 14,182 significant covariates and 650 non-conclusive covariates, which need to be considered again in the next iteration of the process. For the KS statistic, there are 12,411 significant covariates, and 608 non-conclusive covariates. The program is still running for $B = 10^6$ bootstrap resamples.

2. Discussion

Mixture cure models have been usually estimated using parametric or semiparametric methods. A completely nonparametric approach for the estimation in mixture cure models is introduced, and a nonparametric covariate significance test for the probability of cure in mixture cure models is proposed. The methodology, that can be applied to any type of covariates and to high dimensional datasets, is illustrated with medical data. Specifically, the nonparametric incidence and latency estimators are applied to a dataset related to colorectal cancer patients from CHUAC. The incidence in Stages 1 and 2 is higher than in Stages 3 and 4 due to the fact that most of the surgeries in initial stages have healing purposes, whereas in advanced stages, surgeries are usually palliative treatments, and therefore the cure rate is lower. Furthermore, the latency estimation in Stages 3 and 4 is higher for 80 year old patients than for younger patients. The reason is that when a colorectal cancer is diagnosed in a young patient, it is usually in an advanced stage and with worse prognosis, since the cancer cells are more active in young individuals. Regarding the proposed covariate significance test for the incidence with the high dimensional dataset of sarcomas, the results differ for the conservative and the non-conservative approaches.

3. Materials

An R package is being developed with all the techniques proposed, including the implementation of the nonparametric incidence and latency estimators, as well as the covariate significance tests for different types of data: continuous, discrete, binary and qualitative, and for a high dimensional covariate vector. This R package will be uploaded in the Comprehensive R Archive Network (CRAN).

Supplementary Materials: The following are available online at http://www.mdpi.com/2504-3900/2/18/1181/s1. Figure S1: Nonparametric cure rate estimation for the different stages, Figure S2: Nonparametric latency estimation for the different stages.

Author Contributions: Conceptualization, R.C., M.A.J. and I.V.K.; Methodology, A.L.-C., R.C., M.A.J. and I.V.K.; Software, A.L.-C.; Validation, A.L.-C., R.C. and M.A.J.; Investigation, A.L.-C., R.C. and M.A.J.; Writing—Original Draft Preparation, A.L.-C.; Writing—Review & Editing, A.L.-C., R.C. and M.A.J.; Supervision, R.C. and M.A.J.

Funding: The first author was sponsored by the Spanish FPU (Formación de Profesorado Universitario) Grant from MECD (Ministerio de Educación, Cultura y Deporte) with reference FPU13/01371. The work has been partially carried out during two visits at the Université catholique de Louvain. The first stay was financed by INDITEX and the second one was supported by the research group MODES (Modelización, Optimización e Inferencia Estadística). All the authors acknowledge partial support by the MICINN (Ministerio de Ciencia e Innovación) Grant MTM2011-22392 and the MINECO (Ministerio de Economía y Competitividad) Grant MTM2014-52876-R. The first three authors acknowledge partial support of Xunta de Galicia (Grupos de Referencia Competitiva CN2012/130, ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF (European Regional Development Fund). Financial support from the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409) is gratefully acknowledged.

Acknowledgments: The authors are grateful to S. Pértega and S. Pita, at CHUAC, for providing the colorectal cancer dataset; and to A. Díaz-Lagares, at CHUS, for providing the sarcomas dataset.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Xu, J.; Peng, Y. Nonparametric cure rate estimation with covariates. *Can. J. Stat.* **2014**, *42*, 1–17. doi:10.1002/cjs.11197.
2. López-Cheda, A.; Cao, R.; Jácome, M.A.; Van Keilegom, I. Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Comput. Stat. Data Anal.* **2017**, *105*, 144–165. doi:10.1016/j.csda.2016.08.002.
3. López-Cheda, A.; Jácome, M.A.; Cao, R. Nonparametric latency estimation for mixture cure models. *Test* **2017**, *26*, 353–376. doi:10.1007/s11749-016-0515-1.
4. Delgado, M.A.; González-Manteiga, W. Significance testing in nonparametric regression based on the bootstrap. *Ann. Stat.* **2001**, *29*, 1469–1507. doi:10.1214/aos/1013203462.
5. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, *57*, 289–300. doi:10.2307/2346101.
6. Benjamini, Y.; Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **2001**, *29*, 1165–1188. doi:10.1214/aos/1013699998.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).