TEST (2023) 32:499–505
https://doi.org/10.1007/s11749-023-00856-z

DISCUSSION

Comments on: Nonparametric estimation in mixture cure models with covariates

Ricardo Cao[1]

Received: 8 March 2023 / Accepted: 8 March 2023 / Published online: 17 May 2023

Abstract
This paper discusses the invited paper by López-Cheda, Peng and Jácome on nonparametric mixture cure models with covariates. An alternative estimation procedure is proposed in this context. The situation when the two covariate vectors (the one in the incidence and in the latency parts) share some, but not all, their covariates is also considered. Some technical aspects in the assumptions, results and proofs of the invited paper are also discussed. Comments on the simulations and the real-data application are included. Finally, possible interesting topics for further research in this field are briefly discussed.

Keywords Censored data · Cure models · Nonparametric estimation

1 Introduction

The paper by López-Cheda, Peng and Jácome deals with a very interesting topic in Survival Analysis, namely nonparametric mixture cure models with covariates. The authors present a nice overview about the existing literature in the field and new methods to deal with the important practical case when the two key functions in the model (the incidence and the latency) depend on different vectors of covariates. An EM algorithm and a variation of it are proposed to deal with the estimation problem. Some theoretical results show the good behavior of the new methods. Their practical performance is studied in a simulation study, where some other existing methods are compared as well. The methods are illustrated by applying them to a data set concerning bankruptcy among commercial banks insured by the Federal Deposit Insurance Corporation[1].

[1] This comment refers to the invited paper available at https://doi.org/10.1007/s11749-022-00840-z.

✉ Ricardo Cao
rcao@udc.es

[1] Research Group MODES, Department of Mathematics, Research Center for Information and Communication Technologies (CITIC), Universidade da Coruña, Campus de Elviña, s/n, 15071 A Coruña, Spain

Section 2 in these comments deals with some alternative estimation procedure one can think of in this context. The situation when the two covariate vectors, $\mathbf{X}$ and $\mathbf{Z}$, share some, but not all, their covariates is considered in Sect. 3. Section 4 deals with some technical aspects in the assumptions, results or proofs. Comments on the simulations and the real-data application are collected in Sect. 5, while Sect. 6 briefly mentions possible interesting topics for further research in this field.

2 An alternative estimation procedure

Based on model (1) in the paper,

$$S(t|\mathbf{z},\mathbf{x}) = 1 - \pi(\mathbf{z}) + \pi(\mathbf{z})S_u(t|\mathbf{x}), \tag{1}$$

straightforward calculations give:

$$\lim_{t\to\infty} S(t|\mathbf{z},\mathbf{x}) = 1 - \pi(\mathbf{z}). \tag{2}$$

As a consequence, one may plug Beran's estimator, $\hat{S}^B(t|\mathbf{z},\mathbf{x})$, based on the entire set of covariates, $\mathbf{z}$ and $\mathbf{x}$, in (2) to produce an estimator of the incidence function:

$$\tilde{\pi}(\mathbf{z},\mathbf{x}) = 1 - \lim_{t\to\infty} \hat{S}^B(t|\mathbf{z},\mathbf{x}). \tag{3}$$

Although the previous estimator, $\tilde{\pi}(\mathbf{z},\mathbf{x})$, depends on both covariate vectors, the true incidence function, $\pi(\mathbf{z})$, only depends on the vector $\mathbf{z}$, so a reasonable way to produce a final estimator is just averaging expression (3) along $\mathbf{x}$:

$$\int \tilde{\pi}(\mathbf{z},\mathbf{x})dF_{\mathbf{X}}(\mathbf{x}). \tag{4}$$

Expression (4) cannot be used in practice since $F_{\mathbf{X}}$, the true distribution function of the covariate vector $\mathbf{X}$, is not known. However, an empirical version of it can be considered:

$$\hat{\pi}(\mathbf{z}) = \int \tilde{\pi}(\mathbf{z},\mathbf{x})dF_{\mathbf{X},n}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\pi}(\mathbf{z},\mathbf{X}_i), \tag{5}$$

which is a natural estimator of $\pi(\mathbf{z})$.

Once $\pi(\mathbf{z})$ has been estimated, one can solve for $S_u(t|\mathbf{x})$ in (1):

$$S_u(t|\mathbf{x}) = \frac{S(t|\mathbf{z},\mathbf{x}) - 1 + \pi(\mathbf{z})}{\pi(\mathbf{z})}, \tag{6}$$

plug in (6) Beran's estimator, $\hat{S}^B(t|\mathbf{z}, \mathbf{x})$, for $S(t|\mathbf{z}, \mathbf{x})$, and use (5) to obtain

$$\tilde{S}_u(t|\mathbf{z}, \mathbf{x}) = \frac{\hat{S}^B(t|\mathbf{z}, \mathbf{x}) - 1 + \hat{\pi}(\mathbf{z})}{\hat{\pi}(\mathbf{z})}. \tag{7}$$

Once again, although the estimator in (7) depends on $\mathbf{z}$ and $\mathbf{x}$, the true latency function, $S_u(t|\mathbf{x})$, does not depend on the vector $\mathbf{z}$, so a reasonable modification of it is

$$\int \tilde{S}_u(t|\mathbf{z}, \mathbf{x}) dF_{\mathbf{Z}}(\mathbf{z}). \tag{8}$$

Finally, an empirical version of (8) is

$$\hat{S}_u(t|\mathbf{x}) = \int \tilde{S}_u(t|\mathbf{z}, \mathbf{x}) dF_{\mathbf{Z},\mathbf{n}}(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{S}_u(t|\mathbf{Z}_i, \mathbf{x}), \tag{9}$$

which is a natural estimator for $S_u(t|\mathbf{x})$.

It would be interesting to explore the comparative behavior of (5) and (9) with respect to the EM estimator proposed by López-Cheda, Peng and Jácome.

3 X and Z sharing some components

The cases where $\mathbf{X}$ and $\mathbf{Z}$ are the same covariate vectors and where they do not have any component in common are analyzed in Subsections 3.1 and 3.2 of the paper. The situation where $\mathbf{X}$ and $\mathbf{Z}$ share some, but not all, their components is only mentioned in the introduction. This is probably the reason why the estimators proposed in Subsections 3.1 and 3.2 are the ones considered by the authors in the real-data application. However, it is often the case that the two covariate vectors may share some components. This could be the case of the real data application considering $\mathbf{Z} = $ (COREDEP, ROA) and $\mathbf{X} = $ (LOANS+, ROA), according to the p-values found. So it would be nice to propose practical estimators that cover this sharing-some-but-not-all-component setup.

Just to fix the notation, let us assume that there are three subvectors of disjoint covariates: $\mathbf{V}_1$, $\mathbf{V}_2$ and $\mathbf{V}_3$, such that $\mathbf{V}_1$ represents the covariates included in $\mathbf{Z}$ but not in $\mathbf{X}$, $\mathbf{V}_2$ accounts for the shared covariates of $\mathbf{Z}$ and $\mathbf{X}$, and $\mathbf{V}_3$ includes the covariates in $\mathbf{X}$ but not in $\mathbf{Z}$. Mathematically, $\mathbf{Z}^t = (\mathbf{V}_1^t, \mathbf{V}_2^t)$ and $\mathbf{X}^t = (\mathbf{V}_2^t, \mathbf{V}_3^t)$.

With the previous notation, Model (1) can be written as follows:

$$S(t|\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_2) = 1 - \pi(\mathbf{v}_1, \mathbf{v}_2) + \pi(\mathbf{v}_1, \mathbf{v}_2) S_u(t|\mathbf{v}_2, \mathbf{v}_3). \tag{10}$$

Now considering the joint covariate vector, $\mathbf{J}^t = (\mathbf{V}_1^t, \mathbf{V}_2^t, \mathbf{V}_3^t)$, Beran's estimator, $\hat{S}^B(t|\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$, for the covariate vector $\mathbf{J}$, and parallel arguments to those presented in Sect. 2 in these comments, can be used to define

$$\tilde{\pi}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = 1 - \lim_{t\to\infty} \hat{S}^B(t|\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3),$$

$$\hat{\pi}(\mathbf{v}_1, \mathbf{v}_2) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\pi}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{V}_{3,i}), \tag{11}$$

$$\tilde{S}_u(t|\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = \frac{\hat{S}^B(t|\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) - 1 + \hat{\pi}(\mathbf{v}_1, \mathbf{v}_2)}{\hat{\pi}(\mathbf{v}_1, \mathbf{v}_2)},$$

$$\hat{S}_u(t|\mathbf{v}_2, \mathbf{v}_3) = \frac{1}{n}\sum_{i=1}^{n} \tilde{S}_u(t|\mathbf{V}_{1,i}, \mathbf{v}_2, \mathbf{v}_3). \tag{12}$$

So the estimators presented in (11) and (12) can be readily used for the incidence and the latency in this sharing-some-but-not-all-component setting.

4 Technical aspects

A few technical details are considered in this section. These are related to dependence/independence for the data, including conditions on the smoothing parameter to kill the bias, the assumptions listed in Sect. 4 of the paper and in the statement of Theorem 2.

4.1 Dependence/independence setting

In the first lines of Sect. 2 in the paper, the notation for the data is introduced. A sample of n observations, $(\tilde{t}_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, for $i = 1, \ldots, n$ is assumed to be collected. These are n identical realizations of the underlying random variable $(\tilde{T}, \delta, \mathbf{X}, \mathbf{Z})$, but nothing is mentioned about the independence (*iid* setting) or dependence (*did* setting) of these observations. After the list of assumptions, in Sect. 4 of the paper, a comment on an α-mixing type of condition is included, but the data are not assumed α-mixing in Sect. 2. Some *iid* or *did* assumption should be included in Sect. 2. If the data are assumed to be dependent and identically distributed (*did*), some intuition should be given about why the dependence is expected among observations of the lifetime, censoring indicator and covariate vectors in real data sets, as, for instance, the bankruptcy data studied in Section 7 of the paper. In the *did* setting, α-mixing-type conditions are expected to produce a larger variance of the estimators with respect to the *iid* case. However, this seems not to be the case in view of the asymptotic normality result presented in Sect. 2 of the paper. I believe this deserves some comments by the authors.

4.2 Killing the asymptotic variance

The asymptotic normal distribution for the convergence in distribution result in Sect. 2 in the paper has zero mean. This implies that there is no asymptotic contribution coming from the bias of the nonparametric estimator, $\hat{\pi}_h(z)$. A typical condition for this to hold in other settings is that one chooses the smoothing parameter, h, in such a way that the asymptotic bias is killed, i.e., $nh^5 \to 0$ when $n \to \infty$. However, the only conditions on the bandwidth assumed by the authors for this asymptotic normality result are $h \to 0$, $nh/\log n \to \infty$ and $nh^2 \to \infty$ when $n \to \infty$. This is a bit intriguing and it deserves some intuitive explanation about why the condition $nh^5 \to 0$ when $n \to \infty$ is not needed in this setting.

4.3 Assumptions needed for Theorem 2

Section 4 of the paper lists the assumptions needed for Theorem 2. A first paragraph introduces the basic notation and a reference to a vector $\mathbf{e}$ is made. Its components are assumed to be positive and *small*, but this small condition is not stated in a formal way. What does *small* mean here?

A neighborhood of $\mathbf{w}$, $U(\mathbf{w})$ is also mentioned in the introductory paragraph in Sect. 4. It is also mentioned in assumptions (A2') and (A3'). However, it is strange that no condition about how large this neighborhood can be is included in the assumptions. On the other hand, the neighborhood $U(\mathbf{w})$ does not appear in the results. Some explanation about the choice and the role of this neighborhood would be helpful.

In the statement of Theorem 2, some limit conditions on the bandwidths h_{1i} and h_{2i} are assumed: $(\log n)^{-1}nh_{1i}^q \to 0$, $(\log n)^{-1}nh_{2i}^p \to 0$, when $n \to \infty$ for $i = 1, \ldots, n$. However, these conditions have to be stated more carefully. In fact, the banwidths h_{1i} and h_{2i} depend also on n (they are $h_{1,i,n}$ and $h_{2,i,n}$) and since $i = 1, \ldots, n$ and $n \to \infty$, the bandwidths form two triangular arrays. Some questions arise. For instance, are the conditions above required uniformly in $i = 1, \ldots, n$ and then in the limit when $n \to \infty$? More insight is needed in my view.

5 Simulations and real-data application

Some comments about the estimators used in the simulations and the real-data application, as well as the covariate definition for the bankruptcy data, are included in this section.

5.1 NPSJJ estimator

In the context of the simulations and the real-data application (Sects. 6 and 7 in the paper), the estimator NPSXX is considered. This is a special case of Model (1) in the paper, where the covariates in the incidence and latency part coincide. However, when dealing with specific choices for the covariates $\mathbf{X}$ and $\mathbf{Z}$, as it is the case in the simulation study and the real-data application, considering just NPSXX (and not NPSZZ) as a simple competitor seems an arbitrary choice. In fact, a more reasonable competitor of NPSXZ in these two sections would be NPSJJ, where $\mathbf{J}$ denotes, using the same notation as in Sect. 3, the joint covariate vector, $\mathbf{J}^t = (\mathbf{V}_1^t, \mathbf{V}_2^t, \mathbf{V}_3^t)$, which reduces to $\mathbf{J}^t = (\mathbf{Z}^t, \mathbf{X}^t)$ when the two covariate vectors do not share components, as it is the case of Sects. 6 and 7 in the paper. Including NPSJJ in the simulation study and the real-data application would give a more fair comparison.

5.2 Time-dependent covariates

Since the covariates COREDEP, LOANS and ROA in the bankruptcy data set in Section 7 are time-dependent, the authors considered (as Beretta and Heuchene (2019) did) the average of them along the follow-up period to produce time-independent covariates. This is a reasonable way to proceed for an explanatory analysis, but it is not the right thing to do for predictive purposes. If one would like to produce real prediction about the bankrupt probability or the time until bankrupt for a given bank, the value of the covariates at the beginning of the follow-up period would be a more reasonable choice. It would be nice to see the results of the analysis when the time-dependent covariates are transformed to time-independent ones by just considering their initial values at the beginning of the follow-up period.

6 Further topics

As the authors pointed out in the paper, single-index models is a natural extension for cure models when the number of covariates is medium or large, since the curse of dimensionality becomes a real problem. In the context of Model (7) in the paper, it is reasonable to assume that the incidence and the latency functions depend on the covariate vector **X** via a unique projection for both parts of the model. Two different projections could be considered as well if we want to give freedom in the way the two functions depend on the original covariates. However, when Model (1) in the paper is considered, two different projections have to be considered for sure, since the two covariate vectors, **X** and **Z**, are different. They need not to have the same dimension. As a consequence, for Model (7) rather than considering a single-index model, a double-index model needs to be stated.

A relevant topic in cure models is covariate significance tests. This has been mentioned by the authors in the introduction. However, given the possible different covariate vectors in the incidence and latency part in Model (1), this topics becomes even more important for this setting. In fact, covariate significance tests can lead to cure models of the form (1), when one starts from Model (7) by just accepting that some covariates in the vector **X** become significant for the incidence part but not for the latency part or vice versa.

Acknowledgements This research has been supported by MICINN Grant PID2020-113578RBI00 and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01), all of them through the European Regional Development Fund (ERDF). Funding for open access charge: Universidade da Coruña/CISUG.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.