XXXVI National Conference on Statistics and Operations Research
Toledo, September 5th, 6th and 7th, 2016

Summary of the candidate for the Ramiro Melendreras award

Significance testing on nonparametric mixture cure models

Ana López-Cheda[1], M. Amalia Jácome [2] and Ricardo Cao[1]

[1]Grupo MODES, Dep. Matemáticas, INIBIC, Facultade de Informática, Universidade da Coruña
[2]Grupo MODES, Dep. Matemáticas, INIBIC, Facultade de Ciencias, Universidade da Coruña

1 Context

The context of the candidate's work is classified in nonparametric statistics for survival analysis and specifically, in cure models. A common assumption in survival analysis is that all of the subjects will eventually experience the event of interest if they are followed long enough. However, this assumption is becoming more unlikely to be true, for the remarkable progress in cancer treatments, which led to longer patient survival and increased the probability of cure. As a consecuence, often classical survival analysis is not realistic and a new approach is necessary. Consequently, there has been an increasing interest in cure models over the past few years.

Cure model techniques allow modelling and predicting the prognosis of a patient considering, as a novelty, the possibility that the individual is cured. The goal in cure models is to estimate the cure probability of the population (incidence) and the survival function of the susceptible individuals (latency). The existing methods are limited to parametric and semiparametric models, which require some assumptions on the underlying distribution of the data. Nonparametric methods provide an alternative that does not require assumptions on the data. The main contribution of this work is to introduce a completely nonparametric approach with covariates in this context.

Completely nonparametric estimators of the probability of cure and the survival function of the uncured population are proposed and their asymptotic properties are studied, depending on a set of covariates. Besides, a covariate significance test is proposed. The methods are used to analyze a data base related to colorectal cancer patients from CHUAC (Complejo Hospitalario Universitario de A Coruña).

2 State of the art

There are two major types of cure models: mixture and non-mixture cure models. The first papers in non-mixture models were due to [Haybittle, 1959] and [Haybittle, 1965]. One category, belonging to this group, is the proportional hazards (PH) cure model, also known as the promotion time cure model, first proposed by [Yakovlev et al., 1994]. In this work, mixture cure models are studied. They were introduced by [Boag, 1949] and they consider the survival function as a mixture of those of two groups of subjects: those who are cured and those who are not. An important benefit of this model is that it allows the covariates to have different influence on cured and susceptible patients. In mixture cure models, the incidence is usually assumed to have a logistic form and the latency is usually estimated parametrically ([Farewell, 1982], [Farewell, 1986], [Cantor and Shuster, 1992]) or semiparametrically ([Kuk and Chen, 1992], [Peng et al., 1998][Peng and Dear, 2000], [Sy and Taylor, 2000] and [Li and Taylor, 2002]).

Due to the fact that the effects of the covariate on the cure rate cannot always be well approximated using parametric or semiparametric methods, a nonparametric approach is needed. To the present, some nonparametric methods for cure rate estimation have been studied: [Maller and Zhou, 1992] proposed a consistent nonparametric estimator of the incidence, but it cannot handle covariates. In order to overcome this drawback, [Laska and Meisner, 1992] proposed another nonparametric estimator of the cure rate, but it only works for discrete covariates. Furthermore, [Wang et al., 2012] proposed a cure model with a nonparametric form in the cure probability. To ensure model identifiability, they assumed a proportional hazards model for the hazard function. The estimation was carried out by an expectation-maximization algorithm for a penalized likelihood. They defined the smoothing spline function estimates as the minimizers of the penalized likelihood. More recently, [Xu and Peng, 2014] extended the existing work by proposing a nonparametric incidence estimator which allows for a continuous covariate.

Although the above papers have a nonparametric flavor, they fail to consider a completely nonparametric mixture cure model which works for discrete and continuous covariates in both the incidence and the latency. [López-Cheda et al., 2016a] proposed a completely nonparametric mixture cure model, with nonparametric estimators for both the cure probability and the survival function of the uncured individuals. Asymptotic properties and bandwidth selection procedures for the incidence and latency estimators were addressed in [López-Cheda et al., 2016a] and [López-Cheda et al., 2016b], respectively. No significance testing has been proposed yet. In this contribution, this important gap is filled by proposing a covariate significance test for the incidence. Although two cases are considered under the null distribution (with no covariates and with only one covariate), the method can be easily extended to a case with multiple covariates.

3 Main results

Let Y be the time to occurrence of the event (for example, death), X be a set of covariates and ν a binary variable where $\nu = 0$ (or $Y < \infty$) indicates that the individual belongs to the susceptible group and $\nu = 1$ (or $Y = \infty$) indicates that the subject is cured. The conditional probability of not being cured is $p(\mathbf{x}) = P(\nu = 0|\mathbf{X} = \mathbf{x})$ and the conditional survival function of Y is $S_0(t|\mathbf{x}) = P(Y > t|\mathbf{X} = \mathbf{x}, \nu = 0)$. Then, the mixture cure model can be written as:

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

where $1-p(\mathbf{x})$ is the incidence and $S_0(t|\mathbf{x})$ is the latency. We assume that each individual is subject to random right censoring and that the censoring time, C, is independent of Y given the covariates $\mathbf{X}$. Let $T = \min(Y, C)$ be the observed time and $\delta = \mathbf{1}(Y \leq C)$ the uncensoring indicator. Observe that $\delta = 0$ for all the cured patients. This also happens for uncured patients with censored lifetime. The sample is denoted by $\{(X_i, T_i, \delta_i), i = 1, \ldots, n\}$, independent and identically distributed (iid) copies of the random vector (X, T, δ).

To estimate the conditional survival function with covariates, the generalized Kaplan-Meier estimator by [Beran, 1981] is considered:

$$\hat{S}_h(t|x) = \prod_{T_{(i)} \leq t} \left(1 - \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right), \tag{2}$$

where

$$B_{h(i)}(x) = \frac{K_h(x - X_{(i)})}{\sum_{j=1}^{n} K_h(x - X_{(j)})}$$

are the Nadaraya-Watson (NW) weights with $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$, the rescaled kernel, with bandwidth $h \to 0$, $T_{(i)}$ is the i-th ordered statistic for the T-sample and $X_{(i)}$ and $\delta_{(i)}$ are the corresponding i-th concomitants in the X and the δ samples. [Xu and Peng, 2014] introduced the following kernel incidence estimator:

$$1 - \hat{p}_h(x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right) = \hat{S}_h(T^1_{\max}|x), \tag{3}$$

where $T^1_{\max} = \max_{i:\delta_i=1}(T_i)$ is the largest uncensored failure time.

Using (1), [López-Cheda et al., 2016a] proposed the following nonparametric latency estimator:

$$\hat{S}_{0,h}(t|x) = \frac{\hat{S}_h(t|x) - (1 - \hat{p}_h(x))}{\hat{p}_h(x)},$$

where $\hat{S}_h(t|x)$ is the Beran estimator of $S(t|x)$ in (2) and $1 - \hat{p}_h(x)$ is the estimator by [Xu and Peng, 2014] in (3).

It is interesting to test if a covariate has some influence on the cure rate or on the survival time of the susceptible patients. Let $\mathbf{W} = (\mathbf{X}, \mathbf{Z}) = (X_1, ..., X_q, Z_1, .., Z_p)$ be the explanatory covariates. A test (which is based on the significance test by [Delgado and González-Manteiga, 2001]) is proposed to study if the cure probability, as a function of the covariate vector $\mathbf{W}$, only depends on $\mathbf{X}$, but not on $\mathbf{Z}$:

$$H_0 : E(\nu|\mathbf{X}, \mathbf{Z}) = 1 - p(\mathbf{X}).$$

For $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$ the three cases in Table 1 can be considered. For the sake of brevity, only the first two will be studied here, the third one will be addressed in a future work:

	Dimension $\mathbf{W}$	Dimension $\mathbf{X}$	Dimension $\mathbf{Z}$
Case 1: $H_0 : 1 - p(z) = 1 - p$	1	0	1
Case 2: $H_0 : 1 - p(x, \mathbf{z}) = 1 - p(x)$	$1 + p$	1	p
Case 3: $H_0 : 1 - p(\mathbf{x}, \mathbf{z}) = 1 - p(\mathbf{x})$	$q + p$	q	p

Table 1: Different cases of covariate significant testing.

4 Contributions

The work of [Xu and Peng, 2014] for the incidence is extended and, to cover the latency function, the following theoretical results have been obtained ([López-Cheda et al., 2016a],[López-Cheda et al., 2016b]):

- The identifiability of the nonparametric mixture cure model.
- The incidence and latency nonparametric estimators are proved to be the local maximum likelihood estimators.
- An iid representation for each nonparametric estimator.
- The asymptotic mean squared error and the asymptotic normality for the latency estimator.
- Two bootstrap bandwidth selection methods, one for each nonparametric estimator.

The efficiency of the nonparametric incidence and latency estimators is assessed in a simulation study, in which the advantages of both estimators over the semiparametric approach by [Peng and Dear, 2000] is shown. The proposed covariate significance test is assessed in a Monte Carlo simulation study, in which the distribution of the test is approximated using the bootstrap. All the proposed methods are applied to a database of colorectal cancer patients from CHUAC.

In summary, this work addresses cure models, a field of growing interest in applied statistics, as a novelty, with a **completely nonparametric approach**. This enables to carry out a flexible analysis without imposing parametric constraints. **Theoretical results** have been proved, and the good behavior of the proposed estimators and significance test was shown in **computationally extensive simulation studies**.

The **applicability** of the methods for clinical studies was illustrated with real data. These nonparametric estimators are the basis for the proposal of many other nonparametric interesting methods such as comparison and goodness of fit tests.

References

[Beran, 1981] Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley.

[Boag, 1949] Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Stat. Soc. B Met.*, 11:15–53.

[Cantor and Shuster, 1992] Cantor, A. B. and Shuster, J. J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Stat. Med.*, 11:931–937.

[Delgado and González-Manteiga, 2001] Delgado, M. A. and González-Manteiga, W. (2001). Significance testing in nonparametric regression based on the bootstrap. *Ann. Stat.*, 29:1469–1507.

[Farewell, 1982] Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041–1046.

[Farewell, 1986] Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Can. J. Stat.*, 14:257–262.

[Haybittle, 1959] Haybittle, J. L. (1959). The estimation of the proportion of patients cured after treatment for cancer of the breast. *Brit. J. Radiol.*, 32:725–733.

[Haybittle, 1965] Haybittle, J. L. (1965). A two-parameter model for the survival curve of treated cancer patients. *J. Am. Stat. Assoc.*, 60:16–26.

[Kuk and Chen, 1992] Kuk, A. Y. C. and Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79:531–541.

[Laska and Meisner, 1992] Laska, E. M. and Meisner, M. J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, 48:1223–1234.

[Li and Taylor, 2002] Li, C. and Taylor, J. M. G. (2002). A semi-parametric accelerated failure time cure model. *Stat. Med.*, 21:3235–3247.

[López-Cheda et al., 2016a] López-Cheda, A., Cao, R., Jácome, M. A., and Van Keilegom, I. (2016a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. Submitted paper.

[López-Cheda et al., 2016b] López-Cheda, A., Jácome, M. A., and Cao, R. (2016b). Nonparametric latency estimation for mixture cure models. Submitted paper.

[Maller and Zhou, 1992] Maller, R. A. and Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, 79:731–739.

[Peng and Dear, 2000] Peng, Y. and Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56:237–243.

[Peng et al., 1998] Peng, Y., Dear, K. B., and Denham, J. W. (1998). A generalized F mixture model for cure rate estimation. *Stat. Med.*, 17:813–830.

[Sy and Taylor, 2000] Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox Proportional Hazards cure model. *Biometrics*, 56:227–236.

[Wang et al., 2012] Wang, L., Du, P., and Lian, H. (2012). Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*, 68:726–735.

[Xu and Peng, 2014] Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Can. J. Stat.*, 42:1–17.

[Yakovlev et al., 1994] Yakovlev, A. Y., Cantor, A. B., and Shuster, J. J. (1994). Parametric versus nonparametric methods for estimating cure rates based on censored survival data. *Stat. Med.*, 13:983–986.