Estimating the mixture cure model when the cure is partially observed with the npcurePK Package

This version was compiled on May 5, 2023

This document provides a short overview of the npcurePK package.

Introduction

Survival analysis studies the duration of time Y until one event occurs, mainly through the survival function $S(t) = P(Y > t)$. Cure models are a class of time-to-event models where a proportion of individuals will never experience the event of interest, so they are considered *cured* ($Y = \infty$). The mixture cure model (MCM) assumes that the population is a mixture of cured and susceptible individuals (Peng and Yu, 2021). So in a conditional setting with a covariate X, the conditional survival function $S(t \mid x) = P(Y > t \mid X = x)$ can be written as

$$S(t \mid x) = 1 - p(x) + p(x)S_0(t \mid x)$$

where $1 - p(x) = P(Y = \infty \mid X = x)$ is the probability of cure, and $S_0(t \mid x) = P(Y > t \mid Y < \infty, X = x)$ is the conditional survival function of the uncured individuals.

The lifetimes of the so-called cured individuals are always censored. It is usually assumed that one never knows which censored observation are cured and which are uncured, so the cure status is unknown for censored times. The absence of a censored individual's cure status (i.e., cured, uncured) is an important challenge for cure models. It is customary to assume no additional information on the cure status of censored individuals, thus, to model the cure status as a latent variable.

Nonetheless, there are situations where some of the censored individuals can be identified to be immune to the event of interest, that is, to be cured. For example, diagnostic procedures in medical studies can provide further information on whether a subject will not die from a curable illness. Also, for some types of cancer, it is extremely unlikely to have any recurrence later than a given time after treatment, known as cure threshold. In these situations, there are three groups of observations: the *event* times of individuals experiencing the event during the follow-up time; the *regular censored* times of those who neither experienced the event nor were classified as cured; and a new third group, the *cured* (censored) times of those acknowledged as cured from the event. Just modeling the data under the usual cure model framework, that considers the *cured* times as simple *regular censored* times, will not take advantage of this additional cure status information given by the third group.

Few authors have studied cure models from a nonparametric point of view when the cure status is known for some censored observations. Nonparametric cure probability estimation with random cure status partially available was discussed without covariates by Laska and Meisner (1982) when cure is observed based on a cure threshold, and Betensky and Schoenfeld (2001) with random observed cures using a competing risk approach. In a conditional setting with covariates, kernel estimators of the survival function, cure probability and latency functions have been proposed in Safari *et al.* (2021), Safari *et al.* (2022) and Safari *et al.* (2023).

npcurePK

Overview of the package. `npcurePK` is an `R` package that implements the estimators of the survival function, latency function and probability of cure in a mixture cure model when the cure is partially observed (Safari *et al.*, 2021, 2022, 2023). These estimators are based on kernel smoothing ideas, using Nadaraya-Watson weights computed with a bandwidth h, that must be previously selected. In the absence of the optimal value for the bandwidth, a bandwidth selector based on the bootstrap is included in the package.

The package consist of three main `R` functions: `prodlim_curepk()` for the estimation of the conditional survival function $S(t \mid x)$, `prob_curepk()` for the estimation of the cure probability $1 - p(x)$, and `latency_curepk()` for the latency function $S_0(t \mid x)$. Its arguments are the observed covariate values `x`, the observed times `t`, the uncensoring indicator `d`, and the indicator of the individual known to be cured `xinu`. In addition, the argument `x0` specifies the value of the covariate where the functions are to be estimated.

The bandwidth h used in the Nadaraya-Watson weights can be set by the argument `h`. When the functions are estimated in a vector of values `x0`, then argument `local` states if the estimations in each value of `x0` are computed with the corresponding value of `h` (`local = TRUE`) of with all the values of `h` (`local = FALSE`).

If argument `h` is missing, then the bootstrap bandwidth selector for h is used instead. The list of parameters controlling the bootstrap when computing the bootstrap bandwidths includes the number of bootstrap resamples `B` (by default, `B = 100`), the fraction of the sample size that determines the order of the nearest neighbor used for choosing a pilot bandwidth `nnfrac` (by default, `nnfrac = 0.25`), the length of the grid where the bootstrap bandwidth is searched `hl` (by default, `hl = 30`), and the times `hbound` the standardized interquartile range of the covariate values are multiplied to get the search grid of bandwidths (by default, `hbound = c(0.1, 3)`). These default values are returned by the `controlpars()` function called without arguments.

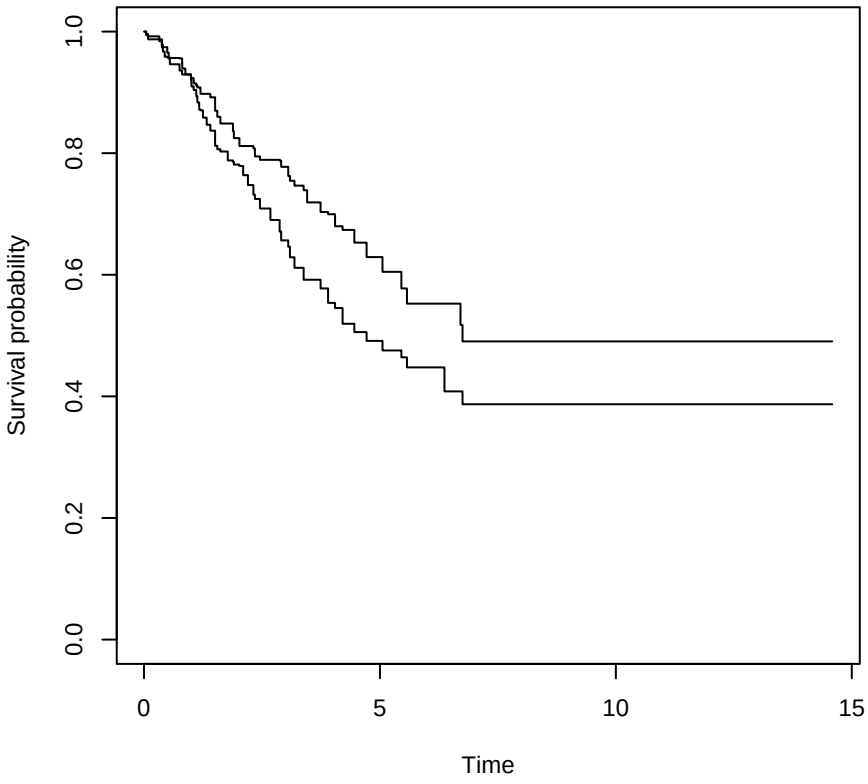
Using the package. For illustration purposes, the package includes the real data set `sarcoma` related to 232 patients diagnosed with sarcoma. Sarcoma is a rare type of cancer that represents 1% of all adult solid malignancies (Choy, 2014). If a tumor can be surgically removed to render the patient with sarcoma free of detectable disease, 5 years is the survival time at which sarcoma oncologists assume long-term remissions. Patients tumor free and alive for more than 5 years were assumed to be long-term survivors. The variables included in the data set are the age at diagnosis `x`, the observed time until death from sarcoma `t`, the censoring status `d` (`0` = censored, `1` = death from sarcoma) and the cure status `xinu` (`0` = dead or unknown, `1` = tumor free and alive for more than 5 years).

The following code illustrates a typical call:

```
library(npcurePK)
```

First, the survival function $S(t \mid x)$ is estimated for patients aged 40 and 90 years old, using the bootstrap bandwidth selector:

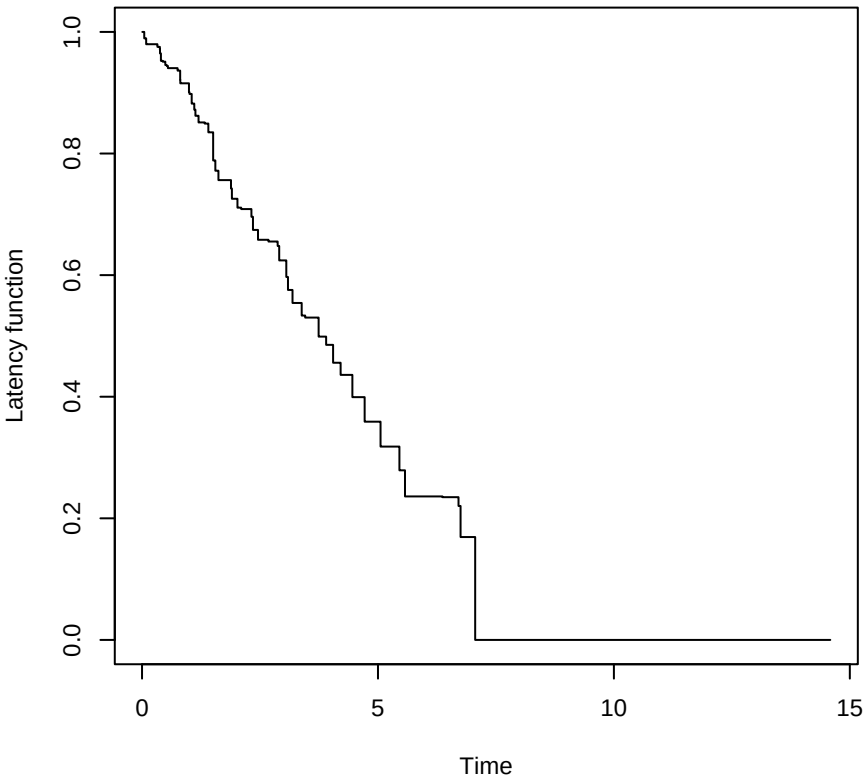
```
S <- prodlim_curepk(x, t, d, xinu, sarcoma, x0 = c(40, 90))
plot(S$t, S$surv[, 1], type = "s", xlab = "Time", ylab = "Survival probability", ylim = c(0, 1))
lines(S$t, S$surv[, 2], type = "s")
```



Next, the latency function $S_0(t \mid x)$ is computed for patients aged 60 years old. The computation of the latency estimator in the covariate value `x0` is based on suitable estimates of $1-p(x0)$ and $S(t \mid x0)$. As a consequence, this latency estimator requires a bivariate bandwidth `h` with dimension (`2 × length(x0)`), so `h[1, ]` is used for estimating $1-p(x)$ at `x0`, and `h[2, ]` is used for estimating $S(t \mid x)$ at `x0`.

In the next lines, the latency estimator is computed using a bootstrap bandwidth `h`. This bandwidth is searched using 2 cores, in a grid of 10 × 10 bandwidths (`hl = 10`) between 0.2 and 2 times the standardized interquartile range of the covariate values (`hbound = c(0.1, 2)`), using 50 bootstrap resamples (`b = 50`). The latency estimates are saved in an array of dimension ($n\times$ `length(x0)`).

```
S0 <- latency_curepk(x, t, d, xinu, sarcoma, x0 = 60,
                     bootpars = controlpars(b = 50, hl = 10, ncores = 2, hbound = c(0.1, 2)))
plot(S0$t, S0$latency[, 1], type = "s", xlab = "Time", ylab = "Latency function", ylim = c(0, 1))
```



Finally, the cure rate $1-p(x)$ is estimated for a vector of 50 values `x0` of the covariate age (X) ranged between the minimum and maximum observed age of the patients:

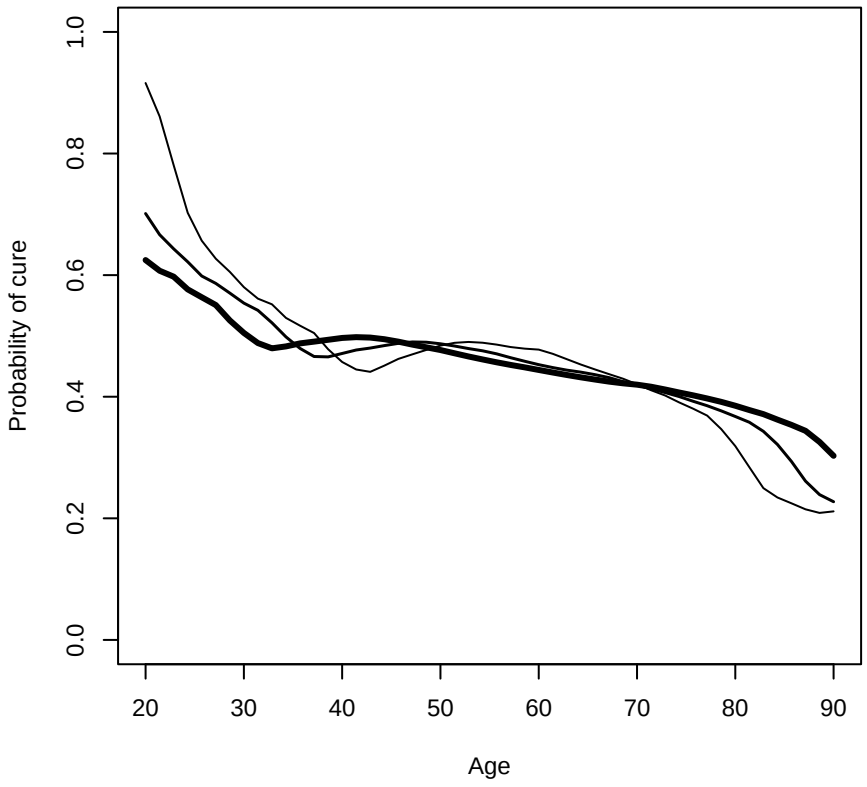
```
x0 <-  seq(from = min(sarcoma$x), to = max(sarcoma$x), length.out = 50)
```

If the estimation of the cure rate $1-p(x)$ in the values `x0` is performed with the same bandwidth h provided by the user for all the values in `x0`, then the bandwidth h is global (argument `local = FALSE`). The code below computes the cure rate estimator with a set of 3 different fixed global bandwidths `h = c(20, 25, 30)`:

```
p <- prob_curepk(x, t, d, xinu, sarcoma, x0 = x0, h = c(20, 25, 30), local = FALSE)
```

The estimated cure rates $1-p(x)$, evaluated in the 50 values of the vector `x0`, and computed with any of the 3 values for the bandwidth h, are given in a matrix `p$prob_cure` of dimension (3 × `length(x0)`). The estimates are represented with a line graph:

```
plot(p$x0, p$prob_cure[1, ], xlab = "Age", type = "l", ylab = "Probability of cure", ylim = c(0, 1))
lines(p$x0, p$prob_cure[2, ], lwd = 1.5)
lines(p$x0, p$prob_cure[3, ], lwd = 3)
```



More efficient results are expected if a different bandwidth h is used for the estimation of the cure rate $1-p(x)$ in each value of the vector `x0`. In this case, a vector of values for the bandwidth `h`, with the same length as `x0`, must be provided by the user, with argument `local = TRUE`.

The cure probability $1-p(x)$ can also be estimated in each value of the vector `x0` using a different bandwidth h for each value of `x0`, if the bandwidth is selected by the bootstrap. The following code illustrates the estimation using 2 cores, when the bootstrap is performed with `b = 50` bootstrap resamples (`seed` is only needed for repeatability):

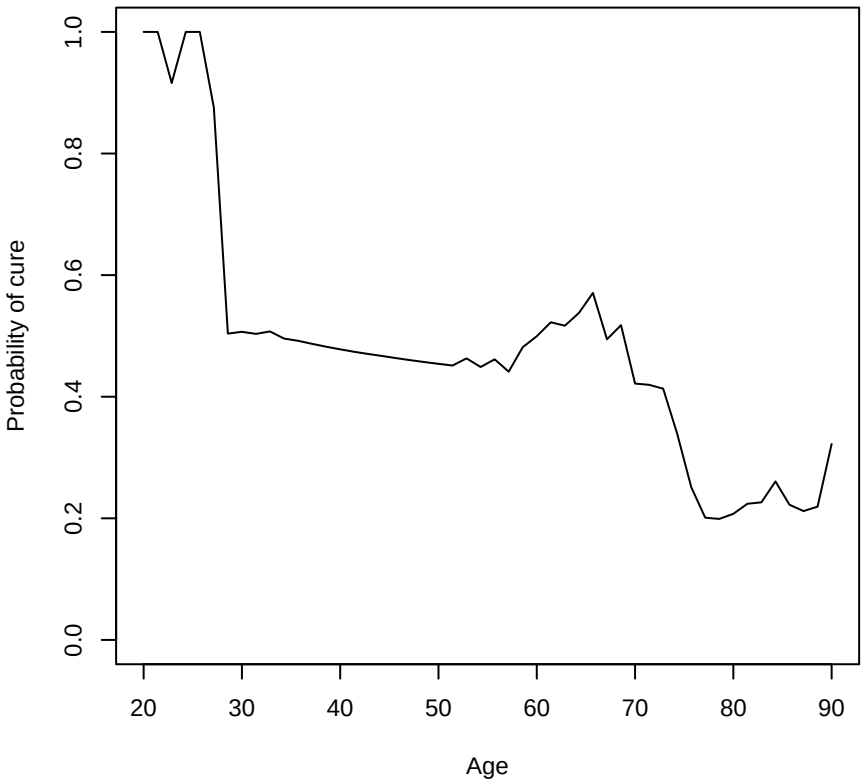
```
library(doParallel)
```

```
#  Loading required package: foreach
```

```
#  Loading required package: iterators
```

```
#  Loading required package: parallel
```

```
p2 <- prob_curepk(x, t, d, xinu, sarcoma, x0 = x0,
                  bootpars = controlpars(b = 50, ncores = 2, seed = 123))
plot(p2$x0, p2$prob_cure, xlab = "Age", type = "l", ylab = "Probability of cure", ylim = c(0, 1))
```



Final comments. When there are no individuals known to be cured (`xinu = 0`), then the usual kernel estimators of the survival function (Beran, 1981), the latency function (López-Cheda *et al.*, 2017b) and the cure rate (López-Cheda *et al.*, 2017a; Xu and Peng, 2014) are computed.

References

Beran R (1981). "Nonparametric regression with randomly censored survival data." *Technical Report. University of California, Berkeley*. URL https://www.jstor.org/stable/4616062.

Betensky R, Schoenfeld D (2001). "Nonparametric estimation in a cure model with random cure times." *Biometrics*, **57**(1), 282–286. doi:https://doi.org/10.1111/j.0006-341x.2001.00282.x.

Choy E (2014). "Sarcoma after 5 years of progression-free survival: Lessons from the French sarcoma group." *Cancer*, **120**(19), 2942–2943. doi:https://doi.org/10.1002/cncr.28834.

Laska EM, Meisner MJ (1982). "Nonparametric estimation and testing in a cure model." *Biometrics*, **48**(4), 1223–1234. doi:https://doi.org/10.2307/2532714.

López-Cheda A, Cao R, Jácome MA, Van Keilegom I (2017a). "Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models." *Computational Statistics & Data Analysis*, **105**, 144–165. doi:https://doi.org/10.1016/j.csda.2016.08.002.

López-Cheda A, Jácome MA, Cao R (2017b). "Nonparametric latency estimation for mixture cure models." *TEST*, **26**, 353–376. doi:https://doi.org/10.1007/s11749-016-0515-1.

Peng Y, Yu B (2021). *Cure models: methods, applications, and implementation*. Boca Raton, Chapman and Hall/CRC, Florida.

Safari WC, López-de-Ullibarri I, Jácome MA (2021). "A product-limit estimator of the conditional survival function when cure status is partially known." *Biometrical Journal*, **63**(5), 984–1005. doi:https://doi.org/10.1002/bimj.202000173.

Safari WC, López-de-Ullibarri I, Jácome MA (2022). "Nonparametric kernel estimation of the probability of cure in a mixture cure model when the cure status is partially observed." *Statistical Methods in Medical Research*, **31**(11), 2164–2188. doi:https://doi.org/10.1177/09622802221115880.

Safari WC, López-de-Ullibarri I, Jácome MA (2023). "Latency function estimation under the mixture cure model when the cure status is available." *Lifetime Data Analysis*. doi:https://doi.org/10.1007/s10985-023-09591-x.

Xu J, Peng Y (2014). "Nonparametric cure rate estimation with covariates." *Canadian Journal of Statistics*, **42**, 1–17. doi:https://doi.org/10.1002/cjs.11197.