R Package: IDPSurvival

Francesca Mangili Alessio Benavoli Marco Zaffalon
Cassio de Campos

November 1, 2024

Abstract

The purpose of this text is to provide a simple explanation about the main features of `IDPSurvival` package for `R` language. In short, we give some examples on how to use the package.

Keywords: IDPSurvival, R package, Imprecise Dirichlet Process, survival curves estimator, survival curves comparison, sum-rank test.

1 Introduction

We assume the reader to be familiar with reliability and/or survival estimation based on the Imprecise Dirichlet Process (IDP). For details we suggest the technical paper: *Imprecise Dirichlet Process for the estimate and comparison of survival functions with censored data*, Mangili et al. 2014.

The problem targeted here can be defined by `time`, an array of times of failure (survival) or components (patients) and `status` the event indicator, that is, `status` equals to one if an event happened at that time, or zero in case of right-censoring. Furthermore, covariates can be present in the data set, which can be used to distinguish different group of data.

Two main functions are included in the IDPSurvival package:

- `isurvfit`: compute the IDP estimator of the survival function for one or more groups of right censored data.
- `isurvdiff`: test the difference between the survival curves of two groups of right censored data.

To derive prosterior inferences about the distribution of the survival function $S(t)$ the IDP model uses a set $\mathcal{T}$ of Dirichlet process (DP) priors obtained by fixing the prior strenght s of the DPs in the set and letting their base measure vary in the set of all distributions. For each prior in $\mathcal{T}$ the method computes the posterior distribution conditioning on the set of right censored data. Posterior inferences are summarized by their upper and lower bounds. For example, when computing the posterior expectation of $S(t)$ in correspondence of each DP prior in $\mathcal{T}$, we obtain an infinite number of different values for $E[S(t)]$. We retain the inf and sup of this set of values as lower and upper bounds for the expectation of $S(t)$.

This document aims to give a simple explanation about the main features of the `isurvfit` `isurvdiff` functions. We refer to the man pages/help and the technical paper for all details about the arguments of these functions.

Before continuing, in case you have not yet done so, the first thing to do before using the functions is to install and load the library.

```
> install.packages("IDPSurvival_VERSION.tar.gz",
+                  repos=NULL,type="source") ## from local file
> install.packages("IDPSurvival")  ## or from CRAN

> library("IDPSurvival")
```

2 Survival Curve Estimator

First, we exemplify the use of the methods without covariates.

Example 1 In this example we simulate a data set with right-censored survival times. The problem regards an experiment with 30 individuals. For simplicity, we define the survival and censoring times by random variables with the exponential distribution.

```
> n <- 30
> lambda <- 5
> X <- rexp(n, rate = lambda) # sample lifetimes
> Y <- rexp(n, rate = lambda) # sample censoring times
> status <- (X<Y)*1
> time <- X*status+Y*(1-status)
> dataset <- cbind(time,status)
> dataset

             time status
 [1,] 0.021255901      0
 [2,] 0.090128496      0
 [3,] 0.005189240      0
 [4,] 0.028221280      0
 [5,] 0.123015973      1
 [6,] 0.184674279      1
 [7,] 0.151357718      0
 [8,] 0.026889316      1
 [9,] 0.109512342      1
[10,] 0.123222765      1
[11,] 0.043943310      0
[12,] 0.006583999      1
[13,] 0.048047766      1
[14,] 0.029426940      1
[15,] 0.077195854      0
[16,] 0.073051506      1
[17,] 0.131040374      1
[18,] 0.040986760      0
[19,] 0.106486486      0
[20,] 0.059435115      1
[21,] 0.008601294      0
[22,] 0.069097351      1
[23,] 0.290616056      1
[24,] 0.016622036      0
[25,] 0.214646595      0
[26,] 0.169934354      0
[27,] 0.081871651      1
[28,] 0.004542977      1
[29,] 0.171288946      1
[30,] 0.146465705      1
```

The function isurvfit can be use to compute and plot the survival curve for the data generated in example 1 based on the IDP model. In order to perform the estimation, following the common practice with other survival analysis packages, the user has to build a formula with time and censoring indication, which is then used to call the estimation method itself:

```
> formula <- Surv(dataset[,1],dataset[,2]) ~ 1
> fit <- isurvfit(formula, s=0.5,
+                  conf.int=0.95,display=FALSE)
> fit
```

```
  n.records n.events n.censored
```

```
1        30       17         13
```

The formula can be also defined using the variable names and specifying the data frame in which to interpret them.

```
> dataset <- data.frame(time,status)
> formula <- Surv(time,status) ~ 1
> fit <- isurvfit(formula,dataset,s=0.5,display=FALSE)
```

The upper and lower bounds of $E[S(t)]$ are stored in `fit$survUP` and `fit$survLOW`; the upper and lower bounds of the 0.95 confidence interval for $S(t)$ are stored in `fit$upper` and `fit$lower`. The value of s defines the strength of the DP prior: larger values of s increase the robustness but also the imprecision (i.e., the gap between the upper and lower bound of $E[S(t)]$).

Here, we compare the IDP with the non-parametric Kaplan-Meier estimator. The result is presented in the Figure 1.

```
> plot(fit)
> # Kaplan-Meier estimation
> library(survival)
> km <- survfit(formula,dataset)
> lines(km,col='red')
> legend('bottomleft',c("IDP","Kaplan-Meier"),lty=c(1,1),
+        col=c('black','red'),pch=c('o','.'))
```

In the following, we present a simple example on how to define different groups to be used with `IDPSurvival`. In fact, we use the same framework of formulas as other survival packages.

```
> # Running isurvfit on lung (from survival package) with
> # two groups: Male and Female
> data(lung,package='survival')
> formula <- Surv(time,status) ~ sex
> fit <- isurvfit(formula, lung)
> legend('topright',c("Male","Female"),
+        lty=c(1,1),col=c(1,2),pch=c(1,2))

> # three groups: ph.ecog = 0, 1, 2
> formula <- Surv(time,status) ~ ph.ecog
> sel =!is.na(match(lung$ph.ecog,c(0,1,2)))
> fit <- isurvfit(formula, lung, subset=sel)
```

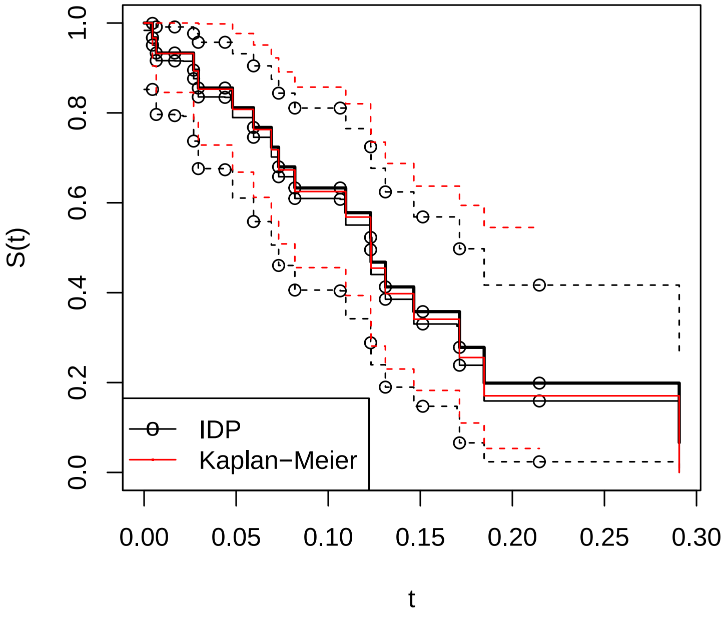

Figure 1: IDP estimator of the survival curve.

```
> legend('topright',names(fit$strata),
+        lty=rep(1,3),col=c(1:3),
+        pch=c(1:3),title='ECOG performance score')
```

The curves are shown in figure 2.

3 Curves comparison

We consider the survival time X and Y for two groups of individuals. Given samples with right censred data for X and Y, the survival curves of the two groups can be compared using the generalized IDP sum rank test implemented by the `isurvdiff` function. The IDP test evaluates posterior upper and lower bounds for the distributions of $P(X < Y)$. It can be one-sided (`alternative="greater"/"less"`) or two-sided (`alternative="two.sided"`).

Consider first the one-sided test with `alternative`"greater"=

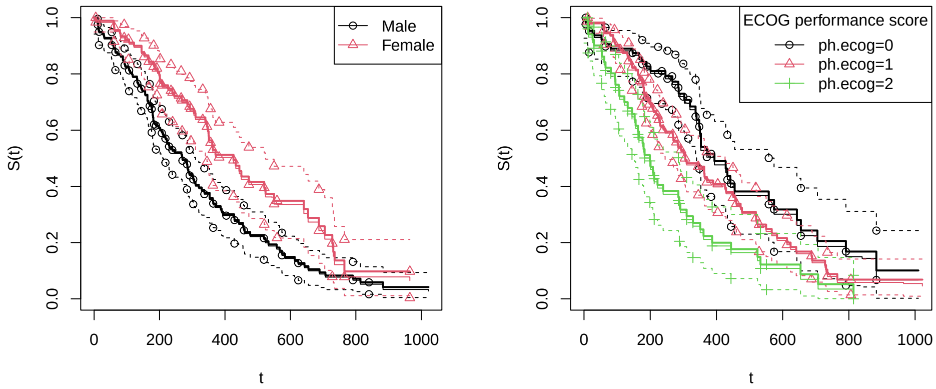

Figure 2: IDP survival curves on lung dataset.

```
> # Tests for the lung cancer dataset if male are
> # more likely to live less than females
> formula <- Surv(time,status) ~ sex
> test <- isurvdiff(formula, lung,
+                     alternative='greater',
+                     nsamples=100000)
> print(test)

---------------------------------------------
Result of the IDP RANK-SUM hypothesis test
h = 1  --> Y is greater than X

Lower Probability of the hypothesis: 0.99978
Upper Probability of the hypothesis: 0.99995

Lower Central credible intervals
[ 0.569 , 0.721 ]
Upper Central credible intervals
[ 0.576 , 0.728 ]
---------------------------------------------
```

By default the function takes as samples for X and Y the data with covariate (at right hand side of ~) equal to, respectively, 1 and 2 (in the lung dataset 1 corresponds to male and 2 to female). Diferent covariate values for the indentification of the groups can be defined using the argument *group*. An example is given in the discussion of the two-sided case. The argument nsamples determines how many Monte Carlo samples are generated to approximate the posterior distributions of $P(X < Y)$.

The test evaluates the posterior upper and lower bounds for the probability that $P(X < Y) > 1/2$ and compares them with the desired value specified by the parameter level (by default equal to 0.95). In the contex of decision making, said K1 and K2 the costs of type I and type II errors, one can minimize the expected loss by choosing level=K1/(K1+K2). If the lower pobability of $P(X < Y) > 1/2$ is larger than 0.95 we can state that *Y is better than X* with probability larger than 0.95 and thus accept the hypothesis. The test returns H=1. If the upper probability is lower than 0.95 we can say that the the probability that *Y is better than X* is smaller than the desired level; the hypothesis is not accepted and test returns H=0. If 0.95 falls between the lowr and upper probabilities, then the decision that minimizes the expected loss depend on the choice of the prior and thus a robut decision cannot be made; the test returns H=2.

The test also provides the plot (figure 3) of the posterior upper and lower distributions of $P(X < Y)$.

The two-sided test only tells if the lifetimes for the two groups are significantly different, i.e., $P(X < Y) \neq P(Y < X)$, without focusing on a single direction. In the next example, we test if the lifetimes X and Y of patients in the lung dataset with ECOG performance score ph.ecog=0 and ph.ecog=1 are diferent.

```
> # Tests for the lung cancer dataset if male are more likely
> # to live less than females
> formula <- Surv(time,status) ~ ph.ecog
> test <- isurvdiff(formula, lung, groups=c(0,1),
+                  alternative='two.sided',
+                  level =0.95, exact=FALSE)
> test

--------------------------------------------
Result of the IDP RANK-SUM hypothesis test
h = 0  --> Y and X are not different


Lower Central credible intervals
```

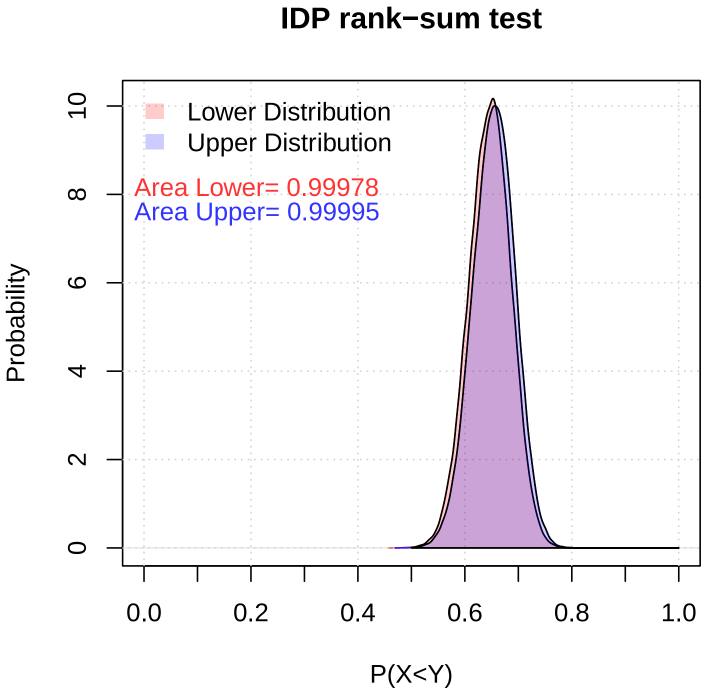

Figure 3: IDP posterior distribution of $P(X < Y)$.

```
[ 0.31 , 0.502 ]
Upper Central credible intervals
[ 0.323 , 0.514 ]
---------------------------------------------
```

The argument *exact* determines whether computing the exact posterior distribution of $P(X < Y)$ by Monte Carlo sampling (`TRUE`), or using the Gaussian approximation (`FALSE`). Clearly, the latter choice is computationally faster but also less accurate if the samples size is small.

The IDP test considers the upper and lower Highest Posterior Density (HPD) credible intervals at the specified level (`level=0.95`) and verifies if they include the value of $1/2$ which corresponds to the null hypothesis $P(X < Y) = P(Y < X)$. In the example, the 0.95 credible intervals do not contain $1/2$; thus, with probability 0.95 either X *is better than* Y or Y *is better than* X and the test returns `H=1` (in practice, since both intervals lay on the right of $1/2$, we can infer that the credible hypothesis is Y *is better than* X). More precisely, the IDP test returns `H=1` if 1/2 is not included between the left bound of the lower and the right bound of the upper HPD credible intervals; it returns, instead, `H\0`= if 1/2 is included in both credible intervals, and `H=2` otherwise.

4 Remarks

This "manual" describes the basics of the IDPSurvival packag. We invite the user to the functions' help pages (available with the package) and to the technical paper mentioned in the beginning of this document for further details.