An R package for Determining Groups in Multiple Survival Curves

Nora M. Villanueva[1], Marta Sestelo[2], Luís Meira-Machado[3]

[1] Dep. Statistics and O.R., University of Vigo, Spain.
[2] SiDOR Research Group and CINBIO, University of Vigo, Spain.
[3] Centre of Molecular and Environmental Biology & Department of Mathematics and Applications, University of Minho, Portugal.

E-mail for correspondence: nmvillanueva@uvigo.es

Abstract: Survival analysis includes a wide variety of methods for analyzing time-to-event data. One basic but important goal in survival analysis is the comparison of survival curves between groups. Several nonparametric methods have been proposed in the literature to test for the equality of survival curves for censored data. When the null hypothesis of equality of curves is rejected, leading to the clear conclusion that at least one curve is different, it can be interesting to ascertain whether curves can be grouped or if all these curves are different from each other. We present the R clustcurv package which allows determining groups with an automatic selection of their number. The applicability of the proposed method is illustrated using real data.

Keywords: Log-rank Test; Multiple Survival Curves; Number of Groups; Survival Analysis

1   Introduction

Survival analysis includes a wide variety of methods for analyzing time-to-event data. One basic but important goal in survival analysis is the comparison of survival curves between groups. For example, in an observational survival study, one may be interested in comparing survival between individuals from different age groups, different genders, racial/ethnic groups, geographic localization, etc.
Several nonparametric methods have been proposed in the literature to test for the equality of survival curves for censored data. The log-rank or Mantel-Haenszel test (Mantel, 1966) is the most well-known and widely used to test the null hypothesis of no difference in survival between two or

more independent groups. An alternative test that is often used is the Peto & Peto (1972) modification of the Gehan-Wilcoxon test (Gehan, 1965).

Though the aforementioned methods can be used to compare multiple survival curves, methods that can be used to determine groups among a series of survival curves are not available, to the best of our knowledge. When the log-rank test (or its analogous) is used to compare three or more survival curves at once, the test reports a single p-value testing the null hypothesis that all the samples come from populations with identical survival. If the null hypothesis of equality of curves is rejected, then, this leads to the clear conclusion that at least one curve is different. However, these methods cannot be used to ascertain whether groups of curves can be performed or if all these curves are different from each other.

One naïve approach would be to perform pairwise comparisons. However, this approach would lead to a large number of comparisons (e.g. 7 groups would lead to 21 pairwise comparisons). One could make it but without the possibility of determining groups with similar survival curves. This can be achieved with the `pairwise_survdiff` of the package `survminer` (Kassambara and Kosinski, 2017) which calculates pairwise comparisons between group levels with corrections for multiple testing.Results for such a test can tell us that all combinations are different, or just one pair. However,as it was mentioned, when the number of curves increases so does the difficult of interpretation.

According to this, the paper introduces `clustcurv`, a software application for R which allows determining groups with an automatic selection of their number based on k-means or k-medians algorithms (Villanueva et al., submitted). It describes the capabilities of the package using a real dataset.

2   The clustcurv package in practice

To illustrate our method we will use one real dataset. It comes from a large clinical trial on Duke's stage III patients, affected by colon cancer, that underwent a curative surgery for colorectal cancer (Moertel et al., 1990). This data set is freely available as part of the R package `condSURV` (Meira-Machado and Sestelo, 2016). From the total of 929 patients, 452 died. For each individual, an indicator of his/her final vital status (censored or not), the survival time (time to death) from the entry of the patient in the study (in days), and a covariate including the number of lymph nodes with detectable cancer (grouped from 1 to ≥ 10 in the dataset colonCSm) were used.

```
> devtools::install_github("noramvillanueva/clustcurv")
> library(clustcurv); library(condSURV)
> head(colonCSm)[1:2, ]
  time status nodes
1 1521      1     5
```

2 3087   0   1

The estimated survival curves after splitting the data according to the number of nodes are shown in Figure 1 (upper panel). When we confront with a dataset like this, with a categorical variable with a high number of levels, maybe a good approximation could be to establish groups with the same risk or survival probability. The unique option until now could be to use first the log-rank test and then, if the result of the application of this test is statistically significant, do a post hoc analysis like a pairwise comparison. The p-value of the log-rank test is < 0.01 and the interpretation of the resulting p-values of the pairwise comparison (not shown) becomes a problem.

```
> survdiff(Surv(time, status) ~ factor(nodes), data = colonCSm)
Call:
survdiff(formula = Surv(time, status) ~ factor(nodes),
+ data = colonCSm)
                    N Observed Expected (O-E)^2/E (O-E)^2/V
factor(nodes)=1   274       94   151.93   22.0901   33.9249
factor(nodes)=2   194       74   102.87    8.1022   10.5979
factor(nodes)=3   125       61    62.56    0.0387    0.0453
factor(nodes)=4    84       43    38.26    0.5868    0.6434
factor(nodes)=5    46       34    17.06   16.8249   17.5428
factor(nodes)=6    43       27    16.43    6.8027    7.0736
factor(nodes)=7    38       25    15.41    5.9636    6.1880
factor(nodes)=8    23       18     7.22   16.0875   16.3765
factor(nodes)=9    20       14     8.05    4.3931    4.4795
factor(nodes)=10   62       49    19.21   46.2239   48.6066
Chisq= 129  on 9 degrees of freedom, p= 0

> survminer::pairwise_survdiff(Surv(time, status) ~ nodes,
+ data = colonCSm, p.adjust.method = "BH")
```

To solve it, we applied the proposed procedure. For a significance level of 0.05 and using the Cramér-von Mises type statistic, the null hypothesis $H_0(1)$ is rejected (p-value of < 0.01) while the null hypothesis $H_0(2)$ is accepted (p-value of 0.19). The assignment of the curves to the two groups can be observed in Figure 1.

```
> res <- clustcurv_surv(time = colonCSm$time,
+ status = colonCSm$status, fac = colonCSm$nodes,
+ algorithm = "kmeans", nboot = 500, cluster = TRUE,
+ seed = 300716)
Checking 1 cluster...
Checking 2 clusters...
Finally, there are 2 clusters.
```

```
> autoplot(res, groups_by_colour = TRUE, xlab = "Time (in days)")
```
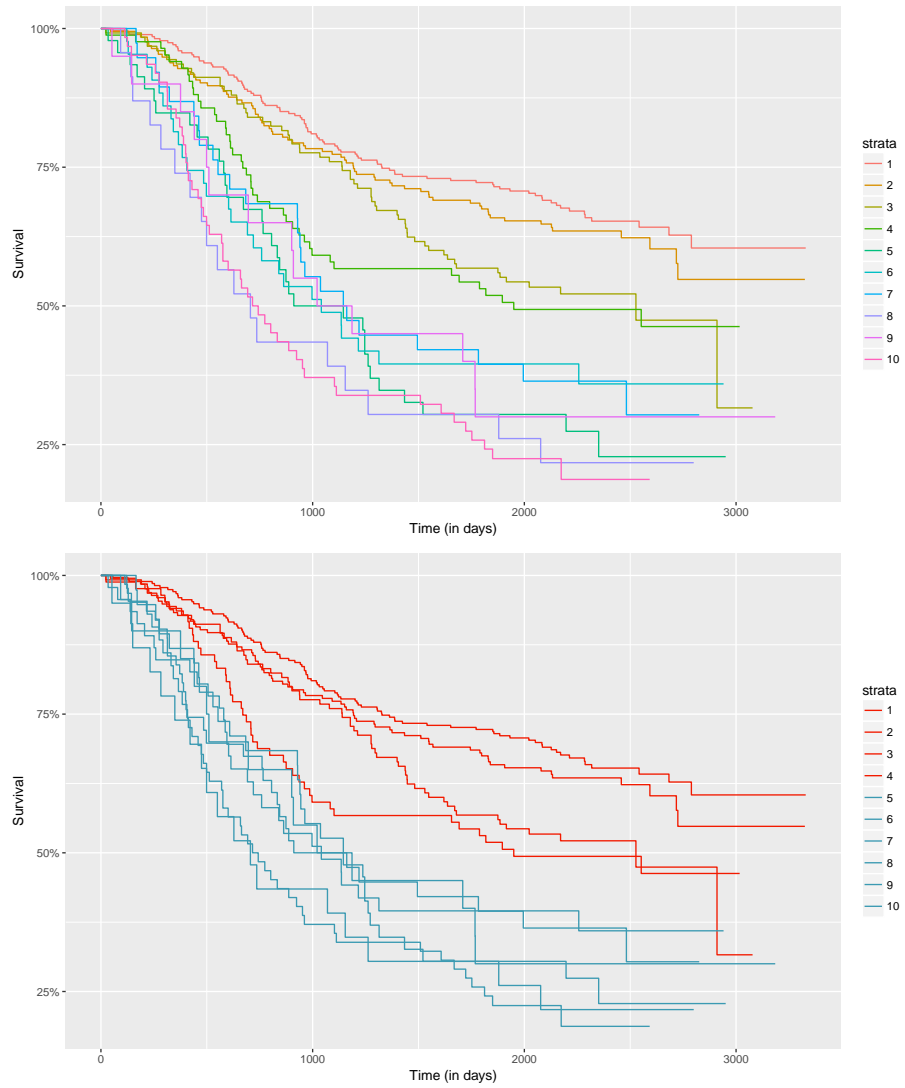

FIGURE 1. Estimated survival curves for each of the levels of the variable "nodes" using the Kaplan-Meier estimator. A specific color is assigned for each curve according to the group to which it belongs (in this case two groups, K = 2).

References

Gehan, E. A. (1965). A generalized wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, **52**, 203 – 223.

Kassambara, A. and Kosinski, M. (2017). survminer: Drawing Survival Curves using 'ggplot2'. *R package version 0.3.1*, version 0.3.1.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, **50**, 163 – 170.

Meira-Machado, L., Sestelo, M. (2016). condSURV: An R Package for the Estimation of the Conditional Survival Function for Ordered Multivariate Failure Time Data. *The R Journal.* , **8(2)**:460–473.

Moertel, C.G., Fleming T.R., Macdonald J.S., et al. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine.* **322(6)**:352–358.

Peto, R. and Peto, J. (1972).Asymptotically efficient rank invariant test procedures (with discussion).*Journal of the Royal Statistical Society, Series A*, **135**, 185 – 206.

Villanueva, N. M., Sestelo, M. and Meira-Machado, L. (2018). A method for determining groups in multiple survival curves. *Statistics in Medicine*, submitted.