Finding groups in time-to-event data by means of the clustcurv package

Nora M. Villanueva[1,2,3], **Marta Sestelo**[1,2,3] **and Luís Meira-Machado**[4]

nmvillanueva@uvigo.es

[1] Department of Statistics and Operation Research, University of Vigo, Spain
[2] SiDOR Research Group and CINBIO, University of Vigo, Spain
[3] Gradiant, Galician Research Center in Advanced Telecommunications, Spain
[4] CMEB and Department of Mathematics and Applications, University of Minho, Portugal.

1. Introduction

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 among independent groups. The log-rank or Mantel-Haenszel test [4] is the most well-known and widely used to test the null hypothesis of no difference in survival between two or more independent groups. Though the aforementioned methods can be used to compare multiple survival curves, to the best of our knowledge, there are none available method that can be used to determine groups among a series of survival curves. In this study we propose an approach that allows determining survival groups with an automatic selection of their number. The proposed method can be used for instances to establish groups with higher risk or with the same risk. It can also be used to check if there is a mono- tonic trend in the survival curves over the levels of ordinal grouping variables (e.g. cancer stages).

2. The algorithm for determining groups

We will assume the J-sample general random censorship model where observations are made on n_j individuals from population $j (j = 1, \dots, J)$. Denote $n = \sum_{j=1}^{J} n_j$ and assume that the observations from the n individuals are mutually independent. Let T_{ij} be an event time corresponding to an event measured from the start of the follow-up of the i-th subject $(i = 1, \dots, n_j)$ in the sample j and assume that T_{ij} is observed subject to a (univariate) random right-censoring variable C_{ij} assumed to be independent of T_{ij}. Due to the censoring, rather than T_{ij} we observe $(\widetilde{T}_{ij}, \Delta_{ij})$ where $\widetilde{T}_{ij} = \min(T_{ij}, C_{ij})$, $\Delta_{ij} = I(T_{ij} \leq C_{ij})$, where $I(\cdot)$ is the indicator function.

Since the censoring time is assumed to be independent of the process, the survival functions, $S_j(t) = P(T_j > t)$, may be consistently estimated by the Kaplan-Meier (KM) estimator [3] based on the $(\widetilde{T}_j, \Delta_j)$.

As we mentioned in the introduction, several nonparametric methods have been proposed in the literature in order to test the equality of multiple survival curves, i.e., to test the null hypothesis $H_0 : S_1 = \ldots = S_J$. However, if this hypothesis is rejected there are no available procedures that let determine groups among these survival curves, that is, to asses if the levels $\{1, \dots, J\}$ can be grouped in K groups $(G_1, \dots, G_K)$ with $K < J$, so that $S_i = S_j$ for all $i, j \in G_k$, for each $k = 1, \dots, K$. Note that $(G_1, \dots, G_K)$ must be a partition of $\{1, \dots, J\}$, and therefore must satisfy $G_1 \cup \ldots \cup G_K = \{1, \dots, J\}$ and $G_i \cap G_j = \emptyset$ for all $i \neq j \in \{1, \dots, K\}$.

Let $(\widetilde{T}_{ij}, \Delta_{ij})$, $i = 1, \dots, n_j$, be a sample from the distribution of $(\widetilde{T}_j, \Delta_j)$, for $j = 1, \dots, J$, a procedure to test, for a given number K, the null hypothesis $H_0(K)$ that at least one partition exists $(G_1, \dots, G_K)$ so that all the conditions above are verified is proposed. The alternative hypothesis $H_1(K)$ is that for any $(G_1, \dots, G_K)$, exists at least a group G_k in which $S_i \neq S_j$ for some $i, j \in G_k$.

The cited testing procedure is based on the J-dimensional process, $\widehat{\mathbf{U}}(t) = (\widehat{U}_1(t), \widehat{U}_2(t), \dots, \widehat{U}_J(t))^t$, where, for $j = 1, \dots, J$, $\widehat{U}_j(t) = \sum_{k=1}^{K} [\widehat{S}_j(t) - \widehat{M}_k(t)] I_{\{j \in G_k\}}$ and $\widehat{M}_k$ function is the pooled Kaplan-Meier estimate based on the combined G_k-partition sample.

The following test statistics were considered in order to test $H_0(K)$: a Cramér-von Mises type statistic and a modification of it based on the L_1 norm proposed in the Kolmogorov-Smirnov test statistic

$$D_{CM} = \min_{G_1,\dots,G_K} \sum_{j=1}^{J} \int_{R_{\widetilde{T}}} \hat{U}_j^2(t) dt, \qquad D_{KS} = \min_{G_1,\dots,G_K} \sum_{j=1}^{J} \int_{R_{\widetilde{T}}} |\hat{U}_j(t)| dt.$$

Note that if $H_0(K)$ is verified, the value of D —which represents D_{CM} and D_{KS}— should be close to zero. The decision rule based on D consists of rejecting the null hypothesis if D is larger than the $(1 - \alpha)$-percentile obtained under the null hypothesis. To approximate the distributions of the test statistic under the null hypothesis, resampling methods such as the bootstrap introduced by Efron [1, 2] can be applied.

Finally, note that repeating this procedure –testing $H_0(K)$– from $K = 1$ onwards until a certain null hypothesis is accepted allows us to determine automatically the number of groups K. The summarized stages of the whole procedure are described in Algorithm 1.

Algorithm 1: *k-survival curves algorithm*

```
1. With $\{(\widetilde{T}_{ij}, \Delta_{ij}), i = 1, \dots, n_j\}$, $j = 1, \dots, J$, and using the KM estimator obtain $\hat{S}_j$.
2. Initialize with $K = 1$ and test $H_0(K)$:
2.1 Obtain the “best” partition $G_1, \dots, G_K$ by means of the $k$-means or $k$-medians algorithm.
2.2 For $k = 1, \dots, K$, estimate $M_k$ and retrieve the test statistic $D$.
2.3 Generate $B$ bootstrap samples and calculate $D^{*b}$, for $b = 1, \dots, B$.
2.4 if $D > D^{*(1-\alpha)}$ then
        reject $H_0(K)$
        $K = K + 1$
        go back to 2.1
    else
        accept $H_0(K)$
    end
3. The number $K$ of groups of equal survival curves is determined.
```

3. Package description

Part of our philosophy is to make easier to others the use of a new statistical methodology. Based on this, we have implemented in a user-friendly and simply *R* package this methodology. Some functions programmed in the **clustcurv** package are shown in Table 1.

Table 1: *Summary of some functions in **cluscurv** package.*

Function	Description
`clustcurv.surv`	Function for grouping survival curves based on the k-means or k-medians algorithm. It returns the number of groups and the assignment.
`kgroups.surv`	Function for grouping survival curves, given a number k, based on the k-means or k-medians algorithm.
`autoplot`	Visualization of ”clustcurv.surv” objects with ggplot2.

4. Application to patients with colon cancer

To illustrate our method we will use a dataset that comes from a large clinical trial on Duke's stage III patients, affected by colon cancer, that underwent a curative surgery for colorectal cancer [6]. This data set is freely available as part of the R package `condSURV` [4, 5]. 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) were used.

The estimated survival curves after splitting the data according to the number of nodes are shown in Figure 1. 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. 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 (bottom panel). Note that having five or more nodes seems to be related with a lower survival than having four or less (group 1: $\leq$ 4 nodes, group 2: > 4 nodes).

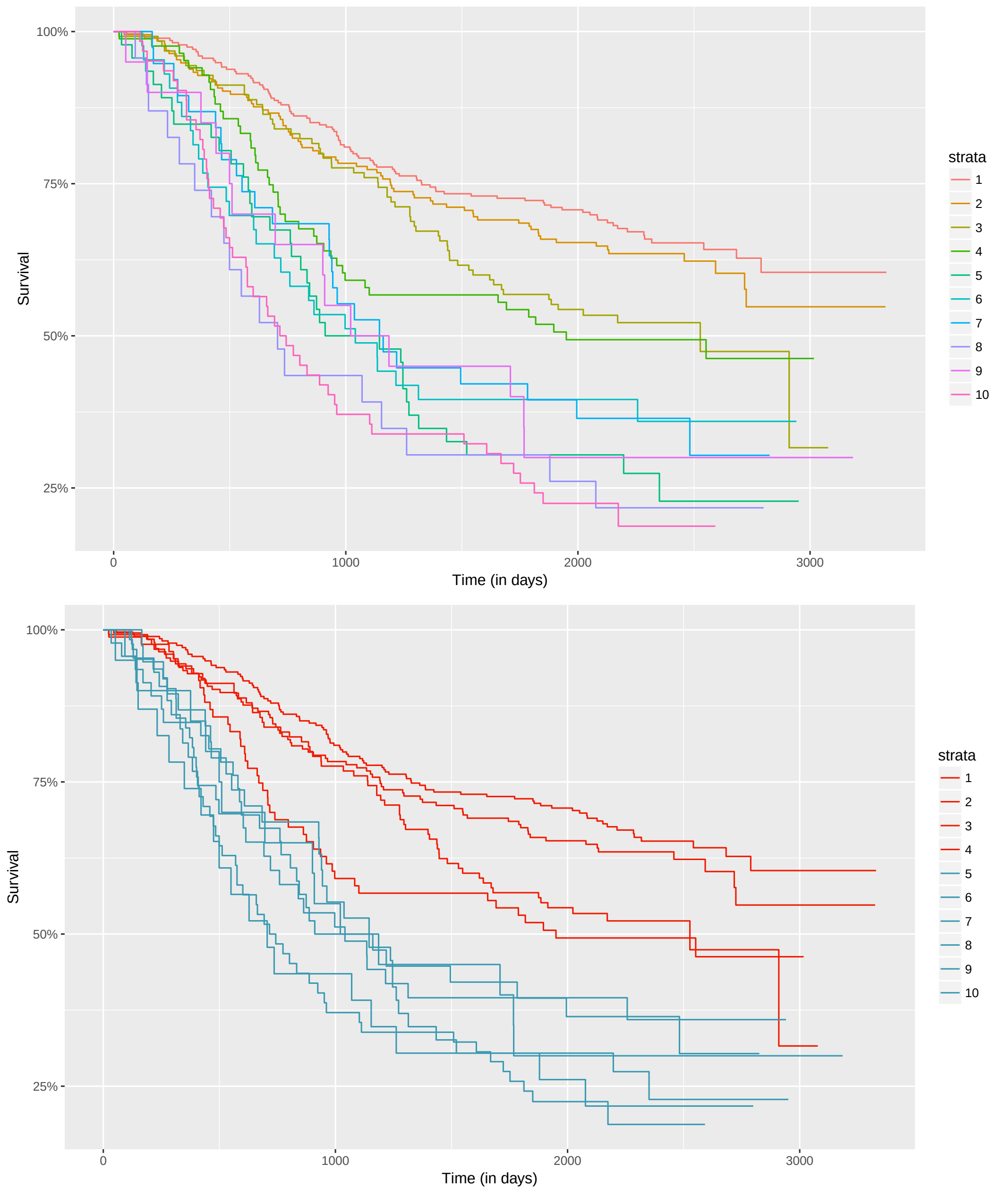

Figure 1: *Estimated survival curves for each of the levels of the variable “nodes” using the Kaplan-Meier estimator. Bottom panel: A specific color is assigned for each curve according to the group to which it belongs (in this case five groups, K = 2).*

References

[1] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

[2] Bradley Efron. Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374):312–319, 1981.

[3] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[4] Luis Meira-Machado and Marta Sestelo. 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, 2016.

[5] Luis Meira-Machado and Marta Sestelo. condSURV: Estimation of the Conditional Survival Function for Ordered Multivariate Failure Time Data. R package version 2.0.1, 2016.

[6] Charles G Moertel, Thomas R Fleming, John S Macdonald, Daniel G Haller, John A Laurie, Phyllis J Goodman, James S Ungerleider, William A Emerson, Douglas C Tormey, John H Glick, et al. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine*, 322(6):352–358, 1990.

Acknowledgements This research was supported by grant MTM2011-23204 (FEDER support included) from the Spanish Ministry of Science and Innovation and by grant CN2012/180 from the Galician Regional Authority (Xunta de Galicia).

European R Users Meeting, May 14-16, 2018, Budapest, Hungary.