Clustering of nonparametric curves by the clustcurv package

Nora M. Villanueva[1,2,*]

[1] Department of Statistics and O. R., SiDOR Group & CINBIO, University of Vigo, Spain
[2] Getlife, https://getlife.es/

Xornadas Usuarios R Galicia
Universidade de Santiago de Compostela

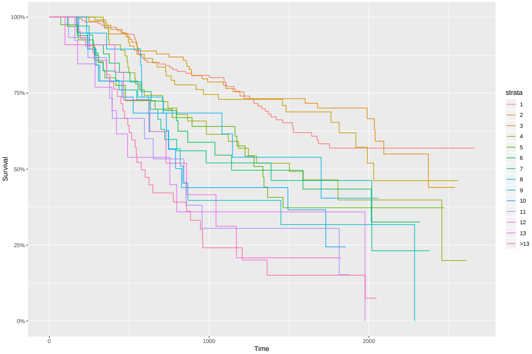
October 20th, 2022

[*] work jointly done with M. Sestelo, L. Meira-Machado and J. Roca-Pardiñas.

SiDOR
Statistical Inference
Decision & Operations Research Group
Universidade Vigo

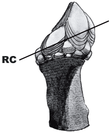
German Breast Cancer study data set*

- 686 patients with primary node positive breast cancer
- 299 patients developed recurrence
- Patients were recruited between July 1984 and December 1989 and 16 variables
  - ∗ Times (in days) to recurrence (`rectime`)
  - ∗ Censoring indicator (`censrec`)
  - ∗ Number of positive nodes with cancer: grouped from 1 to > 13 (`nodes`)



* Sauerbrei W. and Royston P. (1999).

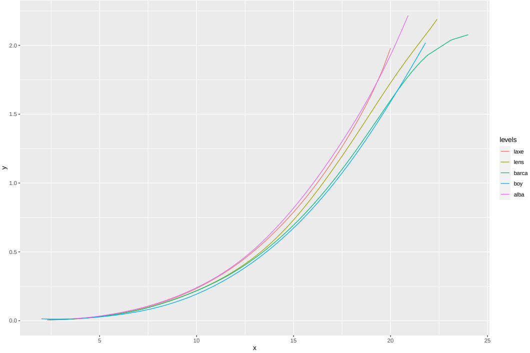
Growth's barnacle study data set*



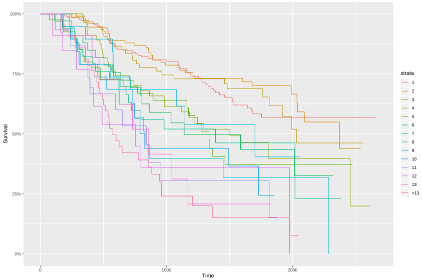
- Five sampling sites of the Galicia's Atlantic coast

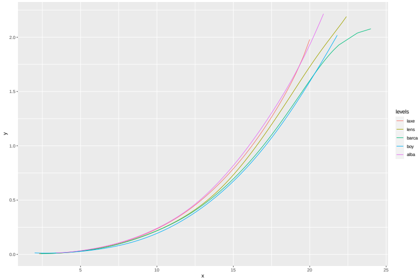
- Two biometric variables:

    * Rostro-carinal length (RC)
    * Dry weight (DW)



* Sestelo, M. (2013).





1. Are all these curves equal?
2. Can we identify groups in some way?

Nonparametric methods to test for the equality of survival curves among independent groups

- Log-rank or Mantel-Haenszel test (Mantel N., 1966)
- Peto & Peto (1972), modification of the Gehan-Wilcoxon test (1965)
- Tarone test (1977), Harrington and Fleming test (1982), Fleming et al. (1987), Chen and Zhang (2016), etc.

The equality of mean functions has been widely investigated in the literature:

- Hall and Hart (1990), King *et al.* (1991), Delgado (1993), Kulasekera (1995), Young and Bowman (1995), Dette and Neumeyer (2001), Pardo-Fernández *et al.* (2007), Srihera and Stute (2010), Park *et al.* (2014), etc.
- González-Manteiga and Crujeiras (2013) offers a good review about this topic.

When the null hypothesis of equality of curves is rejected, at least one curve is different...

- Naïve approach: pairwise comparisons
- Some approaches to determine groups have been developed in
  - functional data context (Abraham et al., 2003; García-Escudero and Gordaliza, 2005; Tarpey, 2007).
  - longitudinal data context (Vogt and Linton, 2017, 2020).

We propose an approach that allows determining survival and regression groups with an automatic selection of their number

Some previous notation - Survival

- J-sample general random censorship model, where observations are made on n_j individuals from population $j(j = 1, \ldots, J)$
- Let T_{ij} be an event time corresponding to an event measured from the start of the follow-up of the i-th subject $(i = 1, \ldots, n_j)$ in the sample j
- Assuming that T_{ij} is observed subject to a (univariate) random right-censoring variable C_{ij} assumed to be independent of T_{ij}
- Due to censoring we only observe $(\widetilde{T}_{ij}, \Delta_{ij})$ where $\widetilde{T}_{ij} = \min(T_{ij}, C_{ij})$, $\Delta_{ij} = I(T_{ij} \leq C_{ij})$

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 estimator (Kaplan and Meier, 1958).

Let $(\widetilde{T}_{ij}, \Delta_{ij})$, $i = 1, \ldots, n_j$, be a sample from the distribution of (T_j, Δ_j), for $j = 1, \ldots, J$, the estimation of the survival function $S_j(t)$ can be obtained as

$$\widehat{S}_j(t) = \prod_{u:t_u \leq t} \left(1 - \frac{d_u}{R_j(t_u)}\right)$$

being d_u the number of events from population j at time t_u and $R_j(t) = \sum_{i=1}^{n_j} I(\widetilde{T}_{ij} \geq t)$ the number of individuals of risk just before time t, among individuals from population j.

Some previous notation - Regression

Let (X_j, Y_j) be J independent random vectors, and assume that they satisfy the following nonparametric regression models, for $j = 1, \dots, J$,

$$Y_j = m_j(X_j) + \varepsilon_j \tag{1}$$

where m_j is a nonparametric smooth function and ε_j is the regression error, which is assumed independent of the covariate X_j with $E(\varepsilon_j) = 0$ and $Var(\varepsilon_j) = \sigma_j^2$.

Explicitly, given J independent random samples, say

$$\left\{\mathcal{P}_1 = \{(X_{i1}, Y_{i1})\}_{i=1}^{n_1}, \ldots, \mathcal{P}_J = \{(X_{iJ}, Y_{iJ})\}_{i=1}^{n_J}\right\}$$

where the random variables $(X_{1j}, Y_{1j}), \ldots, (X_{n_j j}, Y_{n_j j})$ are i.i.d. for each $j = 1, \ldots, J$ and with a total sample size $n = \sum_{j=1}^{J} n_j$, the local linear kernel smoothers[3]

$$\hat{m}_j(x) = \Psi(x, \mathcal{P}_j, h_j, r)$$

at a location x, with $r = 1$, is given by $\hat{m}_j(x) = \hat{\alpha}_{0j}(x)$, where $\hat{\alpha}_{0j}(x)$ is the first element of the vector $(\hat{\alpha}_{0j}(x), \hat{\alpha}_{1j}(x))$ which is the minimiser of

$$\sum_{i=1}^{n_j} \{Y_{ij} - \alpha_{0j}(x) - \alpha_{1j}(x)(X_{ij} - x)\}^2 \cdot \kappa\left(\frac{X_{ij} - x}{h_j}\right),$$

where κ denotes a kernel function (normally, a symmetric density), and $h_j > 0$ is the smoothing parameter or bandwidth selected automatically by cross-validation.

[3] Fan, J. and Gijbels, I. (1996); Wand, M. P. and Jones, M. C. (1995)

If $H_0 : \mathcal{F}_1 = \ldots = \mathcal{F}_J$ is rejected...

- We would like to asses if the levels $\{1, \ldots, J\}$ can be grouped in K groups $\{G_1, \ldots, G_K\}$ with $K < J$, so that
  - $\mathcal{F}_i = \mathcal{F}_j$ for all $i, j \in G_k$, for each $k = 1, \ldots, K$
  - $\{G_1, \ldots, G_K\}$ must be a partition of $\{1, \ldots, J\}$
  - $G_1 \cup \ldots \cup G_K = \{1, \ldots, J\}$ and $G_i \cap G_j = \emptyset$ for all $i \neq j \in \{1, \ldots, K\}$
- A procedure to test, for a given number K, the null hypothesis $H_0(K)$ is that at least exists a partition $\mathbf{G}_0 = \{G_1, \ldots, G_P\}$ with $P \leq K$ so that all the conditions above are verified.
- The alternative hypothesis $H_1(K)$ is that for any partition $\mathbf{G}_1 = \{G_1, \ldots, G_L\}$ with $L > K$, not exists another partition $\mathbf{G}_0$ verifying $\#\mathbf{G}_0 < \#\mathbf{G}_1$ where

$$\#\{G_1, \ldots, G_{\mathcal{K}}\} = 1 + \sum_{k_2=2}^{\mathcal{K}} \left(\prod_{k_1 < k_2} I\{G_{k_1} \neq G_{k_2}\} \right)$$

  and, for definition, $G_{k_1} \neq G_{k_2}$ is verified if $S_i \neq S_j$ for all $(i, j) \in G_{k_1} \times G_{k_2}$.

The testing procedure is based on the J-dimensional process

$$\widehat{\mathbf{U}}(z) = (\widehat{U}_1(z), \widehat{U}_2(z), \ldots, \widehat{U}_J(z))^t,$$

where, for $j = 1, \ldots, J$,

$$\widehat{U}_j(z) = \sum_{k=1}^{K} [\widehat{\mathcal{F}}_j(z) - \widehat{\mathcal{C}}_k(z)] \; I_{\{j \in G_k\}}$$

and $\widehat{\mathcal{C}}_k$ is the pooled nonparametric estimate based on the combined G_k-partition sample

- Statistic tests

$$D_{CM} = \min_{G_1,\ldots,G_K} \sum_{j=1}^{J} \int_R \widehat{U}_j^2(z) dy,$$

$$D_{KS} = \min_{G_1,\ldots,G_K} \sum_{j=1}^{J} \int_R |\widehat{U}_j(z)| dy.$$

$*$ With $J = 30$ and $K = 5$, the total number of distinct assignments[1] is $7.7\ 10^{18}$

[1] Following Jain and Dubes (1988), $R(J, K) = \frac{1}{K!} \sum_{i=1}^{K} (-1)^{K-i} \binom{K}{i} (i)^n$

The testing procedure is based on the J-dimensional process

$$\widehat{\mathbf{U}}(z) = (\widehat{U}_1(z), \widehat{U}_2(z), \ldots, \widehat{U}_J(z))^t,$$

where, for $j = 1, \ldots, J$,

$$\widehat{U}_j(z) = \sum_{k=1}^{K} [\widehat{\mathcal{F}}_j(z) - \widehat{\mathcal{C}}_k(z)]\; I_{\{j \in G_k\}}$$

and $\widehat{\mathcal{C}}_k$ is the pooled nonparametric estimate based on the combined G_k-partition sample

- Statistic tests

$$D_{CM} = \min_{G_1,\ldots,G_K} \sum_{j=1}^{J} \int_R \widehat{U}_j^2(z) dy, \longrightarrow \text{ Kmeans}$$

$$D_{KS} = \min_{G_1,\ldots,G_K} \sum_{j=1}^{J} \int_R |\widehat{U}_j(z)| dy \longrightarrow \text{ Kmedians}$$

$*$ With $J = 30$ and $K = 5$, the total number of distinct assignments[1] is $7.7\ 10^{18}$

[1] Following Jain and Dubes (1988), $R(J, K) = \frac{1}{K!} \sum_{i=1}^{K} (-1)^{K-i} \binom{K}{i} (i)^n$

The testing procedure is based on the J-dimensional process

$$\widehat{\mathbf{U}}(z) = (\widehat{U}_1(z), \widehat{U}_2(z), \ldots, \widehat{U}_J(z))^t,$$

where, for $j = 1, \ldots, J$,

$$\widehat{U}_j(z) = \sum_{k=1}^{K} [\widehat{\mathcal{F}}_j(z) - \widehat{\mathcal{C}}_k(z)] \; I_{\{j \in G_k\}}$$

and $\widehat{\mathcal{C}}_k$ is the pooled nonparametric estimate based on the combined G_k-partition sample

- Statistic tests

$$D_{CM} = \min_{G_1,\ldots,G_K} \sum_{j=1}^{J} \int_R \widehat{U}_j^2(z) dy, \longrightarrow \text{Kmeans}$$

$$D_{KS} = \min_{G_1,\ldots,G_K} \sum_{j=1}^{J} \int_R |\widehat{U}_j(z)| dy \longrightarrow \text{Kmedians}$$

- Decision rule: we reject H_0 for large statistic values.
- Distribution of D? Bootstrap method (Efron, B., 1979, 1981)

The steps of the testing procedure, for a given K, are as follows

Step 1. Using the original sample, for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$, estimate the functions $\mathcal{F}_j$ in a non parametric way and in a common grid, using each sample separately.

Then, using the proposed algorithms, obtain the "best" partition $\{G_1, \ldots, G_K\}$ and with it obtain the estimated curves $\widehat{\mathcal{C}}_k$ using a pooled nonparametric estimator based on the combined partition samples

Step 2. Obtain the D value as explained before.

Step 3. Draw bootstrap samples using a pooled bootstrap procedure (i.e., bootstrap from the pooled combined partition sample given by the null hypothesis $H_0(K)$).

Survival - For $b = 1, \ldots, B$ and for each $j \in G_k$, draw $(\widetilde{T}^{*b}_{1j}, \Delta^{*b}_{1j}), (\widetilde{T}^{*b}_{2j}, \Delta^{*b}_{2j}), \ldots, (\widetilde{T}^{*b}_{n_j j}, \Delta^{*b}_{n_j j})$ by independent sampling n_j times with replacement from $\widehat{F}_k$, the empirical distribution function putting mass n_k^{-1} $(n_k = \sum_{j=1}^{J} n_j I_{\{j \in G_k\}})$ at each point $(\widetilde{T}_{ij}, \Delta_{ij})$, with $j \in G_k$.

Regression - For $b = 1, \ldots, B$, and for each $j \in G_k$, draw $\left\{\left(X_{i1}, Y^{*b}_{i1}\right)\right\}_{i=1}^{n_1}, \ldots, \left\{\left(X_{iJ}, Y^{*b}_{iJ}\right)\right\}_{i=1}^{n_J}$ where

$$Y^{*b}_{ij} = \sum_{k=1}^{K} \widehat{\mathcal{C}}_k(X_{ij}) I_{\{j \in \mathcal{G}_k\}} + \widehat{\varepsilon}_{ij} W^{*b}_i$$

being $\widehat{\varepsilon}_{ij}$ the null errors under the $H_0(K)$ obtained as

$$\widehat{\varepsilon}_{ij} = \sum_{k=1}^{K} \left(Y_{ij} - \widehat{\mathcal{C}}_k(X_{ij})\right) I_{\{j \in \mathcal{G}_k\}}$$

and $W^{*b}_1, \ldots, W^{*b}_n$ i.i.d. random variables with mass $(5+\sqrt{5})/10$ and $(5-\sqrt{5})/10$ at the points $(1-\sqrt{5})/2$ and $(1+\sqrt{5})/2$.

Step 4. Let D^{*b} be the test statistic obtained from the bootstrap samples after applying step 1 and 2 to the cited bootstrap samples.

The decision rule consists of rejecting the null hypothesis if $D > D^{*(1-\alpha)}$, where $D^{*(1-\alpha)}$ is the empirical $(1-\alpha)$-percentile of values D^{*b} $(b = 1, \ldots, B)$ previously obtained.

Algorithm. K-nonparametric curves algorithm

1. With the original sample, for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$, and using the nonparametric estimator obtain $\widehat{\mathcal{F}}_j$.
2. Initialize with $K = 1$ and test $H_0(K)$:
   - 2.1. Obtain the "best" partition $\{G_1, \ldots, G_K\}$ by means of the K-means or K-medians algorithm.
   - 2.2. For $k = 1, \ldots, K$, estimate $\widehat{\mathcal{C}}_k$ and retrieve the test statistic D.
   - 2.3. Generate B bootstrap samples and calculate D^{*b}, for $b = 1, \ldots, B$.
   - 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 nonparametric curves is determined.

- clustcurv package is a shortcut for “clustering curves”
- To provide a procedure that allows users determining groups of multiple curves with an automatic selection of their number
- The package works both for survival and regression curves.
- The design of the clustcurv package has been done in a similar fashion to other R packages
- In view of the high computational cost entailed in these methods, parallelization techniques are included to become feasible and efficient onto real situations
- Several unit tests have been implemented and https://cran.r-project.org/web/packages/clustcurv/vignettes/clustcurv.html
- The package is freely available from the CRAN

- Two main types of functionalities:

  * to determine groups of curves, given a number K, with `kregcurves()` or `ksurvcurves()` functions

  * to determine groups of curves with the automatic selection of their number with `regclustcurves()` or `survclustcurves()` functions

- Numerical and graphical summaries can be obtained by using the generic functions `print()`, `summary()` and `autoplot()`

	survclustcurves() arguments
time	A vector with the variable of interest, i.e. survival time.
status	A vector with the censoring indicator of the survival time of the process; 0 if the total time is censored and 1 otherwise.
x	A vector with the categorical variable indicating the population to which the observations belongs.
kvector	A vector specifying the number of groups of curves to be checked. By default it is NULL.
kbin	Size of the grid over which the survival functions are to be estimated.
nboot	Number of bootstrap repeats.
algorithm	A character string specifying which clustering algorithm is used, i.e., K-means or K-medians.
alpha	A numeric value, particularly, the signification level of the hypothesis test.
cluster	A logical value. If `TRUE` (default) the code is parallelised.
ncores	An integer value specifying the number of cores to be used in the parallelised procedure. If `NULL`, the number of cores to be used is equal to the number of cores of the machine − 1.
seed	Seed to be used in the procedure.
multiple	A logical value. If TRUE (not default), the resulted pvalues are adjusted
multiple.method	Correction method: 'bonferroni', 'holm', 'hochberg', etc.

Table: Arguments of `survclustcurves()`.

Application to real data

German Breast Cancer study data set*

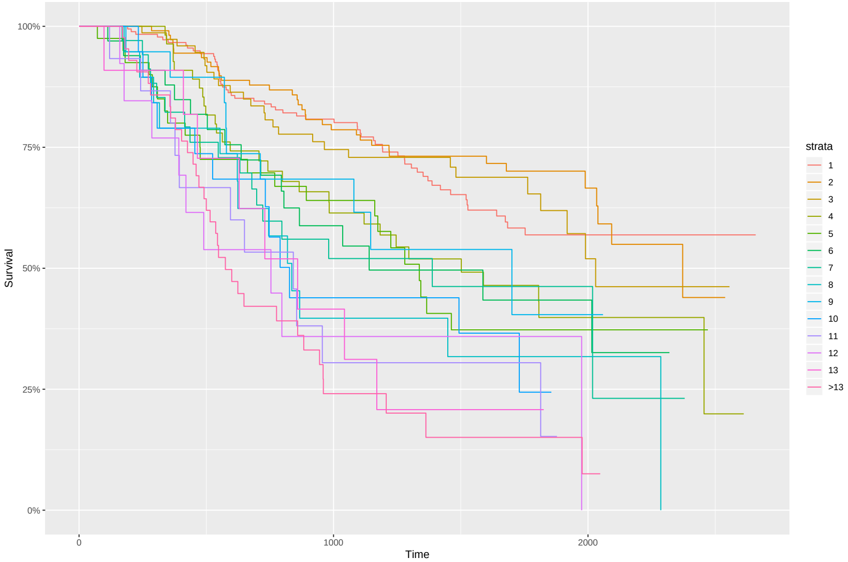
```
> library(clustcurv)
```
[7]
```
> library(condSURV)
> data(gbcsCS)

> head(gbcsCS[, c(5:10, 13, 14)])

  age menopause hormone size grade nodes rectime censrec
1  38         1       1   18     3     5    1337       1
2  52         1       1   20     1     1    1420       1
3  47         1       1   30     2     1    1279       1
4  40         1       1   24     1     3     148       0
5  64         2       2   19     2     1    1863       0
6  49         2       2   56     1     3    1933       0

> table(gbcsCS$nodes)

  1   2   3   4   5   6   7   8   9  10  11  12  13 >13
187 110  79  57  41  33  36  20  20  19  15  13  11  45
```

[7] https://cran.r-project.org/web/packages/clustcurv/vignettes/clustcurv.html

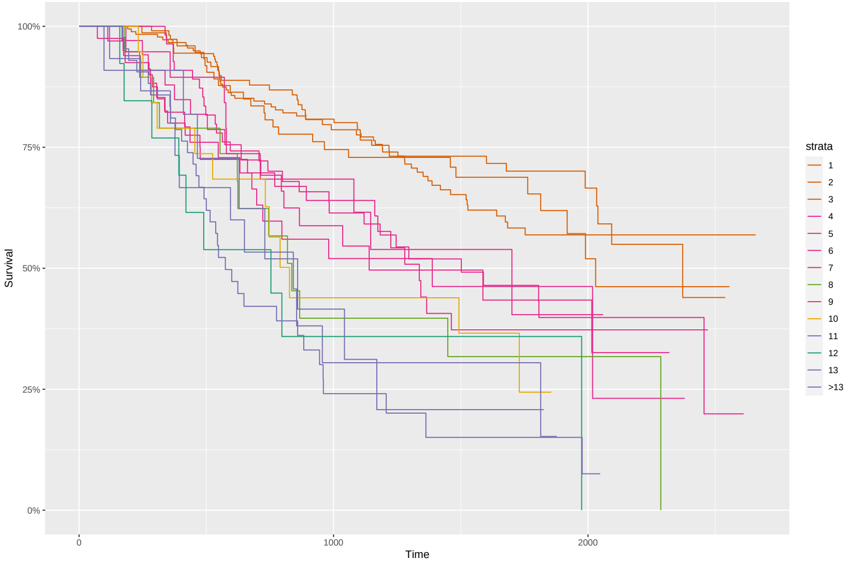
```
> fit.kgbcs <- ksurvcurves(time = gbcsCS$rectime, status = gbcsCS$censrec,
x = gbcsCS$xnodes, seed = 300716, algorithm = "kmedians", k = 6)

> print(fit.kgbcs)
```

```
Call:
ksurvcurves(time = gbcsCS\$rectime, status = gbcsCS\$censrec,
    x = gbcsCS\$nodes, k = 6, algorithm = "kmedians",
     seed = 300716)

Clustering curves in 6 groups
Number of observations:  686
Cluster method:  kmedians
```

```
> autoplot(fit.kgbcs , groups_by_colour = FALSE)
```



```
> autoplot(fit.kgbcs , groups_by_colour = TRUE)
```

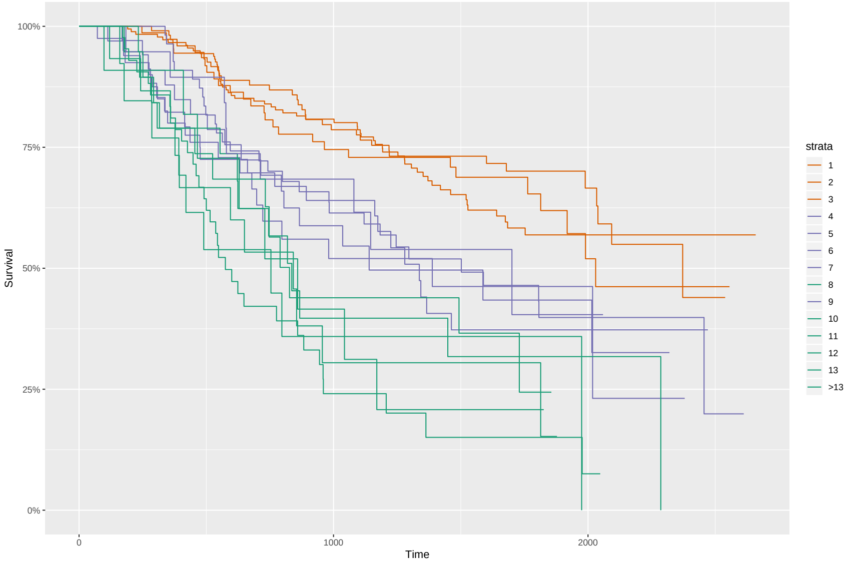


```
> fit.gbcs <- survclustcurves(time = gbcsCS$rectime,
status = gbcsCS$censrec, x = gbcsCS$nodes,
nboot = 500, seed = 300716, cluster = TRUE,
algorithm = 'kmedians')
```

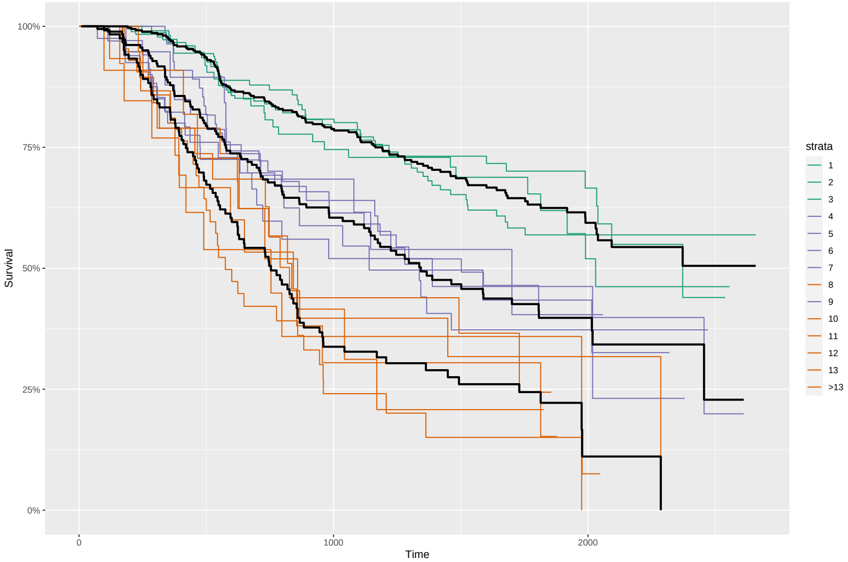
```
Checking 1 cluster...
Checking 2 clusters...
Checking 3 clusters...
Finally, there are 3 clusters.
```

```
> summary(fit.gbcs)
Call:
survclustcurves(time = gbcsCS$rectime, status = gbcsCS$censrec,
x = gbcsCS$nodes, nboot = 500, algorithm = "kmedians",
cluster = TRUE, seed = 300716)

Clustering curves in 3 groups
Number of observations: 686
Cluster method: kmedians

Factor's levels:
[1] "1" "2" "3" "4" "5" "6" "7" "8" "9"
[10]"10" "11" "12" "13" ">13"

Clustering factor's levels:
[1] 1 1 1 3 3 3 3 2 3 2 2 2 2 2

Testing procedure:
H0 Tvalue pvalue
1 1 95.68626 0.000
2 2 56.03966 0.018
3 3 33.63386 0.830

Available components:
[1] "num_groups" "table" "levels" "cluster" "centers"
[6] "curves" "method" "data" "algorithm" "call"
```

```
> autoplot(fit.gbcs, groups_by_colour = TRUE)
```



```
> autoplot(fit.gbcs, groups_by_colour = TRUE, centers = TRUE)
```



- A new package is developed that let us, not only testing the equality of nonparamteric curves but also grouping them if they are not equal.

- It is available from the Comprehensive R Archive Network, CRAN.

- It seems to be stable and computational efficient because of parallezing techniques.

- The contributions of this talk are based on:

Villanueva, N. M., Sestelo, M., Meira-Machado, L. and Roca-Pardiñas, J. (2021). clustcurv: An R package for Determining Groups in Multiple Curves. *The R Journal*, 13 (1), 164-183.

Villanueva, N. M., Sestelo, M. and Meira-Machado, L. (2019). A Method for Determining Groups in Multiple Survival Curves. *Statistics in Medicine*, 38:366–377.

Villanueva, N. M., Sestelo, M., Ordóñez, C. and Roca-Pardiñas, J. (2021). An Automatic Procedure to Determine Groups of Nonparametric Regression Curves. *arXiv*.

- Efron B.(1981) Censored Data and the Bootstrap. *Journal of the American Statistical Association*, 76(374):312–319.

- Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Number 66 in Monographs on statistics and applied probability series. Chapman & Hall.

- González-Manteiga, W. and Crujeiras, R. (2013).Testing the hypothesis of a general linear model using nonparametric regression estimation. *Test*, 2(1), 223–249.

- Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.

- Macqueen J. B. (1967). Some methods of classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

- Mammen, E. (1993). Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *The Annals of Statistics*, 21(1) 255–285.

- Mantel N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3), 163–170.

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

Clustering of nonparametric curves by the clustcurv package

Nora M. Villanueva[1,2,*]

[1] Department of Statistics and O. R., SiDOR Group & CINBIO, University of Vigo, Spain
[2] Getlife, https://getlife.es/

Xornadas Usuarios R Galicia
Universidade de Santiago de Compostela

October 20th, 2022

[*] work jointly done with M. Sestelo, L. Meira-Machado and J. Roca-Pardiñas.

SiDOR
Statistical Inference
Decision & Operations Research Group
Universidade Vigo