Replies to the comments on “clustcurv: An R Package for Determining Groups in Multiple Curves” (R Journal submission 2020-59)

N. M. Villanueva, M. Sestelo, L. Meira-Machado and J. Roca-Pardiñas

December 2020

1 Editor

First of all, we would like to thank the Editor for giving us the opportunity to present a revised version of the manuscript. We also thank the referee for their valuable suggestions and comments. The various points raised are now individually dealt with below. In the hope of making clearer our answers, we have included them in the text in blue color.

1.1 Comments to the Author

Methodologically, the work is well-motivated and for the most part sound. Performing inference in complex analysis is a worthwhile goal. My only hesitations about the methodology are

- The multiple testing strategy is somewhat artificial, as even the authors acknowledge. It treats the learned $\widehat{K}$ as fixed in advance. I am far from an expert in multiple testing, but I am aware of recent work in online control of the false discovery rate that may be of interest to the authors (specifically, work from Ramesh Johari at Stanford and Aditya Ramdas at CMU).

  Our approach deals with the problem of multiple hypothesis testing where a set of K p-values corresponding to the K null hypothesis, $H_0(1), H_0(2), \cdots, H_0(K)$ are given. We note that our proposed method is based on sequential hypothesis testing. It starts with $H_0(1)$ and if this hypothesis (of a single group) is rejected follows with $H_0(2)$ and so on until the null hypothesis is not rejected. To take this issue in consideration we proposed the use of a multiple testing procedure where, after having increased K in the algorithm, the null hypotheses for “smaller” values of K is re-tested simultaneously with $H_0(K)$. In addition to this, as we think that the final user must be able to decide this question, the main functions of the package includes arguments ("multiple = FALSE") and ("multiple.method = 'holm'") where the user can indicate if the correction is applied or not. Users can also use the functions `ksurvcurves` or `kregcurves` (available as part of our package) to make K-groups and make is own decision.

As mentioned in the manuscript, this is an interesting topic in which there are still challenges because there is no information about the number of tests needed to apply these techniques. We are grateful for the referee to have pointed the line of work from Ramesh Johari. These authors propose alternative methods that do not require multiple testing corrections. They adjust for multiple comparisons in a sequential manner. We agree that this issue deserves particular attention and will be considered in future. Unfortunately, we were not able to include them in this version of the package.

Anyway, during the revision process, we have carried out some simulations in order to test if the type I error is hold or if the application of some correction is needed. The scenario and results are shown below.

For $j = 1, \ldots, 120$, the regression model given by

$$Y_j = m_j(X_j) + \varepsilon_j$$

was considered with

$$m_j(X_j) = \begin{cases} 0 & \text{if} \quad j \leq 50, \\ 1 - 2X_j & \text{if} \quad 51 < j \leq 80, \\ 0.75 \tan^{-1} 10(\, X_j - 0.6) & \text{if} \quad 81 < j \leq 100, \\ 2.5\ (1 - X_j^2)^4\, \mathbb{1}(|X_j| \leq 1) & \text{if} \quad 101 < j \leq 110, \\ 1.75 \tan^{-1} 5(X_j - 0.6) + 0.75 & \text{if} \quad 111 < j \leq 120, \end{cases}$$

being X_j the explanatory covariate drawn from a uniform distribution on the interval $[0, 1]$, and ε_j the error distributed in accordance to a normal distribution $N(0, 1.3)$.

The simulation study was carried out under different sample sizes $n_j = 100, 150, 200$ taking into account the test statistic D_{CM}. The remainder parameters of the simulation are 1000 simulation runs and 500 bootstrap replicates.

In order to perform correctly, the proposed method must reject the first null hypothesis, $H_0(1)$, continues, rejects again the second one, $H_0(2)$, and so on until it accepts $H_0(5)$. Results of this simulation are shown in Table 1 which refers to the number of times that the procedure works well (in %) selecting the number of groups K and using a significance level of $\alpha = 0.05$. As can be seen, results reveal a good behaviour of the

Table 1: Number of times in % (of 1000 repetitions) that the proposed method selects the number of groups using a nominal level of 5%.

	Number of groups						
n_j	4	5	6	7	8	9	10
100	0.2	94.3	4.1	1.1	0.1	0.1	0.1
150	0.0	94.5	4.4	0.9	0.1	0.1	0.0
200	0.0	95.2	3.7	0.7	0.4	0.0	0.0

proposed method, with rates of success around 95%, coming quite close to the $(1-\alpha)$ established. Already for the smallest samples size $n_j = 100$, our procedure selects the true number of groups $K = 5$ in 94.3% of the times.

Thus, taking into account these results, we think that the nominal level is hold and, at least in this situation, it is not necessary to apply the correction techniques.

- In the regression context, I did not understand why W_j is needed to shrink what look like bootstrap residuals ϵ^{*b}. Can you explain where this comes from, or at least give a reference that justifies the shrinkage? A simulation that explicitly computes the Type I and II error rates would also alleviate my concerns.

  The use of W_j in the third step of the testing procedure used in the regression context, is a consequence of the type of bootstrap used here. This is known as the wild bootstrap (Wu 1986 [1]) which is suited when the model may exhibit heterocedasticity. This has been clarified in the new version of the paper.

I also find the exposition between the survival and regression contexts repetitive. A more concise approach would describe the main algorithms more abstractly, allowing you to treat the survival and regression contexts as special cases in shorter subsections.

In terms of software, I was able to install the package, and its examples all work. The functions are well documented; however, there are no vignettes. The user-facing functions are supported by a collection of modular utility functions, and though the code could benefit from a thorough linting (mainly to deal with whitespace), it is readable. There is also a comprehensive test suite.

In response to these comments, we have rewritten the methodological section avoiding repetitions in our writing. Moreover, vignettes have been included in the new version of the package.

My main qualms about the software are with respect to the user interface,

- It would have been nice to have a `summary` method associated with the outputs provided by `kclustcurve` and `autoclustcurve`. Right now, the result is a long list, which can be hard to make sense of at first. Even simply reporting the `$table` attribute when calling `summary` would have been useful. A `print` method could also help.

The referee is correct. There are included neither `summary` nor `print` methods associated with the outputs provided by the function `kclustcurve` and `autoclustcurve`. In order to provide a more useful R package, we have included these methods in the new version of the package.

- The regression and survival algorithms do not use the same arguments, but they are implemented in the same functions. This means that some subset of arguments is always being ignored. I would prefer separating functionality for these two contexts.

  With regard to the package functionality, we have followed the reviewer suggestion and we have submitted a new version of the package that considers different functions to implement survival and regression methods, particularly `kregcurves`, `ksurvcurves`, `regclustcurves` and `survclustcurves` functions.

- There should not be a legend entry for every single curve in `autoplot` – this will become messy for even reasonably sized datasets. Perhaps you could (optionally) use `ggrepel` style text labels instead.

  According to the referee's comment, and once tried the `ggrepel` package, we have decided not to include it because we feel that our problem is not to repel the overlapping text labels. In any case, and in order to improve the `autoplot` function, we have include an new argument (interactive = TRUE) that allows the user to use the interactivity afforded by the `plotly` package.

- In theory, there's nothing stopping you from passing in a generic clustering function (with some standardized interface) into the main algorithm. K-means and K-medians are reasonable defaults, but it might expand the reach of your work to allow more arbitrary clustering techniques.

  With regard to the clustering techniques, we are aware that different techniques can be used to tackle this problem (such as Mean-Shift, K-medoids, Clustering using Gaussian Mixture Models (GMM), etc.) however, from a theoretical point of view, some changes must be done in the test statistics. We expect to include some of them when preparing future releases of the package.

- Could you report a T value that has been standardized by the bootstrap null distribution? It is somewhat disorienting to see very large test statistics being assigned large p-values, though I understand it is technically correct.

  While we understand the referee concern we think that the use of a standardization may not be suitable in this case.

In terms of communication, the technical content is easy to follow, but there are many grammatical issues / typos. Most are harmless, but some bothered me,

- Inconsistent use of k vs. K for the number of clusters.
- The $\widehat{U}_j(t)$ is incorrectly defined using $\widehat{M}_k(t)$, rather than $\widehat{Q}_k(t)$.
- There are repeated mentions to testing the "equality" of curves. This is confusing – rather, the test is if the data are consistent with there being exactly K homogeneous sets of curves.

  We appreciate the revision given by the referee. All the corrections have been made.

Also related to communication, I feel that the plots that shade the curves by their sample index are not nearly as informative as those shading by cluster. I would remove them. The fact that you can print all the sample IDs in the legend also highlights the fact that the applications are all performed on relatively simple data: I have seen much more complex collections of curves that would be interesting to cluster and visualize (two that come to mind immediately are home price data from Zillow or genomics time courses). This would also make a more exiciting illustration for your method.

Based on the first referee's suggestion, we have removed the plots without shading by cluster. Accordingly to the second one, the possibility of using the data suggested by the referee was investigated, particularly the home price data from Zillow. However, we have noted that this data is characterized by a temporal dependent variable. We note that, though we can apply our methods to such data, they are not recommended in practice. Regression models with temporal dependence are not yet implemented in this version of the package.

Overall, I recommend a weak accept, with revisions to the exposition. My decision to accept is based primarily on the functionality and modularity of the software package, and secondarily on the validity of the methodology and its relevance in modern statistics / data science. The reason I do not recommend a stronger accept is that the package interface could have been improved, the exposition could be refined, and the illustrations could be more compelling, as detailed above.

We are grateful for the reviewer consideration of this manuscript and software, and we also very much appreciate her/his many suggestions, which have been very helpful in improving the manuscript and the clustcurv software. We have made a considerable effort to take into account most of the interesting proposed suggestions. We hope that the reviewer finds this revised manuscript suitable for publication in R journal.

References

[1] CFJ Wu. Jackknife, bootstrap and other resampling methods in regression analysis (with discussions). *Annals of Statistics*, 14(10):1261–1350, 1986.