Review for R journal-2022-171

RCTS- Robust Clustering of TimeSeries with Interactive Fixed Effects

Thank you for submitting "RCTS: Robust Clustering of Time Series with Interactive Fixed Effects" to the R Journal. The authors present RCTS, an R package that implements a time series clustering method using interactive fixed effects. The package incorporates both the classical Ando and Bai algorithm and the robust Boudt and Heyndels method to cluster time series into specified groups. It also iteratively estimates factors, loadings, and the effects of observable variables, utilizing robust estimators capable of handling various types of outliers. The package addresses an important problem in the field.

However, there are areas in the paper and the package that require clarification. It is necessary to thoroughly check the language and grammar throughout the manuscript and improve the overall flow of the paper. Here are my specific comments for further improvement of the paper and the package:

1. I suggest incorporating a succinct statement within the introduction that explicitly emphasizes the motivation and significance of the work. This will effectively convey the underlying rationale and highlight the importance of the package to readers.

2. **Page 1. Introduction** How do you define outliers in the context of time series? As outliers lack a unified definition, it is crucial to establish a specific definition within the paper that aligns with the discussed context, thus ensuring improved clarity.

3. **Page 1. Introduction** *"... In the best case, only the time series with outliers are misclassified."* This statement is relevant to my second comment as well. However, it is important to note that this outcome depends on the primary objective of the study. In certain classification problems, such as fraud detection, one key objective may be to classify time series with anomalous behavior into a separate class. In such situations, it cannot be considered the best-case scenario if these time series are misclassified. The evaluation of the best-case scenario should align with the specific objectives and requirements of the classification problem at hand.

4. **Page 1. Intorduction** *"... For practical use a robust approach is thus needed. This robust approach has been developed in Boudt and Heyndels (2022)."* It would be beneficial for the reader to understand how this robustness is achieved. I recommend adding a brief explanation in the paper, summarizing the key aspects of the robust approach, and then referring readers to the original paper by Boudt and Heyndels for a more comprehensive understanding. This approach will make the paper self-contained and enhance the reader's comprehension of the robustness techniques employed.

5. **Page 1. Introduction** This package is introduced as a tool with time series clustering capabilities. However, it states that "*RCTS models the input panel data as such that it follows a factor structure:* $y_{it} = \ldots, (i = 1, ..., N, t = 1, ..., T)$,". Define the dimensions i as the spatial dimension and t as the time dimension. To enhance clarity, it would be beneficial to explicitly state whether the package is exclusively intended for handling balanced panel data or if it can also accommodate unbalanced panel data. Providing a clear definition of the package's scope will improve understanding and facilitate proper utilization.

6. Additionally, it is worth noting that even the example provided on page 2 is limited to balanced panel data. This also highlights the need for clarification regarding the package's capability to handle unbalanced panel data.

7. **Page 2 Guide** The default setting of the `create_data_dgp2` function is not explicitly mentioned in the description, making it difficult for the reader to understand the structure of the simulated dataset without referring to the package implementation. To improve the clarity and self-containment of the paper, I recommend describing the default configuration of the function within the paper itself, enabling readers to comprehend the structure of the input dataset in the example on page 2 without external references.

8. In the paper's description, the length of the time series is defined as "T," while in the package implementation, it is denoted as "TT." Ensuring consistency between the terminology used in the paper and the package implementation will greatly enhance the usability of the package.

9. **Page 3 Example** The example provided under the "Guide" section is not fully executable, as the code snippet `estimates <- estimate_algorithm(Y, X, S, k, kg, robust = TRUE)` generates an error message: `"Error in matrix(NA, nrow = (number_of_vars + 1), ncol = S) : object 'S' not found."` By making the example code complete and executable, readers will be able to replicate the analysis and gain a better understanding of the package's functionality.

10. **Page 5:** The presence of extensive code snippets with for loops on page 5 disrupts the flow of the section and makes the article challenging to read. It is only after navigating through the lengthy code that the concept of parallel computing is introduced. To enhance the reader's experience, I recommend providing a clear heads-up in the initial section about the implementation of parallel computing in the code. By doing so, readers can focus on understanding the algorithm's logic in that specific section, without being burdened by considering the time complexity of the code.

11. **Page 12 Conclusion:** To enhance the quality of the paper, it would be beneficial to include key assumptions underlying the proposed methodology and provide a commentary on the extensions of the research. Additionally, suggesting possible directions for further research would greatly contribute to the paper's significance and potential impact.

12. In the package implementation, the package description states that it includes a simulated dataset called "dataset_Y_dgp3." However, upon examination, this specific object is not found within the package. To ensure accuracy and consistency, it is essential to either remove the reference to the nonexistent object from the package description or provide the appropriate dataset within the package. This will help avoid confusion for users who rely on the package documentation for understanding and utilizing its features.

13. In the package implementation you write information messages to the console that cannot be easily suppressed. It is more R like to generate objects that can be used to extract the information a user is interested in, and then print() that object. Instead of print()/cat() rather use message()/warning() or if(verbose)cat(..) (or maybe stop()) if you really have to write text to the console. (except for print, summary, interactive functions)

    (Re `estimate_algorithm`)

14. **Page 6:** "*The required computational time lies an order of magnitude higher then for the classical implementation.* I believe this should be corrected to 'than.' It is important for the authors to carefully proofread the entire manuscript to ensure accuracy and maintain the quality of the paper.