

Reply to Reviewers

Wolfgang Rolke

2026-10-14

I thank the reviewers for their detailed comments and suggestions. Below I have answered them in the order in which they were given. The changes made to the R packages will appear in the new versions as soon as I can upload them to CRAN. Dr. Rolke

Required: 1. fixed line `nx+i` in `src/TS_cont.cpp` 2. Both examples are meant to illustrate the use of the packages, not as full data analyses. Both data sets are widely known and have been analysed many times. I added a short comment to the beginning of the section. I also added the adjusted p-value to the Iris data example and the argument `doMethods="all"` in both examples, so the methods used by `gof_test` and `gof_test_adjusted_pvalue` match. 3. The methods applied depend on the problem (`gof/twosample`) and data (continuous/discrete). The tables included in the paper show what methods are available for the different scenarios. Recommended: 4. All runs are now done with the same seed, so outputs should match 5. the adjusted p-value routines have two default simulation sizes 5000 and 1000, this is now explained in detail in the paper 6. the hybrid route calls `MD2sample` from `MDgof` and requires the routine `rnull` in order to use the parametric bootstrap. This is then used by `MD2sample`, as needed. 7. Including information on the Monte Carlo simulation error, for example providing confidence intervals for the p values, has been in the planning for a while. It has been implemented so far in the `Rgof` package for univariate data and will soon be implemented in `R2sample`, `MDgof` and `MD2sample`. 8. The packages `MDgof` and `MD2sample` use existing R implementations whenever suitable implementations are available. The statistics that I implemented myself were included precisely because suitable implementations were not available in R. Consequently, the proposed comparison with reference R packages is not feasible for these methods. Their mathematical definitions are based on the corresponding published literature, and the implementations follow those definitions. 9. The methods are all cited with the literature in which they were first proposed. Many of them do not have R implementations, and so I cannot give citations for their software. 10. The issue with `j=j-1` and `if(any(is.nan(TS_sim)))` has been fixed. 11. the “dangling else” has been fixed 12. see 4. 13. see 2.

1

14. I added a legend identifying the principal method labels used in the output.
15. I increased the font sizes in the `ggpairs` figure to improve the readability of its labels and correlation values
16. I did not include a separate type I error study because the p-values for the great majority of the implemented procedures are obtained by simulation under the null hypothesis. Thus, their calibration is built into the testing procedure, subject to Monte Carlo error and the correctness of the null simulation mechanism. The manuscript reports dozens of estimated rejection probabilities, including estimates obtained when the null hypothesis is true. With only 250 simulation replications, these estimates are subject to appreciable Monte Carlo variation. In interpreting the results, it is therefore important to consider the collection of estimates rather than treating each observed deviation from the nominal level as an independent indication of miscalibration. The largest deviation noted by the reviewer, 0.124, concerns the chi-square procedure, which is an exception to the simulation-based calibration because it relies on a classical distributional approximation. I agree that a dedicated size study could provide further information, particularly about the chi-square procedures. However, because the primary purpose of the article is to present the computational framework and its simulation and benchmarking capabilities, I have not added a separate size study.

17. The goal data is now included in the `replication_material.Rmd`.
18. I added a paragraph to that effect.
19. Seeded runs are also reproducible when parallel processing is enabled. The supplied seed is passed to the worker processes, so repeated runs with the same seed and parallel configuration reproduce the simulation results. This does however depend on the number of cores used, 23 in my case.
20. I have expanded the computational environment description in Section 8 to report R version 4.6.0, the Windows 11 build, and the versions of `mvtnorm`, `copula`, `Ball`, `energy`, `GGally`, and `Rcpp`. I have also reported the versions of `MDgof` and `MD2sample` used in the computations. Optional
21. The manuscript already explicitly identifies the evaluation of multivariate EDF statistics at observed sample points and in a single orientation as a computational approximation. This limitation means that discrepancies occurring at other evaluation points or in other orientations may not be captured by the computed statistic. I have retained this discussion because it already identifies the limitation raised by the reviewer.
22. I agree that guidance on the choice of test statistic would be valuable to applied users. However, the relative power of the available procedures depends on several aspects of the testing problem, including the null distribution, the nature and magnitude of the alternative, the dimension, and the sample size. The variety of possible null and alternative distributions makes it difficult to formulate reliable general recommendations. Such guidelines would be misleading if interpreted as general guidance for choosing a test.
23. I would like to clarify that the adjusted p-value calculation does not use nested simulation. The two simulation stages are performed serially. The first stage estimates the individual p-values, while the second estimates the null distribution of their minimum, which is used to obtain the adjusted p-value. Thus, the second stage does not rerun the full individual calibration procedure within each replication. The computational cost is associated with these two successive stages, rather than a nested simulation scheme. I have clarified this distinction in the manuscript.
24. I fixed the inconsistency.