Response to the Reviews, Round 2

Power Studies For Two-Sample and Goodness-of-Fit Methods For Multivariate Data

Wolfgang Rolke

November 09, 2026

Dear Matthias,

All required and recommended items have been addressed in the revised version; the optional sensitivity study is left for future work, as the letter allows. The changes are listed below in the order of the decision letter.

Required items

1. [methods-soundness] The MMD recommendation should name the MD2sample implementation.

Done. The recommendation itself now names the implementation in all places where it is made:

- *Abstract:* "In our case studies the aggregated MMD test as implemented in *MD2sample* is the best single method for the two-sample problem; it differs from the published MMDAgg test, and the result should not be assumed to carry over to it."
- *Section 4 (Highlights):* the bullet on the two-sample problem now refers to the aggregated MMD test as implemented in *MD2sample*, states that it differs from MMDAgg (Schrab et al., 2023) in how the statistics for the different bandwidths are combined and how the p value is found, and that the result should not be assumed to carry over to MMDAgg or to other implementations of the MMD test.
- *Section 4 (Recommendations):* a sentence after the list of recommended methods states that MMD refers to the *MD2sample* implementation throughout the paper, for the two-sample as well as the hybrid tests.
- *Section 5 (Methods, MMD test):* the description now states that the p value is found by the permutation method and how the test differs from MMDAgg.

2. [novelty-and-prior-art] Cite the precursor arXiv 2507.16630 and say what this paper adds.

Done. The introduction now cites Rolke (2025), arXiv:2507.16630, and states: "An earlier paper reported power studies for the multivariate two-sample methods. The present paper adds the goodness-of-fit and hybrid studies and the cross-validation of the recommended selections of methods."

Recommended items

3. [novelty-and-prior-art] Cite the companion software paper.

Done. The introduction now cites Rolke (2026), *Journal of Reproducible Statistics*, as the description of the packages: "The packages themselves are described in Rolke (2026a); here we report the power studies carried out with them."

4. [visualisation] Direct labels in Figures 8 and 10.

Done, for all six performance profile figures (Figures 5–10) so that they are consistent. In each figure the four methods with the highest share of case studies within 90% of the best are drawn in the Okabe–Ito

palette, which remains distinguishable under red-green colour vision deficiency, with distinct line types, and are labelled directly at the right end of their curves. All other methods are drawn in light grey, which removes the clutter where curves bunch near the top. The captions explain this. The code is in the function `profile_plot()` of *replication_material.Rmd*.

5. [writing-quality] Define nMC.

Done. Section 3, where the hybrid tests are introduced, now ends with: "We write nMC for the ratio of the size of the MC data set to the size of the real data set, so nMC=1 and nMC=5."

6. Reviewers 1, 2 and 3: the version of R2sample used for the univariate check.

R2sample 5.1.0 has since been released on CRAN, so the replication material no longer needs to carry its own copy. In addition:

- Section 6 now names the version at the check: "with the R package *R2sample*, version 5.1.0 (earlier versions give different results)".
- R2sample 5.1.0 is pinned in *renv.lock*, and Section 6.1 says so.
- *replication_material.Rmd* prints the version of R2sample at the start of the check and stops if it is not 5.1.0, so the knitted output documents which version produced the reported numbers (Kendall correlation 0.85 versus 0.86, best method within three points in 95% versus 97% of the case studies, overlap 79% versus 85%), which are reproduced exactly.

Optional item

7. [statistical-rigor] Sensitivity runs at two or three sample sizes and levels.

Declined for this paper. A full multivariate sensitivity study would require rerunning all case studies of both packages several times, about 23 hours of computing on 23 cores per configuration, and is left for future work. The paper states the invariance of the rankings as a working assumption (Section 6 and the Limitations), supports it with the univariate check, and states the rankings and recommendations as conditional on the configuration studied. The reviewers have accepted this framing.

Production item

8. [consistency-check] Section numbering in Section 7.

This was a numbering slip in the source, not an artefact of the extraction, and it is now fixed. The two-sample results in Section 7.2 now have four sibling subsections:

- 7.2.1 Continuous methods – overall results
- 7.2.2 Continuous methods – with or without equal marginals
- 7.2.3 Discrete methods – overall results
- 7.2.4 Discrete methods – with or without equal marginals

Kind regards,

Wolfgang