Reply to the Reviewers

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

Wolfgang Rolke

November 06, 2026

I thank the editor and the three reviewers for their careful reading and helpful comments. All required items have been addressed. Below, each point is followed by my reply; section and table numbers refer to the revised version. Because of the new Table 1 (type I errors), all other tables are numbered one higher than before.

Required

1. Ranking invariance (Reviewer 1 #1, Reviewer 2 #1, Reviewer 3 optional). The claim in section 6 is now stated as a working assumption: rankings observed for one configuration of sample size, level and alternative are assumed to hold for others, power curves can cross, and this was not verified systematically. The rankings and recommendations are stated to be conditional on the configuration studied, and the Limitations section repeats this. In addition, a check with the univariate package *R2sample* (100 case studies, 14 methods, rerun with $n = 200$, $\alpha = 0.1$ instead of $n = 500$, $\alpha = 0.05$) found rankings nearly as stable as between two replications of the same configuration: mean Kendall rank correlation 0.85 (replications: 0.86), best method within three points in 95% of the case studies (97%). The code is part of the replication material. This check was done with R2sample Ver 5.1.0, which is also part of the submission. CRAN still has Ver 5.0.0, which would yield different results. 5.1.0 will be uploaded to CRAN within a week or so.

2. Monte Carlo error in orderings and selections (Reviewer 2 #2; also Reviewer 1 #3, Reviewer 3 #4, editor). A new paragraph in section 7 gives the Monte Carlo error of single powers, of differences of two powers (about 3 points) and of mean powers (below 0.6 points), states that differences of about three points are not distinguishable, notes the upward bias of the maximum used as the "best" method, and explains how this affects the best selections. The description of the detailed tables (section 11) now says that methods within about three points should be regarded as tied.

While checking this we found that for most methods the power routines estimate the critical value from B null simulations, which adds a second Monte Carlo error of about the same size as the binomial one. The error statements above take this into account (see also the type I error item below).

3. Environment capture and determinism (Reviewer 3 #1). The replication material now includes a file *renv.lock* pinning R 4.6.0 and all package versions (including Rcpp, copula, spatstat, gTests and Ball), and *replication_material.Rmd* prints `sessionInfo()`. Section 6.1 describes the seeds (two-sample case study k: $2026 + k$; goodness-of-fit case study k of family f: $2026 + 1000f + k$). The seed is set inside each case study, so results do not depend on how case studies are distributed over the cores. The replication material reruns three case studies on 23 cores in parallel and compares them with the stored powers; all three are reproduced exactly.

Recommended

Type I errors per case (editor, Reviewer 1 #2, Reviewer 2 #3). New Table 1 gives for each study the range of the method means, the smallest and largest type I error of any method in any case study, and the number of cases above 6.4% together with the number expected by chance. The means are between

4.1% and 5.6% (discrete two-sample tests: 3% to 4.5%). Individual values range from 1.8% to 8.6%, a wider spread than binomial error alone would give, because of the simulated critical values; allowing for this, the number above 6.4% is close to the number expected by chance (238 vs 216 over all continuous studies), and no method or case study stands out. The full list of cases above 7% is in the replication material.

Conservative discrete two-sample tests (Reviewer 3 #5). Section 6.1 now explains that discrete tests cannot achieve the nominal level exactly and that the powers of these tests are, if anything, slightly understated.

qAD (editor). qAD is recommended only for two dimensions; the recommendation now says explicitly that it should not be used in five dimensions, where it is biased for several alternatives.

Bias of the "best" method (editor). Noted in the new Monte Carlo paragraph in section 7.

Scope of categorical claims (Reviewer 2 #5). The abstract, the highlights and the introduction to the recommendations now restrict these statements to our case studies.

Competing interests (editor). A standalone statement has been added.

NN vs NN0 (editor). The continuous test is now labeled NN0 where it is defined; the discrete NN test is identified as a different test.

Figures (editor). The performance profiles distinguish methods by line type as well as color. The captions of Figure 1 and Figure 11 now describe the axes, what each line or point represents, and (Figure 11) the line of equality, which was already drawn. Figures 2 and 4 show data sets, so their axes are the coordinates of the data and carry no units.

MMD vs reference mmdagg (editor). The MMD test in *MD2sample* is not a reimplementation of the MMDAgg test of Schrab et al. (2023); the two methods differ somewhat in how the statistics for the different bandwidths are combined and how the p value is found, so their statistics and p values are not expected to agree. The results for MMD refer to the implementation in *MD2sample*.

Baringhaus-Franz vs energy (editor). The Baringhaus-Franz statistic of *MD2sample* is one half of the energy statistic of `energy::eqdist.e`. For a test data set in the replication material the two agree exactly (1.415198), and both tests find p values by permutation.

Combined test with corrected p-value (Reviewer 1 #4). This is the workflow we recommend, and the packages provide `gof_test_adjusted_pvalue()` and `twosample_test_adjusted_pvalue()` for it. Its power requires a nested simulation for every case study and selection, which is beyond the scope of this revision. The combined test will have somewhat lower power than the best method in the selection, the price for not knowing in advance which method that is. We leave a systematic study to future work.

Equal vs unequal marginals in practice (Reviewer 1 #5). A paragraph has been added to section 7: since the alternative is unknown, users should use the overall selections, which were found from both kinds of case studies. We do not recommend choosing the methods after a preliminary test of the marginals on the same data, as this changes the level of the final test. If there is prior reason to believe the marginals are correct (as in copula modelling), a chi-square test with few bins should be included.

Covering selection vs top-k (Reviewer 2 #4). The cross-validation shows that the best selection does better with many case studies (goodness-of-fit D=2 and discrete), while with few case studies (five dimensions) the methods with the highest mean power do as well or better. Our recommendations combine the two: we start from the best selection, replace a method that enters only by winning a single case study and is weak on average (such as BG) by the method with the next highest mean power, and in five dimensions use the methods with the highest mean power. A paragraph stating this has been added to section 8.

Optional

KSD and FF cross-checks (editor). Left for later work; as the editor notes, FF is applied after the Rosenblatt transform, so an exact match is not expected.

Power curves / sensitivity study (Reviewer 1 #6, Reviewer 2 #6). Left for later work; the working assumption is now stated as such (required item 1).

Production items

- **Ball timing.** The timing table (now Table 2) was recomputed for the revision and now gives 6.4 seconds per data set on one core, in agreement with the text. The 20 seconds are the time per data set when 23 cores run at the same time; the sentence has been reworded to make this clear.
- **ES/EP.** Corrected to "79% for EP and 78% for ES".
- **References.** D'Agostino and Rayner corrected. In addition: the title of Thas (2010) is now *Comparing Distributions*; "Istituto" in the Kolmogorov reference; the journal of Schrab et al. (2023); the Rcpp version; the MMD is now cited with Gretton et al. (2012, JMLR); a missing reference (Dalitz 1953) was added; and the heading "Friedman-Rafski" is now "Friedman-Rafsky".