PEER REVIEW RECORD

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

Wolfgang Rolke

Vol. 1 (2026), Art. 17 · https://doi.org/10.68365/jrepstat.2026.17

The decision letters, the reports the authors read, their answers and the editor's signed decisions, as written. The reports come from the journal's review engine. The authors could have kept this record private until they confirmed the proof.

DECISION PATH

2026-11-06	Major revision
2026-11-06	3 answers
2026-11-09	Minor revision
2026-11-09	2 answers
2026-11-18	Accepted

Round 1: Major revision

SIGNED DECISION

Confirmed and signed by Matthias Templ on 2026-11-06 · ORCID 0000-0002-8638-5276

DECISION LETTER

Thank you for your submission to JRepStat. We are inviting a major revision. The reviewers' points follow, grouped by weight.

This paper compares a large collection of goodness-of-fit and nonparametric two-sample tests for multivariate data (two and five dimensions, continuous and discrete) across many simulated case studies, and distills small recommended method sets that are checked by leave-one-out cross-validation and shipped in two open R packages. It matters to practitioners who need a multivariate fit or two-sample test and want defensible guidance on which methods to run, and to methodologists who want a ready benchmark for new tests. We would be glad to see the revised version.

How to read this letter. The decision rests on the items marked required. If a required item seems wrong to you, or cannot be done within this paper, decline it on the revision page and give the reason. A reasoned decline is a legitimate answer, and the editor weighs it. Reviewers can be mistaken. Items marked recommended or optional are advice from the reviewers. Adopt them, decline them with a sentence, or leave them for later work. None of that counts against the paper. The reviewer reports below carry the same marks. The editor who signs this decision may weigh an item differently in the accompanying message, and that message prevails.

REQUIRED (the decision rests on these):

- [Reviewer 1] The claim in section 6 that method rankings almost never change with sample size, type I error, or the alternative parameter carries the whole single-run design. Either show it for a few representative cases at two or three sample sizes, or restate it as the author's working assumption and limit the conclusions accordingly.
- [Reviewer 2] The detailed tables present fine-grained orderings and the "best" and "within 90% of best" selections from single point estimates, with no account of the Monte Carlo error you already quote (up to 1.6% per estimate, so differences of two methods are uncertain to about 2.3 points). State that orderings within that band are not distinguishable and reflect it in the selection language.

- [Reviewer 3] Add an environment capture (sessionInfo or a lockfile pinning R 4.6.0 and the versions of Rcpp, spatstat, gTests, Ball and other dependencies) and a determinism attestation stating that re-running the replication scripts, including the 23-core parallel run, reproduces the stored powers.

RECOMMENDED:

- [methods-soundness] Type I error control is summarised as a mean over case studies between 4% and 6% per method. A mean can hide individual case studies where a method rejects well above 0.05, which would make its power figures for those cases not comparable to a correctly sized test. Section 10 already records qAD rejecting below the nominal level in five dimensions. Please state whether any method and case study combination has an actual type I error materially above 0.05, so the reader knows the power entries for those cells are on a like-for-like footing. The per-case type I errors are stored in the packages.
- [methods-soundness] qAD is recommended for continuous goodness-of-fit in two dimensions (Section 4) yet Section 10 reports it has power below the nominal level for several five-dimensional alternatives, so it is biased there. Please add one sentence to the recommendation noting the dimensions and alternatives where qAD should not be used, so the recommendation and the limitation agree.
- [software-landscape] The headline conclusion for the two-sample problem is that the MMD test is the best single method, computed by a reimplementation inside MD2sample (MD2sample:::mmdagg) rather than the reference mmdagg of Schrab et al. (2023) that the paper cites. Please add a short table showing the MMD statistic and p value from MD2sample and from the reference mmdagg on one shared data set, so a reader who reaches for the published package knows the recommendation transfers. The nominal type I error table in Section 6.1 shows calibration but does not show that the two implementations agree on the same input.
- [software-landscape] The energy-style two-sample tests (Baringhaus-Franz, Aslan-Zech, Biswas-Ghosh) are reimplemented, while the Baringhaus-Franz energy statistic is available in the CRAN package energy. A one-data-set comparison of the Baringhaus-Franz statistic from MD2sample against energy would place this family against the established implementation at low cost.
- [statistical-rigor] The "best method" used as the reference for "percentage close to best" and "mean shortfall from best" is the maximum power over 11 to 20 methods per case study. A maximum of noisy estimates is biased upward, which inflates every shortfall and makes the 90% threshold a moving target. Note this upward bias where the three summary statistics are defined, and consider reporting the shortfall against the second-best as a check.
- [visualisation] The performance profiles (Figures 5-10) show up to about 17 methods as curves of share against τ. With that many series a single colour palette will not stay distinguishable, and several curves overlap near the top. Please confirm the curves are separated by line style or direct labels as well as colour, and that the palette is legible to readers with red-green colour vision deficiency. The two-sample D=2 and discrete profiles carry the most series and benefit most.
- [visualisation] Figure 1 compares the chi-square test with i×i bins for i=3,5,7,9 over 20 case studies, and the caption states the conclusion rather than the axes. Please state in the caption what the vertical axis is (power in percent) and what each point or line represents (one case study, or the four bin counts), so the figure reads without the body text.
- [visualisation] For Figures 2, 4 and 11, please confirm axes carry labels and units, and that Figure 11 marks the line of equality so the reader can see at a glance where discretization gains or loses power. The diagonal reference supports the claim that discretizing usually costs some power.
- [writing-quality] The simple nearest-neighbor test is defined as "NN" in Section 5.4.8, but the results tables and Table 1 label it "NN0" (while "NN" reappears for the discrete two-sample test in Table 17). State the "NN0" label where the test is defined, or use one name throughout, so readers can map the table columns to the method descriptions.
- [ethics-and-integrity] The author of the paper is also the author of the two R packages (MD2sample, MDgof) being compared and the person who chose the case studies. Section 10 states this and notes that the cross-validation checks only the author's own collection, so the competing interest and its main limitation are already in the text. Add a short competing interests statement making the package

authorship and funding explicit as a standalone declaration, rather than only within the Limitations section.
- [Reviewer 1] Report the spread of the true type I errors per method, not only the 4% to 6% mean, and flag any case where a method sits above nominal, so no recommended method is credited with power that rests on an inflated level.
- [Reviewer 1] Report the power of the combined test with corrected p-value for the recommended selections, since that is the workflow a practitioner follows when told no single method is reliable.
- [Reviewer 1] Add practical guidance on how a user picks among the equal-marginal and unequal-marginal recommendations, given that the true alternative is unknown in practice.
- [Reviewer 2] The cross-validation shows the top mean-power methods match or beat the smallest-covering "best selection" in five dimensions. State which procedure you actually recommend and why.
- [Reviewer 2] Scope the categorical statements in the abstract and highlights (for example "the MMD test is the best single method") to the collection of case studies used here.
- [Reviewer 3] Report the Monte Carlo uncertainty on the summary mean powers so close rankings are read with their spread.
- [Reviewer 3] When comparing power for the conservative discrete two-sample tests, note that their actual type I errors (3% to 4.5%) sit below the others.

OPTIONAL:

- [software-landscape] The Kernel Stein Discrepancy and the Fasano-Franceschini test both have external reference implementations (kgof in Python, fasano.franceschini.test on CRAN). A cross-check against these would further document agreement, but the paper modifies FF by applying it after the Rosenblatt transform, so an exact match is not expected and this can be left for later work.
- [Reviewer 1] A full set of power curves over sample size and effect size would settle the invariance question and show where rankings cross.
- [Reviewer 2] A sensitivity study across two or three sample sizes for the full collection would settle the stability question directly.

We will correct these in production, and you confirm them in the proof:

- [consistency-check] Section 6.1 states the Ball divergence test "takes about 7 seconds per data set on a single core (table 1), and about 20 seconds when 23 cores run in parallel". Table 1 lists the Hybrid nMC=5 D=2 Ball time as 20 seconds, and its caption states all table times are for a single core. The 7 seconds disagrees with the cited table, and the single core vs 23 core figures are reversed relative to the table. Reconcile the text with the 20 second value in Table 1.
- [consistency-check] The narrative following Table 12 reports mean power "79% for ES and 78% for EP" for the equal marginals case. Table 11 lists EP at 79 and ES at 78. The two labels are transposed in the text.
- [writing-quality] Two cited author names appear misspelled relative to the works they name: "D'Agostini, R. B." for the goodness-of-fit volume reads as D'Agostino, and "Raynor, J. C." for Smooth Tests of Goodness of Fit reads as Rayner. Both occur in text and in the reference list. Confirm the correct spellings in the proof.

REVIEWS

Report Reviewer 1, recommends minor

Written by the journal's review engine in the role of the application reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

This is a useful and unusually thorough piece of applied benchmarking. The problem is real: a practitioner with a multivariate data set and a candidate model has little guidance on which fit or two-sample test to use, and the available software is thin for more than one dimension. The paper answers the practical question directly, gives small recommended method sets, checks them by cross-validation, and ships everything in packages on CRAN with seeds and replication scripts. The headline findings read as credible and are stated with appropriate care: MMD as the strongest single two-sample method, chi-square with a coarse grid when marginals match under null and alternative, and a hybrid MMD test overtaking the

goodness-of-fit methods in five dimensions. The limitations section is clear about what the study does and does not cover.

My main concern is the design decision in section 6. One power study is run per case, justified by the statement that a change in the ranking of methods for a different sample size, type I error, or alternative parameter is exceedingly rare. That is a sweeping empirical claim, and it is the foundation for every conclusion in the paper, yet it rests on stated experience. The author also notes that power curves can cross, which is exactly the situation that would overturn a ranking. A benchmarking reader will want either a small demonstration, re-running a handful of representative cases at two or three sample sizes to show the ranking holds, or the claim restated as a working assumption with the conclusions scoped to it. The re-run uses existing code and existing case studies, so the cheap path and the demonstration path are both within reach.

A fair power comparison depends on every method holding the nominal level in each case, and the paper reports only the mean type I error over cases (4% to 6%), plus the note that qAD is conservative in five dimensions and the discrete tests are conservative. The conservative direction understates power and is the safe failure. The unreported risk is a method that is liberal in a specific case and is then credited with power it did not earn there. The full type I errors already sit in the package data, so reporting the spread, or flagging any case and method above nominal, would let a reader trust the specific recommendations. In the same spirit, the rankings are presented as point estimates, and at B=1000 the Monte Carlo standard error is about 1.4 to 1.6 points, so adjacent entries that differ by two or three points (ES 75 against EP 72 in Table 6, for instance) are not separable. One sentence carrying that uncertainty into the ordering would keep readers from over-reading the tables.

The most valuable addition would speak to the paper's own conclusion. The central message is that no single method can be relied on and that a user should run a small selection. The "percentage within 90% of the best" metric rewards a selection in which at least one method does well, but a practitioner cannot know which of the selected methods is the good one for an unknown alternative. The natural response is to run the whole selection and combine the p-values with a multiplicity correction, which the packages already support. Reporting the power of that combined test for the recommended selections would turn the guidance into the workflow a user actually follows, and would test the selection idea on its own terms. Related, the conditioning on equal versus unequal marginals is informative for understanding but harder to act on, because the alternative is unknown. A sentence on how a user navigates this, for example testing the marginals first and then choosing the corresponding recommendation, would help.

The honesty about external validity is a strength. The case studies were chosen by the author, who also wrote the packages, the cross-validation stays within the collection, and the author notes that the non-independence of the cases (shared copulas) means leave-one-out probably overstates how far the selections carry. That framing is right, and it is the correct reason to read the recommendations as "for problems like these" rather than as universal. The reproducibility materials are strong, and I did not find numeric contradictions in the tables I could check against the text.

Numbered suggestions to the authors:

1. REQUIRED The ranking-invariance claim in section 6 carries the single-run design. Either show it holds by re-running a few representative cases at two or three sample sizes, or restate it as your working assumption and scope the conclusions to the one configuration studied.
2. RECOMMENDED Report the spread of the true type I errors per method, not only the 4% to 6% mean, and flag any case where a method exceeds nominal, so no recommendation rests on an inflated level.
3. RECOMMENDED Carry the Monte Carlo uncertainty of the power estimates into the rankings, noting that differences within about two percentage points are not distinguishable at B=1000.
4. RECOMMENDED Report the power of the combined test with corrected p-value for each recommended selection, since that is the workflow implied by the finding that no single method is reliable.
5. RECOMMENDED Add guidance on how a user chooses among the equal-marginal and unequal-marginal recommendations when the alternative is unknown, for example by testing the marginals first.

6. OPTIONAL A set of power curves over sample size and effect size would settle the invariance question directly and show where method rankings cross.

Report Reviewer 2, recommends minor

Written by the journal's review engine in the role of the methodology reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

This is a careful and unusually transparent benchmarking paper. It assembles a large set of goodness-of-fit and two-sample methods for multivariate data, many of them with no other R implementation, runs power studies over a wide range of case studies, and distills the results into small recommended sets of complementary methods. The reproducibility posture is strong: seeds, versions and settings are stored with the results, and scripts that produce every table and figure are shipped. The limitations section is honest about the central threats, that one author chose the case studies and wrote the packages, and that cross-validation checks only within the collection. My comments are about two methodological choices that the conclusions lean on.

The first is the decision to run one power study per case. Section 6 defends this with the claim that a change in ranking for a different sample size, type I error or alternative parameter is "exceedingly rare." This is the load-bearing assumption of the paper, and it is offered as experience, not as evidence. Power curves can cross, as the text itself allows, and the alternative parameters were tuned so the best method sits near 80% to 90% power, which is exactly the region where close competitors can reorder. Please support the claim with a small re-run: take a handful of cases, run them at a second sample size, and show the ranking holds. If that is more than you want to do here, qualify the sentence to an observation and state that the rankings are conditional on the chosen configuration. The existing code makes the small check cheap.

The second is Monte Carlo uncertainty. You quote the standard error correctly, up to 1.6% at B equal to 1000, so the difference between two power estimates carries an uncertainty of roughly 2.3 points. The detailed tables in Section 11 then present methods in a strict order, and the "percentage close to best" and the "best selection" are both decided by comparing single point estimates against a threshold. A method that is "best" in one case study, for example BG in NormalStretchM, may be best by an amount inside that band, and it then enters a selection on the strength of noise. The headline conclusion, that no single method is reliable and a small set is needed, survives this. The specific membership of the selections does not, and you already see the symptom in the cross-validation, where the least stable study is the one with 17 cases. Please state that orderings within Monte Carlo error are not distinguishable and let that caveat govern how the selections are read.

Two smaller points follow from the same concern. Power is only comparable among methods that hold the level, and you report the type I error as a mean between 4% and 6% over cases. A mean can hide a single case study where a method is anticonservative, and the qAD bias in five dimensions shows such cases exist. The per-method range or maximum would let a reader confirm the comparison is fair. Separately, the cross-validation in Table 20 shows that simply taking the top mean-power methods matches or beats the smallest-covering selection in five dimensions (regret 0.0 against 0.2 for the goodness-of-fit study, 2.8 against 3.7 for the two-sample study). This is a useful and honest result, and it raises a question the paper should answer directly: which construction do you recommend a user adopt, the covering selection or the top-k by mean power? The categorical phrasing in the abstract and highlights ("the MMD test is the best single method") should also be read against the collection it was measured on, so a one-clause scope would keep those statements accurate.

None of this unsettles the main message or the value of the packages. The methods, the performance profiles, the cross-validation and the stored type I errors together make a genuinely useful reference.

Numbered suggestions to the authors:

1. REQUIRED Support the Section 6 claim that rankings are stable across sample size, alpha and alternative parameter by re-running a handful of cases at a second sample size and showing the rankings hold, or qualify the claim to an observation and state that the rankings are conditional on the chosen configuration.

2. REQUIRED Add that orderings within Monte Carlo error (about 2.3 points for a difference of two power estimates) are not distinguishable, and let that caveat govern how the "best" and "within 90% of best" selections are read.
3. RECOMMENDED Report the range or maximum type I error per method over the case studies alongside the mean, so power is seen to be compared among methods that hold the level in each case.
4. RECOMMENDED State which construction you recommend in practice, the smallest-covering best selection or the top methods by mean power, given that the cross-validation shows the latter is competitive or better in five dimensions.
5. RECOMMENDED Scope the categorical claims in the abstract and highlights, such as "the MMD test is the best single method," to the collection of case studies used here.
6. OPTIONAL A sensitivity study across two or three sample sizes for the full collection would settle the stability question directly.

Report Reviewer 3, recommends minor

Written by the journal's review engine in the role of the software reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

Read as a software and reproducibility submission, this is a strong piece of work. The methods live in two CRAN packages at stated versions (MDgof 1.3.0, MD2sample 1.5.0) under R 4.6.0, each case study carries its own seed, the seeds and settings are stored with the results, the true type I errors are reported alongside the powers, and the scripts that build every table and figure are declared as part of the replication material. The Monte Carlo standard error is quantified (B=1000, at most 1.6 percentage points on a single estimate), and the computing-time table lets a reader plan a re-run (about 23 hours, or 53 hours with Ball). The deduplication of FR, CF1 and CF3 is handled correctly and stated. This is more reproducibility scaffolding than most simulation papers provide.

My main ask follows from what reproducibility means for a paper whose results are its code's output. The text says a single result can be reproduced from its stored seed, but the headline numbers come from a 23-core parallel run, and parallel scheduling is the usual place where per-task seeding stops being deterministic. Please add an environment capture (sessionInfo or a lockfile that pins the dependency versions, since Rcpp, spatstat, gTests and Ball all affect the numbers) and a short attestation that re-running the scripts, including the parallel run, reproduces the stored powers. You appear to have everything needed for this already.

One number does not match its table. Section 6.1 cites table 1 for "7 seconds per data set" for Ball at nMC=5 in two dimensions, while table 1 lists 20 seconds, and the same sentence then reports "about 20 seconds when 23 cores run in parallel", which reads as slower under parallelism. Please reconcile the figure and the direction of the comparison. Two reference author names also look wrong and are worth a check against the sources: the 1986 "Goodness-of-Fit Techniques" with Stephens reads as D'Agostino rather than D'Agostini, and the 2012 "Smooth Tests of Goodness of Fit" reads as Rayner rather than Raynor.

On the inference the rankings carry, the large gaps are safe. MMD at mean power 89 against the field in two dimensions, and 82 against 69 for the next method in five dimensions, is well clear of Monte Carlo noise, so the headline recommendation of MMD as the best single two-sample method holds. The fine distinctions are the ones a reader should not over-read. ES at 75 and EP at 72, for instance, differ by about the size of the per-estimate Monte Carlo error, and the "percentage close to best" and "best selection" quantities inherit that noise through the arbitrary 90% cutoff. Reporting the spread on the summary mean powers would let a reader see which orderings are firm. The performance profiles already help here. Separately, the discrete two-sample tests run at actual type I errors of 3% to 4.5%, below the others, so their power sits on a slightly different footing; one sentence at the power comparison would flag that. The claim that rankings almost never change with sample size, alpha or the alternative parameter is the one place where a single configuration per case does real work, and it rests on stated experience rather than evidence in the paper; a small re-run at a second configuration would settle it, though that is beyond what this paper needs.

I was given the manuscript alone. The replication scripts, the package vignettes, the stored result data sets and the figures were not provided to me, so I could not run or inspect them, and I have raised no ask about their contents.

Numbered suggestions to the authors:

1. REQUIRED Add an environment capture (sessionInfo or a lockfile pinning R 4.6.0 and the versions of Rcpp, spatstat, gTests, Ball and other dependencies) and a determinism attestation that re-running the replication scripts, including the 23-core parallel run, reproduces the stored powers.
2. [editor_fix] Reconcile the Ball divergence timing: table 1 shows 20 seconds for nMC=5, D=2, while section 6.1 cites table 1 for 7 seconds per data set and then reports 20 seconds under 23-core parallelism.
3. [editor_fix] Check two reference author names against the sources: D'Agostino (not D'Agostini) for the 1986 "Goodness-of-Fit Techniques", and Rayner (not Raynor) for the 2012 "Smooth Tests of Goodness of Fit".
4. RECOMMENDED Report the Monte Carlo uncertainty on the summary mean powers so that close orderings such as ES against EP are read with their spread.
5. RECOMMENDED Note at the discrete two-sample power comparison that those tests run at actual type I errors of 3% to 4.5%, below the other methods.
6. OPTIONAL Re-run a few case studies at a second sample size or alternative parameter to support the statement that method rankings almost never change across configurations.

REVISION, RECEIVED 2026-11-06

Authors confirm each required point as done or give their reason for declining it. Further answers are voluntary, because the review compares every revision with the previous version and the next round's reports say what remains open.

> REQUIRED The claim in section 6 that method rankings almost never change with sample size, type I error, or the alternative parameter carries the whole single-run design. Either show it for a few representative cases at two or three sample sizes, or restate it as the author's working assumption and limit the conclusions accordingly.
>
> **Done.**

> REQUIRED The detailed tables present fine-grained orderings and the "best" and "within 90% of best" selections from single point estimates, with no account of the Monte Carlo error you already quote (up to 1.6% per estimate, so differences of two methods are uncertain to about 2.3 points). State that orderings within that band are not distinguishable and reflect it in the selection language.
>
> **Done.**

> REQUIRED Add an environment capture (sessionInfo or a lockfile pinning R 4.6.0 and the versions of Rcpp, spatstat, gTests, Ball and other dependencies) and a determinism attestation stating that re-running the replication scripts, including the 23-core parallel run, reproduces the stored powers.
>
> **Done.**

The authors' full response letter, as uploaded, follows as Appendix A.

Round 2: Minor revision

SIGNED DECISION

Confirmed and signed by Matthias Templ on 2026-11-09 · ORCID 0000-0002-8638-5276

The editor re-weighted 2 points.

DECISION LETTER

Thank you for the revised version of your manuscript. This is round 2, and we are inviting a further minor revision. Because you were asked for a major revision, this round read the new version whole, and the editor has read your response against it. What follows is our reading of this version. An item carried over from the previous round says so.

This paper compares a large collection of goodness-of-fit and nonparametric two-sample tests for multivariate data (two and five dimensions, continuous and discrete) across many simulated case studies, distills small recommended method sets checked by leave-one-out cross-validation, and ships everything in two open R packages with seeds and scripts. It matters to practitioners who need a multivariate fit or two-sample test and want defensible guidance on which methods to run, and to methodologists who want a ready benchmark for new tests. We would be glad to see the revised version.

How to read this letter. The decision rests on the items marked required. If a required item seems wrong to you, or cannot be done within this paper, decline it on the revision page and give the reason. A reasoned decline is a legitimate answer, and the editor weighs it. Reviewers can be mistaken. Items marked recommended or optional are advice from the reviewers. Adopt them, decline them with a sentence, or leave them for later work. None of that counts against the paper. The reviewer reports below carry the same marks. The editor who signs this decision may weigh an item differently in the accompanying message, and that message prevails.

REQUIRED (the decision rests on these):

- [methods-soundness] The headline "the MMD test is the best single method" rests on MD2sample's aggregated MMD, which the response states differs from the published MMDAgg of Schrab et al. (2023) in how bandwidth statistics are combined and how the p value is found. The abstract and Section 4 should name this as the MD2sample implementation in the recommendation itself, so a reader reaching for the cited package does not assume the result transfers.
- [novelty-and-prior-art] The author's own direct precursor, "Power Studies For Two-sample Methods For Multivariate Data" (Rolke, 2025, arXiv 2507.16630), covers the two-sample multivariate power studies that form roughly half of this manuscript, yet it is absent from the reference list (which cites only Rolke and Gutierrez Gongora, 2021). Cite it and say in one sentence what this paper adds beyond it, namely the goodness-of-fit and hybrid studies and the cross-validation.

RECOMMENDED:

- [novelty-and-prior-art] The companion paper describing the same two packages, "A Computational Framework for Multivariate Goodness-of-Fit and Two-Sample Testing: MDgof and MD2sample" (Rolke, 2026, J. Reproducible Statistics), is also uncited. Cite it so a reader can separate the software contribution from the power results reported here.
- [visualisation] I am reading extracted text, not the rendered figures. I cannot confirm from the images that the performance profile palette is legible under red-green colour vision deficiency, or that the line-style separation holds where curves bunch near the top. Direct-label the three or four highest curves in Figures 8 and 10, the two-sample D=2 and discrete profiles that carry the most series, so the leading methods read without tracing colour or dash pattern.
- [writing-quality] The symbol nMC appears in the headings and body of Tables 1 and 2 and in Section 11 (for example "Hybrid, nMC=1, D=2") without a one-line definition. Section 3 and Section 4 describe the MC data set as "equal to" or "five times" the real data in prose, so the reader can infer nMC is that ratio, but the symbol itself is never stated. Please define nMC where the hybrid tests are introduced (Section 3) so the table labels read without inference.
- [Reviewer 1] Confirm that the R2sample version used for the cross-check (5.1.0) is the copy archived with the paper, since the text notes CRAN still carries 5.0.0 which gives different numbers.
- [Reviewer 2] State near the Section 6 check that the R2sample version carrying it (5.1.0) is bundled in the replication material and not yet on CRAN, so a reader knows which version reproduces the stability numbers.

OPTIONAL:

- [statistical-rigor] A sensitivity run of the full multivariate collection at two or three sample sizes and levels would convert the stability from a working assumption into a direct result, and would show where the power curves of the recommended methods cross; this can be left for later work.

We will correct these in production, and you confirm them in the proof:

- [consistency-check] Subsection 7.3.1 "Discrete methods - overall results" is nested under 7.3 "Continuous methods - with or without equal marginals", so the discrete overall results sit beneath a

continuous-labelled parent heading. This reads as a numbering slip (7.3.1 should be a sibling section, with 7.4 then covering discrete with and without equal marginals). It may be an artefact of heading extraction. Confirm the section numbering in the proof.

REVIEWS

Report Reviewer 1, recommends accept

Written by the journal's review engine in the role of the application reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

Of the one required point and five lighter ones I raised, the required point and four of the recommended or optional ones are met, and one recommended point is declined with a reason I accept.

The required point was the ranking-invariance claim that carried the single-run design. The revision now frames it as a working assumption, scopes the rankings and recommendations to the configuration studied, and supports it with a univariate cross-check: across the 100 R2sample two-sample studies the mean Kendall correlation between the two configurations was 0.85 against 0.86 for two replications of the same configuration, and the best method agreed within three percentage points in 95% of studies against 97% for replications. The cross-check is on univariate methods, not the multivariate methods that are the subject of the paper, and the text says as much. For a working-assumption framing this is the right level of evidence, and the honesty about what was and was not verified is appropriate. This point is settled.

On the type I error spread, Table 1 now reports the range of method means (4.1% to 5.6%), the extreme individual values (1.8% to 8.6%), and the count of method-by-case combinations above 6.4% against the number expected by chance. The revision also identifies a second Monte Carlo source, the simulated critical value, and folds it into both the error statements and the expected counts. Two of the studies sit somewhat above their expected exceedance count (goodness-of-fit continuous D=2, 63 against 45.2; discrete, 50 against 37.8), while the overall total is close (238 against 216). The author states that no method or case study stands out and gives the full list of cases above 7% in the replication material, which lets a reader check the recommended methods directly. That is enough to trust the specific recommendations, and I do not ask for more.

The Monte Carlo uncertainty is now carried into the text: the new paragraph in section 7 gives the standard error of single powers, of differences (about three points), and of mean powers (below 0.6 points), notes the upward bias of the maximum used as the best, and the detailed tables state that methods within about three points should be read as tied. The equal-versus-unequal marginal guidance in section 7 now points users to the overall selections and explains why a preliminary test of the marginals on the same data is not recommended. Both points are addressed.

The one declined point is the power of the combined test with corrected p-value for each recommended selection. The author declines on the ground that this needs a nested simulation for every case study and selection, notes that the combined test will have somewhat lower power than the best method in the selection, and provides the functions that compute it. The cost is real and the point was a strengthening one, so the decline is reasonable and I accept it. A systematic combined-power study would make a good follow-up paper rather than a condition on this one.

The remaining reproducibility item is small. The cross-check result depends on R2sample 5.1.0, which the text says is not yet on CRAN, so a reader reproducing it against the public package would get different numbers. The submission includes 5.1.0 and pins it in renv.lock, so confirming that the archived copy is the one used closes this.

Numbered suggestions to the authors:

1. RECOMMENDED Confirm that the R2sample 5.1.0 used for the univariate cross-check is the version archived with the replication material, since the text notes CRAN still carries 5.0.0 and that it yields different results.

Report Reviewer 2, recommends accept

Written by the journal's review engine in the role of the methodology reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

Of the two required and three lighter points I raised, all are met, and the one optional item is declined with a reason I accept.

The ranking-stability question was the load-bearing one. Section 6 now states the invariance as a working assumption, notes that power curves can cross, and says the multivariate case was not verified systematically, with the rankings and recommendations declared conditional on the studied configuration. That qualification alone was the low-effort route I offered, so the point is settled. The author also went further and ran an empirical check on the univariate R2sample suite (100 case studies, 14 methods) at a second sample size and level: mean Kendall rank correlation 0.85 against 0.86 for two replications of the same configuration, best method within three points in 95% of cases against 97%, and 79% overlap of the within-90% sets against 85%. The check is univariate and the text says so plainly, so it supports the assumption without overclaiming for the multivariate methods. That is the right weight to give it.

The Monte Carlo point is handled more thoroughly than I asked. The new paragraph in Section 7 gives the standard error of a single power, of a difference of two powers, and of a mean power, and the detailed tables in Section 11 now treat methods within about three points as tied. In working through this the author found a second error source I had not flagged: for most methods the critical value is itself the 95% quantile of B simulated statistics, which roughly doubles the variance of the estimated type I error and widens the difference band from about 2.3 to about 3 points. That is a real correction and it is carried consistently into the new Table 1 and into how the best selections are read.

The three recommended points are also in. New Table 1 reports the mean range, the smallest and largest type I error of any method in any case study, and the count above 6.4% against the number expected by chance, so a reader can see that power is compared among methods that hold the level, with the discrete two-sample tests noted as conservative. Section 8 states the recommended construction directly: 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) with the next method by mean power, and in five dimensions use the methods with the highest mean power. The abstract, highlights and recommendations now scope the categorical claims to the case studies used here.

The reproducibility posture is now complete for this journal: renv.lock pinning R 4.6.0 and the dependencies, sessionInfo printed, the seed scheme documented and set inside each case study, and three case studies reproduced exactly on 23 cores. One small gap remains around the univariate stability check, since the R2sample version that produces its numbers (5.1.0) is bundled in the replication material but not yet on CRAN. Naming that where the check is reported would tell a reader which version to use. The paper stands without it.

Numbered suggestions to the authors:

1. RECOMMENDED Note at the Section 6 stability check that it uses R2sample 5.1.0, included in the replication material and not yet on CRAN, so a reader knows which version reproduces the reported rank correlations and overlaps.

Report Reviewer 3, recommends accept

Written by the journal's review engine in the role of the software reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

Of the one required point, the two production fixes and the three lighter items I raised, all are met, and the required reproducibility work is done well. The environment is now captured in renv.lock (R 4.6.0 plus Rcpp, copula, spatstat, gTests and Ball), replication_material.Rmd prints sessionInfo(), the seeds are defined inside each case study so parallel scheduling does not change the result, and three case studies rerun on 23 cores reproduce the stored powers. That seeding design is the right answer to the parallel-determinism concern I had, because it removes the dependence on how tasks are distributed over cores rather than papering over it. The Ball timing is reconciled: Table 2 gives 6.4 seconds per data set on one core, and the text reports about 20 seconds per data set when 23 cores contend for memory, which is now the stated direction. The two reference names are corrected.

The Monte Carlo treatment went further than I asked. Section 7 now separates the binomial error of a single power from the error that the simulated critical value adds for the methods that estimate it, reports about three points for the difference of two powers and below 0.6 points for a mean power, and flags the

upward bias of the maximum used as the best method. The new Table 1 carries this through to the type I errors, and the counts are internally consistent: the "above 6.4%" column sums to 238 over the continuous studies against 216 expected, matching the text, and the discrete two-sample means of 3% to 4.5% are now stated with the note that those powers are, if anything, slightly understated. The claim about ranking invariance is now framed as a working assumption in section 6 and repeated in the Limitations, with the recommendations stated as conditional on the configuration studied. The univariate R2sample check (Kendall rank correlation 0.85 between configurations against 0.86 between replications, best method within three points in 95% of cases against 97%) is a reasoned partial answer to the optional item. It is univariate and the multivariate assumption stays unverified, and the paper says so, so I find the decline to run a full multivariate sensitivity study convincing for this paper.

One point is newly visible and worth a short note. The R2sample check that now supports the working assumption was run with R2sample 5.1.0, while CRAN holds 5.0.0, which the author states would give different results. The package source is in the submission, so the check is reproducible from the replication material, but renv.lock as described lists only the dependencies of the two main packages, and section 6 does not name the R2sample version. Pinning 5.1.0 in the lockfile and naming it in the text closes the gap, so a reader who restores the environment reproduces the 0.85 and the 95% rather than numbers from the CRAN build. This supports a working assumption, not a headline result, so it does not change the grade.

I could read renv.lock, the scripts and the stored data sets only through the manuscript's description of them, so my check of the reproducibility claims is textual. On the evidence in front of me the paper is in good shape.

Numbered suggestions to the authors:

1. RECOMMENDED Pin R2sample 5.1.0 in renv.lock and state the version in section 6, since the cross-configuration check depends on it and the CRAN version (5.0.0) would produce different numbers.

REVISION, RECEIVED 2026-11-09

Authors confirm each required point as done or give their reason for declining it. Further answers are voluntary, because the review compares every revision with the previous version and the next round's reports say what remains open.

> REQUIRED The headline "the MMD test is the best single method" rests on MD2sample's aggregated MMD, which the response states differs from the published MMDAgg of Schrab et al. (2023) in how bandwidth statistics are combined and how the p value is found. The abstract and Section 4 should name this as the MD2sample implementation in the recommendation itself, so a reader reaching for the cited package does not assume the result transfers.

Done.

> REQUIRED The author's own direct precursor, "Power Studies For Two-sample Methods For Multivariate Data" (Rolke, 2025, arXiv 2507.16630), covers the two-sample multivariate power studies that form roughly half of this manuscript, yet it is absent from the reference list (which cites only Rolke and Gutierrez Gongora, 2021). Cite it and say in one sentence what this paper adds beyond it, namely the goodness-of-fit and hybrid studies and the cross-validation.

Done.

The authors' full response letter, as uploaded, follows as Appendix B.

Round 3: Accepted

SIGNED DECISION

Signed by Matthias Templ on 2026-11-18 · ORCID 0000-0002-8638-5276

Accepted. The revision answers the reviewers' required points. The journal reran the replication material: 34 of 35 tables are identical, the 35th holds computing times, and three stored case studies of the full power studies reproduce exactly.

DECISION LETTER

Congratulations, your manuscript has been accepted for publication in JRepStat. Production begins now.

Where it is not on file yet, we still need the following for production.

- a valid ORCID iD for every author
- a captured computational environment (dependencies/lockfile/container)
- a signed determinism attestation for the entry-point command

REVIEWS

Report Reviewer 1, recommends accept

Written by the journal's review engine in the role of the application reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

The single point I left open, the version of R2sample behind the univariate cross-check, is now settled. Section 6 names version 5.1.0 and notes that earlier versions give different numbers, the version is pinned in renv.lock, and replication_material.Rmd prints the version at the start of the check and halts if it is not 5.1.0, so the knitted output records which version produced the reported figures. Since 5.1.0 is now on CRAN a reader can reproduce the check against the public package. That closes the reproducibility concern cleanly.

Nothing else I raised in earlier rounds remains open, and I see no new material issue in this version. I recommend acceptance.

Numbered suggestions to the authors:

(none)

Report Reviewer 2, recommends accept

Written by the journal's review engine in the role of the methodology reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

The single recommended point is settled. Section 6 now names R2sample 5.1.0 at the stability check, notes that earlier versions give different results, pins the version in renv.lock with a line in Section 6.1, and the replication script prints the version and halts if it is not 5.1.0, so the knitted output records which version produced the rank correlations and overlaps. Since 5.1.0 is now on CRAN, the version is available to any reader. This closes the only item I left open.

Nothing further is open from my review.

Numbered suggestions to the authors:

(none)

Report Reviewer 3, recommends accept

Written by the journal's review engine in the role of the software reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

The single recommended point I left open is closed. Section 6 now names R2sample 5.1.0 at the check with the note that earlier versions give different results, renv.lock pins that version with the other dependencies, and replication_material.Rmd prints the version at the start of the check and stops if it is not 5.1.0. The three affected numbers (0.85, 95%, 79%) are now tied to a recorded build, so a reader who restores the environment reproduces them rather than numbers from a different CRAN release. That stop-on-mismatch guard is a better answer than a version note alone, because the knitted output itself documents which build produced the reported figures.

Nothing else was open, and my check remains textual since I could read the lockfile and scripts only through the manuscript's description of them. The paper is in good shape.

Numbered suggestions to the authors:

(none)

APPENDIX A

The authors' full response letter to round 1

As uploaded with the revision received 2026-11-06. Its own pages follow.

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".

APPENDIX B

The authors' full response letter to round 2

As uploaded with the revision received 2026-11-09. Its own pages follow.

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