

PEER REVIEW RECORD

A Computational Framework for Multivariate Goodness-of-Fit and Two-Sample Testing: MDgof and MD2sample

Wolfgang Rolke

Vol. 1 (2026), Art. 8 · <https://doi.org/10.68365/jrepstat.2026.8>

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

2026-10-13	Minor revision
2026-10-15	Revised
2026-10-23	Accepted

Round 1: Minor revision

SIGNED DECISION

Confirmed and signed by Matthias Templ on 2026-10-13 · ORCID 0000-0002-8638-5276

DECISION LETTER

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

The paper presents MDgof and MD2sample, a pair of R packages that put a broad set of multivariate goodness-of-fit and two-sample procedures behind one common interface for testing, parametric-bootstrap calibration, Monte Carlo power estimation, simultaneous inference, and benchmarking, with a slot for user-supplied statistics. It matters to applied analysts choosing among multivariate tests and to methodologists who need established competitors on identical simulation conditions. 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, say so in your response letter 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.

Reproducibility items marked required are the journal's minimum for a paper that ships code with its results: code and a command that runs it, and by acceptance a captured computational environment (dependencies/lockfile/container) and a signed determinism attestation for the entry-point command. They document what you already have. They do not ask for new analysis.

REQUIRED (the decision rests on these):

- [software-quality] In `src/TS_cont.cpp`, the Aslan-Zech accumulator in the y-y loop guards the logarithm with the wrong index: `for(i=0;i<(ny-1);++i) for(j=i+1;j<ny;++j) if(Dz(i,j)>0) tmp(0)=tmp(0)-log(Dz(nx+i,nx+j));`. The condition tests `Dz(i,j)` (an x-x distance) while the term taken is `Dz(nx+i,nx+j)` (a y-y distance). Please confirm the guard should read `Dz(nx+i,nx+j)>0`. As written the AZ statistic can take `log(0)` when a y-y distance is exactly zero, or skip a valid term, which can bias the AZ power figures in Tables and Section 9. Because both observed

and simulated statistics use the same code, level is preserved, so the paper's validity claims stand, but the reported AZ power may shift. Alternatively state that this is the intended computation.

- [statistical-rigor] The iris and EPL applications (Section 9) report bare p-values and draw conclusions from the 0.05 threshold with no effect size or interval. "setosa does not appear to come from a multivariate normal distribution" and "a significant shift in goal-scoring patterns, which may reflect changes in playing styles, tactics, or league dynamics" are substantive claims resting on p-values alone. Either report a distance or discrepancy magnitude for these two examples, or reword both conclusions to present them as illustrations of the interface rather than findings about iris or the Premier League.
- [Reviewer 1] The iris conclusion in Section 9.1 rests on the two smallest of three p-values (qK 0.0014, qCvM 0.0000, qAD 0.1043) while Section 4 warns against exactly that. Report the adjusted p-value the package produces, or qualify the conclusion to reflect the disagreement among the three statistics.
- [Reviewer 1] The EPL example in Section 9.2 drops matches with 5 or more goals (factor levels 0:4 send them to NA) and compares a complete 2020-21 season against a partial 2025-26 season, then reads the result as a change in tactics. State the truncation and the incomplete season, and present the finding as a software illustration rather than a substantive claim about league dynamics.
- [Reviewer 2] In Section 9.2 the adjusted p value combines four tests (ChiSquare, KS, AZ, CvM) while `twosample_test` reports eight. State which tests enter the minimum and why, since the claim that the result is not a multiple-testing artifact rests on that set. Replace "significant shift" and "confirms" with the estimates and their resampling uncertainty.
- [Reviewer 3] The English Premier League example in Section 9.2 reads live CSVs from a URL and sets no seed, so the printed p-values and the adjusted value of 0.018 cannot be reproduced, and the 2025-26 file is a partial, changing season. Ship a frozen snapshot (or give the retrieval date and state the numbers depend on data vintage) and set a seed.
- [Reviewer 3] The set of methods in the printed output changes across functions without explanation: `discrete_gof_test` returns K, AD, KL, P while `discrete_gof_power` returns KS, K, CvM, AD, TV, KL, H, P, and `twosample_test_adjusted_pvalue` reports only four of the eight methods from `twosample_test`. Reconcile these or state in the text why the testing and power/adjustment routines evaluate different method sets.

RECOMMENDED:

- [consistency-check] The English Premier League example reports the same tests twice with different p values. `twosample_test(dta)` in Section 9.2 gives KS = 0.0020, AZ = 0.0042, CvM = 0.0488, while the adjusted-p-value call a few lines later lists KS = 0.0036, AZ = 0.0044, CvM = 0.0508. These are consistent with unseeded Monte Carlo runs, but a reader comparing the two blocks sees conflicting numbers for the same statistics on the same data. Pass a fixed seed to both EPL calls so the individual p values agree across the two outputs, or add one sentence stating that the two calls use independent simulations and so differ by Monte Carlo error.
- [consistency-check] Table 2 marks Chi-square tests as available for continuous data in MD2sample, but the continuous `twosample_test` output in Section 5.2 (KS, K, CvM, AD, NN1, NN5, AZ, BF, BG, MMD, FR, NN0, CF1-CF4, Ball) shows no ChiSquare entry, whereas the discrete output in the same section does include ChiSquare. Reconcile the table with the demonstrated default output, or note that the continuous chi-square test is not run by default.
- [consistency-check] The adjusted-p-value algorithm in Section 4 states "Repeat B(= 1000) times", while the documented default for the testing routines is B = 5000 (Section 5). This is a presentational mismatch in the stated simulation size rather than a result discrepancy. State the default B used by the adjusted-p-value routines, or mark 1000 as illustrative.
- [methods-soundness] The text warns that permutation calibration treating a fitted distribution as fixed "can be inappropriate" when parameters are estimated, and provides parametric-bootstrap calibration for `gof_test`. The hybrid route (generate a Monte Carlo sample from the null and apply a two-sample test) is only illustrated with a simple null. Add one or two sentences stating how `hybrid_test` and `hybrid_power` handle an estimated-parameter null, since generating the reference sample from parameters fitted to the observed data induces dependence that a standard permutation two-sample calibration would not reflect.

- [prose-and-style] The manuscript leans on the “not X but Y” frame several times: “The principal contribution of this work is not any single test statistic, but infrastructure”, “The purpose of including many methods is not to claim that one family dominates the others, but to make procedures available”, “intended not simply as collections of tests but as infrastructure”. State each point directly in its own sentence, for example “The contribution of this work is infrastructure in which existing and newly proposed statistics can be implemented and compared.”
- [prose-and-style] “the packages provide common infrastructure for applying heterogeneous procedures and, importantly, for comparing them under identical simulation conditions” uses “importantly” as emphasis. Drop the word and let the clause stand.
- [simulation-design] Monte Carlo error is never reported or acknowledged. Every displayed example uses $B=250$, and the resulting type I error estimates under the null range from 0 to about 0.13 against a nominal 0.05 (for instance the goodness-of-fit power table first row, and the hybrid power table), which shows the reported figures carry substantial sampling noise. State once that the in-text numbers use a deliberately small $B=250$ for speed and are illustrative, and that the default B (5000 for tests, 1000 for power) should be used for reported comparisons; a single sentence giving the approximate standard error of a rejection-rate estimate at a given B would let readers judge which differences in the rank tables are real.
- [simulation-design] The “average rank” summary from `run.studies` is a reasonable headline metric, but with $B=250$ and no spread reported the ordering of closely ranked methods (for example qK at 6.3 versus qAD at 6.8 versus qCvM at 7.0) is not distinguishable from Monte Carlo noise. If any comparative ranking is to be read as a finding, report it at the default B with an indication of variability.
- [software-landscape] Several built-in statistics are the package’s own reimplementations in Rcpp (maximum mean discrepancy in `mmdagg.cpp`, the Aslan-Zech and Baringhaus-Franz energy statistics and the Biswas-Ghosh statistic in `TS_cont.cpp`, Friedman-Rafsky in `FR.test.R`), while others wrap a reference package directly (`Ball::bd.test`, `gTests::g.tests`, `FNN`, `spatstat.explore::Kest`), where agreement is automatic. For the reimplemented ones, a short table giving the statistic value computed by `MDgof/MD2sample` and by the corresponding reference package (energy or cramer for the energy and Baringhaus-Franz statistics, a kernel package for MMD) on one shared data set would show the built-in statistics match the established definitions and would strengthen the benchmarking claim.
- [software-landscape] The MMD, energy, Friedman-Rafsky and Fasano-Franceschini reimplementations are not accompanied by a citation to any existing software implementation of those methods. Adding one citation each for the packages a reader would otherwise reach for (for example cramer for Baringhaus-Franz, a kernel-MMD package, `fasano.franceschini.test`) would make the relation to existing tools explicit.
- [software-quality] In `R/power_pvals.R` the NaN handling `if(any(is.nan(TS_sim))) {j=j-1;next}` cannot repeat an iteration, since assigning to the loop variable of an R for loop has no effect on iteration. A rejected simulation is therefore dropped rather than re-drawn, so the denominator B no longer matches the number of accumulated statistics for affected rows. Consider a while loop or an explicit resample.
- [software-quality] In `R/twosample_power.R` the message block `if(max(totaltime)<20 | B<=2*ncores) if(!SuppressMessages) message(...) else if(...) message(...)` has a dangling else that binds to the inner `if(!SuppressMessages)`, so the “Using N cores” message prints under the wrong condition. The actual processor counts are set separately below, so only the console message is misleading, but the branch reads as a logic slip.
- [statistical-rigor] Every p-value and power figure in the paper is a Monte Carlo estimate, and none carries an uncertainty measure. The worked examples use $B=250$, so a reported power of 0.404 has a standard error near 0.03 and a p-value near 0.05 has a standard error near 0.014. The average-rank summary in Section 7 inherits this noise and could reorder methods on a re-run. Add the Monte Carlo standard error ($\sqrt{p(1-p)/B}$) to the testing and power output, or state it once in the text, so readers can see which differences the simulation size can resolve.
- [statistical-rigor] The EPL comparison sets the 2020-21 season (played without crowds under pandemic conditions) against 2025-26 and attributes any difference to “playing styles, tactics, or league

dynamics”. The season difference is confounded with the crowd-absence effect. Note this confound where the example is introduced, since the interpretation does not follow from the test alone.

- [writing-quality] The `dnull` argument is used in the Section 5.1 examples (`gof_test(x, pnullC, rnullC, dnull = dnullC, ...)`) and drives the extra BB statistic, but it is absent from the bulleted argument list for `MDgof::gof_test()`. A reader meets `dnull` in the example with no prior definition. Add a line describing it (the null density function) to that argument list, as is already done for `gof_power()`.
- [writing-quality] The abbreviations in the printed output (for example qK, qCvM, qAD, FF, RK, EP, BB, BR, KSD, ES, EP33, NN0/NN1/NN5, CF1–CF4) are not mapped to the method names in Tables 1 and 2. A reader cannot reliably connect, say, ES/EP or RK to a row in the table. A short column of output labels in each table, or one legend, would let readers interpret every result block.
- [visualisation] The density and correlation panels of a default `ggpairs` matrix can be hard to read at print size. Please confirm the axis tick labels and the panel strip labels remain legible at the final column width, and enlarge the base font if they do not.
- [Reviewer 2] Add a size study showing the built-in tests hold the nominal level across representative scenarios. The null rows already shown range from about 0.03 to 0.12 (the discrete chi-square P test reaches 0.124), which at $B=250$ is several Monte Carlo standard errors from 0.05 and warrants a direct check.
- [Reviewer 2] Pin and archive the exact Football-Data.co.uk files used in Section 9.2. The live URLs update, so the reported counts and p values cannot be reproduced from the article as written.
- [Reviewer 2] State that the illustrative runs use $B=250$ and give the resulting Monte Carlo error, so readers do not read the tabulated power values as precise.
- [Reviewer 3] State whether seeded runs reproduce under parallel execution (`maxProcessor` greater than 1), since the packages emphasise parallel processing yet every seeded example uses `maxProcessor = 1`.
- [Reviewer 3] Report the software environment (R version and versions of `mvtnorm`, `copula`, `Ball`, `energy`, `GGally`, `Rcpp`) alongside the CPU and OS already given in Section 8, so the reproduction is pinned.

OPTIONAL:

- [methods-soundness] The continuous multivariate EDF statistics are evaluated only at the observed points along one orientation. The paper already flags this as a computational approximation. A short note on which alternatives this approximation can miss (orientation-dependence of multivariate KS-type statistics) would help users choose among the built-in statistics, and could be deferred to the vignette.
- [Reviewer 1] A compact summary across the benchmark collection, showing which statistic families lead under which alternative geometries, would give applied readers a basis for a default choice.
- [Reviewer 2] Note that the adjusted p value nests simulation inside simulation (each resample reruns every test’s own calibration) and say what that costs.

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

- [writing-quality] In Section 4 the null distribution of the minimum p value is named $F_{\min}$ in `padj = Fmin(pmin, obs)` and then called $\hat{F}_p$ in the algorithm bullets (“empirical distribution function $\hat{F}_p$ ”, “Apply $\hat{F}_p$ ”). The capitalisation of the minimum p value also alternates between $P_{\min}$ and $p_{\min}$. Using one symbol throughout removes the apparent discontinuity.

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](#).

The problem this software addresses is real. Multivariate goodness-of-fit and two-sample testing carry a long list of procedures whose power depends on the alternative, the dimension, the sample size, and the treatment of nuisance parameters, and there is no uniformly best test. An analyst who wants to run several procedures, or a methodologist who needs established competitors on matched conditions, currently

has to stitch together packages with different data conventions and resampling schemes. Putting these procedures behind one interface, with shared calibration, power, and benchmarking machinery and a slot for user statistics, is useful infrastructure. The framing is honest that the contribution is the environment and not any single statistic. The method coverage is broad and the parametric-bootstrap calibration that re-estimates parameters is the right choice for composite nulls. From the applied side the design is sound and the packages are on CRAN.

My main concern is in the two real-data demonstrations, which are the paper's evidence that the framework helps on actual data, and which both state conclusions firmer than the data support. The iris example reports three p-values for multivariate normality of *setosa* (qK 0.0014, qCvM 0.0000, qAD 0.1043) and concludes the data do not come from a normal distribution. That reads the two small values and sets aside the third, which is the move Section 4 is written to prevent. The package already computes an adjusted p-value for this situation, so the fix is to report it, or to qualify the sentence to acknowledge that one of the three statistics is compatible with normality. As written the application contradicts the paper's own message.

The EPL example has two treatment issues. The scoring table uses factor levels 0:4, so any match with five or more goals falls to NA and leaves the counts, which removes real matches from both samples without comment. And the 2020-21 season is complete while the 2025-26 season is in progress at the time of writing, so the two samples differ in completeness as well as in era. On that basis the text reads the test as a shift in goal-scoring that may reflect playing style, tactics, or league dynamics. That interpretation is not supported by a test that compares a full season to a partial one on truncated counts. Please state the truncation and the incomplete season, and present the result as an illustration of the discrete two-sample workflow. The paper also uses "significant" and reads p-values against a threshold in this section. Describing the size of the departure and what it is compatible with would fit the rest of the paper better.

Three smaller points would help a reader use the software. The Football-Data source is live, so the Section 9.2 numbers will not reproduce from the URLs given, and a frozen snapshot in the artifacts would fix that. The output columns use abbreviations (qK, qKS, BR, FF, RK, ES, EP, BB, and the two-sample labels) that a reader cannot map to Tables 1 and 2 without guessing, and a short key would remove the guessing. Finally, the set of methods returned is not stable between the test and power routines: the discrete `gof_test` returned K, AD, KL, P, while the discrete `gof_power` for the same setup returned KS, K, CvM, AD, TV, KL, H, P, and the continuous case shows a similar gap. A sentence explaining why, or a correction if it is unintended, would prevent confusion.

Numbered suggestions to the authors:

1. **REQUIRED** In Section 9.1, report the adjusted p-value for the iris normality test or qualify the conclusion, since qAD (0.1043) disagrees with qK and qCvM and the current wording selects the two small values against Section 4's own caution.
2. **REQUIRED** In Section 9.2, state that matches with five or more goals are excluded by the factor levels 0:4 and that the 2025-26 season is incomplete, and present the result as an illustration of the discrete two-sample workflow rather than a claim about tactics or league dynamics. Replace the threshold language with the size of the departure.
3. **RECOMMENDED** Include a frozen CSV snapshot of the Football-Data.co.uk files in the artifacts so the Section 9.2 outputs reproduce.
4. **RECOMMENDED** Add a mapping from the output abbreviations to the method names in Tables 1 and 2.
5. **RECOMMENDED** Explain, or correct, why the returned method set differs between the test and power routines (for example the discrete `gof_test` returned K, AD, KL, P while the discrete `gof_power` returned KS, K, CvM, AD, TV, KL, H, P).
6. **OPTIONAL** Add a compact summary across the benchmark collection showing which statistic families lead under which alternative geometries, to guide a default choice.

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 software paper, and judged as one the statistical design is sound. The framework separates the observed data, the test statistic, and the null-generation mechanism, and that separation is what lets built-

in and user-defined statistics share one calibration, power, and benchmarking path. Simulation-based calibration is the right choice for a collection this heterogeneous, because most of these multivariate statistics have no usable finite-sample null distribution, and a resampling calibration is valid for each of them by construction. The parametric-bootstrap handling of estimated parameters, with re-estimation inside the null loop, is the correct treatment and is often skipped in practice. The hybrid route, comparing the observed sample to a Monte Carlo reference through a two-sample statistic, is a reasonable device for models that are easy to simulate but hard to evaluate. On the whole the methods are described accurately and matched to the problem.

My main methodological reservation concerns what a benchmarking and calibration framework ought to demonstrate about itself. The central promise is that these tests can be compared on equal terms, and equal terms start with every test holding its stated level. The paper shows null rows inside the power tables rather than a dedicated size study, and some of those entries sit well away from 0.05. The discrete chi-square P statistic reaches 0.124 under the null at $B=250$, which is far enough from nominal to be worth isolating rather than leaving inside a power example. A short size study run through the existing `run.studies` machinery would settle this and would strengthen the paper's own argument about comparability. I have marked it recommended because the simulation calibration is valid by design, so this is confirmation rather than a repair of a claim.

The Section 9.2 application needs attention on two points that bear on a reader's conclusions. The adjusted p value is reported over four tests while the test table lists eight, and the text does not say which tests feed the minimum or why. Because the adjusted value depends entirely on that set, and because the sentence draws a substantive conclusion about Premier League scoring from it, the reader needs the set stated. The surrounding wording ("significant shift", "confirms that this result is not an artifact") reports threshold crossings where it could report the estimates and their resampling spread. The same section relies on live data files that change over time, so the printed numbers are not reproducible as they stand. Pinning the downloaded files would fix that at low cost.

Smaller clarity items would help a reader move between the tables and the printed output. The output uses abbreviations that are not all mapped to the method names, and the set of methods differs between `gof_test` and `gof_power` without explanation. Stating the B used in the examples and the attendant Monte Carlo error would keep readers from over-reading the tabulated power. None of these change the framework, and the framework is the contribution.

Numbered suggestions to the authors:

1. **REQUIRED** In Section 9.2, state which tests enter the adjusted p value (the output shows four, the test table shows eight) and why that subset, and replace "significant shift" and "confirms" with the estimates and their resampling uncertainty.
2. **RECOMMENDED** Add a size study across representative scenarios showing the built-in tests attain the nominal level, and comment on the discrete chi-square P entry of 0.124.
3. **RECOMMENDED** Add a mapping from the output abbreviations to the method names in Tables 1 and 2, and explain why `gof_test` and `gof_power` report different method sets.
4. **RECOMMENDED** Archive the exact Football-Data.co.uk files used in Section 9.2 so the reported counts and p values can be reproduced.
5. **RECOMMENDED** State that the illustrative runs use $B=250$ and give the resulting Monte Carlo error.
6. **OPTIONAL** Note the nested-simulation cost of the adjusted p value, where each resample reruns every test's own calibration.

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](#).

This is a software paper describing two companion R packages that place a large collection of multivariate goodness-of-fit and two-sample procedures behind a common interface, and route both built-in and user-supplied statistics through the same resampling, power, and benchmarking code. From a software-and-reproducibility standpoint the design is coherent. Separating the data, the statistic, and the null-generation mechanism is the right decomposition, it is what lets one Monte Carlo engine serve every statistic, and the parallel naming across the two packages is a sensible interface choice. The package

versions are pinned (MDgof 1.2.0, MD2sample 1.4.0), CRAN distribution is stated, and most examples pass a seed. The contribution as infrastructure is real and the paper presents it clearly.

My concerns are about whether the results shown can be reproduced and whether the printed output can be interpreted. The clearest reproducibility gap is the English Premier League example in Section 9.2. It reads two CSV files over the network, sets no seed, and the 2025-26 file is an in-progress season. The printed statistics, the per-test p-values, and the combined adjusted value of 0.018 are therefore tied to whatever the website served on the day the author ran it, and a reader following the same code today will get different numbers. For a software paper the fix is cheap: include a frozen copy of the two score tables used, or state the retrieval date and note that the numbers depend on the data vintage, and set a seed on the calls. The iris example in Section 9.1 likewise omits a seed, so its p-values are reproducible only up to Monte Carlo error.

The second issue is interpretive consistency across the interface. The paper's central claim is a unified workflow, so it matters that the method set in the printed output changes from one function to the next. Discrete `gof_test` returns K, AD, KL, P, while discrete `gof_power` returns KS, K, CvM, AD, TV, KL, H, P. The continuous `gof_test` returns six or seven statistics, while `gof_power` returns nine. The adjusted-p-value routine reports four of the eight methods that `twosample_test` ran on the same data, and the text does not say which four enter the minimum or why. One or two sentences explaining the rule (for example, that power and adjustment restrict to methods with a particular calibration path) would resolve this and support the unification claim rather than appearing to undercut it.

Interpretation is also limited by the absence of a key for the output labels. The columns qKS, qK, qCvM, qAD, BB, BR, RK, FF, ES, EP, AZ, BF, BG, the NN variants, and CF1 through CF4 are never mapped to the method names in Tables 1 and 2, so a reader cannot tell which row of the table produced which number. A single mapping table or glossary would make every example readable. Two smaller points round this out. The packages advertise parallel processing, yet every seeded example runs at `maxProcessor = 1`, so the manuscript does not establish whether a seeded run reproduces under parallel execution, which is the determinism question a user will hit first. And Section 8 reports the CPU and OS but not the R version or the versions of the dependencies (`mvtnorm`, `copula`, `Ball`, `energy`, `GGally`, `Rcpp`), which is what pins a reproduction.

Numbered suggestions to the authors:

1. **REQUIRED** Make the Section 9.2 example reproducible: include a frozen snapshot of the two score tables (or give the retrieval date and state that the numbers depend on data vintage), and set a seed on `twosample_test` and `twosample_test_adjusted_pvalue` so the printed p-values and the 0.018 adjusted value can be regenerated.
2. **REQUIRED** Reconcile or document the differing method sets across functions, including the discrete `gof_test` versus `gof_power` case (K, AD, KL, P versus KS, K, CvM, AD, TV, KL, H, P) and the four-of-eight methods reported by `twosample_test_adjusted_pvalue`, stating which methods enter each routine and why.
3. **RECOMMENDED** Add a glossary or table mapping every output abbreviation (qKS, qK, qCvM, qAD, BB, BR, RK, FF, ES, EP, AZ, BF, BG, NN0/NN1/NN5, CF1–CF4, Ball) to the method names in Tables 1 and 2.
4. **RECOMMENDED** State whether a seed reproduces results under parallel execution (`maxProcessor` greater than 1), given that all seeded examples use a single core.
5. **RECOMMENDED** Add the R version and the versions of `mvtnorm`, `copula`, `Ball`, `energy`, `GGally`, and `Rcpp` to the environment reported in Section 8.
6. **RECOMMENDED** In Section 9.2, note that factor levels 0:4 drop matches with five or more goals, or widen `max_goals` so no matches are excluded.

REVISION, RECEIVED 2026-10-15

This revision was answered under the earlier rule: a short note on what changed, and a reason for each required point the authors declined.

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

Round 2: Accepted

SIGNED DECISION

Signed by Matthias Templ on 2026-10-23 · ORCID 0000-0002-8638-5276

Final acceptance of version 2, after the editor read the preview of the article. The journal's re-run reproduces the reported values, with one exception: the power estimates of tests that compute their own p-values vary between runs despite the seed, through a fault in the package's random-number handling (calcTS). One AZ entry, printed as 0.236, reproduces as 0.220 and was corrected in the proof.

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](#).

Of the two required points and the lighter ones I raised, the required points are both met and all but one of the rest are settled.

The iris example in Section 9.1 now reports the adjusted p-value alongside the individual statistics, frames the disagreement among qAD (0.1077), BR (0.7747) and the two small values as the motivation for the adjustment, and states that a substantive assessment of normality is not the point. The note that the adjusted value of zero is a Monte Carlo estimate at the resolution used is a good addition. This removes the contradiction with Section 4 that I had flagged.

The EPL example in Section 9.2 now states that matches with five or more goals are excluded, that the 2025-26 season was incomplete when collected, and that no substantive conclusion about scoring patterns is drawn. The frozen 25-row frequency table is in the manuscript and the replication material, so the output reproduces without the live source. The threshold language is gone from that section. Both required points are therefore resolved. The abbreviation legend in Section 7 and the computational environment detail in Section 8 also address the recommended items I raised.

One recommended point is still open. The goodness-of-fit test and power routines return different method sets for the same problem and data: in Section 5.1 the discrete `gof_test` prints K, AD, KL, P, while in Section 6.1 the discrete `gof_power` for the matching setup prints KS, K, CvM, AD, TV, KL, H, P, and the continuous examples show the same pattern. The response addressed why defaults differ between the test and the adjusted-p-value routines, and Section 4 now documents that, but it does not explain the test-versus-power gap, which a reader comparing the two examples will still notice. This does not hold the paper. One sentence saying the power routine reports the full available set by default, or aligning the two defaults, would close it. I accept the reasoned decline on a cross-benchmark summary of which families lead under which alternatives: the dependence on null, alternative, dimension and sample size makes a compact default recommendation hard to state without misleading, and the discussion already says so.

Numbered suggestions to the authors:

1. RECOMMENDED Add one sentence explaining why the goodness-of-fit test and power routines return different method sets for the same problem and data (for example that the power routine reports the full available set by default), or align the default selections so the two examples match.

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 one required point and five lighter ones I raised, all are met except the size study, which the author declined for a stated reason. The required Section 9.2 point is resolved: the adjusted p value now runs over the stated eight discrete methods via `doMethods="all"`, Section 4 lists the default four-method subset and explains why it differs, and the threshold-crossing wording is replaced by the reported p values with the example reframed as a software demonstration.

On the declined size study, the reasoning convinces me. The simulation calibration is valid by construction for every procedure that draws its p value from the null simulation, so the null rows are confirmation rather than a claim the paper rests on. The one genuine outlier, the discrete chi-square P entry at 0.124, is the procedure that relies on a classical distributional approximation, and the revision now names it as the exception and adds the Monte Carlo-variation caveat. Since the item was recommended, not required, the decline does not hold the paper.

The remaining recommended items are in place: the output-abbreviation legend, the pinned goal frequency table in the replication material and in the manuscript body, the $B=250$ statement with its Monte Carlo standard error in Section 6, and the serial two-stage clarification of the adjusted p value that corrects my own nested-simulation framing. Nothing is open.

Numbered suggestions to the authors:

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 two required points and four lighter ones I raised, all are met. The Premier League example now embeds the 25-row frequency table directly in the text (`goals_premier_league`), sets `seed = 1234`, states the retrieval context, and notes the exclusion of matches with five or more goals, so the printed statistics and the 0.013 adjusted value regenerate from the manuscript alone. The iris example now carries a seed as well. Section 4 explains that method availability depends on data type and dimension, documents the default `doMethods` subsets against `doMethods = "all"`, and the iris example uses "all" so the test and adjusted routines evaluate the same five methods. The abbreviation legend at the end of Section 7, the parallel-reproducibility note on the seed argument, and the expanded Section 8 environment (R 4.6.0, the Windows build, and the six dependency versions) all land cleanly.

One residual observation, carrying no ask: the discrete `gof_power` output still lists eight statistics (KS, K, CvM, AD, TV, KL, H, P) while the discrete `gof_test` example above it reports four (K, AD, KL, P). The new Section 4 text covers the test-versus-adjusted default difference and the dimension dependence, which is what I asked for, so this does not hold the paper. If a future revision touches Section 5.1, a half-sentence on why the power default is wider than the test default for discrete data would close the last interpretive gap. The version as it stands is reproducible from its own text and the replication material, and the interface is readable.

Numbered suggestions to the authors: (none)

APPENDIX A

The authors' response letter to round 1

As uploaded with the revision received 2026-10-15. Its own pages follow.

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.