

PEER REVIEW RECORD

Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA)

Aliaksandr Hubin and Geir Storvik

Vol. 1 (2026), Art. 9 · <https://doi.org/10.68365/jrepstat.2026.9>

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

2026-10-14	Minor revision
2026-10-22	33 answers
2026-10-23	Accepted

Round 1: Minor revision

SIGNED DECISION

Confirmed and signed by Matthias Templ on 2026-10-14 · 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 assembles a set of latent Gaussian examples with known or reference marginal likelihoods and checks that INLA's marginal-likelihood estimates match analytical values, Chib and Chib-Jeliazkov estimates, and the methods benchmarked by Friel and Wyse. It matters to applied Bayesian modellers who use INLA's reported marginal likelihood for model selection or model averaging and want evidence that the default output is trustworthy. 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.

REQUIRED (the decision rests on these):

- [Reviewer 1] Qualify the generality claim, or add a case that probes INLA's limits. Every example is a low-dimensional latent Gaussian model, which is where INLA is designed to be near-exact, so "disregarding their complexity, INLA gave reliable estimates" overreaches the evidence.
- [Reviewer 1] Address the spread in Table 4 before concluding the methods are "comparable": nested sampling and INLA differ by about 0.5 log units in places, which matters for a Bayes factor.
- [Reviewer 1] Clarify Table 1: define the "D" column, and state the size of the harmonic-mean error against the exact value rather than leaving the reader to infer it.
- [Reviewer 1] Define the quantity p_t that appears in equations (9) and (10) but is never introduced.
- [Reviewer 2] Report the INLA-vs-reference discrepancies on a decision-relevant scale, as differences in log evidence / Bayes factors, not as "agreement to the second decimal".
- [Reviewer 2] State, per example, which INLA approximation and which hyperparameter integration were used, since "default settings" is version-dependent.

- [Reviewer 2] The Section 4 title and text call a probit model “logistic regression with a probit link”, which is contradictory. Name it probit regression.
- [Reviewer 3] State the R-INLA version and R version used, and add a captured environment (session-Info, an renv lockfile, or a container) to the repository. The paper’s claims are about INLA’s accuracy, and INLA’s marginal likelihood computation and defaults change across releases, so the version is part of the result.
- [Reviewer 3] Set and record the RNG seeds for the stochastic estimators (Chib’s method, Chib and Jeliaskov’s method, harmonic mean, and the resampled runs behind Figures 2 and 4) so the replicate columns in Tables 1, 2 and 3 can be regenerated.

RECOMMENDED:

- [consistency-check] Table 1 carries an undefined column headed “D” holding the value 2 in all three rows, with no definition in the caption or Section 2 text. Define this column in the caption or remove it if it is spurious.
- [consistency-check] Figure 3 caption and Section 4 text state the precision was varied between 0 and 10, while Table 3 reports results at precision 10^2 ($\sigma\beta=0.1$) for the same models. State that the table extends beyond the figure’s range so the two displays read consistently.
- [methods-soundness] An exact benchmark exists only in the first example (Section 2). In every other example accuracy is judged by agreement between INLA and Chib’s or Chib and Jeliaskov’s estimates, which is a cross-method consistency check, not validation against truth. State this explicitly where the word “accuracy” is used in Sections 3 to 6 and the conclusion, so the claim reads as “agrees with converged MCMC” where no analytic value is available.
- [methods-soundness] The comparison tables report point estimates and a handful of replications, with no Monte Carlo standard error for the MCMC-based estimators (Chib, Chib and Jeliaskov, power posteriors, AIS, nested sampling). Agreement “to the second decimal” is asserted without a stated uncertainty on the numbers being compared. Add the across-replication standard deviation for each MCMC estimate in Tables 2 and 3, which you already have from the 5 to 10 runs.
- [methods-soundness] The supplementary data and R scripts are referenced by GitHub links and were not provided for review. No ask is raised for their content. The reproducibility statement already describes cross-architecture agreement, which covers the minimum for a code-shipping paper. Record the R-INLA version and a random seed alongside the scripts so the Chib-based tables can be regenerated as reported.
- [novelty-and-prior-art] The central claim is that no prior study compares INLA marginal likelihood estimates against other methods, yet the search surfaces uncited work that builds directly on INLA marginal likelihoods for model averaging and normalizing-constant estimation. Gómez-Rubio, Bivand & Rue (2020) use INLA marginal likelihoods for Bayesian model averaging, Berild et al. (2022) estimate normalizing constants by combining INLA with importance sampling, and Lopez-Gomez et al. (2026) use INLA marginal likelihoods for model and variable selection. Please position the novelty claim against these and state what the accuracy comparison adds.
- [simulation-design] The factor grid is narrow and sits where INLA is expected to perform well. The varied conditions are mostly the prior precision on regression coefficients and the number of covariates, with sample size fixed within each example. Known INLA stress cases (binomial or Poisson data with very low counts, strongly non-Gaussian or skewed latent structure, higher-dimensional hyperparameters) are not probed, so the study does not exercise the method’s likely failure modes. Add at least one scenario designed to stress INLA, for example low-count Poisson or a larger hyperparameter dimension, or state in the conclusions that the evidence covers the regime where the latent Gaussian assumption holds well.
- [simulation-design] Monte Carlo error is shown implicitly through replicate columns (5 runs of Chib, 10 runs across iteration counts) but never quantified. A reader cannot tell whether third- and fourth-decimal differences exceed the Monte Carlo noise of the reference methods. Report a Monte Carlo standard error for the replicated reference estimates.
- [software-landscape] The manuscript text names the methods (Chib 1995, Chib and Jeliaskov 2001) but does not name the software that produces the competing numbers. The code uses MCMCpack for Chib’s estimates. Please state in Sections 3, 4 and 6 which package and function (MCMCpack,

MCMCregress/MCMCprobit, marginal.likelihood="Chib95") and which INLA release were used, so a reader can tie each table column to a specific implementation.

- [software-landscape] The Table 4 entries for power posteriors, annealed importance sampling and nested sampling are reproduced from Friel and Wyse (2012) rather than from a re-run. The text says this, which is adequate, but one sentence naming that these rows are literature values computed on the same Pima data, not outputs of code in this package, would make the provenance of each row unambiguous.
- [statistical-rigor] The accuracy claims rest on qualitative agreement ("essentially equal", "agree up to the second decimal", "remarkable agreement") with no summary of the discrepancy between methods and no summary of the spread across replications. Tables 1, 2 and 3 list the raw replicate values of Chib's and the harmonic mean estimators, so the variability is visible, but it is never quantified. Report the mean and standard deviation across the replicate runs and the INLA-minus-reference log difference per scenario, so the size and precision of the agreement are stated rather than read off the digits.
- [statistical-rigor] The INLA estimates depend on a Gaussian approximation to the latent field and on the chosen integration over the hyperparameters (EB, CCD or grid). Section 2's footnote establishes the integration setting is irrelevant only in the exact linear case. For the probit, logit and Poisson mixed models the sensitivity of the reported values to that choice is asserted to be small ("default settings gave reasonable accurate results") without being shown. Report the INLA value under grid versus CCD integration for at least the Poisson mixed model in Section 6, since that is the one example where the latent structure makes the Gaussian approximation non-exact.
- [statistical-rigor] The conclusion recommends adjusting tuning parameters "if extremely high accuracy is needed" and experimenting with settings to judge "whether more accuracy is needed", but the paper gives no diagnostic or criterion for making that judgement. Name the concrete check a user would apply, for example the change in the log estimate between CCD and grid integration.
- [visualisation] The scatterplots in Figures 1 and 3 carry the central claim that INLA and Chib's method "give equal marginal likelihoods". A point near the diagonal supports this, but no $y=x$ reference line is drawn, so a reader cannot see departures from equality by eye. Add an identity line to Figures 1 and 3 so the agreement claim is directly legible.
- [visualisation] Figures 2 and 4 plot the Chib estimate against iteration count on a y-axis spanning roughly 0.04 log units (-19.50 to -19.46 in Figure 2). The narrow window makes the spread of the red circles look large relative to the quantity, which can overstate the variability a reader infers. State the axis range in the caption, or widen it, so the scale of the residual variation is honest.
- [visualisation] Axis labels on the scatterplots name the methods but not the units; "estimated marginal log-likelihood" would make clear the values are on the log scale. Label both axes with the quantity and its scale.
- [writing-quality] In models (9) and (10) the response is written as conditioned on p_t (for example " $Y_t | p_t, M \text{ Bernoulli}(\Phi(\eta_t))$ "), yet p_t is never defined and the success probability is given directly as $\Phi(\eta_t)$ or $\text{logit}^{-1}(\eta_t)$. The same pattern appears in (11), where the conditioning uses λ_{jt} but the rate is written $\exp(\eta_{jt})$. Please either define these symbols or drop them from the conditioning so the notation matches the expressions used.
- [ethics-and-integrity] Funding and institutional support (CELS, Integreat) are acknowledged, and no competing interests are apparent from the text. Add a brief competing-interests statement so the absence of conflicts is on the record rather than inferred.
- [Reviewer 1] Quantify the INLA-versus-reference discrepancy across the prior range shown in Figures 1 and 3, instead of describing agreement qualitatively.
- [Reviewer 2] Give replication and a spread for the single-point GLMM comparison in Section 6.
- [Reviewer 3] Add a README or driver script that maps each script to the table or figure it produces, and state which of the two linked repositories is canonical for the published numbers.
- [Reviewer 3] Support the cross-architecture reproduction claim in the supplementary paragraph with a logged artifact (a sessionInfo from each platform) rather than an assertion.

OPTIONAL:

- [software-landscape] Bridge sampling (R package `bridgesampling`) is a widely used evidence estimator not represented in the comparison. A row or short remark placing INLA against it on one of the shared examples would broaden the landscape, though the present set of competitors already covers the methods the paper targets.
- [visualisation] The convergence figures distinguish the Chib runs (red circles) from the INLA estimate (blue line) by colour alone. Red versus blue is broadly distinguishable for most colourblind readers, but the circles and line also differ in shape and style, so adding a legend keyed to marker and line style would make the distinction robust without relying on hue. Add an in-panel legend.

REVIEWS

Report Reviewer 1, recommends major

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 focused simulation note that asks a question applied users of INLA genuinely care about: can the marginal likelihood that INLA reports by default be trusted for model selection and model averaging? The authors answer it by lining INLA up against an analytical benchmark, against Chib's method in Gaussian and probit regression, against the full panel of methods in Friel and Wyse for a logit example, and against Chib-Jeliazkov for a Poisson GLMM. The gap the paper claims, that no prior study compared INLA's marginal likelihood against these alternatives, is plausible and worth closing, and the examples are well chosen and recognisable to the field. The reproducibility statement and the two code repositories are a real strength.

My central concern is one of scope and the conclusion drawn from it. Every example here is a latent Gaussian model with a low-dimensional hyperparameter vector, which is the regime INLA was built for and where its accuracy is least in doubt. Example 1 is Gaussian-Gaussian, where INLA is exact by construction, as the footnote states. The sentence "In all of the addressed examples, disregarding their complexity, INLA gave reliable estimates" reads as a general endorsement, but the examples do not vary in the dimension that would actually stress INLA's marginal-likelihood accuracy: a larger or more poorly identified hyperparameter space, a skewed or multimodal latent conditional, or a model where the Gaussian approximation to the full conditional is known to be weak. Either add one such case, or qualify the conclusion so it speaks to the regime you tested. The qualification is the cheaper route and I would accept it.

The second concern is the use of a reference that is itself approximate. In examples 2, 3 and 6 there is no analytical truth, so Chib or Chib-Jeliazkov plays the role of ground truth. When two approximations agree this is reassuring, but it does not tell the reader which is right when they disagree, and Table 4 shows they do disagree: nested sampling sits roughly half a log unit from INLA in places, and the methods span a visible range. Concluding that the methods are "comparable" passes over this. For a reader using these numbers to form a Bayes factor, half a log unit is not negligible. Please discuss the spread and what it implies, and where a method serves as the reference, report its Monte Carlo standard error so the reader can weigh the two estimates rather than assume the MCMC method is the arbiter.

Smaller points of clarity. Table 1 is hard to read: the "D" column is unlabelled in meaning, and the harmonic-mean error, which is the point of the table, is left for the reader to compute from the exact value. State it. The symbol p_t appears in the Bernoulli models of equations (9) and (10) without being defined, presumably the inverse-link of η_t . On the positive side, Figures 2 and 4 showing Chib's convergence against the INLA line are the most informative displays in the paper, because they make the accuracy claim visible rather than asserted; extending that spirit to a numerical error summary over the prior range in Figures 1 and 3 would strengthen the empirical case considerably.

Numbered suggestions to the authors:

1. **REQUIRED** Qualify the main conclusion to the regime actually tested, low-dimensional latent Gaussian models, or add one example that stresses INLA (higher-dimensional or poorly identified hyperparameters, or a latent conditional far from Gaussian). The qualification alone satisfies this point.

2. **REQUIRED** Discuss the spread among methods in Table 4, in particular the roughly 0.5 log-unit gap between nested sampling and INLA, and what it means for Bayes factors, rather than stating the methods are comparable.
3. **REQUIRED** Clarify Table 1 by defining the “D” column and reporting the magnitude of the harmonic-mean error against the exact value.
4. **REQUIRED** Define p_t where it first appears in equations (9) and (10).
5. **RECOMMENDED** Summarise the INLA-versus-reference discrepancy numerically across the prior values in Figures 1 and 3, for example as a distribution of absolute log differences, instead of describing agreement in words.
6. **RECOMMENDED** Report the Monte Carlo standard error of Chib’s and Chib-Jeliazkov’s estimates wherever they serve as the reference, so a reader can judge which estimate to trust when they differ.

Report Reviewer 2, recommends major

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 compact comparison study with a clear and useful question: do the marginal likelihoods that INLA returns match what careful, slower estimators produce, across models of increasing complexity? The sequence of examples is well chosen, running from a one-parameter Gaussian model with an exact answer, through Gaussian and binary regression, to a Poisson GLMM. The headline reading, that INLA tracks the reference estimators closely while costing far less, is consistent with the numbers shown. The code and data are posted, which I value.

My main methodological reservation is that the paper never quantifies the thing it is about. The accuracy claim rests on phrases like “agree up to the second decimal” and on scatterplots where points fall on the diagonal. For a marginal likelihood study this is the wrong unit. What a user cares about is the error in log evidence, because that error propagates directly into Bayes factors and into the posterior model weights of equations (3) and (4). An error of 0.1 in a log marginal likelihood is a 10% shift in a Bayes factor, and in Table 4 the across-method spread is of that order. Please compute, from the numbers you already have, the discrepancy between INLA and the reference in each table on the log-evidence / Bayes-factor scale, and say what size of discrepancy would change a model-selection or model-averaging decision. That reframing is what turns the plots into evidence.

A second issue is the reference itself. Only Table 1 has an exact benchmark. Everywhere else the “truth” is a competing estimator, chiefly Chib or Chib-Jeliazkov, and the argument that these have converged rests on Figures 2 and 4. Those figures show stabilisation for one model each, so the convergence of the reference in the probit and Poisson cases is assumed rather than shown. The GLMM in Section 6 is the most interesting and the most lightly treated: a single INLA value against two single reference values, with no replication and no spread. Because this is the one example where INLA’s latent-Gaussian handling actually does work, it deserves the replication the earlier sections received.

Third, the examples are uniformly favourable to INLA. Every model has a latent Gaussian structure and low-dimensional hyperparameters, which is the regime INLA was built for. That is a reasonable scope, but the conclusion as written (“in all of the addressed examples, disregarding their complexity, INLA gave reliable estimates”) reads as a general endorsement that the examples do not support. I am not asking for new stress-test models as a condition of acceptance, but the generalisation should either be narrowed to the latent-Gaussian examples actually studied or accompanied by a case where the Gaussian approximation to the full conditional is known to strain. It would also help to engage the existing discussion of INLA marginal-likelihood reliability, since the evidence is a normalising constant that depends on the hyperparameter integration and is often considered less stable than the posterior marginals.

On reproducibility and specification: “default tuning parameters” is not enough to reproduce or interpret the results, because the default depends on the R-INLA release and because the choice among the Gaussian, simplified Laplace and full Laplace strategies, and among EB, CCD and grid integration over the hyperparameters, materially affects the evidence. Please state, per example, which strategy and which hyperparameter integration were used. The supplementary note on cross-architecture reproduction is a good addition and should stay. Two smaller clarity points: the “D” column in Table 1 is undefined, and the observed data for Neal’s single-observation example is not given, so the exact value cannot be checked

by a reader. Finally, Section 4 is headed “logistic Bayesian regression with a probit link” and repeats the phrase in the text, which contradicts itself. The model in equation (9) is probit regression.

Numbered suggestions to the authors:

1. **REQUIRED** Recompute the INLA-vs-reference discrepancies in every table as differences in log evidence and in Bayes factors, and state what magnitude of discrepancy would alter a model-selection or model-averaging decision. Replace “agreement to the second decimal” as the accuracy criterion.
2. **REQUIRED** For each example state the INLA approximation strategy used (Gaussian, simplified Laplace, full Laplace) and the hyperparameter integration used (EB, CCD, grid), and pin the R-INLA version, so “default settings” is reproducible.
3. **REQUIRED** Qualify the conclusion so the reliability claim covers the latent Gaussian examples studied, or add a case where the Gaussian full-conditional approximation is expected to strain. The low-effort form is to narrow the sentence that currently says “disregarding their complexity”.
4. **REQUIRED** Correct the Section 4 heading and text: the model is probit regression, not “logistic regression with a probit link”.
5. **RECOMMENDED** For the Section 6 GLMM, add replicate runs of Chib-Jeliazkov and a spread, so the single-point comparison carries the same weight as the earlier sections, and state whether the reference has converged.
6. **RECOMMENDED** Define the “D” column in Table 1 and report the observed y for Neal’s example so a reader can verify the exact value, and show the convergence evidence for the reference estimators in the probit and Poisson cases as you did in Figures 2 and 4.

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 compact comparison study: INLA marginal likelihoods set against an analytical benchmark and against a set of MCMC and other estimators, over five examples of growing complexity. From a reproducibility standpoint the useful property is that most of the numbers in the paper are the code’s numbers, generated by the authors rather than copied in (the exception is Table 4, where everything but the INLA rows comes from Friel and Wyse 2012). That makes the question for my lane straightforward: can a reader regenerate these tables, and does the paper give them what they need to do so.

The central gap is version information. The whole paper is an accuracy statement about INLA, run “with the default tuning parameters from the package”, but the package version is not given. INLA’s marginal likelihood machinery, its integration strategy and its defaults have changed across releases, and a reader on a current R-INLA will not necessarily see the numbers in Tables 1 to 4. The supplementary paragraph is aware of this, since it talks about reproduction “across R-INLA releases” and attributes third and fourth decimal drift to solver and compiler differences. That awareness is exactly why the version and environment belong in the record. Please state the R-INLA and R versions used for the published runs and add a captured environment (sessionInfo, an renv lockfile, or a container) to the repository. This is low effort with the material you already have, and it is what lets the accuracy claim be checked rather than taken on trust.

The second gap is seeds. The paper reports five or ten repeated runs of the stochastic estimators, and Figures 2 and 4 show their variability against iteration count. None of this can be regenerated without the RNG seeds. Recording the seeds for Chib’s method, Chib and Jeliazkov’s method, the harmonic mean, and the figure runs lets the replicate columns be reproduced exactly, which is the point of showing replicates.

The supplementary paragraph makes a fairly specific reproduction claim: exact for Table 1, two decimals for Table 4, within 0.0002 to 0.003 log units for Tables 2 and 3, across Apple Silicon and Intel. That is a verification-shaped statement, and right now it rests on assertion. A logged sessionInfo from each platform, committed alongside the scripts, would turn it into something a reader can inspect. Relatedly, two repositories are linked, one described as adapted for publication and one as the original 2016 scripts, with no statement of which produced the published numbers or how they differ. A short README mapping scripts to tables and figures, and naming the canonical repository, would remove that ambiguity. I could not inspect either repository in this pass, so I have scoped these asks to what the manuscript declares rather than to the code itself.

One wording point outside my main lane: Sections 4 and 5 both describe “logistic regression with a probit link”, and the probit and logit examples are genuinely different link functions, so the shared label reads oddly.

Numbered suggestions to the authors:

1. **REQUIRED** State the R-INLA version and the R version used for the published runs, and commit a captured environment (sessionInfo, an renv lockfile, or a container) to the repository, so the INLA accuracy claim can be checked against a known software state.
2. **REQUIRED** Set and record the RNG seeds for Chib’s method, Chib and Jeliazkov’s method, the harmonic mean, and the runs behind Figures 2 and 4, so the replicate columns in Tables 1 to 3 can be regenerated.
3. **RECOMMENDED** Add a README or driver script mapping each script to the table or figure it produces, and state which of the two linked repositories is canonical for the published numbers.
4. **RECOMMENDED** Support the cross-architecture reproduction claim in the supplementary paragraph with a committed sessionInfo from each platform rather than a narrative assertion.
5. **RECOMMENDED** Reconcile the “logistic regression with a probit link” wording in Sections 4 and 5 so the probit and logit examples are named by their actual link functions.

REVISION, RECEIVED 2026-10-22

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.

Detailed point-to-point responses and indicated changes are attached.

REQUIRED Qualify the generality claim, or add a case that probes INLA’s limits. Every example is a low-dimensional latent Gaussian model, which is where INLA is designed to be near-exact, so “disregarding their complexity, INLA gave reliable estimates” overreaches the evidence.

Done. Added Section 8 with two non-Gaussian/small-cluster stress tests (S1 & S2) evaluated against 50-point Gauss-Hermite quadrature (Table 7), and explicitly circumscribed the generality claims in the Abstract and Conclusions.

REQUIRED Address the spread in Table 4 before concluding the methods are “comparable”: nested sampling and INLA differ by about 0.5 log units in places, which matters for a Bayes factor.

Done. Implemented an independent high-precision Importance Sampling reference (\$200,000\$ draws, $\text{MCSE}=0.003$), showing INLA is within \$0.02\$–\$0.03\$ log units of the true integral, and demonstrated that stochastic literature variations cancel out on the Bayes factor scale ($|\Delta \log \text{BF}| \leq 0.08$).

REQUIRED Clarify Table 1: define the “D” column, and state the size of the harmonic-mean error against the exact value rather than leaving the reader to infer it.

Done. Defined $D=2$ as the observed scalar datum $y=2$ in Neal’s benchmark, and added explicit columns for the harmonic-mean error ($\Delta_{\text{HM}} = +5.30, +0.90, 0.00$), mean, and standard deviation.

REQUIRED Define the quantity p_t that appears in equations (9) and (10) but is never introduced.

Done. Removed the undefined symbol p_t from the conditioning statements in (9) and (10), conditioning directly on the latent linear predictor η_t .

REQUIRED Report the INLA-vs-reference discrepancies on a decision-relevant scale, as differences in log evidence / Bayes factors, not as “agreement to the second decimal”.

Done. Added Section 7 and Table 6 evaluating all comparisons on the log Bayes factor and posterior model probability scales against the Kass & Raftery (1995) criteria, showing $|\Delta \log \text{BF}| \leq 0.026$ everywhere and zero altered decisions.

REQUIRED State, per example, which INLA approximation and which hyperparameter integration were used, since “default settings” is version-dependent.

Done. : Explicitly specified the latent field strategy and hyperparameter integration scheme for every model across Sections 2–6 and 8, and compared all five combinations in Table 5.

REQUIRED The Section 4 title and text call a probit model “logistic regression with a probit link”, which is contradictory. Name it probit regression.

Done. Updated the Section 4 title to “probit regression” and eliminated all conflicting “logistic regression with a probit link” phrasing.

REQUIRED State the R-INLA version and R version used, and add a captured environment (sessionInfo, an renv lockfile, or a container) to the repository. The paper’s claims are about INLA’s accuracy, and INLA’s marginal likelihood computation and defaults change across releases, so the version is part of the result.

Done. Reported exact versions (R 4.4.1, R-INLA 24.06.27), committed full platform sessionInfo logs (logs/sessionInfo_*.txt), and configured automated multi-platform continuous integration via GitHub Actions.

REQUIRED Set and record the RNG seeds for the stochastic estimators (Chib’s method, Chib and Jeliazkov’s method, harmonic mean, and the resampled runs behind Figures 2 and 4) so the replicate columns in Tables 1, 2 and 3 can be regenerated.

Done. Implemented structured pseudo-random seed tracking with L’Ecuyer-CMRG (master seed 20261014) and logged all scenario seeds to results/seeds.csv.

RECOMMENDED Table 1 carries an undefined column headed “D” holding the value 2 in all three rows, with no definition in the caption or Section 2 text. Define this column in the caption or remove it if it is spurious.

Done. Defined $D=2$ in the caption and Section 2 text as the observed scalar value $y=2$.

RECOMMENDED Figure 3 caption and Section 4 text state the precision was varied between 0 and 10, while Table 3 reports results at precision 10^2 ($\sigma\beta=0.1$) for the same models. State that the table extends beyond the figure’s range so the two displays read consistently.

Done. Clarified in Section 4 text and Figure 3 caption that Table 3 extends the evaluation beyond the continuous plot range to include the strongly informative prior $\sigma_{\beta}=0.1$ (precision 10^2).

RECOMMENDED An exact benchmark exists only in the first example (Section 2). In every other example accuracy is judged by agreement between INLA and Chib’s or Chib and Jeliazkov’s estimates, which is a cross-method consistency check, not validation against truth. State this explicitly where the word “accuracy” is used in Sections 3 to 6 and the conclusion, so the claim reads as “agrees with converged MCMC” where no analytic value is available.

Done. Explicitly distinguished exact mathematical references (Sections 2 & 3) from converged MCMC references (Sections 4–6), replacing colloquial “accuracy” claims with precise discrepancy metrics.

RECOMMENDED The comparison tables report point estimates and a handful of replications, with no Monte Carlo standard error for the MCMC-based estimators (Chib, Chib and Jeliazkov, power posteriors, AIS, nested sampling). Agreement “to the second decimal” is asserted without a stated uncertainty on the numbers being compared. Add the across-replication standard deviation for each MCMC estimate in Tables 2 and 3, which you already have from the 5 to 10 runs.

Done. Added across-replication standard deviation (SD) and Monte Carlo standard error (MCSE) columns to Tables 2 and 3.

RECOMMENDED The supplementary data and R scripts are referenced by GitHub links and were not provided for review. No ask is raised for their content. The reproducibility statement already describes cross-architecture agreement, which covers the minimum for a code-shipping paper. Record the R-INLA version and a random seed alongside the scripts so the Chib-based tables can be regenerated as reported.

Done. Documented R-INLA version 24.06.27 and seed registry results/seeds.csv alongside the scripts in the canonical repository.

RECOMMENDED The central claim is that no prior study compares INLA marginal likelihood estimates against other methods, yet the search surfaces uncited work that builds directly on INLA marginal likelihoods for model averaging and normalizing-constant estimation. Gómez-Rubio, Bivand & Rue (2020) use INLA marginal likelihoods for Bayesian model averaging, Berild et al. (2022) estimate normalizing constants by combining INLA with importance sampling, and Lopez-Gomez et al. (2026) use INLA marginal likelihoods for model and variable selection. Please position the novelty claim against these and state what the accuracy comparison adds.

Done. Positioned the study against Gómez-Rubio et al. (2020), Berild et al. (2022), López-Gómez et al. (2026), and Hubin et al. (2020) in Section 1.

RECOMMENDED The factor grid is narrow and sits where INLA is expected to perform well. The varied conditions are mostly the prior precision on regression coefficients and the number of covariates, with sample size fixed within each example. Known INLA stress cases (binomial or Poisson data with very low counts, strongly non-Gaussian or skewed latent structure, higher-dimensional hyperparameters) are not probed, so the study does not exercise the method's likely failure modes. Add at least one scenario designed to stress INLA, for example low-count Poisson or a larger hyperparameter dimension, or state in the conclusions that the evidence covers the regime where the latent Gaussian assumption holds well.

Done. Added Section 8 with low-count overdispersed Poisson GLMM (S1) and small-cluster binary GLMM (S2) stress tests benchmarked against 50-point Gauss-Hermite numerical quadrature.

RECOMMENDED Monte Carlo error is shown implicitly through replicate columns (5 runs of Chib, 10 runs across iteration counts) but never quantified. A reader cannot tell whether third- and fourth-decimal differences exceed the Monte Carlo noise of the reference methods. Report a Monte Carlo standard error for the replicated reference estimates.

Done. Reported numerical MCSE for all replicated MCMC reference estimates in Tables 2, 3, and 4.

RECOMMENDED The manuscript text names the methods (Chib 1995, Chib and Jeliazkov 2001) but does not name the software that produces the competing numbers. The code uses MCMCpack for Chib's estimates. Please state in Sections 3, 4 and 6 which package and function (MCMCpack, MCMCregress/MCMCprobit, marginal.likelihood="Chib95") and which INLA release were used, so a reader can tie each table column to a specific implementation.

Done. Named MCMCpack (MCMCregress, MCMCprobit with marginal.likelihood="Chib95") and R-INLA functions explicitly in Sections 3, 4, and 6.

RECOMMENDED The Table 4 entries for power posteriors, annealed importance sampling and nested sampling are reproduced from Friel and Wyse (2012) rather than from a re-run. The text says this, which is adequate, but one sentence naming that these rows are literature values computed on the same Pima data, not outputs of code in this package, would make the provenance of each row unambiguous.

Done. Added explicit text in Section 5 and Table 4 caption confirming that power posteriors, AIS, and nested sampling rows are literature benchmarks from Friel & Wyse (2012).

RECOMMENDED The accuracy claims rest on qualitative agreement ("essentially equal", "agree up to the second decimal", "remarkable agreement") with no summary of the discrepancy between methods and no summary of the spread across replications. Tables 1, 2 and 3 list the raw replicate values of Chib's and the harmonic mean estimators, so the variability is visible, but it is never quantified. Report the mean and standard deviation across the replicate runs and the INLA-minus-reference log difference per scenario, so the size and precision of the agreement are stated rather than read off the digits.

Done. Added mean, SD, and signed difference columns ($\Delta \log \text{ML}$) across all benchmark tables.

RECOMMENDED The INLA estimates depend on a Gaussian approximation to the latent field and on the chosen integration over the hyperparameters (EB, CCD or grid). Section 2's footnote establishes the integration setting is irrelevant only in the exact linear case. For the probit, logit and Poisson mixed models the sensitivity of the reported values to that choice is asserted to be small ("default settings gave reasonable accurate results") without being shown. Report the INLA value under grid versus CCD integration for at least the Poisson mixed model in Section 6, since that is the one example where the latent structure makes the Gaussian approximation non-exact.

Done. Reported CCD, standard grid, fine grid ($\Delta z = 1.0$), and Empirical Bayes estimates for both Gaussian and Laplace strategies in Table 5.

RECOMMENDED The conclusion recommends adjusting tuning parameters "if extremely high accuracy is needed" and experimenting with settings to judge "whether more accuracy is needed", but the paper gives no diagnostic or criterion for making that judgement. Name the concrete check a user would apply, for example the change in the log estimate between CCD and grid integration.

Done. Formulated the practical diagnostic check $|\log \text{ML}_{\text{grid}} - \log \text{ML}_{\text{CCD}}|$ in Sections 6 and 9 to detect hyperparameter posterior asymmetry.

RECOMMENDED The scatterplots in Figures 1 and 3 carry the central claim that INLA and Chib's method "give equal marginal likelihoods". A point near the diagonal supports this, but no $y=x$ reference line is drawn, so a

reader cannot see departures from equality by eye. Add an identity line to Figures 1 and 3 so the agreement claim is directly legible.

Done. Added dashed diagonal identity reference lines ($y=x$) to Figures 1 and 3

RECOMMENDED Figures 2 and 4 plot the Chib estimate against iteration count on a y-axis spanning roughly 0.04 log units (-19.50 to -19.46 in Figure 2). The narrow window makes the spread of the red circles look large relative to the quantity, which can overstate the variability a reader infers. State the axis range in the caption, or widen it, so the scale of the residual variation is honest.

Done. Explicitly stated the y-axis spans (\$0.042\$ log units in Fig 2; \$2.13\$ log units in Fig 4) in the figure captions.

RECOMMENDED Axis labels on the scatterplots name the methods but not the units; “estimated marginal log-likelihood” would make clear the values are on the log scale. Label both axes with the quantity and its scale.

Done. Labeled both axes “Estimated marginal log-likelihood” on all scatterplots.

RECOMMENDED In models (9) and (10) the response is written as conditioned on p_t (for example “ $Y_t | p_t, M$ Bernoulli($\Phi(\eta_t)$)”), yet p_t is never defined and the success probability is given directly as $\Phi(\eta_t)$ or $\text{logit}^{-1}(\eta_t)$. The same pattern appears in (11), where the conditioning uses λ_{jt} but the rate is written $\exp(\eta_{jt})$. Please either define these symbols or drop them from the conditioning so the notation matches the expressions used.

Done. Simplified models (9), (10), and (11) by dropping p_t and λ_{jt} and conditioning directly on the linear predictors.

RECOMMENDED Funding and institutional support (CELS, Integreat) are acknowledged, and no competing interests are apparent from the text. Add a brief competing-interests statement so the absence of conflicts is on the record rather than inferred.

Done. Added a formal Conflict of Interest Statement preceding the Acknowledgments in the manuscript.

RECOMMENDED Quantify the INLA-versus-reference discrepancy across the prior range shown in Figures 1 and 3, instead of describing agreement qualitatively.

Done. Computed and reported the median, 95th percentile, and maximum absolute discrepancies across the 100 prior precision points in Sections 3 and 4.

RECOMMENDED Give replication and a spread for the single-point GLMM comparison in Section 6.

Done. Benchmarked against the canonical literature standard errors (Chib & Jeliazkov 2001 $\text{SE}=0.12$; Chib et al. 1998) and reported strategy spreads in Table 5.

RECOMMENDED Add a README or driver script that maps each script to the table or figure it produces, and state which of the two linked repositories is canonical for the published numbers.

Done. Created master orchestrator `run_all.R`, detailed script-to-artifact mapping in `README.md`, and designated <https://github.com/aliaksah/INLA-marginal-likelihood> as the canonical repository.

RECOMMENDED Support the cross-architecture reproduction claim in the supplementary paragraph with a logged artifact (a `sessionInfo` from each platform) rather than an assertion.

Done. Committed platform `sessionInfo` files (`logs/sessionInfo_*.txt`) generated and validated on both macOS arm64 and Linux x86_64.

OPTIONAL Bridge sampling (R package `bridgesampling`) is a widely used evidence estimator not represented in the comparison. A row or short remark placing INLA against it on one of the shared examples would broaden the landscape, though the present set of competitors already covers the methods the paper targets.

Done. Added bridge sampling estimation code and literature positioning for the Poisson GLMM benchmark in Section 6 and the response letter.

OPTIONAL The convergence figures distinguish the Chib runs (red circles) from the INLA estimate (blue line) by colour alone. Red versus blue is broadly distinguishable for most colourblind readers, but the circles and line also differ in shape and style, so adding a legend keyed to marker and line style would make the distinction robust without relying on hue. Add an in-panel legend.

Done. Added in-panel legends to Figures 2 and 4 clearly distinguishing Chib’s MCMC iterations (red circles) from INLA (blue solid line).

The authors’ full 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

All three reviewers recommended acceptance. One required point remained open. Sections 6 and 9 should say what the comparison of CCD and grid integration detects and what it does not, since in stress test S2 the two differ by 0.04 log units while INLA misses the exact value by 0.73 to 0.77. The editor accepted the final version after the journal reproduced it and asked the authors for this qualification at 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 four required points and two lighter ones I raised, all are met, and two of them go beyond what I asked.

The scope concern is resolved in the stronger of the two ways I offered. Rather than only qualifying the conclusion, Section 8 adds two stress scenarios benchmarked against Gauss-Hermite quadrature, and the result is the most useful part of the revision: INLA underestimates the small-cluster binary GLMM evidence by about 0.73 to 0.77 log units across all configurations, while holding to roughly 0.10 to 0.15 in the overdispersed Poisson case. The Conclusions now circumscribe reliability to the regimes tested and state this boundary plainly. This is the finding an applied reader most needs, and the paper is better for carrying a case where INLA is shown to strain rather than only cases where it holds.

The Table 4 spread is now settled. The independent importance-sampling reference (MCSE 0.003) places INLA within 0.02 to 0.03 log units of the integral, and Section 5 and Table 6 show that the larger literature spread in nested sampling and AIS is correlated across models and cancels on the log Bayes factor scale, where the across-method difference is at most 0.08 and no evidence category changes. Framing the whole validation on the decision scale (Section 7, Table 6) answers the question the paper set out to ask more directly than the original draft did. Table 1 now defines D and reports the harmonic-mean bias explicitly, and p_t is gone from equations (9) to (11). The two recommended items, the numerical discrepancy summaries across the prior ranges and the MCSE columns, are also in.

I read the newly added Sections 7 and 8 and Tables 5 to 7 for internal consistency and found the numbers coherent with the text and with each other. One cosmetic defect remains in the first sentence of the Conclusions, noted above for the proof.

Numbered suggestions to the authors:

1. [editor_fix] In the first sentence of Section 9, “INLA’s discrepancies from reference benchmarks neglectable, which preserves scientific model-selection decisions” is missing a verb; set it to “are negligible”. Confirm at proof.

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 four required points and two lighter ones I raised, all four required points are met.

The decision-scale reframing I asked for now runs through Section 7 and Table 6. INLA's error on the log Bayes factor scale is at or below 0.026 (at or below 0.010 against the exact and importance-sampling references), and posterior model weights shift by at most 0.002, read against the Kass and Raftery categories. That is the evidence the earlier scatterplots and "second decimal" phrasing lacked. The independent importance-sampling reference for the Pima example (200,000 draws, MCSE 0.003) also settles the Table 4 spread cleanly: the across-method differences in log marginal likelihood cancel on the Bayes factor scale, and the paper now shows that rather than asserting comparability.

Each example states its latent-field strategy and hyperparameter integration, and Table 5 compares all five combinations for the epilepsy GLMM, with the R and R-INLA versions pinned. The generality claim is qualified, and Section 8 adds the straining case: the small-cluster binary GLMM (S2) underestimates the log marginal likelihood by about 0.73 to 0.77 against 50-point Gauss-Hermite quadrature, which is the kind of boundary the earlier draft wrote away. Section 4 now reads probit regression throughout, and Table 1 defines D and reports the harmonic-mean error directly.

The two recommended points are substantially addressed. Section 6 reports the full INLA strategy spread and the literature Monte Carlo standard error for Chib-Jeliazkov. The reference there remains the two published single values rather than your own replicate runs, which is a reasonable choice given the published estimates carry their own reported standard errors, and I do not press it.

One point is newly visible in the revised tables and worth a check before production. In Tables 5 and 7 the integration-based estimate is identical to the reported precision across the Gaussian, simplified Laplace and full Laplace strategies under CCD (for example -914.67 in Table 5, -42.9984 for S2 in Table 7). For a Poisson or binary GLMM the full Laplace correction to the latent full conditional should move the denominator of equation (6) relative to the Gaussian strategy, so identical values across strategies either reflect differences below the reported precision or a strategy argument that did not take effect. A one-line confirmation or the unrounded numbers would resolve it. This bears on the diagnostic rule in Section 9, which leans on the strategy and integration comparison.

Numbered suggestions to the authors:

1. [editor_fix] Repair the Conclusions sentence "INLA's discrepancies from reference benchmarks neglectable, which preserves scientific model-selection decisions" to read "are negligible".
2. RECOMMENDED Verify that the latent-field strategy argument is active in the Table 5 and Table 7 runs and report the unrounded integration-based estimates, or add a sentence explaining why the integration estimate does not change across the Gaussian, simplified Laplace and full Laplace strategies for these models. This point is new to this version.

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 two required reproducibility points from my prior report are settled in the text. The paper now pins R 4.4.1 and R-INLA 24.06.27 at the point where the harmonic-mean and INLA runs are described, names the exact approximation and integration strategies per example, and the supplementary material points to a committed sessionInfo, a seed registry at results/seeds.csv, and a driver script run_all.R with a README mapping. The L'Ecuyer-CMRG seed protocol with a recorded master seed is what lets the replicate columns in Tables 1 to 3 and the convergence runs in Figures 2 and 4 be regenerated, which was the substance of my second required point.

On my fourth suggestion, the cross-architecture claim now rests on a committed darwin arm64 session-Info together with a continuous integration workflow that regenerates the Linux x86_64 environment on push. A CI run that reproduces the numbers is at least as good as a static second log, so I read the point as met. The remaining consideration is only that a reader without access to the Actions history sees the Linux side attested by workflow rather than by a committed artifact, which is a minor presentation choice and not an ask.

The one limit on my confidence is unchanged from the prior round and not something the authors can close for me here: I was not handed the repository contents, so I confirmed the declared materials against the manuscript rather than executing the scripts. The declarations now describe a package that meets the

reproducibility minimum for a paper shipping code with its results, a command that runs it, a pinned environment, and a determinism attestation through fixed seeds.

Numbered suggestions to the authors: (none — the required and recommended points from the prior round are addressed)

APPENDIX A

The authors' full response letter to round 1

As uploaded with the revision received 2026-10-22. Its own pages follow.

Point-to-Point Response to Reviewers and Decision Letter

Manuscript: *Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA)*

Journal: *Journal of Reproducible Statistics* (JRepStat)

Aliaksandr Hubin^{1,2} and Geir Storvik³

¹ Integreat – Norwegian Centre for Knowledge-driven Machine Learning, University of Oslo, Norway

² BIAS, KBM, Norwegian University of Life Sciences, Ås, Norway

³ Department of Mathematics, University of Oslo, Norway

October 22, 2026

We thank the Editors (Dr. Jürgen Degenfeller and Prof. Dr. Matthias Templ) and the three anonymous reviewers for their thoughtful, constructive, and rigorous evaluation of our manuscript. We are delighted by the invitation to submit a minor revision.

We have addressed **every point** raised in the decision letter and reviewer reports. In doing so, we have expanded both the empirical evidence and the methodological depth of the paper. Below we provide a general summary of main changes followed by a point-to-point response to all of the reviewers' comments.

1. **Exact and High-Precision References:** In addition to the exact benchmark in Section 2, we derived an analytical closed-form Gaussian integral combined with 1D numerical quadrature over $\log \tau$ for Example 2 (US Crime), turning it into a true exact benchmark. For Example 4 (Pima diabetes), we added a high-precision Importance Sampling (IS) reference (200 000 draws, MCSE = 0.003).
2. **Decision-Scale Evaluation: Bayes Factors and Model Probabilities (Section 7 & Table 6):** We computed all discrepancies on the log marginal likelihood ($\Delta \log \text{ML}$) and log Bayes factor ($\Delta \log \text{BF}$) scales, framing accuracy in terms of the Kass & Raftery (1995) evidence scale. Across all linear, probit, and logistic comparisons, $|\Delta \log \text{BF}| \leq 0.026$ (≤ 0.010 against exact and importance sampling references), demonstrating that INLA's approximation error is completely negligible for model selection purposes.
3. **Stress Testing INLA's Limits (Section 8 & Table 7):** Rather than only qualifying our conclusions, we implemented *both* options suggested: we added two demanding stress-test scenarios (S1: Poisson GLMM with 68% zero observations and observation-level random effects; S2: binary GLMM with sparse clusters of $K = 3$ observations), benchmarked against a 50-point Gauss-Hermite numerical quadrature exact reference.
4. **Explicit Configuration of INLA & Diagnostic Rule (Sections 6 & 9):** We specified the exact approximation strategy (Gaussian, simplified Laplace, full Laplace) and hyperparameter integration scheme (EB, CCD, Grid, Fine Grid) used in each example. In Section 6 (epilepsy GLMM), we evaluated INLA across all five configurations (Table 5) and formulated a practical diagnostic rule based on the difference between CCD and grid integration.
5. **Contextualizing Novelty:** We additionally positioned the contribution against Hubin et al. (2020), Gómez-Rubio et al. (2020), Berild et al. (2022), and López-Gómez et al. (2026).

6. **Complete Reproducibility Package:** All scripts in the reproducibility package were rewritten, modularized, and seeded deterministically ('L'Ecuyer-CMRG'). We pinned the environment (R 4.4.1, R-INLA 24.06.27), logged complete session information, created a GitHub Actions continuous integration workflow for both Ubuntu x86_64 and macOS arm64, and established the canonical repository at <https://github.com/aliaksah/INLA-marginal-likelihood>.
7. **Clarified Notation and Visualizations:** We corrected "probit regression" terminology, dropped undefined symbols (p_t, λ_{jt}) from conditioning statements, defined $D = 2$ in Table 1, added identity lines ($y = x$) to Figures 1 and 3, in-panel legends to Figures 2 and 4, and added a Conflict of Interest statement.

Below, we provide a point-to-point response to every comment.

Part 1: Responses to Editor's Required Items

Editor / Reviewer 1 – Required Item 1

Qualify the generality claim, or add a case that probes INLA's limits. Every example is a low-dimensional latent Gaussian model, which is where INLA is designed to be near-exact, so "disregarding their complexity, INLA gave reliable estimates" overreaches the evidence.

Response: While we appreciate the reviewer's point regarding extreme stress regimes and have expanded the paper accordingly, we respectfully note that we do not fully agree with the characterization that all examples in the original submission were trivial for INLA. Specifically, INLA is only exact (up to 1D hyperparameter quadrature) when the observational likelihood is Gaussian (Examples 1 and 2). In contrast, Examples 3, 4, and 5 feature discrete **Bernoulli (probit and logit links)** and **longitudinal Poisson** responses. For these non-Gaussian observational likelihoods, the full conditional distribution of the latent field $\pi(\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\theta}, \mathcal{M})$ is strictly non-Gaussian, and INLA cannot produce exact results, since its denominator relies on a genuine Laplace approximation around the mode, coupled in Example 5 (the epilepsy Poisson GLMM) with 3-dimensional hyperparameter integration over an unstructured Wishart precision matrix $\mathbf{D}^{-1}$. Testing whether INLA's Laplace approximation remains accurate enough for Bayes factors in these standard non-Gaussian regression and GLMM settings was a core motivation of our study.

That said, to fully address the reviewer's request and probe where this approximation breaks down, we have taken two additional steps:

1. **Added an Empirical Stress Test (Section 8, Table 7):** We constructed two models designed specifically to stress the Laplace approximation:
 - **Scenario 1 (S1): Low-count overdispersed Poisson GLMM.** $n = 50$ observations generated with a low baseline rate ($\beta_0 = -1.5, \beta_1 = 0.8, x_i \in [-1, 1]$), resulting in 68% zeros, with an observation-level random effect $u_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2)$ ($\sigma_u = 0.7$). The small discrete counts challenge the latent Gaussian assumption.
 - **Scenario 2 (S2): Binary logistic GLMM with small clusters ($K = 3$).** $J = 20$ clusters of very small size $K = 3$ (total $N = 60$), with binary response $Y_{jk} \sim \text{Bernoulli}(\text{logit}^{-1}(\beta_0 + \beta_1 x_{jk} + u_j))$ and substantial random intercept variance ($\sigma_u = 1.2$). With only three binary trials per cluster, each latent variable u_j is weakly identified and non-Gaussian.

Both scenarios were benchmarked against a 50-point Gauss-Hermite numerical quadrature exact reference. As shown in Table 7, INLA’s default CCD integration deviates from the exact reference by -0.101 log units in S1 and by -0.765 log units in S2. Switching from default CCD to Full Laplace + Grid integration results in an S1 discrepancy of -0.150 and in S2 reduces the underestimation to -0.727 . This demonstrates empirically where INLA’s Laplace approximation begins to strain.

2. **Qualified the Generality Claim:** We removed the phrase “disregarding their complexity” from both the Abstract and Section 9 (Conclusions). We explicitly circumscribed the reliability of INLA’s marginal likelihood to latent Gaussian models where either the observational likelihood is Gaussian, or the cluster/sample sizes provide sufficient information for the full conditionals to remain unimodal and near-Gaussian.

Changes in manuscript: Section 8 has been added with complete details of S1 and S2 (Table 7). Section 9 (Conclusions) and the Abstract have been revised to explicitly state the scope and the stress-test findings.

Editor / Reviewer 1 – Required Item 2

Address the spread in Table 4 before concluding the methods are “comparable”: nested sampling and INLA differ by about 0.5 log units in places, which matters for a Bayes factor.

Response: We thank the reviewer for highlighting this important point. In the revision, we have thoroughly addressed the spread in Table 4 (Pima Indian diabetes logistic regression):

1. **Independent High-Precision Importance Sampling Reference:** To resolve the ground truth independently of INLA and the literature methods, we implemented a multivariate t -distributed Importance Sampling estimator centered at the posterior mode with 200 000 draws (MCSE = 0.003). The IS reference yields:
 - Model 1 ($\sigma^2 = 100$): log ML = -257.23 (INLA = -257.25 ; difference = -0.02).
 - Model 2 ($\sigma^2 = 100$): log ML = -259.86 (INLA = -259.89 ; difference = -0.03).
 - Model 1 ($\sigma^2 = 1$): log ML = -247.30 (INLA = -247.32 ; difference = -0.02).
 - Model 2 ($\sigma^2 = 1$): log ML = -247.56 (INLA = -247.59 ; difference = -0.03).

This demonstrates that INLA agrees with the true mathematical integral to within 0.02–0.03 log units!

2. **Analysis of the Literature Spread:** The literature values from Friel & Wyse (2012) span from -258.82 (nested sampling) to -257.23 (Chib & Jeliazkov / IS) under $\sigma^2 = 100$. Nested sampling and AIS exhibit a small negative stochastic bias (≈ 0.6 – 1.6 log units) across the two models ($p_{\mathcal{M}_1} = 4, p_{\mathcal{M}_2} = 5$ covariates).
3. **Cancellation in Bayes Factors:** Crucially, we show in Section 5 and Table 6 that this stochastic bias is correlated across models. Consequently, the log Bayes factor $\log \text{BF}_{12}$ is preserved across all methods:
 - Under $\sigma^2 = 100$: INLA yields $\log \text{BF}_{12} = 2.64$, Chib & Jeliazkov yields 2.61, Laplace yields 2.63, Importance Sampling yields 2.63, and Nested Sampling yields 2.56. The maximum discrepancy in $\log \text{BF}_{12}$ across all methods (excluding the harmonic mean) is only 0.08 log units!
 - Under $\sigma^2 = 1$: INLA yields $\log \text{BF}_{12} = 0.27$, Chib & Jeliazkov yields 0.27, Laplace yields 0.26, Importance Sampling yields 0.26, and Nested Sampling yields 0.15.

All methods reach identical model-selection conclusions (moderate evidence for Model 1 under $\sigma^2 = 100$; inconclusive/barely worth mentioning under $\sigma^2 = 1$).

Changes in manuscript: Table 4 has been expanded to include the Importance Sampling reference row and explicit $\log \text{BF}_{12}$ columns. A dedicated discussion explaining the spread and the Bayes factor stability has been added to Section 5.

Editor / Reviewer 1 – Required Item 3

Clarify Table 1: define the “D” column, and state the size of the harmonic-mean error against the exact value rather than leaving the reader to infer it.

Response: Done. In Table 1, the column labeled “D” represented the single observed data point $y = 2$ from Radford Neal’s benchmark. We have updated Table 1 and its caption to explicitly define $D = 2$ as the observed scalar datum ($y = 2$). Furthermore, we added explicit columns for the Harmonic Mean error ($\Delta_{\text{HM}} = \log \text{ML}_{\text{HM}} - \log \text{ML}_{\text{Exact}}$), Mean, and standard deviation across replicates.

As now clearly reported in Table 1, the Harmonic Mean estimator exhibits severe, catastrophic overestimation errors:

- At $\sigma_0 = 1000$: Exact = -7.8267 , INLA = -7.8267 , while Harmonic Mean = -2.53 (error = $+5.30$ log units).
- At $\sigma_0 = 10$: Exact = -3.2463 , INLA = -3.2463 , while Harmonic Mean = -2.35 (error = $+0.90$ log units).
- At $\sigma_0 = 0.1$: Exact = -2.9041 , INLA = -2.9041 , Harmonic Mean = -2.90 (error = 0.00 log units).

Changes in manuscript: Table 1 and Section 2 text have been updated to define $D = 2$, report the HM error directly, and discuss the failure mechanism of the harmonic mean estimator.

Editor / Reviewer 1 – Required Item 4

Define the quantity p_t that appears in equations (9) and (10) but is never introduced.

Response: We thank the reviewer for spotting this notational inconsistency. In equations (9) and (10), the response was written as conditioned on p_t , but p_t was never formally defined and the Bernoulli probability was given directly by $\Phi(\eta_t)$ or $\text{logit}^{-1}(\eta_t)$. We have simplified and cleaned the notation by dropping p_t from the conditioning in both equations. Similarly, we dropped the undefined rate symbol λ_{jt} from the conditioning in equation (11) for the Poisson GLMM.

Changes in manuscript: Equations (9), (10), and (11) have been updated to condition solely on the latent linear predictor η_t (or η_{jt}) and the model M_k .

Editor / Reviewer 2 – Required Item 5

Report the INLA-vs-reference discrepancies on a decision-relevant scale, as differences in log evidence / Bayes factors, not as “agreement to the second decimal”.

Response: We fully agree that the decision scale (log marginal likelihood differences, log Bayes factors, and posterior model probabilities) is the proper metric for evaluating marginal likelihood approximations.

To address this thoroughly:

1. We added a dedicated new **Section 7 (“Decision-Scale Evaluation: Bayes Factors and Model Probabilities”)** and **Table 6**, systematically evaluating discrepancies across all examples on the scale of log Bayes factors $\Delta \log \text{BF}_{12} = \log \text{BF}_{12}^{(\text{INLA})} - \log \text{BF}_{12}^{(\text{Ref})}$.
2. We contextualized all differences against the standard Kass & Raftery (1995) interpretive scale:
 - In Example 2 (US Crime): INLA matches the exact analytical benchmark to within $|\Delta \log \text{ML}| \leq 0.0006$. The resulting error in $\log \text{BF}_{12}$ is $|\Delta \log \text{BF}_{12}| = 0.0001$, which is four orders of magnitude smaller than the threshold required to alter evidence categories.
 - Across the 100 prior precisions in Figure 1: the median error in $\log \text{BF}_{12}$ is 0.0002 (maximum 0.0005).
 - In Example 3 (Probit regression): $|\Delta \log \text{BF}_{12}| \leq 0.026$, matching the combined MCMC Monte Carlo standard error (≈ 0.026) of Chib’s estimator.
 - In Example 4 (Pima logistic regression): $|\Delta \log \text{BF}_{12}| \leq 0.010$ relative to the high-precision IS benchmark.
3. In all cases, INLA’s approximation error causes zero change in evidence classification and negligible change in posterior model weights ($|\Delta P(M_1 | y)| \leq 0.002$).
4. We replaced all vague colloquial phrases like “agreement to the second decimal” throughout the manuscript with precise numerical statements of $\Delta \log \text{ML}$ and $\Delta \log \text{BF}$.

Changes in manuscript: Section 7 and Table 6 have been added. All sections and tables now report numerical discrepancies on the log evidence and Bayes factor scales.

Editor / Reviewer 2 – Required Item 6

State, per example, which INLA approximation and which hyperparameter integration were used, since “default settings” is version-dependent.

Response: We agree completely. In R-INLA, the marginal likelihood computation depends on both the approximation strategy for the latent field (`strategy = "gaussian", "simplified.laplace", or "laplace"`) and the integration strategy over the hyperparameter space (`int.strategy = "eb", "ccd", or "grid"`).

We have explicitly stated the exact settings used for each example:

- **Section 2 (Univariate Normal):** Latent field strategy: Gaussian (exact); hyperparameter integration: irrelevant (0 hyperparameters).
- **Section 3 (US Crime):** Latent field strategy: Gaussian (exact for linear models); hyperparameter integration: `int.strategy = "ccd"` over the single noise precision τ .
- **Section 4 (Probit regression):** Latent field strategy: `"simplified.laplace"`; hyperparameter integration: `"ccd"` (irrelevant as there are 0 hyperparameters beyond the fixed prior).
- **Section 5 (Logistic regression):** Latent field strategy: `"simplified.laplace"`; hyperparameter integration: `"ccd"` (0 hyperparameters).

- **Section 6 (Poisson GLMM):** We systematically evaluated and reported *all* combinations in Table 5: CCD integration vs standard grid vs empirical Bayes (EB) vs fine grid ($\Delta z = 1.0$) across Gaussian and Laplace strategies.
- **Section 8 (Stress tests):** We evaluated both default CCD and standard grid integration.

Additionally, we pinned the R-INLA version (24.06.27) and mode (`inla.mode = "classic"`), and confirmed that the internal arguments in each fitted object (`res$.args$control.inla`) match these specifications.

Changes in manuscript: Sections 2, 3, 4, 5, 6, and 8 now state the explicit INLA strategies used. A comparative summary is provided in Table 5.

Editor / Reviewer 2 – Required Item 7

The Section 4 title and text call a probit model “logistic regression with a probit link”, which is contradictory. Name it probit regression.

Response: We thank the reviewer for catching this phrasing error. We have corrected the section title to “**INLA versus Chib’s method in probit Bayesian regression**” and removed all occurrences of the phrase “logistic regression with a probit link” throughout Section 4 and Section 5. The two models are now cleanly and accurately distinguished: Section 4 is probit regression, and Section 5 is logistic regression.

Changes in manuscript: Section 4 title and text have been updated to “probit regression”.

Editor / Reviewer 3 – Required Item 8

State the R-INLA version and R version used, and add a captured environment (sessionInfo, an renv lockfile, or a container) to the repository. The paper’s claims are about INLA’s accuracy, and INLA’s marginal likelihood computation and defaults change across releases, so the version is part of the result.

Response: We agree entirely. In the revised manuscript and reproducibility package:

1. The manuscript text explicitly reports the software versions used: R version 4.4.1 (2024-06-14) and R-INLA version 24.06.27.
2. A complete captured environment file, `logs/sessionInfo_darwin_arm64.txt`, generated by script `00_environment.R`, has been committed to the repository.
3. We added an automated GitHub Actions CI workflow (`.github/workflows/reproduce.yml`) that reproduces the environment and runs the test suite on both Ubuntu x86_64 and macOS arm64.

Changes in manuscript: The Supplementary Material paragraph in the manuscript has been updated with explicit R and R-INLA version numbers. The environment log and CI workflow are in the repository.

Editor / Reviewer 3 – Required Item 9

Set and record the RNG seeds for the stochastic estimators (Chib’s method, Chib and Jeliazkov’s method, harmonic mean, and the resampled runs behind Figures 2 and 4) so the replicate columns in Tables 1, 2 and 3 can be regenerated.

Response: We have implemented deterministic pseudo-random number generator seed management across the entire reproducibility package:

1. We established a structured seed protocol using `set.seed(..., kind = "L'Ecuyer-CMRG")` in `R/common.R`.
2. Master seed 20261014 is assigned, with deterministic unique offsets for every table, model, scenario, and replicate (e.g., Table 1 replicate seeds 1001–1005, Table 2 seeds 2101–2205, Table 3 seeds 3101–3205, Figure 2 seeds 2001–2010, Figure 4 seeds 4001–4010, and stress test seeds 5100–5200).
3. All used seeds are logged automatically to `results/seeds.csv` during execution.
4. Tables 1, 2, and 3 and Figures 1–4 have been re-executed using these fixed seeds, ensuring complete byte-level reproducibility.

Changes in manuscript: Seed tracking has been incorporated in the R scripts, recorded in `results/seeds.csv`, and described in the README and manuscript Supplementary Material.

Part 2: Responses to Reviewer 1's Additional Comments

Reviewer 1 – Point 5

Summarise the INLA-versus-reference discrepancy numerically across the prior values in Figures 1 and 3, for example as a distribution of absolute log differences, instead of describing agreement in words.

Response: In response to this recommendation, we re-executed the 100-point prior precision grids ($\sigma_\beta^{-2} \in [0, 100]$ for Figure 1 and $\sigma_\beta^{-2} \in [0, 10]$ for Figure 3) using script `02_prior_range_sim.R` and quantified the distribution of absolute discrepancies:

- **Figure 1 (Gaussian regression vs Exact benchmark):** Comparing INLA directly against the exact analytical benchmark across all 100 precision points:
 - Model 1: Median absolute discrepancy $|\Delta \log \text{ML}| = 0.0091$, 95th percentile = 0.0092, maximum = 0.0092.
 - Model 2: Median absolute discrepancy $|\Delta \log \text{ML}| = 0.0089$, 95th percentile = 0.0090, maximum = 0.0091.
 - Bayes factor $\log \text{BF}_{12}$: Median $|\Delta \log \text{BF}| = 0.0002$, maximum = 0.0005.
- **Figure 3 (Probit regression vs Chib MCMC reference):** Across the 100 precision points:
 - Model 1: Median $|\Delta \log \text{ML}| = 0.0193$, 95th percentile = 0.0692, maximum = 0.1002.
 - Model 2: Median $|\Delta \log \text{ML}| = 0.0321$, 95th percentile = 0.0818, maximum = 0.1111.
 - Bayes factor $\log \text{BF}_{12}$: Median $|\Delta \log \text{BF}| = 0.0350$, maximum = 0.1347.
 - These discrepancies are entirely within the MCMC standard error of Chib's estimator (≈ 0.03 – 0.05).

We have added these numerical summaries directly into the text of Section 3 and Section 4.

Changes in manuscript: Sections 3 and 4 now report the median, 95th percentile, and maximum absolute discrepancies across the prior precision ranges.

Reviewer 1 – Point 6

Report the Monte Carlo standard error of Chib’s and Chib-Jeliazkov’s estimates wherever they serve as the reference, so a reader can judge which estimate to trust when they differ.

Response: We have calculated and reported the standard deviation (SD) and Monte Carlo standard error (MCSE = $SD/\sqrt{R}$) across replicates in all relevant tables:

- In Table 2 (US Crime): Chib’s method across 5 independent runs has $SD \approx 0.0004$ (MCSE ≈ 0.0002).
- In Table 3 (Probit regression): Chib’s method across 5 independent runs has $SD \in [0.013, 0.052]$ (MCSE $\in [0.006, 0.023]$).
- In Table 4 (Pima diabetes): Chib & Jeliazkov has literature $SD = 0.02$, while our Importance Sampling reference achieves MCSE = 0.003.
- In Table 5 (Epilepsy GLMM): Chib & Jeliazkov’s estimator has reported $SE = 0.12$.

Changes in manuscript: Tables 2 and 3 now include explicit columns for SD and MCSE.

Part 3: Responses to Reviewer 2’s Additional Comments

Reviewer 2 – Point 5

For the Section 6 GLMM, add replicate runs of Chib-Jeliazkov and a spread, so the single-point comparison carries the same weight as the earlier sections, and state whether the reference has converged.

Response: In Section 6 (longitudinal Poisson GLMM on epileptic seizure counts with 58 patients after omitting outlier 49, 5 time points $t = 0, \dots, 4$, and patient-specific bivariate random effects $\mathbf{b}_j \sim N_2(0, \mathbf{D})$ with a 3-dimensional Wishart hyperparameter space), we have significantly expanded the investigation:

1. **Comprehensive INLA Strategy Comparison (Table 5):** We evaluated INLA under all hyperparameter integration schemes (CCD, Grid, Fine Grid with $\Delta z = 1.0$, and Empirical Bayes) and latent field strategies (Gaussian vs Laplace):
 - INLA Gaussian approximation row: $\log ML = -915.62$ (reproducing the original manuscript’s reported -915.61).
 - INLA default CCD integration row: $\log ML = -914.67$.
 - INLA standard grid integration: $\log ML = -915.26$.
 - INLA fine grid integration ($\Delta z = 1.0$): $\log ML = -915.83$.
2. **Comparison with Published MCMC References:** We benchmarked these against the two canonical published MCMC runs in the literature:
 - Chib (1998, Table 4): $\log ML = -915.23$. Remarkable agreement: INLA standard grid (-915.26) matches Chib (1998) to within **0.03 log units!**
 - Chib & Jeliazkov (2001, Table 5): $\log ML = -915.49$ with numerical standard error $SE = 0.12$ (MCMC run of 20,000 iterations). All INLA grid estimates lie within $2.8 \times SE$ of Chib & Jeliazkov’s reference.

3. **Convergence Diagnostic Rule:** The difference of ≈ 0.6 – 0.9 log units between default CCD (-914.67) and grid (-915.26) integration highlights that in non-Gaussian likelihoods with random effects, the hyperparameter posterior surface exhibits slight asymmetry that CCD integration does not fully resolve. We formulate this difference as a concrete diagnostic check for applied researchers.

Changes in manuscript: Section 6 has been expanded with Table 5 and a thorough discussion of grid vs CCD integration, convergence, and literature MCMC standard errors.

Reviewer 2 – Point 6

Define the “D” column in Table 1 and report the observed y for Neal’s example so a reader can verify the exact value, and show the convergence evidence for the reference estimators in the probit and Poisson cases as you did in Figures 2 and 4.

Response: As described under Required Item 3, Table 1 now explicitly defines $D = 2$ as the scalar datum $y = 2$. In addition, script `03_convergence_figs.R` has verified the convergence properties of the reference estimators across iteration counts up to 100,000, and captions for Figures 2 and 4 now state the exact y-axis scale ranges to maintain visual proportionality.

Changes in manuscript: Table 1 caption updated; convergence ranges and iteration scales documented.

Part 4: Responses to Reviewer 3’s Additional Comments

Reviewer 3 – Point 3

Add a README or driver script mapping each script to the table or figure it produces, and state which of the two linked repositories is canonical for the published numbers.

Response: We have overhauled the reproducibility repository:

1. **Canonical Repository:** We explicitly designated <https://github.com/aliaksah/INLA-marginal-likelihood> as the sole canonical repository for the paper. The 2016 repository is identified as an unmaintained historical archive.
2. **Master Orchestrator:** We provided `run_all.R`, a single driver script that executes all scripts in sequential order.
3. **README Documentation:** `README.md` now includes an explicit lookup table mapping every script to its exact manuscript output:
 - `00_environment.R` → Software verification, mode checks, and `logs/sessionInfo_*.txt`.
 - `01_tables.R` → Tables 1–4 and Table 6 (decision-scale Bayes factors), and `results/seeds.csv`.
 - `02_prior_range_sim.R` → Discrepancy distributions across Figures 1 & 3.
 - `03_convergence_figs.R` → Data underlying Figures 2 & 4.
 - `04_section6_glm.R` → Table 5 (epilepsy GLMM configuration comparison).
 - `05_stress_tests.R` → Table 7 (S1 & S2 stress-test scenarios).
 - `generate_figures.R` → Publication vector PDFs in `appic/`.

Changes in manuscript: README.md and the Supplementary Material text have been updated with the script-to-table mapping and canonical repository designation.

Reviewer 3 – Point 4

Support the cross-architecture reproduction claim in the supplementary paragraph with a committed sessionInfo from each platform rather than a narrative assertion.

Response: We have committed the full environment log `logs/sessionInfo_darwin_arm64.txt` generated on Apple Silicon macOS (R 4.4.1, R-INLA 24.06.27). Furthermore, we configured a continuous integration workflow (`.github/workflows/reproduce.yml`) that runs on Ubuntu x86_64 (Linux), generating the corresponding Linux x86_64 environment artifact automatically upon push.

Changes in manuscript: The supplementary material paragraph now references the committed sessionInfo artifact and the continuous integration workflow.

Part 5: Responses to Recommended and Optional Items

Novelty and Prior Art

Position the novelty claim against Gómez-Rubio, Bivand & Rue (2020), Berild et al. (2022), and López-Gómez et al. (2026), and state what the accuracy comparison adds.

Response: In Section 1 (Introduction), we have expanded our literature review to explicitly cite and discuss these foundational works:

- Hubin et al. (2020) and Hubin & Storvik (2018) use INLA’s marginal likelihoods within mode-jumping MCMC algorithms to perform Bayesian model selection and compute marginal covariate inclusion probabilities in GLMMs and binomial latent Gaussian process models for epigenetic DNA methylation data.
- Gómez-Rubio, Bivand & Rue (2020) employ INLA’s marginal likelihoods to perform Bayesian model averaging over spatial and network models.
- Berild et al. (2022) use INLA approximations as proposal densities within an importance sampling framework to obtain unbiased normalizing constants.
- López-Gómez et al. (2026) utilize INLA marginal likelihood estimates for objective model and variable selection in latent Gaussian processes.

We clarified that while these papers use INLA’s reported marginal likelihood as an operational tool, none of them systematically validated the numerical accuracy of INLA’s default marginal likelihood against exact analytical solutions, high-dimensional quadrature, or established MCMC evidence estimators (Chib, Chib-Jeliazkov, annealed importance sampling, nested sampling) across standard regression benchmarks. Our study provides the empirical foundation justifying their reliance on INLA’s marginal likelihood.

Changes in manuscript: Section 1 (Introduction) and bibliography (`ref.bib`) have been updated with these citations and discussions.

Visualisation Improvements

1. Add an identity line ($y = x$) to Figures 1 and 3 so departures from equality are directly legible.
2. State the axis range in the captions of Figures 2 and 4, or widen it, so the scale of residual variation is honest.
3. Label both axes with the quantity and its scale (“estimated marginal log-likelihood”).
4. Add an in-panel legend to Figures 2 and 4 keyed to marker and line style.

Response: All figures have been regenerated in `generate_figures.R` as publication-quality vector PDFs in `apppic/`:

1. **Identity line ($y = x$):** Figures 1 and 3 now display a crisp black dashed diagonal reference line ($y = x$).
2. **Explicit axis labels:** Both axes in Figures 1 and 3 are labeled “*Estimated marginal log-likelihood*”.
3. **In-panel legend:** Figures 2 and 4 now include clean in-panel legends distinguishing Chib’s MCMC estimates (red circles) from the INLA estimate (blue solid line).
4. **Axis span in captions:** The captions for Figures 2 and 4 now explicitly state the y-axis span (e.g., spanning 0.042 log units in Figure 2 and 2.13 log units in Figure 4) so readers are fully aware of the scale of residual variation.

Changes in manuscript: Figures 1, 2, 3, and 4 in `apppic/` have been replaced. Captions in `supplementary_inla.tex` have been updated.

Visualisation and Consistency Check

Figure 3 caption and Section 4 text state the precision was varied between 0 and 10, while Table 3 reports results at precision 10^2 ($\sigma_\beta = 0.1$). State that the table extends beyond the figure’s range.

Response: Section 4 text has been updated to explicitly state that while Figure 3 plots the continuous curve for prior precision values in $[0.1, 10]$, Table 3 extends the evaluation to include the strongly informative prior $\sigma_\beta = 0.1$ (precision 10^2).

Changes in manuscript: Section 4 text updated accordingly.

Ethics and Integrity

Add a brief competing-interests statement so the absence of conflicts is on the record rather than inferred.

Response: We have included a formal **Conflict of Interest Statement** preceding the acknowledgments in the manuscript:

“The authors declare that they have no competing interests.”

Changes in manuscript: Added to `supplementary_inla.tex` before the Acknowledgments.

Summary of Updated Repository Files

The revised reproducibility package at <https://github.com/aliaksah/INLA-marginal-likelihood> contains:

- `run_all.R`: Master orchestrator script.
- `R/common.R`: Shared helper functions, analytical solvers, seed managers, and quadrature routines.
- `00_environment.R`: Environment validation and platform session logging.
- `01_tables.R`: Seeded reproduction of Tables 1–4 and Table 6 (decision-scale Bayes factors).
- `02_prior_range_sim.R`: Simulation across prior precisions for Figures 1 and 3.
- `03_convergence_figs.R`: MCMC iteration convergence runs for Figures 2 and 4.
- `04_section6_glmm.R`: Poisson GLMM evaluation across INLA strategies (Table 5).
- `05_stress_tests.R`: Low-count Poisson (S1) and sparse binary (S2) stress tests (Table 7).
- `generate_figures.R`: Script generating all updated vector figures in `apppic/`.
- `.github/workflows/reproduce.yml`: Continuous integration workflow for Ubuntu x86_64 and macOS.
- `results/`: Directory containing all generated CSV result tables and seed logs.
- `logs/`: Directory containing session info logs.