## Comments to the author

*Although the general feeling is that the paper has the potential to deliver a valuable contribution to the journal, it needs more work to be publishable. In particular, please pay careful attention to the comments of the second reviewer (comparison of classifiers, discrete parameters, more inclusive of other probabilistic languages) that need to be carefully addressed.*

## Reviewer 1

*This is an exciting paper, both conceptually and in its great practical implications. I would accept the paper as is, with only minor comments; see below.  
  
Abstract: "finite sample performance" should be "performance in finite time". In statistics, "finite sample" refers to data, but here you're talking about running MCMC for a finite amount of time.*

We have changed this. *Abstract: change "whether" to "how well"*

Changed. *Abstract: change "instead using" to "instead it uses"*

Changed. *p.2, line 1: change "whether" to "how closely"*

Changed. *p.2, line 5: change "The predominant" to "one"*

*Changed.  
  
p.2, "Unfortunately, anyone who uses MCMC knows that it is full of false dawns: chains can easily become stuck in areas of parameter space, and observation over short intervals mean the sampling distribution appears converged." You can cite this paper from 1992:* [*http://www.stat.columbia.edu/~gelman/research/published/false\_sense.pdf*](http://www.stat.columbia.edu/~gelman/research/published/false_sense.pdf)

Thanks. We’ve now added a reference to this paper. *p.2, "it should not be possible to predict the chain that caused a draw." I would change this to "it should not be possible to predict which draws come from which chain." I think it will just be confusing to introduce causal language here.*

Changed. *p.2, Change "to determine which parameters were most important" to "to guess [or estimate] which parameters were most important"*

We have changed this to, “However, the ML classifier we use can straightforwardly be interrogated to estimate which parameters were most important for generating predictive accuracy.”

*p.11, right before equation (5): why is "alternative" in quotes?*

This is because we are terming this model as “the alternative model” which we refer to later on (as well as just being an alternative / other way of parameterising the Cauchy distribution). Because of this, we would like to keep the quotes.

*p.13, "omniscient": I think this should be "omnibus"*

We agree this terminology was confusing before. We have changed this paragraph to, “Comparing our measure of convergence that requires knowing the actual target distribution (quantile-R^2; in Fig. 10A), with the various heuristic measures, all show a similar pattern: as sample size increases, the various statistics tend towards convergence. The rate at which these converge differs though, and R\* (Fig. 10D) appears at least, qualitatively, most similar to quantile-R^2.” *p.13, 8 schools model: We use the notation N(theta, sigma^2) or normal(theta, sigma). If you write N(theta, sigma), this is not the same as the notation in BDA. Now I usually just write normal(theta, sigma) to be unambiguous. That's for the univariate normal. For the multivariate normal, I will still write N(theta, Sigma) or MVN(theta, Sigma). In any case, N(theta, sigma) could cause confusion, so you should use normal(theta, sigma) or N(theta, sigma^2).*

We agree – this notation was ambiguous before, and we have now changed this to what you propose throughout the manuscript. *p.16, on figure 12 I would plot iteration number on the x-axis as these are time-series plots (the top 2 graphs on figure 12). It seems weird to me to have time on the y-axis. And that would be fine to have time on the x-axis. Just make the lines of the graphs a little bit thinner and they will fit just fine.*

This was an excellent suggestion, thanks. We have changed this figure as suggested and it now looks much better. *p.17, "This is not a sleight on existing measures, more that this illustrates the complementarity of R∗ to them." I don't understand this sentence.*

We agree this was previously diverting and have changed the sentence to, “This indicates the complementarity of R\* to existing measures.”

## Reviewer 2

*General Remarks*

*The authors introduce a novel convergence diagnostic for MCMC by comparing the classification accuracy of gradient-boosted trees, predicting chain ids for drawn samples, to random guessing. I believe that the work is quite relevant but needs some further improvements.*

*Some high level remarks on the draft.*

1. *The text contains various small grammatical errors. I suggest the authors make a careful read through the paper to correct those. Specifically, quite a few sentences seem the have a missing "the".*

We appreciate that our language was not always clear and have sought to address some of our more confusing prose. The use of the definite article “the” before noun phrases is a style choice, rather than a grammatical one (see, for example, Steven Pinker’s “Sense of Style”). Because of this, we have addressed some of these cases; others, we prefer to leave them as they are.

1. *The authors sometimes use "ML" in cases in which this does not make much sense to me. An example is the use of "ML accuracy". I'm not sure what "ML accuracy" refers to if not the "accuracy", which is a term from statistical learning theory. If "ML accuracy" refers to something else, I suggest the authors elaborate on the term.*

We thank the reviewer—we were unaware of this ambiguity before and have changed these.

1. *The authors sometimes make relatively strong claims without much or any evidence. Examples for this can be found in the "detailed comments" section.*

See our line by line responses to individual comments below.

1. *I would ask the authors to have a pass through the cited works. It seem to me that some of the citations to not reference the original work. Examples for this can be found in the "detailed comments" section.*

Again, we have sought to address these and do so in our line by line response below.

*Some additional suggestions:*

1. *It is slightly disappointing that the authors did not evaluate their approach with different ML methods. I would encourage the authors to add a systematic analysis comparing the selected classifier (gradient-boosted trees) against other state of the art classification methods.*

This was a great suggestion and was certainly remiss before. In developing the method, we had, in an ad hoc fashion, trialled a number of classifiers against one another and found that GBMs did well. We did not, however, do so in a systematic fashion—we have now done so, with significant new results that ultimately lead us to recommend both GBM and random forest (RF) classifiers. Ultimately, it also resulted in us changing the paper title to “R\*: A robust MCMC convergence diagnostic with uncertainty using **decision tree classifiers**”.

In the new section “ML sensitivity”, we compared a number of classifiers across four examples taken from the rest of the text. In these, GBMs and RFs consistently had the highest classification accuracy. Because of this, we decided to recommend using both classifiers for calculating R\*: mainly because each works best for different sorts of problems. In the new section, “Comparing GBMs and RFs”, we explicitly compare the performance of each classifier on two tasks: determining lack of convergence in the joint distribution and diagnosing convergence in the tails. We do so across a range of dimensionalities. This (and the various other examples included in the manuscript where we made both GBM and RF results available) demonstrated that GBMs performed best for low dimensional examples; RFs for higher dimensional ones.

With these various systematic comparisons, the manuscript is now much stronger as a result, and we thank again the reviewer.

1. *It is unclear to me how well this approach will work with discrete parameter spaces. I would recommend the authors conduct additional experiments containing draws from a generative process involving discrete variables.*

We agreed this was lacking in the original draft and have added a section “Discrete distributions” that performs experiments using discrete distributions. We used discrete Markov chains as our generative process and evaluated how R\* could detect differences in the target distribution resultant from differences in the transition probability matrix between chains. We used two examples: one with a small state-space consisting of four levels; another, with 20 levels. In both cases, R\* was able to detect differences between chains given a sufficient sample size.

1. *I would generally recommend to use (A), (B), respectively, in the figure captions and use more descriptive captions in some cases.*

We have changed those cases to use alphabetic labels where it made sense to.

*Detailed Comments: Introduction*

1. *The authors mentioned a few selected probabilistic programming languages that implement a dynamic HMC variant. As this list is incomplete, I suggest the authors extend the listing to provide a more complete picture. Examples for missing PPLs are: Turing.jl [Ge et al. 2018], TensorFlow Probability [Dillon et al. 2017] and Pyro [Bingham et al. 2019].*

We have added these references.

1. *"(Stan in particular (Carpenter et al., 2017) is a great exemplar of this)" I would suggest to remove this suggestive mention of a particular library if not really necessary.*

We’ve softened this to, “Stan *(Carpenter et al., 2017)*, for example”.

1. *"MCMC estimates can have negligible bias with only a relatively small number of draws." -> "MCMC estimates can have A negligible bias with only a relatively small number of draws." ... Maybe elaborate when the bias can be expected to be negligible.*

This is a matter of style rather than grammar, and we’d prefer to keep the original.

1. *"easily become stuck in areas of parameter space" -> "easily become stuck in areas of THE parameter space"*

We think it’s clear that the reader would know which parameter space we were referring to, and so would like to keep the original.

1. *I would suggest to rephrase "Unfortunately, anyone who uses MCMC knows that it is full of false dawns: chains can easily become stuck in areas of parameter space, and observation over short intervals mean the sampling distribution appears converged." as this sentence is difficult to understand. Suggestion: "Unfortunately, anyone who uses MCMC algorithms knows that they are full of false dawns. Observing the Markov chain over a short interval may suggest that the sampling distribution is converged, while chains could quickly become stuck in areas of the parameter space."*

We appreciate the suggestion but would like to keep the original.

1. *"high posterior density is, time and again, not evidence" ... Consider removing "time and again".*

We actually like the “time and again” bit! It adds emphasis. It also adds a sense of weariness that we believe will chime with users of MCMC accustomed to these sorts of false dawns.

1. *There seems to be a confusion with punctuation here (and a missing "the"): "To combat this curse of hindsight, running multiple, independent chains, which have been initialised at diverse areas of parameter space is recommended" consider changing to: "To combat this curse of hindsight, running multiple independent chains, which have been initialised at diverse areas of THE parameter space, is recommended."*

Again, we suspect our styles differ, and we would prefer to keep the original.

1. *There are plenty missing "the" before the terms "parameter space". I would suggest the authors make a careful read through the manuscript to correct spelling & grammar errors.*

See above.

1. *"Recently, Stan has adopted more advanced variations on the original" is it necessary to mention Stan here? Maybe consider removing the mention and simply highlight the Detailed Comments Introduction advanced variations.*

Stan is currently the most widely used PPL. The new variations to which we refer were themselves created by developers of Stan (of which one of the authors of this paper is one, for full transparency), and, as far as we are aware, Stan was the first to implement these. As such, we don’t feel like directing our comments at Stan are misplaced.

1. *"Here, we introduce R∗, a new convergence metric." consider writing "diagnostic" instead to stay consistent with the title. Also, maybe explain what is new about the diagnostic. Currently, this sentence does not convey any information.*

We have added “diagnostic” to this sentence. (The rest of the paragraph explains how it works.)

1. *"This statistic is built on the intuition". I would suggest not to use the term "intuition" in a scientific publication. I suppose the idea behind R\* comes, in fact, from the insight that "it should not be possible to discern from a draw’s value the chain that generated it" if the chains are mixed.*

We strongly disagree. Whilst some users are comfortable in the language of mathematics, others (a majority, we’d wager), prefer intuitive reasoning. Indeed, we believe that this is some of the issue that underlies poor adoption of some statistical methods—a lack of good, non-mathematical communication.

1. *"To maximise predictive accuracy, our chosen ML classifier naturally exploits differences in the full joint distributions between chains, which means it’s sensitive to variations across the joint distribution of target model dimensions unlike most existent convergence diagnostics". As far as I can tell, this work does not propose a novel classifier, but only uses existing work. The sentence however, might suggest so. I would, therefore, suggest to rephrase the sentence.*

We think it’s clear from the rest of the work that we do not propose a new classifier: indeed (because of this review), we introduced a new section that compares existing classifiers.

1. *"For our ML classifier," again, I belief this could be confusing to the reader as this work does not propose a classifier.*

See above.

1. *I had a look at the citation "Chollet and Allaire, 2018" but couldn't find any justification for the claim that gradient-boosted trees "are known to perform well for the types of tabular data". Could the authors refer me to the respective section in the book?*

We have removed this citation since, in the reworking of the manuscript, it was not needed.

1. *"For the types of problem we test," maybe "we tested," ?*

We have changed this.

1. *"O(seconds)" I fail to understand this use of the O-notation. Could you elaborate? What is the computational complexity of the mentioned Rhat statistics in O-notation?*

We are just using this to mean that the calculations took between 0-60 seconds approximately. (It’s not a computational complexity.)

1. *"many iterations" what is many?*

We’d prefer to keep this vague as it isn’t as simple as giving a single number here.

1. *"the time taken is longer.". I'm not sure I understand? Is the mentioned the computational complexity or not? I fail to understand this sentence.*

See above.

1. *"we describe in detail the method for calculating R∗ and its uncertainty". Rephrase.*

We think the sentence is fine.

1. *"and elsewhere". Where? Please cite the respective works.*

We have changed this sentence to “across a range of scenarios”.

*Method*

1. *"sampling distribution for any dimension, θ, in the target distribution" -> The introduction of the notation is a bit unclear. It seem theta refers to the value rather than the dimension. So maybe use use "any dimension (j) in the target..." instead to stay consistent with the use in the remaining draft, see Sec. 3.4 for example. Method*

We have removed “theta” altogether.

1. *Figure 1 (caption): Maybe use (A) and (B) instead. This plots also don't show a prediction result, rather the comparison of the marginals for mixed and unmixed chains. I would also recommend to use bolder lines and explain the visualisation a bit more in detail.*

We prefer to keep the labels and the caption. We played around with the line widths and believe the current are best. We think the visualisations are adequately explained in the main text.

1. *Figure 2 (caption): Again, I don't think it is accurate to say "prediction" here, as this is simply a visualisation of the joint rather than the marginals. Also please use (A) and (B) or similar instead. I would also ask to add more descriptive text to the caption as this visualisation effectively illustrates the key insight of the paper.*

We describe the visualisations in the main text, so would like to keep this as is.

1. *"we train a supervised machine learning (ML) model" -> ML has already been introduced in the introduction. Please remove.*

Good spot, thank you. Removed.

1. *"whether classification accuracy" -> "whether THE classification accuracy"*

Again, we’d like to keep the original.

1. *"above the “null” case," -> This is a bit cryptic. I think what the authors mean is that the accuracy should be above random guessing. Could you rephrase or elaborate.*

The second half of the sentence elaborates, so we’d like to keep the original.

1. *"of ML accuracy" -> There is no such thing as ML accuracy. I suppose the authors mean "of the accuracy".*

We have changed this.

1. *"null accuracy," -> Again, unclear what that means.*

See above.

1. *"gives a recipe for calculating" -> I'm not sure what a recipe for calculating something should be. Maybe rephrase to "provides a pseudo-code implementation of...".*

An algorithm is just a recipe. We’d like to keep the original.

1. *"The ML classifier" -> I'm not sure it is necessary to always refer to the classifier as a ML classifier. Maybe consider dropping the ML in the remainder of the draft.*

We think it’s fine to keep it as is.

1. *"which experience has dictated to be a highly predictive framework for use in tabular data" Again I could not find any details about this claim in the book. Could you please refer me to the respective page?*

See above.

1. *"each chain has dimensions: X ∈ RS × RK , where S is the number of ..." Is X really the dimension? What is X? This needs some more careful introduction of the notation.*

We think it’s clear that X is the chain itself and the dimensions are given afterwards.

1. *"70% of draws for training and 30% for testing" -> Why was 70/30 chosen?*

This or an 80/20 split are quite standard in applied problems, so don’t think any further justification is needed.

1. *"taking O(seconds) on a desktop computer" I fail to understand what the O-notation of seconds is. Why not denote the O-notation of gradient-boosted trees? I refer the authors to [Ke et al. 2017] for details. Or maybe I misunderstand what the authors refer to.*

See above.

1. *"This shows that as the number of samples increased, variation in R∗ declined." This is to be expected as the generalisation error of the classifier will decrease.*

Yes. It’s that and sampling variation in forming the test sets. We have added, “…larger sample sizes should lead to higher R\* **values with lower variance**”.

1. *"more training data leads to better ML models." I don't understand this sentence. Maybe rephrase to "more training data can lead to a lower generalisation error of the classifier." Also note that increased training data does not necessarily imply that the performance of the classifier will increase.*

We agree that our language was ambiguous before and have changed this to, “since more training data leads to **ML models with lower generalisation error**”. Thanks to the reviewer for this suggestion.

1. *"randomness in R∗ calculation" -> "randomness in the calculation of R∗"*

We’d prefer to keep ours.

1. *"about convergence being drawn" I don't understand this. Please rephrase.*

“different conclusions about convergence being drawn” makes sense to us.

1. *"The computational cost of doing this may, of course, be unreasonable." Explain what unreasonable means and why it is "unreasonable" or remove this sentence.*

Again, we think this is an issue of style. Unreasonable for us (and for most users) just means a long time. We think this implied meaning is clear, so would like to keep the original.

1. *"variation in R∗ comes from sampling from the probability simplex" Are those class probabilities callibrated?*

We didn’t go through a separate step in our algorithm of calibrating the classifier. Future work could do so but would add to the complexity of the method and add another degree of freedom that would need to be standardised between users of the statistic. As such, we’d like to avoid discussion of this in the manuscript.

1. *"shown in Figure 3B was generated by repeatedly recomputing R∗ using Algorithm 1 " - > Maybe consider using the same scale and x-axis for both plots so that they are easily comparable.*

We changed this: now looks much better, thanks.

1. *"fixed number of chains: four." -> Remove "four".*

We think it’s better keeping as is, since this reinforces what we did before and distinguishes it with the numbers we next consider.

1. *"classification becomes a harder problem when there are more categories," Is this true? Could you please elaborate and provide a reference discussion on the complexity of classification problems with large number of classes.*

You’re right—we were speculating before. We have thus lessened the strength of our statement to be, “Since classification may become a harder problem…”

1. *"where all chains bar one have" -> "where all chains BUT one have"*

“Bar” and “but” mean the same thing, and we prefer the former.

1. *"because it is harder to classify chains when there are more of them." Again, I'm not convinced this is actually true. In fact there is theoretical evidence that the opposite might be the case in some situations. See for example: [Abramovich & Pensky 2019].*

We have lessened this claim to “this may be because”.

1. *"The performance of GBM, like all ML methods, depends on its hyperparameters." This is not entirely true as there are hyperparameter free methods.*

We have changed this part of the manuscript (and added a section on hyperparameter sensitivities to the supplementary), and this sentence has been lost.

1. *"We use Stan’s NUTS algorithm" Maybe consider adding a reference here.*

Done.

1. *"he latter thinned by a factor of 5" Why was this necessary?*

It wasn’t necessary but speeded up the computation. Since we explicitly state our thinning, we think it’s fine to keep it as is.

1. *"remains stubbornly shifted" What does "stubbornly shifted" mean? I suggest you drop the "stubbornly" here.*

We think it is fine as is.

1. *"Rather than run the MCMC" -> "Rather than RUNNING the MCMC"*

We think it is fine as is.

1. *General remark to section 3.2.2. It is difficult to assess the experiment without details about the configuration of the NUTS sampler used. I would suggest the authors provide a more complete description of the experimental setup.*

The default settings were used for the algorithm. (All the examples are backed up by code in the Github repository that we point to in the introduction for the interested reader.)

1. *"and this allows us" -> "which allows us"*

Yes, that is neater—changed. Thank you.

1. *"eq." -> Uppercase*

We use lowercase eq. throughout the text.

1. *"Fig. 8C variable importance" -> "Fig. 8C THE variable importance"*

See above discussion of the definite article.

1. *"high values mean a variable is more important" Please elaborate how feature importance is measured in gradient-boosted trees.*

We’ve provided references, so don’t think this discussion is needed in our manuscript.

1. *"this illustrates that variable importance provides information complementary to R and ESS" Why does it imply this?*

Because there is variation about a straight line meaning that you can’t perfectly map from one measure to another.

1. *"Since the R∗ distribution indicated non-convergence for both parameterisations, we ran each model for sixty-times as long, although thinned by a factor of 3, resulting in 10,000 post-warm-up iterations across each of 4 chains." Please rephrase this sentence.*

We think this is clear.

1. *"remains stubbornly " What does this mean. How is stubborn defined? Please consider removing this.*

Stubborn is defined from the OED as, “having or showing dogged determination not to change one's attitude or position”. As such, we think the sentence makes sense.

1. *"One measure of distributional “closeness” is the KL-divergence, which, in this case, could be used to measure the divergence from target to sampling distribution: if the target distribution is known, fitting a kernel density estimator (KDE) to samples allows an approximate (typically univariate) measure of KL-divergence to be calculated for each dimension. " This is a very long and complicated sentence. Consider rephrasing it.*

We think this is fine.

1. *"fitting a kernel density estimator (KDE) to samples allows an approximate (typically univariate) measure of KL-divergence" Why is this the case if the target distribution is not tractable? If it would be tractable and exists in close-form, why not use it directly?*

Because, whereas the target is tractable, the sampling distribution isn’t.

1. *I appreciate the excursion in 3.3.1 describing why certain measures will fail for heavy- tailed distributions.*
2. *"after c.10,000 iterations" What does the c. stand for?*

Circa—and c. is a common abbreviation of this.

1. *"After c.550 iterations," Again, this notation has not been introduced. What is c.550?*

See above.

1. *"The model can be parameterised in two ways, as described in Vehtari et al. (2020)" I believe those parameterisation have been introduced much earlier. Please cite the original work introducing a reparameterization of the 8-schools model instead.*

We have added a citation to what (we believe at least) is the first use of such a parameterisation:

Van Dyk, D. and X. Meng (2001). The art of data augmentation. Journal of Computational and Graphical Statistics 10 (1),

1–50.

1. *"for the NUTS algorithm (Betancourt, 2017)," Please use the correct citation for NUTS here [Hoffman & Gelman 2014].*

We have changed the sentence so that the citation is now correct (Stan actually implements a dynamic HMC variant).

1. *"To probe the predictive power of the ML classifier, we investigated how predictive accuracy varies across parameter space." -> "To probe the predictive power of the classifier, we investigated how THE predictive accuracy varies across THE parameter space."*

See above discussion of the definite article.

1. *"we group MCMC draws into deciles and draw from the R∗ distribution for each decile." Are you using the whole data set here or only the test set? This is unclear. If it is the test set, how does this relate to the predictive capacities of the classifier?*

This was ambiguous before: thank you for pointing this out. We have changed this to, “…we group MCMC draws in the test set…”.

1. *"eduction in ML classification accuracy" Again, the is not such thing as a "ML classification accuracy". I suppose the authors refer to "classification accuracy", which is a term from statistical learning theory, if so please rephrase.*

We have changed this.

1. *"this was to test that ML classification" Again, there is not such thing as "ML classification". Please rephrase.*

We have changed this.

1. *"didn’t become prone to overfitting in this limit." Maybe instead: "the classifier did not overfitting in this scenario."*

Thank you, but we prefer ours.

1. *"sampling was done using Stan’s NUTS algorithm." Please describe the hyperparameters used.*

This is a short section summarising the results of these addition experiments, so prefer not to mention these specifics here, since they can be found in the supplementary materials or Github repo.

*Discussion*

1. *"we used supervised machine learning (ML) classifiers to quantify" Only one method was used, consider rephrasing.*

We now use multiple classifier types, so think it’s fine.

1. *"a null model (which predicts a chain’s identity uniformly at random)" This is the first time the "null" model is properly described. Maybe consider moving this description into the methods section.*

The null model is defined in the first place it appears (in the Methods section).

1. *"extracting ML-predicted chain" There is no such things as a "ML-predicted chain". Please rephrase.*

We have changed this to “classifier-predicted”.

1. *"includes this information in building a model" I believe the authors refer to "train a model" or "learn a model" not "build a model". Consider rephrasing.*

We think building and training and synonyms in the usage here.

1. *"because the classification boundaries for this task are unlikely to be very complex compared to (say) machine vision tasks. " Why would this be the case?*

We have removed this statement.

1. *"target (RUser4512, 2018)." Please replace this citation with a proper citation of a scientific work, such as [Abramovich & Pensky].*

We have updated the sentence (with a scientific reference for the time complexity of random forests).

1. *"If so, this suggests that larger models" Please rephrase. Discussion*

We have changed this to “larger statistical models”.

1. *"larger models (usually needing more MCMC iterations) may currently be beyond the reach of R∗." I suggest the authors briefly discuss alternative approaches, which would allow the application of R∗ in such scenarios. Examples for this could the the use of deep learning techniques or data reduction techniques.*

We discuss some ideas to allow calculation of R\* in these circumstances below (one of which involves data reduction).

1. *"our ML calculation method" What is a ML calculation method?*

We have changed this to “non-parametric method”.