1 Reviewer 1 Comments

1. The authors propose a broad class of the proposal density, but didnt discuss the choice and tuning within the class. Tuning the proposal is crucial to the mixing and efficiency of the MH algorithm, epecially in the high-dimension space, n ? p-dimensional, that the paper is dealing with. There are two difficulties in tuning within the proposed class. First, since the proposal density is a complicated transformation of p(z?) in Rn?p?1, even if p(z?) belongs to a standard parametric family, p(yp—?) does not and it is not straightforward to assess the properties that are usually used in tuning MCMC algorithm, e.g. mode and tails behaviour. Second, the proposed algorithm (11) is restrictive because it only works for independent proposal, not the random walk proposal. Random walk is more able to explore a non-standard parameter space like the manifold here, while it seems tricky to design the independent proposal that well covers the probability mass of the target density, which again due to the transformation.

2. The simulation studies didnt report the choice of p(z?), acceptance rates of the MH, and mixing of the overall MCMC algorithm. Due to point 1, I think these issues would be of interest to readers.

3. The models studied in both the simulation and real data examples are just simple linear regression which is not realistic. Multiple linear regression with at least three independent variables should be studied.

4. What are the benefits of the proposed method over the classical robust point estimators, like Huber or Tukey estimators compared in the examples, to offset the additional compu- tational cost and tuning efforts? From the simulation studies, whether it has improvement or not depends on the prior distribution, which seems to be from comparing Bayesian and Frequentist methods. Actually, literatures on the asymptotic properties of f(?—T(yobs)) shows that, if T (yobs) is asymptotically unbiased, T (yobs) and E(? — T (yobs)) have the same asymptotic variance. Hence if the prior doesnt matter, both methods have similar performance.

5. I think f(?—T(yobs)) might perform better in the following scenario: when T(yobs) is asymptotically biased, E(?—T(yobs)) can still be asymptotically unbiased as long as the true parameter is identifiable conditioning on the summary statistic, i.e the value of E(T(yobs)) is unique on the true parameter. One possible example is the robust ridge-regression estimator, but it needs further investigation whether this estimator satisfies C1-C8.

6. In page 14, should C?(X) n ? p-dimensional instead of n ? p ? 1-dimensional?

2 Reviewer 2 Comments

1. Is the proposed framework useful for practitioners? The proposed framework targets scenario where (i) we know the data is a mixture of good and bad data, and (ii) we only want to build models that offer good prediction for the good portion of the data (from the 1st paragraph of Sec 5). Authors should provide real examples in which (i) and (ii) hold. In any real examples, (i) often holds, but (ii) doesnt: if we already have had a contaminated training sample, why would we expect the test sample to contain only the good portion? In many real applications, learning the heterogeneity of the good and bad samples is precisely the goal of statistical data

analysis. For the insurance example analyzed in Sec 6, the authors should provide ev- idence, e.g., literature from actural science or white papers from insurance industry, to justify why it makes sense to assume (i) and (ii) hold.

2. As a methodological paper, whats the guideline/recommendation given by the authors? After reading this paper, I dont know how to apply the suggested framework on a simple regression model. Apparently, the choice of the likelihood func- tion (for the good data) f(—?) and the choice of the statistic T are related. To make things more complicated, each has multiple choices. For example, f could be normal, student-t or other heavy-tailed distribution; T could be something named Huber or Tukey although the authors are not even bothered to provide any mathematical expression.

   Instead of just reporting numerical performance, the authors should provide some guide- lines/recommendation on how to apply their framework on linear regression models, e.g., what are the default choices for f and T ? Is there any- way to select which f or T to use based on cross-validation or other empirical methods?

3. Whats the real benefit of this computationally expensive approximation framework? The proposed framework replies on a good summary statistic T(yobs), which has already provided a robust estimate of the target parameter. On the other hand, the proposed MCMC algorithm is computationally expensive. In Sec 4.2, the authors only discussed the computation cost for their proposal dis- tribution, but ignored the computation cost for obtaining T(y) (see Theorem 4.1). Those statistics are M-estimators, i.e., they are maximizers of some objective functions involving n observations and p parameters. The authors should discuss the computation cost for T(y). Whats the gain of all the extra computation? The authors compare KL diver- gence, but 1) few real applications care about KL divergence not mentioning the predictive density f(y?—M,yobs) is not available in closed-form; 2) since f(y?—M,yobs) is not available close-form, the KL divergence is approximated and its not clear to me how accurate the approximation is. The authors should compare prediction and estimation accuracy, common met- rics used for regression models. The simulation setup is too simple. Suggest to add some linear regression models with large p and/or large n.

2.1 Some minor issues.

1. Abstract ... handling data sets with outliers and dealing with model mis- specification. We outline the drawbacks ... and propose a new method as an alternative. The proposed method cannot handle model misspecification, instead it assumes f is the true model (no misspecifica- tion) and whats wrong is part of the data.

2. Sec 2 and Sec 3 can be shortened or merged.

3. Eq(2) in Sec 2.1, the notation f(yi—?,c(yi)) is not introduced yet; whats introduced is f(y—?).

4. Sec 4.1, C3 and C4: a maximizer may not be a continuous function of the data and it may not be unique. Instead of saying Many estimators satisfy the above properties, the authors should list those estimators and explain why those conditions are satisfied.

5. Bottom of p14, Sample z? from a distribution with known density on S. Any distribution? For example, can it be a point mass or its support has to be equal to S? Whats the sampling distribution for z? used in the simulation studies and real data analysis?

6. Eq (18): change dy to dy?

3 AE Comments

This paper proposes a Bayesian approach for making inference based on robust statistics of the data instead of the original observations. The conditioning robust statistics pass the desirable robustness (to outliers) to the posterior distribution of parameters, thus having the potential to improve inference and prediction. This paper addresses an important problem in statistics and contains interesting ideas. However, there are some major concerns. The paper focuses on outliers. Although outliers automatically imply that the model is misspecified, model misspecification is much broader including the misspecification of the density of the good data. The paper does not appear to address model misspecification beyond the case when outliers are present. Please revise the scope of the paper as appropriate or provide more examples to ensure outliers and model misspecification are parallel contributions rather than one nested in another. There are many implementation details that an interested user would like to learn more but feel difficult to find from the current paper. This includes but not limited to the selection of proposal, parameter tuning, recommendations when a practitioner is being faced with a real-world problem, and computational complexity of the entire procedure. See the two referee reports for more details. The authors are also suggested to possibly provide code that is available online with recommended choices as default. It is unclear how the proposed method outperforms existing work, either conceptually or practically. Referees have provided some competing methods for the authors to consider. I'd like to add another existing strategy in addition to the three solutions mentioned in the bottom paragraph of page 2: Bayesian fractional inference, which uses a fractional likelihood function that raises a usual likelihood function to a fractional power. What is the advantage of the proposed restrictive likelihood approach over Bayesian fractional inference, even conceptually? Overall, the paper deserves publication after careful and thorough revision. I hope the authors can address all concerns raised in this report and the two well-grounded referee reports.

4 Editor Comments

Report on Bayesian restricted likelihoods: conditioning on in su cient statistics in Bayesian regression The paper has received a mixed reaction from the two referees and the AE. I have a similar reaction to the paper but agree with the other readers that the content is such that an opportunity to address the issues raised is appropriate. The AEs comments place the paper on the border between Reject with Resubmission and Major Revision. I mention this because it isnt clear that a revision will be successful as there are some significant criticisms in the reviews. My concerns might be a little different than the others and perhaps are less technical in nature. It is generally agreed that *all* models are wrong. The incorrectness of the model can arise in a number of ways and some observations being outliers is one of these. The natural question then is: how are we supposed to deal with that? The answer is surely that we don't unless the discrepancy is so substantial that the inferences would be seriously in error if we proceeded using the assumed model. This part of

a statistical analysis is the model checking aspect and the solution to any issue, whatever it might be, arises there. For example, for the problem being considered in this paper, I would want a model checking procedure that indicated that there is a serious problem because, no matter what distribution was used from the model, some observations are outlying and I would want the methodology to identify the observations in question. In that case there would several ways to deal with the issue, including modifying the model, but also sim- ply discarding the oending observations as part of "data cleaning". It is worth noting though that the answer isn't simple because notable scientific chievements have been obtained by looking carefully at observations that are clearly discrepant. So there are some concerns that have relevance for me and that I think the paper needs to address. What method is used to identify that there is a problem with the model such that the inferences will be strongly affected and does it identify outlying observations? A minor issue is that there may be no need to do modify the analysis but the major issue is that it introduces an arbitrariness into the analysis based on the need to chooseT: The choice of T is clearly subjective and so it needs to be subjected to the same critical analysis that we would apply to the model itself and for that matter, the prior too, and it isnt clear how to do this. I understand that there are problems where reducing the data to some T (y) is necessary, perhaps because of computational problems associated with evaluating a likelihood, but this is clearly a compromised analysis and not one we would recommend unless forced to do so. So I don't agree with the statement made in the paper "that deliberate choice of an insufficient statistic T (y) guided by targeted inference is sound practice". There needs to be a much stronger argument for this at least for me. The paper is well-written and thought-provoking so my hope is that a revision will be able to address the points raised.