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 R^{n-p-1}, even if $p(z^*)$ belongs to a standard parametric family, $p(z^*)$ 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.

   This is an astute observation - we have added the following text after Theorem 3.6:

   'The proposal is governed by the choice of $p(z^*)$ and a poor choice could cause concern about the efficiency of the convergence of the MCMC algorithm. For all the examples in the paper we defined $p(z^*)$ to simply be the uniform distribution on $\mathbb{S}$. The advantage of this choice is that it requires no further tuning parameters and we have noticed good mixing in terms of the ability of the chain to generate new data $\mathbf{y}$ that is accepted with reasonable probabilities. To implement in practice, we simply generate an n-dimensional independent standard normal $\mathbf{y}^*$ for the proposal and transform this via $h(\cdot)$. Theoretically, the random normal vector would be projected onto $\mathcal{C}^{\perp}(X)$ and scaled to unit norm to generate the uniform on $\mathbb{S}$. Using simple algebra and conditions **??-??**, one can show $h(\cdot)$ is invariant to this projection and scaling. Another option for the proposal suggested by a reviewer that the authors have yet to study is generating a random walk. As we are proposing values on a complex manifold, it might be possible to implement this by conducting the random walk on $\mathbf{y}^*$ before transforming via $h(\cdot)$. This could provide some advantages in some situations, though we have yet to run into any serious issues with convergence using the independent proposal we utilize here.'

   We do agree - that there could be cases where our choice is not optimal and more research would be need to explore this. Given the we are generating on a manifold, it's not entirely clear the best way to conduct an random walk - we offered a potential solution that would take additional effort to explore

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.

   Thanks. We report the choice of $p(z^*)$ above and we have added acceptance rates for the simulations.

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.

   For the real data, we have included two additional covariates related to the number of employees at each agency. These were left out originally after some EDA revealed their effect sizes were small, but leaving them in is a good study as well. We have kept the first simulation as we feel it

is a good and simple illustration of the ability to fit hierarchical models with our method. The real data examples now contains a hierarchical example with more covariates. Additionally, we have included a second simulation where the fitting model is a linear regression with many covariates and only 3 of them are active. Additionally there is some correlation between some of the covariates - making it a more realistic variable selection-type problem.

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 computational 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 $f(\theta|T(yobs))$ shows that, if $T(yobs)$ is asymptotically unbiased, $T(yobs)$ and $E(\theta|T(yobs))$ have the same asymptotic variance. Hence if the prior doesnt matter, both methods have similar performance.

   We have added the following to the introduction of the paper to address this comment:

   The advantages and disadvantages of the method are detailed throughout the paper using simulated data and real data. One conceptual advantage is that inferences and predictions are less sensitive to features of the data not captured by the conditioning statistics. Choosing statistics targeting main features of interest allows for more targeted inference on these features. The analysis can help to better understand other features, such as outliers, not capture by the conditioning statistics. The examples in the paper suggest advantages in situations where outliers are concern and there is significant prior information for the non-outlying portion of the model that is not outweighed by the data. The main disadvantage over traditional robust estimate techniques are mainly computational. In Section **??** we detail a data-augmented MCMC algorithm to fit the models proposed in this paper. This requires an additional computational step for each iteration of the chain. Details of the additional burden are given and one must weigh the advantages of these methods with this additional burden. Since it is typically that the restricted likelihood posterior converges to the same expected value as the conditioning statistics as the sample size grows, it may not be practical to implement our method in all situations.

   We have also included in the discussion:

   'We have found the additional computational burden to be beneficial in situation where substantial prior information that will impact the results is available.'

5. I think $f(\theta|T(yobs))$ might perform better in the following scenario: when $T(yobs)$ is asymptotically biased, $E(\theta|T(yobs))$ 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.

   Thank you for this is an interesting idea. We point out here, that the example you gave does not satisfy all of C1-C8. For example, C5 is not satisfied with ridge regression $\hat{\beta} = (X'X + \lambda I)^{-1}X'y$ so that $(X'X + \lambda I)^{-1}X'(y + Xv) \neq (X'X + \lambda I)^{-1}X'y + v$. Thus, new computational capabilities would have to be developed to explore this idea in substantial examples and we feel that would be worthy of another paper.

   Bayesian priors offer a level of shrinkage like many of the bias estimators that would come to mind. So we are unsure how the effect of shrinking using both the prior and the conditional

statistics would manifest itself. Again - a reason for further research.

6. In page 14, should $C^{\perp}(X)$ n - p-dimensional instead of n - p - 1-dimensional? Thanks for the double check - $C^{\perp}(X)$ is indeed $n-p$ dimensional, but we don't see where we mentioned otherwise on page 14. $\mathcal{A}$ is $n-p-1$ dimensional as one degree of freedom is lost for each coefficient estimate and the scale estimate.

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 evidence, e.g., literature from actural science or white papers from insurance industry, to justify why it makes sense to assume (i) and (ii) hold.

   We agree that understanding the heterogeneity in the data is often of concern and this is actually a reason to use robust regression methods. Our argument is that it is not always possible to build a complete model for both 'good' and 'bad' portions of the data. As we stated in our abstract - the method is intended to reduce 'the sensitivity of the analysis to features of the data not captured by the conditioning statistics' (which, in a broad sense is the main argument for robust methods). For example, if outliers govern the inference, they will be less likely to be identified; the proposed model will treat them as 'good' data and inferences will be distorted as a result. Comparing distorted predictions to new data will make it less likely to identify new outliers. So - in this sense, it seems obvious that (ii) is indeed often the case. If we can't (as we argue) build a perfect model for the outlying generating process, we will be better prepared to identify current and future outliers only if we can predict the good portion of the data well.

   We have added the following to the introduction to address part of this comment: The advantages and disadvantages of the method are detailed throughout the paper using simulated data and real data. One conceptual advantage is that inferences and predictions are less sensitive to features of the data not captured by the conditioning statistics. Choosing statistics targeting main features of interest allows for more targeted inference on these features. The analysis can help to better understand other features, such as outliers, not capture by the conditioning statistics.

   For the real data example: The goal for the insurance example was derived from personal collaborations and we do not know of specific literature. However we have provided more detail of the goal. It is of concern to the company to predict closures and future performance for agencies that remain open. It is important for planning purposes that the predictions are not overly influenced by a handful of over/underperforming agencies. Our analysis focuses on one aspect of the business problem - the prediction of future performance for agencies, given they remain open.

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 function (for the good data) $f(y|\theta)$ 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 guidelines/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?

   Thanks for pointing out this need. We have taken the following steps:

   (a) Point out that M-Estimators have been our default choice in the very first illustrative example: Details of these estimators can be found in many places, including (**?**). We return to the two M-estimators throughout this paper as we have found them to offer good default choices for practitioners dealing with outliers. A short review of these estimators is provided in the Supplementary Material.

   (b) Included a review of M-Estimators in the Supplement

   (c) Included a 'Practical Considerations for Using the Restricted Likelihood' to the Supplement.

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 distribution, 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?

   At the end of section 4 we have included a paragraph:

   ' Finally, it is clear the estimators themselves must be computed for every iteration of the Markov Chain. We have found this burden to be marginal in relation with respect to computing the needed Jacobian. In the simulations and real data analyses presented below, we will see that the additional burden is often of added value over traditional robust regression when substantial prior information is available that is not swamped by current data. ' Additionally we have included a paragraph in the introduction discussing advantages and disadvantages.

4. 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.

We have changed to the metric to address the estimation concern you bring up. The metric is $\log(\hat{\sigma}^2) - \log(\sigma^2) + \frac{1}{2\hat{\sigma}^2}(\sigma^2 + (\theta - \hat{\theta}_i)^2$ which combines estimation error for both the θ_i and σ. While we couldn't compute this metric in a real analysis since we wouldn't know the θ_i and σ, we feel it is best to make this direct comparison since we know them in the simulation.

5. The simulation setup is too simple. Suggest to add some linear regression models with large p and/or large n.

   We feel the first hierarchical simulation example is a good introduction to how the method is conducted in a hierarchical setting. However, we have added 2 additional regression covariates to the real data analysis which also contains a hierarchical example. We have also included an additional simulation studying variable selection where only a few variables are active out of many and hope this satisfies your suggestion.

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 misspecification) and whats wrong is part of the data.

   We are taking a different view: the 'true' model would also have the correct model for the outliers or any other type of misspecification. f is only a model for part of the overall data generating mechanism - hence, even it were exactly correct for this part, we would still have model misspecification

2. Sec 2 and Sec 3 can be shortened or merged. We have merged the sections.

3. Eq(2) in Sec 2.1, the notation f(yi—?,c(yi)) is not introduced yet; whats introduced is f(y—?). We have included a sentence for this notation after it's first use.

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.

   We have added:

   ' These M-estimators satisfy C3 and C4 since they are optimizers of (almost surely) continuous and differentiable objective functions. Constraints C5-C8 are often satisfied by location and scale estimators but should be checked on a case by case basis. '

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?

   Thanks for noticing this detail. It should have full support on S since the transformation is 1-1 and onto A. We have now mentioned this. We have also included in discussion of what we used for p throughout based on a previous comment. This comment is after Theorem 4.6.

6. Eq (18): change dy to d$\tilde{y}$? Thanks for noticing this typo - based on your other comment we are using a different metric then KL divergence and so this equation is no longer in the paper.

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.

HOW TO ADDRESS?

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.

Thank you - we believe we have addressed each of these concerns based on the the 2 reviewer's comments. Detailed responses are given above

The authors are also suggested to possibly provide code that is available online with recommended choices as default.

We have included a default choices in the paper and in the Supplement. Information for where to obtain code on Github (an R package as well as data and code for the examples) is now provided in the Introduction

It is unclear how the proposed method outperforms existing work, either conceptually or practically.

we believe we have addressed similar concerns from the two reviewers - specifically we have included in several areas the advantages and disadvantages (e.g., see Introduction) we have found as well as practical recommendations (e.g. See Supplement)

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?

HOW TO ADDRESS?

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

unknown response at this point ???

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 AE's comments place the paper on the border

between Reject with Resubmission and Major Revision. I mention this because it isn't 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 simply discarding the offending observations as part of "data cleaning". It is worth noting though that the answer isn't simple because notable scientific achievements 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 choose T: 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 isn't 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.