1 Reviewer 1 Comments

1. The authors propose a broad class of the proposal density, but didn't 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 lead to an inefficient MCMC algorithm. For all examples in this paper we defined $p(z^*)$ to be the uniform distribution on $\mathbb{S}$. The advantage of this choice is that it requires no further tuning parameters. We have noticed good mixing in terms of the ability of the chain to generate new data $\mathbf{y}$ that is accepted with a reasonable probability. To implement the method in practice, we 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 that $h(\cdot)$ 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 independence proposal we utilize here.'

   There are undoubtedly cases where our choice is not optimal. Here, we generate proposals for the complete data on a manifold that (in our examples and typically) is of much higher dimension than the conditioning statistic. The independence proposals have reasonably large acceptance probabilities. Random walk proposals are generally used to ensure that proposals are "near" the current state of the Markov chain, reducing proposed movement in exchange for a greater acceptance probability. Our take on our algorithm is that we offer one solution that has provided effective MCMC in the cases we have explored. The method applies to a large class of models that are in common use and to a large class of conditioning statistics. We note that the independence proposals for the complete data are but one part of a larger algorithm. Given the complete data, generation of the regression parameters may or may not make use of random walk proposals, block proposals, Gibbs steps involving generation from full conditionals, or other techniques.

   With the basic paradigm and one implementation in place, we look forward to many future developments. Further research is needed to produce these developments. For computation, the development of random walk proposals on (nonlinear) manifolds is an interesting problem.

2. The simulation studies didn't 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 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 contain a hierarchical example with more covariates. Additionally, we have included a second simulation where the ground truth is a linear regression model with 3 active covariates. A model with 30 covariates (27 inactive) is fit. There is correlation amongst some of the covariates, making this problem more realistic.

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 doesn't matter, both methods have similar performance.

   As you note, one major advantage is that our method is *exactly* a Bayesian method, with $T(y_{obs})$ used in Bayes Theorem to move from prior to posterior. In a typical setting, under relatively mild conditions on the prior distribution and estimator, Bayesian and frequentist methods will agree asymptotically. This is reflected in the large-sample central limit theorems (Bayesian and frequentist versions) and is not specially connected to our work. In this paper, our interest is not in these asymptotic settings where we could simply use a limiting normal distribution as the posterior over the parameters, but rather for the finite sample setting where Bayesian and frequentist methods yield different answers. The simulations and document this differential performance. The standard advantages of Bayes apply – a guarantee of admissibility from a decision-theoretic view; the ability to combine information across a collection of problems (e.g., the hierarchical setting); and the transition from moment-based methods to likelihood-based methods, for example. The weaknesses are also there – a really bad prior distribution produces poor finite-sample performance. There is a greater computational cost (standard for Bayes vs frequentist), but, given our MCMC routine, no real effort is needed to tune the computation.

   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 and real data. One conceptual advantage of our method is that inferences and predictions are less sensitive to features of the data not captured by the conditioning statistics than are methods based on the complete likelihood. Choosing statistics targeting the main

features of interest allows for inference that focuses on these features. The analysis can help to better understand other features which may not be captured by the conditioning statistics, such as outliers.

The examples in the paper provide a Bayesian analog of classical robust estimators. The main disadvantage of our methods relative to the classical estimators is computational. In Section **??** we detail a data-augmentation MCMC algorithm to fit the models proposed in this paper. The advantages are those of Bayesian methods. As is standard for Bayes-classical comparisons, the Bayesian method requires greater computational effort while providing better inference. As a referee notes, asymptotically, the Bayesian and classical parameter estimates are often very close and have the same limiting posterior variance / sampling variance. In situations where asymptotic approximation suffices, there is no need to use the computational techniques developed in this paper.'

We have also included in the discussion:

'We have found the benefits of using our Bayesian technique to outweigh the additional computational burden (relative to a classical estimator) in the situation where substantive 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.

   This is idea is in perfect alignment with our work. You are right that $T(yobs)$ need not be unbiased – in fact, it need not even be an estimator (think of sums of squares in an ANOVA setting as opposed to mean squares). It is merely a statistic upon which we condition when using Bayes Theorem.

   Our computational strategy provides an implementation for a big variety of conditioning statistics that are also estimators. The ridge regression estimator does not satisfy property C5 and so, on the surface, would need a new computational development ($\hat{\beta} = (X'X + \lambda I)^{-1}X'y$: $(X'X + \lambda I)^{-1}X'(y + Xv) \neq (X'X + \lambda I)^{-1}X'y + v$). But the ridge regression estimator for full rank $X'X$ is equivalent to least squares regression for conditioning. And the least squares estimator does satisfy C1-C8.

   For any hope of consistency, the models would need to distinguish between different values of θ, and so a condition along the lines of uniqueness of $E(T(yobs))$ in θ is important.

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 another degree of freedom is lost for 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) doesn't: 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.

I think we are in agreement that it is common to encounter data sets where a portion of the data is 'good' and a portion of the data is 'bad' (your point (i)). The 'bad' data may be bad for many reasons, including a breakdown of the measurement process, misrecorded data, and cases that simply aren't relevant to the problem at hand. The ability to perform a decent analysis for such data sets is important. And the typical decent analysis focuses on (ii). We would have little interest in predicting misrecorded test data in a typical situation or in predicting the recorded values when the measurement process has broken down.

Developing an understanding of sources of variation in the data is important. In settings where the mechanism that produces bad data is stable, we might even entertain building a model for the bad data as is one tradition in Bayesian statistics. But there are many instances where this is of little interest.

When we are in a situation where (i) and (ii) hold – we would argue that this is the typical situation and not the exception – and we have test data available, there would rarely be any reason to believe that the test data would be pure, consisting only of good data. This is an important motivator of this work as described in the introduction to our paper.

We have added the following to the introduction in response to your comment and a comment from the other referee.

'The advantages and disadvantages of the method are detailed throughout the paper using simulated and real data. One conceptual advantage of our method is that inferences and predictions are less sensitive to features of the data not captured by the conditioning statistics than are methods based on the complete likelihood. Choosing statistics targeting the main features of interest allows for inference that focuses on these features. The analysis can help to better understand other features which may not be captured by the conditioning statistics, such as outliers.

For the real data example: The goal for the insurance example was derived from personal collaborations with those in the insurance industry. Privacy concerns and a turnover of executives in the company from which the data come limit our ability to provide a full description and limit our ability to provide the data themselves. However we have expanded slightly on the goal of the analysis. 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, what's the guideline/recommendation given by the authors? After reading this paper, I don't 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 the wish by some to have default choices for f and T. We consider regression analysis to be so fundamental to the discipline of statsitics and to the practice of data analysis that we feel it inappropriate to specify defaults so heavy-handedly. This would be much like specifying a default likelihood if one wishes to build a regression model. As you note, choice of the likelihood is part of the analysis when building a Bayesian model. To this, we add choice of the conditioning statistic T and illustrate the impact of different choices. Omission of the formal description of the particular M-estimators we use is an oversight.

We have taken the following steps to address your comment:

(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 short review of M-Estimators in the Supplement

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

3. What's 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). What's 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 relative to computation of the needed Jacobian. In the simulations and real data analyses presented below, we will see that the additional computational expense needed to fit the Bayesian model is often worthwhile, leading to better performance compared to traditional, non-Bayesian robust regression estimators. This is most evident when substantive prior information is available and information in the data is limited.'

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 it's not clear to me how accurate the approximation is. The authors should compare prediction and estimation accuracy, common met- rics used for regression models.

   To address the concern you have raised, we have changed our evaluation metric in the simulation to a closed form version of KL divergence. We use the KL divergence between two normals: $\log(\hat{\sigma}^2) - \log(\sigma^2) + \frac{1}{\hat{\sigma}^2}(\sigma^2 + (\theta - \hat{\theta}_i)^2$ While we wouldn't be able to compute this in a real application, it does combines estimation error for both the θ_i and σ and is a decent metric to compare the methods under the simulation setup. John - I'm puzzled by this one. I'll have to check the paper when I get dropbox up and running again. (Steve)

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

   We feel that the first hierarchical simulation example shows how to use the method in a hierarchical setting. We have added 2 more 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. We 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 what's wrong is part of the data.

   We stand by the abstract. The insurance company example is along the lines of a number of examples in the literature on robust regression. Outliers – best defined as cases which do not reflect the mechanism under study – often have large residuals. This does not mean that outliers are independent of one another. They often reflect model misspecification. The tradition of robust regression is to provide a model that discounts the outliers (that discounts the deficiencies in the model) and results in a sensible, practical analysis of the data set. The robust analysis often points the way to an understanding of why the model fits some cases well and others poorly. Our paper is about model misspecification as well as what some feel the word 'outlier' captures.

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; what's 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 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? What's the sampling distribution for z? used in the simulation studies and real data analysis?

   Thanks for noticing this detail. It is most natural if it is mutually absolutely continuous with respect to the distribution induced by the model on S. For the cases we discuss, this is mutually absolutely continuous with respect to the uniform distribution on S. We now mention the importance of the support. We have also included a description of what we use for p throughout based on a previous comment. This comment is after Theorem 4.6. John - a little puzzled by what was intended here. (Steve)

6. Eq (18): change dy to d$\tilde{y}$?

   Thanks for noticing this typo - based on your other comment on KL divergence, this formula was taken out and now is a closed form version.

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.

Model misspecification covers broad territory. It includes traditional outliers (retaining the independence of observations, but assuming some have greater variance, for example), misspecification of the mean function in a regression setting, missed covariates for regression, missed dependence among cases, misspecified functional form for the 'good' data (one reason for the growth of nonparametric Bayesian methods), and much, much more. Box' famous aphorism is one direct statement that all models are misspecified.

Our own belief is that it would be impossible to address *all* forms of model misspecification, yet discussion of the issue is important. Our paper illustrates misspecification in the real data example on the insurance companies. The inclusion of closed agencies and agencies with different types of contracts represent model misspecification. These 'outliers' are not merely the result of a huge chance error, a faulty measurement process, or a failed experimental run. They are unusual

for a reason and the analysis helps to uncover this reason (we are limited in the commentary that we can attach to the analysis). As with many of the examples used to illustrate robust regression techniques, these data are best described as having a misspecified model.

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 some '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. As noted in our response to Referee 2, we hesitate to push the notion of a default likelihood and default conditioning statistic. The paper provides a conceptual framework for data analysis and develops the methodology and computation needed to implement the analysis. We do not view the work as developing an algorithm into which one should just feed data.

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?

The central advantage of our method compared to the fractional Bayes methods is that our method uses Bayes Theorem! It is exactly Bayesian in that the posterior distribution is a rescaled version of prior distribution times likelihood (of a statistic). This means that the method inherits a host of properties from Bayesian methods.

In contrast, the fractional Bayesian methods do not actually follow Bayes Theorem (there are exceptions in special cases). The use the theorem as an analogy and so miss many of the benefits of Bayes. The fractional methods are interesting and they deserve study, but they are not the only solution to Bayesian model misspecification.

In the long run, we suspect that robust and effective Bayesian analyses will make use of ideas from a variety of perspectives that are currently under study. But the various lines of research need time and effort to develop, and they need visibility for the community to appreciate the perspectives and to develop an understanding of what each brings to the table.

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.