Response to Referees' Comments on "Bayesian Restricted Likelihood Methods" by John Lewis, Steven M. MacEachern, and Yoonkyung Lee.

The authors thank the AE and the two referees for their thoughtful comments and suggestions. Our responses are listed as bullets.

Comments from the AE

Summary: This is a thought provoking paper that pursues a novel idea how to perform Bayesian inference under the assumption that only part of the data are so- called good data, while the rest are bad data or outliers in the sense that they are generated by a probability law different from the good data. I agree with both reviewers that this is an interesting issue and that the paper provides a novel and ingenious contribution in this context. However, both referees raise several concerns which I share:

1. The exposition of the material in the paper needs to be improved considerably and reading of the paper has to be made easier. One of the reviewer made very detailed suggestions how this could be achieved. Like this reviewer, I found the exposition of the MH algorithm on p. 12 particularly obscure, see also Comment 7 below.

- A substantial revision/reorganization of the entire manuscript has been done including special attention paid to the explanation of the MCMC algorithm

2. I agree that the term "restricted is misleading. How about a term such as Bayesian inference for data with incompletely specified likeilihood, or something similar?

   - In efforts to come up with a different name for the method, we feel similar concerns arise. We have used restricted likelihood for some time now and so we hope to stick with it. In an effort to reduce confusion we have given the paper a more descriptive title.

3. I found the discussion how the suggested approach is related to ABC methods very vague and would appreciate more details.

   - We have provide more details.

4. There are several references to unpublished material such as Jung et al. (2014) which makes it hard to evaluate the value added by the corresponding material in the present paper.

   -

5. As for ABC methods, the suggested approach heavily relies on choosing appropriate statistics of the data. Since robustness to "bad data is the major concern of the authors, they should pursue one reviewers advice go beyond M-estimators and consider high-break down estimators such as LMS and LTS. Also the choice of a suitable statistic for the scale is an important issue that should be discussed in more detail.

- We agree that the choice of statistic is paramount and doesn't have to be restricted to M-estimation. We included two simple examples utilizing LMS and LTS. And added dialogue of the choice of statistic to the discussion.

6. As noted by one referee, the gain in performance of the suggested approach for the two models discussed in the paper is not terribly convincing. Hence, I strongly encourage the authors to include an additional example that shows a significant edge of their method over classical ones, or at least to provide artificial data, where this is the case.

    - We have added a simulation study showing the enhance performance of our method in a situation where the data generating model is known. We have also made some refinements to the data analysis section to help highlight some of the advantages of out approach. In particular, we present results for individual state-level regressions for states of various sizes. We also streamline the presentation of the hierarchical model by including a single training sample size, as the differences before were not that meaningful with the different training sample sizes.

7. One reason for this performance could be that, strictly speaking, the model is not only misspecified for the bad part of the data, but also for the good part of the data which are count data, rather than genuinely Gaussian. In particular, if these outcomes are small counts (there is not scale on Figure 3, so it is difficult to say), the square root transformation might not be enough to ensure normality of the outcomes in regression (14) and its generalization in Section 4.3. Apart from non-Gaussianity, also the assumption of homoscedasticity of the error term might be misspecified. I wonder, how robust the whole approach is to model

misspecification also for the good part of the data, since the acceptance rate of the MH algorithm in (7) and (8) seems to be based on the (potentially misspecified) outcome regression.

- Yes, homeoscedasticity is certainly one piece contributing to model misspecification and we added this comment to the data analysis section. We have not seen problems in the MH acceptance rates across many examples and added the acceptance rates we've observed in the data analyses sections.

Comments from Referee 1

Major Points: The writing in this paper is very smooth; it reads easily. This is good but Im worried this is achieved by not providing enough detail and reasoning. Hence the bulk of my comments below are efforts to quantify what further details the authors almost certainly know and would be well advised to include.

1. The Introduction and Sec. 2 are too long and detract from the focus of the paper: The authors basically want to give (3) and two examples of it, namely (1) and (2). Then they want to explain how they generate $(y|T(y), \theta)$. The central idea is strong enough to stand on its own without the elaborate discussion. The discussion points seem more appropriate for a concluding section on the conceptual implications of using a restricted likelihood. I suggest this because the conceptual implications are not the focus of the paper, they result from the methodology. The discussion material in the early part of the paper could be moved to the last section and made more focused. There are also a good number of references that the authors might want to read and possibly cite in the general area of model uncertainty outside the across-model prior setting. In no particular order, some are: Zhu, Ibrahim, Tang 2011 BKA, Barnardo and Smith 2000 (for M-closed, complete, and open), Gustafson and Clarke 2004 CJS, Clarke 2010 BA, Draper 1996 JRSSB, Yuan and Clarke 1999 CJS, Clyde and Iverson 2013 (M-open, proceedings volume) and Clarke, Clarke and Yu 2013 (M-complete) BA.

    - We have re-organized and shortened Sections 1 and 2 and moved some of the commentary to the discussion. Several of the suggested references were added in the introduction.

2) The paper should be more self-contained: Even after several readings I do not understand the procedure on p. 12 satisfactorily. Details are left out, e.g., how (7) follows from (8); the one sentence explanation on l. 8-9 is not enough. Why is it OK to drop the conditioning on $T(y_p) = T(y_c) = T(y_{obs})$? Some notation seems odd e.g., what is the relationship amongst yp, yc, and yobs? Where is the expression p()? Or do the authors mean p(—?)? Also, I thought the idea of Metropolis-Hastings was to generate a sample from a distribution for which direct generation was difficult. If evaluation of $f(y|\theta)$ is straighforward (as stated on p. 12, l. 10) and $f(y|T(y), \theta)$ is available, why is the authors procedure important? I think the answer is that the authors want to compare $f(y|\theta)$ to the use of a $f(T(y)|\theta)$ corrected by an estimate of $f(y|T(y), \theta)$ so that one can use the authors procedure without having $f(y|\theta)$ but this should be stated explicitly. A related point is that the discussion in Sec. 3.2.1 was also hard to follow because too many details were left out. The role of z? needs to be explained better. While the words in the last two lines of p. 12 may be obvious to those who know the methodology, giving the mathematical formalities and the motivations for the various steps would make them and p. 13, 14 easier to follow. In particular, where does y? come from? Why would one want to transform y? into y as in Theorem 1? How does the expression in Theorem 1 lead to the set in (9)? I have no doubts that the authors can provide more formality and motivation for each step in the procedure; its just a matter of doing it.

- We have made substantial revisions to the paper, especially the MH algorithm in an attempt to make the details more clear including: more detail on the simplification of the the MH ratio R in (7) and (8); the fact that $f(y|T(y), \theta)$ is not available in closed form (only up to a proportionality

constant), so direct generation is difficult.

3) The intuition should be better explicated: The authors have clearly thought carefully about the intuition behind their procedure, but some how it doesnt 2 come across in Figs. 1 and 2 which I am unable to link to the formalities they do present. The authors have not yet found a good way to express their ideas so that they can be readily assimilated by some one who does not already understand. I dont understand the steps in the reasoning in Sec. 3.2.2. Its OK to relegate proofs to an Appendix, but the reasoning should be clearer in the main text.

- We agree that this section was difficult to follow. We have made substantial effort to re-organize the section and explain the steps and reasonings more clearly.

4) The results for the application need better labeling/exposition: It would be helpful if the authors could write out exactly and explicitly what cross-validation type of evaluation is being calculated. This is described at the bottom of p. 24, top of p. 25 in words, but the mathematical expressions should be given fully as well. This would clarify the role of the base method and the comment on p. 27, l. 3-4. The same sort of comment applies to p. 27-28: Please give the mathematical expressions rather than just the words. (This would make the last lines on p. 29 easier to follow.) Maybe its just the way that my copy printed, but in Figs. 3, 4, and 5 the ranges/dots are hard to distinguish from each other so its hard to tell which is what. Could the different cases have different symbols that would be easier to tell apart? Even making the figures larger would help. To make it easier for the reader could the discussion of Table 1 p. 31 refer to the rows in the table? That would make it easier for the reader to tell which

comparisons the discussion was commenting on. A related question: What does Fig. 1 show that is essential to the discussion?

- The results sections has been refined substantially. The models were simplified and more focused to elucidate some of the advantages of our method. A more complete description with mathematical expressions is provided for how the TLM is calculated. We paid more attention to the clarity of the new figures. The purpose of Figure 1 (now Figure 3) is explained better in Section on the MCMC algorithm (now section 4).

5) Gather and organize your thoughts on model misspecification in Sec. 5: This is a sort of repeat of item #1, but there are many places where the authors comment on various aspects of their analysis, the comments are valid and helpful, but they detract from the flow of the paper. If these were gathered and organized in the Discussion section the authors points would be more effectively made. For instance, the paragraph beginning In addition to... on p. 4, the first part of the paragraph beginning Further examples... on p. 6, middle paragraph on p.8, the discussion of instability in the first two paragraphs of Sec. 3.2, the paragraph beginning We digress...on p. 28. There are other cases where remarks are valid but maybe they are not in the right place and would be more effectively made in the Discussion.

- We have moved many of the remarks originally in the beginning to the discussion.

Minor Points: Here are a few specific points. Some are picky, some are more substantive, but they are all smaller than those above. If I had more time, Id probably

find more suggestions, but Im sure the authors can see the general tenor of my points is to encourage them to be clearer about the details.

1. p. 3, para. beginning The focus of this work...: State this more formally with mathematical notation.

    - We've added $f(\boldsymbol{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})$ and some commentary on how such imperfections have traditionally been dealt with.

2. p. 3, l. -2: The word restricted likelihood is OK, but if the authors could use some other word it would be better. The term restricted is already used in the sense of restricted MLE and thats not the same as the authors mean here. This would help avoid possible confusion.

    - Addressed above.

3. p. 4, l. 11: Wong and Clarke CJS 2004 shows a case where using less but well chosen data gives better prediction.

    - Thanks for the reference, it was added to the literature review.

4. p. 4, l. -8: The authors might want to have a look at Bernardo and Smith 2000 for their discussion on M-closed, -complete, -open problems. Maybe mention variance-bias decompositions for MSE?

    - We have added this reference.

5. p. 4, l. -4,-5: Yes, mis-specification and outliers somehow represent the same problem...Could this point be made with more clarity?

- The first sentence of the literature review section now reads 'Our motivation for the use of summary statistics in Bayesian inference is concern about outliers or, more generally, model misspecification. Specifically, the likelihood is not specified correctly and concentrating on using well chosen parts of the data can help improve the analysis'

6. p. 5, l. -4: term should be factor.

   - done

7. p. 6, l. 7: Alternatively should be Hence (for instance).

   - Changed

8. Keep the ordering of conditioning variables the same e.g., (3) p. 6 uses (y—T (y), ?) while p. 1, l. -2 uses (y—?, T (y)).

   - We made this change.

9. p. 23, l. -1: tunning should be tuning

   - done.

Comments from Referee 2

The manuscript develops a new methodology for using a restricted likelihood in a Bayesian context. It derives an ingenious way of generating complete data samples that match the observed data as far as the conditioning statistics is concerned. The resulting algorithm is presented in sufficient detail to be useful to the practitioner. In view of these qualities, the manuscript deserves to be published. However, in the example chosen by the authors the gain in performance as compared to classical methods is marginal at best, at the price of a presum- ably much larger computational complexity. In the simple regression model the restricted likelihood is outperformed by the heavy-tailed model for the small sample size, whereas for larger sample sizes its performance is virtually identical to the classical robust estimators. In the hierarchical model, it is indeed slightly better than the classical estimators, but I wonder whether this slight gain justifies the additional effort of finding reasonable priors for the hyperparameters and the much more involved computation of the estimates. If the priors are not chosen carefully, the results could even be worse than the ones from the classical estimators. The authors should therefore provide an additional example where the superiority of their approach shows more clearly.

- We have provided additional examples as well as revised the data analysis section to better explicate the advantages of our method.

It would also be useful to tell the reader whether the code used in the example is already publicly available, and how it compares in terms of computational load with the classical M-estimators and also with other robust estimators such as LMS (least median of squares) and LTS (least trimmed squares).

- information on acquiring code and current capabilities of it are now given in Section 4.1. Currently the MCMC algorithm is coded for the M-Estimators dis-

cussed. LMS, LTS, and others can theoretically be implemented using MCMC, but the software currently isn't generalizable enough for it. The package implements methods effective in lower dimensional settings that can be used for these estimators.

The estimator must be computed at every MCMC iteration; and an orthonormal bases must be found. These do add cost to the MCMC for our method - we have added this to the discussion.

In addition to the usual concerns about the sensitivity of the results on the prior in the Bayesian context, the method as presented here raises the problem of choosing the type of the M-estimator, to which the example data seem to be quite sensitive. It is not difficult to understand why Tukeys redescending M-estimator performs better than the Huber-type M-estimator, but maybe not so easy why the difference is much more pronounced in the hierarchical model as compared to the simple model. If the authors have an explanation for this effect it should be given in the paper. For other data sets, however, a good choice of the M-estimator could well be different, so some guidelines would be welcome to the practitioner. The authors should also discuss in more detail whether it is possible to go beyond M-estimators and to use high breakdown-point estimators such as LMS and LTS as the conditioning statistic, and which scale estimators could or should be used as companions. As the LMS regression has no tunable parameters, and the LTS regression has only a single one, the user would be relieved from studying the sensitivity to the choice of the M-estimator from an infinite number of possibilities.

- Added to the discussion some commentary/recommendations for choosing an estimator. Also, as stated above we have added examples of LMS/LTS.

Finally, I have to remark that the first sentence of the Discussion is dangerously

close to an insult to the numerous researchers and practitioners who have applied Bayesian methods in data analysis for years, even decades. The approach chosen by the authors is certainly innovative and in my opinion potentially fruitful, but it can hardly claim to begin to reconcile the two fields. I think it is fair to ask that the authors find a slightly more modest expression of their enthusiasm.

- We certainly did not intend for this statement to be interpreted this way and we have removed it. We have also updated the discussion to reflect the rest of your comments.

In summary, I recommend publication of the manuscript under the provision that the authors submit a revised version which addresses the points raised above. It should contain an additional example that shows a significant edge of their method over the classical ones. In this new example robust estimators with high breakdown point (LMS and LTS) should be studied along with suitable M-estimators. As is well known, the LMS-estimator has worse asymptotic properties than M-estimators and LTS, and it will be interesting to see whether this is visible in the results.

- We have included additional examples including a simulation. The simple example include LMS and LTS and we added some discussion on the choice of estimators.

Minor comment:

1. Figure 4 is very hard to read, the panels should have the same size as the ones in Figure 5.

    - The hard-to-read figures have been updated.