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

The comments from the referees/AE are paraphrased in *italics*

Comments from the AE

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 reviewers 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.*

   See response to all the major points from Referee 1 as well as to comment 7.

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?*

   See response to Referee 1 Minor point number 2.

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

   We can add more detail here. Including what ABC actually does and how it 'looks' like our method but with different motivation . . . as well as it's use of an observed data dependent 'statistic' which does not partition the sample space.

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 *There are several references to unpublished material such as Jung et al.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.*

   Jung et al. 2014, Lee J. and MacEachern 2014, and my thesis were unpublished when we wrote this. My thesis is openly available now on OhioLink. Should we add the website https://etd.ohiolink.edu/ to the citation? Also, are the other manuscripts available somewhere? I couldn't find them with a google search.

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.*

   Yes, robustness to "bad" is the major concern. Going beyond M-estimators is something I always thought of as a possible route to take next with this research. It would involve a detailed study of conditions C1-C8 to see if such estimators satisfy these assumptions as the validity of the sampling strategy depends on them. Otherwise a potentially new sampling strategy would need to be developed. Beyond this, my personal concern with pursuing this for this paper is that my code is only robust enough to handle the standard M-estimators

(Huber's and Tukey's at this point). A great deal more work on the coding side would be needed to attempt other estimators even if they satisfy C1-C8.

As a seminal paper I don't think our goal is really to have a comprehensive study of different conditioning statistics. I think the main goal is to present the idea and give a few examples. In my mind, a comprehensive comparison of different conditioning statistics could be its own stand alone paper. Also, Huber himself has a little discussion in the newest version of his book arguing that researchers have spent far too much time trying to achieve high-breakdown estimators (e.g. a breakdown of .5) when in reality datasets will not have that much contamination. I think this gives some merit to our choice of statistics we present here.

A suitable scale statistic is definitely important. Again, I think a comprehensive study of this could be part of another stand alone paper. With all that said, I think we could at least add a discussion on the possibility of going beyond M-estimators as Reviewer 2 suggests.

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.*

These comments on the results not being terribly convincing are echoed by referee 2. There is definitely merit to these comments. It would be nice to have another example. Though not too terribly interesting, the analysis of Newcomb's data and the Belgium calls data I have in my dissertation are two potential examples that show substantial differences in posterior distributions

between the heavy-tailed model and the restricted models. It doesn't address the comparison between the classical fits and the restricted fits though.

However, my take is this: For the standard regression model, we expect the classical and Bayesian methods to perform similarly for large sample sizes. The priors should only make a difference for small sample sizes. Indeed, for n=25 the improvement is marginal as long as the trimming fraction is large enough. For smaller trimming fractions (less than .15) the differences are more substantial. This demonstrates some robustness to the trimming fraction. Also, for n=25, our methods are outperformed by the heavy-tailed model, but our method still performs well. Perhaps this is an area where another choice of statistic (or tuning parameters) could improve the fits.

For the hierarchical regression model, we see a substantial reduction in variance of the TLM for n=1000 which I think justifies the extra effort. For n=2000, The similar performance between the classical fits to each group separately and our method reflects the substantial differences between the groups. The hierarchical model is able to capture this as well and is a demonstration of the flexibility of modeling hierarchically — even with our new method. This is a worthy observation and I think computational complexity is a secondary issue (at least from a philosophical point of view)

Steve: I know you had the idea to try a few applications. Did anything fruitful ever come out of that?

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.

The observations we centered and scaled before taking square roots. The center and scaling was done to mask the data (the reason why there are no scales on the axes) but also helps to 'normalize' the data after taking the square roots. The homoscedasticity assumption may not be justified for the simple regression model. This is not something I even considered and it is a noteworthy point. For the hierarchical model, a separate variance was assumed for each group. I am assuming this helps in terms of the assumptions, but again I did not do any diagnostics within a each group to check the assumption that a homoscedasticity held within each group.

With robust regression, observations with large residuals (large variances) are trimmed (either fully or partially). Hence assuming homoscedasticity when it is not valid could cause the method to trim 'good' data. The estimate of the mean regression line, which is the major concern here, should not be effected much by assuming homoscedasticity. This is heuristic and deserves research.

Comments from Referee 1

1. *The Introduction and Sec. 2 are too long and detract from the focus of the paper.*

   I am still wondering how to address this comment in ways other than trying to cut down some of the wording. I believe the introduction should address the following points (which are already in the paper):

   (a) The difficulty in eliciting a really good data model for certain situations

   (b) The usual approaches using mixture/heavy-tailed models and their drawbacks. This will lead to pointing to the 'restricted likelihood' as an alternative

   (c) A high level description of the new method.

   With that said, it may be appropriate to move the discussion on model misspecification (2nd to last paragraph of intro) to the concluding section. I say this because the paper indeed focuses on handling outliers and stating that the method also addresses model misspecification at the end might be a nice 'take-away' that isn't forgotten.

   I believe Section 2 should address the following points (which are already in the paper):

   (a) Give the two data analysis examples (subset of known bad data and censored data) to introduce equation (3) and the fact that our method is often used informally in practical ways.

   (b) Explicit form of $\pi(\boldsymbol{\theta}|T(y))$ as well as $f(y_{n+1}|T(\boldsymbol{y}))$ should be given

   (c) lit review on the previous uses of this method

I am unsure what the referee means by 'model-uncertainty outside the across-model prior setting.' However, I am familiar with some of the references given. Perhaps a discussion in the concluding section on the possible extensions to model averaging and it's relationship to the M-closed/complete/open concepts of Bernardo and Smith 2000 is in order... though I am not sure what this relationship is exactly at this point. I'd have to sit down with several of these papers to be see how they can be related to our situation.

2. *The paper should be more self-contained*

I can definitely see how the disposition on page 12 is confusion. Before discussing the MH step for proposing data, I think it would help the casual reader to explicitly state that we are using a data-augmented algorithm sampling from $[\boldsymbol{\theta}, \boldsymbol{y}|T(\boldsymbol{y})]$ and using Gibbs sampling to iteratively sample from $[\boldsymbol{\theta}|\boldsymbol{y}, T(\boldsymbol{y})] = [\boldsymbol{\theta}|\boldsymbol{y}]$ and $[\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y})]$. The second step, sampling from $[\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y})]$, is the difficult part and requires the MH algorithm.

Before the expression for R (or perhaps after) I think something along the lines of the following may help:

The proposed data is sampled in such a way that $T(\boldsymbol{y}_p) = T(\boldsymbol{y}_c) = T(\boldsymbol{y}_{obs})$. Further, $f(\boldsymbol{y}_p|\boldsymbol{\theta}, T(\boldsymbol{y}_p) = T(\boldsymbol{y}_{obs})) = [T(\boldsymbol{y}_p) = T(\boldsymbol{y}_{obs})|\boldsymbol{\theta}, \boldsymbol{y}_{obs}] f(\boldsymbol{y}_p|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/[\boldsymbol{\theta}, T(\boldsymbol{y}_p) = T(\boldsymbol{y}_{obs})]$. The first expression in brackets is one since $\boldsymbol{y}_p$ was sampled such that $T(\boldsymbol{y}_p) = T(\boldsymbol{y}_{obs})$. Further, the expressions $\pi(\boldsymbol{\theta})$ and $[\boldsymbol{\theta}, T(\boldsymbol{y}_p) = T(\boldsymbol{y}_{obs})]$ are fixed in this step of the Gibbs sampler. Similar expansions can be made of the the proposal distribution. These fixed quantities cancel in the Metropolis-Hastings acceptance ratio, and consequently, the ratio reduces to (8).

After the above explanation, an additional line of algebra between (7) and (8) may not be necessary. Also, the notation $p(\cdot)$ was confusing to referee 1.

Perhaps a better notation is $p(\cdot|\boldsymbol{\theta}, T(\cdot) = T(\boldsymbol{y}_{obs}))$.

The additional comments of the referee here indicate confusion (and perhaps this is our fault?). However, we did state clearly that $[\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y})]$ is NOT available in closed form. Perhaps adding explicitly that we are sampling from $[\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y})]$, as a suggested above, will alleviate this confusion.

Referee 1 also comments that too many details are left out of section 3.2.1:

I think what is missing here is an explicit statement that it is not trivial to directly sample a new data vector satisfying $T(\boldsymbol{y}_{obs})$. This is the motivation for first proposing an intermediate data vector and then mapping it a vector satisfying the summary statistics. After the sentence ending in "say, $(b(X, \boldsymbol{y}), s(X, \boldsymbol{y}))$" I suggest adding something to clarify this ... perhaps:

'It is not a simple matter to directly sample a data vector satisfying $T(\boldsymbol{y}_{obs})$. However, given any data vector $\boldsymbol{y}^* \in R^n$ and our assumptions, we will see in Theorem 3.1 below that it is possible to scale and shift $\boldsymbol{y}^*$ in a particular way resulting in a vector $\boldsymbol{y}$ satisfying $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$. Hence, to obtain such data $\boldsymbol{y}$, we proceed in two steps.'

From here we pick up with "First a vector, $\boldsymbol{z}^*$ is generated..." However, we might change $\boldsymbol{z}^*$ to $\boldsymbol{y}^*$ here and also leave out any detail about proposing from a reduced space until after Theorem 3.1. In particular... here is how I think the rest of this paragraph should read:

'First, a vector $\boldsymbol{y}^*$ is sampled from a known distribution (to be described later). This vector has summary statistic $T(\boldsymbol{y}^*) = (b(X, \boldsymbol{y}^*), s(X, \boldsymbol{y}^*))$ which (with probability one) will not match the observed summary $T(\boldsymbol{y}_{obs})$. The vector $\boldsymbol{y}^*$ is mapped into a vector $\boldsymbol{y}$ by rescaling and shifting appropriately using Theorem 3.1 below to match the observed conditioning statistic. To evaluate

the proposal density, we need to adjust the known density of $\boldsymbol{y}^*$ with a Jacobian. We choose the distribution of the intermediate data vector $\boldsymbol{y}^*$ to have support so the transformation to $\boldsymbol{y}$ is one-to-one and the Jacobian can be computed using first principles. In the sequel, we use this artificially low-dimensional example to illustrate the method.'

The other option I think would be to change the $\boldsymbol{y}^*$ in theorem 3.1 to $\boldsymbol{z}^*$. I know we have discussed this notional issue before . . . but I think we may be able to cut down a little on the number of 'stars' we introduce.

Referee 1 one asks, 'How does the expression in the Theorem lead to the set in (9). After set (9) we should state that an examination of the proof of Theorem 3.1 in the appendix shows that any $\boldsymbol{y}$ expressed as in the Theorem will have $\boldsymbol{b}(X, \boldsymbol{y}) = \boldsymbol{b}(X, \boldsymbol{y}_{obs})$ and $s(X, \boldsymbol{y}) = s(X, \boldsymbol{y}_{obs})$ (i.e. $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$). Further any $\boldsymbol{y}$ in $\mathcal{A}$ can be expressed as in the theorem by simply replacing $\boldsymbol{y}^*$ with $\boldsymbol{y}$. This should make clear that the range of the transformation is indeed $\mathcal{A}$.

3. *The intuition should be better explicated*

Well. . . I am a little bummed after reading, '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' given the amount of time I have thought about this sampling process and how to explain the reasoning. However, maybe there should be a little more explanation as to why the Jacobian is what it is. For example, we say on page 17 that 'This deformation contributes an attenuation to the Jacobian equal to the ratio of infinitesimal volumes in the tangent spaces of the sphere and $\Pi(\mathcal{A})$ at $\boldsymbol{z}$.' Maybe at the beginning of Section 3.2.2 we can state that from first principles Jacobians are simply ratios of infinitesimal volumes and that in the construction of the proposal we must understand what

these ratios are and efficiently compute them.

4. *The results of the application need better labeling/expostion*

   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.

   Sure. We give the expression for TLM and we could also add more explicit notation for which observations are kept.

   The same sort of comment applies to p. 27-28: Please give the mathematical expressions rather than just the words.

   This involves giving detailed expressions for how the priors were constructed. I didn't think this we was worth doing when writing the original draft but if the reviewers want it then I don't see why we couldn't do it. Perhaps adding this to the appendix would be better. It would involve some detail.

   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.

   Yes, the symbols are already different (but could be made more different) and the figures should be made larger.

   To make it easier for the reader could the discussion of Table 1 p. 31 refer to the rows in the table?

   Yes. We can add this.

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

Figure 1 is notional and used in an attempt to explain the sampling process.

5. *Gather and organize your thoughts on model mis-specification in Sec. 5* Listed in order with my comments are:

   For instance, the paragraph beginning In addition to... on p. 4 Addressed as in comment 1 by moving the discussion of model misspecification to the discussion.

   the first part of the paragraph beginning Further examples... on p. 6 I am okay with moving this directly to the discussion.

   middle paragraph on p.8, I think this is an important part of the lit review is should probably remain where it is.

   the discussion of instability in the first two paragraphs of Sec. 3.2, These two paragraphs motivate the need for the sampling method. I don't think we can wait until the discussion for this. However, the sentence starting with 'A variety of variance techniques...' in the first paragraph is is vague and we may think about removing it.

   We digress...on p. 28 This paragraph explains that no new computational strategies are needed for the hierarchical model. I think this is an important point to make since we describe our method for standard regression. It could potentially wait until the discussion.

6. Minor comments

   1) p. 3, para. beginning The focus of this work...: State this more formally with mathematical notation. Not sure how to do this?

   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.

This point is well taken and also addressed by the AE's comment number 2 who gives the suggestion "Bayesian inference for data with incompletely specified likelihood.' I am certainly okay with giving a more descriptive name or perhaps going back to 'the blended paradigm.' I am not really sure what made me stick with the 'restricted likelihood'. . . it just propagated through my dissertation and at some point just became too difficult to change with deadlines looming :).

3) p. 4, l. 11: Wong and Clarke CJS 2004 shows a case where using less but well chosen data gives better prediction. We can mention this in the list 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? Already addressed this a bit above

5) p. 4, l. -4,-5: Yes, mis-specification and outliers somehow represent the same problem...Could this point be made with more clarity? More clarification may be warranted. The discussion of model misspecification may be moved to the Discussion as noted above.

6) p. 5, l. -4: term should be factor. Okay. Can change here as well as if 'term' it is used elsewhere in similar context

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

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)). Yes, we should be careful about this.

9) p. 23, l. -1: "tunning" should be "tuning" Yes, yes it should.

Comments from Referee 2

1. *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. 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).*

   Comments in the second paragraph: These refer to the marginal gain in performance of our method and the suggestion to explore other conditioning statistics. These comments were addressed above in the AE's comments.

   Code for implementing the method can be made available.

2. *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 Tukey's 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.

Comments in the third paragraph: The referee wonders if we have intuition as to why the difference between Tukey's and Huber's estimator is much more pronounced in the hierarchical model. I believe this is due to the fact that the conditioning statistics are estimated separately within each group, and some groups have small sample sizes and/or a large number of outliers. Huber's method doesn't completely trim outliers and can be more affected by them especially with small sample sizes. Thus, the conditioning statistic may not be as good in these groups.

As far as guidelines go, I don't think these would be much different from the guidelines users of classical robust estimators use, but it is something to think about.

3. *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.*

   Comments in the fourth paragraph: These comment was addressed in the AE

section.

4. *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.*

   Comments in the fifth paragraph: Yes, we could tone down this first sentence a bit.

5. *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.*

   Comments in the sixth paragraph: We should have an additional example even if it is simple, and it could potentially study other estimators.

6. Minor Comment: Figures will be made larger.