Referee Report on "Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression"

The authors proposed a Bayesian approach for making inference based on some robust statistics of the data $T(\mathbf{y}_{obs})$, instead of the original data $\mathbf{y}_{obs}$, to avoid the influence of outliers. Computation is carried out by iteratively 1) sampling "fake" data that has the same statistics as the observed data from $f(\mathbf{y}|\theta, T(\mathbf{y}) = T(\mathbf{y}_{obs}))$ and 2) sampling parameter θ given "fake" data from $f(\theta|\mathbf{y})$. The first step is not trivial and the authors proposed an MH algorithm utilizing conditions C1-C8 on summary statistics in the context of a linear regression model.

The paper contains some interesting and nontrivial contribution to statistics and would be of interest to some readers of Bayesian Analysis, but I feel the current draft needs some major revision.

Below are my main concerns.

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.

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) $f(\cdot|\theta)$ 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 anyway to select which f or T to use based on cross-validation or other empirical methods?

3. What's the real benefit of this computationally expensive approximation framework?
   The proposed framework replies on a good summary statistic $T(\mathbf{y}_{obs})$, 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 $T(\mathbf{y})$ (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 $T(\mathbf{y})$.

   What's the gain of all the extra computation? The authors compare KL divergence, but 1) few real applications care about KL divergence not mentioning the predictive density $f(\tilde{y}|M, \mathbf{y}_{obs})$ is not available in closed-form; 2) since $f(\tilde{y}|M, \mathbf{y}_{obs})$ 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 metrics used for regression models.

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

Some minor issues.

1. Abstract "... handling data sets with outliers and dealing with model misspecification. 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.

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

3. Eq(2) in Sec 2.1, the notation $f(y_i|\theta, c(y_i))$ is not introduced yet; what's introduced is $f(y|\theta)$.

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

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

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