Response letter

Dear Editor,

many thanks for your comments on our paper "**The augmented Hat-matrix of Hierarchical Generalised Linear Models and its use in leverage diagnostics**". In the next pages, you can find detailed point-by-point answers to all the editor's and reviewers' comments.

Looking forward to hearing from you,
the authors.

Answers to Editor

I like many aspects of this paper. I think it addresses an interesting problem (the generalization of the concept of the hat matrix to hierarchical glms including mixed models) and then identification of better thresholds to identify influential clusters and individual observations within clusters.
There is a reasonable review of the relatively scant literature on the topic and then an description of the proposed method which seems to borrow much from Hodges and Sargent's work on the hat matrix for linear mixed models using the data augmentation representation.
Curiously this work was not cited in this current paper but I did find another paper written by one of these authors for linear mixed models which does cite the Hodges and Sargent work.
I believe the description of the models and methods is clearly written with enough background information to make this material accessible to a broad readership.
Re: We thank you for your positive feedback and thoughtful comments on our paper. We also value your observation regarding the work of the Hodges and Sargent, we have revised the manuscript to include the proper citations.

Editor: When getting to the motivation of the dual thresholding, I do feel a bit more motivation could be given for using the averaging approach that's being proposed. Why this over other types of approaches?
RE: The classical single-threshold formula works well in fixed-effects models due to the predictable trace of the hat matrix (p). In contrast, in hierarchical models, the trace is partitioned between the fixed and random components ($p + r$). However, the sub-matrix containing the hat values of the fixed component does not have a trace equal to p, and similarly, the part containing the random components does not have a trace equal to r. This means that, not only a single threshold is inappropriate, but also that the use of two distinct thresholds should be based on the empirical hat values.
Given this, the method of using the empirical averages of the sub-matrices' diagonal elements directly mirrors the logic of the classical threshold ($2 \cdot$ mean diagonal element), while adapting to the hierarchical structure of the data. We believe that this approach suits well the hierarchical structure of the data, preserving simplicity and interpretability. It ensures that high-leverage diagnostics remain robust without introducing additional assumptions or complexities that may hinder practitioners from using them.
This reasoning is discussed at the beginning of Section 4.2, and we have added further motivation in the revised manuscript.

Editor: With regards to the toy examples, first of all, there are too many of them. Please reduce the number to 2-3 at most. Also, a more full description of the data generating mechanisms for the toy examples needs to be given. I also think revising the text to avoid presentation like "Setting" and "Outcome" would be better. Just have the text flow from description of the setting to description of the outcome.
RE: In response to your comment, we have reduced the number of toy examples in the main text to the most relevant four, while moving the remaining two to the appendix. Moreover, following one of the referee's requests, we have added another toy example (Example 7 in the appendix) to explore an additional scenario.
We also agree with your suggestion regarding the description of the data-generating mechanisms, and have provided a more thorough explanation in the revised manuscript. Specifically, at the beginning of section 4.1, we added the model specification and provided more details on the implementation of each scenario. Finally, we have removed the “Setting” and “Outcome” labels, and have ensured that the text now flows more smoothly.

Editor: The simulation results on the thresholds was pretty impressive. Clearly the more naive approaches break down and the proposed method works well. For the few scenarios where that does not happen, could the authors provide more insights in the text description of the findings? Stating something more than what happened but why they believe it happened would be useful.
RE: We appreciate the editor's positive feedback regarding the performance of our proposed thresholds. Empirically, it can be observed that the traditional threshold $\frac{p+r}{n+r}$ is slightly larger than t_s

but notably smaller than t_c. As a result, the traditional threshold tends to produce slightly more conservative results when identifying high-leverage observations, and noticeably more liberal results when identifying high-leverage clusters (see Table 6).

Scenario 1 is particularly unique due to the small number of clusters ($m = 5$, $n = 75$) and the very low proportion of cluster containing a high-leverage observation (0.2). In such sparse cases, the traditional threshold performs better compared to the proposed one because the risk of false negatives is inherently very low, given that there is essentially only *one* true high-leverage observation to identify. However, in more complex scenarios with higher proportions of high-leverage points, the traditional threshold becomes more prone to false negatives. In these cases, our proposed threshold t_s proves more effective, as it provides a better balance between precision and recall, leading to more accurate identification of high-leverage observations.

We have now clarified this concept in the manuscript.

Editor: I like the demonstration of differences in the real data analysis when comparing naive to new approaches. Clearly the naive method is too liberal. Could there be some further details exposed about the particular cluster which showed up as highly influential using the new threshold? Also, what is the explanation for so many influential points being discovered by either threshold? Is there something unusual about this dataset?

I also agree with the referee that additional details on the real data set description is needed to be fully self-contained as well as the specific mixed models that were fit.

Overall, I thought this was a nice piece of work that would require minor revisions.

RE: We thank the Editor for the positive and constructive feedback on our work. In response, we have added further details on the real dataset, including a more thorough description and the specification of the mixed models that were fitted.

Regarding the high-leverage cluster, its hat value is particularly large due to higher values in the predictor variable (*nn_sqft_lot*) compared to the others. This deviation in the predictor values amplified the leverage of that cluster in the mixed model structure, which was why it was identified by the proposed threshold. Importantly, we emphasize that high leverage, as identified in our analysis, does not necessarily imply that an observation or cluster is influential on the regression coefficients. Leverage merely indicates a greater potential to influence the results, which warrants further diagnostic investigation.

Regarding the large number of high-leverage points identified by both thresholds, we attribute this to the structure of the dataset. Specifically, the dataset exhibits substantial variability in the explanatory variable. It is not unusual for datasets with hierarchical structures or clustered designs to contain multiple high-leverage points, as the variance components and mixed effects can amplify leverage, particularly for data points which sensibly differ from other observations in the same cluster. As previously mentioned, high leverage indicates a greater potential to influence the model, but further diagnostic checks are needed to determine if these points actually exert significant influence on the results.

We have added more details on this matter in Section 4.3 of the paper.

Answers to Reviewer 1

You derive very general expressions for individual- and cluster-level influence diagnostics and propose thresholds for 'high influence'. The framework is hierarchical generalized linear models, but that is a large class of models that includes the more commonly used generalized linear mixed models. The derived expressions are very easy to use. The threshold expressions make sense and more accurately indicate influential observations in most of your simulation evaluations, especially for clusters. The exposition is very easy to follow throughout.

Details:

Reviewer: Tp 14, section 4.1. Aren't there 10 clusters in your examples, not 5?
RE: Thank you for noticing it! It was a typo, and we have corrected it.

Reviewer: The toy examples are well chosen to illustrate the behaviour of your indices. Is there any interesting behaviour when cluster 1 has X = 100 for all observations and cluster 2 has 1 observation at X = -100? This contrasts with all your current examples that have only positive values in different patterns.
Response to Reviewer: Thank you for your suggestion. Following the editor's comment, some of the toy examples have now been moved to the appendix for better organization and readability. Moreover, we have added the example you recommended in the appendix (Example 7). In this new toy example, all the observations in Cluster 1 are assigned an X value of 100, while Cluster 2 contains a single observation with an extreme X value of -100. The results of this example highlight interesting behaviour: the individual hat value for the observation in Cluster 2, $x_{2,1}$, is considerably high due to its extreme X value, while Cluster 1 shows high leverage due to the uniformity of its observations.

Reviewer: Please provide more details about the house price data set and the model you have fit: How many observations are in the house price data set? What model are you fitting? linear in nn_sqft_lot with a random intercept? Or a random-coefficient regression (correlated random intercept and slope). I am struck by the relatively large number of influential points, especially when there is only 1 covariate.
RE: To address your concerns, we have clarified the details of the house price dataset and the model in the revised manuscript. Specifically, we have added the following sentences:
"Given the nature of the data and the results obtained from the scenarios in the simulation study, we decided to fit a Normal-Gamma model as an example of non-conjugated HGLM. Specifically, the model assumes a linear relationship between the expected price and the predictor *nn_sqft_lot*, and includes a random intercept for each *zip_code*, modeled using a Gamma distribution with a logarithmic link function. The response is assumed to be normally distributed, with expectation:

$$E[\text{price}_{ij}] = \beta_0 + \beta_1 \cdot \text{nn\_sqft\_living}_{ij} + u_j,$$
$$u_j \sim \mathcal{G}\left(\frac{1}{\lambda}, \frac{1}{\lambda}\right)."$$

Regarding the large number of high-leverage points identified by both thresholds, we attribute this to the structure of the dataset. Specifically, the dataset exhibits substantial variability in the explanatory variable. It is not unusual for datasets with hierarchical structures or clustered designs to contain multiple high-leverage points, as the variance components and mixed effects can amplify leverage, especially for points that lie far from the center of their respective clusters. However, we highlight that high leverage, as identified in our analysis, does not necessarily imply that an observation or cluster is influential on the regression coefficients. Leverage merely indicates a greater potential to influence the results, which warrants further diagnostic investigation. We have now added new sentences in Section 4.3 of the paper to make this concept clearer.