Author's Response to Review

We appreciate the helpful reviews. The paper has been revised following these suggestions as closely as possible. Specific changes made in response to each reviewer comment are shown in blue below.

Response to Reviewer #1

Visual Inference is a very interesting topic. The authors have presented their ideas about how to use visual inference for HDLSS data in a clear way, and the results they had are indeed interesting. My main comments about this paper are as follows:

1. In the abstract, the authors refer to "large dimension small sample size" data, and then use the acronym HDLSS, which would stand for "High Dimensional Low Sample Size". You probably need to stick to one of them.

Author response: This is changed in the abstract. "large dimension small sample size" is changed to "high dimension low sample size". This is checked throughout the paper and changed if needed to.

2. Almost all of the graphs in this paper are not color blind friendly. I would suggest to use color combinations other than red and green (I believe it is blue, but on print it looks green). You can refer to http://colorbrewer2.org/ for the choice of good graphics colors. This issue might have affected the experiment results!

Author response: The plots in the paper are made color blind friendly now. The color scheme was selected using http://colorbrewer2.org/.
All the plots in the lineups have separation among the groups. The least separation can be obtained from dimension = 20, separation = No. In this case, neither the null plots or the observed plots will have any separation. So the subjects will not be able to identify the observed plot. For the other dimensions, the separation is always present and the subjects has to identify the largest separation among the groups. Hence it is safe to say this issue did not affect the experimental results.

3. In the caption of Figure 2, the author refers to "m=20". What is m? I did not see any reference for that in the text. And Is there any reason for choosing 20? Please clarify if so.

Author response: The clarification is added to the paper.

4. Table 2 is not explained in a very clear way in the text or the caption.

Author response: Table 2 is explained in details in the text of the paper.

5. In the section 4.4 (data collection), the authors mentioned that their subjects were recruited through Amazon Mechanical Turk. Could you please elaborate in one or two sentences what is AMT?

Author response: A paragraph is added to the paper about Amazon Mechanical Turk.

6. In Figure 7, I suppose that there should be three dots for each dimension. Some of the dimensions (e.g. p = 0 and Projection = 1D), has two dots. Does this mean that there are some dots on top of each other? Please elaborate.

Author response: Figure 7 has been changed to include alpha blending for the points now. The transparency gives an idea about the frequency of dots on a single point. A line has also been added to the text explaining this figure.

7. It would interesting to compare between different groups of subjects based on gender and age.

Author response: This is added to the paper. The effect of gender, age and education on the response of the subjects was studied but no significant effect was noticed.

8. What software did you use for producing these graphics?

Author response: The reference is added to the paper.

Reviewer #2 - no review returned

Response to Reviewer #3

My overall impression of this paper, and particularly the research area of visual inference, is very positive. The authors are applying empirical methods to assess not so much the quality of visual inference but rather the quality of a particular visual inference (jittered dotplots/scatterplots of PDA projections) for HDLSS data. It seems to me that the present title is therefore overstating the contribution of the results.

Author response: The title of the paper has been changed to better suit the content of the paper.

A real difficulty with an empirical study on visual inference methods is that in the end, we come away with results that simply confirm our prior assessments. In the present case, that it becomes more difficult to correctly detect fixed differences (which do not grow with dimensionality p) and that this is worse when the dimensionality of the separation is greater (d = 2 rather than d = 1). Im pretty certain that the authors had this view before undertaking the study.

The real punch of the paper is stated in lines 20-21 of page 2 of the manuscript: There is no conventional inferential methods (sic) which enables us to conclude . . . statistically significant or not. I suggest that the paper early on state this as the principal contribution. The intent of the paper would then be twofold:

1. Visual inference methods may be used where conventional methods are unavailable. The case in point is any test applied to the separation of groups that does not take into account that the dimensions used were themselves determined empirically. In this sense, a conventional test is

conditional on the derived dimensions and so interpreting its significance level as unconditional is inappropriate.

The visual test can incorporate the whole process. (Although a more proper comparison might be with a bootstrap distribution of the formal tests significance level, one which incorporated the whole of the process including the PDA.) The discussion of HDLSS is not as central as the present manuscript would suggest.

2. The visual tests are consistent in behaviour with what we might expect vis-a‘-vis the effects of increased dimension on the ability to detect separation as well as the time taken to do so. Of these, the first is by far the more important.

Following a clear statement of this intent, the Mechanical Turk experiments could be introduced to demonstrate the first point (i.e. many of the detailed results need not be discussed at this point). Then the scientific problem of the wasp types could be introduced and discussed as a con- crete application. Again, to me this is the punch of the paper and deserves this kind of emphasis.

I would then delay the other interesting results of the Mechanical Turk experiments to a new section, one where the characteristics of the visual test are being investigated. These include:

- The visual test reflects our prior view of the effect of increasing dimensionality (fixed n), projection pursuit, etc. The results relating proportion correct, and time taken to select would appear here.
- Effects of visualization choices.
  - rotation effect
  - connection with WBratio
  - effect of particular null plots
- speculation for future research on this particular type of visual testing

As I see it, this is mainly a re-organization of the material in the paper. But is one that would make the paper more effective . . . I think.

Author response: The material in the paper is re-organized as suggested by Reviewer #3. The wasp experiment is explained first and then the follow up simulation experiment is described followed by the results.

Some detailed comments

These are mostly questions for possible elaboration.

- Page 6, last paragraph of Section 4.1, Hence the probability . . . . I would write this as it gets harder to detect real differences. Also, would it not be more meaningful to look at the individual absolute differences, or the average. And what are these differences? That is which

means? (on each variable? on the PDA direction?)

Author response: Necessary changes are made. A paragraph is added to the paper explaining this clearly.

- Last sentence of Section 4.3. Isnt this biasing the sample towards selecting plots where the observed effect is minimal?

  Author response: The procedure does not bias the samples. In fact the procedure makes it fair as it makes sure that the optimization procedure works even in the null plots. Otherwise the null plots will show no distinct clusters.

- Last sentence, page 10, . . . but the opposite . . . What does this mean? Users pick out real data (significantly) when there is no separation better and better as the dimension increases? (Seems to be indicated by bottom row of plots in Fig. 7. But this makes no sense)

  Author response: This is real data which is obtained from the Amazon Turk Experiment. The slope of the line is not significant. If we could do a lineup by permuting the dimension for projection:1D and 2D separately for separation = No, it could be seen that the real data will not be picked. The proportion correct is close to 0 for all dimensions when there is no separation. This can be seen in Fig 11 as well.

- Section 5.3, page 11. Im not sure this is the best test of the rotation effect. Why not always colour the points the same in a clockwise order?

  Author response: The groups for one dimensional projections are colored systematically. But the groups for 2D projections are randomly colored. The effect of projection is not significant. Hence it can be assumed that the rotation effect is not significant. But this is surely not the best test as the rotation effect is confounded with the color effect. Coloring the points in a clockwise order will be hard to do. But this can be thought as a future work where for some lineups, the groups will be colored in a clockwise order and for some lineups, the groups will be randomly colored and then the rotation effect in the plots could be tested.

- Figure 8. These points should be jittered, rather than rely on alpha blending.

  Author response: Figure 8 has been changed. The mean time taken for each dimension is plotted on the log scale against the dimension with colors representing presence of separation. Bootstrap confidence bands are also drawn.

- Page 12, Section 5.4 This suggests that as the number of dimension (sic) . . . when the data has some real separation.
  Could it not be that giving people however much time they like is confounded with their ability to pick the true plot? That is, as they have more time, they second guess themselves and so choose the wrong plot? I think psychologists much prefer short fixed time responses for this reason. Too late now, but I think it would be interesting to have also considered and experiment where the time viewing each plot was fixed (short) so that no response would be indicative of indecision.

  Author response: Providing fixed time to evaluate is an interesting suggestion which will be considered in a future experiment.

- Page 13, end of topmost paragraph. Investigating these lineups further may reveal why this is. This should be done now. For example Fig 9(a), row 3 from top, dim = 100. What does the real data config look like compared to those with small WBratios? Either the WBratio is a poor measure, or there is something interesting about these configurations. Worth examining and presenting.

  Author response: WBratio is surely not the best measure to describe separation. A more meaningful measure is used. Separation between the groups is calculated and the average separation is considered. For two groups, the average separation is equal to the separation.

- Figure 11, page 15, and surrounding discussion.
  Isnt part of the problem here that separation of groups visually is not well defined. For example, it might be the case that the human visual system favours symmetry in the separation. In Figure 11, this might correspond to a preference of the first plot in the top row. It would be interesting to know if there were some favoured configurations and what they actually looked like.

  Author response: The subjects were asked to identify the plots with the most separation among the groups. This relates to the maximum average separation. A plot (not included in the paper) like Figure 11 clearly indicates that the subjects pick the plots with the most separation among the colored groups.

Some corrective comments There are numerous typographical/grammatical errors in the paper. For example (not an exhaustive list):

- poor English/style:
  - second sentence of the abstract should read We often seek low-dimensional projections of high-dimensional data that reveal . . .

    Author response: This is changed in the paper.

  - in the abstract, line 20, authors should use exactly the phrase high dimensional low sample size instead of large dimension small sample size since this is where the acronym HDLSS is introduced

    Author response: This is changed in the paper.

  - line 2, page 2. You might use variances rather than distances between means so as to have the between group description parallel the within group (which uses within group variance)
  - line 22, page 3. A simpler first sentence might be: For example, suppose we have data on the concentration . . .

    Author response: This is changed in the paper.

  - line 37, page 3. Last sentence should read A comparison of this visual test with the conventional test is shown in Table 1.

    Author response: This is changed in the paper.

- punctuation: First sentence of Section 5.5 should have an apostrophe on subjects to indicate the possessive.
- mismatched verb tenses
  - page 3, last two sentences of topmost paragraph. analyzes, provides. Technically, data is plural in the last sentence so dataset might be a better choice.
  - page 12, title of Section 5.5. affects
- latex suggestion
  - page 4 X-bars. Try using widebar instead of bar to make the overline more prominent.

    Author response: This is changed in the paper.