Referee's Report

Visual Statistical Inference for High Dimension, Small Sample Size Data

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.

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 $d = 1$). I'm 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 test's 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-à-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 concrete 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.

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?)
- Last sentence of Section 4.3. Isn't this biasing the sample towards selecting plots where the observed effect is minimal?
- 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)
- Section 5.3, page 11. I'm not sure this is the best test of the rotation effect. Why not always colour the points the same in a clockwise order?
- Figure 8. These points should be jittered, rather than rely on alpha blending.
- 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.

- 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.

- Figure 11, page 15, and surrounding discussion.

  Isn't 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.

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 . . . "
  - 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
  - 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 . . . "
  - line 37, page 3. Last sentence should read "A comparison of **this** visual test with **the** conventional test is shown in Table 1."
- 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. "analyze**s**", "provide**s**". Technically, data is plural in the last sentence so "dataset" might be a better choice.
  - page 12, title of Section 5.5. "affect**s**"

- latex suggestion
  - page 4 X-bars. Try using “widebar” instead of “bar” to make the overline more prominent.