February 7, 2024

Co-Editor, *Journal of Marketing Research*

Dear Professor Gordon,

We thank you and the review team for continued helpful feedback that has helped refine the paper along several dimensions.

We include a separate response to the issues raised by each member of the review team, including AE, and R1. We thank you for the specificity of your prior suggestions, as we have implemented and responded to all of them. We believe that we have addressed all concerns raised in the review process, and trust the present revision will be acceptable to *JMR*.

Best,
The Authors

Summary of Major Changes in the Revised Paper

We include below a summary of the major changes in the present revision, with most changes arising from the review team's feedback.

1.

2. Exposition:

    (a)

Response to the Associate Editor

Thank you very much for your positive and constructive evaluation of our paper and for the opportunity to submit a revision of the paper. We provide our point-by-point responses to your comments below. To provide context, we reproduce your original comments as appropriate.

AE 1: *R1 and I both previously asked why some supervisory variables work better than others and why adding more supervisory variables doesn't help. The explanations in the paper are still a bit ad-hoc. Have you considered the possibility that good supervisory variables are good because they are correlated with the "look" of the product? For example, maybe brand and circa are good supervisory variables for vintage watches because the visual design of watches is correlated with brand and the era. Perhaps sneaker brands all provide a wide range of designs, but price is correlated with visual design. If you agree with this hypothesis, then this overarching explanation could be developed more in the discussion of the results, e.g. on page 28, and mentioned as an area for future research in the discussion.*

Reply: Thank you for

AE 2: *Like R1.2, I found the new distinction made between "structured product characteristics" and "visual product characteristics" in the opening paragraph distracting. I suggest you cut the first paragraph and lead with the statement of the importance of visual product characteristics in the second paragraph.*

Reply: Thank you for

AE 3: *I am familiar with variational Bayes, but not variational autoencoders and found the description of the VAE on page 11 very hard-to-follow. I do not understand why AE is not generative. I do not understand why it is useful to have a full posterior for the visual features in VAE. Is the multivariate Gaussian mentioned on page 11 a part of the specification of the encoder and decoder models? Or is it used to approximate the posterior in Variational Bayes? What precisely does it mean to "model uncertainty of the visual characteristics" (page 12)? How are the full posteriors used? Please make edits to both the introduction and the methodology sections. Your target JMR reader would be familiar with Bayeisan inference, but perhaps not VAE or variational Bayes.*

Reply: We really appreciate your

AE 4: *Thanks for adding the new description of the UDR. It is now clearer to me how UDR is computed. But there are some issues with this new section.*

1. *I found the first paragraph beginning "Evaluating Disentanglement Performance" to be too high-level to be meaningful. I suggest you move this to the end of the section to sum up the reasons for using UDR. (BTW "UDR is a metric based on heuristics of good disentanglement" doesn't make sense.)*
2. *Please explain earlier that UDR is based on different fitted versions of the same VAE architecture. Are these different sampled training data sets? Or do the different random number seeds affect the variational Bayes algorithm in other ways, too?*

3. *The variables a and b are overloaded in equation 6. Suggest using $\tilde{a}$ and $\tilde{b}$ in the summations in the numerators.*

4. *The definitions of r_a^2 and R(a,b) are difficult to follow. Please define them more directly. Why is r_a^2 squared and R(a,b) is not?*

5. *I could not figure out how you aggregate the UDR across the 45 pairs of fitted models. Do you just average?*

Reply:

Here are

1.
2.
3.
4.
5.

AE 5: *Conjoint study:*

1. *The paper finds empirically that the standard hierarchical Bayes logit model with the disentangled visual features seems to work well for predicting choices relative to some newly-introduced ML alternatives. But we don't have a lot of evidence that this will always hold; this comparison is done for one relatively small data set. So, I see this as an empirical "side note" for this paper, but a potentially important one. (I expect the next decade will bring more research into when and how ML models outperform HB. See Smith, Seiler and Aggarwal 2022 https://doi.org/10.1287/mksc.2022.1387, for another comparison of structured HB versus ML approaches.) So, I would leave Table 7 as-is, i.e. do not change the reported analysis. But downplay the finding that HB does well in the introduction (page 3) and bring this up as an area for future research in the discussion.*

2. *It is not clear to me whether Figure 10b is showing estimates of the elements of Ω_β or the observed covariance of the estimated β_i. Those should be similar, but please clarify in the Figure caption for completeness.*

3. *The new discussion of observed and unobserved heterogeneity in the alternative conjoint models was difficult to follow. What does "does not account for uncertainty nor heterogeneity" (referring to the homogeneous logit) mean? What does "modeling uncertainties" mean when discussing the HB MNL? When you say the interactions were modeled as "homogeneous (not conditional on covariates", does that mean they also had no unexplained heterogeneity? You describe the Random Forest as not heterogeneous, but if you pass in the user covariates, it could pick up some explained heterogeneity and if you pass the respondent ID in, it may pick up some unobserved heterogeneity. When discussing heterogeneity, please be very clear about observed and unobserved heterogeneity. In most cases, you can replace the term uncertainty with observed or unobserved heterogeneity. It would be helpful to explain the input data for each model more precisely (but still concisely).*

4. *Is it possible that the conjoint design favors linear utility models? Each visual characteristic was presented only at high, medium and low levels, which might only weakly identify non-linearities in preference. Please address this limitation in the discussion of Table 7.*

Reply:

Here are

1.
2.
3.
4.
5.

AE 6: *The "ideal point" analysis makes some big assumptions which I'm not sure I believe. First, it assumes that it makes sense to segment the population based on the SVD of the Θ matrix (equation 9). Why $\bar{\Theta}$ and not β_i? Second, to generate an interior "ideal point" when utility is linear in the visual, the analysis requires some sort of "cost" assumption. The analysis assumes a very specific penalty – the Euclidean norm of existing products – that seems to bear no relationship to real production costs. Asking the reader to understand and accept these assumptions is getting in the way. This point of the conjoint analysis is to show one practical use of the disentangled visual features and the generative model. I think it might be just as effective to show the ideal product specification for a segment defined by your observed user characteristics like high-income men versus low-income women. You could explain it more quickly and it would be a whole lot easier on the reader.*

Reply:

Here are

AE 7: *Minor wording suggestions:*

1. *Abstract: "characteristics constituting an object's visual representation" ¿¿ "visual characteristics".*
2. *Abstract "readily available structured product characteristics as supervisory signals to enable disentanglement" >> "readily available product characteristics such as brand and price as supervisory signals to discover disentangled visual characteristics".*
3. *Page 3: "since we are able to vary the visual design along one one (or a subset of) discovered dimensions" >> "since we are able to automatically generate images that vary the visual design along only one of discovered dimensions"*
4. *Page 6: strike "that supervised learning may not be a panacea and,"*

5. *Page 6: "human interpretable" >> "human-interpretable" [Note, a hyphen should generally be used for compound adjectives, but not compound nouns.]*
6. *Page 11, it appears that you are using both the greek letter omega and a boldface lower-case "w" to mean the same thing.*
7. *Page 25: "we do no artificially" >> "we do not artificially"*
8. *Page 36: There is no need to report your random number seeds. The exact values don't mean much to the reader.*

Reply:

Here are

1.
2.
3.
4.
5.

AE 8: *R1.3-5 are good suggestions and should be incorporated in the revision.*

Reply:

Here are

AE 9: *Any additional edits you have to simplify the exposition and make points A-C really stand our for the reader are very welcome. I want readers to really see the contributions really clearly, so that they can appreciate your work.*

Reply:

Here are

Response to Reviewer 1

R 1.1: *I did not follow the authors' response to my question about why different supervisory signals seem to produce the same visual features. Their reply to my comment 1.1.ii was: "However, if we have sufficiently good signals with valuable visual information... we would expect to find the same set of visual characteristics that are present in ground truth" (up to order/sign/etc.). This was exactly my question though: why should different supervisory signals guarantee some correlation with exactly the same set of visual features? If brand is correlated with some set of visual features, and price a different set, wouldn't those supervisory signals lead to different disentangled representations? (This question relates to my overall original question of, "how do we know we've found the right signal to get the right disentangled representation?")*

Reply: Here are

R 1.2: *I'm not sure if this has changed from the last round to this round, but in the introduction, there is a lot of discussion about structured product characteristics, where the argument seems to be that the visual product characteristics are not structured. I would argue that they are structured characteristics – things like color are very classic structured features. This is a minor choice of words, but I think it starts the paper off on the wrong foot, so to speak.*

Reply: Here

R 1.3: *With the addition of more details of vanilla VAE, I find the transition to beta-VAE a bit unclear. When beta-VAE is introduced on p. 14, there isn't much intuition as to why it induces disentangling. I would suggest the authors perhaps foreshadow that the explanation will come later, when the loss terms are described. It may be worth it to divide this long section into several subsections.*

Reply: Here

R 1.4: *It took me a few passes through to understand the new section, "Empirical Evidence for Supervisory Signal Effectiveness in Disentanglement." I would suggest putting the brand / price intuition at the start, and then using that to motivate the JS metric.*

Reply: Here

R 1.5: *Typos and small errors:*

- *There are two periods around footnote 5 on p. 6*
- *The second sentence of the section "Representation Learning and Disentanglement" is an incomplete sentence (p. 9)*

- *Basically the same sentence is repeated twice, before and after the "Methodology" header on p. 9*
- *I don't understand this sentence on p. 42: "This interesting case can and did happen given our less strong assumptions on segmentation than often conventional (e.g., hardthreshold latent class or mixture models), while at the same time taking a fully Bayesian approach to estimate individual-level preference distributions when modeling consumer heterogeneity."*

Reply: Here

References