Response to Reviews: JCGS-25-420

Jayani P. Gamage, Dianne Cook, Paul Harrison, Michael Lydeamore

2025-08-19

We thank the anonymous associate editor and reviewers for their careful reviews. What follows is our point-by-point response to reviewer comments.

Note that the original reviewer comments are in normal and *our response is in italic* text.

Reviewer 1

Summary:

This manuscript proposes an algorithm to evaluate nonlinear dimension reduction (NLDR) techniques. The authors identify some characteristics and limitations of NLDR and develop an algorithm to analyze NLDR results' quality of how well they match the original high-dimensional data.

Strengths:

1. This manuscript is well-organized, providing sufficient background knowledge and examples which makes this paper easy to follow.
2. This paper contains sufficient technical details in the introduction of the algorithm.

Weakness:

1. The main concern is whether the proposed algorithm can properly evaluate the NLDR's performance. Using the RMSE plot with different binwidth values (e.g., in Figure 8 and Figure 11) to evaluate NLDR's performance is problematic for not considering the density within each cluster. For NLDR results that have denser clusters, for example layout a,c,g,h in Figure 11, a smaller binwidth needs be used to achieve the same average

count in each bin, which will cause their lines in Figure 11 to be more upper left. However, denser clusters are not worse NLDR results.

2. In Figure 7, as RMSE steadily increases with the binwidth, it's not clear why the three binwidths 0.03, 0.05, and 0.07 are chosen. In addition, we shouldn't use a single simulated dataset to determine the default hyperparameters without any real analysis.
3. For several figures that contains results from multiple NLDR algorithms, it would be nice to add the information of which algorithm/ what hyperparameter values for each of the subfigure. Many readers may have tried several NLDR algorithms themselves, and they will benefit from those figures to gain insights about behaviors of different NLDRs. Also it's nice to mention which data is used to generate each figure (e.g., in Figure 7).
4. In section 3.6.1, it says "Values of b1 between 2 and b1=sqrt(n/r2) are allowed", why the choice of b1 should consider r2?

Reviewer 2

Summary

The manuscript proposes a novel, mostly graphical diagnostic to evaluate nonlinear dimension reduction (NLDR) layouts by effectively "lifting" the 2-D embedding back into the original high-dimensional space and inspecting the fit visually via tours (i.e., sequences of 2D-linear projections of the original high-D space). The goal is to help analysts discern which 2-D projection is the most "honest" (faithful) representation of the true structure in the data. The paper's contributions are:

1. an algorithm that overlays a geometric model on an NLDR layout using hexagonal binning and Delaunay triangulation in the embedding, then maps this model into the original large-D data space for visual assessment;
2. a new goodness-of-fit metric, the RMSE in original space, computed as the root mean squared distance between each point and its bin centroid in high-D (Eq 2) (3) interactive visualization tools (implemented in published software) to compare/investigate different NLDRs

It is mostly well written, adresses an important problem with critical real-world relevance, and provides a very well done, extensively documented R implementation. However, this reviewer is skeptical about the suitability of the newly proposed GOF metric to choose the most "reasonable" or "faithful" NLDR, finds the empirical evaluation of it to be severely lacking without comparison to alternative measures from the literature, and would request some restructuring and shortening of the paper.

Strengths

Relevance, Accessibility, and Intuitiveness

The way this method visualizes the spatial distortions induced by NLDR seems uniquely accessible to me, and, coupled with the modern interactive/animated visualizations implemented for it, seems likely to become a very useful tool to investigate and understand high-D data structures while avoiding over-reliance on / misinterpretation of misleading 2D patterns so prevalent in many areas at the moment (e.g. Izarry (2024), Chari & Pachter (2023)).

Background Material & Examples

The paper is fully fleshed out with illuminating examples, and I commend the authors for the effort they put into documenting their package and providing very informative videos etc of the method in action.

Reproducibility and Code Availability

The paper is exemplary in this regard - the underlying software is a CRAN compliant R package, and the Github repository itself allows to recompile the entire paper and appendix with all figures etc from scratch.

Major Issues

No Standard Embedding Quality Metrics or Comparisons

My primary concern is the absence of established quantitative metrics

1. to evaluate to what extent the newly proposed "RMSE" metric corresponds to established notion of embedding faithfulness and
2. to validate the different embeddings under consideration.

The paper's only evaluation of "honesty" is done via the proposed custom "RMSE" in high dimensional space (defined in Section 3.4) and visual heuristics, but it does not report any conventional measures like trustworthiness/continuity scores, k-Neighborhood preservation rates (as derived from the co-ranking matrix - e.g. R_{NX}-curves for different neighborhood sizes k, c.f. (Lee et al, 2015), (Lueks et al, 2011)), or Shepard diagrams of original space vs. embedding space distances. This is a significant omission because NLDR almost always faces a trade-off between optimizing preservation of local structures vs preservation of global structures, which these metrics are (mostly) designed to resolve / analyze. For example, the authors could have computed such measures of trustworthiness for each NLDR layout to see if the one with lowest

"RMSE" indeed had the highest (local and or global) neighborhood structure preservation but no such analysis was provided. The manuscript doesn't even mention or discuss these metrics, which is not acceptable in a study focused on faithful representation, in my opinion.

Questions a revised version should aim to resolve:

1. does "RMSE" measure local or global structure preservation (or a mixture of both)? From the definition, it seems to be focused on how well small neighborhoods in embedding space (the bins)) correspond to small neighborhoods in high-D space, with, IIUC, particularly high values if there are lots of "hard intrusions" (i.e., points that are embedded in the same 2D bin despite being far apart in high-D space)?
2. how strongly are "RMSE" values correlated to other established measures of (local or global) structure preservation? and (how) are these correlations affected by the binning hyperparameters?
3. are the NLDRs selected by lowest "RMSE" the ones that would also be selected by other, more established measures of embedding faithfulness whose characteristics are well understood? at the very least, such measures should be provided, compared and discussed for all the embeddings under consideration in the case studies.

The reliance of "RMSE" on euclidean distances in high-D space also seems limiting:

1. at least for coarser bins and/or highly non-linear data structures, geodesic distances along the data manifold à la ISOMAP are likely to measure a more relevant kind of dissimilarity
2. a definition based on a more general notion of distance would make this applicable to investigating 2D embeddings of data types for which the euclidean distance of their numerical data representations is not a suitable/applicable measure of dissimilarity (e.g. images, text).

Insufficient Validation of the Proposed Diagnostic (RMSE) Across Methods

Related to the above points - the paper has no theoretical or empirical demonstration that minimizing this "RMSE" aligns with minimizing other distortion measures like stress or maximizing neighborhoood preservation (as measured by the co-ranking matrix). Furthermore, the claim that RMSE can be used to compare different NLDR methods on equal footing deserves scrutiny. Different methods optimize different criteria; it's plausible that one method might (tends to) achieve lower "RMSE" simply by construction, especially if my reading that it mostly rewards embedings that manage to avoid hard intrusions is correct. Overall, much more critical validation of the RMSE diagnostic is needed to convince readers that "honesty" as defined here aligns with meaningful goodness-of-fit in embedding.

Limitations of the Evaluation (Simulation Study Design):

The manuscript includes a simulation to illustrate the approach (the “2NC7” dataset with two nonlinear clusters in 7-D) and two real-data applications (single-cell RNA-seq and an image dataset of handwritten digits). While these examples are appropriate, the scope of simulation scenarios is narrow, and the analysis lacks any notion of variability. The simulation seems to be a single realization of a specific two-cluster scenario. The authors did not explore, for instance, varying the overlap between clusters, adding more clusters or different manifold shapes, or introducing different noise levels. Ideally, the authors would systematically vary factors like: number of clusters or pattern types, intrinsic dimensionality of structures, sample size, etc., and show that their approach reliably selects the truth (or at least, the most interpretable view) in each case, and how well alternative measures would do so. Because only one run is shown, there is also no information how stable the RMSE curves in Figures 8 and 13 are across replications. I encourage the authors to perform more systematic, replicated simulations, report measures of variability for their performance evaluations, compare against relevant alternatives, and take much more care not to overstate their conclusions beyond the tested scenarios (e.g. the claim that it “identify bad representations so they can be avoided” is based on a very limited set of examples).

Minor Issues

1. First part of the title seems inappropriate for a scientific paper to me, as does the implied (and not substantiated, IMO) claim that the proposal is (primarily) useful to reliably select one of many NLDRs (as opposed to primarily useful for visually investigating details about a specific 2D NLDR).
2. For marketing and clarity reasons, I would strongly suggest to not name the metric “RMSE”, but something more distinctive/descriptive.
3. Flow from 3.1 to 3.2 (with too many sub-subsections) is difficult. Sections 3.2.1 (Scale the data) and 3.2.2 (Construct hexagon grid) are just short paragraphs, would be clearer if combined into a single numbered list of steps for the model construction. Much of 3 might be clearer and more compact if simply replaced by a direct step-by-step description of the algorithm in algorithmic pseudocode form with suitable short explanations.
4. Section 3.5 is a claim without supporting evidence – many out-of-sample embedding methods exist (some based on rather similar ideas, which should be referenced), you would need to show that this particular one actually works reasonably well for this to deserve space in the paper. I would strongly suggest to simply mention this in the discussion section as a possible avenue for further development instead.
5. Section 3.6 should be drastically shortened and details moved into an appendix.

6. p. 18: "A particular pattern that we commonly see is that analysts tend to pick layouts with clusters that have big separations between them." The literature/internet contains many lively discussions of the perils of mis-interpreting and post-hoc'ing NLDRs (e.g. (Izarry, 2024), (Chari & Pachter, 2023)), it might be good to cite/summarize these instead of/in addition to basing your motivation on personal anecdotal evidence? The tone here seems overly conversational to me (also: "we almost always see there are no big separations in the data") and should be made more formal.

7. Section 4, esp. p.19: IMO, "to compare and assess a range of representations" an analyst really should look at their respective RNX curves, stress values, and Shepard diagrams, think about whether they care more about local or global structure preservation, and then decide based on those factors. The method proposed here seems to me to be mostly useful to then investigate specific regions / slices of high-D space to see for which sets of data points the high-D structure is (not) preserved well in the NLDR. I find this whole section to be vastly overselling the proposal – it ignores and implicitly discards most of the prior work in this field. Its intro paragraphs are also highly redundant with other content in the paper and should be cut.

8. Section 5 should be moved into an appendix to shorten the paper somewhat. Figure in Section 5.1. could benefit from coloring the clusters in different colors so global structure preservation is also somewhat legible from the figures for the NLDRs (e.g. "is orange cluster between green and blue as in high-D or do they switch positions"). Section 5.2 might also have benefited from a corresponding ISOMAP or similar embedding of this data suitable for truthfully (isometrically!) unrolling the high-D manifold, and then showcasing that such a successful "unrolling" can be discerned from the way the projections look? tSNE just isn't isometric, so the distortions visible here are entirely expected, IMO?

Typos and grammar:

Revised version should be much more carefully proofread: "appropraite", "doen't", "it is has two separated nonlinear clusters", "the methods tNSE"