Jeremy R. Manning
Dartmouth College
Department of Psychological & Brain Sciences
HB 6207 Moore Hall
Hanover, NH 03755

June 10, 2021

To the editors of *Nature Communications*:

We have enclosed our revised manuscript entitled *High-level cognition during story listening is reflected in high-order dynamic correlations in neural activity patterns* (previous submission: NCOMMS-19-32008A). We appreciate the reviewers' insightful comments on both of our previous submissions.

Reviewers 1 and 3 indicated that they had no remaining concerns about our manuscript. Reviewer 2 asked us to provide additional detail about the new simulations we added in our previous submission, and to include references to some additional work. We have made the requested changes in our latest revision, as detailed in our point-by-point responses to the reviewers' comments. The reviewers' comments are *italicized* and our responses are in **bold**.

Thank you for considering our revised manuscript.

Sincerely,

Jeremy R. Manning
Jeremy.R.Manning@Dartmouth.edu

Reviewer #1 (Remarks to the Author):

I am satisfied with the authors' responses to my comments

(Nothing to address)

Reviewer #2 (Remarks to the Author):

I thank the authors for their substantive revisions to the manuscript; I have just a few points remaining.

[1] Thank you for for adding the new simulations (Figure 3), which help greatly to concretize the main claims and illustrate the strengths and weaknesses of the analytic approach. I do have two remaining question about these simulation results:

(a): what is the limiting factor in achieving higher correlations between the ground truth values in the simulation and the practically recovered values? Is is just the length of the data samples? If this is the case, it could be helpful to make this clear to the reader, and it could be illustrative to show how the match with the ground truth approaches ceiling values as sources of noise are decreased and sample size is increased. [I recognize that this may render the synthetic data less realistic, but it seems important to demonstrate that the true underlying values can be recovered in the limit of infinite noiseless data.] On the other hand, if I am misunderstanding the reason for the (seemingly) low correlation between ground-truth and recovered values (e.g. please explain the source of this discrepancy), or the reason why the "ceiling" on these correlations is low.

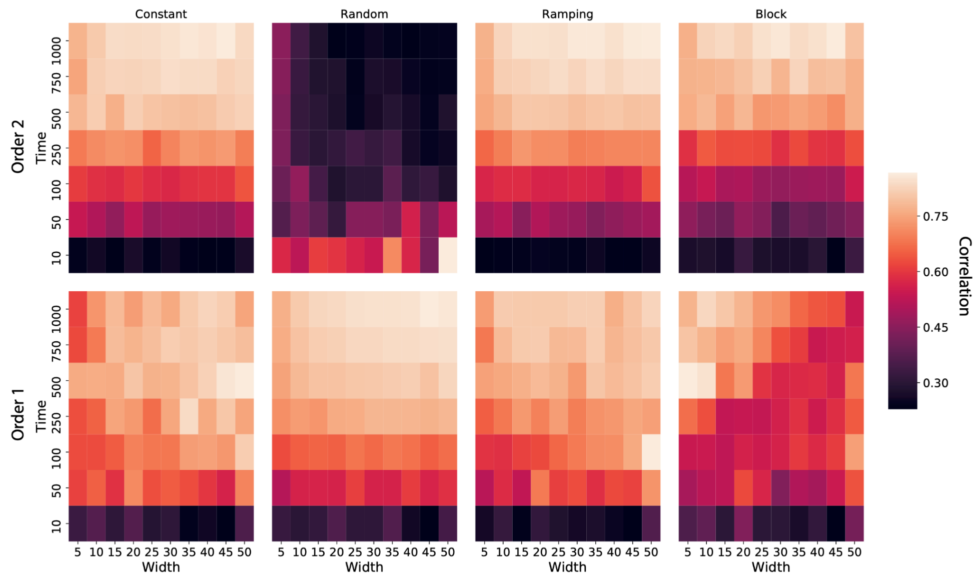
There are (at least) three main factors that negatively affect the correlations between the ground truth and recovered dynamic correlations (we have added a note to this effect to page 7 of our revised manuscript, along with two additional supplemental analyses summarized in Figs. S3 and S4):

A. **First, the procedure that is used to generate the correlations is itself noisy (based on random draws), so in practice what we refer to as the "ground truth" reflects the expected values from the data generation process, rather than the sampled values. Our characterization of the "ground truth" at each order becomes less accurate as the order decreases. This follows from how we generate synthetic data. The highest-order timeseries is generated first (via a noisy process), and each successively lower-order timeseries is generated via a noisy process from the next-higher-order timeseries. In this way, noise in our synthetic data *generation* procedure propagates from higher orders to lower orders. We think**

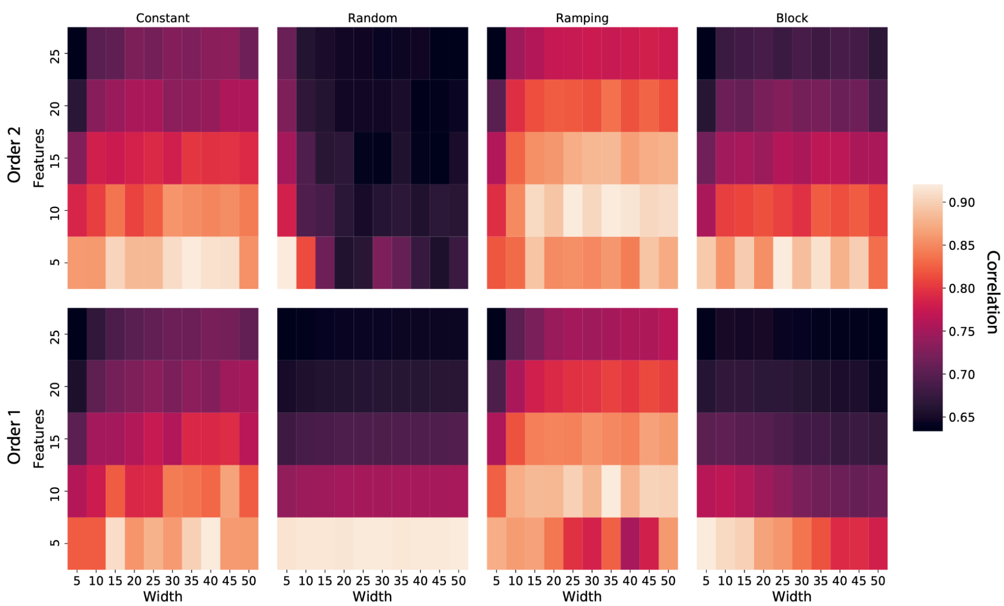
this is the primary reason that the recovered second-order correlations appear to be lower than the recovered first-order correlations.

B. **Second, our procedure for recovering dynamic (low-order and high-order) correlations is inexact. Because each order (*n*) of correlations is estimated using the next-lower order (*n* - 1), noise in our estimation procedure is compounded with each new order. In this way, *estimation* errors propagate from lower orders to higher orders. This means that the best possible recovery performance we can achieve at a given order is upper-bounded by the reconstruction accuracy of the next-lowest order.**
C. **Finally, as the reviewer notes, our simulations are data-limited (e.g., we cannot generate timeseries data embedded with high-order correlations for very large values of T or K). The main challenge we face is that our procedure for generating timeseries data embedded with high-order correlations is very computationally expensive. The approach we used in our previous submission required storing *O*(T * K^(2^(n+1))) values (where *T* is the number of timepoints, *K* is the number of features, and *n* is the order). After tweaking our synthetic data generation procedure (revised methods, pages 23--24), we were able to reduce the memory footprint to *O*(T * K^(2^n)). While our revised procedure is substantially more efficient (while still achieving similar performance), it is still not scalable to very large values of K or n. For example, generating second-order timeseries data of the same size as our (reduced) fMRI data from a single participant would require nearly 300 TB of memory, which is well beyond the capacity of the computing resources available to our lab. This limits our ability to directly test our ability to recover ground truth patterns in datasets of a similar size to the fMRI dataset we examine in the main part of the manuscript.**

In addition to adding the above discussion points, we have added two supplemental analyses (Figs. S3 and S4) to directly examine the impacts of the number of samples (T) and the number of features (K) on our ability to recover ground truth first-order and second-order dynamic correlations from synthetic data. In summary, we found that as the number of samples increases, our ability to recover ground truth patterns improves (Fig. S3). This prediction is in line with the reviewer's prediction. Of note, when T is greater than between 250 and 500 samples (i.e., the range encompassed by our fMRI datasets), reconstruction accuracy starts to asymptote:



We also explored how reconstruction accuracy changes with the number of zero-order features (K). Up to the relatively small number of features we were able to test, our ability to recover ground truth patterns appears to be slightly impaired as the number of features increases (Fig. S4). We were somewhat surprised by this latter finding, since we had naively expected that recovery accuracy would *improve* with the number of features:



Exploring this issue more fully would require further optimizations and improvements to our synthetic data generation procedure. While this is beyond the scope of our current manuscript, it would be interesting to address in future work.

(b) What is the cause of the "dip" at the start and "ramp" at the end of the Order-2 curves in Figure 3—"Constant" and Figure 3—"Ramping"? Are these finite-sample effects arising from the change in size of a sliding window / kernel at the start and end of the synthetic data? If

so, please make this clear to the reader. Also, if these effects are present, might it make sense to "trim" the Order-2 empirical results to remove these finite-sample edge effects?

We have added a note to the manuscript (Fig. 3 caption) clarifying that the dips and ramps observed at the sharp transition points in our synthetic datasets reflect finite-sample "artifacts" as the reviewer notes. Nevertheless, due to the reasons outlined above (in response to the reviewer's previous comment), our sense is that we are still under-estimating recovery performance, even at sharp transition points. If these finite-sample issues were simply adding "noise" to our reconstruction procedure, then we would have expected reconstruction accuracy to be *worse* at boundaries (whereas in practice we observe that reconstruction accuracy is *higher* near boundaries).

A second set of reasons for not trimming the recovered high-order timeseries estimates in the empirical data relate to what we are trying to learn from the fMRI data. We are specifically interested in how decoding accuracy varies when we consider different orders of features. To fairly compare across features, all of the features should be based on the same underlying data, contain the same numbers of samples, etc. Further, we would run into a practical limitation whereby trimming the beginning and end of the recovered timeseries at each successively increasing order would lead us to "run out of data" at higher orders.

In acknowledgement of noise in our estimation procedure, we have included a note (p. 8--9, emphasis added) that "[t]his suggests that our modeling approach provides a meaningful (*if noisy*) estimate of high-order correlations".

[2] In my original review I asked two related questions [these were both part of Question 2 from my original review]:

(2a) Could it be the case that the "higher order correlation" method does better when there is a larger number of inter-subject reliable voxels? And (2b) Could the total number of reliable voxels per condition be confounded with the "optimal order" for decoding at that level?
In their response, the authors (as far as I can tell) answered (2a) but did not provide an answer to (2b). As they stated in their response: "For experimental condition, we want to know which orders of neural dynamics are reliable and stable across people." So my question in (2b) remains: whether the set of "which orders are reliable and stable" across conditions could be confounded (across conditions) with the total number of Order-1 reliable voxels. If the answer is "yes", I guess that's OK, but it seems that this potential confound should be made clear to the reader.

We have added a note to our discussion section to clarify this point (p. 15):

"One limitation of our approach relates to how noise propagates in our estimation procedure. Specifically, our procedure for estimating high-order dynamic correlations depends on estimates of lower-order dynamic correlations. This means that our measures of which higher-order patterns are reliable and stable across experimental conditions are partially confounded with the stability of lower-order patterns. Prior work suggests that the stability of what we refer to here as first-order dynamics likely varies across the experimental conditions we examined (Simony et al., 2016). Therefore a caveat to our claim that richer stimuli evoke more stable higher-order dynamics is that our approach assumes that those high-order dynamics reflect relations or interactions between lower-order features."

[3] The Introduction is still a little light on some of the (by now) classic work on network interactions and higher order cognition — the authors may wish to examine the two papers below (and related literature), to see whether the se papers usefully support or contextualize the central claim of this manuscript — that more complex inter-regional interaction patterns coincide with more complex cognitive function:

Coordination dynamics and cognition:
Bressler, S. L., & Kelso, J. S. (2001). Cortical coordination dynamics and cognition. Trends in cognitive sciences, 5(1), 26-36.

The concept of "neural context" [i.e. brain network context] for understanding high-level cognition:
McIntosh, A. R. (2000). Towards a network theory of cognition. Neural Networks, 13(8-9), 861-870.

We appreciate the pointers to these papers and have added citations to both of them in our revised introduction (p. 2) and discussion (p. 14).

Reviewer #3 (Remarks to the Author):

I thank the authors for thoroughly addressing my comments. The manuscript is much clearer as a result of the changes made. I do continue to have some reservations about the strength of the claims drawn from relatively small differences in classifier accuracy between models including features of different correlation orders (in Fig. 4). However, the authors have done a good job demonstrating the validity and theoretical contribution of their new approach, and I think this manuscript will make an important contribution to the literature.

(Nothing to address)