Comparing different criteria for neural dimensionality estimation

F.E. Vaccari (francesco.vaccari6@unibo.it)
Dept. of Biomedical and Neuromotor Sciences, University of Bologna, Italy
S. Diomedi (sdiomedi.work@gmail.com)
Institute of Cognitive Sciences and Technologies, National Research Council, Padova, Italy
E. Bettazzi (edoardo.bettazzi3@unibo.it)
Dept. of Biomedical and Neuromotor Sciences, University of Bologna, Italy
P. Fattori (patrizia.fattori@unibo.it)
Dept. of Biomedical and Neuromotor Sciences, University of Bologna, Italy

Abstract

Despite dimensionality reduction is essential in modern Neuroscience and Principal Component Analysis (PCA) continues to serve as the standard approach, in the field it is still missing a widely accepted criterion for choosing the number of components to retain. To fill this gap, we aimed to compare the performance of different retention criteria. We designed a data simulation procedure to generate data matrices with a ground-truth latent structure. Simulation parameters were varied to compare the different retention criteria in several scenarios. Among the tested criteria, Parallel Analysis and a cross-validation scheme, specifically conceived for dimensionality reduction, resulted to be the most effective methods. Finally, by applying these criteria to real spiking activity, we show that different criteria can lead to significantly different results in the estimation of dimensionality and noise. Our study highlights the need for an explicit definition of “dimensionality” in the analysis of population spike activity and a consequent careful choice of the retention criterion to be used, as this can lead to important biases and non-comparable results between studies.

Keywords: dimensionality reduction; neural dynamics; data simulation; parietal cortex; spiking activity

Introduction

To analyse the high-dimensional neural datasets available today, dimensionality reduction techniques are powerful tools to obtain information on latent neural dynamics and ongoing brain computations. In this regard, although non-linear algorithms provide the best performance, in many cases their complexity limits a widespread use and linear techniques, such as PCA or its derivations, remain a popular choice due to simplicity and interpretability.
However, the choice of the number of latent components to consider is far from trivial and still not supported by robust consensus in the field literature.

To address this issue, we compared different criteria for choosing how many latent dynamics to retain (thus excluding those that depend only on noise) when applying PCA derived from the neuroscience literature, but also from other fields.

Methods

Neural data.

Two types of data were considered: I) simulated data, providing a ground-truth used to compute various performance metrics and directly compare the different component retention criteria; II) real-world spiking data recorded from macaque cortex during a reaching task used to highlight the differences in results.

Simulations. A data simulation procedure was designed based on linear combinations of latent variables (correlated signals across synthetic units) and Gaussian noise addition (uncorrelated signals). The initial set of latent variables was either generated with an iterative random process or starting from real-world principal components calculated on parietal spiking activity (see next paragraph). Many parameters were varied to simulate different scenarios (number of synthetic units, noise amount etc). Note that the criteria to be tested and PCA were based on linear models, we did not add any non-linear step in the simulation procedure.

Real spiking activity. An online dataset was the source of real-world neural activity (Diomedi et al., 2024). Spikes were collected during a reaching task from the posterior parietal cortex of macaque. Activity was binned every 50ms, averaged for each target, soft-normalized and mean-centered. Only data collected during the initial rest period and during movement execution were considered.

Criteria comparison.

Retention criteria. Hard explained variance thresholds (80 and 90%) and Participation Ratio (PR; Gao et al., 2017) were considered for the widespread usage in Neuroscience. From other fields' literature, the Kaiser rule (K1, Kaiser, 1960) and the Parallel Analysis (PA, Horn, 1965) are known to perform well. Finally, we implemented a specific cross-validation (CV) scheme by removing elements along both rows (time points) and columns (neurons) of the data matrix.

Performance metrics. To account for properties useful in real-world cases, the following performance metrics we computed for each criteria: i) Dimensionality error, i.e. difference between the dimensionality estimated by the criteria and the ground-truth from the simulation; ii) Reconstruction accuracy i.e R2 between the data matrix reconstructed with a chosen number of components and the matrix of correlated signals, generated during the simulation procedure (before noise addition); iii) Estimated noise error, i.e. difference between the variance unexplained by the generated latent variables (ground-truth noise amount) and the variance unexplained by the chosen PCs (noise estimated by the used criterion).

Results

Simulation results.

Parallel Analysis and cross-validation resulted to be the most effective methods. Indeed, they scored the lowest error in both dimensionality and noise amount estimation as well as the most accurate reconstruction of the correlated activity (Figure 1A, C and B respectively).

Both were poorly affected by the noise regime, i.e. they did not tend to incorporate more components when the data became noisier as the other criteria did (see almost flat vs increasing lines in Figure 1A).

Real data results.

Different criteria led to different trends when applied to real-world spiking activity. Indeed, the dimensionality assessed using PR decreased from rest to movement phase from 18 to 11 components for V6A area and from 11 to 7 for PEc. Instead, when the dimensionality was assessed using PA, the trend was much less marked or even absent: for V6A the dimensionality decreased only from 7 to 6 components, while for PEc was stable at 4 components.

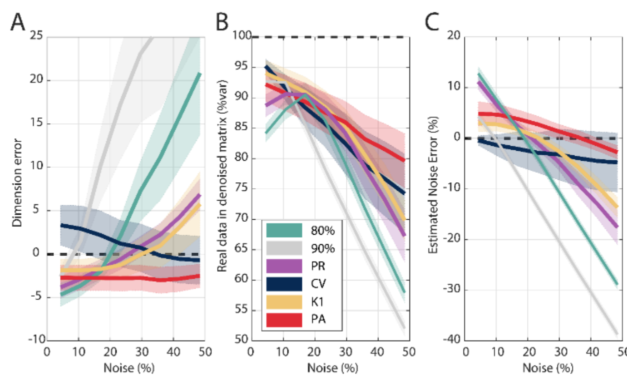

Figure 1: Performance of retention criteria tested on simulated data as a function of noise. Horizontal dashed lines represent optimal performance.

Discussion and Conclusion

The study aimed at comparing different criteria that estimate the number of principal components and showing how much this choice can bias neural analysis. Hard explained variance thresholds and PR are the most used in Neuroscience, but their results were extremely affected by the noise as well as other factors, such as the number of units in the population (data not shown here). Among the other tested criteria, Parallel Analysis and cross validation were the least influenced by noise and the number of neurons, representing promising tools to further applications to real-world data. Our findings add support for the future use of more robust methods that could be routinely used in neural data analysis.

Related work

The preprint of the study is available at: F. E. Vaccari, S. Diomedi, E. Bettazzi, M. Filippini, M. De Vitis, K. Hadjidimitrakis, P. Fattori. (2024). bioRxiv https://doi.org/10.1101/2024.11.28.625854

Acknowledgements

The work was supported by: #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) – A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022); Ministry of University and Research (MUR), PRIN2022-2022BK2NPS; grant H2020-EIC-FETPROACT-2019-951910-MAIA.

References

Diomedi, S., Vaccari, F. E., Gamberini, M., De Vitis, M., Filippini, M., & Fattori, P. (2024). Neurophysiological recordings from parietal areas of macaque brain during an instructed-delay reaching task. *Scientific data, 11(1), 624.* https://doi.org/10.1038/s41597-024-03479-7

Gao, P., Trautmann, E., Yu, B., Santhanam, G., Ryu, S., Shenoy, K., & Ganguli, S. (2017). A theory of multineuronal dimensionality, dynamics, and measurement. *bioRxiv.* https://doi.org/10.1101/214262

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30(2),179–185.* https://doi.org/10.1007/BF02289447

Kaiser, H. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20, 141–151.*