LETTER TO THE EDITOR

Principal component analysis-based latent-space dimensionality under-estimation, with uncorrelated latent variables

Thomas M. H. Hope,[1,2] Ajay Halai,[3] Jenny Crinion,[4] Paola Castelli,[2] Cathy J. Price[1] and Howard Bowman[5]

Author affiliations:

1 Wellcome Centre for Human Neuroimaging, Department of Imaging Neuroscience, Institute of Neurology, University College London, London, WC1N 3AR, UK

2 Department of Psychology, John Cabot University, 00165, Rome, Italy

3 MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge CB2 7EF, UK

4 Institute of Cognitive Science, Department of Experimental Psychology, University College London, London, WC1N 3AR, UK

5 School of Psychology, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

Correspondence to: Thomas Hope

Wellcome Centre for Human Neuroimaging, Department of Imaging Neuroscience, Institute of Neurology, University College London, 12 Queen Square, London, WC1N 3AR, UK

E-mail:t.hope@ucl.ac.uk

In many scientific disciplines, features of interest cannot be observed directly, so must instead be inferred from observed behaviour. In the study of the damaged brain, those 'features of interest' might be the function or disruption of dissociable cognitive sub-systems, and the 'observed behaviour' might be accuracies and / or reaction times recorded in standardised, behavioural tasks. This inverse inference from observed data to features of interest is increasingly approached using latent variable analyses[1–4]. One of the simplest and most popular of these methods, is Principal Components Analysis (PCA). During the last decade, stroke outcomes research, using PCA, has yielded a surprising result: latent spaces appear lower-dimensional than expected. These analyses typically find no more than 5 latent variables, and sometimes just 1[1–4], even when applied to scores from wide-ranging batteries of tasks, which could potentially capture impairments to many more dissociable sensory, motor and cognitive sub-systems.

Recently, this apparent 'dimensionality under-estimation problem' has been explained as potentially arising from spatial correlations in natural stroke lesion distributions[5]. The authors used simulated data derived from real stroke induced lesions, in which impairment severity scores were assigned based on the extent of damage to non-overlapping brain regions. Since the regions were independent, the impairments should have been independent: i.e., the latent-space dimensionality should have been the same as the number of simulated scores. But instead, the authors observed that PCA typically found lower-dimensional latent spaces, because natural stroke-induced lesions tend to damage neighbouring (non-overlapping) brain regions together, causing the impairments to be correlated in practice even though they need not have been correlated in theory[5]. The implication is that PCA-based analyses of stroke outcomes data might tell us as much about lesion distributions as they ever can about the fundamental organisation of cognition.

Here, we show that dimensionality under-estimation can occur entirely regardless of lesion distributions – even when post-stroke impairments are independent by construction. We show that this effect is partly a function of task impurity, the extent to which behavioural performance in individual tasks is thought to emerge from the interaction of many different cognitive skills. And we show that dimensionality under-estimation can be ameliorated by employing more multivariate behavioural data (i.e., more tasks).

Materials and methods

We use PCA to analyse synthetic, multivariate behavioural data, which are linear mixtures of known latent variable values. No lesion data were included. Following the approach employed by Sperber and colleagues, we count the components derived by PCA as those whose eigenvalues surpass a threshold and employ two different thresholds: the Kaiser criterion (threshold eigenvalue = 1)[6], and the Jolliffe criterion (threshold eigenvalue = 0.7)[7]. The more conservative Kaiser criterion is more popular, in our experience, but the more permissive

Jolliffe criterion might be more appropriate when we expect to observe dimensionality under-estimation. Both latent variable values and latent-to-behaviour weights are defined as random uniform numbers in the range 0-1 (e.g., imagining both numbers to represent percentages of maximum function / influence). We did consider other types of random distribution, but none made any substantive difference to our results. And following the prior report, we employ a sample size of 300. Our simulations vary the number of latent variables in the range 1-22, and the number of behavioural scores in the range 22-100. We ran 1,000 simulations per parameter configuration, randomly re-specifying latent variable values and latent-to-behavioural weights each time, and report summary results.

Results

Analysis 1: Under-estimation with uncorrelated impairments

Figure 1A illustrates how the derived dimensionality of the system varies with its real dimensionality, for a fixed number of (22) behavioural scores. Estimated dimensionality is mostly accurate for systems with just 1 or 2 latent variables, but then grows less quickly than real dimensionality – and indeed begins to fall again for higher-dimensional systems. Naturally, the effect is more pronounced when using the more conservative, Kaiser criterion to count derived components. Figure 1B illustrates dimensionality estimation in identical circumstances to those considered in Figure 1A, with one exception: 95% of the latent-to-behavioural weights are set to 10^{-6}. This change effectively ensures that most behavioural variables are mediated by fewer latent variables than before: i.e., task impurity is reduced. In this case, the relationship between derived and real latent system dimensionality is more intuitive, in that both grow together. However, derived dimensionality still only grows about half as quickly as real dimensionality, so dimensionality under-estimation still occurs.

Analysis 2: Under-estimation is avoided if we use many more behavioural tasks

Figure 2 illustrates how dimensionality estimation changes as the number of behavioural tasks grows. With ~6 times as many behavioural tasks as real latent variables, PCA can enumerate the real latent system accurately, when using the Joliffe criterion (~10 times for the Kaiser criterion). The under-estimation problem returns as the number of latent variables increases further.

Discussion

Our results suggest that dimensionality under-estimation might occur, in stroke research and beyond, as a simple artefact of task impurity – or more generally, the notion that important relationships between latent and observed variables might be many-to-many – even when the latent variables themselves are uncorrelated. This problem appears to worsen as real latent system dimensionality increases. In data derived from three or more latent variables, PCA was an unreliable way to enumerate those latent variables unless: (a) task impurity was low (i.e., latent-to-behavioural weight matrices were sparse); or (b) there were many more behavioural tasks than real latent variables in the system.

Quite how these results apply in practice, is hard to judge. First, if the latent system is 'cognition', then we should probably allow that it might be higher-dimensional than any latent system considered here. But at the same time, the *effective* dimensionality of post-stroke impairments, as represented in any given sample of stroke patients, might be much lower than this theoretical maximum. For example, if all of the patients in a sample have the same, selective impairment, then the sample is one-dimensional. Moreover, the absolute range of impairment severity might appear smaller in stroke patient samples than in the wider patient population, either because standardised measures of that severity lack sensitivity, or because the most severely impaired patients might struggle to travel to study sites, or tolerate the testing process itself[8]. These factors make lower-dimensional estimates plausible. But since we might observe lower-dimensional estimates even when they are wrong (Analysis 1), our only rational choice is to treat the results of all analyses like this with caution. And further caution is called for because task impurity is likely to be the norm rather than the exception in these studies[9].

Our results also point to a practical solution for reducing dimensionality under-estimation: using many more tasks than there are latent variables to find (Analysis 2). In practice, the required ratio of tasks to variables might well vary from that observed here. From the perspective of latent-space dimensionality estimation, real latent-to-behavioural weights might be less efficient than those used here, so that more tasks are needed to count latent dimensions accurately. On the other hand, latent-to-behaviour weights might be more informative than those considered here, because batteries of tasks used in stroke research are often expressly designed to vary the engagement of cognitive functions systematically and informatively. Since our simulations only very rarely over-estimated real system dimensionality, one might navigate this issue by adding tasks incrementally to PCA, until further additions no longer yield more principal components. But of course, this approach is the opposite of what many researchers might prefer, given the costs and effort required to employ extra tasks[8].

Notably, our results are also at least potentially consistent with Sperber and colleagues' account of latent space dimensionality under-estimation, at least in stroke outcomes research. They highlighted spatial correlations in natural stroke-induced lesion distributions. This factor might operate in tandem with the task impurity on which our analyses are focused. We hope that our results will encourage caution in the interpretation of results, derived from post-stroke impairment severity score batteries, via PCA.

Data availability

No empirical data were used in the reported analyses. MATLAB scripts used to run the analyses can be downloaded from: https://github.com/tmhhopegit/pca_dim_under-estimation

Competing interests

The authors report no competing interests.

Funding

This work was supported by the Wellcome Trust (205103/Z/16/Z and 203147/Z/16/Z to C.J.P.); and the Medical Research Council (MR/V031481/1 to A.D.H.).

References

1. Halai AD, Woollams AM, Lambon Ralph MA. Using principal component analysis to capture individual differences within a unified neuropsychological model of chronic post-stroke aphasia: Revealing the unique neural correlates of speech fluency, phonology and

semantics. *Single Brain Model Sufficient*. 2017;86:275-289. doi:10.1016/j.cortex.2016.04.016

2. Ramsey LE, Siegel JS, Lang CE, Strube M, Shulman GL, Corbetta M. Behavioural clusters and predictors of performance during recovery from stroke. *Nat Hum Behav*. 2017;1(3):0038. doi:10.1038/s41562-016-0038

3. Schumacher R, Halai AD, Lambon Ralph MA. Assessing and mapping language, attention and executive multidimensional deficits in stroke aphasia. *Brain J Neurol*. 2019;142(10):3202-3216. doi:10.1093/brain/awz258

4. Akkad H, Hope TMH, Howland C, et al. Mapping spoken language and cognitive deficits in post-stroke aphasia. *NeuroImage Clin*. 2023;39:103452. doi:10.1016/j.nicl.2023.103452

5. Sperber C, Gallucci L, Umarova R. The low dimensionality of post-stroke cognitive deficits: it's the lesion anatomy! *Brain*. 2023;146(6):2443-2452. doi:10.1093/brain/awac443

6. Kaiser HF. The application of electronic computers to factor analysis. *Educ Psychol Meas*. 1960;20(1):141-151.

7. Jolliffe IT. Discarding variables in a principal component analysis. I: Artificial data. *J R Stat Soc Ser C Appl Stat*. 1972;21(2):160-173.

8. Halai AD, De Dios Perez B, Stefaniak JD, Lambon Ralph MA. Efficient and effective assessment of deficits and their neural bases in stroke aphasia. *Cortex*. 2022;155:333-346. doi:10.1016/j.cortex.2022.07.014

9. Burgess PW. Theory and methodology in executive function research. In: *Methodology of Frontal and Executive Function*. Routledge; 2004:87-121.

Figure legends

Figure 1 Dimensionality under-estimation. Both panels illustrate the mean and standard deviation of derived dimensionality when analysing 22 behavioural scores, derived from latent systems including 1-22 latent variables. Accurate dimensionality estimation occurs when the lines intersect the diagonal of either panel, where estimated dimensionality equals real system dimensionality. In panel A, latent-to-behaviour weights are random uniform numbers in the range 0-1. In this case, dimensionality estimation is mostly accurate for systems with 1-3 latent variables, but then becomes less accurate as real dimensionality increases. In panel B, 95% of the weights are set to 10^{-6}, reducing task impurity, so that estimated dimensionality grows as real dimensionality grows (albeit that the former only grows about half as quickly as the latter).

Figure 2 Latent space dimensionality estimation as the numbers of tasks increases from 40 to 100. As in Figure 1, derived dimensionality equals the real dimensionality of the system along the dotted diagonal line. The number of systems that can be accurately estimated increases with the number of task scores. Once the real dimensionality increases beyond ~1/6 of the number of task scores (~1/10 for Kaiser criterion), derived dimensionality again under-estimates the true the dimensionality of the system: i.e., the estimation problem returns. Panels A and B illustrate the effect when using the Joliffe and Kaiser criteria, respectively: for a given real system dimensionality, more tasks are required for accurate dimensionality estimation when using the more conservative (Kaiser) criterion to count those estimated dimensions.

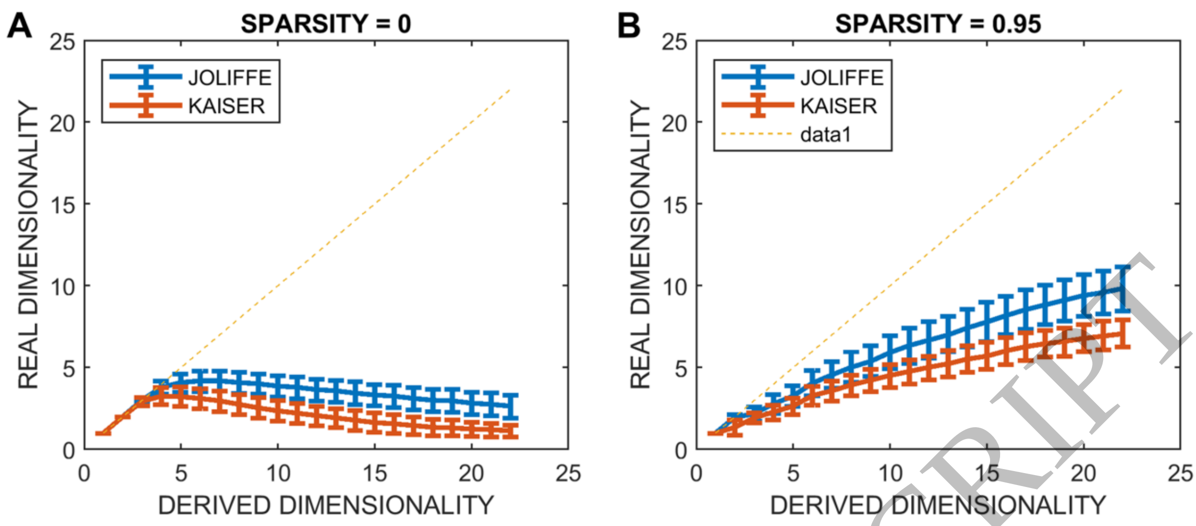

Figure 1
159x69 mm (x DPI)

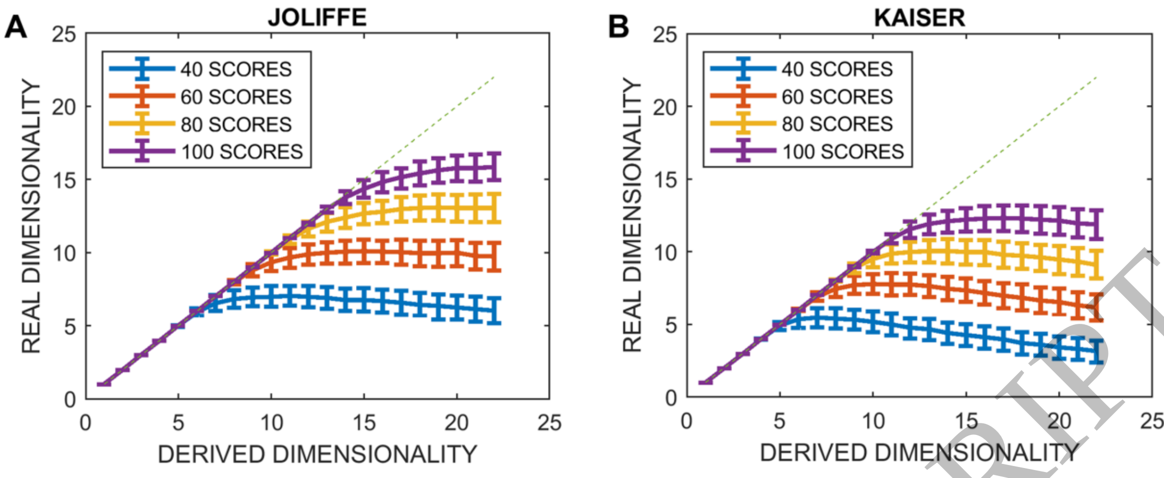

Figure 2
159x62 mm (x DPI)