Validating the impact of sample size for modelling brain and behaviour interactions with canonical correlation analysis

Michelle Wang, Brent McPherson, Jérôme Dockès, Nikhil Bhagwat, Jean-Baptiste Poline

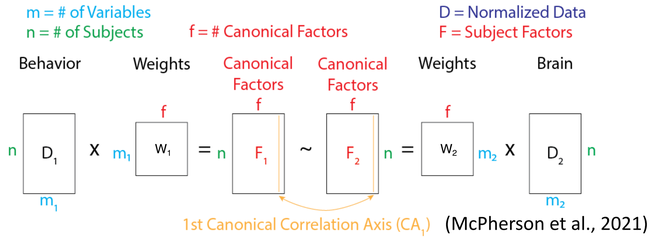
NeuroDataScience - ORIGAMI laboratory, McGill University, Montreal, Canada

Introduction

Reproducibility crisis in neuroimaging

- Large-scale studies to increase statistical power
- Deep phenotyping of behaviour and genetic data

Canonical correlation analysis (CCA)



(McPherson et al., 2021)

- Useful for describing how measures from different domains (e.g., brain/behaviour) covary
- Recent work suggests that 1000s of participants are needed for to obtain reproducible results (Marek et al., 2022)

Methods

Data

- ~40,000 subjects from the UK Biobank
- **Behavioural measures**: cognitive function
  - E.g., numeric memory, fluid intelligence
- **Brain measures**: phenotypes derived from diffusion magnetic resonance imaging (dMRI)
  - E.g., mean fractional anisotropy in main white matter structures

Bootstrap sampling

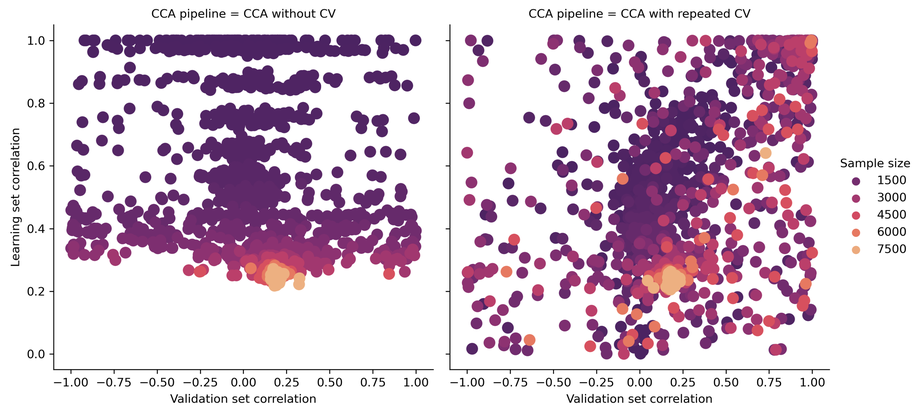
- 100 bootstrap repetitions
- Split dataset into Learning and Validation halves
  - Use subsets of Learning half with 16 increasing sample sizes (203 to 7686, log-spaced)

Preliminary results

Comparison between different CCA analysis pipelines

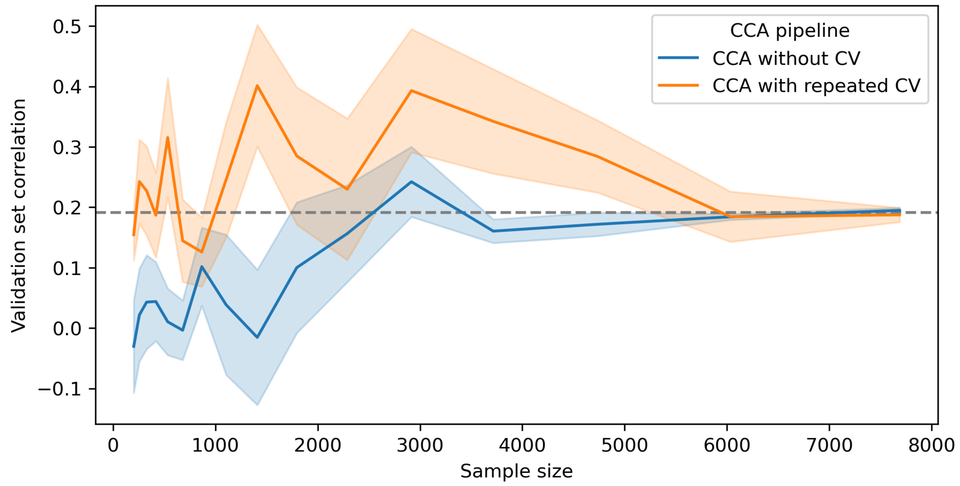
Tested two CCA pipelines

1. CCA without cross-validation (CV): fit a single model on the Learning set
2. CCA with repeated cross-validation: fit 50 models, take central tendency of canonical factors (McPherson et al., 2021)



Validation correlations stabilize as sample size increases

At low sample sizes, CCA with repeated CV yields higher validation correlations than CCA without CV.



Error band: 2 standard errors of the mean

Discussion

Impact of sample size on brain-behaviour CCA results

- As sample size increases, there is less overfitting of the data: Learning set correlation decreases, validation set correlation variability decreases

Future directions

- Investigate why CCA with repeated cross-validation is noisier and sometimes produces Learning/Validation correlations close to 1
- Analyze variable loadings from CCA models with different sample sizes for each of the pipelines: is the relative strength of CCA loadings preserved?

References

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654–660. https://doi.org/10.1038/s41586-022-04492-9

McPherson, B. C., & Pestilli, F. (2021). A single mode of population covariation associates brain networks structure and behavior and predicts individual subjects' age. *Communications Biology*, *4*(1), 1–16. https://doi.org/10.1038/s42003-021-02451-0

Acknowledgements

This work was partially funded by the Michael J. Fox Foundation, National Institutes of Health (NIH), the National Institute Of Mental Health, the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains for Healthy Lives initiative and the Brain Canada Foundation with support from Health Canada, through the Canada Brain Research Fund in partnership with the Montreal Neurological Institute.