Canonical Correlation Analysis in high dimensions with structured regularization

Elena Tuzhilina

(joint work with Trevor Hastie and Leonardo Tozzi)

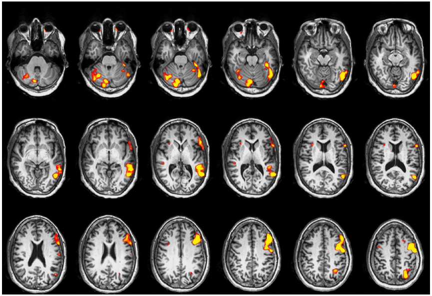
Stanford University, Department of Statistics

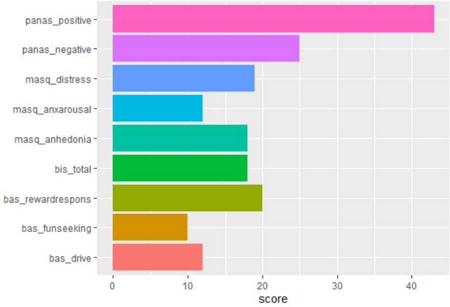
elenatuz@stanford.edu

August 12, 2021

Data

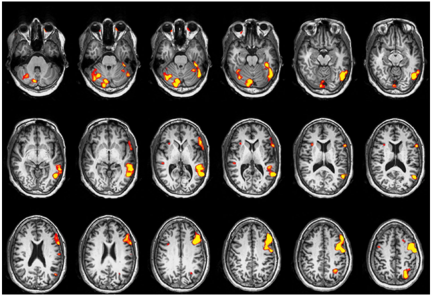
1. **brain activations**: magnetic resonance imaging obtained during a gambling task designed to probe the brain circuits underlying reward
2. **behavioral performance measures**: self-reports assessing various aspects of reward-related behaviors, depression symptoms and positive as well as negative affective states



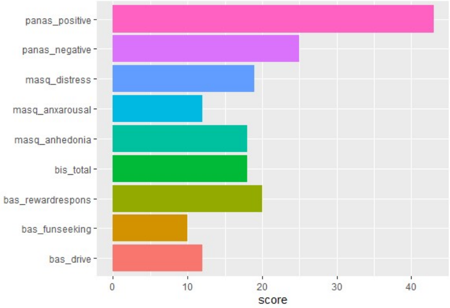


Question: is there any correlation between brain activity and behavioral measures of performance during the gambling tasks?

Notations



$X \in \mathbb{R}^{n \times p}$ – brain activations



$Y \in \mathbb{R}^{n \times q}$ – behavior scores

Dimensions:

- $n = 153$ participants
- $p = 90,368$ greyordinates
- $q = 9$ scores

Canonical Correlation Analysis

Goal: find $\alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^q$ such that

$$\text{maximize } \rho(\alpha, \beta) = \text{cor}(X\alpha, Y\beta) = \frac{\alpha^T \Sigma_{XY} \beta}{\sqrt{\alpha^T \Sigma_{XX} \alpha}\,\sqrt{\beta^T \Sigma_{YY} \beta}}$$

- canonical coefficients α and β
- canonical correlation $\rho(\alpha, \beta) = \text{cor}(u, v)$

Find maximum via Singular Value Decomposition of $\Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}}$

Regularization

Problem: Σ_{XX} is singular when $p > n$, so CCA does not work.

Solution: adjust the diagonal of Σ_{XX}.

Modified correlation coefficient

$$\rho(\alpha, \beta) = \frac{\alpha^T \Sigma_{XY} \beta}{\sqrt{\alpha^T (\Sigma_{XX} + \lambda_1 I) \alpha} \sqrt{\beta^T \Sigma_{YY} \beta}}$$

Find maximum via Singular Value Decomposition of $(\Sigma_{XX} + \lambda_1 I)^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}}$

RCCA shrinkage property: $\lambda_1 \uparrow$ shrinks $\alpha \to 0$, similar to ridge regression.

Kernel trick

```
library(CCA)
rcc(X = activation, Y = behavior, lambda1 = 10, lambda2 = 0)
```

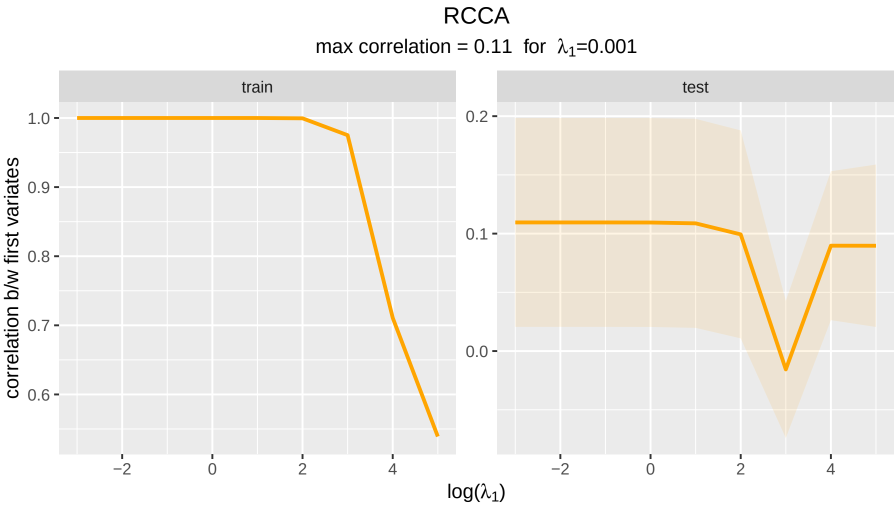
```
Error: cannot allocate vector of size 62.1 Gb
Traceback:

1. rcc(X = activation, Y = behavior, lambda1 = 10, lambda2 = 0)
2. var(X, na.rm = TRUE, use = "pairwise")
```

Problem: Σ_{XX} is $p \times p$, so cannot store for $p \approx 90K$.

Solution: find a linear transformation such that RCCA for (X, Y) is equivalent to RCCA for (R, Y) and

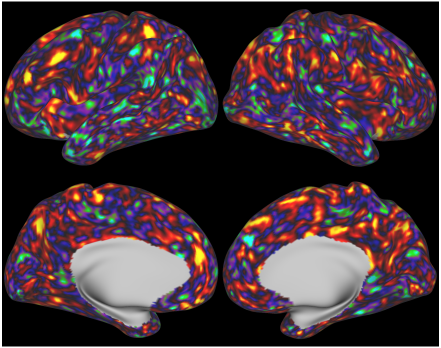
$$V = \boxed{p \times n} \qquad R = XV = \boxed{n \times p}\,\boxed{p \times n} = \boxed{n \times n}$$

RCCA preformance

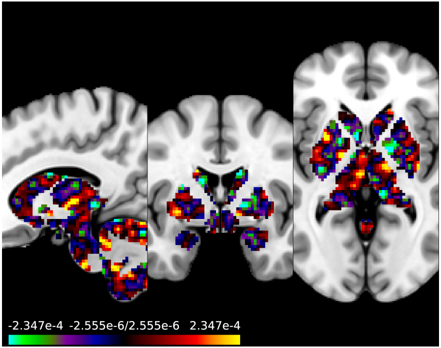


RCCA interpretability

Visualization: plot canonical coefficients α for the optimal RCCA model.



(a) Cortical coefficients.

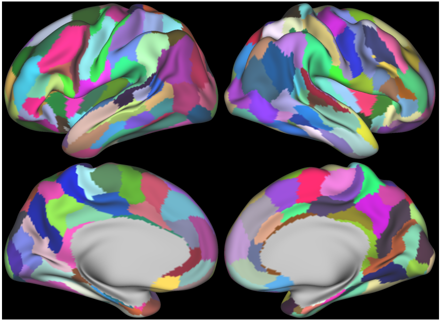


(b) Subcortical coefficients.

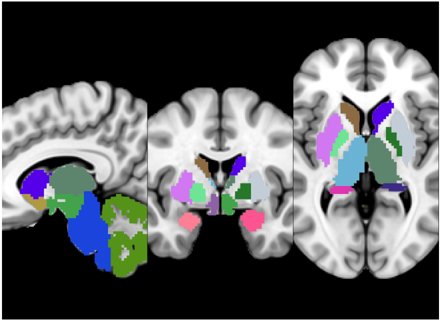
Problem: There is a lot of variation, so this plot is hard to interpret.

Group structure

Motivation: brain features come in groups (aka brain regions). How to take into account the group structure?



(a) Cortical parcellation (210 regions).



(b) Subcortical parcellation (19 regions).

Grouped structure

Block structure:

$$X = (\underbrace{X_1}_{region_1}, \ldots, \underbrace{X_K}_{region_K}) \text{ and } \alpha = (\underbrace{\alpha_1}_{region_1}, \ldots, \underbrace{\alpha_K}_{region_K})$$

Assumptions:

1. group homogeneity $\alpha_k \approx \bar{\alpha}_k$
2. sparsity on a group level $\bar{\alpha}_k \approx 0$

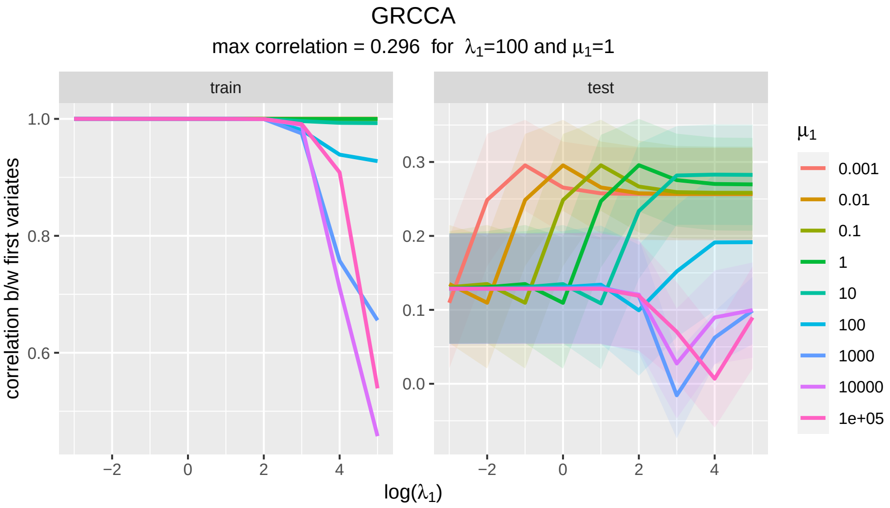
$$C = \begin{bmatrix} \frac{11^T}{p_1} & 0 & \ldots & 0 \\ 0 & \frac{11^T}{p_2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \frac{11^T}{p_K} \end{bmatrix}$$

Modified correlation coefficient

$$\rho(\alpha, \beta) = \frac{\alpha^T \Sigma_{XY} \beta}{\sqrt{\alpha^T(\Sigma_{XX} + \lambda_1(I - C) + \mu_1 C)\alpha} \sqrt{\beta^T \Sigma_{YY} \beta}}$$

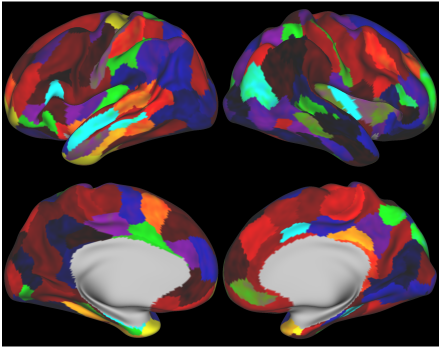
GRCCA shrinkage property: $\lambda_1 \uparrow$ shrinks $\alpha_k \to \bar{\alpha}_k$,
$\mu_1 \uparrow$ shrinks $\bar{\alpha}_k \to 0$.

GRCCA performance

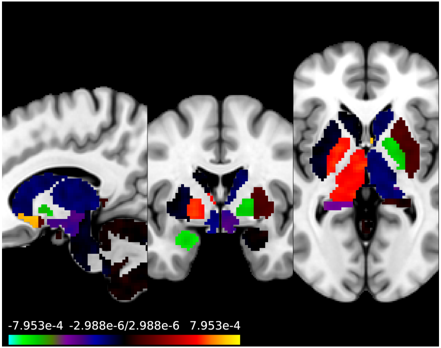


GRCCA interpretability

Visualization: plot canonical coefficients α for the optimal GRCCA model.



(a) Cortical coefficients.



(b) Subcortical coefficients.

References

D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, D. Nolan, E. Bryant, T. Hartley, O. Footer, J. M. Bjork, R. Poldrack, S. Smith, H. Johansen-Berg, A. Z. Snyder, and D. C. Van Essen, *Function in the human connectome: Task-fMRI and individual differences in behavior*, NeuroImage, 80 (2013), pp. 169–189.

D. C. V. Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, and W.-M. H. Consortium, *The WU-Minn Human Connectome Project: an overview*, Neuroimage, 80 (2013), pp. 62–79.

M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. V. Essen, *A multi-modal parcellation of human cerebral cortex*, Nature, 536 (2016), pp. 171–178.

I. Gonzalez, D. S, P. Martin, and A. Baccini, *CCA: An R Package to Extend Canonical Correlation Analysis*, Journal of Statistical Software, 23(12) (2008), pp. 1–14.

H. Hotelling, *Relations between two sets of variables*, Biometrika, 28 (1936), pp. 321—-377.

W. Härdle and L. Simar, *Applied Multivariate Statistical Analysis: Second edition*, Springer, 2007.

S. Leurgans, R. Moyeed, and B. Silverman, *Canonical Correlation Analysis when the Data are Curves*, Journal of the Royal Statistical Society, 55(3) (1993), pp. 725—-740.

K. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, New York, NY, 1979.

H. Vinod, *Canonical Ridge and Econometrics of Joint Production*, Journal of Econometrics, 4(2) (1976), pp. 147—-166.

S. Weintraub, S. S. Dikmen, R. K. Heaton, D. S. Tulsky, P. D. Zelazo, P. J. Bauer, N. E. Carlozzi, J. Slotkin, D. Blitz, K. Wallner-Allen, N. A. Fox, J. L. Beaumont, D. Mungas, C. J. Nowinski, J. Richler, J. A. Deocampo, J. E. Anderson, J. J. Manly, B. Borosh, R. Havlik, K. Conway, E. Edwards, L. Freund, J. W. King, C. Moy, E. Witt, and R. C. Gershon, *Cognition assessment using the NIH Toolbox*, Neurology, 80 (2013), pp. S54–64.

Thank you for your attention!