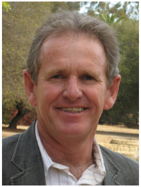
Canonical Correlation Analysis in high dimensions with structured regularization

Elena Tuzhilina

Department of Statistical Sciences, University of Toronto

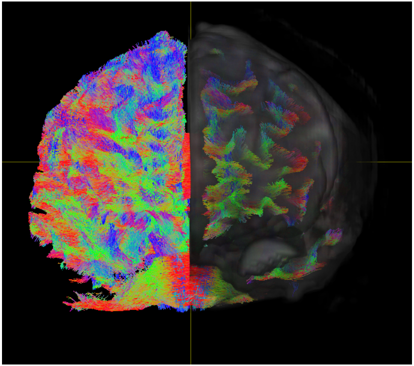
elena.tuzhilina@utoronto.ca



Trevor Hastie

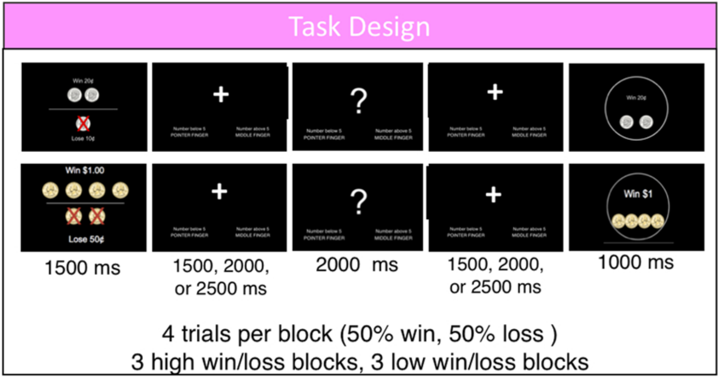
Leonardo Tozzi

Human Connectome Project for Disordered Emotional States aims to link the function of macroscopic human brain circuits to self-reports of emotional well-being using magnetic resonance imaging.



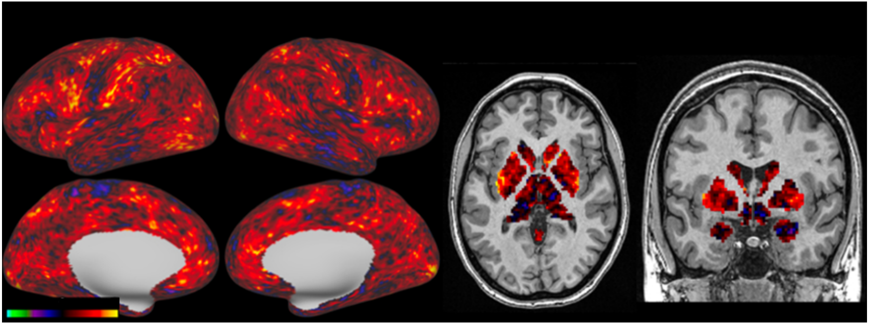
Leanne Williams

Gambling task: A question mark is displayed on the screen and the participant must guess whether a number is greater than or less than five. If the participant identifies correctly, they win money, and if they guess incorrectly, they lose money. At the end of the task, 5 trials are randomly selected and summed together to determine the participant's payment



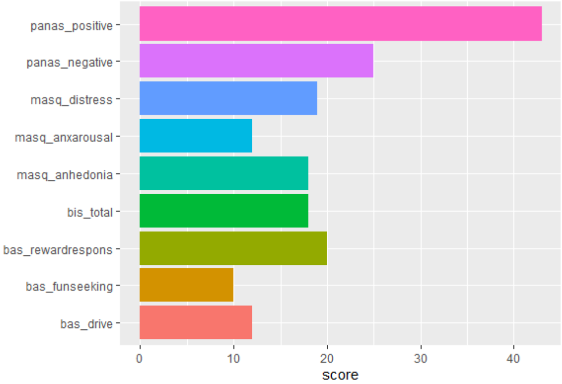
Data

Brain activations (90,368 greyordinates, cortical and sub-cortical): magnetic resonance imaging obtained during a Gambling task designed to probe the brain circuits underlying reward

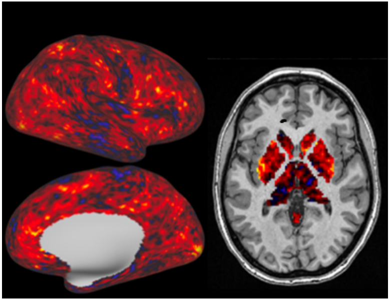


Data

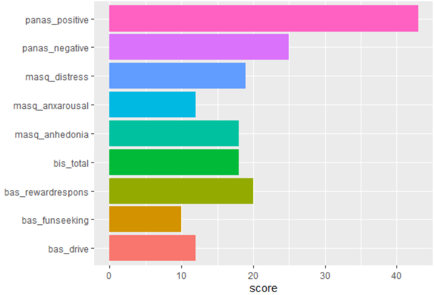
Behavioral performance measures (9 scores): self-reports assessing various aspects of reward-related behaviors (Behavioral Approach System/Behavioral Inhibition Scale), depression symptoms (Mood and Anxiety Symptom Questionnaire) and positive as well as negative affective states (Positive and Negative Affect Schedule)



Notations



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

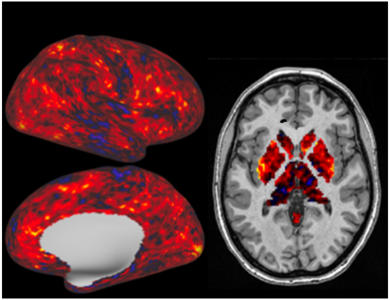


$Y \in \mathbb{R}^{n \times q}$ – behavioral scores

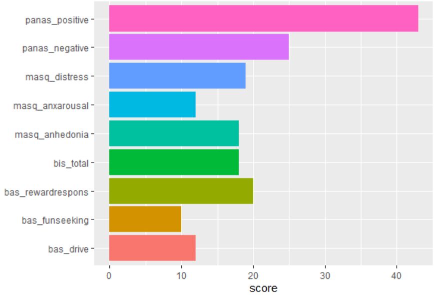
Dimensions:

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

Notations



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



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

Dimensions:

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

Question: is there any correlation between brain activations and behavioral scores?

Canonical Correlation Analysis

CCA is a classic method commonly used in statistics for finding association between two sets of variables.

Canonical Correlation Analysis

CCA is a classic method commonly used in statistics for finding association between two sets of variables.

Goal: given two random vectors $x = (x_1, \ldots, x_p)$ and $y = (y_1, \ldots, y_q)$

$$\text{maximize } \operatorname{cor}(\alpha^\top x, \beta^\top y) \text{ w.r.t. } \alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^q$$

Canonical Correlation Analysis

CCA is a classic method commonly used in statistics for finding association between two sets of variables.

Goal: given two random vectors $x = (x_1, \dots, x_p)$ and $y = (y_1, \dots, y_q)$

$$\text{maximize } \operatorname{cor}(\alpha^\top x, \beta^\top y) \text{ w.r.t. } \alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^q$$

- canonical coefficients α and β
- canonical variates $u = \alpha^\top x$ and $v = \beta^\top y$
- canonical correlation $\operatorname{cor}(u, v)$

Canonical Correlation Analysis

CCA is a classic method commonly used in statistics for finding association between two sets of variables.

Goal: given two random vectors $x = (x_1, \ldots, x_p)$ and $y = (y_1, \ldots, y_q)$

$$\text{maximize } \text{cor}(\alpha^\top x, \beta^\top y) \text{ w.r.t. } \alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^q$$

Actually, it is possible to find a sequence...

For $i = 1, \ldots, \min(p, q)$ find $\alpha_i \in \mathbb{R}^p$ and $\beta_i \in \mathbb{R}^q$ that

1. maximize $\text{cor}(\alpha_i^\top x, \beta_i^\top y)$
2. have independent variates $u_i = \alpha_i^\top x$ and $v_i = \beta_i^\top y$,
   i.e. $u_i \perp\!\!\!\perp (u_1, \ldots, u_{i-1})$ and $v_i \perp\!\!\!\perp (v_1, \ldots, v_{i-1})$

CCA solution

Correlation coefficient

$$\mathrm{cor}(\alpha^\top x, \beta^\top y) = \frac{\alpha^\top \mathrm{cov}(x, y)\beta}{\sqrt{\alpha^\top \mathrm{var}(x)\alpha}\,\sqrt{\beta^\top \mathrm{var}(y)\beta}}$$

CCA solution

Correlation coefficient

$$\text{cor}(\alpha^\top x, \beta^\top y) \approx \rho(\alpha, \beta) = \frac{\alpha^\top \Sigma_{XY} \beta}{\sqrt{\alpha^\top \Sigma_{XX} \alpha}\,\sqrt{\beta^\top \Sigma_{YY} \beta}}$$

CCA solution

Correlation coefficient

$$\text{cor}(\alpha^\top x, \beta^\top y) \approx \rho(\alpha, \beta) = \frac{\alpha^\top \Sigma_{XY} \beta}{\sqrt{\alpha^\top \Sigma_{XX} \alpha}\,\sqrt{\beta^\top \Sigma_{YY} \beta}}$$

CCA optimization problem:

$$\text{maximize } \alpha^\top \Sigma_{XY} \beta \text{ w.r.t. } \alpha \in \mathbb{R}^p \text{ and } \beta \in \mathbb{R}^q$$
$$\text{s.t. } \alpha^\top \Sigma_{XX} \alpha = 1 \text{ and } \beta^\top \Sigma_{YY} \beta = 1$$

CCA solution

Correlation coefficient

$$\mathrm{cor}(\alpha^\top x, \beta^\top y) \approx \rho(\alpha, \beta) = \frac{\alpha^\top \Sigma_{XY} \beta}{\sqrt{\alpha^\top \Sigma_{XX} \alpha}\,\sqrt{\beta^\top \Sigma_{YY} \beta}}$$

CCA optimization problem:

$$\begin{aligned} &\text{maximize } \widetilde{\alpha}^\top \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}} \widetilde{\beta} \text{ w.r.t. } \widetilde{\alpha} \in \mathbb{R}^p \text{ and } \widetilde{\beta} \in \mathbb{R}^q \\ &\text{s.t. } \|\widetilde{\alpha}\| = 1 \text{ and } \|\widetilde{\beta}\| = 1 \end{aligned}$$

CCA solution

Correlation coefficient

$$\mathrm{cor}(\alpha^\top x, \beta^\top y) \approx \rho(\alpha, \beta) = \frac{\alpha^\top \Sigma_{XY} \beta}{\sqrt{\alpha^\top \Sigma_{XX} \alpha}\,\sqrt{\beta^\top \Sigma_{YY} \beta}}$$

CCA optimization problem:

$$\text{maximize } \widetilde{\alpha}^\top \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}} \widetilde{\beta} \text{ w.r.t. } \widetilde{\alpha} \in \mathbb{R}^p \text{ and } \widetilde{\beta} \in \mathbb{R}^q$$
$$\text{s.t. } \|\widetilde{\alpha}\| = 1 \text{ and } \|\widetilde{\beta}\| = 1$$

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

CCA solution

Correlation coefficient

$$\text{cor}(\alpha^\top x, \beta^\top y) \approx \rho(\alpha, \beta) = \frac{\alpha^\top \Sigma_{XY} \beta}{\sqrt{\alpha^\top \Sigma_{XX} \alpha} \sqrt{\beta^\top \Sigma_{YY} \beta}}$$

CCA optimization problem:

$$\text{maximize } \widetilde{\alpha}^\top \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}} \widetilde{\beta} \text{ w.r.t. } \widetilde{\alpha} \in \mathbb{R}^p \text{ and } \widetilde{\beta} \in \mathbb{R}^q$$
$$\text{s.t. } \|\widetilde{\alpha}\| = 1 \text{ and } \|\widetilde{\beta}\| = 1$$

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

Problem: does not work for $p > n$!

Regularization

Modified correlation coefficient

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

RCCA optimization problem:

$$\begin{aligned} \text{maximize } & \alpha^\top \Sigma_{XY} \beta \text{ w.r.t. } \alpha \in \mathbb{R}^p \text{ and } \beta \in \mathbb{R}^q \\ \text{s.t. } & \alpha^\top \Sigma_{XX} \alpha = 1,\ \beta^\top \Sigma_{YY} \beta = 1 \text{ and } \|\alpha\| \leq t_1 \end{aligned}$$

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

CCA package

rcc {CCA} R Documentation

Regularized Canonical Correlation Analysis

Description

The function performs the Regularized extension of the Canonical Correlation Analysis to seek correlations between two data matrices when the number of columns (variables) exceeds the number of rows (observations)

Usage

```
rcc(X, Y, lambda1, lambda2)
```

Arguments

`X`	numeric matrix (n * p), containing the X coordinates.
`Y`	numeric matrix (n * q), containing the Y coordinates.
`lambda1`	Regularization parameter for X
`lambda2`	Regularization parameter for Y

Details

When the number of columns is greater than the number of rows, the matrice X'X (and/or Y'Y) may be ill-conditioned. The regularization allows the inversion by adding a term on the diagonal.

Value

A list containing the following components:

`corr`	canonical correlations
`names`	a list containing the names to be used for individuals and variables for graphical outputs
`xcoef`	estimated coefficients for the 'X' variables as returned by `cancor()`
`ycoef`	estimated coefficients for the 'Y' variables as returned by `cancor()`
`scores`	a list returned by the internal function comput() containing individuals and variables coordinates on the canonical variates basis.

Author(s)

Sébastien Déjean, Ignacio González

CCA package

```
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")
```

CCA package

```
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")
```

```
"rcc" <-
function (X, Y, lambda1, lambda2)
{
    Xnames <- dimnames(X)[[2]]
    Ynames <- dimnames(Y)[[2]]
    ind.names <- dimnames(X)[[1]]
    Cxx <- var(X, na.rm = TRUE, use = "pairwise") + diag(lambda1,
        ncol(X))
    Cyy <- var(Y, na.rm = TRUE, use = "pairwise") + diag(lambda2,
        ncol(Y))
    Cxy <- cov(X, Y, use = "pairwise")
    res <- geigen(Cxy, Cxx, Cyy)
    names(res) <- c("cor", "xcoef", "ycoef")
    scores <- comput(X, Y, res)
    return(list(cor = res$cor, names = list(Xnames = Xnames,
        Ynames = Ynames, ind.names = ind.names), xcoef = res$xcoef,
        ycoef = res$ycoef, scores = scores))
}
```

CCA package

```r
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")
```

```r
"rcc" <-
function (X, Y, lambda1, lambda2)
{
    Xnames <- dimnames(X)[[2]]
    Ynames <- dimnames(Y)[[2]]
    ind.names <- dimnames(X)[[1]]
    Cxx <- var(X, na.rm = TRUE, use = "pairwise") + diag(lambda1,
        ncol(X))
    Cyy <- var(Y, na.rm = TRUE, use = "pairwise") + diag(lambda2,
        ncol(Y))
    Cxy <- cov(X, Y, use = "pairwise")
    res <- geigen(Cxy, Cxx, Cyy)
    names(res) <- c("cor", "xcoef", "ycoef")
    scores <- comput(X, Y, res)
    return(list(cor = res$cor, names = list(Xnames = Xnames,
        Ynames = Ynames, ind.names = ind.names), xcoef = res$xcoef,
        ycoef = res$ycoef, scores = scores))
}
```

$$C_{XX} = \boxed{p \times p}$$

$$C_{YY} = \boxed{q \times q}$$

$$C_{XY} = \boxed{p \times q}$$

Problem: C_{XX}, C_{XY} are large for $p \gg n$

"Kernel trick"

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

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

"Kernel trick"

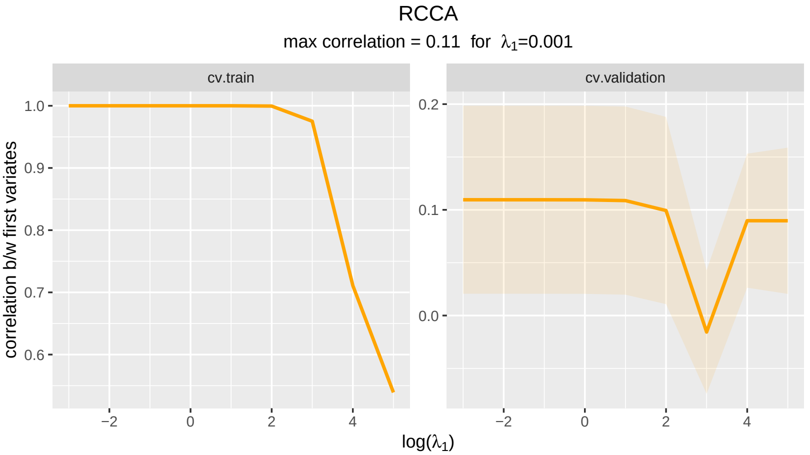
Goal: find a linear transformation such that RCCA for (X, Y) is equivalent to RCCA for (R, Y)

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

Solution:

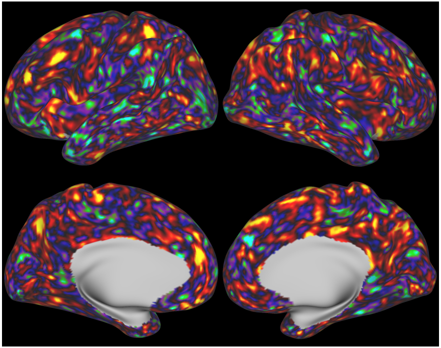
1. $X = UDV^\top = \boxed{n \times n}\ \boxed{n \times n}\ \boxed{n \times p}$
2. set $R = XV = UD$ and solve RCCA problem for $(R, Y) \Longrightarrow$ canonical coefficients α_R, β_R
3. apply inverse transformation $\alpha_X = V\alpha_R$ and $\beta_X = \beta_R$
4. the variates stay the same $v_R = R\alpha_R = X\alpha_X = v_X$ and $u_R = u_X$

RCCA results

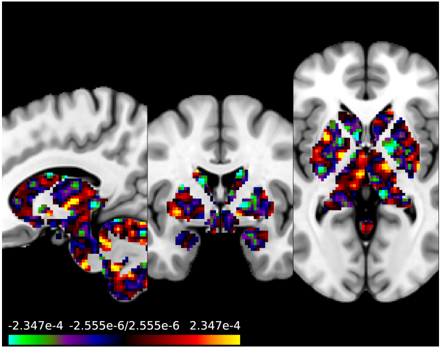


Details: adjust for sex variable, run 10-fold cross-validation to tune the hyperparameter λ_1, plot unpenalized correlation between canonical variates

Visualization: plot canonical coefficients α for the optimal RCCA model with $\lambda_1 = 0.001$



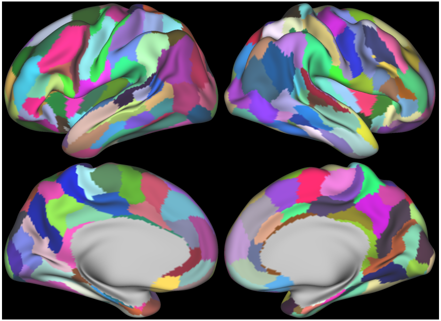
(a) Cortical coefficients



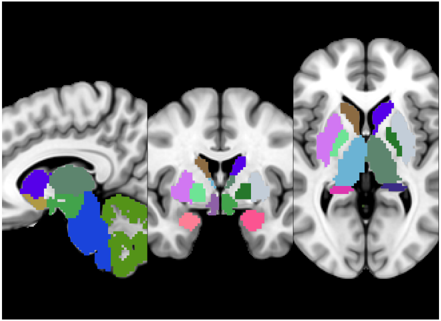
(b) Subcortical coefficients

Brain regions

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)

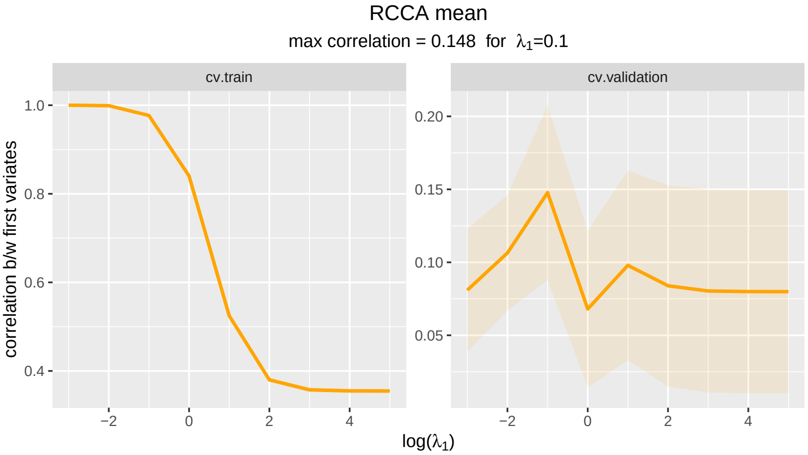
Grouped structure

Notations:

- $K = 229$ groups
- $p_k = \#$ features in group k
- X_k – set of features in group k
- α_k – set of coefficients in group k

$$X = (\underbrace{X_1}_{p_1}, \dots, \underbrace{X_K}_{p_K}) \text{ and } \alpha = (\underbrace{\alpha_1}_{p_1}, \dots, \underbrace{\alpha_K}_{p_K})$$

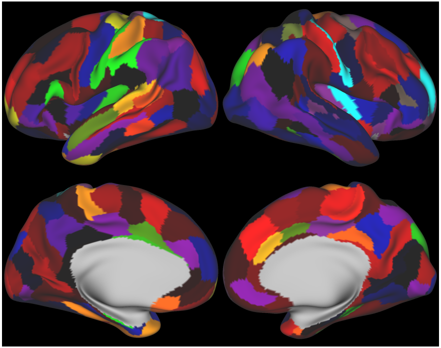
First attempt: RCCA mean



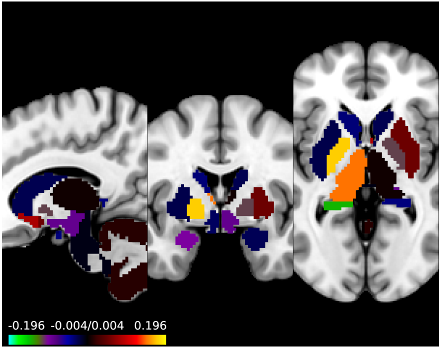
Details: replace $X = (\underbrace{X_1}_{p_1}, \dots, \underbrace{X_K}_{p_K})$ by $\bar{X} = (\underbrace{\bar{X}_1, \dots, \bar{X}_K}_{K})$

First attempt: RCCA mean

Visualization: plot canonical coefficients α for the optimal RCCA mean model with $\lambda_1 = 0.1$



(a) Cortical coefficients



(b) Subcortical coefficients

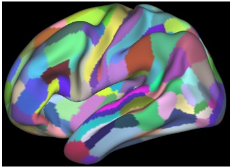
Can we do better?

Grouping constraints

$$X = (\underbrace{X_1}_{p_1}, \ldots, \underbrace{X_K}_{p_K}) \text{ and } \alpha = (\underbrace{\alpha_1}_{p_1}, \ldots, \underbrace{\alpha_K}_{p_K})$$

Assumptions:

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

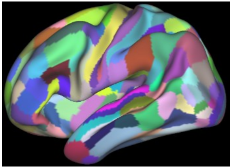


Grouping constraints

$$X = (\underbrace{X_1}_{p_1}, \dots, \underbrace{X_K}_{p_K}) \text{ and } \alpha = (\underbrace{\alpha_1}_{p_1}, \dots, \underbrace{\alpha_K}_{p_K})$$

Assumptions:

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



GRCCA optimization problem:

$$\text{maximize } \alpha^\top \Sigma_{XY} \beta \text{ w.r.t. } \alpha \in \mathbb{R}^p \text{ and } \beta \in \mathbb{R}^q$$
$$\text{s.t } \alpha^\top \Sigma_{XX} \alpha = 1, \; \beta^\top \Sigma_{YY} \beta = 1,$$
$$\sum_{k=1}^{K} \|\alpha_k - \bar{\alpha}_k\|^2 \leq t_1 \textit{ and } \sum_{k=1}^{K} p_k \bar{\alpha}_k^2 \leq s_1$$

GRCCA optimization problem:

$$\text{maximize } \alpha^\top \Sigma_{XY}\beta \text{ w.r.t. } \alpha \in \mathbb{R}^p \text{ and } \beta \in \mathbb{R}^q$$

$$\text{s.t } \alpha^\top \Sigma_{XX}\alpha = 1,\ \beta^\top \Sigma_{YY}\beta = 1,$$

$$\sum_{k=1}^{K} \|\alpha_k - \bar{\alpha}_k\|^2 \le t_1 \ \textit{and} \ \sum_{k=1}^{K} p_k \bar{\alpha}_k^2 \le s_1$$

Modified correlation coefficient

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

$$C = \begin{bmatrix} \frac{\mathbf{1}\mathbf{1}^\top}{p_1} & 0 & \dots & 0 \\ 0 & \frac{\mathbf{1}\mathbf{1}^\top}{p_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{\mathbf{1}\mathbf{1}^\top}{p_K} \end{bmatrix}$$

GRCCA properties

- when $\lambda_1 = \mu_1$ GRCCA is equivalent to RCCA
- when $\lambda_1 \to \infty$ we it becomes RCCA mean

GRCCA properties

- when $\lambda_1 = \mu_1$ GRCCA is equivalent to RCCA
- when $\lambda_1 \to \infty$ we it becomes RCCA mean

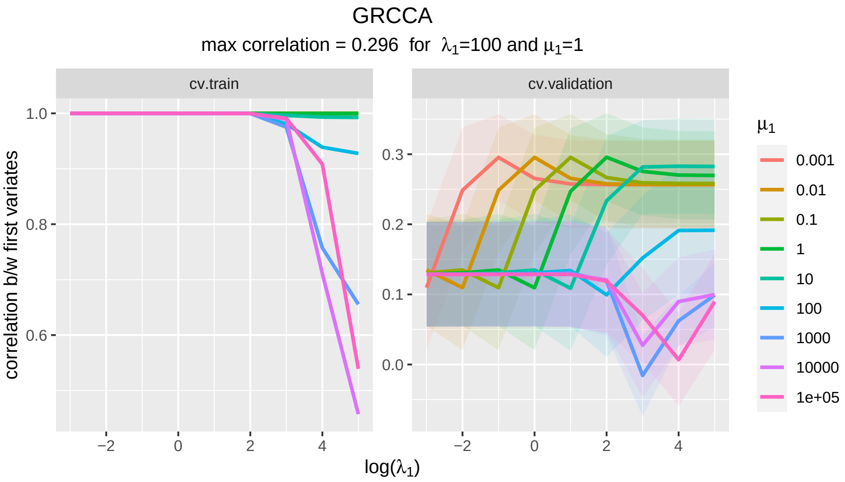
Lemma

GRCCA for (X, Y) is equivalent to RCCA for $(\widetilde{X}, Y)$ where

$$\widetilde{X} = \left(\mathbf{X}_1 - \bar{\mathbf{X}}_1, \sqrt{\frac{p_1\lambda_1}{\mu_1}}\bar{\mathbf{X}}_1, \ldots, \mathbf{X}_K - \bar{\mathbf{X}}_K, \sqrt{\frac{p_K\lambda_1}{\mu_1}}\bar{\mathbf{X}}_K\right)$$

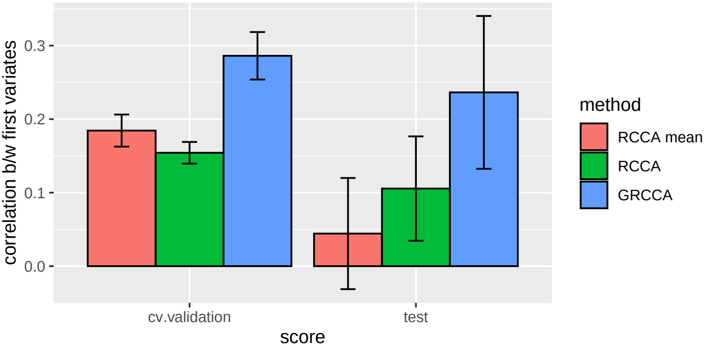
Can use the "kernel trick"!

GRCCA results



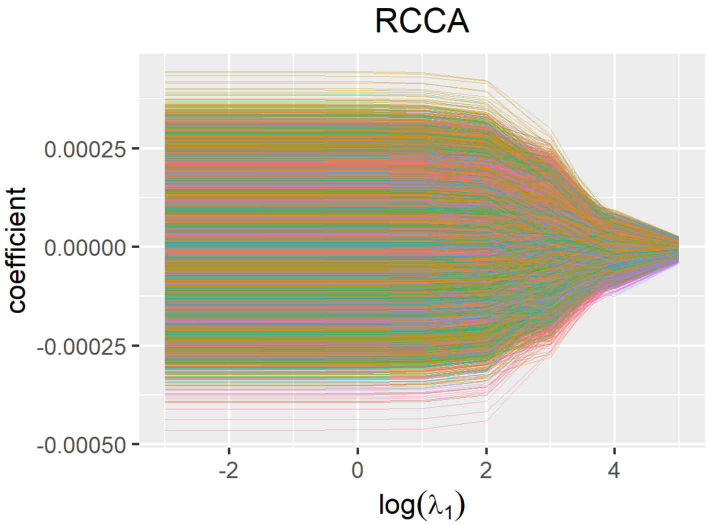
Details: λ_1 controls variation within each region, μ_1 controls variation across the regions; we observe the pattern $\frac{\lambda_1}{\mu_1} = 100$.

Significance of the CCA correlation

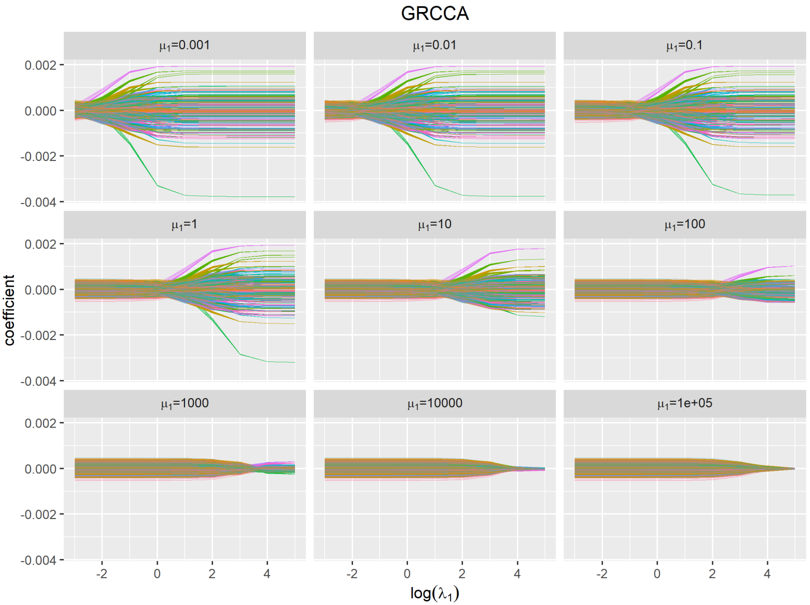


Procedure: split into 11 folds; *cv.validation* = maximum score obtained via 10-fold cross-validation, averaged across 11 NCV folds; *test* = score computed on independent test set, averaged across 11 NCV folds.

Using GRCCA for visualization

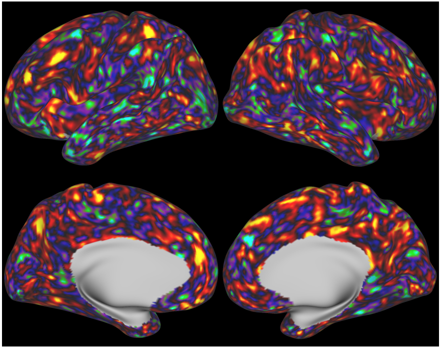


Using GRCCA for visualization

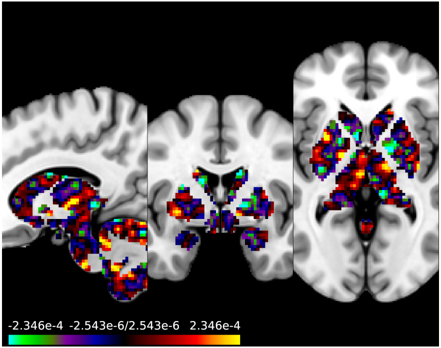


Improved interpretability of coefficients

Visualization: plot canonical coefficients α for GRCCA model with $\lambda_1 = 1$ and $\mu_1 = 1$



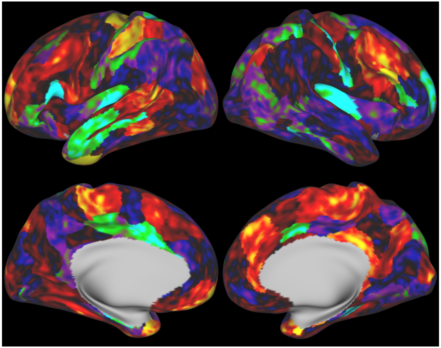
(a) Cortical coefficients.



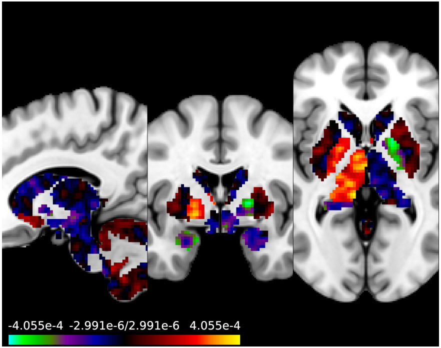
(b) Subcortical coefficients.

Improved interpretability of coefficients

Visualization: plot canonical coefficients α for GRCCA model with $\lambda_1 = 10$ and $\mu_1 = 1$

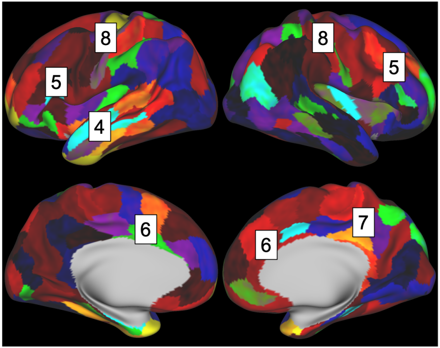


(a) Cortical coefficients.

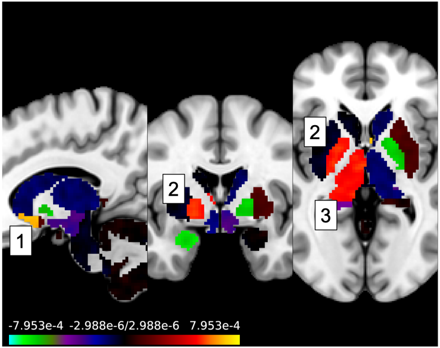


(b) Subcortical coefficients.

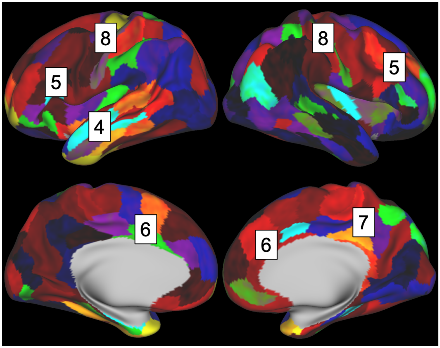
Visualization: plot canonical coefficients α for GRCCA model with optimal $\lambda_1 = 100$ and $\mu_1 = 1$



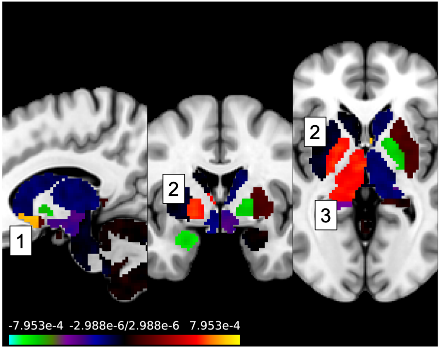
(a) Cortical coefficients.



(b) Subcortical coefficients.



(a) Cortical coefficients.



(b) Subcortical coefficients.

Annotation of brain regions: [1] nucleus accumbens, [2] putamen, [3] thalamus, [4] temporal lobe, [5] dorsolateral prefrontal cortex, [6] dorsomedial prefrontal cortex, [7] posterior cingulate cortex, [8] precentral cortex. High positive loadings in subcortical regions involved in reward processing, such as the striatum [1-2] and thalamus [3], and on a cortical network encompassing the temporal lobe [4-8]. Most of these regions have been shown to be connected to the striatum and to be part of key reward-processing.

Appendix: general approach to regularization

Constraint: $\alpha^\top K \alpha \leq t$

Modified correlation coefficient:

$$\rho(\alpha, \beta; K_X) \approx \frac{\alpha^\top \Sigma_{XY} \beta}{\sqrt{\alpha^\top (\Sigma_{XX} + K)\alpha}\sqrt{\beta^\top (\Sigma_{YY})\beta}}$$

Examples:

Full	$K = \lambda_1 I$	$\|\alpha\|^2 \leq t_1$	$\alpha = (\alpha_1, \dots, \alpha_q)$
Partial	$K = \lambda_1 \left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$	$\|\alpha_1\|^2 \leq t_1$	$\alpha = (\underbrace{\alpha_1}_{p_1}, \underbrace{\alpha_2}_{p_2})$
Groups	$K = \lambda_1 (I - C)$	$\sum_{k=1}^{K} \|\alpha_k - \bar{\alpha}_k\|^2 \leq t_1$	$\alpha = (\underbrace{\alpha_1}_{p_1}, \dots, \underbrace{\alpha_K}_{p_K})$
Sparse groups	$K = \mu_1 C$	$\sum_{k=1}^{K} p_k \bar{\alpha}_k^2 \leq s_1$	$\alpha = (\underbrace{\alpha_1}_{p_1}, \dots, \underbrace{\alpha_K}_{p_K})$

Appendix: simulation

- Assume equal group size $p_k = \frac{p}{K}$.
- Generate $Y \in \mathbb{R}^q$ and centroids $X^c \in \mathbb{R}^K$

$$(Y, X^c) \sim \mathcal{N}_{q+K}(0, \Sigma) \quad \text{with} \quad \Sigma = \begin{pmatrix} I_q & 11^\top \sigma_{XY}^2 \\ 11^\top \sigma_{XY}^2 & I_K \end{pmatrix}.$$

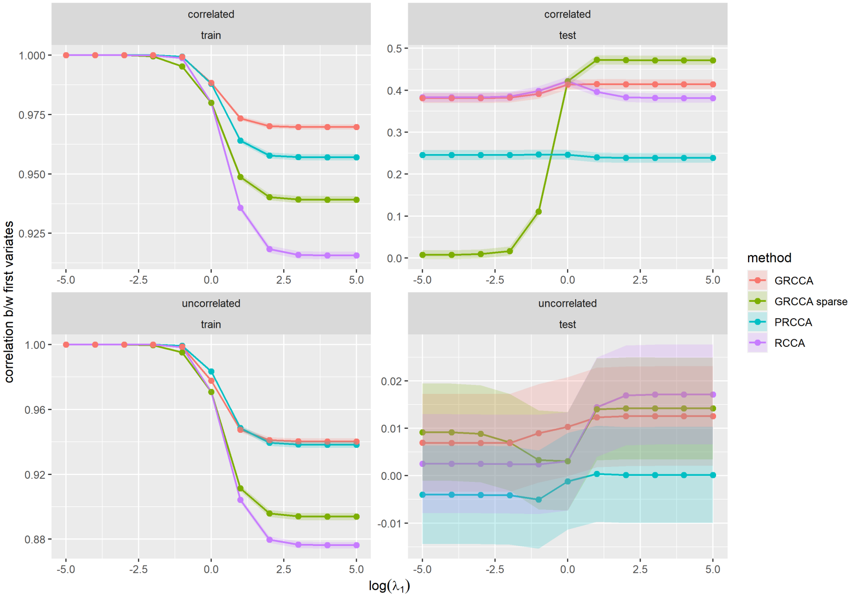
- Generate blocks $X_k \in \mathbb{R}^{p_k}$

$$X_k | X_k^c \sim \mathcal{N}_{p_k}(1X_k^c, \sigma_X^2 I).$$

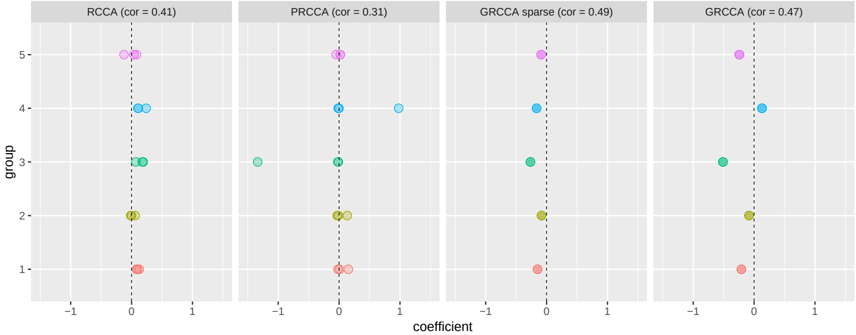
- Concatenate blocks $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K) \in \mathbb{R}^{n \times p}$.

In our experiments we use $n = 10$, $p = 15$ and $q = 3$, the number of groups is $K = 5$. We set $\sigma_X = 1$ and test two settings: $\sigma_{XY} = 0.5$ for correlated data and $\sigma_{XY} = 0$ for independent data.

Appendix: simulation correlation



Appendix: simulation 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.

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.

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.

Elena Tuzhilina, L. Tozzi, and T. Hastie, *Canonical correlation analysis in high dimensions with structured regularization*, Statistical Modelling, (2021).

L. Tozzi, B. Staveland, B. Holt-Gosselin, M. Chesnut, S. E. Chang, D. Choi, M. Shiner, H. Wu, G. Lerma-Usabiaga, O. Sporns, D. M. Barch, I. H. Gotlib, T. J. Hastie, A. B. Kerr, R. A. Poldrack, B. A. Wandell, M. Wintermark, and L. M. Williams, *The human connectome project for disordered emotional states: Protocol and rationale for a research domain criteria study of brain connectivity in young adult anxiety and depression*, NeuroImage, 214 (2020), p. 116715.

L. Tozzi, **Elena Tuzhilina**, M. Glasser, T. Hastie, and L. Williams, *Relating whole-brain functional connectivity to self-reported negative emotion in a large sample of young adults using group regularized canonical correlation analysis*, NeuroImage, (2021).

R package is available at: github.com/ElenaTuzhilina/RCCA

Thank you for your attention!