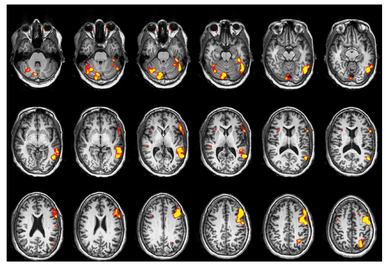
Canonical Correlation Analysis in high dimensions with structured regularization

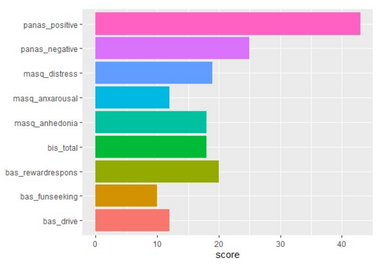
[Elena Tuzhilina] University of Toronto, Department of Statistical Sciences
Joint work with C. Donnat, L.Tozzi and T.Hastie

Motivation

Canonical correlation analysis (CCA) is a technique for measuring the association between two multivariate data matrices. A *regularized modification* of canonical correlation analysis (RCCA), imposing an ℓ_2 penalty on the CCA coefficients is widely used in applications with high-dimensional data. One limitation of such regularization is that it ignores any data structure, which can be ill-suited for some applications. Here we introduce a novel approach that takes the underlying data structure into account. The proposed *group regularized canonical correlation analysis (GRCCA)*, is especially useful when the variables are correlated in groups. We illustrate some computational strategies to avoid excessive computations with regularized CCA in high dimensions. We demonstrate the application of GRCCA method in our motivating application from neuroscience.

Data





Brain activations $X \in \mathbb{R}^{n\times p}$: magnetic resonance imaging obtained during a gambling task.

Behavioral scores $Y \in \mathbb{R}^{n\times q}$: self-reports assessing various aspects of reward-related behaviors.

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

Canonical Correlation Analysis

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

maximize $\text{cor}(\alpha^T x, \beta^T y)$ w.r.t. α, β

- **canonical coefficients** α and β
- **canonical variates** $\alpha^T x$ and $\beta^T y$
- **canonical correlation** $\text{cor}(\alpha^T x, \beta^T y)$

Correlation coefficient:

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

CCA optimization problem:

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

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

Standard regularization

Motivation: CCA doesn't work for $p > n$

Modified correlation coefficient:

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

Shrinkage inequality:

$$\|\alpha\| \le t_1$$

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

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: $\Sigma_{XX} \in \mathbb{R}^{p\times p}$, $\Sigma_{XY} \in \mathbb{R}^{p\times q}$ are too large when $p \approx 90$K

Idea: find a linear transformation

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

such that RCCA for $(X, Y) \iff$ RCCA for (R, Y)

Step-by-step procedure:

1. $X = UDV^T = \boxed{n\times n}\ \boxed{n\times n}\ \boxed{n\times p}$
2. set $R = XV = UD$ and solve RCCA problem for $(R, Y) \Longrightarrow$ get α_R, β_R
3. recover coefficients $\alpha_X = V\alpha_R$
4. variates stay the same $R\alpha_R = X\alpha_X$

Group structure

Motivation: brain features come in $K = 229$ groups (aka brain regions).

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

Assumptions:

- group homogeneity $\alpha_k \approx \bar{\alpha}_k$
- differential group sparsity $\bar{\alpha}_k \approx 0$

Group RCCA

Shrinkage inequalities:

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

Modified correlation coefficient:

$$\frac{\alpha^T \Sigma_{XY} \beta}{\sqrt{\alpha^T (\Sigma_{XX} + K(\lambda_1,\mu_1))\alpha}\sqrt{\beta^T \Sigma_{YY} \beta}}$$

where $K(\lambda_1, \mu_1) = \lambda_1(I - C) + \mu_1 C$

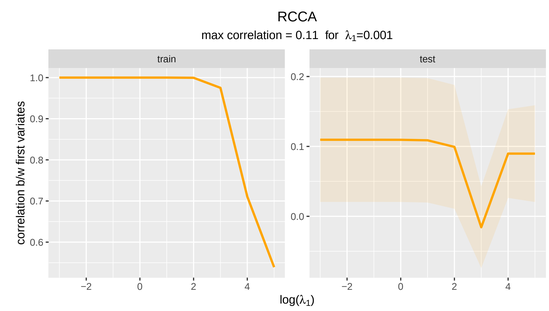
and $C = \begin{bmatrix} \frac{\mathbf{1}\mathbf{1}^T}{p_1} & 0 & \dots & 0 \\ 0 & \frac{\mathbf{1}\mathbf{1}^T}{p_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{\mathbf{1}\mathbf{1}^T}{p_K} \end{bmatrix}$

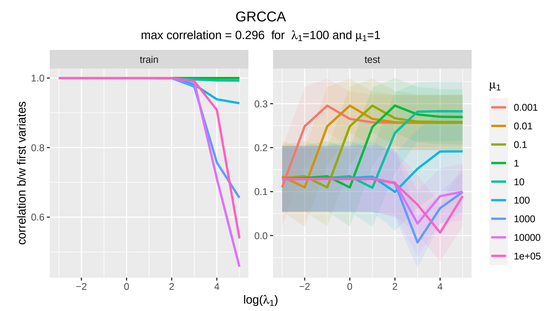
Lemma

GRCCA for $(X, Y) \iff$ RCCA for $(\tilde{X}, Y)$

$$\tilde{X} = (\mathbf{X}_1 - \bar{\mathbf{X}}_1, \dots, \mathbf{X}_K - \bar{\mathbf{X}}_K, \sqrt{\frac{p_1\lambda_1}{\mu_1}}\bar{\mathbf{X}}_1, \dots, \sqrt{\frac{p_K\lambda_1}{\mu_1}}\bar{\mathbf{X}}_K)$$

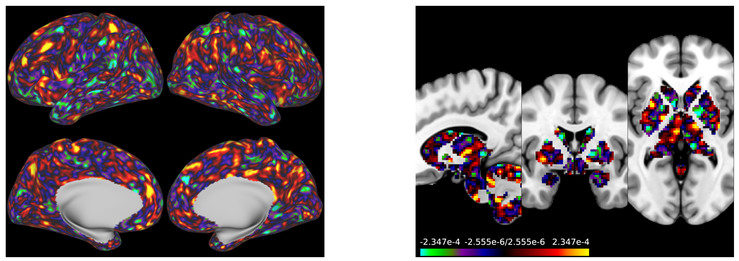
Cross validation results



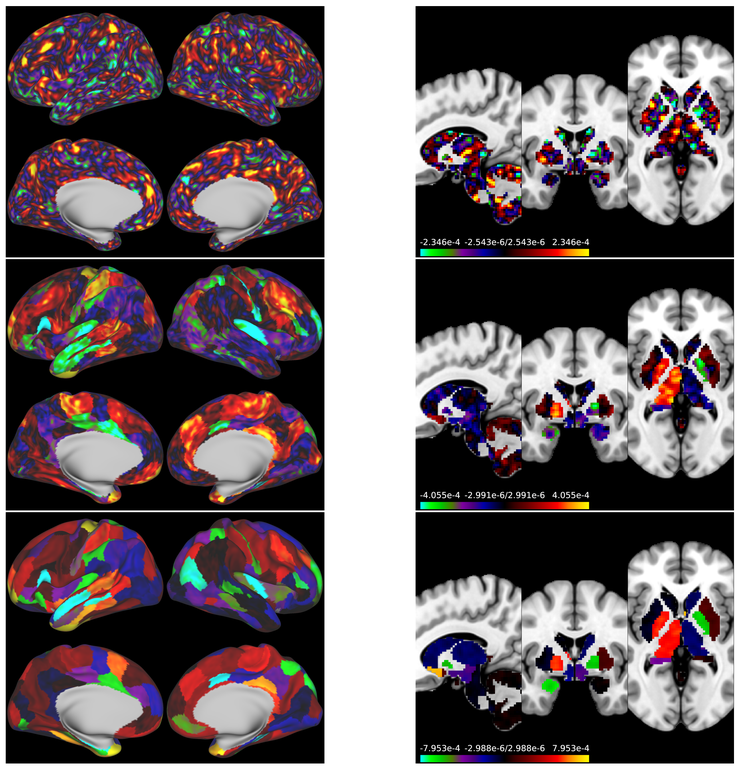


Canonical coefficients

RCCA ($\lambda_1 = 0.001$)



GRCCA ($\lambda_1 = 1, 10, 100$ and $\mu_1 = 1$)



Graph regularization

Motivation: brain features can be represented as a graph (V, E).

$V = \{1, \dots, p\}$ and $E \subseteq V \times V$

Assumptions:

edge homogeneity $\alpha_i \approx \alpha_j$ for $(i,j) \in E$

Shrinkage inequality:

$$\sum_{(i,j)\in E} (\alpha_i - \alpha_j)^2 \le t_1$$

Modified correlation coefficient:

$$\frac{\alpha^T \Sigma_{XY} \beta}{\sqrt{\alpha^T (\Sigma_{XX} + \lambda_1 L)\alpha}\sqrt{\beta^T \Sigma_{YY} \beta}}$$

where $L = D - A$ is Laplacian matrix.

Question: Can we automatically detect clusters of X features (brain regions)?

Idea: We can encourage the edges to be sparse using

$$\sum_{(i,j)\in E} |\alpha_i - \alpha_j| \le t_2$$

References

[1] Hotelling. Relations between two sets of variables. *Biometrika*, 28, 1936.

[2] Gonzalez et al. CCA: An R Package to Extend Canonical Correlation Analysis. *Journal of Statistical Software*, 23(12), 2008.

[3] Leurgans et al. Canonical Correlation Analysis when the Data are Curves. *Journal of the Royal Statistical Society*, 55(3), 1993.