

Canonical correlation analysis in high dimensions with structured regularization

Elena Tuzhilina¹, Leonardo Tozzi² and Trevor Hastie³

¹Department of Statistics, Stanford University, Stanford, CA, USA

²Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA

³Department of Statistics, Stanford University, Stanford, CA, USA

Abstract: Canonical correlation analysis (CCA) is a technique for measuring the association between two multivariate data matrices. A regularized modification of canonical correlation analysis (RCCA) which imposes 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, treating all the features equally, which can be ill-suited for some applications. In this article we introduce several approaches to regularizing CCA that take the underlying data structure into account. In particular, the proposed group regularized canonical correlation analysis (GRCCA) is 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 these methods in our motivating application from neuroscience, as well as in a small simulation example.

Key words: canonical correlation analysis, group penalty, high dimensions, regularization, structured data

Received May 2021; revised August 2021; accepted August 2021

1 Introduction

Canonical correlation analysis (CCA) is a classic method commonly used in statistics for studying complex multivariate data. CCA was first introduced by Hotelling (1936) as a tool for finding relationships between two sets of variables. It remains relevant in many domains including, but not limited to, genetics (see, for example, Waaijenborg et al., 2008; Parkhomenko et al., 2009; Cao et al., 2009) and neuroscience (see, for example, Wang et al., 2020; Zhuang et al., 2020). In many applications the number of available observations is significantly smaller than the number of features under consideration, so some form of regularization is essential. Numerous regularized CCA extensions have been proposed (see, for example, Lykou and Whittaker, 2010; Hardoon and Shawe-Taylor, 2011; Witten and Tibshirani, 2009). The most popular existing approach, called Regularized CCA (RCCA),

Address for correspondence: Elena Tuzhilina, Department of Statistics, Stanford University, Sequoia Hall, Stanford, CA-94305, USA.
E-mail: elenatuz@stanford.edu

imposes an ℓ_2 -penalty on the canonical coefficients (see, for example, Vinod, 1976; Leurgans et al., 1993). Like any other standard regularization method based on the ℓ_2 penalty, it has the property of treating all the coefficients equally and shrinking them towards zero. Although RCCA is well suited to data with general structure, in some applications the structure of the data can play an important role when investigating the association between variables. In this article, we develop several regularized extensions of CCA. These extensions were originally motivated by brain imaging applications, but the scope of applications can readily be extended to other fields.

The article is organized as follows. In Section 2, we introduce the necessary background for both CCA and RCCA methods. In Section 3 we propose several approaches to regularization that account for the underlying structure of the data. In particular, in Section 4, we introduce partially regularized canonical correlation analysis (PRCCA) that imposes an ℓ_2 penalty only on a subset of the features. Although both RCCA and PRCCA problems have simple explicit solutions that can be computed via singular value decomposition, they require us to work in terms of sample covariance matrices. This can be infeasible when the number of features is very large. In Sections 2.4 and 4.2, we cover the ‘kernel’ trick that allows to escape excessive computations while conducting CCA with regularization. In Section 5, we introduce group regularized CCA (GRCCA), a novel method that exploits group structure in the data.

Since all the methods under consideration have similar structure, they can be considered as special cases of CCA with a general regularization penalty discussed in Sections 6. All the technical details and proofs for these methods are covered in the supplemental material.

We illustrate the proposed methods on our motivating study example involving functional brain imaging data in Section 3 as well as on a small simulation example in Section 7. We conclude with a discussion, including some ideas for future work.

2 Canonical correlation analysis with regularization

2.1 Canonical correlation analysis

Consider two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$. The goal of CCA is to find a linear combination of X variables and a linear combination of Y variables with the maximum possible correlation. Typically, we find a sequence of such linear combinations. Namely, for $i = 1, \dots, \min(p, q)$ define a sequence of pairs of random variables (U_i, V_i) as follows (see, for example, Härdle and Simar, 2007)

1. Random variables U_i and V_i are linear combinations of X and Y , respectively, that is,

$$U_i = \alpha_i^\top X \text{ and } V_i = \beta_i^\top Y.$$

2. Coefficient vectors $\alpha_i \in \mathbb{R}^p$ and $\beta_i \in \mathbb{R}^q$ maximize the correlation

$$\rho(\alpha_i, \beta_i) = \text{cor}(\alpha_i^\top X, \beta_i^\top Y).$$

3. Pair (U_i, V_i) is uncorrelated with previous pairs, i.e

$$\text{cor}(U_i, U_j) = \text{cor}(V_i, V_j) = 0 \text{ for } j < i.$$

The pair (U_i, V_i) is called *i-th pair of canonical variates*; the corresponding optimal correlation value $\rho_i = \rho(\alpha_i, \beta_i)$ is called *i-th canonical correlation*.

Note that correlation coefficient $\rho(\alpha, \beta)$ can be rewritten as

$$\rho_{CCA}(\alpha, \beta) = \frac{\alpha^\top \Sigma_{XY} \beta}{\sqrt{\alpha^\top \Sigma_{XX} \alpha} \sqrt{\beta^\top \Sigma_{YY} \beta}}, \quad (2.1)$$

where Σ_{XX} , Σ_{YY} and Σ_{XY} refer to the covariance matrices $\text{cov}(X)$, $\text{cov}(Y)$ and $\text{cov}(X, Y)$, respectively. It is easy to restate maximization of $\rho_{CCA}(\alpha, \beta)$ w.r.t. α and β in terms of a constrained optimization problem

$$\begin{aligned} & \text{maximize } \alpha^\top \Sigma_{XY} \beta \text{ w.r.t. } \alpha \in \mathbb{R}^p \text{ and } \beta \in \mathbb{R}^q \\ & \text{subject to } \alpha^\top \Sigma_{XX} \alpha = 1 \text{ and } \beta^\top \Sigma_{YY} \beta = 1. \end{aligned} \quad (2.2)$$

Thus, finding the *i*-th canonical pair is equivalent to solving the problem:

$$\begin{aligned} & \text{maximize } \alpha_i^\top \Sigma_{XY} \beta_i \text{ w.r.t. } \alpha_i \in \mathbb{R}^p \text{ and } \beta_i \in \mathbb{R}^q \\ & \text{subject to } \alpha_i^\top \Sigma_{XX} \alpha_i = 1 \text{ and } \beta_i^\top \Sigma_{YY} \beta_i = 1 \\ & \alpha_i^\top \Sigma_{XX} \alpha_j = 0 \text{ and } \beta_i^\top \Sigma_{YY} \beta_j = 0 \text{ for } j < i. \end{aligned}$$

One can show that the canonical variates can be found via a singular value decomposition of the matrix $\Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XY} \Sigma_{YY}^{-\frac{1}{2}}$, and that the canonical correlations coincide with the singular values of this matrix (see, for example, Mardia et al., 1979).

2.2 Dealing with high dimensions

In practice, we replace covariance matrices Σ_{XX} , Σ_{YY} and Σ_{XY} by the sample covariance matrices $\widehat{\Sigma}_{XX}$, $\widehat{\Sigma}_{YY}$ and $\widehat{\Sigma}_{XY}$. Specifically, suppose $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$ refer to matrices of n observations for random vectors X and Y , respectively. Without loss of generality, assume that the columns of \mathbf{X} and \mathbf{Y} are centred (mean 0), then

$$\widehat{\Sigma}_{XX} = \frac{1}{n} \mathbf{X}^\top \mathbf{X}, \quad \widehat{\Sigma}_{YY} = \frac{1}{n} \mathbf{Y}^\top \mathbf{Y} \quad \text{and} \quad \widehat{\Sigma}_{XY} = \frac{1}{n} \mathbf{X}^\top \mathbf{Y}.$$

If the number of observations n is smaller than p and/or q , the corresponding sample covariance matrices are singular and the inverses $\widehat{\Sigma}_{XX}^{-\frac{1}{2}}$ and/or $\widehat{\Sigma}_{YY}^{-\frac{1}{2}}$ do not exist. *Regularized canonical correlation analysis (RCCA)* resolves this problem by adding diagonal matrices to the sample covariance matrices of X and Y (see, for example, Leurgans et al., 1993; Gonzalez et al., 2008):

$$\widehat{\Sigma}_{XX}(\lambda_1) = \widehat{\Sigma}_{XX} + \lambda_1 I_p \quad \text{and} \quad \widehat{\Sigma}_{YY}(\lambda_2) = \widehat{\Sigma}_{YY} + \lambda_2 I_q. \quad (2.3)$$

Here I_p refers to the $p \times p$ identity matrix. The *modified correlation coefficient* that is maximized while seeking pairs of canonical variates is, hence,

$$\rho_{RCCA}(\alpha, \beta; \lambda_1, \lambda_2) = \frac{\alpha^\top \widehat{\Sigma}_{XY} \beta}{\sqrt{\alpha^\top (\widehat{\Sigma}_{XX} + \lambda_1 I) \alpha} \sqrt{\beta^\top (\widehat{\Sigma}_{YY} + \lambda_2 I) \beta}}. \quad (2.4)$$

By analogy with CCA, it is easy to show that the RCCA variates can be found via the singular value decomposition of the matrix $(\widehat{\Sigma}_{XX} + \lambda_1 I)^{-\frac{1}{2}} \widehat{\Sigma}_{XY} (\widehat{\Sigma}_{YY} + \lambda_2 I)^{-\frac{1}{2}}$ and that RCCA modified correlations are equal to the singular values of this matrix.

2.3 Shrinkage property

Similar to ridge regression, regularization shrinks the CCA coefficients α and β towards zero, where the *penalty parameters* λ_1 and λ_2 control the strength of the shrinkage of α and β , respectively. This can be supported by the following reasoning. As in the case of CCA, maximization of the modified correlation $\rho_{RCCA}(\alpha, \beta; \lambda_1, \lambda_2)$ w.r.t. α and β can be restated as a constrained optimization problem

$$\begin{aligned} & \text{maximize } \alpha^\top \widehat{\Sigma}_{XY} \beta \text{ w.r.t. } \alpha \in \mathbb{R}^p \text{ and } \beta \in \mathbb{R}^q \\ & \text{subject to } \alpha^\top (\widehat{\Sigma}_{XX} + \lambda_1 I) \alpha = 1 \text{ and } \beta^\top (\widehat{\Sigma}_{YY} + \lambda_2 I) \beta = 1. \end{aligned}$$

Note that the constraints can be rewritten as

$$\alpha^\top \widehat{\Sigma}_{XX} \alpha + \lambda_1 \|\alpha\|^2 = 1 \quad \text{and} \quad \beta^\top \widehat{\Sigma}_{YY} \beta + \lambda_2 \|\beta\|^2 = 1,$$

where $\|\cdot\|$ refers to the vector Euclidean norm. Finally, one can interpret λ_1 and λ_2 as Lagrangian dual variables for constraints $\|\alpha\|^2 \leq t_1$ and $\|\beta\|^2 \leq t_2$ which brings us to the optimization problem

$$\begin{aligned} & \text{maximize } \alpha^\top \widehat{\Sigma}_{XY} \beta \text{ w.r.t. } \alpha \in \mathbb{R}^p \text{ and } \beta \in \mathbb{R}^q \\ & \text{subject to } \alpha^\top \widehat{\Sigma}_{XX} \alpha = 1, \quad \|\alpha\|^2 \leq t_1 \quad \text{and} \quad \beta^\top \widehat{\Sigma}_{YY} \beta = 1, \quad \|\beta\|^2 \leq t_2. \end{aligned} \quad (2.5)$$

One can show that for some appropriately chosen t_1 and t_2 this problem is equivalent to maximizing objective (2.4). Moreover, increasing λ_1 and λ_2 is equivalent to decreasing thresholds t_1 and t_2 which leads us to the shrinkage property. Finally,

increasing λ_1 and λ_2 increases the denominator of (2.4) thereby shrinking the modified correlation coefficient to zero as well.

2.4 RCCA kernel trick

In some applications we need to deal with a very high-dimensional feature space. For instance, analysing functional magnetic resonance imaging (fMRI) data, where the dimension refers to the number of brain regions (or voxels), the number of features can reach hundreds of thousands. If one of p and q is very large it can be problematic to store matrices $\widehat{\Sigma}_{XX}$ and $\widehat{\Sigma}_{YY}$. In this section we illustrate a simple trick based on the invariance of the RCCA problem under orthogonal transformations, that allows one to handle high-dimensional data when computing the RCCA solution. The idea to reduce a high dimensional CCA problem to a low-dimensional one via the kernel trick was previously introduced by Kuss (2003) and Hardoon et al. (2005). Below we demonstrate the practical application of this idea to the RCCA problem that we subsequently use in the implementation.

For simplicity, we assume that regularization is imposed on the X part only, that is, we assume $q < n$ and set $\lambda_2 = 0$. The same reasoning applies if we regularize Y part as well. First we use the fact that any $n \times p$ matrix X with $p \gg n$ can be decomposed (e.g., via SVD) into a product $X = RV^\top$, where $R \in \mathbb{R}^{n \times n}$ is a square matrix, and $V \in \mathbb{R}^{p \times n}$ is a matrix with orthonormal columns, that is, $V^\top V = I$.

Lemma 2.1. [RCCA kernel trick] The original RCCA problem stated for X and Y can be reduced to solving the RCCA problem for R and Y . The resulting canonical correlations and variates for these two problems coincide. The canonical coefficients for the original problem can be recovered via the linear transformation $\alpha_X = V\alpha_R$.

See Supplement Section 1 for the proof. Note that for $p \gg n$ the above trick allows us to avoid manipulating large $p \times p$ and $p \times q$ covariance matrices $\widehat{\Sigma}_{XX}$ and $\widehat{\Sigma}_{XY}$ and to operate in terms of smaller $n \times n$ and $n \times q$ matrices $\widehat{\Sigma}_{RR}$ and $\widehat{\Sigma}_{RY}$. Of course, exactly the same trick can be applied to the Y part if $q \gg n$.

2.5 Hyperparameter tuning

Before proceeding to our first example, let us discuss how one can tune the hyperparameters. There are two hyperparameters for RCCA, that is, λ_1 and λ_2 . Let us denote the vector of hyperparameters by θ . The values for these hyperparameters can be chosen via cross-validation. Below we present the outline for hold-out cross-validation; however, it can be naturally extended to the case of k -fold cross-validation.

First we split all available observations into train (X_{train}, Y_{train}) and validation (X_{val}, Y_{val}) sets. We use the former set to fit the model and compute canonical coefficients $\alpha(\theta)$ and $\beta(\theta)$. Further, we use the latter set to estimate the model performance, that is, we calculate $\rho_{val}(\theta) = \text{cor}(X_{val}\alpha(\theta), Y_{val}\beta(\theta))$. Note that here

we utilize simple correlation instead of the modified correlation as a measure of performance. We pick the values of the hyperparameters maximizing the validation correlation, that is, $\theta^* = \text{argmax}_\theta (\rho_{val}(\theta))$, which can be done by means of grid search.

3 Example. Human Connectome data study

In this section we present an application of regularized CCA to data from a neuroscience study: the Human Connectome Project for Disordered Emotional States (HCP-DES) Tozzi et al. (2020). One aim of HCP-DES is to link the function of macroscopic human brain circuits to self-reports of emotional well-being using magnetic resonance imaging. Here, we focused on brain activations during a Gambling task designed to probe the brain circuits underlying reward (described in detail in Barch et al., 2013; Tozzi et al., 2020). We linked this neuroimaging data with self-reports assessing various aspects of reward-related behaviours (Behavioural Approach System/Behavioural Inhibition Scale (BIS/BAS), Carver and White, 1994), depression symptoms (Mood and Anxiety Symptom Questionnaire (MASQ), Wardenaar et al., 2010) and positive as well as negative affective states (Positive and Negative Affect Schedule (PANAS), Watson et al., 1988). We selected participants who had complete self-report and imaging data as well as no quality control issues, for a total of 153 participants (94 females, 59 males, mean age 25.91, sd 4.85). For details on the preprocessing and subject-level modelling used to derive brain activations in response to the Gambling task, see Section 6 of the Supplement. We used for our analysis the activations for the monetary reward compared to monetary loss during the task. For each subject, the activation at each greyordinate (grey matter coordinate) in the brain was extracted, yielding a matrix \mathbf{X} of $n = 153$ rows (subjects) and $p = 90\,368$ columns (greyordinates). The self-report data consisted of 9 variables: drive, fun seeking, reward responsiveness (from the BAS), total behavioural inhibition (from the BIS), distress, anhedonia, anxious arousal (from the MASQ), positive and negative affective states (from the PANAS). These were entered in a matrix \mathbf{Y} of $n = 153$ rows (subjects) and $q = 9$ columns.

To test our structured methods, the 90 368 greyordinates were grouped into 229 regions depending on their functional and anatomical properties, corresponding to the 210 regions of the Brainnetome atlas (Fan et al., 2016) with the addition of the 19 subcortical regions of the Desikan-Killiany Atlas (Desikan et al., 2006). The resulting brain regions are presented in Figure 1.

3.1 RCCA results

We start with a relatively simple model averaging activation inside each brain region and using the averaged values as features. Thus \mathbf{X} becomes of dimension 153×229 . To remove the effect of sex on the resulting correlation we adjusted both activation and behavioural data for the binary sex variable by means of

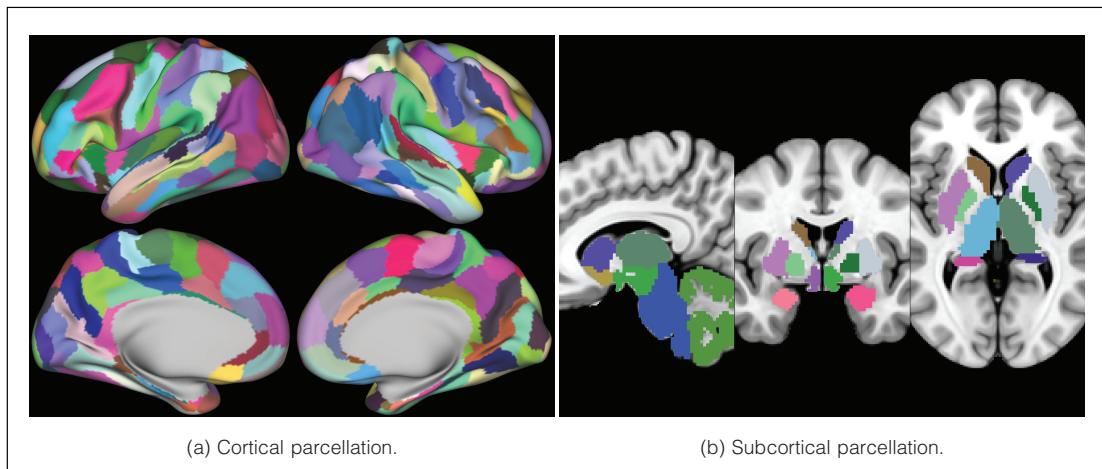


Figure 1 229 brain regions: 210 cortical regions of the Brainnetome atlas and 19 subcortical regions of the Desikan-Killiany Atlas

simple linear regression (mean adjustment). Since the \mathbf{Y} matrix is relatively small we imposed no penalty on \mathbf{Y} ($\lambda_2 = 0$). To pick the hyperparameter λ_1 we ran ten fold cross-validation on the adjusted data pairs with the penalty factor varying over the grid $\lambda_1 = 10^{-3}, 10^{-2}, \dots, 10^4, 10^5$.

The resulting cross-validation curves represent the unpenalized correlation between canonical variates (computed on the 9 folds of train set and one fold of validation set) averaged across 10 folds (see Figure 2a). Note that although larger λ_1 shrinks the modified correlation ρ_{RCCA} towards zero, the unpenalized correlation is not guaranteed to be monotonically decreasing in λ_1 . According to the plot, the highest score is achieved for $\lambda_1 = 0.1$ with the corresponding test correlation equal to 0.148. Using the kernel trick we now run RCCA for the original activation data (90 368 features adjusted for the sex effect). According to Figure 2b, the maximum score is equal to 0.11 ($\lambda_1 = 0.001$). In general, to compare two models and check that the cross-validation score does not reflect spurious findings we should validate the performance of the models on an independent test set. However, the small sample size of the data (only 153 observations) makes the test correlation estimates unreliable. We use nested cross-validation (NCV) to overcome the problem of overfitting to the dataset that we use for tuning. Specifically, we split the data in 11 folds. Each of the 11 folds is given an opportunity to be used as an independent test set, while all other 10 folds are used to tune the hyperparameters via ten fold cross-validation. Therefore, we report 11 cross-validation scores along with 11 test scores, and we present the average for both as well as 1SE confidence intervals (see Figure 7). According to the NCV scores, the cross-validation procedure for the full RCCA model discovered significant correlation (independent test set correlation averaged across 11 folds is 0.105). However, the correlation value obtained by the smaller mean RCCA model was way too optimistic (average test score is 0.044 with wide error bands).

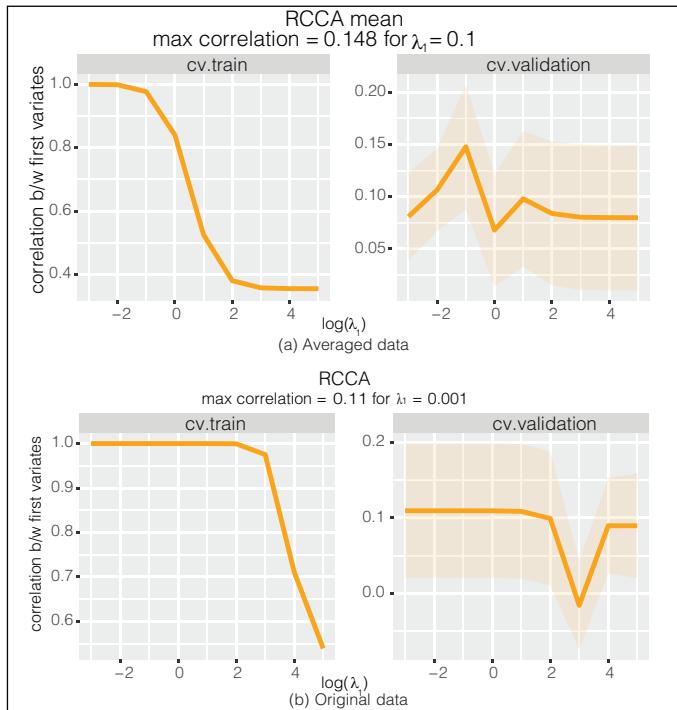


Figure 2 The cross-validation curves obtained via RCCA with ten fold cross-validation for the Human Connectome dataset. Left panel: (unpenalized) correlation between train canonical variates. Right panel: (unpenalized) correlation between validation canonical variates

It is worth noting the computational speed of the proposed method. Unlike the `rcc()` function from the popular CCA R package (see Gonzalez et al. (2008)), which is not able to handle such a large number of features ($\approx 90K$ for the X side), our implementation of RCCA with the kernel trick completes the calculations in 20 seconds.

3.2 Interpretability of canonical coefficients

In this section we visualize the RCCA coefficients α corresponding to the hyperparameters chosen by cross-validation. Recall that the original X features represent the brain activation detected at each brain greyordinate, so we can map the resulting RCCA coefficients back to the brain surface (see Figure 3). There is an apparent trade-off between the interpretability and flexibility of the model. Namely, although full data RCCA is more flexible, there is quite a lot of variation in the resulting canonical coefficients. This makes the corresponding brain image harder to interpret. On the other hand, the reduced model allows us to identify the brain regions that have the highest impact on the resulting correlation. However, it loses in terms of flexibility (and, potentially, performance). In what follows, we aim to develop the

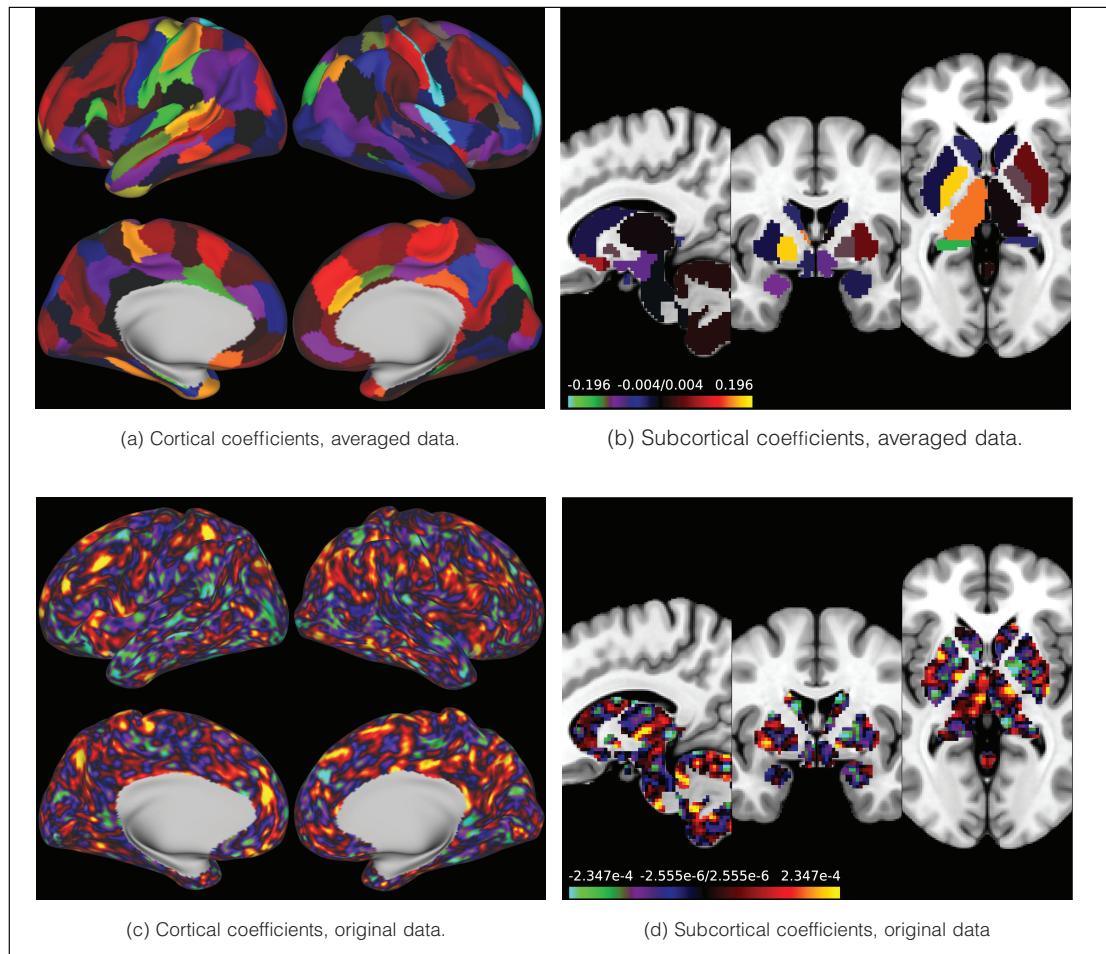


Figure 3 RCCA coefficients computed for the averaged data (229 features) with hyperparameter $\lambda_1 = 0.1$ chosen by cross-validation and for the original data (90 368 features) with hyperparameter $\lambda_1 = 10$ chosen by cross-validation

model that links these two extremes enabling us to control the interpretability vs. flexibility trade-off.

4 CCA with partial regularization

4.1 Penalizing a subset of canonical coefficients

Suppose you are interested in the influence of a specific brain region on the resulting CCA correlation, however, you do not want to completely eliminate the remaining

brain regions from the data. Recall that the inequality constraints in the RCCA optimization problem (2.5) control the deviation of all canonical coefficients from zero. PRCCA is a modification of RCCA that allows one to shrink only a subset of the CCA coefficients leaving the complement unpenalized. The proposed PRCCA method is a key building block for our final group RCCA approach and also admits some independent interesting applications.

Suppose that both α and β are split into two parts

$$\begin{aligned}\alpha &= \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \text{ where } \alpha_1 \in \mathbb{R}^{p_1} \text{ and } \alpha_2 \in \mathbb{R}^{p_2} \text{ with } p_1 + p_2 = p, \\ \beta &= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \text{ where } \beta_1 \in \mathbb{R}^{q_1} \text{ and } \beta_2 \in \mathbb{R}^{q_2} \text{ with } q_1 + q_2 = q.\end{aligned}$$

Replacing the constraints $\|\alpha\|^2 \leq t_1$ and $\|\beta\|^2 \leq t_2$ in the optimization problem (2.5) by $\|\alpha_1\|^2 \leq t_1$ and $\|\beta_1\|^2 \leq t_2$, respectively, we get the PRCCA optimization problem

$$\begin{aligned}&\text{maximize } \alpha^\top \widehat{\Sigma}_{XY} \beta \text{ w.r.t. } \alpha \in \mathbb{R}^p \text{ and } \beta \in \mathbb{R}^q \\ &\text{subject to } \alpha^\top \widehat{\Sigma}_{XX} \alpha = 1, \quad \|\alpha_1\|^2 \leq t_1 \text{ and } \beta^\top \widehat{\Sigma}_{YY} \beta = 1, \quad \|\beta_1\|^2 \leq t_2.\end{aligned}\quad (4.1)$$

Re-expressing the constraints for α and β in terms of dual variables, we get

$$\begin{aligned}\alpha^\top \widehat{\Sigma}_{XX} \alpha + \lambda_1 \|\alpha_1\|^2 &= \alpha^\top \left(\widehat{\Sigma}_{XX} + \lambda_1 \begin{pmatrix} I_{p_1} & 0 \\ 0 & 0 \end{pmatrix} \right) \alpha \\ \beta^\top \widehat{\Sigma}_{YY} \beta + \lambda_2 \|\beta_1\|^2 &= \beta^\top \left(\widehat{\Sigma}_{YY} + \lambda_2 \begin{pmatrix} I_{q_1} & 0 \\ 0 & 0 \end{pmatrix} \right) \beta.\end{aligned}$$

This leads us to the PRCCA modification of the correlation coefficient as follows:

$$\rho_{PRCCA}(\alpha, \beta; \lambda_1, \lambda_2) = \frac{\alpha^\top \widehat{\Sigma}_{XY} \beta}{\sqrt{\alpha^\top (\widehat{\Sigma}_{XX} + \lambda_1 \begin{pmatrix} I_{p_1} & 0 \\ 0 & 0 \end{pmatrix}) \alpha} \sqrt{\beta^\top (\widehat{\Sigma}_{YY} + \lambda_2 \begin{pmatrix} I_{q_1} & 0 \\ 0 & 0 \end{pmatrix}) \beta}}. \quad (4.2)$$

As usual, PRCCA variates and coefficients obtained by maximizing (4.2) can be found by means of SVD of the matrix $(\widehat{\Sigma}_{XX} + \lambda_1 \begin{pmatrix} I_{p_1} & 0 \\ 0 & 0 \end{pmatrix})^{-\frac{1}{2}} \widehat{\Sigma}_{XY} (\widehat{\Sigma}_{YY} + \lambda_2 \begin{pmatrix} I_{q_1} & 0 \\ 0 & 0 \end{pmatrix})^{-\frac{1}{2}}$.

4.2 PRCCA kernel trick

In this section we extend the kernel trick to the PRCCA problem set up. Note that, because I was replaced by block matrix $\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$ in the denominator of the modified correlation coefficient (2.4), the PRCCA problem does not preserve the property of invariance under orthogonal transformations. Thus, the mathematics used in Section 2.4 does not work anymore. There are two main ingredients for the PRCCA kernel trick. First, if the feature matrix consists of two orthogonal blocks, then the kernel trick can be applied to each block independently. Second, there

exists a non-orthogonal transformation of the feature matrix making the two blocks orthogonal to each other, while resulting in an equivalent PRCCA problem.

We again assume for simplicity that the regularization penalty is imposed on the X part only. Suppose $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, where random vectors $X_1 \in \mathbb{R}^{p_1}$ and $X_2 \in \mathbb{R}^{p_2}$ correspond to penalized coefficients part α_1 and unpenalized part α_2 , respectively. Let $\mathbf{X}_1 \in \mathbb{R}^{n \times p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{n \times p_2}$ represent the corresponding matrices of observations, so $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. To make the PRCCA solution identifiable we require \mathbf{X}_2 to be tall and full rank, that is, $p_2 < n$ and $\text{rank}(\mathbf{X}_2) = p_2$. We can also assume that $p_1 \gg n$.

As the first step we find a linear transformation $A \in \mathbb{R}^{p \times p}$ such that matrix $\tilde{\mathbf{X}} = \mathbf{X}A = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2) \in \mathbb{R}^{n \times p}$ has orthogonal blocks $\tilde{\mathbf{X}}_1 \in \mathbb{R}^{n \times p_1}$ and $\tilde{\mathbf{X}}_2 \in \mathbb{R}^{n \times p_2}$, that is, $\tilde{\mathbf{X}}_1^\top \tilde{\mathbf{X}}_2 = 0$, and that preserves the second block, that is, $\tilde{\mathbf{X}}_2 = \mathbf{X}_2$. This can be easily done by linear regression (see Supplement for details). This linear transformation maps the original PRCCA problem to an equivalent one (equivariant in terms of the coefficients and invariant in terms of the objective).

Note that the above transformation forces the sample covariance matrix $\widehat{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}$ to be block-diagonal with blocks $\widehat{\Sigma}_{\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1}$ and $\widehat{\Sigma}_{\tilde{\mathbf{X}}_2\tilde{\mathbf{X}}_2}$, which enables us to apply the kernel trick to the first and the second block of $\tilde{\mathbf{X}}$ independently. Specifically, consider the decomposition of $\tilde{\mathbf{X}}_1$, as in the RCCA kernel lemma, that is, $\tilde{\mathbf{X}}_1 = \mathbf{R}_1 V_1^\top$ where $V_1 \in \mathbb{R}^{p_1 \times n}$ is a matrix with orthonormal columns and $\mathbf{R}_1 \in \mathbb{R}^{n \times n}$ is some square matrix. Then the following lemma holds.

Lemma 4.2. [PRCCA kernel trick] The original PRCCA problem stated for \mathbf{X} and \mathbf{Y} can be reduced to solving the smaller PRCCA problem for $\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{X}_2 \end{pmatrix}$ and \mathbf{Y} . The resulting canonical correlations and variates for these two problems coincide. The canonical coefficients for the original problem can be recovered via the linear transformation $\alpha_X = A \begin{pmatrix} V_1 & 0 \\ 0 & I \end{pmatrix} \alpha_R$.

See Supplement Section 2 for the proof. Note that, according to the lemma, instead of working with large matrices $\widehat{\Sigma}_{XX} \in \mathbb{R}^{p \times p}$ and $\widehat{\Sigma}_{XY} \in \mathbb{R}^{p \times q}$ we can operate in terms of smaller matrices $\widehat{\Sigma}_{RR} \in \mathbb{R}^{(n+p_2) \times (n+p_2)}$ and $\widehat{\Sigma}_{RY} \in \mathbb{R}^{(n+p_2) \times q}$ thereby avoiding excessive computations.

4.3 Testing PRCCA on the Human Connectome data

First, we chose the brain region of interest that we aim to release from the regularization penalty. To do so for each column of averaged activation data we compute Cohen's d , which measures the effect size for a one-sample t-test comparing the population mean to zero, and pick the regions with at least a medium effect ($d > 0.3$). The resulting 26 regions demonstrate the largest activation during the Gambling task. Then we run PRCCA on the averaged data imposing the penalty on all but these

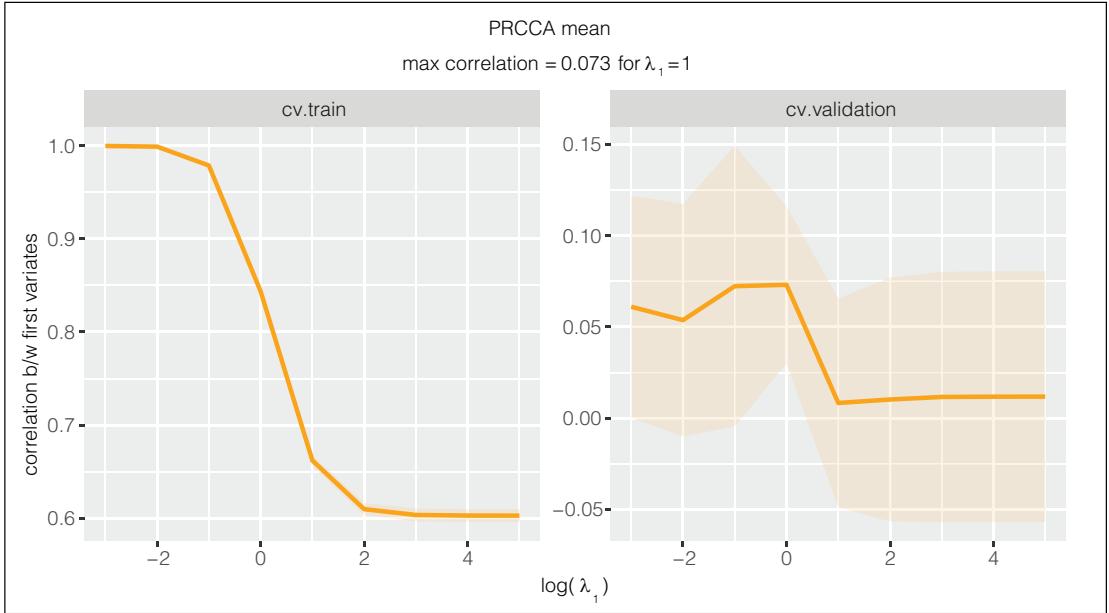


Figure 4 The cross-validation curves obtained via PRCCA with ten fold cross-validation for the Human Connectome dataset. Left panel: (unpenalized) correlation between train canonical variates. Right panel: (unpenalized) correlation between validation canonical variates

26 regions. We consider the same grid of values $\lambda_1 = 10^{-3}, 10^{-2}, \dots, 10^4, 10^5$ and choose the hyperparameter according to the maximum cross-validation score (see Section 2.5 for the details). The highest validation correlation is equal to 0.073 and is achieved when $\lambda_1 = 1$ (see Figure 4). Figure 5 represents the canonical coefficients computed for λ_1 chosen by cross-validation. As expected, the coefficients for most regions were shrunk to zero leaving only a few standing out. Again, it is the kernel trick which enables running cross-validation for the extremely high-dimensional feature matrix in just a few minutes. Although PRCCA does not perform well in this application, it will play an important role in developing subsequent methods.

5 Canonical correlation analysis for grouped data

5.1 Handling data with a group structure

The main critique of applying standard RCCA approach to the fMRI data is that, in fact, RCCA completely ignores the brain geometry treating all the features equally. Recall that in the Human Connectome data the features representing particular greyordinates are grouped into macro regions according to the function and anatomy. The goal of the *group regularized canonical correlation analysis (GRCCA)* is to incorporate this underlying data structure into the regularization penalty.

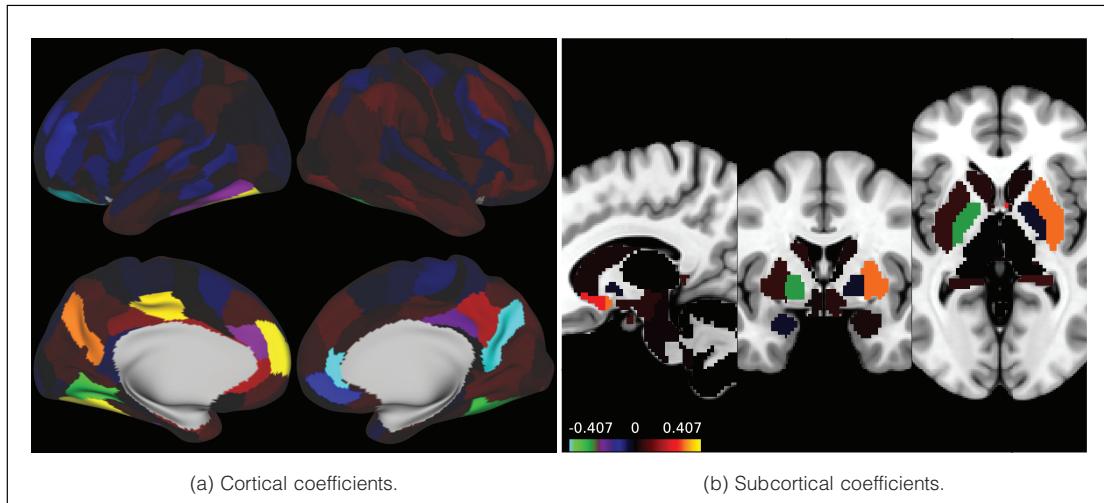


Figure 5 PRCCA coefficients computed for the averaged data (229 features) with hyperparameter $\lambda_1 = 1$ chosen by cross-validation

There are some group extensions of CCA based on elastic net and group lasso penalizations suggested in the literature (see, for example, Chen et al., 2012; Lin et al., 2013). Unlike the existing methods the proposed GRCCA approach does not require an iterative algorithm and has a simple explicit solution. Equipped with the kernel trick it also allows working with data in a very high-dimensional feature space.

GRCCA solves the CCA problem under the following two natural assumptions. First, we assume *homogeneity of groups* and expect that the features within each group have approximately equal contribution to the canonical variates. In other words, the corresponding CCA coefficients do not vary significantly inside each group. Second, we assume the *differentiating sparsity on a group level* and expect that the coefficients will be shrunk towards zero all together for some groups. In terms of brain imaging applications these two assumptions mean that greyordinates ‘act in concert’ within each macro region and that some regions have a weaker effect on the studied phenomenon.

To state the GRCCA optimization problem rigorously we need to introduce some further notation. Suppose that the elements of random vectors X and Y are known to come in K and L groups, respectively. For each $k = 1, \dots, K$ let p_k be the number of elements from X that belong to group k , so $\sum_{k=1}^K p_k = p$, and let $X_k \in \mathbb{R}^{p_k}$ be the random vector that consists of these elements. Without loss of generality, we can assume that $X = \begin{pmatrix} X_1 \\ \vdots \\ X_K \end{pmatrix}$. Similarly, one can divide the CCA coefficients α into blocks corresponding to different groups, that is, $\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{pmatrix}$. If $\bar{\alpha}_k = \frac{1^\top \alpha_k}{p_k}$ denotes the mean of the CCA coefficients in group k , then the group homogeneity assumption implies

that all values of α_k do not deviate significantly from the mean value $\bar{\alpha}_k$, whereas, the differentiating sparsity on a group level implies that the average deviation of $\bar{\alpha}_k$ from zero is small. This can be characterized by the following two constraints: $\sum_{k=1}^K \|\alpha_k - \mathbb{1}\bar{\alpha}_k\|^2 \leq t_1$ and $\sum_{k=1}^K p_k \bar{\alpha}_k^2 \leq s_1$. Note that these two equations can be interpreted as bounds on within- and between- group variation, respectively. One can derive similar constraints for β coefficients as well. Adding all the constraints to the CCA optimization problem we end up with the GRCCA optimization problem

$$\begin{aligned} & \text{maximize } \alpha^\top \widehat{\Sigma}_{XY} \beta \text{ w.r.t. } \alpha \in \mathbb{R}^p \text{ and } \beta \in \mathbb{R}^q \\ & \text{subject to } \alpha^\top \widehat{\Sigma}_{XX} \alpha = 1, \quad \sum_{k=1}^K \|\alpha_k - \mathbb{1}\bar{\alpha}_k\|^2 \leq t_1, \quad \sum_{k=1}^K p_k \bar{\alpha}_k^2 \leq s_1 \\ & \text{and } \beta^\top \widehat{\Sigma}_{YY} \beta = 1, \quad \sum_{\ell=1}^L \|\beta_\ell - \mathbb{1}\bar{\beta}_\ell\|^2 \leq t_2, \quad \sum_{\ell=1}^L q_\ell \bar{\beta}_\ell^2 \leq s_2. \end{aligned} \quad (5.1)$$

Next, denote $C_m = \frac{\mathbb{1}\mathbb{1}^\top}{m} \in \mathbb{R}^{m \times m}$. Let $C_X = C_{p_1} \oplus \dots \oplus C_{p_K} \in \mathbb{R}^{p \times p}$ refer to the block diagonal matrix with blocks C_{p_1}, \dots, C_{p_K} . Thus, the constraints on α can be rewritten in terms of a regularization penalty as

$$\begin{aligned} & \alpha^\top \widehat{\Sigma}_{XX} \alpha + \lambda_1 \sum_{k=1}^K \|\alpha_k - \mathbb{1}\bar{\alpha}_k\|^2 + \mu_1 \sum_{k=1}^K p_k \bar{\alpha}_k^2 = \\ & \alpha^\top \widehat{\Sigma}_{XX} \alpha + \lambda_1 \sum_{k=1}^K \alpha_k^\top (I - C_{p_k}) \alpha_k + \mu_1 \sum_{k=1}^K \alpha_k^\top C_{p_k} \alpha_k = \\ & \alpha^\top (\widehat{\Sigma}_{XX} + \lambda_1(I - C_X) + \mu_1 C_X) \alpha = \alpha^\top (\widehat{\Sigma}_{XX} + K_X(\lambda_1, \mu_1)) \alpha. \end{aligned}$$

Here, $K_X(\lambda_1, \mu_1) = \lambda_1(I - C_X) + \mu_1 C_X$ is the penalty matrix. The constraints for β can be combined in a similar way leading to the GRCCA modified correlation coefficient

$$\rho_{GRCCA}(\alpha, \beta; \lambda_1, \mu_1, \lambda_2, \mu_2) = \frac{\alpha^\top \widehat{\Sigma}_{XY} \beta}{\sqrt{\alpha^\top (\widehat{\Sigma}_{XX} + K_X(\lambda_1, \mu_1)) \alpha} \sqrt{\beta^\top (\widehat{\Sigma}_{YY} + K_Y(\lambda_2, \mu_2)) \beta}}.$$

Note that this correlation coefficient has similar structure to the RCCA coefficient (2.4) and PRCCA coefficient (4.2), but now the covariance matrices in the denominator are adjusted by block diagonal matrices $K_X(\lambda_1, \mu_1)$ and $K_Y(\lambda_2, \mu_2)$.

Similar to RCCA and PRCCA, the explicit solution to the GRCCA problem can be found via the SVD of matrix $(\widehat{\Sigma}_{XX} + K_X(\lambda_1, \mu_1))^{-\frac{1}{2}} \widehat{\Sigma}_{XY} (\widehat{\Sigma}_{YY} + K_Y(\lambda_2, \mu_2))^{-\frac{1}{2}}$, which can be problematic in high dimensions. It turns out that there is a simple linear transformation that converts the GRCCA problem to an equivalent RCCA/PRCCA

problem. In Supplement Sections 4 and 5 we give two ways: via the SVD of the penalty matrix and via feature matrix extension. This link can be subsequently used to establish the kernel trick for group-structured data thereby reducing computations in high dimensions.

5.2 Link to the flexibility vs. performance trade-off

There are several important properties of the proposed penalty matrix explaining the motivation for the GRCCA method. First, one can show that for $K_X(\lambda_1, \lambda_1) = \lambda_1 I$, so RCCA is the special case of GRCCA. Second, increasing λ_1 restrains the variability of coefficients within each brain region and, in limit, makes all the coefficients that belong to the same brain region equal to each other (and equal to the region mean). This is essentially equivalent to replacing features in each brain region by the average. Therefore, when $\lambda_1 \rightarrow \infty$ the GRCCA problem becomes equivalent to the RCCA problem solved for the reduced data (see Section 3.1 for the details). To sum up, varying λ_1 and μ_1 allows us to approach the RCCA method conducted for either full or reduced data thereby controlling the flexibility vs interpretability trade-off described in Section 3.2.

5.3 GRCCA for the Human Connectome study

In this section, we apply the GRCCA method to the Human Connectome study data grouping activation features according to the brain regions. We again adjusted X and Y for the sex effect and ran ten fold cross-validation on the adjusted data pair with the penalty factors varying in the grid $\lambda_1, \mu_1 = 10^{-3}, 10^{-2}, \dots, 10^4, 10^5$ (see Section 2.5 for the details). The resulting cross-validation curves are presented in Figure 6. According to the plot, the highest cross-validation score is attained for $\lambda_1 = 100$ and $\mu_1 = 1$ leading to the correlation of 0.296. To validate these significant findings, we again run nested cross-validation (see Section 3.1). According to Figure 7, the NCV average test set score is equal to 0.236. Thus, although slightly optimistic (by a modest value of 0.06), the GRCCA cross-validation correlation is not a spurious finding and is a significant improvement comparing to the RCCA method.

In addition to better performance, the GRCCA method allows us to track the effect of the variation inside each brain region (controlled by the λ_1 penalty factor) on the resulting canonical correlation separately from the effect of the variation across the regions (controlled by μ_1 penalty factor). For example, the spikes for small μ_1 values suggest that it is more beneficial to reduce within group variation than between group one. In other words, shrinking all brain region coefficients towards the group means improves the performance more than shrinking group means towards zero. Moreover, for small μ_1 it is the ratio of hyperparameters that plays the key role: the highest score is always achieved when $\frac{\lambda_1}{\mu_1} = 100$. For large μ_1 this pattern disappears as we start to over-penalize both between and within group variations.

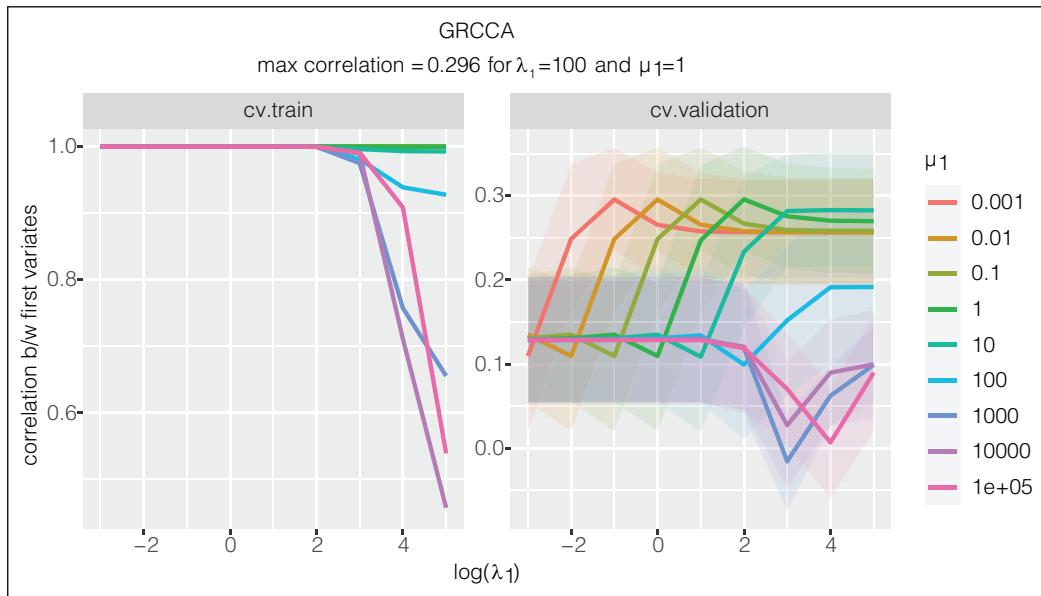


Figure 6 The cross-validation curves obtained via GRCCA with ten fold cross-validation for the Human Connectome Project dataset. Left panel: (unpenalized) correlation between train canonical variates. Right panel: (unpenalized) correlation between validation canonical variates

Note: For colour figure, please refer to the online version.

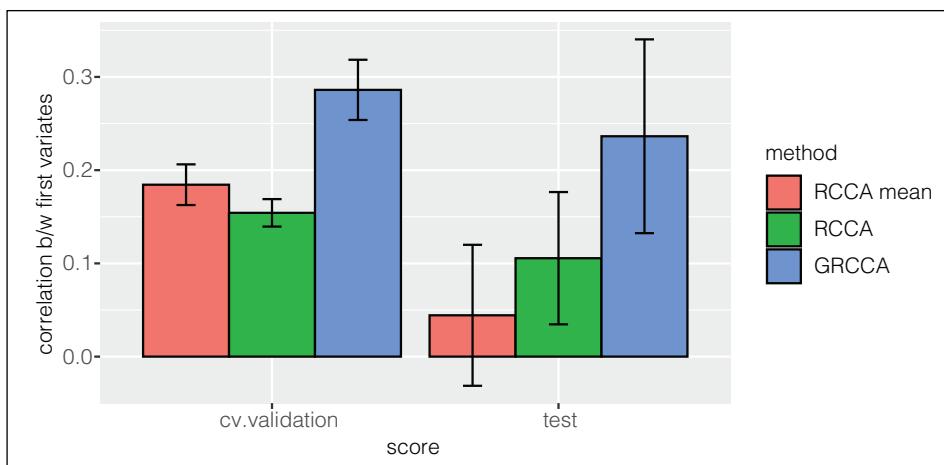


Figure 7 Nested cross-validation scores computed for three models: RCCA fitted on the averaged data (red, 229 features), RCCA fitted on the original data (green, 90 368 features), GRCCA (blue). Two scores are reported: *cv.validation* = maximum score obtained via ten fold cross-validation, averaged across 11 NCV folds; *test* = score computed on independent test set, averaged across 11 NCV folds

5.4 Using GRCCA for visualization

In this section we demonstrate another advantage of GRCCA in the context of visualization and interpretability. In Figure 8, we present the coefficient paths (α vs. λ_1) produced by the RCCA method as well as the group modification. Here different colours represent different brain regions. According to the plot, we observe the following behaviour of the coefficients. For the RCCA method the canonical coefficients are shrunk towards zero all together with the growth of λ_1 . On the contrary, for large λ_1 all the GRCCA coefficient paths become horizontal, which implies the convergence of the coefficients to the group means. Finally, increasing μ_1 shrinks all the group means towards zero encouraging differentiating sparsity on a group level.

In Figure 9, we present the brain images computed for $\mu_1 = 1$ and $\lambda_1 = 1, 10, 100, 1\,000$. Note that larger λ_1 makes the brain region pattern more obvious (similar to Figures 3a–3b). Moreover, for $\lambda_1 = 100$ we get the plot corresponding to the best GRCCA model from Section 5.3. Thus, the GRCCA model chosen by cross-validation has not only better performance on the validation set than the RCCA model, but it is also more interpretable in the context of the importance of each brain region. Specifically, the canonical component had especially high positive loadings in subcortical regions involved in reward processing, such as the striatum (nucleus accumbens, putamen) and thalamus (Haber, 2017). It also loaded positively on a cortical network encompassing the temporal lobe, dorsolateral prefrontal, dorsomedial prefrontal, posterior cingulate and precentral cortices (see Figure 9 for an annotated visualization of these results). Most of these regions have been shown to be connected to the striatum and to be part of key reward-processing pathways as well (Haber, 2017).

6 General approach to regularization

It turns out that all RCCA, PRCCA and GRCCA methods are similar in nature: they perform regularization by means of adjusting covariance matrices $\widehat{\Sigma}_{XX}$ and/or $\widehat{\Sigma}_{YY}$ in the denominator of the modified correlation coefficient. In this section, we consider the class of CCA problems with general weighted ℓ_2 regularization.

If $K_X, K_Y \in \mathbb{R}^{p \times p}$ are some positive semi-definite *penalty matrices*, then the general modified correlation coefficient can be written as

$$\rho(\alpha, \beta; K_X, K_Y) = \frac{\alpha^\top \widehat{\Sigma}_{XY} \beta}{\sqrt{\alpha^\top (\widehat{\Sigma}_{XX} + K_X) \alpha} \sqrt{\beta^\top (\widehat{\Sigma}_{YY} + K_Y) \beta}}. \quad (6.1)$$

The accompanying *general RCCA* optimization problem is therefore

$$\begin{aligned} & \text{maximize } \alpha^\top \widehat{\Sigma}_{XY} \beta \text{ w.r.t. } \alpha \in \mathbb{R}^p \text{ and } \beta \in \mathbb{R}^q \\ & \text{subject to } \alpha^\top \widehat{\Sigma}_{XX} \alpha = 1, \alpha^\top K_X \alpha \leq t_1 \text{ and } \beta^\top \widehat{\Sigma}_{YY} \beta = 1, \beta^\top K_Y \beta \leq t_2. \end{aligned} \quad (6.2)$$

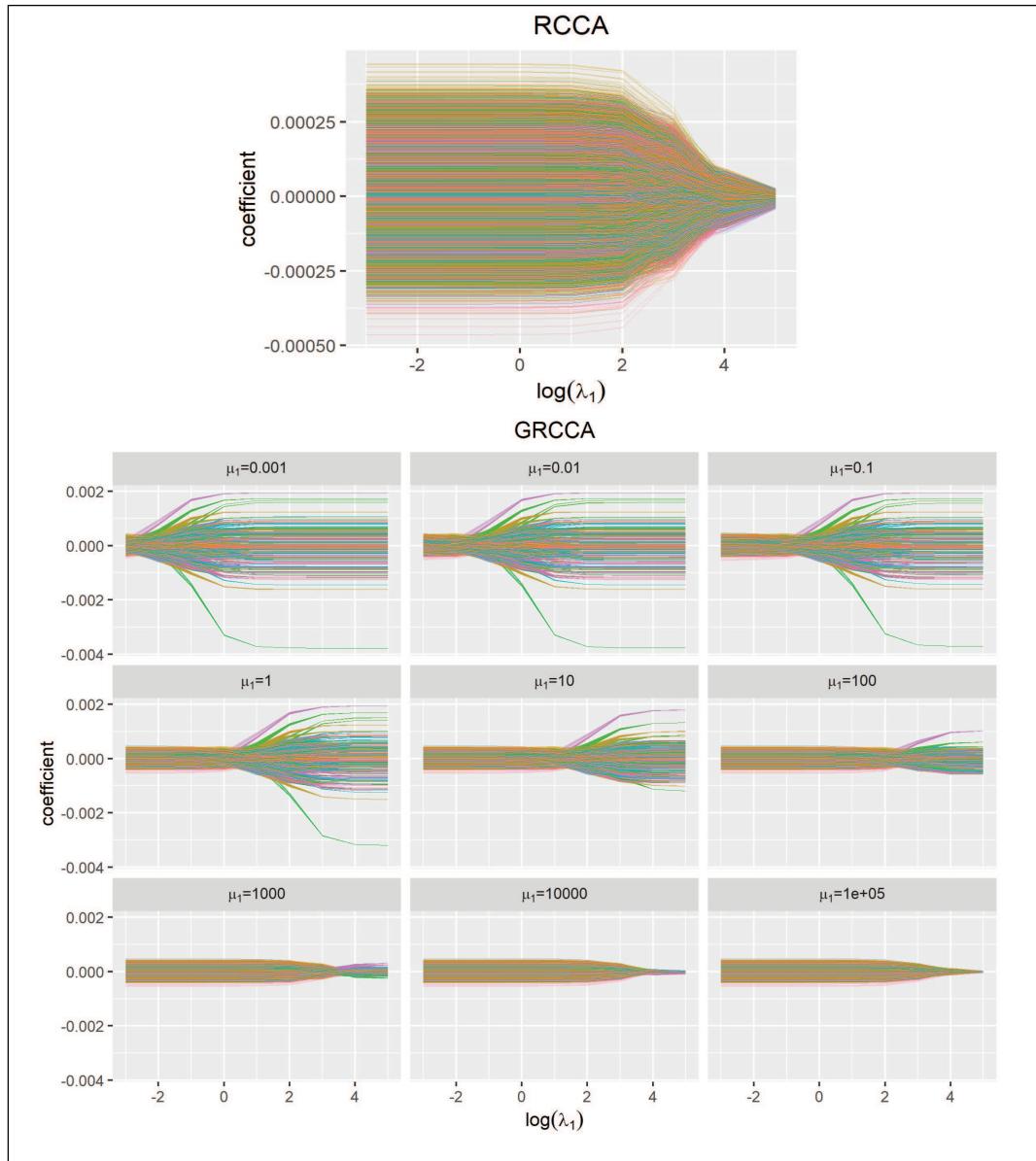


Figure 8 Coefficient paths for RCCA and GRCCA with colours representing the regions. The RCCA method shrinks canonical coefficients towards zero all together with the growth of λ_1 . The GRCCA method shrinks them towards the group means with the growth of λ_1 , whereas increasing μ_1 shrinks all the group means towards zero

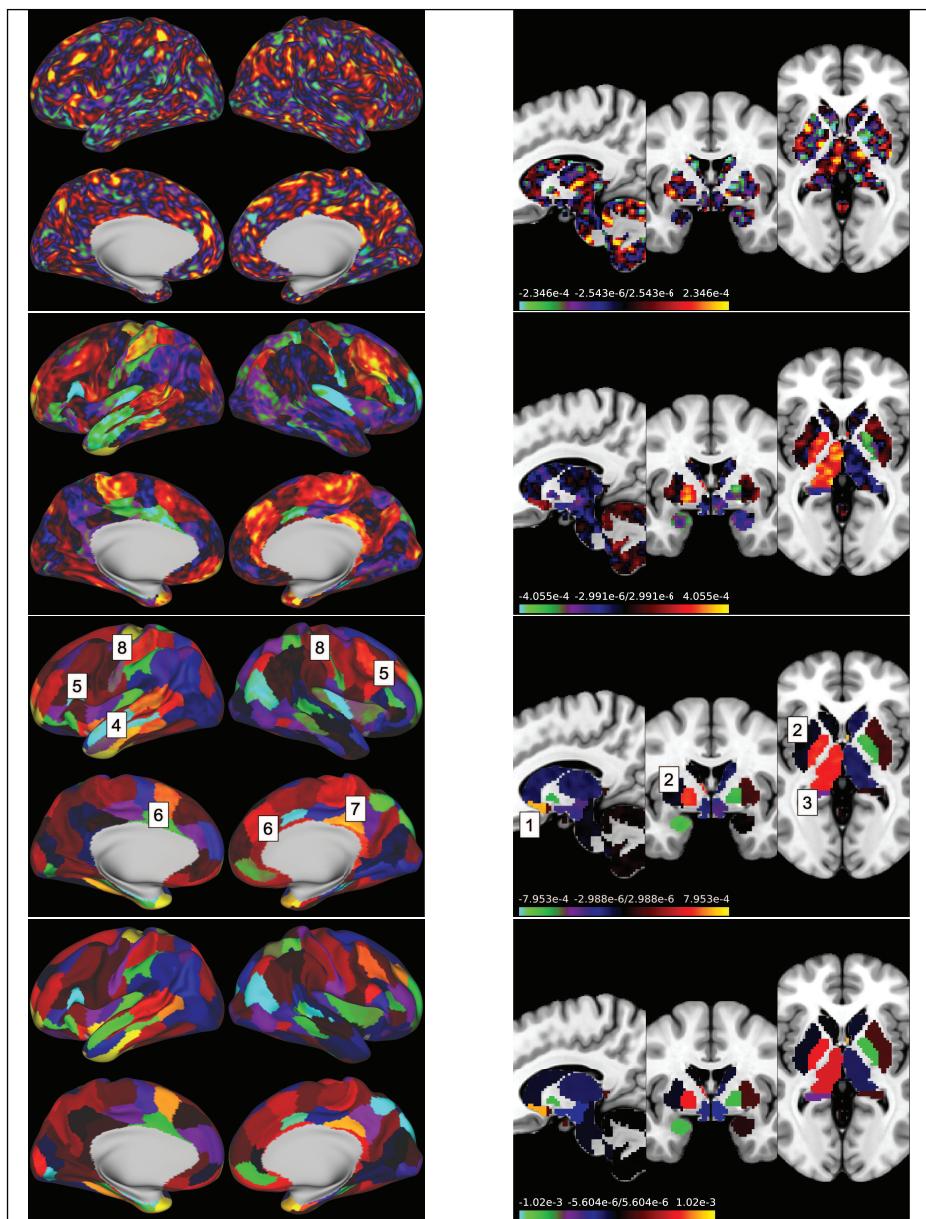


Figure 9 From top to bottom: GRCCA coefficients for $\mu_1 = 1$ and $\lambda_1 = 1, 10, 100, 1000$. The third row represents the solution produced by the cross-validation procedure. Annotation of brain regions: [1] nucleus accumbens, [2] putamen, [3] thalamus, [4] temporal lobe, [5] dorsolateral prefrontal cortex, [6] dorsomedial prefrontal cortex, [7] posterior cingulate cortex, [8] precentral cortex

Note that the inequality constraints in (6.2) can be rewritten as $\|\alpha\|_{K_X} \leq t_1$ and $\|\beta\|_{K_Y} \leq t_2$, where $\|\cdot\|_A$ is weighted Euclidean norm defined as $\|x\|_A = x^\top A x$. The resulting canonical variates and coefficients can be found via the singular value decomposition of the matrix $(\widehat{\Sigma}_{XX} + K_X)^{-\frac{1}{2}} \widehat{\Sigma}_{XY} (\widehat{\Sigma}_{YY} + K_Y)^{-\frac{1}{2}}$. To handle General RCCA in high dimensions, one can link it to the two methods for which we already established the kernel trick. In the Supplement Section 3 we provide the proof of the following lemma.

Lemma 6.3. [General RCCA to RCCA/PRCCA] If both K_X and K_Y are positive definite then, by some proper change of basis, the general RCCA problem can be reduced to the RCCA one. Alternatively, if one of K_X and K_Y has zero eigenvalues then general RCCA boils down to solving the PRCCA problem with number of unpenalized coefficients equal to the multiplicity of the zero eigenvalue.

7 Simulation study

7.1 Generating data with a group structure

In this section, we set up a small simulation experiment where we compare performance of all the above methods on the data with group structure. We generate the data as follows. For random vector X we assume that it is grouped into K groups of equal size, thereby having $p_k = \frac{p}{K}$ variables in group k . Each group of X is generated by one of K centroid random variables and is obtained by adding some Gaussian noise to the centroid. Moreover, we assume the presence of some correlation between the centroids and Y . To be precise, to generate the data we exploit the multivariate normal distribution as a joint distribution of random vector $Y \in \mathbb{R}^q$ and random vector of centroids $X^c \in \mathbb{R}^K$:

$$(Y, X^c) \sim \mathcal{N}_{q+K}(0, \Sigma) \quad \text{with} \quad \Sigma = \begin{pmatrix} I_q & \mathbf{1}\mathbf{1}^\top \sigma_{XY}^2 \\ \mathbf{1}\mathbf{1}^\top \sigma_{XY}^2 & I_K \end{pmatrix}.$$

Next, we generate random vector $X_k \in \mathbb{R}^{p_k}$ corresponding to groups $k = 1, \dots, K$ from the distribution $X_k | X^c \sim \mathcal{N}_{p_k}(\mathbf{1}X_k^c, \sigma_X^2 I)$, where X_k^c is the k th component of the centroid vector. Finally, we obtain matrices $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_K) \in \mathbb{R}^{n \times p}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$ by drawing n samples from the above distributions. 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.

As the next step, we run RCCA, PRCCA and GRCCA on the generated data imposing the regularization on the X part only and using the following hyperparameters. For all methods the penalty factor is chosen to be $\lambda_1 = 10^{-5}, 10^{-4}, \dots, 10^4, 10^5$. For PRCCA we penalize $p_1 = 10$ variables only leaving $p_2 = 5$ variables untouched; these five unpenalized variables correspond to the first features in each group. For GRCCA we again try two versions. First, we run GRCCA

with $\mu_1 = 0$, that is, without differentiating sparsity on a group level assumption. Next, we add differentiating sparsity to GRCCA and vary the penalty factor in the range $\mu_1 = 10^{-4}, 10^{-3}, \dots, 1, 10$.

We compare all methods in terms of resulting correlations. For this purpose, we generate 1 000 train and test sets, fit models on train and evaluate canonical correlation value on test. We plot average train and test correlations as well as their one standard error intervals vs. penalty factor λ_1 (see Figure 10). According to the plot, for the correlated data the best test score is achieved by sparse version of GRCCA, which significantly outperforms RCCA. Better performance can be explained by the presence of the groups structure in the data. Note that non-sparse GRCCA also looses in terms of the test score. The possible reason is that number of observations ($n = 10$) is only twice as large as the number of groups ($K = 5$), so regularization on a group level helps to prevent overfitting to train data. In the case of independent data, all the competitors perform in a similar way: The average

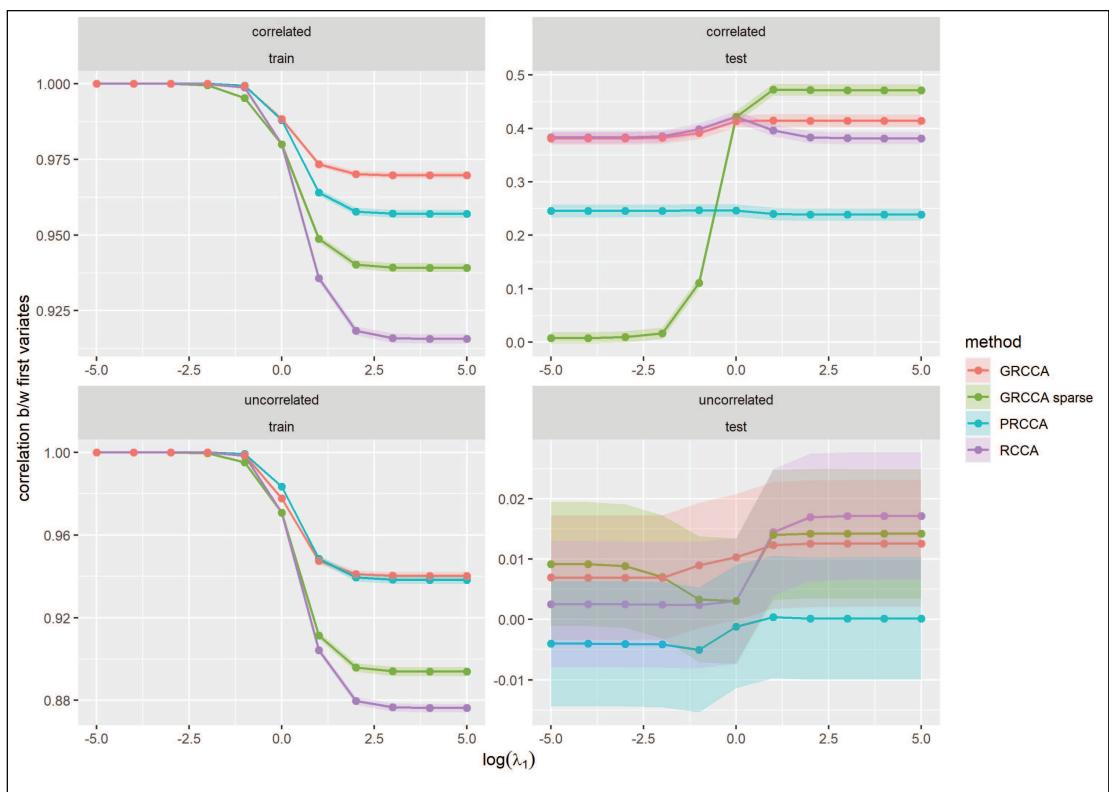


Figure 10 Train and test curves computed via simulation. Four models presented: RCCA, PRCCA and GRCCA with zero ($\mu_1 = 0$) and non-zero sparsity ($\mu_1 = 1$). First row: train and test correlation obtained for data with correlation ($\sigma_{XY} = 0.5$). Second row: train and test correlation obtained for uncorrelated data ($\sigma_{XY} = 0$)

Note: For colour figure, please refer to the online version.

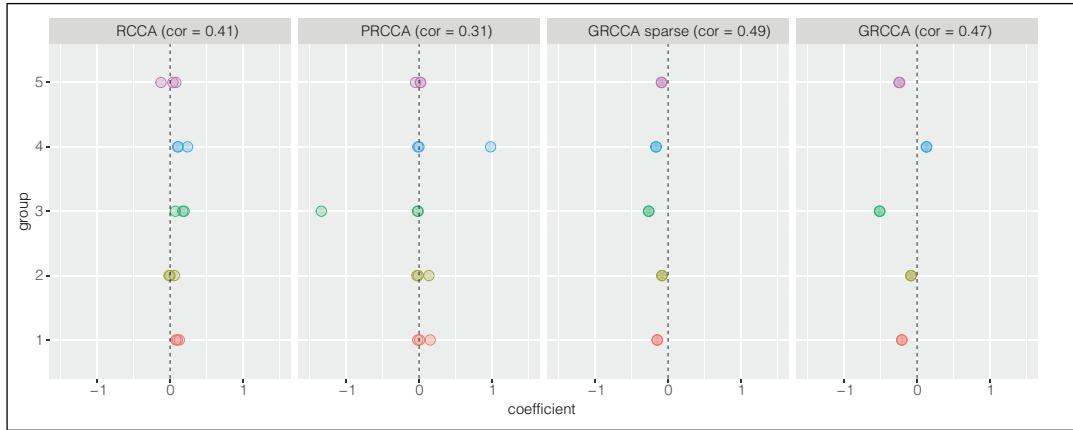


Figure 11 Coefficient values obtained via RCCA, PRCCA and two versions of GRCCA with hyperparameter λ_1 chosen to maximize the test correlation. In this plot colour corresponds to the group number

test correlation is very close to zero regardless the hyperparameter value; the test correlation curves are almost flat.

Finally, for each model we pick the value of λ_1 according to the maximum test score and compare the CCA coefficients α for the chosen models. Figure 11 displays the main difference between the RCCA and GRCCA methods. Although both techniques aim to reduce the data dimensionality, the reduction is achieved in a different way. Unlike RCCA, which treats all the coefficients equally and confines their deviation from zero, GRCCA carries out the reduction treating equally the coefficients inside each group and removing the within group noise. To sum up, in the presence of a group structure the group modification of RCCA allows for dimensionality reduction in a more efficient and interpretable way.

8 Discussion

In this article, we proposed several approaches to the CCA regularization. The introduced PRCCA technique has a similar flavour as RCCA, but it penalizes only a subset of canonical coefficients. Both of these methods combined with the proposed kernel trick allows us to find the CCA solution even in case of extremely high data dimensionality. We further present the GRCCA method, which is based on the underlying group structure of the data and which, therefore, can be useful in some applications, and extend regularization to the case of a more general regularization penalty thereby proposing General RCCA. The close connection between the latter techniques with RCCA and PRCCA methods enables to utilize the kernel trick in the general case thus providing a powerful tool for regularizing CCA in the high-dimensional framework.

There is still much scope for future work. One interesting direction for further research is to consider other applications of the proposed group RCCA technique. For

example, there are many problems in genetics where genes are grouped by functional similarity. Further, in this article we cover only two types of penalties: partial and group; although the proposed kernel trick can handle any ℓ_2 -type penalty (see Section 6 for general RCCA). Thus we can study other structured modifications of RCCA that can be beneficial for applications. As an example, it may be interesting to explore hierarchical group structure, where not only brain loci are combined in some regions, but also regions are combined in some groups (e.g., we have cortical and subcortical groups in the HCP study).

From the computational point of view, it would be useful to investigate how one can optimize the choice of the hyperparameters. The following idea is inspired by ridge regression, which also uses the ℓ_2 penalty. Note that currently it is not necessary to apply any data normalization before running regularized CCA (all the fMRI features have the same scale). However, the overall scale of X influences the choice of the hyperparameters, that is, multiplying X by some number a implies increasing the penalty factors by a^2 times as well. Therefore, it would be beneficial to develop some recommendations for the grid of hyperparameters the user should search through. For instance, we can use the ridge regression heuristic and introduce a concept of degrees-of-freedom (aka df) for CCA with regularization. Then, we can base the hyperparameter recommendation on the df value.

Software

Proposed methods are implemented in the R package RCCA; the software is available from Github (<https://github.com/ElenaTuzhilina/RCCA>).

Supplementary material

Supplementary material is available online <http://www.statmod.org/smij/archive.html>.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship and/or publication of this article: Leonardo Tozzi was supported by grant U01MH109985 under PAR-14-281 from the National Institutes of Health. Trevor

Hastie was partially supported by grants DMS-2013736 and IIS 1837931 from the National Science Foundation, and grant 5R01 EB 001988-21 from the National Institutes of Health.

References

- Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, Glasser MF, Curtiss S, Dixit S, Feldt C, Nolan D, Bryant E, Hartley T, Footer O, Bjork JM, Poldrack R, Smith S, Johansen-Berg H, Snyder AZ and Van Essen DC (2013) Function in the Human Connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, **80**, 169–89.
- Cao K-AL, Martin P, Robert-Granié C and Besse P (2009) Sparse canonical methods for biological data integration: Application to a cross-platform study. *BMC Bioinformatics*, **10**, 1–17.
- Carver CS and White TL (1994) Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, **67**, 319–33.
- Chen X, Han L and Carbonell J (2012) Structured sparse canonical correlation analysis. *Proceedings of Machine Learning Research*, **22**, 199–207.
- Desikan RS, Segonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS and Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, **31**, 968–80.
- Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, Yang Z, Chu C, Xie S, Laird AR, Fox PT, Eickho SB, Yu C and Jiang T (2016) The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral Cortex*, **26**, 3508–26.
- Gonzalez I, Déjean S, Martin P and Baccini A (2008) CCA: An R Package to extend canonical correlation analysis. *Journal of Statistical Software*, **23**, 1–14.
- Haber SN (2017) Chapter 1: Anatomy and connectivity of the reward Circuit. In *Decision Neuroscience*, edited by JC Dreher and L Tremblay, pages 3–19. San Diego, CA: Academic Press.
- Härdle, W. and Simar, L. (2007). *Applied Multivariate Statistical Analysis*, 2nd edition. Berlin: Springer.
- Hardoon DR and Shawe-Taylor J (2011) Sparse canonical correlation analysis. *Machine Learning*, **83**, 331–53.
- Hardoon D, Szekely S and Shawe-Taylor J (2005) Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, **16**, 2639–64.
- Hotelling H (1936) Relations between two sets of variables. *Biometrika*, **28**, 321–77.
- Kuss M (2003) The geometry of kernel canonical correlation analysis (Technical report).
- Leurgans S, Moyeed R and Silverman B (1993) Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society*, **55**, 725–40.
- Lin D, Zhang J, Li J, Calhoun V, Deng H-W and Wang Y-P (2013) Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*, **14**, Article 245.
- Lykou A and Whittaker J (2010) Sparse CCA using a Lasso with positivity constraints. *Computational Statistics and Data Analysis*, **54**, 3144–57.
- Mardia K, Kent JT and Bibby JM (1979) *Multivariate Analysis*. New York, NY: Academic Press.
- Parkhomenko E, Tritchler D and Beyene J (2009) Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, **8**, 1–34.
- Tozzi L, Staveland B, Holt-Gosselin B, Chesnut M, Chang SE, Choi D, Shiner M, Wu H,

- Lerma-Usabiaga G, Sporns O, Barch DM, Gotlib IH, Hastie TJ, Kerr AB, Poldrack RA, Wandell BA, Wintermark M and Williams LM (2020) 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**, 116715.
- Vinod H (1976) Canonical ridge and econometrics of joint production. *Journal of Econometrics*, **4**, 147–166 of *Econometrics*, **4**(2), 147–66.
- Waaijenborg S, de Witt Hamer PCV and Zwinderman A (2008) Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, **7**, Article 3.
- Wang H-T, Smallwood J and Mourao-Miranda J (2020) Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. *Neuroimage*, **216**, 116745.
- Wardenaar KJ, van Veen T, Giltay EJ, de Beurs E, Penninx BWJH and Zitman FG (2010) Development and validation of a 30-item short adaptation of the Mood and Anxiety Symptoms Questionnaire (MASQ). *Psychiatry Research*, **179**, 101–6.
- Watson D, Clark LA and Tellegen A (1988) Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, **54**, 1063–70.
- Witten D and Tibshirani R (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, **8**, Article 28.
- Zhuang X, Yang Z and Cordes D (2020) A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, **41**, 3807–33.