

## Multi-output predictions from neuroimaging: assessing reduced-rank linear models

Mehdi Rahim, Bertrand Thirion, Gaël Varoquaux

► **To cite this version:**

Mehdi Rahim, Bertrand Thirion, Gaël Varoquaux. Multi-output predictions from neuroimaging: assessing reduced-rank linear models. PRNI 2017 - The 7th International Workshop on Pattern Recognition in Neuroimaging, Jun 2017, Toronto, Canada. pp.1 - 4, 2017, <<https://sites.google.com/site/prni2017/home>>. <10.1109/PRNI.2017.7981504>. <hal-01547572v2>

**HAL Id: hal-01547572**

**<https://hal.inria.fr/hal-01547572v2>**

Submitted on 30 Oct 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Multi-output predictions from neuroimaging: assessing reduced-rank linear models

Mehdi Rahim, Bertrand Thirion, Gaël Varoquaux  
 Parietal - Inria / CEA. Paris-Saclay University. France  
 mehdi.rahim@inria.fr

**Abstract**—Typical neuroimaging studies analyze associations between physiological or behavioral traits and brain structure or function. Some rely on predicting these scores from neuroimaging data. To explain association between brain features and multiple traits, reduced-rank regression (RRR) models are often used, such as canonical correlation analysis (CCA) and partial least squares (PLS). These methods estimate latent variables, or canonical modes, that maximize the covariations between neuroimaging features and behavioral scores. Here, we investigate theoretically and empirically the extent to which reduced-rank models predict out-of-sample clinical scores from functional connectivity. Experiments on a schizophrenia dataset show that *i)* significant correlations between canonical modes do not necessarily mean accurate generalization on unseen data, and *ii)* better accuracy is achieved when taking into account regularized covariance between scores.

## I. INTRODUCTION

Large-scale population neuroimaging studies collect brain imaging data together with a rich description of subject behavior, genetics, and health status. Analyzing these data is done through models that relate brain activity to personal traits or neuropsychiatric profiles. They explain phenotypic differences by brain variations measured from various imaging modalities. The goal is to better understand inter-subject differences [1]. In particular, extracting biomarkers with predictive models is a key element to better characterize brain disorders [2]. In this context, the accuracy of predictive models on independent data measures how strong the brain-behavior relationship is [3]. Joint analyses of multiple behavioral scores typically rely on multi-output methods such as canonical correlation analysis (CCA) or partial least squares (PLS). Such models find maximal correlations between pairs of latent variables of imaging descriptors and subject profiles. CCA and PLS are variants of reduced-rank regression (RRR) models, a general formulation for low-rank linear models. These models have been used to link brain connectivity and behavior [4] [5], or brain and genetics [6] [7], or to characterize patients with schizophrenia [8] or Alzheimer’s Disease [9]. Such studies evaluate the model accuracy by measuring correlations between the computed canonical modes. Yet, the predictive capacity of those models needs to be assessed.

In this paper, we review different variants of RRR including PLS and CCA for out-of-sample prediction of clinical scores from functional-connectivity measures. First, we describe the relationship between RRR, CCA, and PLS. Then, we present experiments on resting-state functional magnetic resonance

imaging (rfMRI) data to characterize multiple clinical scores of subjects and patients with schizophrenia. Results suggest that high canonical correlations may not be a sufficient indicator of the model accuracy, and that the metric –Euclidean versus Mahalanobis– used on the outputs to measure the quality of the prediction has a strong impact on the conclusions.

## II. REDUCED RANK LINEAR MODELS

We first detail the widely used linear models for regression, from the ordinary least squares formulation to the low-rank models. In cross-subject prediction settings, we denote  $n, p, q$  the number of subjects, the number of imaging features, and the output size, respectively.  $\mathbf{X} \in \mathbb{R}^{n \times p}$  is the feature matrix: each row represents the vectorized functional connectivity for a subject.  $\mathbf{Y} \in \mathbb{R}^{n \times q}$  is the score matrix: each row represents multiple scores (e.g.: age, IQ, ...) for a subject. We assume that  $\mathbf{X}$  and  $\mathbf{Y}$  are centered.

### A. Generalized least squares

The general linear model estimates the relationship between the measurements  $\mathbf{X}$  and the responses  $\mathbf{Y}$ ,

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{E}, \quad (1)$$

$\mathbf{W} \in \mathbb{R}^{p \times q}$  is the coefficient matrix that relates  $\mathbf{X}$  and  $\mathbf{Y}$ .  $\mathbf{E} \in \mathbb{R}^{n \times q}$  is the model error, or residuals. Ordinary least squares solution is :  $\widehat{\mathbf{W}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ . Generalized least squares minimize squared Mahalanobis distance  $\text{tr}((\mathbf{Y} - \mathbf{X}\mathbf{W})\mathbf{\Gamma}(\mathbf{Y} - \mathbf{X}\mathbf{W})^T)$  where  $\mathbf{\Gamma} \in \mathbb{R}^{q \times q}$  is the whitening matrix.

### B. Reduced-rank regression

RRR [10] considers a low-rank approximation of the linear model coefficients  $\mathbf{W}_k = \mathbf{B}\mathbf{A}$ ,

$$\mathbf{Y} = \mathbf{X}\mathbf{B}\mathbf{A} + \mathbf{E}, \quad (2)$$

with  $\mathbf{B} \in \mathbb{R}^{p \times k}$  and  $\mathbf{A} \in \mathbb{R}^{k \times q}$ .  $\mathbf{A}$  and  $\mathbf{B}$  are obtained by minimizing generalized least squares such that  $\mathbf{\Gamma} \in \mathbb{R}^{q \times q}$  is a whitening matrix on the outputs  $\mathbf{Y}$ . The inverse empirical covariance is often taken as  $\mathbf{\Gamma} = (\mathbf{Y}^T \mathbf{Y})^{-1}$ . We show in the following subsections that depending on how  $\mathbf{\Gamma}$  is set, RRR can be a generalization of PLS and CCA. RRR solutions are estimated by computing  $\mathbf{H}_{\text{RRR}} \in \mathbb{R}^{p \times k}$ , the first  $k$  normalized eigenvectors associated with the  $k$  first eigenvalues of  $\mathbf{R}_{\text{RRR}} = \mathbf{\Gamma}^{\frac{1}{2}} \mathbf{Y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \mathbf{\Gamma}^{\frac{1}{2}}$ . The solution is:

$$\widehat{\mathbf{A}}_{\text{RRR}} = \mathbf{H}_{\text{RRR}}^T \mathbf{\Gamma}^{-\frac{1}{2}}, \quad (3)$$

$$\widehat{\mathbf{B}}_{\text{RRR}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \mathbf{\Gamma}^{\frac{1}{2}} \mathbf{H}_{\text{RRR}}. \quad (4)$$

### C. Canonical correlation analysis

CCA [11] uses a low-rank model that maximizes the correlation between the latent variables  $\mathbf{X}\mathbf{u}_j$  and  $\mathbf{Y}\mathbf{v}_j$  ( $j = 1..k$ ):

$$\max_{\|\mathbf{u}_j\|_2=\|\mathbf{v}_j\|_2=1} \frac{\mathbf{u}_j^T \mathbf{X}^T \mathbf{Y} \mathbf{v}_j}{\sqrt{\mathbf{u}_j^T \mathbf{X}^T \mathbf{X} \mathbf{u}_j} \sqrt{\mathbf{v}_j^T \mathbf{Y}^T \mathbf{Y} \mathbf{v}_j}}. \quad (5)$$

This formulation can be seen as a variant of RRR such that the whitening matrix  $\mathbf{\Gamma} = (\mathbf{Y}^T \mathbf{Y})^{-1}$ . Similar to RRR, CCA coefficients  $\mathbf{u}_j, \mathbf{v}_j$  are derived from singular value decomposition of  $\mathbf{R}_{\text{CCA}} = (\mathbf{Y}^T \mathbf{Y})^{\frac{1}{2}} \mathbf{Y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{\frac{1}{2}}$ . The CCA coefficients are

$$\widehat{\mathbf{U}}_{\text{CCA}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}} \mathbf{H}_{\text{CCA}} \mathbf{M}_{\text{CCA}}^{-1}, \quad (6)$$

$$\widehat{\mathbf{V}}_{\text{CCA}} = (\mathbf{Y}^T \mathbf{Y})^{-\frac{1}{2}} \mathbf{H}_{\text{CCA}}, \quad (7)$$

where  $\mathbf{M}_{\text{CCA}}$  is a diagonal matrix of the square roots of the corresponding  $k$  eigenvalues,  $\mathbf{H}_{\text{CCA}}$  consists of the first  $k$  normalized eigenvectors of  $\mathbf{R}_{\text{CCA}}$ . We notice that  $\widehat{\mathbf{U}}_{\text{CCA}}$  in (6) is proportional to  $\widehat{\mathbf{B}}_{\text{RRR}}$  in (4), and  $\widehat{\mathbf{V}}_{\text{CCA}}$  in (7) is proportional to  $\widehat{\mathbf{A}}_{\text{RRR}}$  in (3).

CCA is not a complete predictive model since it only estimates coefficients that maximize the correlations between  $\mathbf{X}$  and  $\mathbf{Y}$ , and discards the explained variance of input and output variables by whitening them. In order to perform out-of-sample prediction, latent variables on test folds are rescaled w.r.t the train set on which the CCA is computed.

### D. Partial least squares

PLS [12] is another low-rank regression model that maximizes the covariance between the latent variables  $\mathbf{X}\mathbf{u}_j$  and  $\mathbf{Y}\mathbf{v}_j$  ( $j = 1..k$ ):

$$\max_{\|\mathbf{u}_j\|_2=\|\mathbf{v}_j\|_2=1} \mathbf{u}_j^T \mathbf{X}^T \mathbf{Y} \mathbf{v}_j. \quad (8)$$

PLS considers  $\mathbf{\Gamma} = \text{Id}$ . Thus,  $\mathbf{R}_{\text{PLS}} = \mathbf{Y}^T \mathbf{X} \mathbf{X}^T \mathbf{Y}$ . The PLS coefficients are :

$$\widehat{\mathbf{U}}_{\text{PLS}} = \mathbf{X}^T \mathbf{Y} \mathbf{H}_{\text{PLS}} \mathbf{M}_{\text{PLS}}^{-1}, \quad (9)$$

$$\widehat{\mathbf{V}}_{\text{PLS}} = \mathbf{H}_{\text{PLS}}, \quad (10)$$

where  $\mathbf{M}_{\text{PLS}}$  is a diagonal matrix of the square roots of the corresponding  $k$  eigenvalues,  $\mathbf{H}_{\text{PLS}}$  consists of the first  $k$  normalized eigenvectors of  $\mathbf{R}_{\text{PLS}}$ . We note the relationship between CCA and PLS coefficients in equations (6, 9) and equations (7, 10), respectively. PLS does not take into account the covariations between the responses  $\mathbf{Y}$  explicitly. This property may be desirable when dealing with related scores, such as different measures of IQ.

TABLE I  
SELECTED SCORES FROM COBRE. OUR EXPERIMENTS USE 14 SCORES; THE TABLE DESCRIBES THE MOST SIGNIFICANT SCORES.

Score	Description
BACS	Brief Assess of Cognition in Schizophrenia
MD ProcSpeed	MatricsDomain Processing Speed Index
WAIS Coding	Wechsler Intelligence on Symbol Coding
WAIS PSI	Wechsler Intelligence Processing Speed
WAIS SymSearch	Wechsler Intelligence on Symbol Search

### E. Regularized reduced-rank models

Many variants of RRR, CCA, and PLS have been proposed, such as using an  $\ell_2$  regularization on  $\mathbf{X}$  and  $\mathbf{Y}$  [13], or using an  $\ell_1$  regularization on the reduced-rank coefficients [6], [7]. Here, we propose to experiment low-rank regularizations of the whitening matrix  $\mathbf{\Gamma} = (\mathbf{Y}^T \mathbf{Y})^{-1}$ . We consider the rank- $l$  whitening matrix  $\mathbf{\Gamma}_l = \mathbf{\Theta}_l \mathbf{\Sigma}_l \mathbf{\Psi}_l^T$ , where  $\mathbf{\Theta}_l \in \mathbb{R}^{q \times l}$ ,  $\mathbf{\Sigma}_l \in \mathbb{R}^{l \times l}$ ,  $\mathbf{\Psi}_l \in \mathbb{R}^{q \times l}$  are the truncated singular value decomposition of  $\mathbf{\Gamma} \in \mathbb{R}^{q \times q}$ . Such approximation avoids numerical and statistical instabilities due to ill-conditioning of  $\mathbf{Y}$ .

## III. PREDICTING SCHIZOPHRENIA WITH RRR

We compare reduced-rank linear models presented above to predict clinical scores from connectivity profiles from rfMRI data that study schizophrenia.

### A. Connectivity and clinical scores from COBRE

The COBRE study aims at characterizing the brain in schizophrenia ([cobre.mrn.org](http://cobre.mrn.org)). It comprises anatomical and functional MRI data from 160 subjects. Diagnostic information and clinical scores are collected using the Structured Clinical Interview used for DSM Disorders (SCID). Table I lists some clinical scores used in our study. rfMRI scans are preprocessed and normalized following the recommendations in [14]. Spatial preprocessing with SPM12 includes motion correction, fMRI coregistration to T1-MRI, normalization to MNI template followed by spatial smoothing (6mm FWHM). Temporal preprocessing with Nilearn v0.2.6 [15] includes linear detrending and band filtering (0.01–0.1Hz). Confounds are then removed from timeseries, they are composed of age, white matter, 12 motion components, and 6 noise components from CompCor [16]. Functional connectivity for each subject is computed through pairwise correlations between timeseries of 64 regions of interest extracted from BASC atlas [17]. Feature matrix  $\mathbf{X}$  is then built where each row consists of vectorized subject connectivity measures.  $\mathbf{Y}$  consists of 14 continuous clinical scores.

### B. Experiment settings

We present two experiments studying the capacity of the reduced-rank models to predict multiple clinical scores from functional connectivity. The experiments evaluate: *i) Prediction performance of RRR, PLS, and CCA*: The three models are built on COBRE dataset; we compare their accuracy for out-of-sample prediction of continuous clinical scores related

## Out-of-sample regression performances on COBRE

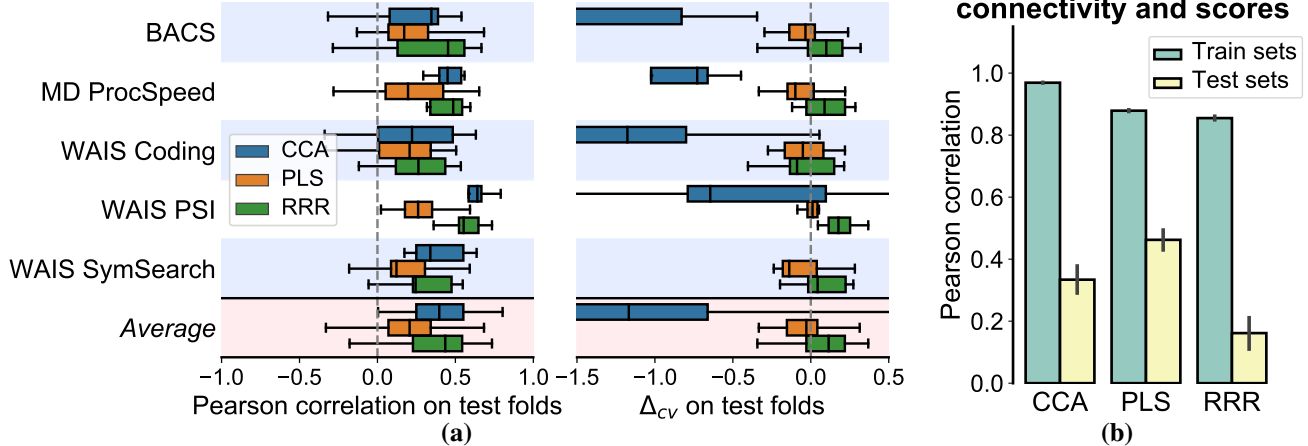


Fig. 1. **Prediction performances on COBRE clinical scores.** (a) CCA, PLS and RRR rank is set to 10 ( $k = 10$ ). While Pearson correlations on test folds are between 0.3 and 0.6 for all models,  $\Delta_{cv}$  on test folds show that RRR is outperforming PLS and CCA that cannot predict accurately the clinical scores. (b) Despite higher CCA correlations between first modes compared to RRR and PLS, CCA accuracies for clinical score regression on test fold are low.

to schizophrenia. *ii) Sensitivity of RRR to its parameters:* We study the impact of the rank  $k$  of the RRR, and the impact of the whitening matrix  $\Gamma$  on the prediction accuracy of RRR model, either by taking the identity matrix, or by varying the low-rank approximation  $l$  of  $\Gamma$ .

All models are assessed with 100 randomized train and test splits. Test sets represent 25% of the whole dataset. Regression accuracies of clinical scores on each test fold is measured with the cross-validated determination coefficient  $\Delta_{cv}$ , and Pearson's correlation between true and predicted values.  $\Delta_{cv}$  –or r-squared– is a relative distance between the prediction values and the true values:

$$\Delta_{cv} = 1 - \frac{\sum_{k=1}^{|\text{test}|} (y_k - \hat{y}_k)^2}{\sum_{k=1}^{|\text{test}|} (y_k - \bar{y}_{\text{test}})^2}, \quad (11)$$

where  $y_k, \hat{y}_k, \bar{y}_{\text{test}}$  are respectively the true score, the predicted score, and the mean score on the test fold. Since CCA and PLS are not explicit predictive models, we rescale test set predictions :

$$\hat{\mathbf{Y}}_{\text{test}} = \left( \frac{\mathbf{X}_{\text{test}} - \bar{\mathbf{X}}_{\text{train}}}{\Sigma_{\text{train}}} \right) \hat{\mathbf{U}}_{\text{train}} \hat{\mathbf{V}}_{\text{train}}^T - 1 + \bar{\mathbf{Y}}_{\text{train}}, \quad (12)$$

where  $\bar{\mathbf{X}}_{\text{train}}$  and  $\bar{\mathbf{Y}}_{\text{train}}$  are the columnwise mean matrices of  $\mathbf{X}_{\text{train}}$  and  $\mathbf{Y}_{\text{train}}$ , respectively.  $\Sigma_{\text{train}}$  is the columnwise variance matrix of  $\mathbf{X}_{\text{train}}$ . Experiments are implemented in Python using Scikit-Learn v0.17 [18].

## IV. RESULTS AND DISCUSSION

**4.1. RRR is more suitable for prediction.** Figure 1 shows the distribution of the accuracy when predicting clinical scores by RRR with a low-rank approximated whitening matrix, CCA, and PLS. Rank of each method is set to 10 ( $k = 10$ ). We observe that :

– *RRR outperforms CCA and PLS* Figure 1-(a) shows that RRR's  $\Delta_{cv}$  are higher and more stable than CCA and PLS  $\Delta_{cv}$ , for all selected clinical scores, average values are between 0.05 for WAIS Coding and 0.20 for BACS. This demonstrates that RRR is more adequate for score prediction compared to CCA and PLS.

–  *$\Delta_{cv}$  is more reliable than correlation* Figure 1-(a) also shows discrepancies between  $\Delta_{cv}$  and Pearson's correlation as a regression assessment metric. Correlation values of CCA are higher than RRR and PLS, but  $\Delta_{cv}$  are the lowest. This is explained by the fact that correlations measure the linear accordance, but discard scaling and offsets. As for clinical scores, predicting well scaling and offsets is important, using  $\Delta_{cv}$  brings a better assessment of regression quality.

– *Modes correlation may not imply prediction* Figure 1-(b) shows the average correlations between first modes of CCA, PLS, and RRR ( $\mathbf{X}\mathbf{u}_1$  and  $\mathbf{Y}\mathbf{v}_1$ ,  $\mathbf{X}\mathbf{b}_1$  and  $\mathbf{Y}(1/\mathbf{a}_1)^T$ ) for the train and the test sets. Results show that high correlations of the first modes do not always imply high prediction accuracy. In particular, CCA first modes are highly correlated even for on test sets, but the prediction of each individual score is poor because CCA discards the explained variance of  $\mathbf{X}$  and  $\mathbf{Y}$ .

**4.2. Which RRR whitening matrix to use ?** Figure 2 summarizes averaged  $\Delta_{cv}$  on all test folds for all clinical scores, by varying RRR rank  $k$  and the whitening matrix. Figure 2-(a) shows that the RRR prediction accuracy is improved as the rank increases. The optimal accuracy is achieved when considering a full rank ( $k = 14$ ). In figure 2-(b), we compare different whitening matrices using full rank RRR ( $k = 14$ ). Using a whitening matrix regularized to a rank of  $l=2$  or  $l=5$  gives the best prediction accuracy. Interestingly, it is suboptimal to use full-rank unregularized matrix  $\Gamma = (\mathbf{Y}^T \mathbf{Y})^{-1}$ , in which case RRR boils down to CCA. The other end of the spectrum, equivalent to a PLS, with an identity whitening matrix is also suboptimal.

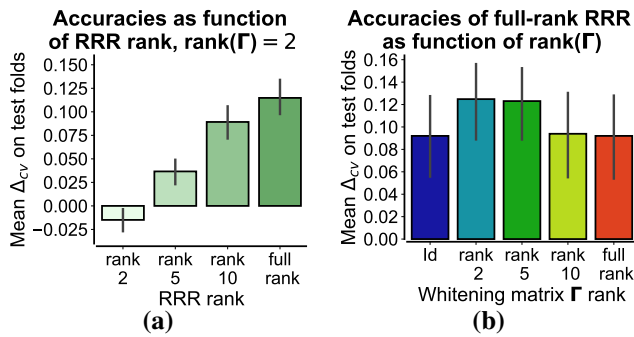


Fig. 2. **RRR sensitivity to the rank and the whitening matrix.** (a) Average  $\Delta_{CV}$  on test folds show that full rank RRR gives the most accurate predictions. (b) For full rank RRR ( $k = 14$ ), using a low-rank approximation of the whitening matrix is better than full-rank or identity whitening matrices.

## V. CONCLUSION

We assessed different reduced-rank models to predict multiple schizophrenia-related scores from functional connectivity features. Results of RRR on COBRE dataset suggest that a low-rank regularization of co-variations between clinical scores in the whitening matrix  $\Gamma$  helps to improve overall prediction accuracy. Also, obtaining a high correlation of the first modes between imaging and outputs in CCA and PLS does not imply that these outputs can be predicted accurately from connectivity with these models. Finally, we advocate to use  $\Delta_{CV}$  instead of using Pearson correlation to assess regression accuracy. Unlike correlations, it is sensitive to absolute differences between the prediction values and the true values, which is essential in the perspective of individualized predictions.

## ACKNOWLEDGMENTS

This project has received funding from the European Union’s Horizon 2020 Framework Programme for Research and Innovation under Grant Agreement No 720270 (Human Brain Project SGA1).

## REFERENCES

- [1] J. Dubois and R. Adolphs, “Building a science of individual differences from fmri,” *Trends in cognitive sciences*, vol. 20, no. 6, pp. 425–443, 2016.
- [2] C.-W. Woo *et al.*, “Building better biomarkers: brain models in translational neuroimaging,” *Nature Neuroscience*, vol. 20, no. 3, pp. 365–377, 2017.
- [3] G. Varoquaux *et al.*, “Assessing and tuning brain decoders: cross-validation, caveats, and guidelines,” *NeuroImage*, vol. 145, 2017.
- [4] S. M. Smith *et al.*, “A positive-negative mode of population covariation links brain connectivity, demographics and behavior,” *Nature Neuroscience*, 2015.
- [5] K. L. Miller, *et al.*, “Multimodal population brain imaging in the UK biobank prospective epidemiological study,” *Nature Neuroscience*, 2016.
- [6] E. L. Floch *et al.*, “Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares,” *NeuroImage*, 2012.
- [7] M. Vounou *et al.*, “Discovering genetic assoc. with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach,” *NeuroImage*, 2010.
- [8] J. Sui *et al.*, “A CCA+ICA based model for multi-task brain imaging data fusion and its application to schizophrenia,” *NeuroImage*, 2010.
- [9] J. M. Monteiro *et al.*, “A multiple hold-out framework for sparse partial least squares,” *J. Neurosci. M.*, 2016.

- [10] A. J. Izenman, “Reduced-rank regression for the multivariate linear model,” *J. of multivariate analysis*, 1975.
- [11] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.
- [12] A. Krishnan *et al.*, “Pls methods for neuroimaging: A tutorial and review,” *NeuroImage*, 2011.
- [13] L. Breiman and J. H. Friedman, “Predicting multivariate responses in multiple linear regression,” *JRSS*, 1997.
- [14] R. A. Poldrack, T. Nichols, and J. Mumford, *Handbook of Functional MRI Data Analysis*, 2009.
- [15] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, “Machine learning for neuroimaging with scikit-learn,” *Frontiers in Neuroinformatics*, vol. 8, p. 14, 2014.
- [16] Y. Behzadi, K. Restom, J. Liao, and T. T. Liu, “A component based noise correction method (CompCor) for BOLD and perfusion based fMRI,” *NeuroImage*, vol. 37, no. 1, pp. 90–101, 2007.
- [17] P. Bellec *et al.*, “Multi-level bootstrap analysis of stable clusters in resting-state fMRI,” *NeuroImage*, 2010.
- [18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.