

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/141190>

Copyright and reuse:

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk



**Latent Variable Modelling of Population
Neuroimaging and Behavioural Data**

by

Zhangdaihong Liu

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Mathematics Institute

April 2020

Contents

List of Tables	v
List of Figures	ix
Acknowledgments	xxiv
Declarations	xxv
Abbreviations	xxvi
Chapter 1 Introduction	1
1.1 Neuroimaging	2
1.2 From MRI to Brain Connectome	4
1.2.1 Generating functional connectomes	5
1.3 Linking Brain and Behaviour	8
1.3.1 Challenges	8
1.4 Outline and Contributions	9
Chapter 2 Background Methods	12
2.1 PCA and SVD	12
2.1.1 SVD	13
2.1.2 Principal loadings	14
2.2 Variations of PCA	15
2.2.1 PCA in high-dimensional situation	15
2.2.2 PCA versus covariance in subject domain	16
2.2.3 PCA with missing data	17
2.2.4 PCA on data with mixed data types	18
2.3 CCA	18
2.3.1 CCA as generalised eigenvalue problem	21
2.3.2 Multi-view CCA	21
2.4 Factor Analysis	23
2.4.1 Interpretation of FA	24

2.5	Group Factor Analysis	24
2.5.1	Model interpretation	26
2.5.2	Comparison with multi-view CCA	27
2.6	Indeterminacy and Non-identifiability Issues	27
2.7	Factor Rotation	28
2.7.1	Factor rotation in PCA/SVD and CCA	30
2.8	General Data Pre-processing Techniques	31
2.8.1	Quality control	31
2.8.2	Normalisation	32
2.8.3	Missing data imputation	33
2.8.4	De-confounding	38
2.9	Model Validation and Assessment	39
2.9.1	Cross-validation	39
2.9.2	Permutation testing	43
2.10	Predictive Modelling	43
2.10.1	Linear regression	44
2.10.2	Support vector machines	46
2.10.3	Making predictions	49
2.11	Conclusion	50
Chapter 3 Supervised Dimension Reduction		51
3.1	Sign Flipping	51
3.1.1	Effects on PCA/SVD	52
3.1.2	Effects on covariance matrix and eigen decomposition	52
3.1.3	Effects on CCA	53
3.2	Supervised Dimension Reduction	53
3.2.1	Evaluating the stability of SDR	57
3.3	Conclusion	58
Chapter 4 Human Connectome Project		59
4.1	Introduction	59
4.2	Data	60
4.2.1	Data pre-processing	61
4.3	Analysis Pipeline	62
4.3.1	Comparison between PCA and SDR	64
4.4	Results	64
4.4.1	Sign alignment	64
4.4.2	PCA CCA results	64

CONTENTS	iii
4.4.3 Sub-domain analysis and SDR results	66
4.4.4 Results for SDR CCA	70
4.4.5 Stability of SDR CCA	74
4.5 Discussion	78
4.5.1 Interpretation of CCA loadings	80
Chapter 5 UK Biobank Project	82
5.1 Introduction	82
5.2 Data	83
5.2.1 Sub-domain grouping	83
5.2.2 Data issues and fixes	84
5.2.3 Data Pre-processing	86
5.2.4 Sign-flipping	86
5.3 Method	87
5.3.1 Canonical correlation analysis pipeline	87
5.3.2 Group factor analysis pipeline	92
5.4 CCA Results	93
5.4.1 SDR results	94
5.4.2 Pairwise CCA on non-reduced data	94
5.4.3 Multi-view CCA on non-reduced data	100
5.4.4 Pairwise CCA on SDR-reduced data	106
5.4.5 Multi-view CCA for SDR reduced data	118
5.4.6 Stability study on SDR CCA	124
5.5 GFA Results	132
5.6 Conclusion	142
Chapter 6 Predicting Personality with Functional Connectivity	145
6.1 Introduction	145
6.2 Method	146
6.2.1 Data	146
6.2.2 fMRI data pre-processing and functional connectivity matrix	147
6.2.3 Personality related functional brain network	147
6.2.4 Prediction using personality related brain network	147
6.2.5 Common network for all subjects	148
6.3 Results	149
6.3.1 Linear regression with Power atlas	149
6.3.2 Linear regression with AAL2 atlas	151
6.3.3 Comparison between Power and AAL2 parcellations	155

6.3.4	Analysis using Support Vector Regression	155
6.4	Personality Prediction on HCP data	157
6.4.1	Prediction using the original connectivity matrix	158
6.4.2	Prediction using the de-confounded connectivity matrix	159
6.4.3	Effects of confounders	160
6.5	Conclusion	161
Chapter 7	Conclusions	163
Bibliography		167
Appendix A Appendix for Human Connectome Project		181
A.1	Confounds for the HCP project	181
A.2	Full list of 234 SMs (HCP 1200 release)	182
A.3	Sub-domain summary reports	184
A.4	Stability of SDR CCA canonical loading	192
Appendix B Appendix for UK Biobank Project		195
B.1	Rotated Loadings in SM Sub-domains	195
B.2	CCA Results for Non-reduced Data	203
B.2.1	Canonical loadings between non-reduced SM and FC	203
B.2.2	Canonical loadings between non-reduced SM and IDP	205
B.2.3	Canonical loadings between non-reduced FC and IDP	207
B.3	CCA Results for SDR-reduced Data	209
B.3.1	Canonical loadings between SDR SM and SDR FC	209
B.3.2	Canonical loadings between SDR SM and SDR IDP	211
B.3.3	Canonical loadings between SDR FC and SDR IDP	215
B.3.4	Canonical loadings for multi-view SDR CCA	217
B.4	Stability of SDR CCA	221
B.4.1	CCA between SDR IDP and SDR SM	221
B.4.2	CCA between SDR FC and SDR IDP	224
B.4.3	Pairwise and multi-view CCA on PCA-reduced data	227
B.4.4	Comparing SDR with PCA	230
Appendix C Appendix for Predicting Personality Project		231
C.1	Significant network links	231

List of Tables

4.1	Summary table for PCA CCA with 5 different input dimensions of CCA (first column). Second and third columns show the variance explained (VE) by the SM and BM canonical variables in the observed SM and BM sets for significant canonical pairs respectively; the fourth column shows the canonical correlation for the canonical pairs; the last column shows the number of significant canonical pairs obtained by permutation testing.	66
4.2	Summary of SM sub-domains. The factors are orthogonally rotated principal components and ordered by R-squared values in the original sub-domain. The second column shows the factor names summarised from panel E in each of the sub-domain report like the top figure in Fig. 4.4. The third column shows the variance explained by the dimension reduced sub-domain in the original sub-domain. The numbers in the brackets are the two-way CV estimated dimension verses the total number of variables in the sub-domain. The Personality is represented by the Big Five personality traits: Neuroticism (N), Agreeableness (A), Extraversion (E), Conscientiousness (C), Openness to experience (O).	69
4.3	Summary table for SDR CCA. The ‘Input Dimensions’ column shows the dimensions of SM and BM as CCA inputs; the ‘VE by SM’ and ‘VE by BM’ parts of the table represent the variance explained (VE) by the SM and BM canonical variables in the observed SM and BM set for significant canonical pairs (pairs of canonical variables) respectively; ‘Canonical Correlation’ part shows the canonical correlation between the canonical pairs; the last part shows the number of significant canonical pairs given by permutation testing.	70
4.4	5-fold CV on 62 dimensional SM and 100-dimensional BM in SDR CCA analysis. Mean variance explained (MVE) and mean canonical correlations (MCC) are shown for the first 3 pairs of canonical variables with standard deviation (std) in brackets.	76

5.1	Functional interpretation for the positive and negative brain maps in the CCA between FC and SM (shown in Fig. B.10). The keywords are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.	98
5.2	Functional interpretation for the positive and negative brain maps in the CCA between FC and IDP (shown in Fig. B.14). The keywords are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.	101
5.3	Functional interpretation for the positive and negative brain maps in the multi-view CCA (shown in Fig. 5.16). The keywords are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.	106
5.4	Summary of SM sub-domains. The factors are orthogonally rotated principal components and ordered by R-squared values in the original sub-domain. Second column shows the factor names summarised from figures like Fig. 5.22.	109
5.5	Functional interpretation for the positive and negative brain maps in the CCA between SDR FC and SDR SM (shown in Fig. B.16). The keywords are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.	116
5.6	Functional interpretation for the positive and negative brain maps in the CCA between SDR FC and SDR IDP (shown in Fig. B.20). The keywords are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.	117
5.7	Functional interpretation for the positive and negative brain maps in the multi-view CCA on SDR SM, SDR FC and SDR IDP (corresponding to the maps shown in Fig. B.23). The key words are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.	123

6.1	Prediction power on Big Five personality factors using Power atlas. Significance of individual links is 0.01. Values with * are significant.	149
6.2	Summary statistics using AAL2 atlas for SVR modelling.	156
6.3	Summary statistics using Power atlas for SVR modelling.	157
6.4	Prediction results on the five personality factors using for the positive network extracted from the non-de-confounded connectivity matrix. The second columns shows the mean correlation between predicted and true personality scores over 10 simulations with standard deviation in brackets; the third column shows the number of significant simulations out of the 10 splits; the last column is the deviation of the predicted score from the true score in percentage.	159
6.5	Prediction results on the five personality factors using for the negative network extracted from the non-de-confounded connectivity matrix. The second columns shows the mean correlation between predicted and true personality scores over 10 simulations with standard deviation in brackets; the third column shows the number of significant simulations out of the 10 splits; the last column is the deviation of the predicted score from the true score in percentage.	160
6.6	Prediction results on the five personality factors using for the positive network extracted from the de-confounded connectivity matrix. The second columns shows the mean correlation between predicted and true personality scores over 10 simulations with standard deviation in brackets; the third column shows the number of significant simulations out of the 10 splits; the last column is the deviation of the predicted score from the true score in percentage.	160
6.7	Prediction results on the five personality factors using for the negative network extracted from the de-confounded connectivity matrix. The second columns shows the mean correlation between predicted and true personality scores over 10 simulations with standard deviation in brackets; the third column shows the number of significant simulations out of the 10 splits; the last column is the deviation of the predicted score from the true score in percentage.	161

C.1	All significant links in the common network that is positively related to Openness to Experience using Power atlas. A link is the connection between the strength of the link and the Openness to Experience score. The links are ordered by R-value. * indicates no matching Brodmann area found by the MNI coordinates.	231
C.2	All significant links in the common negative network of Extraversion using Power atlas. A link is the connection between Brodmann Area 1 to Brodmann Area 2. R-value is the Pearson's correlation between the strength of the link and the Extraversion score. The links are ordered by R-value. * indicates no matching Brodmann area found by the MNI coordinates.	233

List of Figures

1.1 Examples of MRI sequences. (a) is T1-weighted structural MRI sequence, and (b) is T2-weighted MRI sequence (both figures are imported from [15]). (c) is another anatomical sequence, diffusion tensor imaging, and the figure imported from [1]. (d) is Blood-Oxygen-Level-Dependent fMRI, one of the functional sequences (figure is imported from [2]).	3
1.2 Illustration of the slice timing problem. Suppose that an identical signal was present at all slices (top), all slices were acquired at the same time would lead to an incorrect shift in the timing of the signal at other slices. This figure is imported from Fig. 1 in [113], and TRs stands for repetition time which is a sequence parameter.	6
1.3 Examples of MRI sequences. (a) is the Brodmann atlas which parcelates the brain into 52 regions. This figure is imported from [100] (b) is the AAL atlas, parcellating the brain into 90 regions. This figure is imported from [16].	7
2.1 GFA illustration with three views/groups of data (imported from [71]). GFA finds latent factors accounted by all three views as well as by subsets of the three views.	25
2.2 An example of Simpson's paradox. This figure is imported from [77]. Salary and Neuroticism appear to show positive correlation in the data. However, after regressing out Education, they show negative correlations.	38
2.3 Illustration of 5-fold cross-validation (CV) in the analysis of PCA followed by CCA.	41

2.4 Illustration of SVM (Fig. 12.1 in [60]). The left panel shows the linearly separable case; the right panel is the more generalised case (non-separable). The solid line in the middle is the hyperplane separating two classes. The dotted lines form the margin of the SVM and only depend on the nearest data points to the hyperplane. In the right panel, ξ_i represents the distance of the points fall on the wrong side of their margin.	46
3.1 SDR overview	54
3.2 Illustration of the 5-fold two-way cross-validation. It minimises PRESS and estimates the dimensionality in an automated fashion. Yellow blocks represent the training data and light blue blocks represent the test data. Two-way CV includes a subject-way (CV over subject direction) and a variable-way (CV over variable direction). Prediction error is calculated by the reconstruction error using different numbers of principal components.	55
4.1 Method overview of SDR CCA. SM and BM are first grouped into sub-domains. PCA is applied to each sub-domain while a two-way CV method (* see Fig. 3.2) is used to estimate the dimension. Then the rotated principal components from all sub-domains are concatenated to form the reduced SM and BM. Finally the reduced SM and BM are fed into CCA and for further CV (** see Fig. 2.3) and permutation testing.	63
4.2 Pairwise correlation among 234 subject measures grouped by 14 functional domains. On the top is the correlation matrix among original variables; on the bottom is the correlation matrix after sign-flipping which aims to align pairwise correlations between and within domains.	65
4.3 Top 20 canonical loadings for 4 significant SM canonical variables in PCA CCA method using 62 dimensional SM and 100 dimensional BM. Variable names with a '-' sign means the values have been flipped.	67
4.4 Top figure shows the rotated principal loadings; figure at the bottom shows the error curves calculated by Eqn. (3.11) (dotted line) and Eqn. (3.13) (red line), with the minimal error circled at the second component. The naive way of calculating PRESS (dotted line) is monotonically decreasing, while the two-way CV method (red line) offers a minimum point.	68

4.5	Top 20 SM canonical loadings for 3 significant canonical variables. Variable name with '-' sign shows that it was flipped in the original dataset. Canonical loadings of CCA 1 are very similar to the first set of PCA CCA, heavily cognition dominated; the second set is mixed with cognition, drug use etc; The third set is combination of tobacco use and cognition variables. The labels are the exact variables name given on the HCP official websites.	72
4.6	SM canonical loadings on the CCA input for the 3 significant canonical variables. The left set of figures shows the top 20 loadings for the 3 significant canonical variables respectively; the right set of figures show the mean of all positive loadings (red bars) and the mean of all negative loadings (blue bars) within each sub-domain for the 3 significant canonical variables.	73
4.7	Positive and negative CCA strengths on brain surface (left) and volume (right) for the 3 significant canonical variables. The visualisation is cut by 80 percentile. Positive (red maps) and negative (blue maps) CCA strengths are generated by mapping the canonical loading with the sign of population mean correlation between each pair of ICA regions, then average the top 20 positive and negative modulated loadings respectively.	75
4.8	Stability of SM canonical loadings on CCA input. Bar plot shows the occurrence frequency in CV out of the 5 folds. Variables are chosen by selecting the top 20 mostly weighted ones in each fold. The ones appeared at least twice are shown above. The right axis shows the mean and the standard deviation over all occurred loadings. Top and bottom plots are the canonical loadings for the first and second canonical variables respectively.	77
5.1	UK Biobank CCA analysis pipeline. Due to the higher complexity of SMs, they go through more data cleaning steps. Then all FC, SM and IDP undergo the same QC and pre-processing procedures. Two parallel CCA pipelines follow after, with and without dimension reduction.	88

5.2	Cross-validation method overview. We apply SDR (dotted box on the top) to the held-in (training) set after the split of the data. Within SDR the principal loadings are rotated to construct rotated components (RCs), and then the RCs of the held-in set are fed into CCA. Cross-validated canonical variables and correlations are obtained by multiplying the held-in canonical weights with the RCs from the held-out (test) set.	90
5.3	The left set of figures are some of the rotated principal loadings in a SM sub-domain in 10-fold CV; The right set of figures are the same group of rotated loadings but after applying the matching algorithm and ordered by R-squared values.	91
5.4	Correlation between every pair of SM variables before (left) and after (right) sign-flipping. The most noticeable change happen in the blocks of ‘Alcohol Use’, ‘Tobacco Use’ and ‘Cognition’ where many of the variables are sign-flipped. For sub-domains ‘Mental Health’ and ‘Health & Medical History’, almost all variables are flipped. This is reflected by the reversed colour pattern of the correlation between these two sub-domains and the rest of the sub-domains. ‘Food & Drink’ and ‘Physical Measures’ are left un-flipped.	93
5.5	The original number of variables (blue bar) in each SM (left) and IDP (right) sub-domain and the number of latent components (orange bar) SDR reduces to in each sub-domain.	94
5.6	Pairwise canonical correlations for the CCA of FC and SM (orange), IDP and SM (green), FC and IDP (blue). The canonical correlations between FC and IDP are significantly stronger than the other two pairs, with the correlations between IDP and SM being the weakest.	95
5.7	Observed canonical correlations (blue line) versus the distribution of canonical correlation between the permuted canonical pairs (box plot). The grey shaded area is the 5 to 95 percentile from the distribution of first permuted canonical pair (first box plot) which is used to define the significance of the canonical pairs (canonical correlation falls under the upper bound of this band is defined as insignificant). From left to right are the CCA between FC and SM, IDP and SM, and FC and IDP respectively.	96
5.8	Variance explained by the significant canonical variables in their original dataset. From left to right are the CCA between FC and SM (7 pairs), IDP and SM (6 pairs), IDP and FC (31 pairs).	96

5.9	Top 30 canonical loadings for the first three sets of significant SM canonical variables in the CCA of FC and SM. From top left to bottom right are the first to the seventh canonical loadings respectively. Variables that are sign-flipped have a ‘-’ sign in front of their names.	97
5.10	Canonical loaded maps for the first set significant FC canonical variables in the CCA with IDP. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).	100
5.11	The left subplot shows the first 20 multi-view canonical correlations. The right subplot shows the corresponding variance explained in the original datasets.	102
5.12	The first set of permutation testing in multi-view setting. Plots on the first row are obtained by permuting FC only. Bottom plot is obtained by permuting SM only. Blue lines are the true canonical correlations between FC and SM (top left), IDP and SM (top right) and FC and IDP (bottom). Box plots are the distributions of the canonical correlation between permuted data. Grey band is plotted from the 5th to the 95th percentile of the permuted distribution with the largest mean.	103
5.13	The second set of permutation testing, testing the significance of the sum of the correlations with FC and SM permuted at the same time. Blue line is the sum of the true canonical correlations between all three pairs of modalities. Box plots are the distributions of the sum of the canonical correlations between permuted data. Grey band is plotted from the 5th to the 95th percentile of the permuted distribution with the largest mean.	104
5.14	Top 30 SM canonical loadings for the first four sets in multi-view CCA. From top left to bottom right are the first to the fourth set respectively.	104
5.15	Top 30 IDP canonical loadings for the first four sets in multi-view CCA. From top left to bottom right are the first to the fourth set respectively.	105
5.16	Canonical loaded maps for the first four significant FC canonical variables in the multi-view CCA. From top to bottom are the first to the fourth canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).	105

5.17 Canonical correlation for the first ten canonical pairs in pairwise CCA for SDR-reduced datasets.	107
5.18 Variance explained by the significant SDR canonical variables in their original dataset. From left to right are the CCA between FC and SM (7 pairs), IDP and SM (8 pairs), IDP and FC (19 pairs).	108
5.19 Permutation testing on the SDR reduced data for 1000 permutes. Observed canonical correlations (blue line) versus the distribution of canonical correlation between the permuted canonical pairs (box plot). The grey shaded area is the 5 to 95 percentile from the distribution of first permuted canonical pair (first box plot) which is used to define the significance of the canonical pairs (canonical correlation falls under the upper bound of this band is defined as insignificant). From left to right are the CCA between FC and SM, IDP and SM, and FC and IDP respectively.	108
5.20 Top 20 canonical loadings for the first three significant SM canonical variables in the CCA of SDR FC and SDR SM. The number after the domain name indicates the n th latent component in the sub-domain. Canonical loadings for all seven significant canonical variables can be found in Fig. B.9.	112
5.21 Mean squared SM loadings summarised from each sub-domain in the CCA of SDR FC and SDR SM. Blue bars are the mean squared positive loadings and orange bar are the mean squared negative loadings. From top left to bottom right are the first to the seventh set respectively.	113
5.22 First seven latent factor (rotated) loadings from sub-domain Physical Measures. There are in total 20 such rotated loadings in the ‘Physical Measures’ sub-domain.	114
5.23 Brain volume maps for the first three significant SDR FC canonical loadings in the CCA with SDR SM. From top to bottom are the first to the third canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1). Maps for all seven sets significant canonical loadings can be found in Fig. B.15.	115

5.24 First set of Mean squared SDR SM loadings (left) and SDR IDP loadings (right) summarised from each sub-domain. Orange bars are the mean squared positive loadings and blue bars are the mean squared negative loadings. Loading for all significant sets for SDR SM and SDR IDP canonical variables can be found in Fig. B.17 and B.18 respectively.	116
5.25 Mean squared IDP loadings for the first (left) and second (right) canonical variables summarised from each sub-domain in the CCA of SDR IDP and SDR FC. Orange bars are the positive loadings and blue bars are the negative loadings. The first eight sets of summarised loadings can be found in Fig. B.19.	117
5.26 Left: top 20 canonical correlation between every pair of the modalities plotted together with the sum of the correlation (red line). Noticing here the sum of the correlation is not monotonic anymore. Right: corresponding variance explained for each modality.	118
5.27 The first set of permutation testing: testing the significance of the canonical correlation between individual pairs. Plots on the first two plots are obtained by permuting FC only. Bottom plot is obtained by permuting SM only. Blue lines are the true canonical correlations between FC and SM (top), FC and IDP (middle), and IDP and SM (bottom) and. Box plots are the distributions of the canonical correlation between permuted data. Grey band is plotted from the 5th to the 95th percentile of the permuted distribution with the largest mean.	120
5.28 The second set of permutation testing on SDR reduced data, testing the significance of the sum of the correlations with FC and SM permuted at the same time. Blue line is the sum of the true canonical correlations between all three pairs of modalities. Box plots are the distributions of the sum of the canonical correlations between permuted data. Grey band is plotted from the 5th to the 95th percentile of the permuted distribution with the largest mean.	121
5.29 The second mode of multi-view SDR CCA. The top two plots are mean squared SM loadings (left) and IDP loadings (right) summarised from each sub-domain. Orange bars are the positive loadings and blue bars are the negative loadings. Bottom plots are positive brain map (left) and negative brain map (right) for the SDR FC loadings.	122

- 5.30 The eighth mode of multi-view SDR CCA. The top two plots are mean squared SM loadings (left) and IDP loadings (right) summarised from each sub-domain. Orange bars are the positive loadings and blue bars are the negative loadings. Bottom plots are positive brain map (left) and negative brain map (right) for the SDR FC loadings. 122
- 5.31 Stability of SDR SM canonical loadings for the seven significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than three times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than seven out of ten times are shadowed with dark grey, the rest is shadowed by light grey. 125
- 5.32 Stability of SDR FC canonical loadings for the seven significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than three times out of the 10 folds). The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than seven out of ten times are shadowed with dark grey, the rest is shadowed by light grey. 126
- 5.33 Stability of SDR SM canonical loadings in the multi-view CCA of SDR FC, SDR IDP and SDR SM for the first eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than three times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than seven out of ten times are shadowed with dark grey, the rest is shadowed by light grey. 129

5.34 Stability of SDR IDP canonical loadings in the multi-view CCA of SDR FC, SDR IDP and SDR SM for the first eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than three times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than seven out of ten times are shadowed with dark grey, the rest is shadowed by light grey.	130
5.35 Stability of SDR FC canonical loadings in the multi-view CCA of SDR FC, SDR IDP and SDR SM for the first eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than three times out of the 10 folds). The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than seven out of ten times are shadowed with dark grey, the rest is shadowed by light grey.	131
5.36 Distributions of the training set canonical correlations (blue box plot on the top) compared with the cross-validated canonical correlations (bottom box plot) in the CV study of CCA between SDR FC and SDR SM.	132
5.37 Distributions of the training set canonical correlations (blue box plot on the top) compared with the cross-validated canonical correlations (bottom box plot) in the CV study of CCA between SDR IDP and SDR SM.	133
5.38 Distributions of the training set canonical correlations (blue box plot on the top) compared with the cross-validated canonical correlations (bottom box plot) in the CV study of CCA between SDR FC and SDR IDP.	133
5.39 Distributions of the training set canonical correlations (blue box plot on the top) compared with the cross-validated canonical correlations (bottom box plot) in the CV study of multi-view CCA between SDR FC, SDR SM and SDR IDP.	134
5.40 Loading matrices for all 78 GFA components.	136

5.41	Significant GFA loadings (loadings with mean absolute values larger than 0.05) for the three modalities. Subplot (a), (b) and (c) are filtered from (a), (b) and (c) in Fig. 5.40 respectively.	137
5.42	GFA loading matrix summarised by modality sub-domains for all GFA components. For a component in a sub-domain, the summarised loading is calculated by taking the mean of absolute loadings that are larger than 0.01.	138
5.43	GFA loading matrix for components shared by at least two modalities. Loadings are summarised by sub-domains (same as shown in Fig. 5.42).	140
5.44	Thumbnail for ICA regions 9, 10, 11 and 3.	140
5.45	GFA loading matrix for components shared by all three modalities. Loadings are summarised by sub-domains (same as shown in Fig. 5.42).	141
6.1	Predicted values versus true values for the positive network of Openness to Experience, negative network of Extraversion and negative network of Conscientiousness (from left to right, top to bottom).	150
6.2	On the left is the circular graph of links in the common negative network of Openness to Experience; on the right is the circular graph of links in the common negative network of Extraversion. Links in both graphs were generated using threshold p-value smaller than 0.01.	152
6.3	Matrix graph of links in the common negative network of Openness to Experience (top) and Extraversion (bottom). The colour stands for the significance level of the link in predicting respective personality factor, the logarithm of p-values of the correlation between each of the edges and personality score.	153
6.4	Circular graph of the negative network of Extraversion using threshold 0.05.	154
6.5	On the left is the predicted values versus true values for the negative network of Extraversion using threshold 0.02 with AAL2 atlas; On the right is the matrix graph showing the common links across all subjects in this network. The colour stands for the logarithm of p-values of the correlation between each of the edges and Extraversion score.	155
6.6	Predicted Extraversion score against true Extraversion score using Power atlas and Support Vector Regression (SVR) with RBF kernel (left) and linear kernel (right).	156
A.1	Summary report of Family History.	184

A.2	Tobacco Use sub-domain summary report.	185
A.3	Sensory sub-domain summary report.	185
A.4	Psychiatry sub-domain summary report.	186
A.5	Physical Health sub-domain summary report.	186
A.6	Personality sub-domain summary report.	187
A.7	Motor sub-domain summary report.	187
A.8	Feminine Health sub-domain summary report.	188
A.9	Emotion sub-domain summary report.	188
A.10	Drug Use sub-domain summary report.	189
A.11	Demographics and SES sub-domain summary report.	189
A.12	Cognition sub-domain summary report.	190
A.13	Alcohol Use sub-domain summary report.	190
A.14	Alertness sub-domain summary report.	191
A.15	Stability of SM canonical loadings on observed variables in SDR CCA. Bar plot shows the occurrence frequency in CV out of the 5 folds. Variables are chosen by selecting the top 20 mostly weighted ones in each fold. The ones appeared at least twice are shown above. Right axis shows the mean and the standard deviation over all occurred loadings. Top and bottom plots are the canonical loadings for the first and second canonical variables respectively. It is obvious that the second canonical loadings are less stable than the first set.	193
A.16	Stability of BM canonical loadings on observed data in SDR CCA. Bar plot shows the occurrence frequency in CV out of the 5 folds. The positive (top plots) and negative (bottom plots) maps are chosen by first averaging the top 20 positive and negative canonical loadings within each region respectively; then select the top 20 nodes with the highest positive and negative mean loadings in each fold. The ones occurred at least three times are shown above. Right axis shows the mean and the standard deviation over all occurred loadings. Similar to SM canonical loadings, the first set shows better stability than the second set.	194
B.1	Mental Health sub-domain rotated loadings.	196
B.2	Health & Medical History sub-domain rotated loadings.	197
B.3	Alcohol Use sub-domain rotated loadings.	198
B.4	Tobacco Use sub-domain rotated loadings.	199
B.5	Cognition sub-domain rotated loadings.	199

B.6 Lifestyle & Environment sub-domain rotated loadings.	200
B.7 Food & Drink sub-domain rotated loadings.	201
B.8 Exercise & Work sub-domain rotated loadings.	202
B.9 Top 30 canonical loadings for the 7 significant SM canonical variables in the CCA of FC and SM. From top left to bottom right are the first to the seventh canonical loadings respectively. Variables that are sign-flipped have a ‘-’ sign in front of their names.	203
B.10 Canonical loaded maps for the 7 significant FC canonical variables in the CCA with SM. From top to bottom are the first to the seventh canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).	204
B.11 Top 30 canonical loadings for the 6 significant SM canonical variables in the CCA of IDP and SM. From top left to bottom right are the first to the sixth canonical loadings respectively. Variables that are sign-flipped have a ‘-’ sign in front of their names.	205
B.12 Top 30 canonical loadings for the 6 significant IDP canonical variables in the CCA of IDP and SM. From top left to bottom right are the first to the sixth canonical loadings respectively.	206
B.13 Top 30 canonical loadings for the first 6 significant IDP canonical variables in the CCA of IDP and FC. From top left to bottom right are the first to the sixth canonical loadings respectively.	207
B.14 Canonical loaded maps for the first 6 significant FC canonical variables in the CCA with IDP. From top to bottom are the first to the sixth canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).	208
B.15 Top 20 canonical loadings for the 7 significant SM canonical variables in the CCA of SDR FC and SDR SM. From top left to bottom right are the first to the seventh canonical loadings respectively. The number after the domain name means the nth latent component in the sub-domain.	209
B.16 Canonical loaded maps for the 7 significant SDR FC canonical variables in the CCA with SDR SM. From top to bottom are the first to the seventh canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).	210

B.17 Mean sum of squared SM loadings summarised from each sub-domain in the CCA of SDR IDP and SDR SM. Blue bars are the mean sum of squared positive loadings and orange bar are the mean sum of squared negative loadings. From top left to bottom right are the first to the eighth set respectively.	212
B.18 Mean sum of squared IDP loadings summarised from each sub-domain in the CCA of SDR IDP and SDR SM. Blue bars are the mean sum of squared positive loadings and orange bar are the mean sum of squared negative loadings. From top left to bottom right are the first to the eighth set respectively.	213
B.19 Mean squared value of IDP loadings summarised from each sub-domain in the CCA of SDR IDP and SDR FC. Blue bars are the mean sum of squared positive loadings and orange bar are the mean sum of squared negative loadings. From top left to bottom right are the first to the eighth set respectively.	215
B.20 Canonical loaded maps for the first 8 significant SDR FC canonical variables in the CCA with SDR IDP. From top to bottom are the first to the eighth canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).	216
B.21 Mean sum of squared SM loadings summarised from each sub-domain in the SDR multi-view CCA. Blue bars are the mean sum of squared positive loadings and orange bar are the mean sum of squared negative loadings. From top left to bottom right are the first to the eighth set respectively.	218
B.22 Mean sum of squared IDP loadings summarised from each sub-domain in the SDR multi-view CCA. Blue bars are the mean sum of squared positive loadings and orange bar are the mean sum of squared negative loadings. From top left to bottom right are the first to the eighth set respectively.	219
B.23 Canonical loaded maps for the first 8 significant SDR FC canonical variables in the multi-view CCA. From top to bottom are the first to the eighth canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).	220

B.24 Stability of SDR SM canonical loadings in the CCA of SDR SM and SDR IDP for the eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than 3 times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than 7 out of 10 times are shadowed with dark grey, the rest is shadowed by light grey.	222
B.25 Stability of SDR IDP canonical loadings in the CCA of SDR SM and SDR IDP for the eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than 3 times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than 7 out of 10 times are shadowed with dark grey, the rest is shadowed by light grey.	223
B.26 Stability of SDR IDP canonical loadings in the CCA of SDR FC and SDR IDP for the eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than 3 times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than 7 out of 10 times are shadowed with dark grey, the rest is shadowed by light grey.	225
B.27 Stability of SDR FC canonical loadings in the CCA of SDR FC and SDR IDP for the eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than 3 times out of the 10 folds). The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than 7 out of 10 times are shadowed with dark grey, the rest is shadowed by light grey.	226
B.28 Comparison between SDR and PCA on canonical correlations in pairwise CCA analysis.	227

B.29 Permutation testing on the PCA reduced data for 1000 permutes. True canonical correlations (blue line) versus the distribution of canonical correlation between the permuted canonical pairs (box plot). The dotted line is the 95 percentile of the distribution of the first permuted canonical pair (first box plot), which is used to define the significance of the canonical pairs (canonical correlation falls under this line is defined as insignificant). From left to right are the CCA between FC and SM, IDP and SM, and FC and IDP respectively.	228
B.30 Comparison between SDR and PCA on variance explained by the first 10 canonical variables in pairwise CCA analysis.	228
B.31 Comparison of multi-view CCA between SDR and PCA on canonical correlation and variance explained.	229
B.32 Permutation testing on the sum of the canonical correlations for the multi-view PCA CCA.	230

Acknowledgments

Undertaking this PhD has been an extraordinary experience for me. It would not have been possible to do without the support and guidance that I received from many people.

Firstly, I would like to express my sincere gratitude to my supervisors Prof. Thomas E. Nichols and Prof. Jianfeng Feng for their support of my PhD research. I would particularly like to thank Prof. Thomas E. Nichols, for his patience, motivation, immense knowledge and continuous guidance throughout my PhD. I would also like to thank Prof. Edmund Rolls for his kind and insightful help for one of the projects, and Dr. Kirstie Whitaker for her enthusiastic mentorship and support when I was an enrichment student at the Alan Turing Institute. I am deeply touched by all my supervisors and mentors passion and devotion towards scientific research.

I greatly appreciate the support received from my coursemates, colleagues and friends, especially Iliana Peneva and Dragana Pavlovic for their valuable feedback on this thesis, Ayman Boustati, Jim Skinner, Tim Pearce and everyone in the NISOx group, Soroosh Afyouni, Alex Bowring and Tom Maullin for their academic help and amazing friendships. My special thanks go to Matthew Groves, my coursemate, colleague, as well as best friend, who has accompanied my low times, shared the happy moments with me and offered invaluable encouragement and advice when I felt lost, and made my PhD journey particularly enjoyable.

I am also extremely grateful to my parents, my grandparents and the whole family for their unconditional love throughout my life. Only with their enormous support, I am able to follow my heart and accomplish my study in the UK.

I gratefully acknowledge the funding received towards my PhD from the China Scholarships Council (CSC)-University of Warwick Scholarships, Mathematics for Real-world Systems CDT, the Alan Turing Institute, and the Guarantor of the Brain to undertake my PhD and other academic activities.

Declarations

I hereby declare that except where specific reference is made to the work of others, this thesis has been composed by myself and has not been submitted for any other degree or professional qualification.

- Chapters 3 and 4 has been presented in the 2017 Organization for Human Brain Mapping Annual Meeting, and is based on a manuscript which is currently under a revision for submission.
- The pre-processing steps mentioned in Chapter 5 have been turned into Python package ‘funpack’ ([87]; <https://git.fmrib.ox.ac.uk/fsl/funpack>) which is used to process UK Biobank data specifically.
- Part of the work in Chapter 5 has been presented in the 2018 Organization for Human Brain Mapping Annual Meeting, and is in preparation for publication.

Abbreviations

- a.k.a: also known as
- BA: Brodmann Area
- BM: Brain Measure
- BMI: Body Mass Index
- BOLD: Blood-Oxygen-Level-Dependent
- CCA: Canonical Correlation Analysis
- CV: Cross-Validation
- FA: Factor Analysis
- FC: Functional Connectivity
- fMRI: functional Magnetic Resonance Imaging
- GFA: Group Factor Analysis
- GLRM: Generalised Low Rank Model
- HCP: Human Connectome Project
- ICA: Independent Component Analysis
- ICVF: Intra-Cellular Volume Fraction
- IDP: Imaging Derived Phenotype
- INT: Inverse Normal Transformation
- ISOVF: ISO-tropic or free water Volume Fraction KNN: K-Nearest-Neighbour
- LOOCV: Leave-One-Out Cross-Validation
- MAD: Median Absolute Deviation
- MCCA: Multi-view Canonical Correlation Analysis
- MD: Mean Diffusivity
- MO: diffusion tensor mode
- MRI: Magnetic Resonance Imaging
- OD: Orientation Dispersion index

- PCA: Principal Component Analysis
PC(s): Principal Component(s)
PRESS: Predicted Residual Error Sums of Squares
QC: Quality Control
RBF: Radial Basis Function
RC(s): Rotated Component(s)
RL(s): Rotated Loading(s)
rfMRI: resting-state functional Magnetic Resonance Imaging
SDR: Supervised Dimension Reduction
SES: Social Economic Status
SM: Subject Measure
SVD: Singular Value Decomposition
SVM: Support Vector Machines
SVR: Support Vector Regression
tfMRI: task functional Magnetic Resonance Imaging
VE: Variance Explained

Abstract

Neuroimaging has aroused much interest in recent years due to the growth of Magnetic Resonance Imaging (MRI) technology and data acquisition techniques. This has led to an increase in interest for work that links neuroscience to behavioural research – using neuroimaging data to reveal the interplay between brain and behaviours. Latent variable models are popular tools to investigate such relationships, with many studies exploring links between functional MRI and various behavioural and demographic measures. However, a common challenge is the interpretability of the latent variable models, in particular, their applications to large datasets with thousands of variables. In this thesis, we first introduced the basic concepts in neuroimaging and the challenges faced when linking it to behaviours. Then, we introduced the background methods applied in the thesis including latent variable models, predictive models and some widely applied data processing techniques. The discussion focused on clarifying easily confused and misused concepts, the theory and application of some rare model extensions, and the demonstration of cross-validation in chained latent variable models. Many of these notes, to our knowledge, have not been discussed elsewhere. One of the main focuses and contributions of this thesis is the proposal of a dimension reduction method, namely Supervised Dimension Reduction. It aims to improve the interpretation of latent variable models, especially in the application of chaining multiple models together. We applied Supervised Dimension Reduction together with other latent variable models to the Human Connectome Project and the UK Biobank project to study the relationships between neuroimaging and behavioural data. We revealed many interesting patterns between brain and behaviours. Moreover, we further clarified the interpretation of a commonly applied latent variable model, Canonical Correlation Analysis. In particular, the multi-view extension and their applications in brain-behaviour study. In the end, we attempted to use functional MRI to predict a specific behavioural measure: personality. However, no results turned out to be significant under the analysis pipeline we applied.

CHAPTER 1

Introduction

There remain vast gaps in our understanding of the human brain. How the brain processes information; what causes mental illnesses like Alzheimer’s disease and depression; what is the relationship between brain and behaviour. However, before the 21st century, due to the limits of computing technology, our understanding of the brain remained elementary and on a small scale. With the development of technology like magnetic resonance imaging (MRI), and the advancement of machine learning, cross-disciplinary interactions among mathematics, statistics, computer science, it has become realistic and necessary to acquire a deeper insight of the brain on a population level.

MRI as a non-invasive imaging acquisition technique, has allowed researchers to acquire 3D brain images with high resolution. By collecting those images on a population level, it opens many lines of data-driven research. Especially in recent years, the dynamics of the resting brain, i.e. brain in the absence of any tasks, has emerged as an exciting research subject in brain science. Many empirical studies, for example [29] and [138] have observed aberrant dynamics of diseased brains compared with normal controls. However, the cause of such alterations is still unclear and it is currently believed that signals from resting brain contain invaluable information to the understanding of general human behaviours and psychological disorders. Resting-state functional MRI (rfMRI) is a promising approach to unravel the fundamental mechanisms underpinning the normal and diseased brain.

Statistical and machine learning methods have proved successful in analysing large-scale neuroimaging data for various purposes. In particular, latent factor/variable models are extensively used for understanding the inherent structures of brain functionality and mental disorders ([117], [57], [118]), and finding relationships between brain and other health related data modalities ([115], [92], [71], [53]); supervised/unsupervised learning and regressions are getting increasingly popular and accurate in predicting all kinds of behaviours, demographics and disease using neu-

roimaging data ([32], [43], [111], [64]).

Lately, brain modelling has garnered considerable attention by many governmental organisations and industrial companies. As a result, large-scale health data/research projects with focuses on multiple brain imaging modalities, mental disorders, human demographic and behavioural measurements emerge as required. Notable backing has included the Alzheimers Disease Neuroimaging Initiative (ADNI; [141]), Human Connectome Project (HCP; [142], [131]) and UK Biobank (UKB; [143], [116]). Theses projects offer unprecedented opportunities for researchers to unveil brain mechanisms and its links to behaviours and mental disorders.

In the rest of this chapter, we will introduce the basics of neuroimaging including the pre-processing and the derivation of brain functional connectivity, both of which are not the focus of this thesis but provided as background reading. Then, we will introduce the interesting questions and methods used to address them. In the end, there will be the structure and outline of this thesis.

1.1 Neuroimaging

There are many types of imaging techniques that allow brain image acquisition in a non-invasive way, for example Electroencephalography (EEG; [109]), Computed Tomography (CT; [67]), Positron Emission Tomography (PET; [11]), Magnetoencephalography (MEG; [56]) Magnetic Resonance Imaging (MRI; [98]).

Compared with the rest of the imaging techniques, MRI has the advantages of having higher spatial resolution and not using ionising radiation so that it is better at localising signals. Because of this MRI is the dominant neuroimaging technique in research and produces most of the neuroimaging data ([94], [81], [131], [3]). Since MRI is the source of neuroimaging data in this thesis, the rest of the section focuses on the introduction of MRI. For details of the other modalities, readers can refer to the references provided above.

MRI scanners use a strong magnetic field to stimulate atomic nuclei, more specifically hydrogen, to emit radio frequency signal. Hydrogen abundantly exists in water and fat, and possesses a magnetic moment conceptualised as spin, like that of a spinning top. When a magnetic field is applied, hydrogen aligns to the external field and then slowly decays back to the relaxed state. The rate of relaxation differs in different tissues of the brain. MRI makes use of this property to produce contrast between tissues and thus images. There are different types of MRI depending on the acquisition sequence. Widely applied in neuroimaging includes MRI with T1-

Introduction

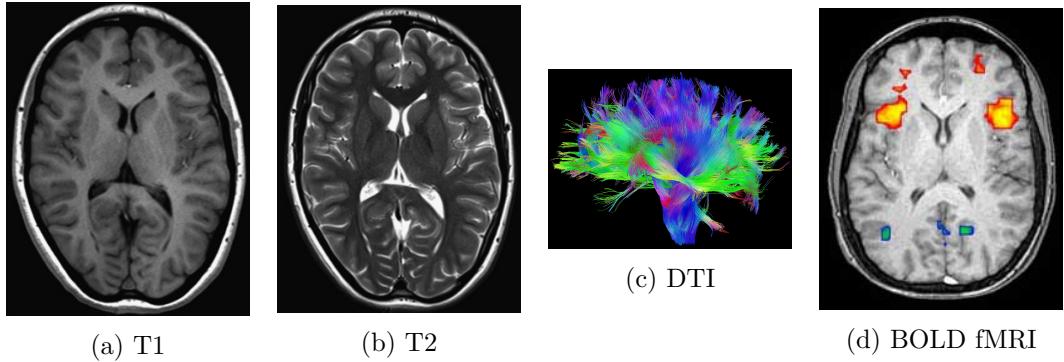


Figure 1.1: Examples of MRI sequences. (a) is T1-weighted structural MRI sequence, and (b) is T2-weighted MRI sequence (both figures are imported from [15]). (c) is another anatomical sequence, diffusion tensor imaging, and the figure imported from [1]. (d) is Blood-Oxygen-Level-Dependent fMRI, one of the functional sequences (figure is imported from [2]).

and T2-weighted contrast, as well as functional MRI (fMRI) and diffusion MRI.

T1 weighted MRI (Fig. 1.1a) is valuable for obtaining structural images as it shows clear boundaries between different brain tissues, white matter (WM; light), grey matter (GM; darker) and cerebrospinal fluid (CSF; darkest) ([89]). Similar to T1, *T2 weighted MRI* also reflects the structure of the brain by providing a different contrast between matters, WM darker than GM and CSF is the lightest (Fig. 1.1b).

The goal of diffusion MRI and specifically *diffusion tensor imaging* (DTI) is to measure the Brownian motion of water molecules in the directions of nerve fibres. In the ventricles, voids filled with cerebrospinal fluid, water diffuses homogeneously in all directions. Water inside of WM diffuses along the same direction as the actual WM fibres and myelin, yet GM is more dense and water cannot diffuse in many directions. Therefore, DTI is extensively used for WM tractography. In addition to strong magnetic fields, different gradient fields are also applied to the brain so that the direction of water diffusion can be captured accurately ([39]). An example image is shown in Fig 1.1c.

While the MRI sequences mentioned so far are all concerned with the anatomy of the brain, *fMRI* measures the activity in the brain. fMRI is generally referred to as the BOLD (blood-oxygen-level dependent) imaging which detects the neural activity by tracking the changes of blood flow related to energy consumption by the brain cells. More specifically, an increased supply of oxygen is carried by haemoglobin in red blood cells to provide energy to active neurons, and oxygenated haemoglobin becomes less magnetic than deoxygenated haemoglobin. BOLD fMRI

makes use of this phenomenon to localise changes in brain activity. During a fMRI scan session, data is collected over units named *voxels*. A 3D MRI image is built out of voxels, a small cube (typically 2mm edge length). The scanner keeps track of the BOLD signal in each voxel to form the raw fMRI data, which is time series of voxels. There are two sub-types of fMRI, *rfMRI* and *task fMRI* (tfMRI). In tfMRI, subjects get scanned while undertaking some task, for example, performing cognitive tests, listening to audio stimuli or watching video clips. tfMRI is used to study the association between brain regions and specific tasks. In contrast, for rfMRI, data is collected while the participant is awake but without any task. There are then no systematic changes in the signal, and instead the interest is in correlations between different brain regions in this ‘default’ mental state. Both tfMRI and rfMRI are extensively used in research to study the functional activity of the brain.

1.2 From MRI to Brain Connectome

The intricate and complex connectivity pattern in the brain is considered as the main attribution of the presence of human psychology and physiology such as emotion, personality, as well as the ability of performing difficult cognitive tasks like carrying out scientific research. Such connectivity pattern can be tracked from microscopic (neuron) level to macroscopic (brain region) level. For each level, it can be further segregated into structural connectivity, functional connectivity and effective connectivity. Studying microscopic connection between neurons has been a historical focus of neuroscience. With the advances of MRI technology and data analysis techniques, complex inter-regional patterns can be observed at the macroscale, allowing researchers to build structural and functional brain connectome at population levels. Connectomes constructed at a regional level are much less computationally heavy than those posed at the microscopic level. Moreover, brain network system has shown more similarities in macroscopic level despite the differences the microscopic details hold ([26]). Therefore, in this thesis, we focus on the macroscopic connectome.

Structural connectivity relates to anatomical connections between neurons, collections of neurons or brain regions which helps us to understand the fundamental architecture of the brain. Neurons or brain regions that are topologically close to each other have relatively higher chance to be structurally connected ([26]). Structural connectivity is mediated by WM which is mainly consisted of myelin. Therefore, DTI is normally used to measure the structural connectivity.

Unlike structural connectivity, *functional connectivity* reflects the functional

Introduction

dependencies between different brain regions, and is a statistical concept. Functional connectivity is complementary to structural connectivity, revealing neurophysiological dynamics on top of anatomical architecture ([26]). Functional connectivity may occur between structurally unconnected brain regions and is more temporally dependent than structural connectivity ([106]). The relationship between the two remains as a challenging topic in neuroimaging research. The majority of connectivity related studies are carried out using functional connectivity, and more calculations and processing steps need to be done to acquire functional connectivity. Therefore, the rest of this section will review the generation of functional connectivity.

Effective connectivity represents the causal relationship between brain regions. With respect to functional connectivity which is un-directed and symmetric measurement between brain regions, effective connectivity forms a directed network. However this kind of connectivity is not studied and used anywhere in this thesis, therefore will not be discussed; for more details see [45].

1.2.1 Generating functional connectomes

Functional connectivity can be estimated by calculating the correlation or covariance between time-series of the BOLD signal measured by the MRI scanner. However, raw data from the scanner cannot be directly used for generating functional connectivity and statistical analysis, and certain pre-processing steps need to be taken. The purpose of pre-processing is to standardise the data to suit mass-univariate modelling. Standard processing steps include:

- Slice timing correction: a 3D MRI image is acquired by taking 2D slices successively following a certain order (e.g. bottom up or top down) and normally takes about 2 seconds to complete the whole brain scanning. Slice time correction is used to correct the temporal misalignments that occur between individual slices due to this sequential acquisition of 2D slices ([62]). An illustration is shown in Fig. 1.2.
- Spatial realignment: this step is to correct for head motion during the scan; each scan is realigned to a reference scan, e.g. first time point ([9]).
- Spatial normalisation: this step is needed due to the different topological shapes between individual brains. It registers individual brain images to a common template using linear and/or non-linear transformation. Such transformation is normally estimated using structural MRI which has better anatomical details, then same transformation can be applied in fMRI ([46]).

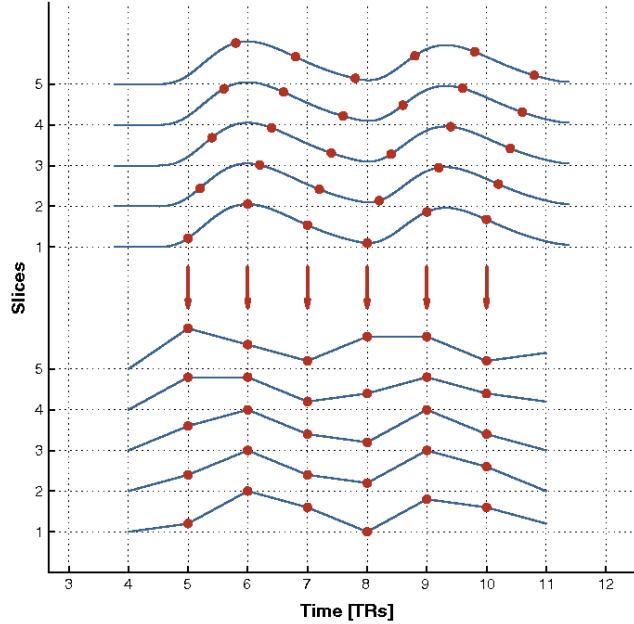


Figure 1.2: Illustration of the slice timing problem. Suppose that an identical signal was present at all slices (top), all slices were acquired at the same time would lead to an incorrect shift in the timing of the signal at other slices. This figure is imported from Fig. 1 in [113], and TRs stands for repetition time which is a sequence parameter.

- Spatial smoothing: it is used to further alleviate the anatomical variability that is not resolved by spatial normalisation. Spatial smoothing is achieved by convoluting MRI images with a Gaussian kernel in the frequency domain. Although this decreases the spatial resolution by removing the high frequency spatial signals, it increases the signal-to-noise ratio and generally improves statistical power. Moreover, it makes the data more Gaussian-like, therefore more compatible with many model assumptions.

Pre-processed raw MRI can then be used to build functional connectivity. Macroscopic connectivity is constructed over brain regions and linked by functional ‘strength’. Therefore, it is important to define those regions and quantify the links. After the pre-processing, MRI data is still at voxel-level, and there are hundreds of thousands of voxels in a whole brain scan. Instead of working with voxel-level MRI, it can be converted to a region-level with brain parcellations, and these parcellations can be based on a single brain or a collection of brains. There are many parcellation schemes based on different approaches, however the general goal is to divide the brain into regions with coherent patterns of anatomical/functional con-

Introduction

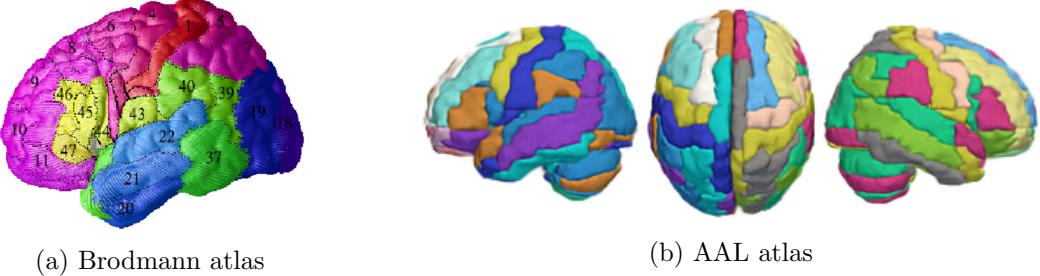


Figure 1.3: Examples of MRI sequences. (a) is the Brodmann atlas which parcellates the brain into 52 regions. This figure is imported from [100] (b) is the AAL atlas, parcellating the brain into 90 regions. This figure is imported from [16].

nections ([106]). Most of the parcellation schemes are based on anatomy like the famous Brodmann’s atlas [25] (Fig. 1.3a) or the Automatic Anatomic Labelling (AAL) atlas [128] (Fig. 1.3b). Such schemes are often only based on a single brain or a few brains, therefore cannot capture individual differences in where coherent structure/function exists. In recent decades, data driven methods using BOLD signal and machine learning techniques have become popular. One such parcellation is the Group-ICA parcellation introduced in [114] and applied in the HCP ([142]) and UK Biobank projects ([143]). Group-ICA uses a data-driven approach to decompose a multi-subject rfMRI dataset into a common set of spatial components and subject-specific time series. Without anatomical data, this approach essentially finds parcellations of space that have maximally homogeneous functional signals. Notably different parcellation schemes parcellate the brain into different number of regions and can lead to significantly distinct functional connectivities. Therefore it is not meaningful to compare results obtained via different brain parcellations.

After the parcellation is chosen, a time series is calculated for each region in the parcellation by aggregating the time series of the voxels in the region, e.g. taking the mean of the voxel-level time series or applying regression. Finally, the functional connectivity is computed by calculating the Pearson’s correlation between every pair of the regions. Normally, the pairwise correlation is further transformed into z-scores using Fisher’s transformation

$$z_{ij} = \frac{1}{2} \log \left[\frac{1 + r_{ij}}{1 - r_{ij}} \right], \quad (1.1)$$

where r_{ij} is the Pearson’s correlation between region i and j .

1.3 Linking Brain and Behaviour

With the availability of complex, large-scale health projects such as the HCP and UK Biobank, which collect health related data from cohorts representative of a broad population, it becomes possible to extend neuroimaging research to population level. With such data, a central goal is to understand the interplay between the brain imaging and health related non-imaging variables. [115] and [92] have demonstrated this using the HCP and UK Biobank data respectively. Both studies discovered certain latent patterns between functional connectivity obtained from rfMRI and behavioural/demographic measures by applying latent variable models, in particular, Canonical Correlation Analysis (CCA) which maximises the correlations between two data modalities. In fact, CCA and its closely related method Partial Least Squares (PLS) have been extensively applied to investigate the links between neuroimaging data and other modalities ([117], [134], [44], [53] and [140]). For example, [91] uses CCA to uncover the differences between depressed and healthy young people by linking their functional connectivity to a set of self-report questionnaires including IQ and demographic data. [41] also explores brain-behaviour relationships by applying sparse PLS. There also have been extensive studies on the proper guidelines of applying those techniques to neuroimaging and behavioural data, as well as extensions of latent variable models ([90], [93], and [57]). However, most of the current studies involving neuroimaging data are limited to two data modalities.

1.3.1 Challenges

Linking neuroimaging and human features such as behavioural and demographic measures is often challenging. A large part of the difficulty is embedded in the nature of the data which can be attributed to several factors. First of all, the disparity between data modalities. Different modalities are measured for different purposes and over different units which makes it hard to compare and carry the same analysis over. Secondly, the data structures and types are especially complex for human features. Unlike neuroimaging data, which is in general continuous and dense, human features are often obtained from questionnaires or human interviews, therefore, the data types are highly mixed (continuous, categorical, binary etc.) with high levels of inconsistency and missingness. Thirdly, the data is very high-dimensional. For functional connectivity, the dimension varies due to different parcellation schemes. For example, if the brain is parcellated into 200 regions, the connectivity dimension

Introduction

for a subject would be at the magnitude of 200^2 . For non-connectivity measures, such as brain structural and behavioural measures, data can be collected to very fine details and over various life aspects leading the data to tens of thousands of dimensions. Not limiting to the factors illustrated above, data can also be confounded by nuisance such as head movements for neuroimaging data, age and gender for human features and so on. Therefore, extensive and careful processing needs to be done to the data prior to any analysis.

Furthermore, to avoid introducing noise to the models which are used to find the relationship (such as CCA), a common technique is to reduce the dimensionality of the observed datasets and then use the lower-dimensional representation in the further analysis. PCA is commonly applied to achieve such purpose ([115], [92], [76]). However, there is normally no (rigorous) justification for the choice of the dimension of the reduced data. It is often a randomly picked number or by looking at the ‘elbow’ of the eigenspectrum.

One of the biggest challenges of applying latent variable model is the interpretation of model results. Model interpretability is of vital importance, especially in health-related research. Linear models are often more preferable in these areas for such reason. However, in the case of the data with ultra-high dimensionality, even linear models become hard to make sense of. Although there is no unified definition of the model/result interpretability, in this thesis, we aim to improve it in the sense of providing more human interpretable results; in other words, to make the composition of latent variable loadings/components easily understood. This includes being able to track the contribution of the components, and the contribution has stable and clear functional meanings.

Finally, fMRI studies have been facing the stability and reproducibility issues ([17], [40]). This is partially due to the small sample sizes. With the access to large-sample datasets, it is essential to cross-validate the results and make sure they are stable, reproducible and generalisable. However, large datasets also raise the complexity of the analysis pipeline, making it difficult to cross-validate.

1.4 Outline and Contributions

This thesis explores the relationships between neuroimaging especially functional connectivity and human features including demographics and behavioural measures with latent variable models. In particular, we apply CCA to investigate such relationships, and its variational methods to extend the study to more than two data

modalities. In addition, the study is then restructured into a prediction problem focusing on predicting a specific trait, personality with functional connectivity. Alongside the analysis, this thesis will focus on addressing the challenges outlined in the previous section, with the particular emphasis on developing a dimension reduction method which improves interpretability and complete analysis pipelines with thorough cross-validation, which guarantee the stability of the results.

The rest of the thesis is organised as follows:

- Chapter 2 will introduce the background statistical and machine learning methods applied in this thesis including latent variable models such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), CCA, and their variations; data processing techniques including several missing data imputation methods; model validation methods such as permutation and cross-validation and in the end, two predictive models. Furthermore, we will clarify some easily confused concepts which may be misused in practice and even published work ¹, as well as clarification on details of rare model extension. For example, the non-monotonicity of canonical correlations in multi-view CCA, which to our knowledge, not explicitly discussed anywhere else. Last but not least, we demonstrate the implementation of model validation in chained latent variable models. This chapter serves as a reference for the future chapters, however when models are applied to specific datasets, the updates and modifications will be introduced in relevant sections. This chapter also provides general guidance on how to implement, cross-validate and interpret particular latent variable models.
- In Chapter 3, we propose a new dimension reduction method that is refined from PCA, called Supervised Dimension Reduction (SDR). It aims to improve the interpretability of PCA on high-dimensional data by imposing variable grouping supervised by human knowledge. At the same time, SDR achieves dimension estimation of the data automatically by a two-way cross-validating algorithm. In addition, we include sign-flipping on the observed variables into the method to further improve the interpretation of the model. SDR will be used as the main dimension reduction method for the HCP and UK Biobank projects which will be presented in Chapter 4 and 5 respectively.
- Chapter 4 explores the links between functional connectivity and human fea-

¹E.g. in the code of [115], the pairwise covariance during a PCA variation is mis-calculated and this variation is clarified in Section 2.2.2

Introduction

tures on the Human Connectome Project by applying SDR introduced in Chapter 3 and CCA introduced in Chapter 2. It will describe the analysis pipeline specifically adopted for the HCP dataset and offer insights on the HCP data structure. We investigate the significant latent patterns between brain and human features and their stability, as well as compare the performance of SDR with traditional PCA on the HCP data. In the end we clarify the interpretation of CCA in particular canonical loadings, and rise caution especially when using them for inference.

- Chapter 5 extends the study to the UK Biobank project and include one more modality, image derived phenotypes (IDPs) into the analysis pipeline. We first explore pairwise relationships between functional connectivity, human features and IDPs. Then we apply multi-view CCA and Group Factor Analysis (GFA) to consider the three modalities at the same time. Due to the data complexity, we will also describe the data pre-processing steps in detail, as well as the adapted analysis pipeline. This chapter will also introduce a permutation method on multi-view CCA which can incorporate three modalities. To our knowledge, this is the first piece of work exploring latent patterns across functional connectivity, human features and IDPs all together. Due to the high volume of results, only highlights will be shown in the main body and full results can be found in appendix.
- Chapter 6 focuses on predicting personality with functional connectivity using two different datasets. We will consider several factors that may affect the prediction including different brain atlases (parcellation schemes), linear and non-linear models and predicting with certain confounds included and removed. This was a novel application and had not been studied (up to when our work was finished).
- The last chapter, Chapter 7 concludes the thesis and discusses some possible future directions.

CHAPTER 2

Background Methods

This chapter introduces the statistical and machine learning approaches used in this thesis, including methods for finding the latent structures in the data, and multivariate methods applied to explore the relationships between different data modalities. This mainly involves the basic model and some variations of PCA, SVD, CCA and factor analysis in general. This chapter also discusses some data processing techniques and predictive models.

2.1 PCA and SVD

PCA is a traditional machine learning technique that is mostly used to reduce the dimension of the data. It decomposes the data along the variable dimension into uncorrelated orthogonal components (principal components). The first principal component identifies the direction with the largest variance; the second principal component lies on the direction of the second greatest variance, and so on. It is equivalent to fitting a k -dimensional ellipsoid to the data and every axis of the ellipsoid represents a principal direction. All these axes form an orthogonal coordinate system. A dataset with k dimensions could have k principal components. PCA can reduce the dimension of the data by projecting the data to the first d ($d < k$) largest axis, i.e. mapping the row vectors of X , \mathbf{x}_i , into the new orthogonal coordinate system. This is achieved by right-multiplying a set of d-dimensional vectors called the *weights* or *loadings*:

$$\mathbf{t}_{ij} = \mathbf{x}_i \cdot \mathbf{w}_j, \quad (2.1)$$

where \mathbf{t}_{ij} is the coordinates of the i th observation transformed by the j th loading vector, i.e. the representation of the original data in the principal space and it is called *principal component* or *score*. \mathbf{w}_j is the j th loading.

Background Methods

Mathematically, PCA is a linear orthogonal transformation and achieved by decomposing the empirical covariance matrix of X . PCA does not treat data with different variances or means differently, so that it is sensitive to the scales of measurements. Therefore, very importantly, as prerequisites for PCA, the data X needs to be column mean centred and standardised to unit variance. The empirical covariance matrix of X is therefore defined as:

$$S = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top, \quad (2.2)$$

where \mathbf{x}_i , $i = 1\dots N$ are row vectors of X and $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$. For simplicity, we assume that data matrix X is column-mean centralised so that the covariance matrix can be written as

$$S = \frac{1}{N} X^\top X. \quad (2.3)$$

The principal loadings and components are determined by the eigenvectors of the covariance matrix S (the deduction can be referred to [20]). We also have $S \propto X^\top X$. Without loss of generality, the scaling factor in Eqn. (2.3) is neglected, and the eigenvectors are normalised to unit length. The eigen-decomposition of $X^\top X$ is

$$X^\top X = V \Lambda V^\top. \quad (2.4)$$

Here V is the eigenvector matrix of $X^\top X$ and Λ is a diagonal matrix with eigenvalues for $X^\top X$ on the diagonal. The columns of V are the *principal loadings*. The product of X and the first d columns of V forms the first d *principal components* (PCs) or *principal scores*, i.e. $\text{PC} = XV$.

2.1.1 SVD

SVD is another technique for dimension reduction and has form:

$$X = U \Sigma V^\top, \quad (2.5)$$

where V is a $k \times k$ orthogonal matrix called the *right eigenvector matrix* of X , and is equivalent to V in Eqn. (2.4). Σ is the *singular value matrix* of X , and $\Sigma = \Lambda^{1/2}$. U is the *left eigenvector matrix* of X . It is also orthogonal and the eigenvector matrix of XX^\top , i.e.

$$XX^\top = U \Lambda U^\top. \quad (2.6)$$

In the context of PCA, V in Eqn. (2.5) is also known as the *principal loadings*

or *weights*. From Eqn. (2.1), we know that PCs are obtained by right-multiply X with the loadings. Rearranging Eqn. (2.5), we get

$$PC = XV = U\Sigma. \quad (2.7)$$

PC is simply singular value scaled left-eigenvector matrix.

In the rest of the chapter, to simplify notation, we will use U^{PC} to stand for PC. Then SVD becomes

$$X = U^{PC}V^\top. \quad (2.8)$$

The advantage of using SVD is that it avoids calculating the covariance matrix of X which has complexity of $O(k^3)$. By applying SVD, we get the left and right eigenvectors at the same time. It can be much faster than the eigen-decomposition of the covariance matrix when the dimension of the data is high.

2.1.2 Principal loadings

Principal loadings play an important role in model interpretation: they illustrate the weights/importance of variables that span the principal space. Principal loadings are defined as the right eigenvectors, V in Eqns. (2.4) and (2.5), and it is an orthogonal matrix. Notably, the orthogonal property only holds on the left and right eigenvectors (U and V in Eqn. (2.5)), not on the principal components (eigenvalue scaled left eigenvectors $U\Sigma$ or U^{PC} in Eqn. (2.8)).

PCA is often used as a method of dimension reduction. As mentioned earlier, this is achieved by keeping the first d (where d is smaller than the original dimension of the data) principal loadings and use them to project the data to a d -dimensional principal space. Although the original loading matrix V in Eqns. (2.4) and (2.5) is orthogonal, after abandoning some of the loadings, the orthogonal property is naturally lost. They however still preserve the principal components being uncorrelated.

One of the biggest challenges in the application of PCA is to choose the number of principal loadings/components to keep. There are many options. One of the most popular approaches is to plot the scree plot (eigenvalue spectrum) and find the ‘elbow’, which is the tuning point on the eigen-spectrum. This is the place where the principal components start to explain considerably less variance. However often in reality, this ‘elbow’ is very obscure therefore hard to identify. Of course one can solve this problem in a probabilistic fashion using Bayesian PCA ([20]), however the price to pay is the implementation complexity and time.

Another difficulty in PCA is the interpretation of the principal loadings. When the dimensionality of the data is high (with hundreds, thousands or even more variables), the understanding of the principal loadings gets extremely challenging. Later on in this chapter, we will introduce a technique, factor rotation to alleviate this problem. In the next chapter, we will introduce a refined method of PCA which incorporates both dimension choosing and loading interpretation solutions.

2.2 Variations of PCA

2.2.1 PCA in high-dimensional situation

When the dimension of the dataset is much higher than the number of subjects, i.e. $k \gg N$, computing covariance matrix is computationally heavy. Moreover, the eigenvalues of $X^\top X$ are the same with the eigenvalues of XX^\top . We can reconstruct the eigenvectors of $X^\top X$ by the eigenvectors of XX^\top . First of all, X has to be column-mean centralised. Then we have

$$X^\top X \mathbf{v}_i = \lambda_i \mathbf{v}_i, \quad (2.9)$$

where λ_i is the eigenvalue and \mathbf{v}_i is the corresponding eigenvector. We left-multiply both sides of Eqn. (2.9) by X to get

$$XX^\top X \mathbf{v}_i = \lambda_i X \mathbf{v}_i. \quad (2.10)$$

If we denote $\mathbf{u}_i = X \mathbf{v}_i$, then Eqn. (2.10) becomes

$$XX^\top \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad (2.11)$$

which is the eigenvector equation for XX^\top . Once we have calculated the eigenvectors for XX^\top , to match back, we multiply both sides of Eqn. (2.11) by X^\top to get

$$(X^\top X)(X^\top \mathbf{u}_i) = \lambda_i (X^\top \mathbf{u}_i). \quad (2.12)$$

Thus, \mathbf{v}_i can be recovered by $X^\top \mathbf{u}_i$. Therefore, when the dimension of the data is higher than the number of observations, we can calculate XX^\top instead of the covariance matrix and reconstruct the principal loadings for PCA.

2.2.2 PCA versus covariance in subject domain

In real-world applications, sometimes we do not need the principal loadings but only the PCs are needed. Besides, with missing data, implementing SVD can be problematic. In such cases, implementing PCA in the subject domain offers an solution. This is almost the same with PCA in high dimensional space, apart from the fact that we do not need to recover the eigenvectors for the covariance matrix of X . Instead we take the eigenvectors and eigenvalues in Eqn. (2.11) to directly serve as the PCs.

However, there is a critical point here we are going to address, the difference between PCA in the subject domain and covariance matrix along the subject domain. The difference is conceptual therefore subtle: in the high dimensional case/PCA in the subject domain, we only calculate XX^\top which is actually a Gram matrix and scaling factor is ignored. Notably here X is pre-centred by column means. However if the covariance matrix in the subject direction is being calculated, i.e. one is calculating the covariance of X^\top . By definition, X^\top needs to be column mean-centred first. We can take it as centring X by row means first and then taking the transpose. This gives a different matrix from X^\top (if the column means and rows means of X are different which is in the majority of cases). Therefore, XX^\top will not be needed to be computed anymore.

To make it more precise, we express these two cases in mathematical forms. Assume X is a $N \times D$ dimensional matrix and not centred by any means. For PCA in high dimensional case or subject domain, the target matrix is calculated as

$$S = \frac{1}{N} \sum_{i=1}^D (\mathbf{x}_i - \bar{\mathbf{x}}_{col})(\mathbf{x}_i - \bar{\mathbf{x}}_{col})^\top, \quad (2.13)$$

where \mathbf{x}_i is a **row** vector of X and $\bar{\mathbf{x}}_{col}$ is the column mean vector. S here is not a strict covariance matrix. For the covariance matrix in the subject domain,

$$S = \frac{1}{D} \sum_{i=1}^D (\mathbf{x}_i - \bar{\mathbf{x}}_i)(\mathbf{x}_i - \bar{\mathbf{x}}_i)^\top, \quad (2.14)$$

where $\bar{\mathbf{x}}_i$ is a $D \times 1$ vector with mean value of $\bar{\mathbf{x}}$ replicated for D times. We notice the differences are the centring factors as well as the scaling factors. Although the scaling factors can be ignored here, when the data has missing values, it may be critical (see Section 2.2.3).

The reason we emphasise these differences here is that it is so trivial and often neglected, therefore it is very easy to make mistakes especially during the

programming stage. Therefore, we recommend when computing covariance in the subject domain, one should centralise data by rows and scale data by row size; be very careful of applying built-in covariance functions (in any software) to calculate covariance in the subject domain, especially when there are missing data existed. They may use the wrong scaling factors.

2.2.3 PCA with missing data

There are various ways of dealing with missing data in PCA. One can apply data imputation methods to fill in missing data. We treat this as data processing technique and will discuss it later. One could also apply Probabilistic PCA to obtain principal components without filling in the missing data. However, this method does not provide principal spaces without missingness (respective entries in the data are still missing in the principal space) [20]. Since this approach involves an iterative algorithm, the computation might be heavy. It causes inconvenience for the further analysis if the following method requires no missingness. We can get an unbiased estimation of the covariance matrix (assuming missing at random) by calculating the pairwise covariance matrix.

Assuming X has N rows (subjects) and D columns (dimensions), and is column-mean centred. Pairwise covariance is computed as follows: for every pair of columns, \mathbf{x}_i^\top and \mathbf{x}_j in X , $i, j \in [1, D]$, we take rows with no missing value in either column i or j , and then take their product to be the respective entry in the covariance matrix, S_{ij} . Note here we are taking the product instead of the covariance of the two column-vectors. For matrix with no missing data, taking product would be equivalent with taking covariance for pair-wise covariance computation. However, with missing data, taking covariance of the column vectors would lead to different scaling factors for different pairs of columns since every pair may have different number of rows with no missing data, and therefore, each entry in the pairwise covariance matrix would be scaled by different factors. This is also the reason to distinguish Gram matrix from the covariance matrix in the subject domain mentioned in the section above (the former has no scalar the latter has). This mistake is again very easy to make especially during coding in software. The disadvantage of this method is that it does not guarantee a positive definite covariance matrix, one may need to apply other techniques to find the nearest positive definite matrix.

2.2.4 PCA on data with mixed data types

PCA is designed for continuous data. Now we will discuss the application of PCA to other data types, categorical ordinal (where data has natural ordering), binary (a special type of ordinal with only two values), and categorical nominal (no particular relationship between different numbers). There are different ways of treating non-continuous data. One can apply multiple correspondence analysis (MCA) which is a special case of correspondence analysis and represent data as points in a low dimensional Euclidean space ([79], [5], [6]); one can apply polychoric PCA which is a maximum likelihood method ([73]), or the Filmer and Pritchett procedure (also known as the dummy coding) which binarises discrete variables ([42]). Each of the method has its own pros and cons. Polychoric PCA does not violate the PCA assumption on the data distribution and is asymptotically efficient. However, it requires iterative procedure thus can be very slow, especially for high dimensional data. The MCA method is designated for nominal data and may not outperform ordinary PCA on other data types. Filmer and Pritchett procedure assumes the experimental group (coded as 1) only compares with the control group (coded as 0), therefore imposes fewer assumption on the data. If it is applied to ordinal data, the order will be lost, therefore, Filmer and Pritchett procedure would not perform well.

In fact, for binary and ordinal data types, one can ignore the discreteness and extend PCA naturally to them ([73]). The advantage of doing so is that it is very easy to apply and does not change the objective. The disadvantage would be that it violates the normal distribution assumption that PCA has. It turns out in real applications, applying ordinary PCA to ordinal data does not have much difference compared with other methods, and the principal components obtained from ordinary and polychoric PCA are very similar ([73]). Therefore in the rest of this thesis, during PCA we ignore the discreteness of ordinal and binary variables. For categorical nominal, we simply apply the Filmer and Pritchett procedure, turning a nominal variable with n categories to $n - 1$ dummy variables and then apply ordinary PCA (since all dummy variables would be binary).

2.3 CCA

CCA is way of measuring linear relationship between two sets of multidimensional data. If we have two vectors of random variables X_1 and X_2 , CCA will seek for vectors a and b such that random variables $a'X_1$ and $b'X_2$ maximise the correlation $\rho = \text{corr}(a^\top X_1, b^\top X_2)$. The linear combinations $a'X_1$ and $b'X_2$ are called the

Background Methods

canonical variables/variates, and denoted as P and Q . The objective function in CCA is

$$\begin{aligned} \max_{a,b} \quad & \text{corr}(P, Q) = \frac{\text{cov}(P, Q)}{\sigma_P \sigma_Q} = \frac{a^\top \Sigma_{X_1 X_2} b}{\sqrt{a^\top \Sigma_{X_1 X_1} a} \sqrt{b^\top \Sigma_{X_2 X_2} b}} \\ \text{s.t.} \quad & a^\top \Sigma_{X_1 X_1} a = 1 \quad \text{and} \quad b^\top \Sigma_{X_2 X_2} b = 1, \end{aligned} \quad (2.15)$$

where σ is standard deviation, Σ is covariance matrix.

In matrix form,

$$P = X_1 A, \quad Q = X_2 B. \quad (2.16)$$

A and B are called the *canonical weights* of data matrices X_1 and X_2 respectively. The correlation that has been maximised is called the *canonical correlation* and denoted by R . R is a diagonal matrix with the column-wise correlations of P and Q on diagonal. P and Q are orthogonal matrices, and also satisfy $\text{corr}(p_i, q_k) = 0$ where $i \neq k$.

The first columns of P and Q are called the first pair of canonical variables, and they have the largest correlation. Then CCA seeks another pair of variables maximising the same correlation subject to the constraint that the second pair is uncorrelated with the first one. This procedure goes on until dimension equals to $\min\{\text{rank}(X_1), \text{rank}(X_2)\}$.

CCA is another dimension reduction method. Unlike PCA which happens under a single dataset setting, CCA is applied to two sets of variables. It reduces both sets of data to their most correlated latent spaces.

CCA has the following widely used terminologies:

- Canonical loadings: they are simply the Pearson's correlations between the canonical variables and the observed/input variables, i.e. $\text{corr}(P, X_1)$ and $\text{corr}(Q, X_2)$. We use this measure instead of canonical weights to assess the variable importance due to the instability of canonical weights ([12], [121] and [59]).
- Variance explained by the canonical variables (P or Q) in the original data (X_1 or X_2): this is computed by taking the average of the R-squared value of the canonical and the observed variables.
- Canonical cross-loadings: $\text{corr}(X_1, Q)$ or $\text{corr}(X_2, P)$. It tells about the correlation between canonical variable of one set and the other observed variable.
- Canonical functions: equations in (2.16), describing the relationships between the canonical variables and the inputs of CCA.
- Canonical roots: R^2 . This is also called the *Amount of Explained Variance*. It

provides an estimate of the amount of shared variance between the canonical variates, i.e. the variance of $P(Q)$ explained by $Q(P)$.

- The amount of shared variance: This is the amount of shared variance in X_1/X_2 included in P/Q . Such information can be obtained by canonical weights A and B which tells us the correlation between each input variable and its own canonical variable. By taking the average of squared each of the loadings, we get the amount of shared variance between the observed variable and canonical variable.
- Redundancy index: This is the amount of variance in a canonical variate explained by the other canonical variate in the canonical function. e.g. a redundancy index of P represents the amount of variance in X_1 explained by Q . This is also the product of the *amount of explained variance* and *amount of shared variance*.
- Variance explained: in this thesis, when we refer to the variance explained, we specifically mean the R-squared value (also known as the coefficient of determination) between canonical variables and the observed variables or the inputs of CCA, i.e. the amount of variance in the observed variables or CCA inputs accounted by canonical variables. It is calculated as the squared value of Pearson's correlation between the observed and canonical variables.

In the end, there are some notation issue and concepts need to be clarified. In our analysis, we distinguish the concepts of *canonical loadings* from *canonical weights*. Canonical weights are the coefficients direct output from CCA, i.e. A and B . Canonical loadings (also known as *structural coefficients*) are defined as the correlation between canonical variable and the observed data which in many cases also the inputs of CCA. However, when the observed data is of very high dimensionality, especially $D > N$, to reduce the noise in the data and solve the degeneration issue, we often reduce the dimensionality of the observed data and then feed the reduced data to CCA. PCA is often used in this case. Therefore, the inputs of CCA would be the principal components of the observed datasets. However, this gives us two types of canonical loadings: the correlations between canonical variables and the observed datasets, i.e. $\text{corr}(P, X_1)$, $\text{corr}(Q, X_2)$; the correlation between canonical variables and the inputs of CCA, principal components, i.e. $\text{corr}(P, U_1^{PC})$, $\text{corr}(Q, U_2^{PC})$.

As the principal components are linear combinations of the observed data, the second type of the canonical loadings are generally uninterpretable. The first type is therefore used in most cases to assess CCA results.

2.3.1 CCA as generalised eigenvalue problem

Eqn. (2.15) can be solved with the help of Lagrangian multiplier (λ). The problem of CCA becomes

$$\mathcal{L}(\lambda; a, b) = a^\top \Sigma_{X_1 X_2} b - \frac{\lambda_1}{2} (a^\top \Sigma_{X_1 X_1} a - 1) - \frac{\lambda_2}{2} (b^\top \Sigma_{X_2 X_2} b - 1). \quad (2.17)$$

To solve the above problem, we take derivatives with respect to a and b and get the following system of equations:

$$\frac{\partial \mathcal{L}}{\partial a} = \Sigma_{X_1 X_2} b - \lambda_1 \Sigma_{X_1 X_1} a = 0 \quad (2.18)$$

$$\frac{\partial \mathcal{L}}{\partial b} = \Sigma_{X_2 X_1} a - \lambda_2 \Sigma_{X_2 X_2} b = 0. \quad (2.19)$$

We can easily show that $\lambda_1 = \lambda_2 = a^\top \Sigma_{X_1 X_2} b$ which also is the correlation $\text{corr}(P, Q)$. We can write (2.17) as a generalised eigenvalue problem:

$$S_c v = \lambda S_w v, \quad (2.20)$$

where S_c is the cross covariance super-matrix having form $\begin{pmatrix} 0 & \Sigma_{X_1 X_2} \\ \Sigma_{X_2 X_1} & 0 \end{pmatrix}$; S_w is a block-diagonal matrix with the within-set covariance on diagonal, $\begin{pmatrix} \Sigma_{X_1 X_2} & 0 \\ 0 & \Sigma_{X_2 X_1} \end{pmatrix}$; v is eigenvector of the system and also the canonical weights super-matrix $\begin{pmatrix} a \\ b \end{pmatrix}$, and λ is the eigenvalue of the system and also the canonical correlation we are trying to maximise between P and Q . Therefore, solving CCA would be essentially solving the eigenvectors of the system (2.20).

2.3.2 Multi-view CCA

Multi-view Canonical Correlation Analysis (MCCA) is an extension of the traditional CCA to more than two datasets ([65], [70]). It is also known as multi-set or multi-way CCA ([82], [107]). MCCA tries to optimise certain objective between more than two sets of data with co-occurring samples. There are several objective functions we can follow in the multi-view setting. For example, one can try to maximise the sum of between sets variance (known as SUMVAR); or minimise the variance between canonical variables (known as MINVAR) ([70]). In this study, we consider the most common and intuitive extension, maximising the sum of correlations between canonical variables (known as SUMCOR; [70]), and specifically focus

on the three-view scenario.

For the three-view SUMCOR extension, we denote the original datasets as X_1 , X_2 and X_3 , the latent canonical variables as P , Q and R respectively. The corresponding canonical weights are denoted as a , b and c . The equivalent objective function with Eqn. (2.15) becomes

$$\begin{aligned} \max_{a,b,c} \quad & \text{corr}(P, Q) + \text{corr}(P, R) + \text{corr}(Q, R) = \frac{a^\top \Sigma_{X_1 X_2} b}{\sqrt{a^\top \Sigma_{X_1 X_1} a} \sqrt{b^\top \Sigma_{X_2 X_2} b}} \\ & + \frac{a^\top \Sigma_{X_1 X_3} c}{\sqrt{a^\top \Sigma_{X_1 X_1} a} \sqrt{c^\top \Sigma_{X_3 X_3} c}} + \frac{b^\top \Sigma_{X_2 X_3} c}{\sqrt{b^\top \Sigma_{X_2 X_2} b} \sqrt{c^\top \Sigma_{X_3 X_3} c}} \\ \text{s.t.} \quad & a^\top \Sigma_{X_1 X_1} a = 1, \quad b^\top \Sigma_{X_2 X_2} b = 1, \quad \text{and} \quad c^\top \Sigma_{X_3 X_3} c = 1. \end{aligned} \quad (2.21)$$

Similar to Eqns. (2.17) to (2.18), this optimisation problem can be transformed into the following Lagrange function and partial system with the help of Lagrangian multipliers λ_1 , λ_2 and λ_3 :

$$\begin{aligned} \mathcal{L}(\lambda; a, b, c) = & a^\top \Sigma_{X_1 X_2} b + a^\top \Sigma_{X_1 X_3} c + b^\top \Sigma_{X_2 X_3} c - \frac{\lambda_1}{2} (a^\top \Sigma_{X_1 X_1} a - 1) \\ & - \frac{\lambda_2}{2} (b^\top \Sigma_{X_2 X_2} b - 1) - \frac{\lambda_3}{2} (c^\top \Sigma_{X_3 X_3} c - 1), \end{aligned} \quad (2.22)$$

and

$$\frac{\partial \mathcal{L}}{\partial a} = \Sigma_{X_1 X_2} b + \Sigma_{X_1 X_3} c - \lambda_1 \Sigma_{X_1 X_1} a = 0 \quad (2.23)$$

$$\frac{\partial \mathcal{L}}{\partial b} = \Sigma_{X_2 X_1} a + \Sigma_{X_2 X_3} c - \lambda_2 \Sigma_{X_2 X_2} b = 0 \quad (2.24)$$

$$\frac{\partial \mathcal{L}}{\partial c} = \Sigma_{X_3 X_1} a + \Sigma_{X_3 X_2} b - \lambda_3 \Sigma_{X_3 X_3} c = 0. \quad (2.25)$$

Notably here the λ s do not hold the equal relations anymore. Actually if we try to solve for λ s in the above system, we get

$$\lambda_1 = a^\top \Sigma_{X_1 X_2} b + a^\top \Sigma_{X_1 X_3} c = \text{corr}(P, Q) + \text{corr}(P, R), \quad (2.26)$$

$$\lambda_2 = a^\top \Sigma_{X_1 X_2} b + b^\top \Sigma_{X_2 X_3} c = \text{corr}(P, R) + \text{corr}(Q, R), \quad (2.27)$$

$$\lambda_3 = a^\top \Sigma_{X_1 X_3} c + b^\top \Sigma_{X_2 X_3} c = \text{corr}(P, Q) + \text{corr}(Q, R), \quad (2.28)$$

subjecting to the unit variance constraints. In matrix form, the equivalent gener-

alised eigenvalue problem becomes

$$\begin{pmatrix} 0 & \Sigma_{X_1 X_2} & \Sigma_{X_1 X_3} \\ \Sigma_{X_2 X_1} & 0 & \Sigma_{X_2 X_3} \\ \Sigma_{X_3 X_1} & \Sigma_{X_3 X_2} & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \lambda \begin{pmatrix} \Sigma_{X_1 X_1} & 0 & 0 \\ 0 & \Sigma_{X_2 X_2} & 0 \\ 0 & 0 & \Sigma_{X_3 X_3} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}. \quad (2.29)$$

The solution to multi-view CCA in Eqn. (2.21) is still the eigenvectors in problem (2.29). However unlike the two-view scenario, the eigenvalues here will not be equal to the sum of the canonical correlations in Eqn. (2.21).

2.4 Factor Analysis

Factor analysis (FA) is a branch of multivariate analysis used to describe variability of observed data using lower dimensional latent factors ([122], [58]). Although in this thesis, the model of factor analysis is not explicitly used, it is very closely related to other methods discussed in this thesis, especially its relationship with PCA. We will provide a short description of FA, highlighting the differences between FA and PCA.

The general model of FA is (assuming observed data X is column-mean centralised)

$$X = ZW^\top + \Sigma, \quad (2.30)$$

where Z is the *latent factor* matrix or the *factor scores*, W is the *factor loadings/weights* and Σ is the noise term with uncorrelated columns.

The goal of FA is to find latent factors Z and its loadings W that minimises the mean square error Σ . There are two types of factor analysis, explanatory factor analysis (EFA) and confirmatory factor analysis (CFA). The latter tests some specific hypothesis that the observed variables are associated with certain factors, whereas there is no a priori assumption made between the observed data and latent factors in the former case. EFA is used to identify the interrelationship between the observed data and latent variables.

Both FA and PCA can be used to reduce the dimensionality of the data in a linear fashion. The most subtle and important difference between PCA and FA is that in FA, especially CFA, the assumption of a causal model underlying the observed data and latent factors is made, which takes into account the random error that is inherent in the observed data (Σ in Eqn. (2.30)), whereas PCA only extracts linear components from the observed data with no randomness accounted for and no hypothetical model assumed. Moreover, PCA decomposes the covariance ma-

trix of the observed data with the diagonal value on the covariance matrix being 1; FA would account for the common variance in the data and the diagonal values are adjusted to different factors ([8]). This readdresses the prerequisite of PCA, the data needs to be measured on the same scale. It is important to normalise data to unit variance before applying PCA. However for FA, it can capture the independent variances between variables therefore does not require this normalisation.

2.4.1 Interpretation of FA

Similarly to PCA and CCA, the interpretation of FA is mainly done via factor loadings (W in Eqn. (2.30)) since the latent factors themselves may not have intrinsic meanings. Factors loadings are analogous to principal/canonical loadings, and represent the relationship/correlation between the observed variable and the unobserved latent factor. The factor loadings can be interpreted as the standardised regression coefficient. The larger the factor loadings, the heavier the observed variable is weighted in the respective latent factor. Furthermore, the squared factor loading can be interpreted as the variance shared between the latent factor and the observed variable.

The stability of factor loadings is affected by the sample size, the variables considered etc. It is generally recommended to use large sample size to improve the reliability, as sample size has been considered being most influential in FA ([63], [85], [54]). Besides, the inclusion of new variables may change the loading drastically. Moreover, the smaller the factor loadings are, the higher the variability in general. Therefore, researchers usually focus on the variables with large loadings to interpret the model.

2.5 Group Factor Analysis

Group factor analysis (GFA) is a factor analysis model that can learn the latent structure of the data when there are more than two views/groups of data with the same samples ([71]). In addition, it can identify latent structures underlying a subset of the input data, i.e. there will be some latent factors only active in one or two views of the data (in a three-view scenario). An illustration of the method with three views as an example is shown in Fig. 2.1.

The model is based on the general FA model in Eqn. (2.30), assuming the noise (Σ in Eqn. (2.30)) is Gaussian with diagonal covariance but separate variance

Background Methods

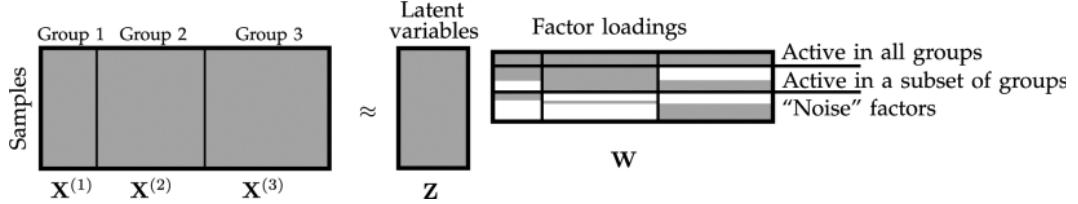


Figure 2.1: GFA illustration with three views/groups of data (imported from [71]). GFA finds latent factors accounted by all three views as well as by subsets of the three views.

for each group. The likelihood of the data is

$$\mathbf{x}_i^{(m)} \sim \mathcal{N}(\mathbf{W}^{(m)\top} \mathbf{z}_i, \tau_m^{-1} \mathbf{I}), \quad (2.31)$$

where $\mathbf{x}_i^{(m)}$ is the i th sample in the m th view; $\mathbf{W}^{(m)}$ is the loading matrix for view m ; $\mathbf{z}_i \in \mathbb{R}^K$ where K is the latent dimension and τ_m is the noise precision. The latent variable \mathbf{z}_i is assumed to have unit Gaussian prior $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I})$, and τ_m has a Gamma prior with both shape and rate parameters set to 10^{-14} ([71]). One novel improvement of GFA is that it imposes an advanced structural sparsity prior to the loading matrix W , taking into account all possible dependencies between different views of data (Eqn. (2.32)).

$$p(\mathbf{W}|\alpha) = \prod_{m=1}^M \prod_{k=1}^K \prod_{d=1}^{D_m} \mathcal{N}(\mathbf{w}_{k,d}^{(m)} | 0, \alpha_{m,k}^{-1}), \quad (2.32)$$

where $\alpha_{m,k}^{-1}$ represents the inverse strength of association between the k th factor and the m th view of the data. α is further decomposed to two low rank matrices U and V in the log-space. This enables us to model explicitly the associations between specific latent factors and to uncover the latent structure across subset of views.

GFA solves the above model by using mean-field variational inference. It approximates the posterior with a factorised distribution

$$q(\Theta) = q(\mathbf{Z})q(\mathbf{W})q(\tau)q(\mathbf{U})q(\mathbf{V}), \quad (2.33)$$

where $\Theta = \mathbf{Z}, \mathbf{W}, \tau, \mathbf{U}, \mathbf{V}$. The approximation is found by minimising the Kullback-Leibler (KL) divergence from $q(\Theta)$ to $p(\Theta|\mathbf{Y})$. KL divergence measures the difference between two probability distributions ([75]), and is defined as

$$D_{KL}(P||Q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx. \quad (2.34)$$

Plugging $q(\Theta)$ and $p(\Theta|\mathbf{X})$ into Eqn. (2.34), and Baye's rule we have

$$\begin{aligned} D_{KL}(q(\Theta)||p(\Theta|\mathbf{X})) &= \int q(\Theta) \log \left(\frac{q(\Theta)}{p(\Theta|\mathbf{X})} \right) d\Theta \\ &= \int q(\Theta) \log \left(\frac{q(\Theta)p(\mathbf{X})}{p(\mathbf{X}|\Theta)p(\Theta)} \right) d\Theta \\ &= \int q(\Theta) \log p(\mathbf{X}) d\Theta - \int q(\Theta) \log \left(\frac{p(\mathbf{X}|\Theta)p(\Theta)}{q(\Theta)} \right) d\Theta \\ &= \log p(\mathbf{X}) - \int q(\Theta) \log \left(\frac{p(\mathbf{X}, \Theta)}{q(\Theta)} \right) d\Theta. \end{aligned} \quad (2.35)$$

The KL divergence is always non-negative, therefore,

$$\log p(\mathbf{X}) \geq L(\Theta) = \int q(\Theta) \log \left(\frac{p(\mathbf{X}, \Theta)}{q(\Theta)} \right) d\Theta, \quad (2.36)$$

where $L(\Theta)$ is defined as the expected lower bound. To minimise $D_{KL}(q(\Theta)||p(\Theta|\mathbf{X}))$, the algorithm will find the set of parameters that maximise the expected lower bound $L(\Theta)$, and this is achieved numerically by applying a second order approximate gradient method, L-BFGS ([97]).

2.5.1 Model interpretation

In Eqn. (2.32), we notice there is a free parameter K which specifies the dimension of the latent space. A different K defines a different model. A recommended practice by [135] is to run the model with incremental K until empty factors (factor loadings are all 0s) are found. Solutions with all factors in use (all loadings being non-zero) may suggest a found factor actually represents multiple true factors, therefore is a sign of insufficient K being specified.

Like other latent factor models, the interpretation of GFA also focuses on the loading matrix W which describes the variable importance in the latent space. After K has been determined, the loading matrix W would be sparse. Where factors are active on more than one view, they depict the interplay between the views, and reveal the shared latent structure of the data. Factors specific to only one view illustrate the complexity of the residual variance within that view of the data, which has not been explained by the shared factors.

2.5.2 Comparison with multi-view CCA

Both multi-view CCA (MCCA) and GFA are linear models that can capture latent structures across all views/groups of the input data. The main difference between them is that GFA can capture relationships between a subset of the views and view specific variations at the same time, whereas MCCA cannot.

GFA is setup in a totally different framework compared with MCCA, the former being probabilistic parametric and the latter is non-parametric. To my knowledge, there has not been a multi-view CCA method extended in a Bayesian fashion. This implies that MCCA will offer a definitive solution whereas the solution from GFA is based on likelihood and finding a unique optimal solution may not be possible. Moreover, MCCA is much faster to run than GFA especially in high-dimensional cases. In fact, it is mentioned in [71] that GFA inference scales linearly with respect to D , M and N , however has cubic complexity with respect to K . If the true number of latent factors is high (e.g. hundreds), the method may even become unpractical to apply.

2.6 Indeterminacy and Non-identifiability Issues

All latent variable have the indeterminacy and non-identifiability issues which means their solutions cannot be uniquely identified and the signs of the solution are indeterminate (can be randomly flipped). As the goal of PCA is to find a new orthogonal coordinate system that spans the most variance in the data, only the principal axis matter and not the signs or magnitudes. Therefore, without loss of generality, all the columns of the loading matrices can be assumed to have unit length. Moreover, when decomposing the covariance matrix and taking the eigenvectors, the direction of the eigenvectors can point to either way, i.e. the signs of eigenvectors are indeterminate. For example, if \mathbf{v}_i is one of the eigenvectors, $-\mathbf{v}_i$ can be the eigenvector corresponding to the same eigenvalue. If D is a diagonal matrix with either 1 or -1 on diagonal, then sign changing in columns of the the eigenvector matrix V is equivalent with right-multiply a matrix D with corresponding entries on diagonal being -1 . Therefore, from Eqn. (2.4), we can infer

$$X^\top X = (VD)\Lambda(VD)^\top = VD\Lambda D^\top V^\top = V\Lambda V^\top. \quad (2.37)$$

Since D is symmetric and orthogonal, $D\Lambda D^\top$ is just Λ . Eqn. (2.37) implies that V and VD can be both eigenvector matrix for X .

Similarly in the high dimensional case, we need to calculate subject covari-

ance XX^\top (scalar is ignored), which is equal to

$$XX^\top = (UD)\Lambda(UD)^\top = UDA D^\top U^\top = U\Lambda U^\top. \quad (2.38)$$

Therefore, the left eigenvector matrix U is also invariant under column sign-flipping.

Similar to PCA, SVD also has the sign indeterminacy problem. As shown in (2.37), SVD is invariant under the same column flipping to U and V ,

$$X = U^{PC}D(VD)^\top = U^{PC}DD^\top V^\top = U^{PC}V^\top. \quad (2.39)$$

In fact, matrix D in Eqns. (2.37) and (2.38) can be any orthogonal matrix and the equality still holds. Therefore, the sign indeterminacy is extended to the model being invariant under orthogonal transformation. This is referred as the *unidentifiable issue*. There is no unique solution to Eqns. (2.37) and (2.38). In CCA, assuming that R is an orthogonal matrix, Eqn. (2.16) is equal to

$$X_1 = PA^\top, \quad X_2 = QB^\top. \quad (2.40)$$

We have

$$\begin{aligned} X_1 &= PR(AR)^\top = PRR^\top A^\top = PA^\top, \\ X_2 &= QR(BR)^\top = QRR^\top B^\top = QB^\top, \end{aligned} \quad (2.41)$$

since for any orthogonal matrix R , $RR^\top = I$. This does not change the covariance between P and Q , and we still have PR and QR being mostly correlated (AR and BR as a solution of Eqn. (2.15)). Analogously, for FA, Eqn. (2.30) becomes

$$X = ZR(WR)^\top + \Sigma = ZRR^\top W^\top + \Sigma = ZW^\top + \Sigma. \quad (2.42)$$

Unidentifiable issue generally exists in latent factor models. It increases the difficulty of the interpretation of the model. The indeterminacy of the loadings suggests one should never interpret the absolute sign of the loadings.

2.7 Factor Rotation

Factor rotation is a very commonly applied technique in factor analysis aiming to improve the interpretability of latent factors, and is usually applied to the factor loadings. It achieves the goal by searching a so-called ‘simple structure’ which

Background Methods

maximises the number of zero/near-zero loadings and only has a small number of non-vanishing loadings ([124]) within each factor. There are several approaches to find the ‘simple structure’, which are mainly divided into two types: the orthogonal factor rotation and oblique factor rotation. As the name suggests, orthogonal rotation keeps the latent axis orthogonal to each other, and thus leaving the latent sub-space invariant; oblique rotation does not preserve such orthogonality ([33], [4]).

In general, factor rotation is carried out by multiplying the loading matrix with a rotation matrix, and is only applied to a sub-space of the original data. The general rotation step is

$$L_{rot} = L_0 R, \quad (2.43)$$

where L_0 is the un-rotated loading matrix; R is the rotation matrix, and L_{rot} is the rotated loading matrix. The rotated latent factors of data X can be obtained by

$$F_{rot} = X L_{rot}. \quad (2.44)$$

One of the most popular methods for orthogonal rotation is named *Varimax* ([69]). It aims to maximise the sum of the variances of the squared loadings. Assume R in Eqn.(2.43) is a $k \times k$ orthogonal matrix. Both L_{rot} and L_0 are $p \times k$ matrices. Suppose the entry in L_{rot} at row i and column j is l_{ij} . We can write the objective function as

$$\max \sum_{j=1}^k \sum_{i=1}^p (l_{ij}^2 - \frac{d_j}{p})^2 \quad \text{subject to} \quad r_j^\top r_j = 1, r_j^\top r_s = 0, \quad j \neq s, \quad (2.45)$$

where $l_{ij} = l_i^0 r_j$, l_i^0 being the i th row in L_0 , r_j being the j th column in R , and $d_j = \sum_{i=1}^p l_{ij}^2$, sum of squared loadings in the j th column of L_{rot} .

There are two limitations of Varimax. First, the rotated loadings may change considerably when additional factors are included. Therefore, picking the rotation sub-space needs additional caution. Second, this method does not work very efficiently when there is a dominant factor. For example, loadings for one factor are evenly high across all variables whereas the loadings for all the other factors are relatively small. This would cause the algorithm calculating R to converge very slowly.

There are other orthogonal rotation methods including Quartimax ([27]), Equimax which are less widely used than Varimax and are not used in this thesis, therefore will not be discussed. Notably, orthogonal factor rotation does not change the total amount of variance of the factors been rotated. However, it changes the

distribution of the variance on those factors. From PCA point of view, after rotating the loadings, the principal components are not necessarily ordered by variance explained anymore. Arguably, rotated principal loadings/components are not principal loadings/components. For future reference, we name them as *rotated loadings/components*.

When the latent factors are believed to be orthogonal or one wants to preserve the orthogonality of the latent space, orthogonal rotation is normally applied to improve the interpretation. In other cases, oblique rotation can be more helpful. Orthogonal rotation can be thought as a special case of oblique rotation. Therefore, in many occasions, oblique rotation is preferred since it provides higher freedom in the form of the latent axis. Orthogonal rotation does not change the coordinate system whereas oblique does. The assumption of the latent factors being uncorrelated is violated in oblique rotation. However, in general the correlation between the latent factors is still low ([4]).

Direct Oblimin and *Promax* ([61]) are the most popular oblique rotation methods. They aim to minimise the covariance of the squared factor loadings with the objective function being

$$\min \sum_{p \neq q} \left(\sum_i s_{ip}^2 s_{iq}^2 - \frac{\gamma}{n} \left(\sum_i s_{ip}^2 \right) n \left(\sum_i s_{iq}^2 \right) \right), \quad 0 \leq \gamma \leq 1, \quad (2.46)$$

where s s are the respective entries in the rotated loading matrix, and γ is a normalising factor; n equals to the number of rows in the original matrix. For more details of the oblique rotation see [58] and [68]. The results from these two methods are often similar but Promax is computationally faster ([4], [61]).

Factor rotation to some degree alleviates the unidentifiable issue of latent factor models by transforming the solution to a more interpretable form.

2.7.1 Factor rotation in PCA/SVD and CCA

Factor rotation can be applied to PCA/SVD and CCA as well. As in the case of factor analysis, the principal/canonical loadings are rotated.

In PCA/SVD, when a rotation matrix R is applied to the loading matrix V in Eqn. (2.5) or (2.4), we get that

$$X = U^{PC} R (VR)^\top = U_{rot}^{PC} V_{rot}^\top. \quad (2.47)$$

In CCA, the rotation becomes more complex since two sets of factors/loadings are involved. Note that in CCA the rotation is not applied to the canonical weights

(A and B in Eqn. (2.16)), but to the canonical loadings or structural coefficients defined as the correlation between canonical variables and the observed data (Section 2.3). [31] shows that both sets of loadings should be transformed by the same transformation matrix, and [38] shows that the implementation of such procedure is possible. Applying rotation matrix R to canonical loading matrices L_1 and L_2 is equivalent to applying R to the canonical weights. Therefore, the correlations between rotated canonical variables stay the same (Eqn. (2.48)).

$$S = \frac{1}{N} P_{rot} Q_{rot}^\top = \frac{1}{N} X_1 A R (X_2 B R)^\top = \frac{1}{N} X_1 A R R^\top (X_2 B)^\top = \frac{1}{N} P Q^\top \quad (2.48)$$

Notably, the distribution of variance explained by the rotated canonical variables will be different and tend to be more evenly spread over the rotated canonical variables ([38]). In addition, the loading matrices in these two methods are originally orthogonal. As the rotations are applied only to a subspace of the loadings, the rotated space is not orthogonal anymore. The rotated subspace can be found by rotating only the significant canonical variables. Methods for finding such will be discussed in Section 2.9.

2.8 General Data Pre-processing Techniques

Data pre-processing can form a standalone subject in the area of data science due to its complexity and broadness of topics. Real-world data often is not suitable to feed into analysis directly. In this section, we will focus on four topics which play important roles in analysing brain imaging and behavioural data, quality control (QC), data normalisation, missing data imputation and data de-confounding.

2.8.1 Quality control

QC is often the very first step to filtering out ill-conditioned variables. There are a few commonly used criteria:

- Missingness of the variables: it shows the proportion of missing data within a variable. Normally variables with missingness higher than a certain threshold (e.g. 50%) are removed from the study. This threshold can be dependent on the sample size. Since with high missingness, the representativeness of the variable is not guaranteed, and missing data imputation methods will not necessarily perform well.

Background Methods 2.8. GENERAL DATA PRE-PROCESSING TECHNIQUES

- Standard deviation (std): normally variables with std equal to 0 are removed since they are not providing any extra information about the data. It can also be used to remove outliers. Within a variable, data points that fall out of two or three std can either be removed (set to NA) or set to a certain number (the maximum, minimum or mean of the variable). This is to prevent outliers drive the analysis. For example, PCA is very sensitive to outliers.
- Pervasiveness of single values: similarly to having 0 std, when a variable lacks distinct values, it often gets removed as it is not very informative. The threshold for removal is generally set fairly high (e.g. 95%, i.e. a variable will be removed if it has more than 95% duplicated values).

2.8.2 Normalisation

There are various methods to normalise data to a desired shape, typically to a Gaussian distribution. This is because many statistical models are based on the assumption that the data or residuals follow a Gaussian distribution [18]. However, often in health data, especially in behavioural data, the variable distributions are not perfectly Gaussian or not Gaussian at all. This requires researchers to normalise the data prior to the analysis.

The most common method is to subtract the column means from the columns and then divide by the column standard deviations. Another approach is to subtract the column means from the columns and divide by the overall standard deviation. This preserves the variation proportion between the columns. It is sometimes preferred in neuroimaging data, since brain regions with high variation tend to be more informative.

Another method is to normalise by the median absolute deviation (MAD) instead of the standard deviation. MAD is a robust estimator of statistical dispersion of univariate quantitative data, and is defined as the median of the absolute deviations from the median of the data ([55]):

$$\text{MAD} = \text{median}(|X_i - \text{median}(X)|). \quad (2.49)$$

MAD is more robust to outliers as it does not involve the computation of the squared distance to the mean. One can also use MAD to infer the sample standard deviation using

$$\hat{\sigma} = k \cdot \text{MAD}, \quad (2.50)$$

where k is a constant and depends on the distribution of the data ([105]). For data

following Normal distribution, $k \approx 1.4826$, or

$$\text{MAD} \approx 0.67449\sigma. \quad (2.51)$$

Rank-based inverse normal transformation is a widely applied method in neuroimaging. It is a type of inverse normal transformation (INT) first proposed by [137]. INTs transform sample distribution of continuous variables into a normal distribution. [13] gives a overview of different INT methods. In brief, rank-based INT first converts a variable to ranks and then applies the probit function, which is the inverse of the cumulative distribution function of the standard normal distribution. The most commonly applied rank-based INT has the form ([13]):

$$\tilde{y}_i = \Phi^{-1} \left\{ \frac{r_i + c}{N - 2c + 1} \right\}, \quad (2.52)$$

where \tilde{y}_i is the transformed value for observation i , $\Phi(\cdot)$ is the probit function, r_i is the rank of the i th observation among the N samples, and c is a constant. Different c represent different variations of the rank-based INT. The most widely used c is recommended by [22] with the value equal to $3/8$, and this is the value used in this thesis. Other values of c also have been proposed: the original proposer of the method, [137] uses $c = 0$; [127] suggests $c = 1/3$ and [21] suggests $c = 1/2$. However all these variations are linear transformations of each other and produce similar results all very close to the normal scores.

2.8.3 Missing data imputation

Missing data arise in many real-world applications. Many statistical models cannot handle data with missing values which has motivated the creation of data imputation methods. Discarding observations/variables with moderate missingness might lead to a substantial reduction in the sample size. Therefore impute missing values with substitutes is necessary for many datasets. It is one of the most challenging issues in data pre-processing especially for discrete and mixed data. There are a few very widely used simple methods, including filling missing values with the mean, median, or a random draw from the data. They work fairly well in univariate cases since they preserve certain statistics of the data. One can also apply simple regressions on the data to fill the missing values. However, such methods generally do not provide uncertainties on the fitted values and tend to reduce the variability of the data.

There are extensive studies on more advanced methods in multivariate anal-

ysis, and comparisons of the methods on different types of data ([83], [10], [14], [130], [66]). We will focus on three popular multivariate imputation methods, **k-nearest-neighbour** (kNN; [36]), **generalised low rank model** (GLRM; [129]), and **soft-impute** ([86]) which have been used in the applications of this thesis.

2.8.3.1 kNN

kNN is best known as a classification and regression method. It is a non-parametric method that classifies or regresses data based on certain metric. For continuous data, the metric is the Euclidean distance and for discrete data, the Hamming distance is normally used. In the classification case, for an unclassified data point, one first selects the k closest neighbours, then this point takes the majority class from the k neighbours as its own class. In the regression case, the value for a target data point is simply the average of the k nearest neighbours.

We can easily adjust this concept to impute missing data. kNN for missing data imputation can be divided into two steps, a selection step and an aggregation step. For a data point with missing value(s), the selection step finds its k nearest non-missing neighbours, and the aggregation step fills the missing values with some aggregated value(s), traditionally the average of all k neighbours. The aggregation step can be extended to use a weighted average instead since closer (based on certain metric/criterion) data points may have more influence on the missing point. A popular way of weighting health data introduced by [119] is to incorporate correlations between data points:

$$w_i = \left(\frac{r_i^2}{1 - r_i^2 + \epsilon} \right)^2, \quad (2.53)$$

where w_i is the weight for the i th neighbour, r_i is the correlation between the i th neighbour and the missing data point, and ϵ is an arbitrary constant term (set to 10^{-6}) to avoid denominator being 0.

There are two instinctive variations of the above kNN for missing value imputation: impute by the k nearest variables (kNN-V) and impute by the k nearest subjects (kNN-S). The distinction is the direction/domain of selecting the neighbours. [83] extends these two variations to accommodate different data types by specifying correlation measures between them. The selection step of kNN-V is to include variables that have the highest k absolute correlations, and with no missing value for the same subject(s) with the target variable. The aggregation step is to apply regression with the method being data type dependent. For binary and nominal data types, the imputed value is the weighted majority vote using weight in Eqn. (2.53). For continuous and ordinal types, it is the weighted average.

Background Methods

The idea for kNN-S is basically the same as kNN-V apart from it seeking similarity between subjects using Gower's distance ([52]). The distance between the n th subject with the target subject is defined by

$$d_n = \frac{\sum_{v=1}^p \delta_{nv} d_{nv}}{\sum_{v=1}^p \delta_{nv}}, \quad (2.54)$$

where d_{nv} is the dissimilarity between the two subjects for variable v , and δ_{nv} indicates whether variable v is available for both subjects (1 for both being available; 0 otherwise). d_{nv} is variable type dependent. For binary and nominal types, d_{nv} is equal to 1 if two subjects agree on this variable and 0 otherwise. For other data types, it is the absolute difference between the values standardised by the total range of the variable.

There are other more advanced extensions of kNN introduced in [83] including the Hybrid imputation by nearest subjects and variables (KNN-H) and Hybrid imputation using adaptive weight (KNN-A) which will not be introduced here (details see [83]). One of the biggest advantages of kNN is that it can deal with different data types whereas most well studied missing data imputation methods like Bayesian PCA only be applied to continuous data. The concept of kNN is simple and flexible in terms of the choices of distance metric and aggregation method. However, the challenges of kNN include the choice of k and its high computational load. The number of nearest neighbours k is a pre-determined parameter in kNN and it often takes several simulations to find the optimal k . For every simulation, kNN needs to compute the distance between the target data point with every other point in the data and store them, and then implement the aggregation method which leads to high computational complexity.

2.8.3.2 Generalised low rank models

Generalised low rank models (GLRM) are essentially a generalisation of PCA to handle different data types. [129] gives detailed introduction for such models. The underlying idea is to approximate the original dataset with two low rank matrices by minimising an objective function. In PCA, this objective function is the least-squares errors. GLRM extends this approach with a general loss function, and uses different loss functions for the different data types. The generalised objective function is

$$\min \sum_{(i,j) \in \Omega} L_{ij}(x_i y_j, A_{ij}) + \sum_{i=1}^m r_i(x_i) + \sum_{j=1}^n \tilde{r}_j(y_j), \quad (2.55)$$

Background Methods 2.8. GENERAL DATA PRE-PROCESSING TECHNIQUES

where $L_{ij}(\cdot) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ is the loss function; A is the original data matrix; XY forms the low rank factorisation of A with x_i being the i th row of X and y_j being the j th column of Y ; $r(\cdot)$ and $\tilde{r}(\cdot)$ are the regularisers for x and y respectively. If we take the loss function in (2.55) to be the squared Frobenius norm, i.e. $L_{ij}(\cdot) = \|\cdot\|_F^2$, this is known as the generalised PCA. Note that (2.55) generalises both loss function and regularisers.

To deal with binary data, the loss function can take the form of $L(u, a) = (1 - au)_+$ which is the hinge loss (takes the larger value between 0 and $1 - au$). This becomes equivalent to *Boolean PCA*. It can also be set to $L(u, a) = \log(1 + \exp(-ua))$, and this refers to *Logistic PCA*. For non-negative integers, $L(u, a) = \exp(u) - au + a \log a - a$. This is known as *Poisson PCA*. Finally for ordinal data which encodes levels of some variable (e.g. the degree of happiness) and takes values from 1 to d , (2.55) becomes the ordinal PCA, and $L(u, a) = \sum_{a'=1}^{a-1} (1 - u + a')_+ + \sum_{a'=a+1}^d (1 + u - a')_+$ ([129]).

To impute the missing data, we fill the missing value A_{ij} with the approximated value \tilde{A}_{ij} , where \tilde{A}_{ij} minimises the loss of the low rank factorisation $x_i y_j$:

$$\tilde{A}_{ij} = \operatorname{argmin}_a L_{ij}(x_i y_j, a). \quad (2.56)$$

GLRM is an elegantly designed method which does low rank approximation of the data as well as missing data imputation. The model offers flexibility in the choice of loss functions and regularisers. However, at the same time, it adds complexity to the analysis since finding the optimal regularisers and loss function can be time consuming. In addition, it is a relatively new method ([129]), and the implementation is not well optimised in the available software and is either non-implementable due to out of maintenance or takes very long time to run ([84]). Therefore, there are few studies using this method and comparing it to other methods.

2.8.3.3 Soft-impute

Soft-impute is a well studied method for missing data imputation and commonly used in neuroimaging ([86], [125], [115], [92]). It replaces missing values with the results from soft-thresholded SVD using an iterative procedure. The general model of soft-impute uses nuclear norm to regularise the squared Frobenius norm between the target data matrix X and its low rank approximation Z . Therefore, it can be thought as a sub-type of GLRM. Suppose target matrix X has rank r , then the

objective function is given by:

$$\min_Z \frac{1}{2} \|X - Z\|_F^2 + \lambda \|Z\|_*, \quad (2.57)$$

where $\|\cdot\|_F$ is the Frobenius norm and $\|\cdot\|_*$ is the nuclear norm. The solution of (2.57) is given by the soft-thresholding of X :

$$\hat{Z} = S_\lambda(X) \equiv U D_\lambda V^\top, \quad (2.58)$$

where D_λ is a diagonal matrix with $(d_1 - \lambda)_+, \dots, (d_r - \lambda)_+$ on the diagonal, and $d_i, i \in [1, r]$ are the r singular values of X .

The imputation procedure is done by initialising the original missing values as 0, and iteratively solving (2.57), and updating the missing values with its solution until Z converges for a given λ . The procedure is repeated for different λ s to find the best λ . [86] has proved that the convergence is guaranteed.

Soft-impute has several advantages: it outperforms many other methods like hard impute, maximum-margin matrix factorisation, and singular value thresholding ([86]); it is minimally parameterised with only one parameter λ ; the objective function uses a convex relaxation, nuclear norm, whereas in hard-impute, the regularisation term is the rank of the data which is non-convex. However, soft-impute is limited for the application of continuous data with missing at random, and due to the nature of iterative procedure, it can be very time-consuming.

In conclusion, based on our experience, when the data are mostly continuous and ordinal, both soft-impute and kNN are easy to implement and reasonably quick to run. However, kNN provide more flexibility in terms of the choice of similarity function and direction of imputation (row-wise or column-wise). For example, when subject-wise correlations exist in the data, one can choose to use kNN-S (row-wise imputation). GLRM is powerful and can be adapted to all kinds of data types. However, there lacks well-written libraries in softwares, therefore, rarely used in practice.

There are other data imputation methods that are designed for multivariate analysis and can incorporate different data types. One such method is Multiple Imputation by Chained Equations which iteratively uses multiple regression on other variables to replace missing values ([10]). Each type of variable can be modelled by different regressions, for example, binary variables are modelled by logistic regression and continuous variables are modelled by linear regression. However, the actual implementation of this method turns out to be very time consuming for high

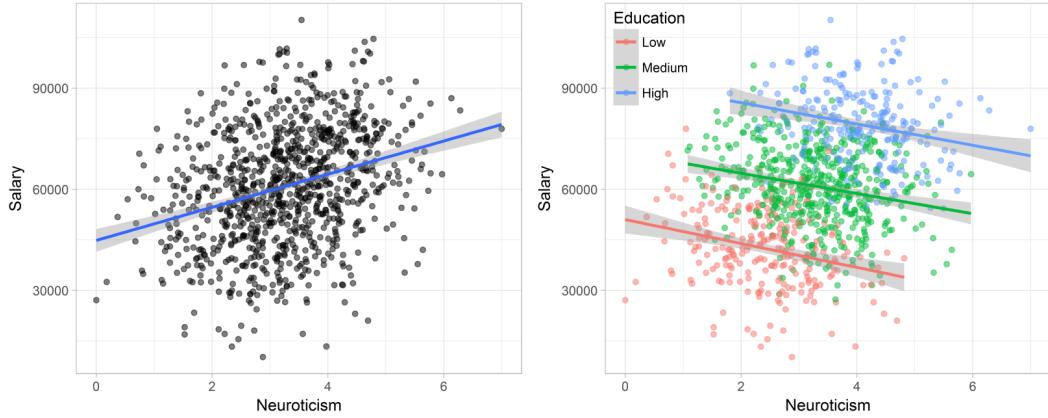


Figure 2.2: An example of Simpson’s paradox. This figure is imported from [77]. Salary and Neuroticism appear to show positive correlation in the data. However, after regressing out Education, they show negative correlations.

dimensional data with hundreds or thousands of variables, therefore is not applied in this thesis.

To compare the accuracy of different imputation methods, one can generate synthetic data. However, in real-world data, the true structure is very rarely known, and therefore very hard to compare between methods. A practical solution is to use the final objective to assess the imputation method. For example, in prediction, use the final prediction accuracy.

2.8.4 De-confounding

In statistics, a confounder is a variable that affects or is correlated with both explanatory and response variables, and their effects are not of interest to the researchers. Without removing confounders, they prevent the identification of the true relationship between the variables of interest. This phenomenon is well-known as the Simpson’s paradox ([112]) and Fig. 2.2 gives an example. Salary appears to be positively correlated with Neuroticism on the whole population, however after controlling for (regressing out) education, they show negative correlations.

There are different ways of controlling confounders when designing the studies ([139]). For example, one can set control groups (this method is not used in this thesis, and details can be referred to [139]). In our analysis, we choose to use linear regression, regressing out the effects of the confounding variables. Suppose the confounding set is C , and the target data matrix is Y . The effects of C on Y

Background Methods

expressed as a linear model is:

$$Y = C\beta + \epsilon, \quad (2.59)$$

where ϵ is the residual variance unexplained by C . The best estimator of β is

$$\hat{\beta} = (C^\top C)^{-1} C^\top Y. \quad (2.60)$$

Removing the effect of C from Y gives

$$Y - C\hat{\beta} = Y - C(C^\top C)^{-1} C^\top Y. \quad (2.61)$$

The more challenging step in the de-confounding is to select the confounders. One can hardly control for all the nuisances in a study. One of the main reasons is that not all confounding variables can be collected/measured. It is also partly because researchers do not have complete knowledge of the data they are analysing. Therefore, de-confounding is done with the hope that no obvious or significant nuisance confounders the results severely. It is often the case that the confounding set starts with variables selected based on researcher's prior knowledge of the data. As the analysis goes, confounders reveal themselves at different stages, and are added to the confounding set for the further analysis.

Unless the variables of interests include age, gender, race, these basic demographic measures are normally taken as confounders. Commonly, the effect of their interactions (e.g. age \times gender) or higher order of these variables (e.g. age 2) are removed from the study ([115], [92]). In the area of neuroimaging, head size and head movement during scan are generally considered as confounders.

2.9 Model Validation and Assessment

Once the model is built and the results are obtained, it is very important to make sure the model performs consistently and the results of the analysis are reproducible and generalisable. Therefore, it is necessary to validate the stability of the results, assess model accuracy and check whether the model has overfitting problem.

2.9.1 Cross-validation

Cross-validation (CV) probably is the most well-known method for model validation in statistics/data science. It is widely used in predictive modelling for estimating prediction accuracy/errors. However, it can be extended to assess performance

for non-predictive models. The main idea is to train the model on a training set (sometimes known as the held-in set), and validate it on an independent test set (also known as the held-out set) to compare the results with the training set. There are several types of CV: K -fold, leave- k -out ([72]), repeated random sub-sampling validation [120], all of which are applied in this thesis.

K -fold CV splits the data into K equal-sized (or as equal as they can be) sets, using $K - 1$ sets as training data, and the other set as the test set. The whole CV procedure is repeated K times until every fold of the data is served as the test set. The widely used K s are 5, 10 and 2. When K is equal to 2, it is equivalent with split-half CV (using half of the data to train the model and the other half to validate).

Leave- k -out CV is to set k observations/variables out as the test set, and train the model on the rest. Again, the procedure is iterated until all observations/variables have been used as a test set. The most frequently used k is one, leave-one-out CV (LOOCV). When $k = 1$, it also becomes a special case of K -fold for K taking the sample size.

Repeated random sub-sampling validation is a type of non-exhaustive CV method. It splits the data into random sets, taking one subset as test set and the rest as training set. It can be repeated as many times as one wants. Therefore, it is also known as Monte Carlo CV.

The final question is: what is the best type of CV to use? K -fold and leave- k -out CV are exhaustive methods, and are generally considered as unbiased. However, recently LOOCV has been severely criticised in the machine learning community for its instability issues ([133]). The argument is that when testing the model on a single observation, with the training sets being very similar to each other, it maximises the test variance therefore produce unstable and biased results. The computational burden is also considerable since the validation need to be iterated for as many times as the sample size. For such reasons, K -fold is the most widely used among exhaustive methods ([23], [72], [133], [60]). The biggest advantage of Monte Carlo CV is that the number of splits does not depend on the division of the data. The disadvantage is that some of the data may never get tested/trained on. Although Monte Carlo CV is a non-exhaustive method in theory, it has become popular in recent years. Because the procedure can be repeated infinite number of times, researchers can produce more training sets than other methods at a price of being computationally heavy. It is beneficial for hyper-parameter tuning and model selection. Model performance can be tested with more times and then the average is taken to select the best model ([133]).

In practice, the choice of CV method is very case dependent. The general rules of thumb are first of all, the test set should be independent of the training set. Secondly, the training and test sets should not be disproportionately big or small. If the sample size is too small, K -fold loses its power for the small training sets therefore, cannot train the model properly. In such case, LOOCV would be a better choice. However under the circumstances of large datasets, LOOCV should be avoided ([133]), with K-fold or Monte Carlo CV being preferable.

2.9.1.1 A demonstration of CV in the analysis of PCA followed by CCA

One common mistake in implementing CV is not being able to keep test and training sets independently especially when more than one model is involved. This happens, for example, if we split data after selecting features/reduce the dimensions. In such case, the features are extracted based on the whole dataset which will lead to bias during CV. We will use PCA followed by CCA analysis as an example to demonstrate a CV procedure using 5-fold CV (all notations are the same with in Section 2.3).

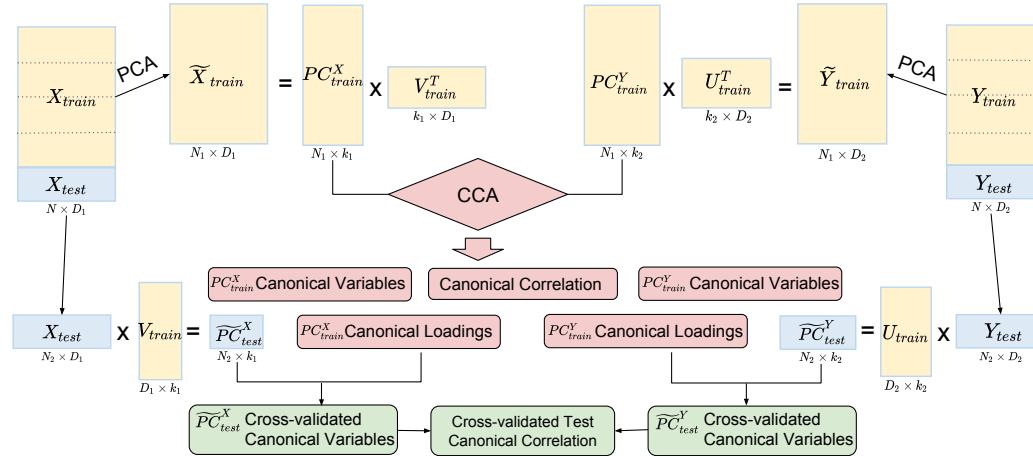


Figure 2.3: Illustration of 5-fold cross-validation (CV) in the analysis of PCA followed by CCA.

The goal of implementing CV in this analysis is to assess CCA performance on the test set when the inputs of CCA are reduced by PCA. Algorithm 1 illustrates this CV procedure, and is further summarised by Fig. 2.3.

Algorithm 1 CV procedure in the analysis of PCA followed by CCA

Step 1 Split the original X_1 and X_2 into 5 roughly equal-sized folds, each using 4 as the training sets (X_1^{in} , X_2^{in}) and the other one as the test set (X_1^{out} and X_2^{out}).

Step 2 For each fold, apply PCA to the training sets of X_1 and X_2 , and use the principal loadings from the training set V^{in} to construct the principal components for the test set

$$\begin{aligned}\tilde{U}_{1,PC}^{out} &= X_1^{out} V_{2,k}^{in}, \\ \tilde{U}_{2,PC}^{out} &= X_2^{out} V_{2,k}^{in},\end{aligned}\quad (2.62)$$

where V_k takes the first k principal loadings to achieve dimension reduction.

Step 3 Feed the PCA-reduced training sets from X_1^{in} , X_2^{in} to CCA to obtain canonical variables P^{in} and Q^{in} , canonical weights A^{in} and B^{in} and canonical correlations R^{in} .

Step 4 Construct *cross-validated canonical variables* for the test set using canonical weights from the training set

$$\begin{aligned}\tilde{P}^{out} &= \tilde{U}_{1,PC}^{out} A^{in}, \\ \tilde{Q}^{out} &= \tilde{U}_{2,PC}^{out} B^{in}.\end{aligned}\quad (2.63)$$

Step 5 Calculate correlations between \tilde{P}^{out} and \tilde{Q}^{out} as the *reconstructed/cross-validated canonical correlation* \tilde{R}^{out} .

Step 6 Go to step 1 to start the next fold.

2.9.2 Permutation testing

Permutation tests are a type of non-parametric statistical significance test [96]. They can be used to test the significance for any test statistics by computing the sampling distribution under the null hypothesis. The sampling distribution under the null hypothesis is generally computed by relabelling/resampling/permuting the observed data. If the null hypothesis is true, the distribution of the test statistic on the relabelled/resampled/permuted data should be indifferent from the one of the original data. At the end of permutation tests, a p-value is computed for the permuted data and is used to assess the significance of the test statistic.

There are several advantages of permutation tests: there is no limitation on the type of the test statistics; the distribution of test statistic does not need to be known, and few assumptions are required; it is easy to implement with high flexibility. Being computationally intensive is one of its main drawbacks. Besides, there is one crucial assumption embedded in permutation testing: the observed data is exchangeable, i.e. shuffling the original data does not affect the test statistic. However, this assumption is often violated in real-world datasets. In other words, we can often detect correlations among observations. For example, the data has temporal/spatial structure, or is collected over subjects who are from the same family (siblings, twins etc.). In such cases, permutation testing is still applicable with extra stratification of the data [48]. For example, for data collected over siblings, subjects can be permuted without separating the siblings, i.e. siblings from the same family need to be permuted as a whole and not to be mixed with the other families. Permutation testing on correlated data unfortunately tends to lose power compared with independent observations. However, other benefits still preserve ([48]).

Permutation tests can be intuitively extended to the framework of latent variable models like CCA, and they are often used to test the number of significant canonical variables ([115], [92]). The procedure is shown in Algorithm 2 (notations are the same as in Section 2.3).

2.10 Predictive Modelling

Prediction is one of the most important and significant purposes for data science, especially in the area of health data. Researchers often would like to predict patients survival rate, risks for developing certain diseases, how certain behaviour affects the brain or the other way round. One golden criterion for model performance is the prediction accuracy on unseen data. Most of the statistical/machine learning

Algorithm 2 Permutation procedure on CCA.

Step 1 Permute one of the two input sets of CCA. Suppose we permute the rows of X_1 and denote the permuted data as X_1^p , keeping X_2 unchanged.

Step 2 Run CCA on X_1^p and X_2 , and record the largest canonical correlation in \mathbf{r}^p .

Step 3 Repeat the previous steps for K times. K is usually a large number, e.g. 1000 or 10000, with $\max K = N!$.

Step 4 The final p-value for the i th pair of canonical variables is calculated as:

$$p_i = \frac{1 + \sum_{k=1}^K \mathbb{1}_{\mathbf{r}_k^p \geq r_i}}{K}, \quad (2.64)$$

where r_i is the canonical correlation between the i th pair of non-permuted data, and $\mathbb{1}$ is the indicator function.

Step 5 The i th pair of canonical variables is considered as significant if $p_i < \alpha$, α generally takes the value of 0.05.

models can be used or generalised to predictive modelling. There are many factors affecting the prediction accuracy including data quality, features extraction and the choice of models. There are mainly two types of predictive models, linear and non-linear. Although in recent years, non-linear models have become very popular in machine learning and data science, linear models still have their own non-replaceable advantages. The most valued one is that they offer better interpretability, which is vital for health data. Moreover, linear models have lower complexity and overfitting can be more easily avoided. In many cases, linear models can be generalised to non-linear data. It is recommended to always start the analysis with some type of linear models.

In this section, we will focus on two widely applied linear models often have fairly good performance, linear regression and support vector machines and its non-linear generalisation.

2.10.1 Linear regression

We have already encountered linear regression models in the de-confounding section (Section 2.8.4). Generalising the confounding set C in Eqn. (2.59) to a normal explanatory variable \mathbf{x} gives the multivariate form of the linear regression model

$$y_i = \beta_0 + \beta_i \mathbf{x}_i + \epsilon_i, \quad i = 1, \dots, N, \quad (2.65)$$

Background Methods

where ϵ is the residual term. In matrix form, this can be expressed as follows:

$$Y = X\beta + \epsilon, \quad (2.66)$$

where Y is the matrix with N response variables, X is a $N \times p$ matrix.

β can be found using the *least squares* method, which finds the beta that minimises the sum of squared residuals (ϵ)

$$\min_{\beta} \text{RSS} = \sum_{i=1}^N (y_i - \beta_0 - \mathbf{x}_i \beta_i)^2. \quad (2.67)$$

Or in matrix form:

$$\min_{\beta} \epsilon^\top \epsilon = (Y - X\beta)^\top (Y - X\beta) \quad (2.68)$$

β has an unique solution for (2.68) as introduced in Eqn. (2.60), $\hat{\beta} = (X^\top X)^{-1} X^\top Y$. Therefore, the predicted value from (2.65) is

$$\hat{Y} = X\hat{\beta} = X(X^\top X)^{-1} X^\top Y. \quad (2.69)$$

For more details about linear regression models including their generalisations, the reader can refer to [20], [60], [110], [95] and [88]. The main point we would like to address here is fitting linear model with confounds. Recall from Section 2.8.4 that with the presence of confounds (denoted as C), researchers can not explore the real relationship between explanatory and response variables. After removing the effects of confounds, linear regression models can be refit between the response variable Y and the explanatory variable X to study their relationships

$$Y - C(C^\top C)^{-1} C^\top Y = X\beta + \epsilon. \quad (2.70)$$

However in the context of predictive modelling, there is confusion sometimes on the definition of confounding variables. As mentioned in Section 2.8.4, age and gender are often treated as confounders, whereas during predictions, they are often informative predictors. If the focus of the predictive models is to improve prediction accuracy, those factors should be included with the other explanatory variables.

$$Y = C\beta_c + X\beta_x + \epsilon. \quad (2.71)$$

If C and X are dependent, which is usually the case (this was the reason of de-confounding), β in Eqn. (2.70) is not equal to β_x in Eqn. (2.71) and β_c in Eqn.

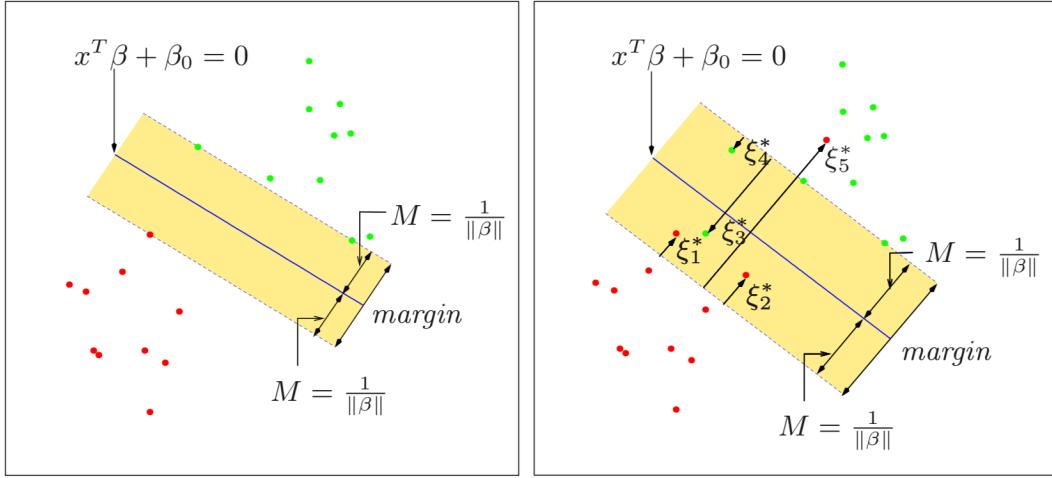


Figure 2.4: Illustration of SVM (Fig. 12.1 in [60]). The left panel shows the linearly separable case; the right panel is the more generalised case (non-separable). The solid line in the middle is the hyperplane separating two classes. The dotted lines form the margin of the SVM and only depend on the nearest data points to the hyperplane. In the right panel, ξ_i represents the distance of the points fall on the wrong side of their margin.

(2.70) is not equal to $(C^\top C)^{-1} C^\top Y$. The final remark is that the importance of the predictors/explanatory variables is interpreted by the sign and value of β .

2.10.2 Support vector machines

Support vector machine (SVM) is a non-probabilistic binary linear classification method. It tries to classify data points in d -dimensional space with a $d - 1$ -dimensional hyperplane into two classes ([60]). Since such linear separation is not unique, the objective of SVM is to maximise the separation between two classes, i.e. the distance of the nearest data points to either side of the hyperplane is maximised.

Suppose the training set of data is $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Then the hyperplane can be expressed as

$$\mathbf{x}^\top \boldsymbol{\beta} + \beta_0 = 0. \quad (2.72)$$

The left plot in Fig. 2.4 illustrates SVM in the linearly separable case ([60]). From this figure we can see that the goal is to maximise the margin $\frac{2}{\|\boldsymbol{\beta}\|}$ subject to data points on the same side of the hyperplane having the same label. One popular

Background Methods

objective is to minimise the squared 2-norm of β ([37] and [74]). Therefore, the objective function is

$$\begin{aligned} & \min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 \\ \text{s.t. } & y_i(\mathbf{x}^\top \beta + \beta_0) \geq 1, \quad i = 1, \dots, n. \end{aligned} \quad (2.73)$$

In the linearly non-separable case, SVM maximises the margin and at the same time minimises the total distance of the data points fall to the wrong side of their margin. Note that misclassification only occurs when the distance to the margin is larger than 1. Thus, the optimisation problem becomes

$$\begin{aligned} & \min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t. } & \xi_i \geq 0, \quad y_i(\mathbf{x}^\top \beta + \beta_0) \geq 1 - \xi_i \quad \forall i, \end{aligned} \quad (2.74)$$

where C is a constant constraining the total number of misclassified points. (2.74) can be solved with the help of Lagrange multiplier and becomes equivalent to

$$L(\beta, \beta_0, \xi, \alpha, \mu) = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [y_i(\mathbf{x}^\top \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^n \mu_i \xi_i \quad (2.75)$$

Taking the derivatives of L with respect to β , β_0 , and ξ_i and plugging them into Eqn. (2.75), we get the dual representation

$$L_D = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} y_i y_{i'} \mathbf{x}_i^\top \mathbf{x}_{i'}. \quad (2.76)$$

(2.75) then can be solved by

$$\hat{\beta} = \sum_{i=1}^n \hat{\alpha}_i y_i \mathbf{x}_i, \quad (2.77)$$

where $\alpha_i \geq 0$. Finally the solution hyperplane (2.72) is given by

$$f(\mathbf{x}) = \sum_{i=1}^n \hat{\alpha}_i y_i \mathbf{x}_i^\top \mathbf{x} + \hat{\beta}_0, \quad (2.78)$$

and the predicted class using this solution is $\text{sign}(f(\mathbf{x}))$.

Fitting a linear hyperplane to separate linearly non-separable data will always

have misclassified data points, and generally does not achieve as high performance as using a non-linear separation boundary. The non-linear extension of SVM is accomplished by applying kernel functions. A kernel function maps the original data space to a higher (generally speaking) dimensional feature space so that the problem can be solved linearly in the kernel feature space. The kernel function is given by

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle, \quad (2.79)$$

where $\langle \cdot \rangle$ computes the inner product in the feature space, and function ϕ does not need to be specified since only the inner product is required. The function k should be a symmetric and positive (semi-) definite.

Three widely applied kernel functions for SVM are:

$$\text{Polynomial kernel: } k(\mathbf{x}, \mathbf{x}') = (1 + \langle \mathbf{x}, \mathbf{x}' \rangle)^d, \quad (2.80)$$

$$\text{Radial basis kernel: } k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2), \quad (2.81)$$

$$\text{Sigmoid kernel: } k(\mathbf{x}, \mathbf{x}') = \tanh(\alpha \langle \mathbf{x}, \mathbf{x}' \rangle + \beta) \quad (2.82)$$

Replacing the dot product in (2.76) with the kernel function $k(\mathbf{x}, \mathbf{x}')$, the non-linear problem can be solved by replacing \mathbf{x} in (2.78) with $k(\mathbf{x}, \mathbf{x}')$:

$$f(\mathbf{x}) = \sum_{i=1}^n \hat{\alpha}_i y_i k(\mathbf{x}, \mathbf{x}') + \beta_0. \quad (2.83)$$

SVM can also be generalised to solve regression problems. Recall from Section 2.10.1 that the goal of linear regression models is to minimise the differences between actual data value and the predicted value from the model based on some loss function. Now, we consider a regularised objective function of (2.67), and replace the residual least squares with a different loss function $E(\cdot)$, using y_i as the actual data value and $f(x_i)$ as the predicted value from the model. The objective function becomes

$$\min_{\boldsymbol{\beta}, \beta_0} \sum_{i=1}^n E(y_i - f(x_i))^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2, \quad (2.84)$$

where $E(\cdot)$ is generally taken as the ϵ -insensitive error function $E_\epsilon(\cdot)$ ([34])

$$E_\epsilon(r) = \begin{cases} 0 & \text{if } |r| < \epsilon, \\ |r| - \epsilon & \text{otherwise,} \end{cases}$$

which allows the loss be 0 if the predicted value is within the ‘ ϵ -tube’ with the true value. This form of the support vector regression (SVR) is generally known as the

ϵ -SVR. Similar to SVM, ϵ -SVR tries to optimise the same objective function as in (2.74) with updated constraints:

$$\begin{aligned} \min_{\beta, \beta_0} \quad & \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & \xi_i \geq 0, |y_i - \mathbf{x}_i^\top \beta + \beta_0| \leq \epsilon - \xi_i \quad \forall i. \end{aligned} \quad (2.85)$$

Similarly to the classification case, the problem can be solved with the help of Lagrange multipliers. The solution (predicted value) generalised with kernel function has form

$$f(\mathbf{x}) = \sum_{i=1}^n (\alpha_i - \hat{\alpha}_i) k(\mathbf{x}, \mathbf{x}_i) + \beta_0, \quad (2.86)$$

where α_i and $\hat{\alpha}_i$ are the Lagrange multipliers satisfying non-negativity (detailed solution can be found in [132], [20] and [60]).

ϵ -SVR does not penalise errors smaller than ϵ . There is a widely applied variation of ϵ -SVR that accommodates this problem, called ν -SVR ([108]). ν -SVR adds an additional constant to (2.85):

$$\begin{aligned} \min_{\beta, \beta_0} \quad & \frac{1}{2} \|\beta\|^2 + C(\nu\epsilon + \sum_{i=1}^n \xi_i) \\ \text{s.t.} \quad & |y_i - \mathbf{x}_i^\top \beta + \beta_0| \leq \epsilon - \xi_i \\ & \xi_i \geq 0, \quad \epsilon \geq 0, \quad \forall i. \end{aligned} \quad (2.87)$$

The addition of ν allows one to control the number of support vectors and optimises ϵ in (2.85) automatically. The solution to ν -SVR has the same form with (2.86), and the detailed derivation can be referred to [108].

2.10.3 Making predictions

The predictive models are trained and assessed in a CV framework to avoid overfitting. The most commonly used CV framework is k-fold CV. Taking linear regression as an example, the data is firstly divided into the training and test sets. Within each fold, a linear model is fitted to the training set to obtain the β in Eqn. (2.66). Then this β is applied to the test set using Eqn. (2.66) again to obtain predicted values for the test subjects. The model accuracy can be measured by the difference between the predicted and true values in the test sets. There are several ways to

quantify this ‘difference’, including the mean-squared errors, mean-absolute errors and Pearson’s correlation between true and predicted values.

One final important remark is that, similar to the general CV process, we should keep training and test sets independent. In the context of prediction, model building on the training set should be independent of model testing on the test set. This includes all pre-processing steps on the data such as de-confounding and normalisation, as well as feature selections. In other words, we should always split data first and then de-confound, normalise and select features independently within the training and test sets.

2.11 Conclusion

In this chapter, we have introduced the background methods that are used in this thesis: latent variable models, predictive models and some data pre-processing techniques. Many of them are well studied and established methods and can be found with more details and depth in other literature. [60], [20], [124] and [78] are the main textbooks used to gather the information. The topics covered in this chapter are too broad for us to give a thorough overview, therefore, we focused on the aspects that caused confusion during our analysis, and clarified the parts where we struggled to find clear definitions in references.

CHAPTER 3

Supervised Dimension Reduction

In this chapter, we present a novel dimension reduction method refined from PCA which aims to improve the interpretability of latent variable models, namely supervised dimension reduction (SDR). SDR estimates the dimension of the data automatically from a prediction perspective. It is a non-parametric method and supervises the dimension reduction by imposing prior knowledge of the data from human researchers. Moreover, there will also be introduction on the setup for SDR, in particular sign-slipping, which is an important component for embedding SDR into the analysis pipeline of the latent variable models, and help to improve the interpretability of them.

3.1 Sign Flipping

This step is specifically targeted for behavioural data. Behavioural data is commonly collected via questionnaires where the answer to a question is often quantified later on. The way of recording data can be inconsistent and question dependent. For example, if we consider two alcohol drinking measurements, one is the ‘weekly frequency of drinking’ quantified from 1 to 5 meaning from ‘very frequent’ to ‘do not drink at all’; the other variable is ‘weekly intake of alcohol by units’, 0 means ‘do not drink at all’ and the positive integers correspond to the units drank. For the variable ‘weekly frequency of drinking’, higher value means lighter drinker, whereas for ‘weekly intake of alcohol by units’, it is the opposite way round. Therefore, two variables measuring similar behaviour can end up having opposite result from latent variable models due to being oppositely recorded. Especially for CCA, one of the main outcomes is the canonical loadings which is the correlation between canonical variable and the observed data. If two observed variables are oppositely recorded,

they are going to have opposite signs of the loading, and mislead the interpretation.

To facilitate such issue, we flip the signs of some behavioural measures to provide a consistent meaning: specifically so that more positive values corresponded to better life outcomes. This is implemented by first selecting a benchmark variable, e.g. ‘income’, and then flipping variables that have negative correlations with it. Then we carry out a sanity check, examining the pairwise correlations and going through all variable meanings to correct for missed/mis-flipped ones.

3.1.1 Effects on PCA/SVD

Suppose the data matrix is X , flipping the sign of variables (columns) in X is equivalent to right-multiply a diagonal matrix D , where the diagonal entries are either 1 or -1 (-1 corresponds to the columns being flipped). D is orthogonal and symmetric, i.e. $DD^\top = I$ and $D = D^\top$. Flipping the columns of X is equivalent to right-multiply D both sides of Eqn. (2.5)

$$XD = U\Sigma V^\top D = U\Sigma(DV)^\top. \quad (3.1)$$

Therefore, U and PCs remain the same, but V has the respective rows flipped in signs. As introduced in Section 2.6, SVD has the inherent indeterminacy issue, i.e. the column signs of U and V in Eqn. (3.1) can be randomly flipped without changing the reconstruction of X . Taking this into consideration and let the inherent sign flipping transformation matrix be D_0 . Using Eqn. (2.39), Eqn. (3.1) becomes

$$XD = UD_0\Sigma(VD_0)^\top D = UD_0\Sigma(DVD_0)^\top. \quad (3.2)$$

This shows that under column sign-flipping of X , the left eigenvector matrix U (or the principal component $U\Sigma$) is only subject to inherent column sign flipping D_0 and not affected by D , and the right eigenvector matrix V is affected by both D and D_0 , with column sign-flipped by D_0 and row sign-flipped by D .

3.1.2 Effects on covariance matrix and eigen decomposition

After sign flipping the covariance matrix in Eqn. (2.4) becomes

$$(XD)^\top(XD) = DX^\top XD = DV\Lambda(DV)^\top \quad (3.3)$$

Therefore, the magnitudes of this covariance matrix stay the same, but have the corresponding rows and columns sign-flipped simultaneously, i.e. having predictable

signs flipped at certain entries. Again, V still has the sign invariance issue.

The covariance matrix in the subject domain has form XX^\top (ignoring the scalar). After sign flipping, it becomes

$$(XD)(XD)^\top = XDD^\top X^\top = XX^\top = U\Lambda U^\top = U^{PC}(U^{PC})^\top. \quad (3.4)$$

The covariance matrix in the subject domain is not subject to any changes after column sign-flipping.

3.1.3 Effects on CCA

Suppose the inputs of CCA are X_1 and X_2 . Flipping column signs of X_1 and/or X_2 does not change the canonical variables P and Q in (2.16), due to the objective of CCA in (2.15). The sign-flipping is reflected in canonical weights A and B

$$\begin{aligned} P &= X_1 D(DA) \\ Q &= X_2 D(DB). \end{aligned} \quad (3.5)$$

As we can see from (3.5), the canonical weights will have the respective rows sign-flipped.

For the canonical loadings which are given by the correlations between canonical variable and the observed data $\text{corr}(X_1, P)$ and $\text{corr}(X_2, Q)$, flipping column signs of X_1 and/or X_2 , and the respective loadings for X_1 and/or X_2 are sign-flipped.

3.2 Supervised Dimension Reduction

Supervised Dimension Reduction (SDR) is a refined application of principal component analysis (PCA). Instead of reducing the dimension on the whole variable space, the data is grouped into sub-domains based on prior knowledge of the data. In general, behavioural data is grouped by variable functions, e.g. cognition related, alcohol related and demographics. Brain imaging like functional MRI can be grouped by brain parcellations or the brain atlases being used; other brain imaging derived measures can be grouped by function or imaging purpose such as diffusion measures and volume related measures.

PCA is then applied to each sub-domain in turn. The PCs from the sub-domain analysis are concatenated to form the reduced data space. To further improve the interpretability, factor rotation can be applied to the principal loadings in all sub-domains. The dimensionality of the sub-domains are automatically esti-

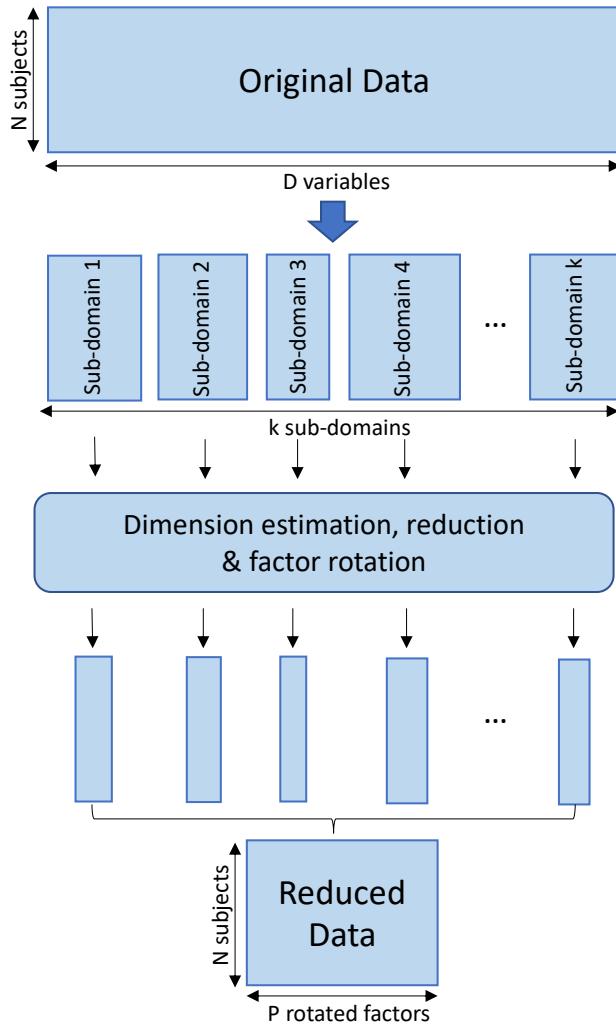


Figure 3.1: SDR overview

mated by minimising the Predicted Residual Error Sum of Squares (PRESS) using a two-way CV method (Fig. 3.2; [24], [7])). The method overview is shown in Fig. 3.1.

3.2.0.1 Two-way cross-validation

The two ways of CV are variable-wise and subject-wise. The subject-wise CV is a K -fold CV. Taking 5-fold as an example, within each of the 5 folds, a leave-one-out variable-wise CV is implemented, i.e. in each fold, one variable is left out at a time and predicted by different numbers of PCs extracted from the rest of the variables.

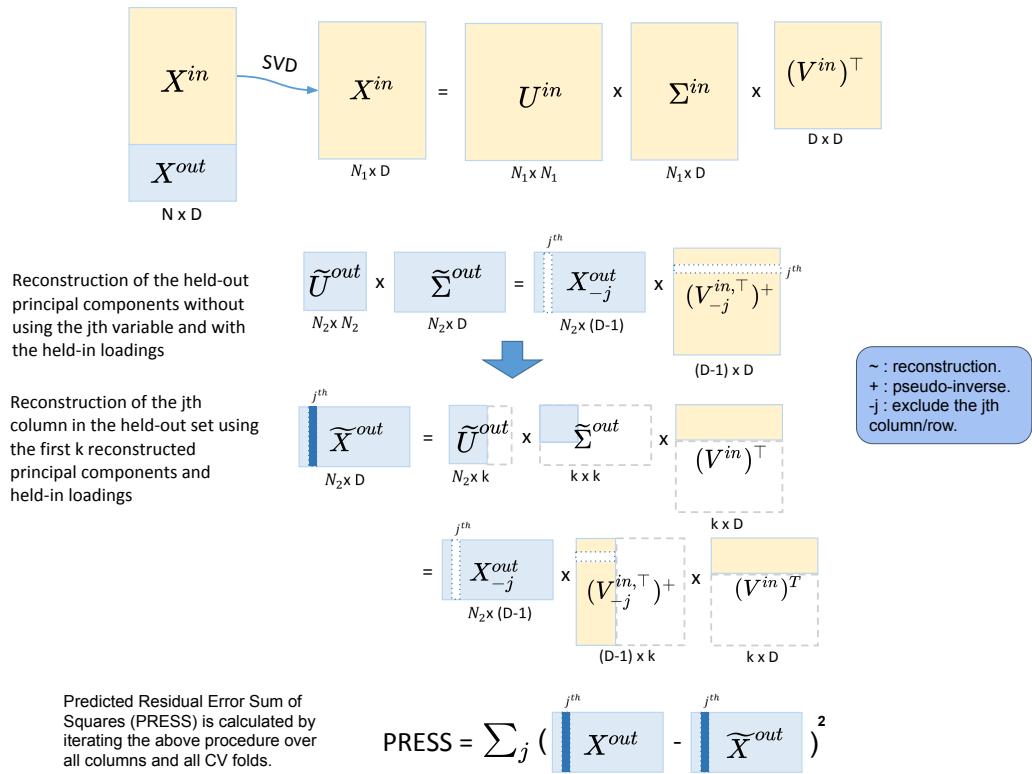


Figure 3.2: Illustration of the 5-fold two-way cross-validation. It minimises PRESS and estimates the dimensionality in an automated fashion. Yellow blocks represent the training data and light blue blocks represent the test data. Two-way CV includes a subject-way (CV over subject direction) and a variable-way (CV over variable direction). Prediction error is calculated by the reconstruction error using different numbers of principal components.

The dimension of the latent space is estimated by calculating PRESS for predictions made by a range of number of PCs (predicted by 1, 2, 3... PC(s), and more details will be introduced below). The number with the lowest PRESS is selected as the optimal dimension.

For example, for the 5-fold subject-way CV, we split the data into the held-in set (4/5 of the cohort), X^{in} and the held-out set (1/5 of the cohort), X^{out} . Let P be the number of total variables in the sub-domain. Then we apply SVD to X^{in} , i.e.

$$X^{in} = U^{in}\Sigma^{in}(V^{in})^\top, \quad (3.6)$$

where U^{in} is the left-eigenvector matrix of X^{in} , eigenvectors of the subject covariance ($\frac{1}{P}X^{in}(X^{in})^\top$); V^{in} is the right-eigenvector matrix, eigenvectors of the variable covariance ($\frac{1}{N}(X^{in})^\top X^{in}$); Σ^{in} is the singular value matrix. The PCs of X^{in} are the singular-value-scaled left-eigenvectors, which we denote $U_{PC}^{in} = U^{in}\Sigma^{in}$, an Eqn. (3.6) becomes

$$X^{in} = U_{PC}^{in}(V^{in})^\top. \quad (3.7)$$

Noting that $U_{PC}^{in} = X^{in}V^{in}$ are the observations in the PC space, in order to reduce the dimensionality in the PC space to k where $k < P$, we can apply the following transformation

$$U_{k,PC}^{in} = X^{in}V_k^{in} \quad (3.8)$$

where $U_{k,PC}^{in}$ and V_k^{in} are the first k columns in U_{PC}^{in} and V^{in} respectively.

Then we can likewise rearrange the held-out data to reconstruct the first k held-out PCs with the k-dimensional held-in principal loadings:

$$\tilde{U}_{k,PC}^{out} = X^{out}V_k^{in}. \quad (3.9)$$

Thus, a lower dimensional reconstruction of the held-out data is

$$\tilde{X}^{out} = \tilde{U}_{k,PC}^{out}(V_k^{in})^\top = X^{out}V_k^{in}(V_k^{in})^\top. \quad (3.10)$$

We can now calculate the prediction error as the difference between X^{out} and \tilde{X}^{out} . Iterating this algorithm over all 5 folds gives the subject-wise action of our

two-way CV method, and we get a PRESS of

$$\sum_{i=1}^N \|\mathbf{x}_i^{out} - \tilde{\mathbf{x}}_i^{out}\|^2 = \sum_{i=1}^N \|\mathbf{x}_i^{out} - \mathbf{x}_i^{out} V^{in} (V^{in})^\top\|^2, \quad (3.11)$$

where \mathbf{x}_i is a row vector which represents the i th subject.

However, the PRESS in Eqn. (3.11) monotonically decreases as k (the number of PC) increases and so is not suitable for dimensionality estimation. This is because the reconstruction of X^{out} in Eqn. (3.11) uses X^{out} itself. To address this we modify the reconstruction of \tilde{X}^{out} in Eqn. (3.10), predicting the j th column of X^{out} using the rest columns in X^{out} :

$$\tilde{X}^{out} = X_{-j}^{out} [V_{-j,k}^{in,T}]^+ (V_k^{in})^T, \quad (3.12)$$

where X_{-j}^{out} is X^{out} with the j th column removed; $[V_{-j,k}^{in,T}]^+$ is the pseudo-inverse of the transpose of $V_{-j,k}^{in}$, where $V_{-j,k}^{in}$ takes the first k columns of V^{in} and then removes the j th row. The pseudo-inverse is required since removing a row of V^{in} breaks its orthogonality. The j th column in \tilde{X}^{out} is now reconstructed without using the j th column in X^{out} , and we denote this column as $\tilde{\mathbf{x}}_j^{out}$. If we iterate j from 1 to P , we reconstruct the whole held-out set in turn. This is the variable-wise action in the two-way CV method. For each of the held-out in a CV fold, the corresponding PRESS can be calculated as:

$$\sum_{j=1}^P \|\mathbf{x}_j^{out} - \tilde{\mathbf{x}}_j^{out}\|, \quad (3.13)$$

where \mathbf{x}_j^{out} is the j th column in X^{out} , and $\tilde{\mathbf{x}}_j^{out}$ is as described above. Finally, the total PRESS for all subjects is calculated by summing PRESS in Eqn. (3.13) over all CV folds, completing the subject-wise action of the method.

Finding the dimensionality k with the minimum PRESS over dimensions completes the method for a given sub-domain. The reduced dataset is then obtained by the concatenation of the selected PCs from each of the sub-domains.

3.2.1 Evaluating the stability of SDR

SDR is based on k -fold CV. However, PRESS varies between the different folds. To address this issue, we repeat SDR for N times and take the mode of the estimated dimension for each sub-domain.

To further test the accuracy of the dimension SDR estimates, we can compare them with the eigen-spectrum and null eigen-spectrum on each of the sub-domains. The eigen-spectrum provides information on the variance explained by each of the eigenvectors. The null eigen-spectrum is obtained by shuffling the row values for each column in the original matrix, and calculating its eigen-spectrum. This is essentially permutation testing on eigenvalues (see Section 2.9.2). It shows the amount of ‘background noise’ exist in the dataset. When the null eigen-spectrum exceeds the eigen-spectrum, we can interpret this as the background noise taking over the information. Ideally the estimated dimension from SDR falls near where the null eigen-spectrum crosses the eigen-spectrum.

3.3 Conclusion

In PCA, the PCs are linear combinations of all variables. It makes the interpretation difficult, especially when the number of variables is high. To alleviate this problem, we proposed a new approach of grouping variables by functions or criteria that is easily understood by humans. The next challenge was to represent each sub-domains reasonably. Since different sub-domains consist of different numbers of variables, a unified standard is needed to justify the dimensions in sub-domains. The two-way CV method serves this purpose from a prediction point of view, which also fits in the non-parametric framework CCA is based on. To further improve the stability of the loadings in the intermediate steps, factor rotation can be applied optionally. The latent factors in SDR can be tracked back to the sub-domains which were previously grouped by some informative measure. Therefore, using SDR factors for further analysis like CCA and GFA will also improve the interpretability for their outcomes. In conclusion, there are two major advantages of SDR: it incorporates auto-dimension decision method which provides a unified way of choosing a representative dimension for meaningful sub-domains; it improves the interpretability of the results from latent variable models.

CHAPTER 4

Human Connectome Project

4.1 Introduction

In this project, we studied the relationship between functional neuroimaging and behavioural data by making use of the Human Connectome Project (HCP; [131]). This study is based on the published work of [115] which applies CCA to explore such relationships. As mentioned in Chapter 1, PCA components are commonly used as inputs of CCA. However, this can result in CCA results that are difficult to interpret. Moreover, [115] applies PCA to reduce the dimension of the data to 100, which is selected without justification. In this project, we replaced PCA with SDR (introduced in Section 3.2), reducing the dimensionality of the original datasets using prior knowledge of the structure of the variables studied. SDR improved the interpretability of the CCA results as well as offering justified dimensionalities for the reduced datasets. As mentioned in Section 1.3.1, being able to understand the composition of latent factors is of vital importance, especially for health-related datasets. SDR improves the interpretability by making the latent factors more functionally meaningful and stable. Thus, we were able to gain deeper insights on the data structures in the original space as well as latent space.

We considered two data modalities in this study, subject measures (SM) and brain measures (BM). SM consists of behavioural and demographic variables, and BM is the functional connectivity of the subjects as introduced in Section 1.2. SDR firstly grouped SM and BM into sub-domains and reduced dimensionality by the sub-domains with an automatic dimension estimation method. We applied CCA to the SDR reduced SM and BM to study the correlations between brain and behaviour. To further improve the interpretability, factor rotation was applied during SDR.

This analysis pipeline was applied to the HCP data. The performance was assessed by examining canonical correlations, significant canonical variables and canonical loadings (also known as the structural coefficients). SDR offers us insights

on the structure of sub-domains of SM and BM, and more interpretable CCA results. We then carefully validated the stability of SDR and CCA by applying 5-fold CV.

For comparison, and to test the stability of the results, we replicated the analysis pipeline used by [115] (PCA then CCA) on the larger S1200 data, and then compared its results with SDR CCA.

It is worth noting that in this analysis, sign-flipping was applied to the data to maximise the pair-wise correlation between variables (Section 3.1). As discussed in Section 3.1, sign-flipping does not affect the results of CCA.

4.2 Data

Human Connectome Project (HCP) is a health data project launched in 2009 and led by Washington University, University of Minnesota and Oxford University. The data is collected over young (age 22-35) healthy adults in the US, most of which have family structures, being siblings or twins. The goal of HCP is to ‘map the human brain, aim to connect its structure to function and behaviour’ on a large population level ([131]).

There are several data releases at different time points from 2013 to 2017. Three of them are analysed in this project, HCP 500 release, 900 release and 1200 release. The main difference between releases is the number of subjects, with release-specific changes including new behavioural measures, corrections or updates on the previous measures, data pre-process pipeline, and data acquisition updates (see <https://wiki.humanconnectome.org/display/PublicData/HCP+Data+Release+Updates>).

The HCP 500 release is mainly used for the replication of the published results in [115] and carrying out exploratory analysis on the data. HCP 900 release is used to develop SDR and the new analysis pipeline, and the final HCP 1200 release is used for validating the analysis pipeline on a larger dataset. In this chapter, we focus on showing the results from the largest and most recent data release HCP 1200.

The HCP 1200 cohort consists of $N = 1003$ subjects. For the brain measures (BM), we used connectivity matrix (partial correlation) generated from resting-state fMRI data. Details about data acquisition can be found at the HCP database website (<http://humanconnectome.org/data>) and in [114] and [131]. It is worth mentioning that the resting-fMRI was collected over four 15-minute scans. For non-imaging data, we considered 234 (after QC) behavioural and demographic measures, and denoted them as subject measures (SM).

4.2.1 Data pre-processing

BM was pre-processed in the same way from resting-state fMRI (rfMRI) as described in [115]. In brief, Group-ICA (Independent Component Analysis) was performed to parcellate the brain into 200 independent nodes (regions). This number was arbitrarily selected as stated in [115]. However, [115] also commented that this number of parcellation did not affect the results significantly. Therefore, we used the 200 ICA parcellation for better comparison. Each subject's rfMRI data was then mapped with this ICA brain map to obtain one time series per ICA region. A functional connectivity matrix for each subject was generated by calculating the Tikhonov-regularized (a.k.a ridge regression; [126]) partial correlation for every pair of the time series. This resulted in a 200×200 connectivity matrix or 19900 ($200 \times (200 - 1)/2$) brain measures for every subject, and each of the entries represents a connectivity edge between two ICA regions.

For SM, most of the variables were neatly collected and properly quantified with few categorical nominal variables existed in the data. We turned the nominal ones into dummy variables (as discussed in Section 2.2.4). Then we applied sign-flipping prior to other pre-processing steps to align the variables so that higher value corresponds to better life outcomes. We removed ill-conditioned SMs according to three criteria: if they had more than 50% missing values; if the standard deviation was 0; if more than 95% of the total entries were identical values. This left us with 234 SM variables. (see Appendix A.2 for a full list of SMs). Then the missing data was imputed using soft-impute (Section 2.8.3).

Both datasets were normalised by rank-based inverse normal transformation (see Section 2.8.2) and then de-confounded (see Section 2.8.4). Fifteen confounding variables were carefully chosen as they could potentially affect the relationship between brain and behaviour, including age, gender, height, weight, rfMRI movement etc. We also de-confounded the squared values for some of these variables like age, BMI etc. (see Appendix A.1 for the full list of confounds).

Grouping of SM and BM into sub-domains

The pre-processed 234 SMs were grouped into 14 sub-domains based on their functions: Alcohol use, Alertness, Psychiatric history, Tobacco use, Drug use, Emotion, Cognition, Family history, Physical health, Motor, Personality, Sensory, Feminine health and Demographics (including SES). This grouping followed the official HCP variable dictionary (<https://wiki.humanconnectome.org/display/PublicData/HCP+Data+Dictionary+Public+-+Updated+for+the+1200+Subject+Release>). BMs

were grouped based on the 200 different ICA regions. Thus, there are 200 BM sub-domains, each with 200 brain edges.

4.3 Analysis Pipeline

In this work, we applied CCA to the SDR reduced SM and BM to explore the relationship. In general, applying dimension reduction before CCA is not necessary, however, to reduce the impact of noise and to avoid a degenerate solution when the number of subjects is less than the number of variables, a dimension reduction is often applied to each dataset and the reduced data are fed into CCA. In this study we chose to apply SDR (introduced in Section 3.2) to achieve the dimension reduction so that the functional interpretation of CCA results can be better understood.

After applying SDR, we found that BM was reduced to a much higher dimension than the number of subjects. This is a result of the large dimension of the original BM space (19900). Due to the limitation of the number of subjects in this study, we applied PCA to SDR reduced BM to further reduce its dimension to avoid the degeneration problem in CCA. We chose to reduce the dimension of BM to 100 to match the method in [115], and to the same dimension as the SDR reduced SM for a more fair comparison. To further improve the interpretation of these PCs, we applied Varimax factor rotation (Section 2.7) to the principal loadings in the sub-domains, and then use the rotated components (RCs) as the inputs of CCA. Notably, orthogonal rotation is an invariant transformation on CCA inputs (Section 2.7), therefore, does not affect CCA outputs.

We examined the number of significant pairs of canonical variables using permutation testing for 10,000 permutations. Since the data has family structures, only families (not individual subjects) were permuted. The variable importance was evaluated by two different measures, canonical loadings (structural coefficients) for observed variables and canonical loadings for SDR factors (CCA inputs) (Section 2.3). We also calculated variance explained by each of the significant canonical variables in the original datasets of SM and BM. This is achieved by taking the average of R-squared values over all subjects (see Section 2.3 for more details). Method overview is illustrated in Fig. 4.1.

In the end we carried out a stability study on this pipeline by using 5-fold CV (Fig. 2.3). In CV, we do not break the family structures, i.e. subjects from the same family would either all go into the training set or the test set.

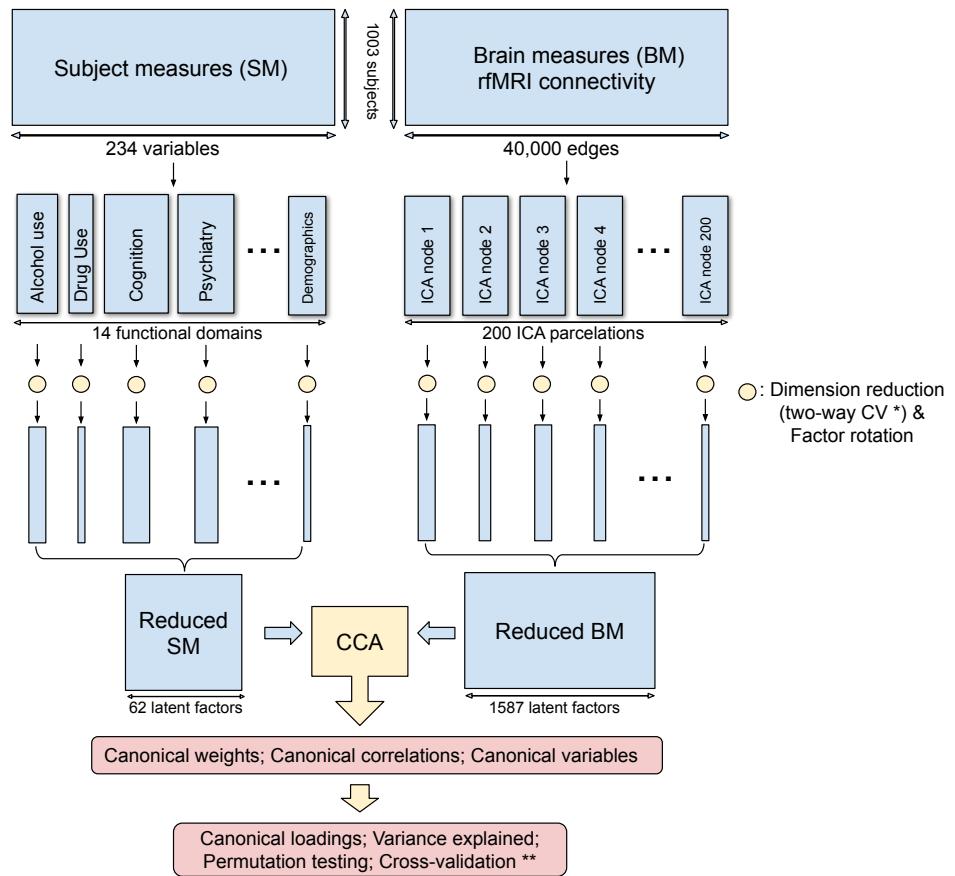


Figure 4.1: Method overview of SDR CCA. SM and BM are first grouped into sub-domains. PCA is applied to each sub-domain while a two-way CV method (* see Fig. 3.2) is used to estimate the dimension. Then the rotated principal components from all sub-domains are concatenated to form the reduced SM and BM. Finally the reduced SM and BM are fed into CCA and for further CV (** see Fig. 2.3) and permutation testing.

4.3.1 Comparison between PCA and SDR

To assess the performance of SDR in comparison with PCA, we applied the same analysis pipeline using PCA (instead of SDR) on the same datasets with sign-flipping on SM. [115] applied PCA to reduce the dimensions of SM and BM to both 100. However, our SDR method automatically reduces the dimension of SM to under 100, and BM to over 100. Trying to make the dimensions consistent, we apply PCA to match the dimension of SDR reduced SM dataset. For BM, we further applied PCA after SDR to reduce BM dataset to 100, and this was to avoid the degenerated solution from CCA.

We compared the variance explained by the latent components obtained by PCA and SDR respectively in the original SM and BM spaces. We also examined the canonical variables, canonical correlations and canonical loadings output by using datasets reduced by PCA and SDR respectively, as well as the variance explained by canonical variables in the original SM and BM datasets. In the end, we applied the same CV to compare the stability on PCA and SDR in preserving canonical correlations, number of significant canonical variables etc.

4.4 Results

4.4.1 Sign alignment

Fig. 4.2 shows the pairwise correlations on SM before (left subplot) and after (right subplot) sign-flipping. Compared with the un-flipped correlation matrix, the flipped correlation matrix appears to be more aligned, especially in the sub-domain of ‘Alcohol Use’ and ‘Emotion’. Moreover, there is noticeable better alignment between the ‘Psychiatry’ and ‘Emotion’ sub-domains. Notably, if most/all the variables in a sub-domain get flipped, the pairwise correlations would hardly be changed. This is the case in the ‘Psychiatry’ sub-domain where most of the variables are sign-flipped since normally a higher psychiatric score means a more severe mental symptom (see Appendix A.2 for a full list of flipped variables).

4.4.2 PCA CCA results

We applied the method used in [115] to the HCP 1200 dataset as benchmark.

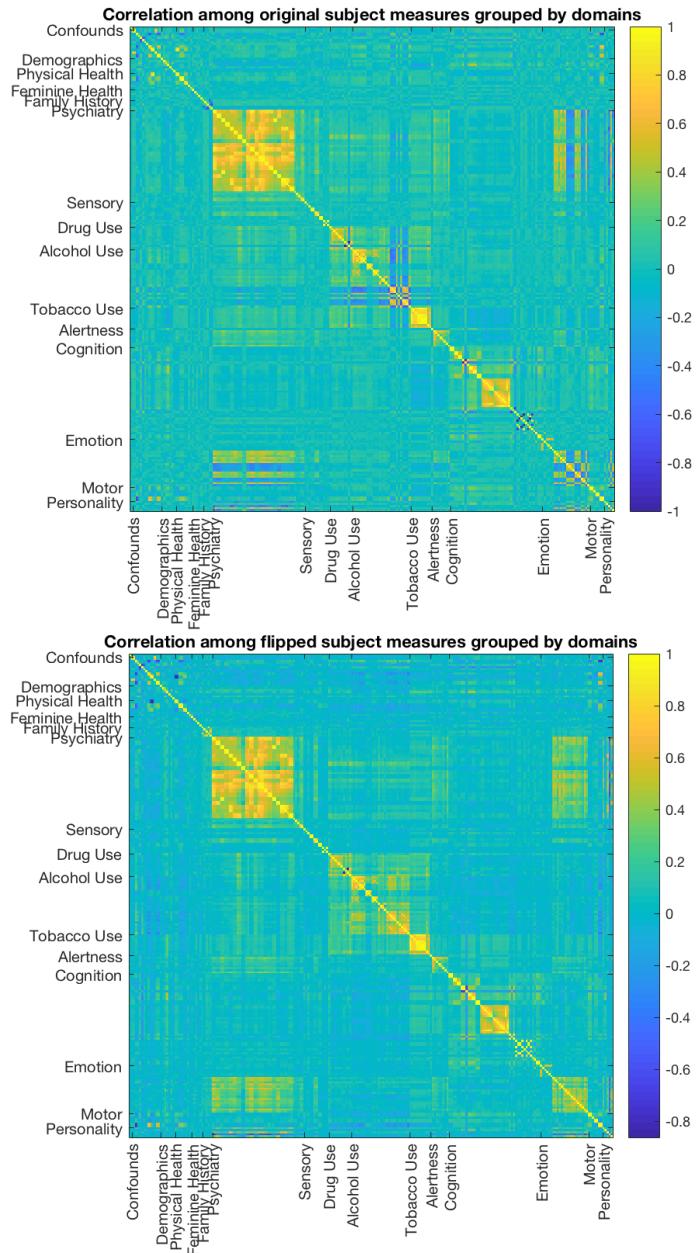


Figure 4.2: Pairwise correlation among 234 subject measures grouped by 14 functional domains. On the top is the correlation matrix among original variables; on the bottom is the correlation matrix after sign-flipping which aims to align pairwise correlations between and within domains.

Tab. 4.1 shows the variance explained by significant canonical variables of SM and BM, canonical correlations and number of significant CCA modes given by permutation testing at 5 sets of different input dimensions. Interestingly, if we keep

Table 4.1: Summary table for PCA CCA with 5 different input dimensions of CCA (first column). Second and third columns show the variance explained (VE) by the SM and BM canonical variables in the observed SM and BM sets for significant canonical pairs respectively; the fourth column shows the canonical correlation for the canonical pairs; the last column shows the number of significant canonical pairs obtained by permutation testing.

Input Dimensions	VE (%) by SM Canonical Variable				VE (%) by BM Canonical Variable				Canonical Correlation				# of Sig. Mode	
	SM	BM	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
62	100	3.586	3.817	1.622	0.920	0.205	0.203	0.229	0.199	0.674	0.637	0.604	0.588	4
62	62	3.296	2.333	1.501	1.207	0.233	0.208	0.266	0.298	0.625	0.543	0.531	0.511	4
30	100	3.943	4.358	2.216	1.482	0.207	0.191	0.221	0.214	0.649	0.603	0.548	0.514	4
30	62	3.663	3.534	1.840	1.754	0.234	0.231	0.280	0.278	0.596	0.505	0.458	0.446	4
30	30	2.869	2.384			0.308	0.370			0.466	0.387			2

SM or BM fixed, and reduce the dimension of the other, the canonical variables would explain more variance in the dimension fixed observed dataset. However the strength of canonical correlation would decrease as the dimensions decrease (observe the first two or last three rows in Tab. 4.1).

For the rest of the study, we are going to focus on the 62 dimensional SM and 100 dimensional BM since 62 is the SDR estimated dimension for SM and 100 was selected in previous studies ([115]). There are 4 significant canonical pairs in this setting. The canonical loadings of SM (Fig. 4.3) for these 4 canonical variables display 4 behavioural/demographic modes. The first set is mainly loaded on cognition variables; the second set is dominated by tobacco variables; the most of the top loadings in the third set are alcohol variables; the fourth set is more of mixture with cognition, emotion and motor.

Note that most of the SM variable loadings shown in Fig. 4.3 have the same sign, and this is in contrast to previous CCA results with the HCP data. For example, [115] found a mode with tobacco use and education measures having opposing signs, while here, after flipping the signs of the observed variables, they are now on the same side of the axis (CCA mode 2 in Fig. 4.3). While the canonical variable found by CCA is invariant to sign flips of the variables, the canonical loadings of course reflect any sign flips (the theory behind this can be found in Sections 3.1 and 2.6).

4.4.3 Sub-domain analysis and SDR results

We grouped the 234 SM variables into 14 sub-domains based on their functions provided by HCP (all 14 sub-domains are listed later in Tab. 4.2). For each of

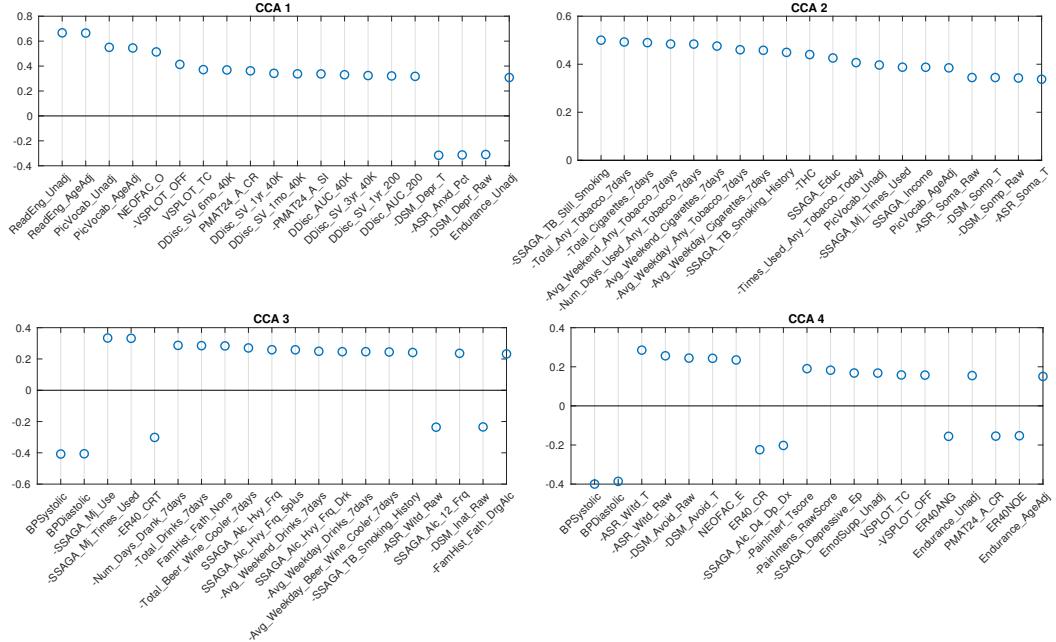


Figure 4.3: Top 20 canonical loadings for 4 significant SM canonical variables in PCA CCA method using 62 dimensional SM and 100 dimensional BM. Variable names with a ‘-’ sign means the values have been flipped.

the sub-domains, we generated a summary report which helps us understand the structure of each sub-domain.

Two of the panels in the Family History report are shown in Fig. 4.4 to show as an example. Top subplot in Fig. 4.4 shows the rotated principal loadings, i.e. the variable importance on generating the latent factors for the sub-domain. We observed higher interpretability on the rotated loadings, latent variables loaded on fewer observed variables. Therefore, we used the rotated loadings to summarise the meaning of latent factors in the sub-domain, as shown in Tab. 4.2. The dimension of the sub-domains is decided by the minima of the red line in the bottom subplot in Fig. 4.4, which is calculated by Eqn. (3.13). These SDR estimations also generally corresponded to where the actual and null eigenspectrum cross (Panel A of the figures in Appendix A.3).

Moreover, by investigating the sub-domain structures, we observed strong stability of SDR factors and understood better the composition of the latent factors in each sub-domain.

In total, SDR selected 62 factors from 14 sub-domains. The selection criterion is not based on a single cut-off point of the variance explained, but from

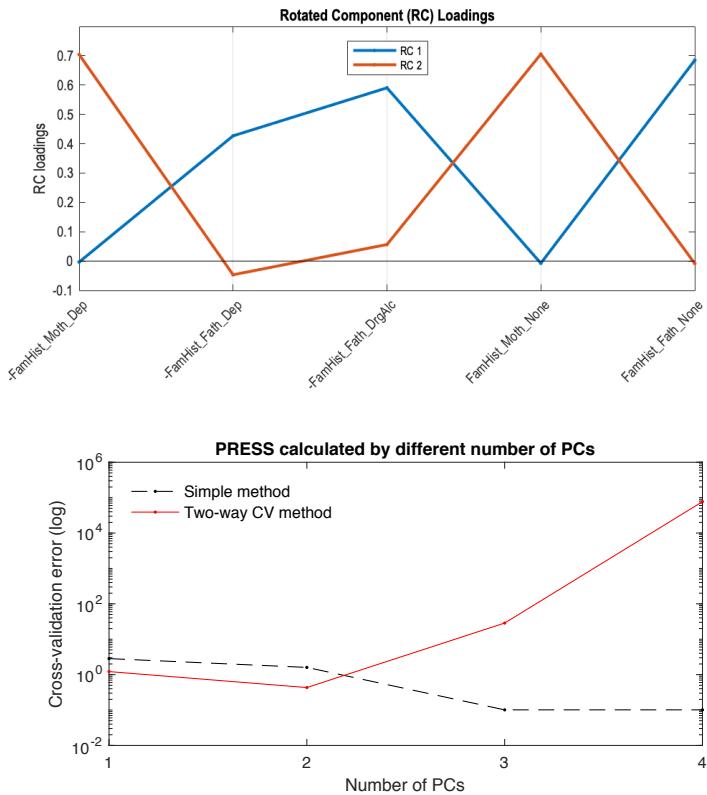


Figure 4.4: Top figure shows the rotated principal loadings; figure at the bottom shows the error curves calculated by Eqn. (3.11) (dotted line) and Eqn. (3.13) (red line), with the minimal error circled at the second component. The naive way of calculating PRESS (dotted line) is monotonically decreasing, while the two-way CV method (red line) offers a minimum point.

Table 4.2: Summary of SM sub-domains. The factors are orthogonally rotated principal components and ordered by R-squared values in the original sub-domain. The second column shows the factor names summarised from panel E in each of the sub-domain report like the top figure in Fig. 4.4. The third column shows the variance explained by the dimension reduced sub-domain in the original sub-domain. The numbers in the brackets are the two-way CV estimated dimension verses the total number of variables in the sub-domain. The Personality is represented by the Big Five personality traits: Neuroticism (N), Agreeableness (A), Extraversion (E), Conscientiousness (C), Openness to experience (O).

Sub-domain Factors	Factor Summary	Variance explained & SDR estimation
Demographics	SES	23.85% (1/7)
Physical Health 1 Physical Health 2 Physical Health 3	Hematocrit Blood pressure BMI (-)	62.53% (3/8)
Feminine Health 1 Feminine Health 2 Feminine Health 3 Feminine Health 4 Feminine Health 5	Regular cycle Days since last cycle Cycle length Age began menstruation Using birth control	100% (5/5)
Family History 1 Family History 2	History of mental health disorder - Father (-) History of mental health disorder - Mother (-)	77.05% (2/5)
Psychiatry 1 Psychiatry 2 Psychiatry 3 Psychiatry 4 Psychiatry 5 Psychiatry 6 Psychiatry 7 Psychiatry 8 Psychiatry 9 Psychiatry 10	Anxiety (-) Attention deficit (-) Thought problems (-) Aggressive behaviour (-) Anti-social behaviour (-) Withdrawn/avoidant behaviour (-) Somatic (-) Intrusive behaviour (-) Depression (-) Panic/phobia (-)	86.88% (10/44)
Sensory 1 Sensory 2 Sensory 3 Sensory 4 Sensory 5 Sensory 6 Sensory 7	Visual acuity (number of errors) Visual and auditory acuity (-) Taste intensity (-) Olfactory ability Subjective pain experience (-) Eyesight Visual acuity and Audition	90.56% (7/12)
Drug Use	All drug use (-)	46.14% (1/11)
Alcohol Use 1 Alcohol Use 2 Alcohol Use 3 Alcohol Use 4 Alcohol Use 5	Alcohol abuse and dependence Heavy alcohol consumption Alcohol consumption (-) Hard liquor consumption (-) Wine consumption (-)	68.36% (5/28)
Tobacco Use	Smokes tobacco (-)	80.54% (1/10)
Alertness	Sleep quality	35.28% (1/9)
Cognition 1 Cognition 2 Cognition 3 Cognition 4 Cognition 5 Cognition 6 Cognition 7 Cognition 8 Cognition 9 Cognition 10 Cognition 11 Cognition 12 Cognition 13 Cognition 14	Delay discounting (small amount) Delay discounting (large amount) Delay discounting (short term) Language Fluid intelligence Sustained attention (specificity) Sustained attention (sensitivity) Executive function - set shifting Visuospatial processing Executive function - inhibition (Flanker) Working memory Processing speed Visual episodic memory Verbal episodic memory	85.01% (14/44)
Emotion 1 Emotion 2 Emotion 3 Emotion 4 Emotion 5 Emotion 6 Emotion 7 Emotion 8	Social support Negative affect (angry, fearful, sad) Positive affect Aggressive behaviour Emotion recognition (fear and sad) Emotion recognition (anger against fear) Emotion recognition (neutral against sad) Emotion recognition (fast response time)	69.96% (8/23)
Motor 1 Motor 2 Motor 3	Endurance Strength Dexterity	86.84% (3/7)
Personality	N against ACE	39.70% (1/5)

minimising the prediction error point of view. We can see factors in different domains explain different amount of variances (third column, Tab. 4.2). For example, the first PC (out of 10) in Tobacco Use explains 80.54% variance in the whole sub-domain, and the first PC (out of 11) in Drug Use explains less than 50%. However, with only 1 PC in these two sub-domains, they achieved the lowest prediction errors in the test set. For BM, SDR reduced 200 ICA regions/sub-domains from total dimension of 40,000 to 1587. Proportionally, SDR reduced BM to a larger degree compared with SM, therefore, it implies that the noise level in BM is higher than SM.

4.4.4 Results for SDR CCA

We applied CCA to SDR reduced SM (SDR SM) and SDR followed by PCA reduced BM (SDR+PCA BM). Like in PCA CCA, we applied the analysis to different input dimensions of CCA (Tab. 4.3). Due to the nature of SDR, the dimension for SDR SM was fixed at 62. To obtain lower-dimensional SM as input for CCA, we applied PCA after SDR to further reduce the dimension. We found that similarly to PCA CCA, lower CCA input dimension leads to canonical variables explaining more variance, whereas the canonical correlations get weaker. Comparing each setting with results in PCA CCA (Tab. 4.1), we found that the number of significant canonical variables was always one lower.

Table 4.3: Summary table for SDR CCA. The ‘Input Dimensions’ column shows the dimensions of SM and BM as CCA inputs; the ‘VE by SM’ and ‘VE by BM’ parts of the table represent the variance explained (VE) by the SM and BM canonical variables in the observed SM and BM set for significant canonical pairs (pairs of canonical variables) respectively; ‘Canonical Correlation’ part shows the canonical correlation between the canonical pairs; the last part shows the number of significant canonical pairs given by permutation testing.

Input Dimensions		VE by SM			VE by BM			Canonical Correlation			# of Sig. Mode
SM	BM	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	
62	100	2.629	1.727	1.698	0.202	0.183	0.211	0.632	0.582	0.574	3
62	62	2.318	1.386	2.361	0.203	0.302	0.271	0.555	0.519	0.495	3
30	100	3.323	3.928	1.737	0.194	0.181	0.267	0.586	0.541	0.503	3
30	62	2.776	2.867	1.290	0.306	0.260	0.244	0.475	0.445	0.440	3
30	30	2.716			0.376			0.419			1

Canonical loadings for SM

For 62 SM and 100 BM, we found three significant canonical pairs, and their top 20 canonical loadings are shown in Fig. 4.5. Noticeably, all top 20 loadings for these three canonical variables were positive after sign-flipping of the observed variables.

With the help of SDR, we were able to explore the contributions of CCA inputs directly. Apart from the canonical loadings shown in Fig. 4.5, we could also correlate the canonical variables with the CCA inputs directly. The inputs of CCA are the latent factors of SM and BM, and canonical loadings on the inputs are much less interpretable using PCA compared with SDR, since PCs are loaded on too many observed variables. However, with SDR, we are able to interpret not only the latent factors but also the canonical loadings on the inputs (Fig. 4.6). Using the summarised latent factors in Tab. 4.2, we were able to conclude, for example, in the first canonical loadings (the first subplot in Fig. 4.6), Language factor (Cognition 4) has the largest loading. The second and third largest loadings are Cognition 3 and 1, and they are Delay Discounting factors.

The right set of figures in Fig. 4.6 offers us insight in the overall contribution of each sub-domain, by averaging all positive (blue bars) and negative (red bars) loadings within each domain. We notice here the top loadings and overall loadings are not mono-signed anymore even with the sign-flipping in effect. Interestingly, the pattern presented in the first set of overall canonical loadings (top right, Fig. 4.6) is driven by subjects with good cognition and motor ability, who do not smoke, but take drugs, have some kind of mental disorder and drink. The second and third sets are displaying good well-being patterns. In particular, the second set of loadings are dominated by high SES (social-economic status) and no drug use; the third set shows the alignment between no tobacco use and high SES.

We can essentially do the same thing in PCA: multiplying canonical weights with principal loadings and then group the variables into sub-domains. However, this weight backtrack is not stable due to the instability of canonical weights. Therefore, the sub-domain level patterns we observed via SDR cannot be observed in PCA.

Canonical loadings for BM

Each set of canonical loadings for BM is a 200×200 symmetric matrix. Each entry represents a CCA connection (edge) between two ICA regions. We first map this loading matrix with the signs of the group mean correlations between the ICA regions, i.e. if two ICA regions were negatively correlated at resting-state, it would

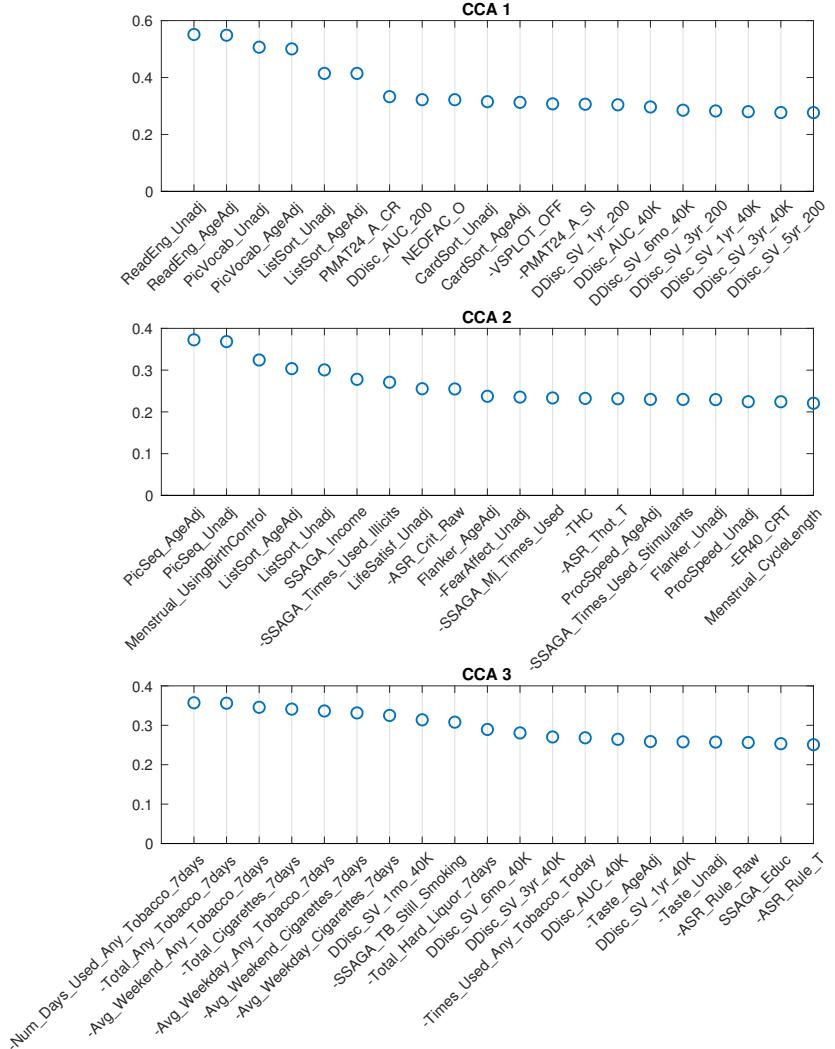


Figure 4.5: Top 20 SM canonical loadings for 3 significant canonical variables. Variable name with '-' sign shows that it was flipped in the original dataset. Canonical loadings of CCA 1 are very similar to the first set of PCA CCA, heavily cognition dominated; the second set is mixed with cognition, drug use etc; The third set is combination of tobacco use and cognition variables. The labels are the exact variables name given on the HCP official websites.

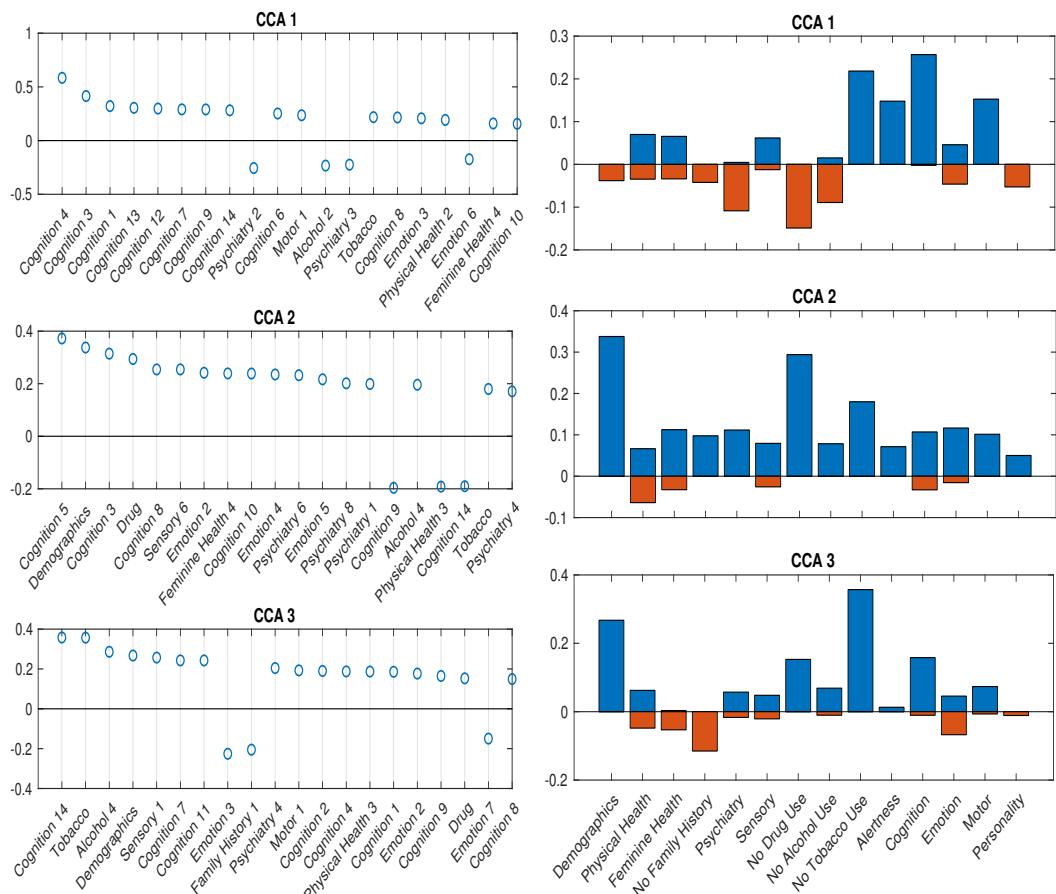


Figure 4.6: SM canonical loadings on the CCA input for the 3 significant canonical variables. The left set of figures shows the top 20 loadings for the 3 significant canonical variables respectively; the right set of figures show the mean of all positive loadings (red bars) and the mean of all negative loadings (blue bars) within each sub-domain for the 3 significant canonical variables.

decrease the positive CCA strength but enhance the negative CCA strength. Due to the difficulty of interpreting each of these 19,900 ($200 * 199/2$) edges, we came up with the following summary statistics. We averaged the top 20 (10%) positive and negative modulated canonical loadings for each ICA region (in each column/row), and denote them as the *positive CCA strength* and *negative CCA strength* respectively. They are shown in Fig. 4.7.

The CCA strengths for CCA 1¹ illustrate a weak contrast between language, sentences, semantic areas (positive strength) against premotor, motor, primary areas (negative strength); positive and negative strengths for CCA 2² are much less distinguishable, both overlapping with parietal and intraparietal which are arguably linked to working memory and default mode network. The positive CCA strength maps for CCA 3³ overlap considerably with CCA 2. The positive map shows weak connection with the default mode network whereas the negative map activates in occipital and pre-motor areas.

Combining results from both SM and BM sides, CCA mode 1 reveals an interesting pattern: language and comprehension related brain areas associate positively with no tobacco use, no psychiatric illnesses, better alertness and cognitive ability, and negatively with drug use. Whereas drug use is positively correlated with the motor areas in the brain.

4.4.5 Stability of SDR CCA

We applied 5-fold CV to SDR CCA with the 62-dimensional SDR SM and the 100-dimensional SDR+PCA BM. Tab. 4.4 shows that in many folds, the first canonical variable does not explain the most variance in the observed datasets. We also ran permutation testing on the training and cross-validated sets for 10000 simulations to get significant canonical pairs. Permutation testing resulted in mostly two significant canonical pairs on the training sets, and ranged from 0 to 4 on the CV sets, with 0 or 1 being the common numbers.

In general, for canonical correlation, as the sample size gets larger, the correlation gets weaker. The first canonical correlation is 0.632 for 1003 subjects, and 0.662 on average for four-fifths of those subjects. This also applies to the number

¹Positive map: <http://neurosynth.org/decode/?neurovault=108956>; negative map: <http://neurosynth.org/decode/?neurovault=108957>

²Positive map: <http://neurosynth.org/decode/?neurovault=108976>; negative map: <http://neurosynth.org/decode/?neurovault=108977>

³Positive map: <http://neurosynth.org/decode/?neurovault=108978>; negative map: <http://neurosynth.org/decode/?neurovault=108979>

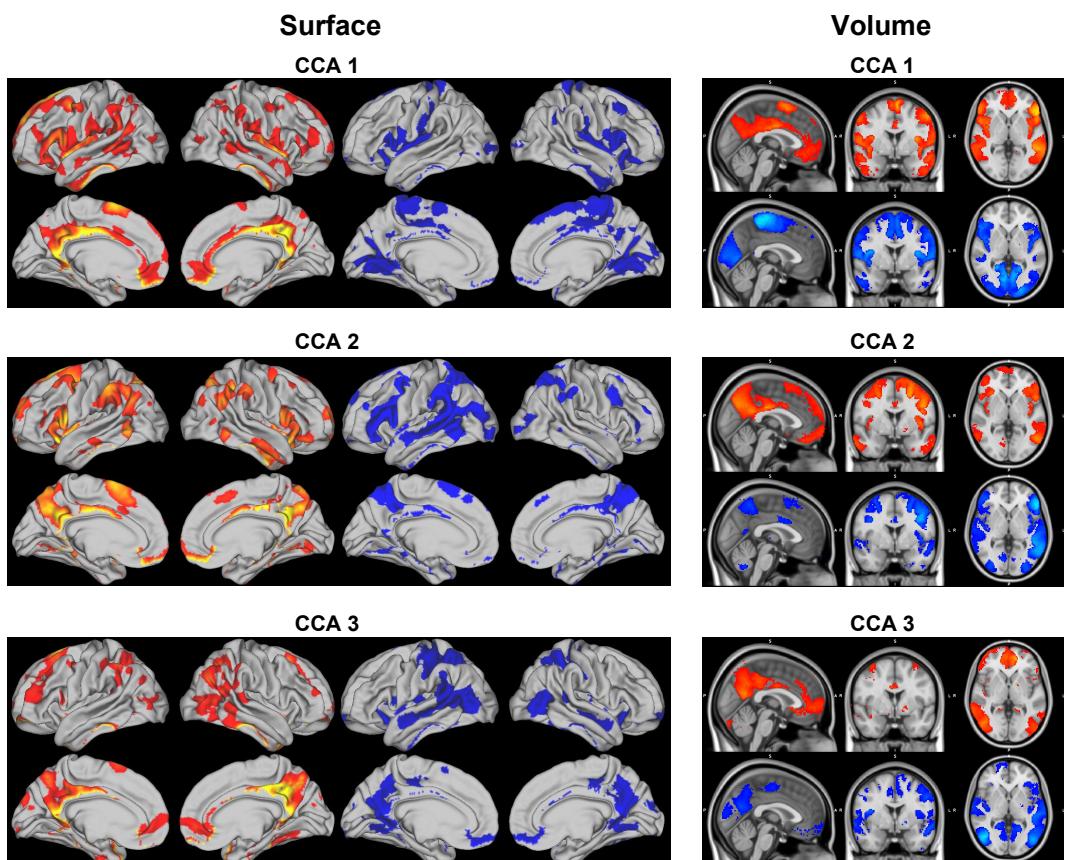


Figure 4.7: Positive and negative CCA strengths on brain surface (left) and volume (right) for the 3 significant canonical variables. The visualisation is cut by 80 percentile. Positive (red maps) and negative (blue maps) CCA strengths are generated by mapping the canonical loading with the sign of population mean correlation between each pair of ICA regions, then average the top 20 positive and negative modulated loadings respectively.

Table 4.4: 5-fold CV on 62 dimensional SM and 100-dimensional BM in SDR CCA analysis. Mean variance explained (MVE) and mean canonical correlations (MCC) are shown for the first 3 pairs of canonical variables with standard deviation (std) in brackets.

	CCA 1	CCA 2	CCA 3
MVE in training SM set (std)	2.34% (0.50)	2.13% (0.85)	1.95% (0.84)
MVE in training BM set (std)	0.23% (0.027)	0.22% (0.028)	0.22% (0.017)
MVE in CV SM set (std)	2.80% (0.73)	2.60% (0.85)	2.37% (0.57)
MVE in CV BM set (std)	0.63% (0.051)	0.65% (0.086)	0.65% (0.059)
MCC for held-in set (std)	0.662 (0.011)	0.639 (0.016)	0.611 (0.0084)
MCC for CV set (std)	0.228 (0.037)	0.108 (0.057)	0.174 (0.039)

of significant CCA pairs permutation test detects. With the HCP 500 release, only one significant pair was detected ([115]); we replicated the study with HCP 900 release and found two pairs; In this study, three pairs were discovered using the whole cohort. However in CV, the training set consists four-fifth of the subjects (the same amount as in the 900 release), and 2 was the most common significant number which is consistent with the previous results. Hence, we are going to look at the results on the two significant pairs on training sets in CV in more detail.

We have selected the top 20 SM canonical loadings in every fold and took the ones that occurred at least two times out of the five CV folds for the two significant canonical variables. The stability of the first (left) and second (right) SM canonical loadings on observed SM data is shown in Appendix A.4. The language variables tend to be the most stable and heavily weighted in the first set, appearing on the top in every fold. Besides, most of the variables in the first set in CV appeared in the first set of canonical loadings for the whole cohort. The second set of canonical loadings turn to be less stable than the first set with the most occurrence being 3 out of 5. The second set of canonical loadings (Appendix A.4) looks like a combination of the second and third sets in the one-off analysis on the whole cohort.

We focus on the stability of canonical loadings on SM inputs of CCA (Fig. 4.8). The top two most stable latent factors are Language factor (Cognition 4) and Delay Discounting (Cognition 3), and they are the top two in the one-off analysis (Fig. 4.6). Similar to the canonical loadings on the observed data, the second loadings here are combined from the second and third loadings in the one-off analysis (analysis on the whole cohort). Comparing the stability of the canonical loadings on CCA input (Fig. 4.6) with the canonical loadings on observed variables (Appendix A.4), there

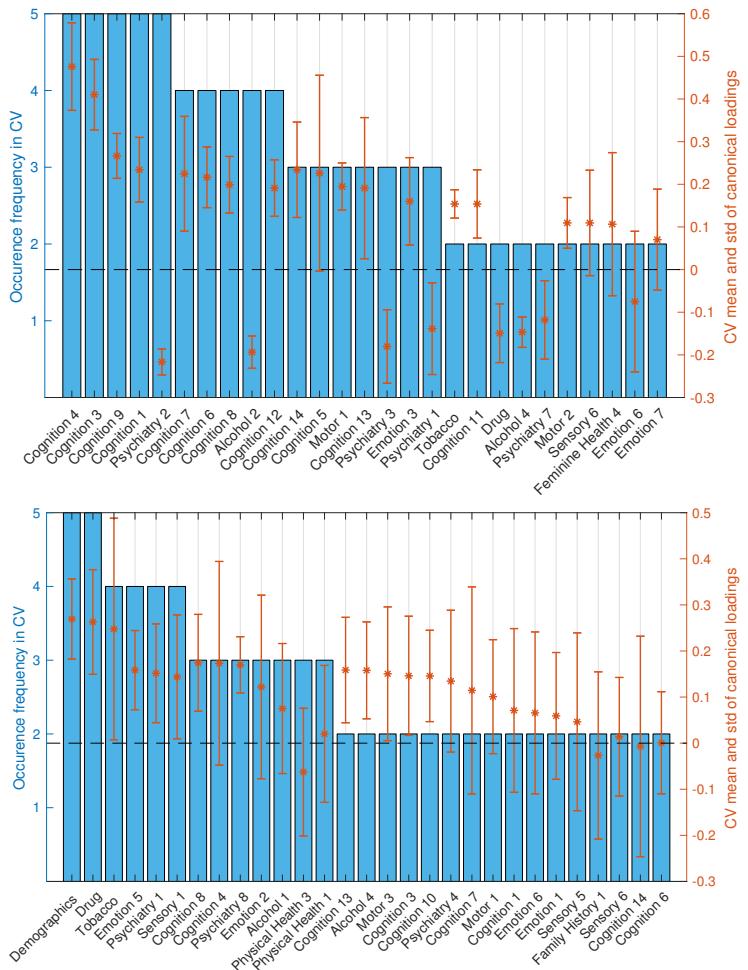


Figure 4.8: Stability of SM canonical loadings on CCA input. Bar plot shows the occurrence frequency in CV out of the 5 folds. Variables are chosen by selecting the top 20 mostly weighted ones in each fold. The ones appeared at least twice are shown above. The right axis shows the mean and the standard deviation over all occurred loadings. Top and bottom plots are the canonical loadings for the first and second canonical variables respectively.

are improvements from the occurrence frequency point of view, as well as in the variance of the canonical loadings across different folds. Particularly for the second set of canonical loadings, loadings on the observed variables have the highest occurrence being 3 whereas 5 on the CCA input. The only SDR factor in Demographics and Drug Use appeared in every single fold with high and stable canonical loadings, which we cannot observe from the loadings on the observed variables. Moreover, similar to the one-off analysis on the whole cohort, the stability of canonical loadings on the CCA input presents the contrast of relationship again. Psychiatry, Drug, Alcohol factors have opposite contributions to Cognition and Motor ones in the first set of canonical loadings.

The BM canonical loadings also exhibit high stability (Fig. A.16). Similar to SM, the second CCA mode is less stable than the first mode. Some ICA regions, such as ICA 18 and 15, appear to be highly stable in both modes as well as both positive and negative maps; ICA 64, 37, 29, 54 and 51 also appear in both positive and negative maps, but only for the first mode. We can also observe the overlaps between the positive and negative maps from Fig. A.16. For example, out of the eight highly stable ICA regions in the second CCA mode, four of them are the same.

4.5 Discussion

In this project, we carefully replicated the study in [115] with a modified analysis pipeline for the HCP S1200 release. We replaced PCA with SDR, a refined dimension reduction technique that automatically estimates the dimensionality of the data, particularly for the function-specific sub-domains of the SM and independent regions of BM. It is often quite challenging to interpret the results from applying CCA to behavioural and brain imaging data, in particular, the canonical variables and the canonical loadings. The primary motivation for using SDR instead is to improve the interpretability of these results.

Sign-flipping and de-confounding

Sign-flipping in SM plays an essential role in this study. The reason of flipping is to assure the results of CCA are not affected by the inconsistency of variable recording. As expected, after flipping the signs of negatively recorded variables, the contrast between canonical loadings presented in the original study ([115]) is gone. For example, the canonical loadings for Picture Vocabulary Test, Oral Reading Recognition

Test, Fluid Intelligence (correct response) do not have opposite behaviour with many of the tobacco and alcohol measures. After sign-flipping, they are now laying on the same side of the axis, i.e. the canonical loadings of those variables have the same sign (Fig. 4.3 and Fig. 4.5).

Apart from the impact of the sign-flipping, the PCA replication in this work overlaps considerably with the previous study. The difference in results may arise, in part, from the different variable sets used, as this work considered a wider range of variables. Also, the confounders are slightly different between the two studies with the current work including racial factors, release versions, age and gender as extra confounders.

Comparing SDR with PCA

By grouping the variables of SM into sub-domains based on their functions, we are able to interpret the canonical loadings of the inputs of CCA as shown in Fig. 4.6. This would not be straightforward when using principal components of the whole data space as inputs. By doing so, we have also saved the effort on manually selecting relevant variables to use in the analysis. The dimension estimating algorithm we applied minimises the prediction errors in each sub-domain, therefore achieves the same goal as picking important information manually from each functional domain. Although the SDR-reduced space would explain less variance than the same dimensional PCA space as PCA is designed to maximise variances explained in the original datasets, SDR focuses more on the structure within variables that share the same functionality, making sure each functional domain has a representative number of components feeding into CCA.

Both PCA and SDR have their own advantages and disadvantages. Using data that is reduced by PCA as inputs, the canonical correlations are higher than using SDR (Tab. 4.1 and Tab. 4.3). Permutation testing gives more significant canonical variables for PCA and those explain slightly higher variance in the original datasets. This is due to the fact that PCA-reduced sub-space is still orthogonal, whereas SDR-reduced sub-space is not. This allows PCA to capture more variance in the observed dataset than SDR (with the same dimensionality). However, SDR saves the effort of selecting relevant variables manually and it automatically estimates the dimensionality. One of the largest drawbacks of PCA is that the results are not as interpretable as SDR. With SDR, we could track the contribution of each sub-domain and directly interpret the canonical loadings of CCA inputs which cannot be easily interpreted in the PCA case. Moreover, these loadings are not subject to the signs of the observed variables. We applied the same stability analysis

to canonical loadings on the SDR factors (CCA inputs) and found higher stability than the loadings on the observed variables (Appendix A.4 and Fig. 4.8).

What we learnt from CCA

In general, we have found that with larger sample size, CCA tends to identify less correlated canonical variables (Tab. 4.1 and 4.3). There is evidence showing that correlations tend to have higher bias with smaller sample size ([80]). When we try to interpret canonical correlation using small samples, we should be extremely cautious and depend on out of sample validation to obtain unbiased estimates of canonical correlation. Additionally, canonical correlations get weaker if we use lower dimensional data as inputs. This is explained by higher dimensional data having greater flexibility to maximise the correlation. We observe that canonical variables constructed by lower dimensional data actually have increased average variance explained in the observed datasets (Tabs. 4.1 and 4.3). Further analysis shows that the amount of variance explained in the original dataset has a non-linear behaviour against CCA input dimension, and it peaks at around dimension 30 in this study.

Further, we found that mean, median and 90th percentile of the distribution of canonical loadings also reduced with increased CCA input dimension. Hence we postulate that higher dimensional inputs may overfit and produce canonical variables that are less related to the original variables.

Since CCA maximises the correlation between two sets of data rather than the variance canonical variables explain in their original datasets, it is important to be aware that variance explained can be an informative measure, however, cannot become the sole measure used to assess CCA performance. Other measures should be considered such as canonical loadings and canonical correlations.

4.5.1 Interpretation of CCA loadings

Variable importance is always a major challenge in interpreting CCA results. The canonical weights are the most direct measures of the importance of CCA inputs. However, they are sensitive to the inputs: small perturbation in inputs can lead to significant change in canonical weights, thus not ideal for variable importance evaluation ([24] and [49]). Different studies ([24] and [123]) have suggested using structural coefficients which are also known as canonical loadings to measure the variable importance. In this study, we have shown that it is a stable measurement with the canonical loadings on SDR factors being more stable than on observed

variables (Fig. 4.8 and Appendix A.4).

Notably, canonical loadings are sign-subjective since they are just correlations between canonical variables and observed variables/CCA inputs. However, principal components which are often used as the inputs of CCA and canonical variables are sign-invariant, i.e. they may take arbitrary signs between different simulations. Therefore, one should not interpret the absolute sign of canonical loadings as the positive/negative contribution of the variable. We should interpret the loadings as what they are in contrast with, and what is the picture on the other set of canonical variables, in our case, linking SM and BM canonical loadings.

Interpretation of latent factor models like CCA still remains challenging. Researchers often need to trade between model performance and interpretability. However, in the areas of medical/public health research, being able to interpret the results of any statistical/machine learning model is of vital importance. The SDR method proposed here tries to combine prior knowledge on the data collected with mathematical models, to improve the understanding of the intermediate and final results of a CCA pipeline, at the same time, reducing the arbitrary choices researchers have to make and increase analysis automation. To total understand the mechanism between brain and behaviour, and fully interpret the results of all different kinds of latent factor models, much additional research across disciplines is still required.

CHAPTER 5

UK Biobank Project

5.1 Introduction

UK Biobank is a population-based longitudinal epidemiological study consisting of 500,000 participants aged 40 to 69 in the UK. It aims to ‘improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses’ [3]. UK Biobank has a wide variety of data modalities including behavioural, demographics, hospital admissions, brain and body imaging, genetic data, as well as long term follow-ups for the participants. Therefore it offers a unique opportunity for researchers to study a wide range of health-related topics.

We consider UK Biobank as an excellent opportunity for applying the previous analysis pipeline to an independent and much larger dataset, with the possibility to extend the analysis to other data modalities. With datasets as large as the UK Biobank, we should have high power to detect associations present in the data. Moreover, in the previous chapter, we only studied the relationship between functional connectivity and behavioural measures. However, structural measures are significant constituents of the brain, and they have been shown to be closely related to many health-related measures as well as the functional MRI. With the availability of the rich data modalities in the UK Biobank, we are able to investigate the interplay between all three parties.

There are three main stages of this project. The first stage is data acquisition, cleaning and pre-processing; the second stage is analysis replication on the new datasets; the last stage is the analysis extension. We extend the analysis pipeline used in the HCP project (Section 4.3) to include one more modality, *image-derived phenotype* (IDP). IDP includes measures like structural MRI, diffusion MRI and susceptibility-weighted brain MRI. Together with the *functional connectivity* (FC) from resting-state fMRI (rfMRI) and *subject measure* (SM; behavioural and demographic measures), we aim to explore the latent structures shared by all three

modalities, as well as the unique patterns embedded into each pair of them. To achieve this, we apply MCCA (introduced in Section 2.3.2) and GFA (introduced in Section 2.5) to examine such relationships.

5.2 Data

The subject measures (SM) were collected over all 500,000 participants at the baseline assessment during recruitment. Among those subjects, 200,000 were chosen to have repeat assessments. The repeat assessments were scheduled to be carried out every two to three years. So far, there has been two such visits, and the last visit is also the imaging visit when the participants have the MRI data collected. Collecting imaging data is very time consuming and therefore is still an ongoing process. It aims to collect imaging data from 100,000 of the participants who attend repeat assessment. The data is released on a rolling basis. Up to June 2019, just under 40,000 participants have been scanned. When this present research was started, brain imaging data for around 13,000 subjects was released, of which 9932 also have IDP data available. Finally, this project consists of 9301 subjects who have all three modalities available for analysis.

Functional connectivity (FC) was constructed from 100-dimensional group-ICA parcellation using the same processing pipeline with the HCP project (refer to Section 4.2.1 for more details). Out of the 100 ICA regions, 45 were not neuronally driven which were discarded. The remaining 55 regions give 1485 ($55 \times 54/2$) functional connectivities for each subject. For non-imaging modalities (SM and IDP), there are in total 13,803 variables, 887 of which are selected as image derived phenotypes (IDPs), and the rest are subject measures (SMs).

5.2.1 Sub-domain grouping

To prepare the data for supervised dimension reduction (SDR; introduced in Section 3.2), the variables need to be grouped into sub-domains.

However, instead of manually grouping over 10,000 variables, we opted to define sub-domains of interests. The definitions of the sub-domains are guided by the categories provided by the UK Biobank official website (<http://biobank.ndph.ox.ac.uk/showcase/cats.cgi>).

We selected nine SM sub-domains, including ‘Mental Health’, ‘Health & Medical History’, ‘Alcohol Use’, ‘Tobacco Use’, ‘Cognitive Phenotypes’, ‘Lifestyle & Environment’, ‘Exercise & Work’, ‘Food & Drink’ and ‘Physical measures’. This

grouping gives 7712 SMs. The IDPs are grouped into ten sub-domains based on the official website categories. They are ‘Cerebral Volume’, ‘T2 & Bold’, ‘FA’ (fractional anisotropy), ‘MD’ (mean diffusivity), ‘MO’ (diffusion tensor mode), ‘L1-3’, ‘ICVF’ (intra-cellular volume fraction), ‘OD’ (orientation dispersion index), ‘ISOVF’ (isotropic or free water volume fraction) and ‘Cerebellum volume’. The FCs grouping is again based on ICA regions (same as the HCP project). Therefore, there are 55 FC sub-domains.

5.2.2 Data issues and fixes

SMs in the UK Biobank project were collected over a vast range of health and lifestyle factors. Due to its enormous scale and excessive details, SMs have a few special characteristics as well as challenges.

1. The data is longitudinal and has multiple repeated assessments, i.e. many variables are collected multiple times at different time points. Note that in this analysis, variables from all visits are included.
2. The data has hierarchical structures, different answers to the same question may lead to different subsequent questions. It causes hierarchical missingness in the data. For example, if a parent question asks how long has the participant been using their mobile phone, and they answer that they’ve never used a mobile phone, the participant will not be asked the children questions like ‘What is the average weekly usage of mobile phone in the past three months?’ Therefore, the answers for the ‘children’ questions would be missing for this group of participants.
3. The data is extremely sparse. If the questions are not relevant to the participants, e.g. hospital admission or disease related, or the questions are too trivial to notice in life, e.g. the thickness of butter/margarine spread on bread rolls, they might be left unanswered.
4. The rules for data recording are heterogeneous and non-intuitive. For example, the daily intake of avocados could be recorded as one or two, meaning that a participant has one or two avocado(s) daily. However, over three avocados might be recorded by 300 while a half or a quarter of avocado may be recorded by 444 and 555, respectively.

To resolve the above issues, as well as reduce the sparsity, minimise the uncertainty and improve the consistency of the data, we apply the following steps to the raw non-imaging data.

1. **Data re-coding:** Concerning the heterogeneity and interpretability of data entries, we re-code the data into a monotonic or intuitive order. For the avocado example, we re-code more than three avocados as 3; a quarter, and a half avocados as 0.25 and 0.5, respectively. This procedure entails screening all data codings and then re-coding the non-intuitive ones.
2. **NA insertion:** Many of the questions are answered as ‘Do not know’ and/or ‘Prefer not to answer’, and they are recorded by negative values (e.g., -1 and -3). We can primarily treat such answers as missing (‘NA’). Although at this stage, this would boost the sparsity, it avoids distorting quantitative data and allows imputation methods to impute those values later on.
3. **Hierarchical missing data impute:** To resolve the hierarchical missingness, we applied structural imputation. In the example mentioned in the second data issue, it is not hard to see that the child question weekly mobile phone usage should be 0 if the answer to the parent question is never used mobile phone before. Therefore, for cases like these, we can impute children questions from NA to 0 and reduce the missingness of the data. Notably, child value imputation only works for quantitative data. If the variable is nominal or it has been coded in a non-straightforward way, this imputation should not be implemented. There are 88.50% of entries missing in the whole SM data matrix. With the child value imputation implemented, the missingness is reduced to 88.24%, with the algorithm impute 188,542 entries.
4. **Nominal variable removal:** UK Biobank labels all non-binary categorical variable as ‘categorical (multiple)’, without distinguishing nominal and ordinal variables. In the analysis, nominal variables are often treated differently from ordinal variables. As discussed in Section 2.2.4, nominal variables are transformed into dummy variables to enter the analysis. There is no automatic way of labelling them apart from going through the variable definitions manually. Most of the categorical (multiple) variables are nominal, and the majority of them are medical codes (e.g. ICD codes) which have very finely defined sub-categories. After turning them into dummy variables, they become too sparse. Therefore, we removed all nominal variables.

The above steps were implemented in Python and the UK Biobank data pre-processing Python package ‘*funpack*’ [87]. Although the order of data cleaning is not unique it may still be very important. For example, in ‘*funpack*’ the last step in NA insertion is to further turn all negative values to NA. This step needs to be done

after the data re-coding step, and data re-coding should make sure all informative negative codings are re-coded into non-negative numbers.

5.2.3 Data Pre-processing

All three modalities (FC, SM and IDP) went through the same quality control (QC) procedure. This procedure excludes variables with more than 50% missingness, 0 standard deviation, or/and having a dominant value (over 95% of the values are the same). Also, for variables with correlations higher than 0.99, the one with higher missingness is excluded. If both variables have the same missing rate, one variable is removed at random.

Variables which passed QC then went through the following steps. SMs and IDPs firstly underwent rank-based inverse normalisation (Section 2.8.2). After missing value imputation, they were de-confounded (Section 2.8.4), mean-centred and normalised by standard deviations. For FC, the process was the same apart from that the rank-based inverse normalisation was replaced by global standard deviation normalisation (Section 2.8.2).

We carefully selected the following confounders: age, sex, scan date, head size, rfMRI head motion and tfMRI head motion, age^2 , $age \times sex$ and $age^2 \times sex$.

5.2.3.1 Missing value imputation

There were several imputation methods attempted in the UK Biobank project including GLRM (Section 2.8.3.2), variable-wise kNN (Section 2.8.3.1) and soft-impute (Section 2.8.3.3). GLRM was implemented by Python package ‘GLRM’ [129] which is out of maintenance and therefore cannot run properly on the data. kNN and soft-impute gave similar results, therefore, we picked the same method as the one used on the HCP data, soft-impute.

Finally, after all above pre-processing, SM was reduced to 482 variables on the nine sub-domains; IDPs were reduced to 869 variables. Since FC is dense, it is still of dimensionality 1485.

5.2.4 Sign-flipping

Recall that sign-flipping aims to align all SM variables so that they all have co-ordinated meaning: higher value means better life outcome. More specifically, we first picked a benchmark variable, ‘Average total household income before tax’, then

flipped the sign of the variables with Pearson correlation less than -0.01 . Finally we went through all variable meanings for sanity check.

This automatic flipping worked fairly well on the HCP dataset, however it only worked well on the ‘Mental Health’, ‘Alcohol Use’ and ‘Tobacco Use’ sub-domains of the UK Biobank data. For other sub-domains, especially the ‘Health & Medical History’, many variables did not flip automatically according to their meanings. Therefore, we replaced the automatic correlation method with pure human knowledge. For ‘Physical Measures’ and ‘Food & Drink’ sub-domains, we realised there is no clear way of associating most variables with ‘good life outcomes’ (for example, both high and low BMI can indicate poor health); therefore these domains were left un-flipped.

Again, sign-flipping is only performed for SMs, since the variable meanings in IDP and FC do not easily take on a definition of ‘good’ or ‘bad’.

5.3 Method

5.3.1 Canonical correlation analysis pipeline

Fig. 5.1 illustrates the general CCA pipeline. There are two parallel pipelines being carried out, with and without dimension reduction prior to CCA. Unlike in the HCP, the number of subjects in the UK Biobank is much larger than any modality dimensions, and so it is feasible to apply CCA directly to the non-reduced data.

To assess the SDR performance, we also applied PCA to reduce the dimensionality of the three data modalities, FC, SM and IDP (see Appendix B.4.3 for the results of CCA on PCA-reduced data). For the pipeline without dimension reduction, we fed the pre-processed data directly into CCA/multi-view CCA. For the pipeline with dimension reduction, we fed the reduced data (latent factors) into CCA/multi-view CCA. As noted in Fig. 5.1, during the CCA step, we implemented both pairwise CCA for all pairs of the three modalities and multi-view CCA to investigate latent structures shared by all three modalities.

In the end, we tested the significance of canonical pairs by applying permutation testing (Section 2.9.2) for 1000 permutations and 10-fold CV (Section 2.9.1). Permutation testing tests the significance from statistical re-sampling perspective, whereas CV provides insights from the prediction point of view.

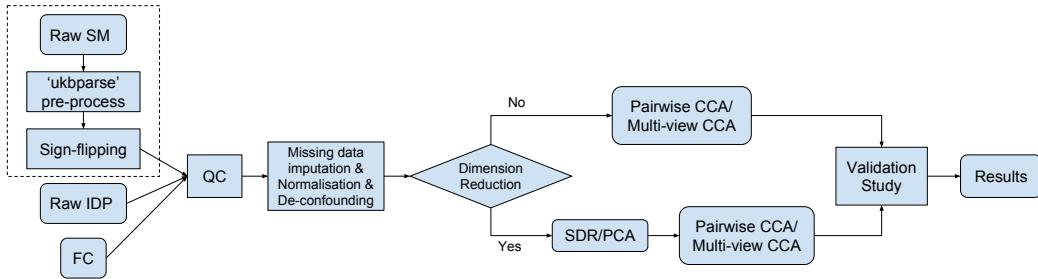


Figure 5.1: UK Biobank CCA analysis pipeline. Due to the higher complexity of SMs, they go through more data cleaning steps. Then all FC, SM and IDP undergo the same QC and pre-processing procedures. Two parallel CCA pipelines follow after, with and without dimension reduction.

5.3.1.1 Model interpretation

To assess the performance of CCA, we looked at the canonical correlations, canonical loadings and variance explained which were introduced in Section 2.3.

Canonical loadings are simply the Pearson’s correlations between the canonical variables and the observed/input variables. While the canonical loadings are straightforward to visualise for SM and IDP, they are far more challenging to visualise for FC since these are generated over the 1485 measures. Given that each of the 1485 FC measures represents a connection between two ICA regions, it is very difficult to interpret every single loading. Therefore, we proposed some summary statistics to represent the canonical loadings for each ICA region and visualised them on brain volumes. The following procedure describes brain maps are generated:

1. Map each set of 1485 canonical loadings back to a 55×55 symmetrical loading matrix.
2. Apply Fisher’s z transformation to the loading matrix.
3. Multiply the transformed loading with the signs of the original connectivity matrix (i.e. group mean (partial) correlation).
4. Sum up the 10 largest loadings for each region (row/column) as the *positive CCA score* for this region; Sum up the 10 smallest loadings for each region (row/column) as the *negative CCA score* for this region.
5. Load the 55 group-ICA volume maps. Note that these maps store z-transformed ICA weights.
6. Flip all ICA volume maps that have negative peaks, and covert all negative values in every volume map to 0. This is equivalent with silencing voxels that

do not have significant contributions in forming the ICA regions, and $z = 0$ is an arbitrary threshold for visualisation purposes.

7. Multiply the positive and negative CCA scores by the voxel-wise z-transformed weights, and average over all voxels within each ICA region.
8. Finally, sum up all 55 ICA regions, then normalise by the mean of the 55 ICA regions to get one volume map for each of the positive and negative scenarios, and name them as the *positive brain map* and *negative brain map*.

For the rest of the study, we use this procedure to visualise FC canonical loadings.

We used Neurovault’s ([51]) NeuroSynth image decoder ([145]) to assist the interpretation of the brain map. The NeuroSynth image decoder uses a large meta-analytic database of abstract text and peak coordinate data to find the words that are most likely to appear with a given statistic map.

5.3.1.2 Cross-validation of SDR CCA

We cross-validated the SDR CCA results to examine the stability and generalisability. The CV method overview is shown in Fig. 5.2. It is modified from Fig. 2.3 in Section 2.9 with SDR added to the procedure, and only shows two views/datasets as an illustration. CV with three views is implemented in the same fashion.

In Fig. 5.2, we fed rotated components ($RC_{in}^{X_i}$ and $RC_{in}^{Y_i}$) from SDR into CCA in each fold of CV. To be able to compare the results of CCA between different folds, we need to make sure that the inputs of CCA represent the same components between folds, i.e. the meanings of rotated components between different folds are the same. Therefore, we investigated the stability of the rotated components (RCs) from the SDR step (RC in Fig. 5.2) by a separate 10-fold CV. Note that factor rotation is applied to the principal loadings as shown in Fig. 5.2, and RCs are obtained by multiplying the original data with the rotated principal loadings. Essentially, we are testing the stability of the rotated loadings (RLs). The stability of the RLs is affected by two factors: the factor rotation step and inherent sign indeterminacy issue. In PCA, principal components are ordered by the variance explained (or eigenvalues). However, as discussed in Section 2.7, factor rotation will change the amount of variance explained by each rotated component, and it does not preserve the order, i.e. the RLs can be output in any order. This implies that even if the RLs are very stable across different folds, we lose the stability if they are output by different orders. In addition, the signs of principal loadings/components

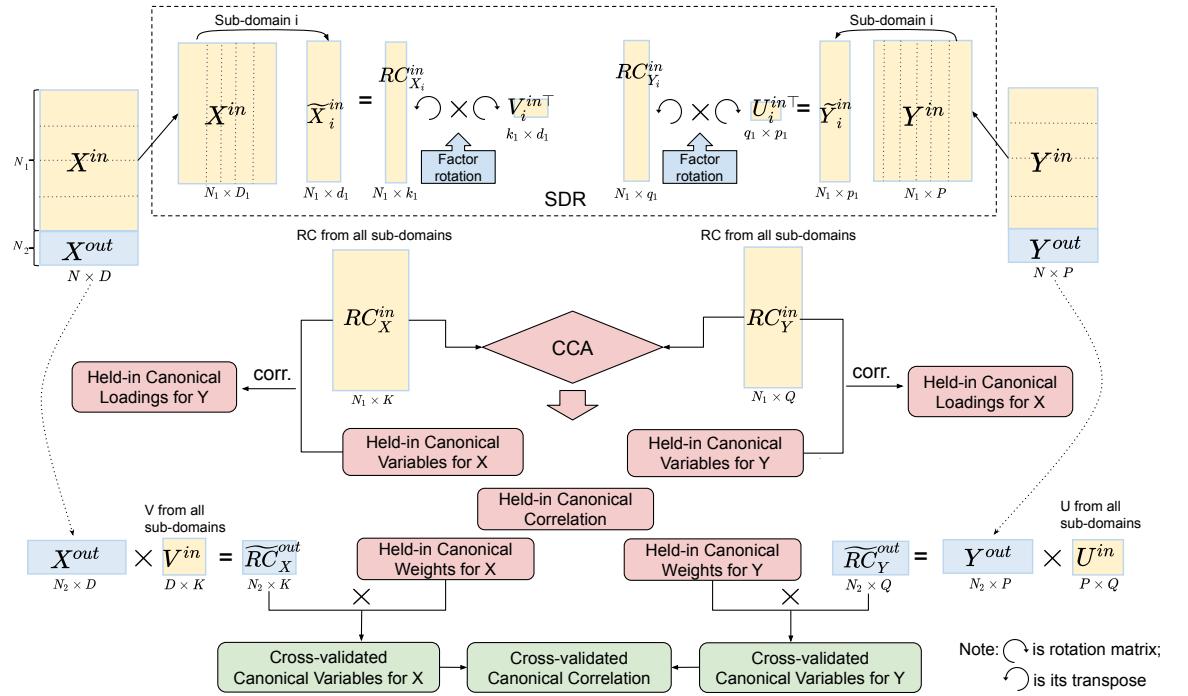


Figure 5.2: Cross-validation method overview. We apply SDR (dotted box on the top) to the held-in (training) set after the split of the data. Within SDR the principal loadings are rotated to construct rotated components (RCs), and then the RCs of the held-in set are fed into CCA. Cross-validated canonical variables and correlations are obtained by multiplying the held-in canonical weights with the RCs from the held-out (test) set.

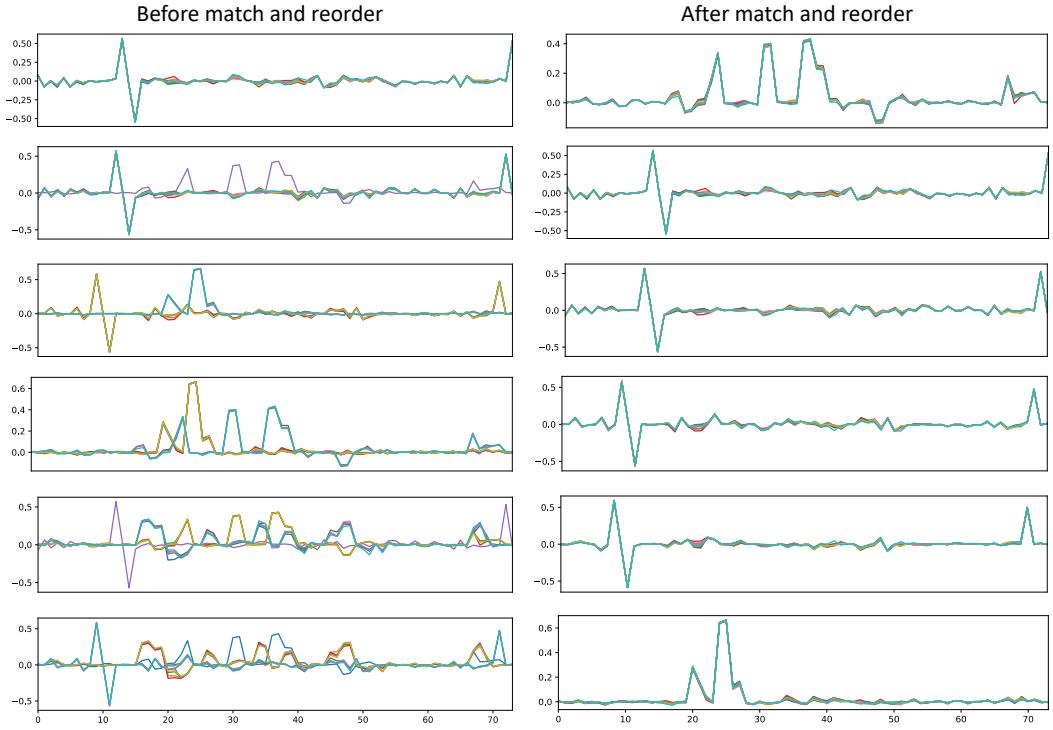


Figure 5.3: The left set of figures are some of the rotated principal loadings in a SM sub-domain in 10-fold CV; The right set of figures are the same group of rotated loadings but after applying the matching algorithm and ordered by R-squared values.

can be randomly flipped in different simulations/repetitions. Therefore, the signs of RLs/RCs can be random as well.

To resolve these issues, during the first fold of the RL CV, we ordered all RLs by R-squared values (fixing order), and made sure the mean of the top values in each RL is positive (fixing sign). In the following folds, we reordered the RLs to maximise the pair-wise correlation between RLs in the current fold with the RLs in the first fold (using the Gale-Shapley algorithm¹ [47]). In the end, we visualised the RLs in every fold to see stability. Fig. 5.3 shows an example group of RLs before and after realignment. As we can see from Fig. 5.3, RLs show very nice stability across different folds after the realignment. However, in the larger CV for the analysis pipeline, we still needed the RLs to be in this stable order. Therefore, we set benchmark RLs for all sub-domains and all data modalities. The benchmark RLs were calculated by taking the mean of all folds in the RLs CV, i.e. the mean

¹We try to match the target list with the reference list. For the first item in the reference list, we find the best match (highest correlation) from the target list, place it first and remove it from the target list, and so on.

of all realisations in the right set of Fig. 5.3. We then summarised the RCs for all sub-domains in SM and IDP based on these benchmark RLs (an example will be shown in Section 5.4.4.1, and the rest is shown in Appendix B.1).

Once RLs are set and stable, we can implement the CV procedure shown in Fig. 5.2. Every time after the rotating the principal loadings (U and V in Fig. 5.2) within the SDR step, we matched the current RLs with the benchmark RLs to make sure the RCs in each fold of CV have the same meanings. In the end, we analysed the stability of the training set canonical loadings and the cross-validated canonical correlations.

All the above analysis is implemented in Python with the use of ‘pyrcca’ package [19].

5.3.2 Group factor analysis pipeline

We applied GFA as a complementary method for multi-view CCA. Multi-view CCA only seeks the latent structure shared by all input views/datasets, whereas GFA also finds the latent structures underlying a subset of the input datasets (Section 2.5).

The GFA pipeline is almost the same with the CCA pipeline (see Fig. 5.1), only with ‘Pairwise CCA/Multi-view CCA’ replaced by ‘GFA’. However, after the preliminary analysis of GFA on non-reduced and PCA-reduced datasets, we realised that the results are hardly interpretable. Therefore, we have decided to focus on the analysis of GFA on SDR-reduced datasets.

GFA analysis is implemented in R with package ‘CCAGFA’ [136]. As discussed in Section 2.5.1, we need to specify the model by initialising a parameter K which is an estimate of the dimension of the latent space. The model will start from K latent components and drop the unloaded components during inference. As noted in Section 2.5.1, if K is set to a value which is lower than the true latent space dimension, then the model will be mis-specified, and no components will be dropped. However, if K is initialised too high, i.e. too many unnecessary components to start with, it slows the inference down significantly.

Once a proper K is chosen, to improve the stability of the model, it is recommended to run the model several times and pick the best set of parameters which offers the lowest lower bound $L(\Theta)$ in Eqn. (2.36) for model interpretation. This could potentially stop the algorithm stay trapped in a local minimum.

5.4 CCA Results

Fig. 5.4 displays the pairwise correlations before and after sign-flipping of SMs. The two correlation heat maps present somewhat similar patterns with the most noticeable change in the ‘Cognition’ block. Notably, if all variables were flipped in a sub-domain, the pairwise correlation within sub-domain would stay the same; however, the correlations pattern with other functional sub-domains would be reversed. This is the case for sub-domains ‘Mental Health’ and ‘Health & Medical History’ where most variables were subject to sign flipping. Thus the top left corner between the left and right plot in Fig. 5.4 look very similar, however, the colour patterns almost flipped across other sub-domains. In contrast, in ‘Food & Drink’ and ‘Physical Measures’ sub-domains, no variable was flipped. Besides, the strongest correlation pattern is shown within the ‘Physical Measures’ sub-domain, and this domain seems to have the strongest cross-domain correlations as well.

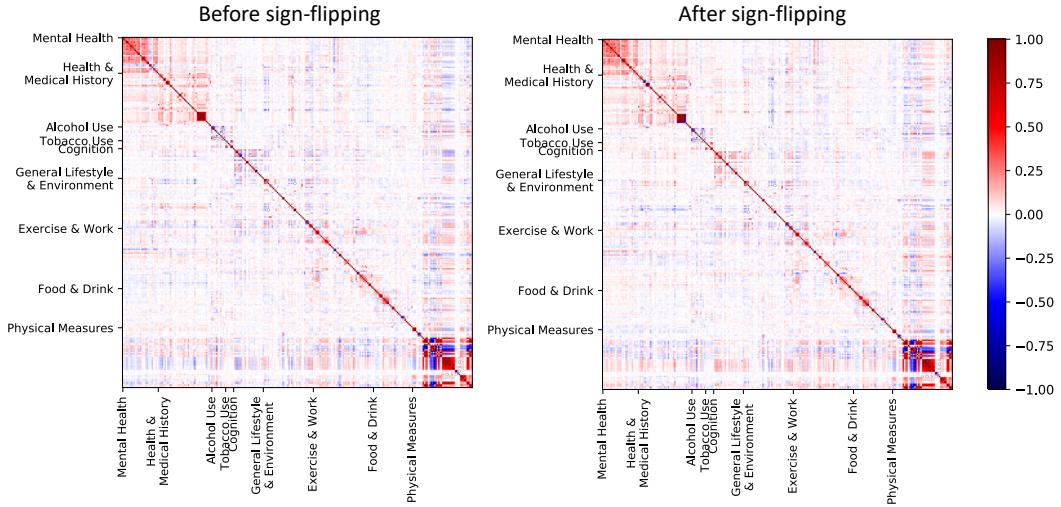


Figure 5.4: Correlation between every pair of SM variables before (left) and after (right) sign-flipping. The most noticeable change happen in the blocks of ‘Alcohol Use’, ‘Tobacco Use’ and ‘Cognition’ where many of the variables are sign-flipped. For sub-domains ‘Mental Health’ and ‘Health & Medical History’, almost all variables are flipped. This is reflected by the reversed colour pattern of the correlation between these two sub-domains and the rest of the sub-domains. ‘Food & Drink’ and ‘Physical Measures’ are left un-flipped.

5.4.1 SDR results

SDR reduces FC from 1485 to 57 dimensions, SM from 482 to 107 dimensions and IDP from 869 to 205 dimensions. A sub-domain breakdown of SDR SM and SDR IDP are shown in Fig. 5.5. We notice that sub-domains are reduced by different proportions which reflects different noise ratios contained within each sub-domain.

For FC, SDR reduces almost all ICA regions to one dimension apart from ICA region 5 and 22 which are reduced to two dimensions. This significant reduction implicates that the noise level in FC is the largest among all three modalities.

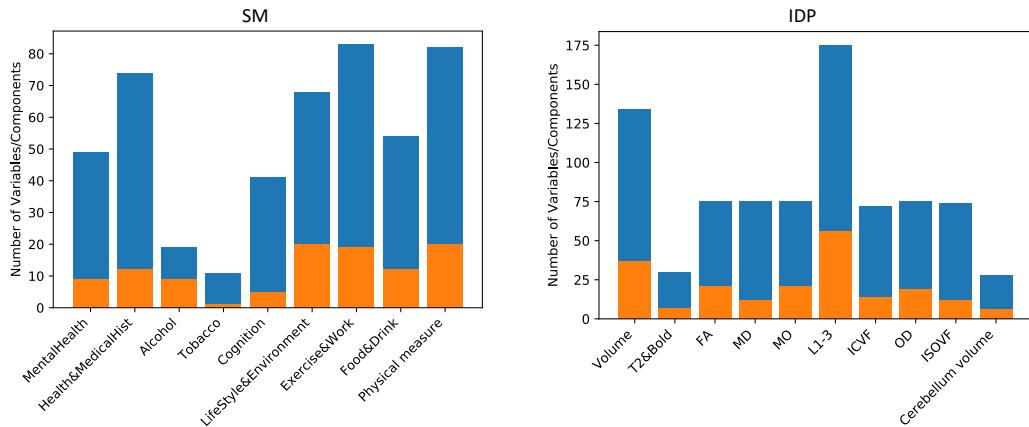


Figure 5.5: The original number of variables (blue bar) in each SM (left) and IDP (right) sub-domain and the number of latent components (orange bar) SDR reduces to in each sub-domain.

5.4.2 Pairwise CCA on non-reduced data

Pairwise canonical correlations for the first ten canonical pairs are shown in Fig. 5.6. Between the three pairs of CCA, FC and IDP have the strongest canonical correlations, then comes FC and SM; IDP and SM have the weakest canonical relationship. Moreover, the canonical correlations between FC and IDP are considerably higher than the other two. This may be explained by the fact that brain-related measures are more correlated among themselves, as well as the high dimensionalities of the non-reduced FC (1485) and IDP (869) so that CCA has more freedom to produce more correlated linear latent space.

Permutation testing gives seven significant pairs of FC and SM canonical variables, six pairs of IDP and SM canonical variables and 31 pairs of FC and IDP canonical variables (Fig. 5.7). These results also suggest the order of correlation

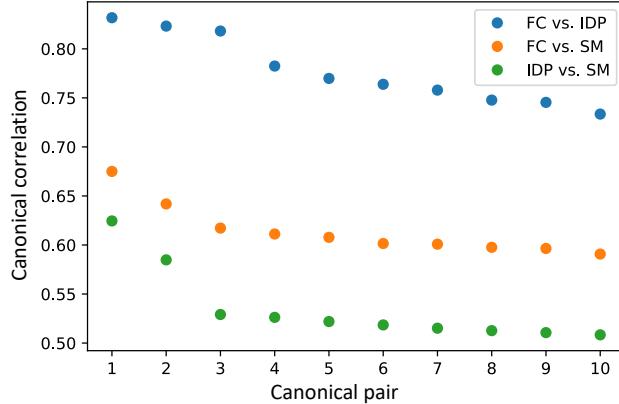


Figure 5.6: Pairwise canonical correlations for the CCA of FC and SM (orange), IDP and SM (green), FC and IDP (blue). The canonical correlations between FC and IDP are significantly stronger than the other two pairs, with the correlations between IDP and SM being the weakest.

between three modalities, i.e. FC and IDP are the most correlated. Since both FC and IDP are brain related modalities and derived from brain imaging, it is not surprising that they have the strongest correlations and more significant canonical pairs.

Fig. 5.8 displays the amount of variance explained (defined in Section 2.3) by the significant canonical variables in their original dataset for all three CCAs. FC canonical variables explain the least variance in its original dataset due to the original high dimensionality of FC. We also notice that variance explained is not a monotonic measure, which addresses the point that canonical variables that maximise between sets correlation do not necessarily explain the most variance in the original data space.

5.4.2.1 Canonical loadings

Due to the high volume of the results, we focus on the top 30 canonical loadings for the significant canonical variables, and only present the highlights within each set of the analysis.

CCA between FC and SM. Most of the top SM canonical loadings for the seven significant canonical variables are from the ‘Physical Measures’ sub-domain. Fig. 5.9 shows a snapshot of the first three significant canonical variables. The first set of SM canonical loadings (leftmost figure in Fig. 5.9) are almost all physical mea-

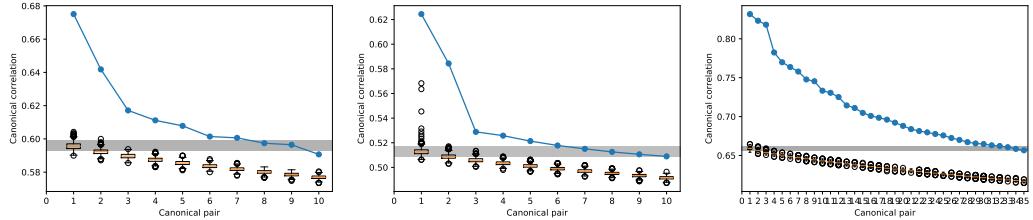


Figure 5.7: Observed canonical correlations (blue line) versus the distribution of canonical correlation between the permuted canonical pairs (box plot). The grey shaded area is the 5 to 95 percentile from the distribution of first permuted canonical pair (first box plot) which is used to define the significance of the canonical pairs (canonical correlation falls under the upper bound of this band is defined as insignificant). From left to right are the CCA between FC and SM, IDP and SM, and FC and IDP respectively.

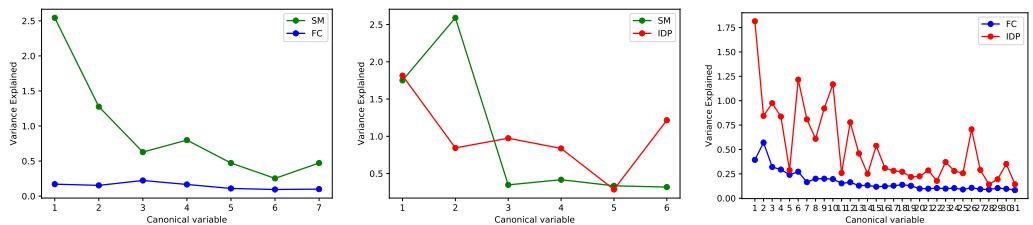


Figure 5.8: Variance explained by the significant canonical variables in their original dataset. From left to right are the CCA between FC and SM (7 pairs), IDP and SM (6 pairs), IDP and FC (31 pairs).

UK Biobank Project

Loading	Name	Loading	Name	Loading	Name
0 0.734214	Body mass index (BMI) visit-3 2.0	0 0.406918	Fluid intelligence score visit-3 2.0	0 -0.252215	Pulse rate, automated reading 0.0
1 0.670712	Body mass index (BMI) 0.0	1 0.376907	Weight visit-3 2.0	1 -0.234901	Pulse rate, automated reading visit-5 2.0
2 0.667755	Leg fat mass (right) 0.0	2 0.375939	-Qualifications 0.0	2 -0.231920	Pulse rate, automated reading visit-2 0.1
3 0.662667	Waist circumference visit-3 2.0	3 0.370264	Fluid intelligence score 0.0	3 0.228062	Number of vehicles in household 2.0
4 0.656368	Leg fat percentage (right) 0.0	4 0.366746	Leg fat-free mass (right) 0.0	4 -0.225194	Pulse rate, automated reading visit-6 2.1
5 0.637466	Whole body fat mass 0.0	5 0.365383	Weight 0.0	5 0.212439	Number of vehicles in household 0.0
6 0.632968	Weight visit-3 2.0	6 0.362716	Year ended full time education 0.0	6 -0.209509	Handedness (chirality/laterality) 0.0
7 0.626562	Arm fat mass (left) 0.0	7 0.360578	-Qualifications 2.0	7 0.201202	Peak expiratory flow (PEF) visit-2 0.1
8 0.625501	Arm fat mass (right) 0.0	8 0.338298	Whole body fat-free mass 0.0	8 -0.197953	Handedness (chirality/laterality) 2.0
9 0.614080	Body fat percentage 0.0	9 0.329695	Waist circumference visit-3 2.0	9 0.195635	-Time to complete round 0.0
10 0.609484	Waist circumference 0.0	10 0.327374	Waist circumference 0.0	10 -0.194716	-Time to complete round 0.1
11 0.594852	Trunk fat mass 0.0	11 -0.324994	Time spend outdoors in summer 0.0	11 -0.187661	Impedance of arm (right) 0.0
12 0.590770	Arm fat percentage (left) 0.0	12 0.315371	Arm fat mass (right) 0.0	12 0.184256	Hand grip strength (right) 0.0
13 0.587975	Arm fat percentage (right) 0.0	13 0.312250	Arm fat mass (left) 0.0	13 0.181378	Hand grip strength (right) visit-3 2.0
14 0.565113	Weight 0.0	14 -0.308995	Time spent outdoors in summer 2.0	14 0.181040	Peak expiratory flow (PEF) visit-7 2.0
15 0.560306	Hip circumference visit-3 2.0	15 0.300668	Arm fat-free mass (right) 0.0	15 0.180443	-Trail making completion status 0.0
16 0.555584	Trunk fat percentage 0.0	16 0.298401	Hip circumference 0.0	16 0.178027	-Duration screen displayed visit-3 2.0
17 0.493437	Hip circumference 0.0	17 0.296339	Whole body fat mass 0.0	17 0.177898	Average total household income before tax 2.0
18 0.367626	Arm fat-free mass (right) 0.0	18 -0.294957	Time spent outdoors in winter 2.0	18 0.176490	Hand grip strength (left) visit-3 2.0
19 0.341210	Time spent watching television (TV) 2.0	19 0.293452	Body mass index (BMI) visit-3 2.0	19 -0.174416	Impedance of leg (right) 0.0
20 0.339498	Diastolic blood pressure, automated reading 0.0	20 0.292969	Trunk fat mass 0.0	20 -0.173634	Impedance of whole body 0.0
21 0.337385	Systolic blood pressure, automated reading vis...	21 0.283603	Hip circumference visit-3 2.0	21 0.173105	Number in household 2.0
22 0.334840	Leg fat-free mass (right) 0.0	22 0.282885	Body mass index (BMI) 0.0	22 0.172054	-Nervous feelings 2.0
23 0.334266	Time spent watching television (TV) 0.0	23 0.281762	Leg fat mass (right) 0.0	23 -0.167952	-Time to complete round 0.1
24 0.329739	Systolic blood pressure, automated reading vis...	24 -0.273620	Job involves mainly walking or standing 0.0	24 0.167318	-Worrier / anxious feelings 2.0
25 0.325669	Whole body fat-free mass 0.0	25 -0.260419	Impedance of leg (right) 0.0	25 0.166978	-Time to complete round 2.1
26 0.323231	Diastolic blood pressure, automated reading vi...	26 0.256721	-Job involves heavy manual or physical work 0.0	26 0.166001	Peak expiratory flow (PEF) visit-8 2.1
27 0.322341	Diastolic blood pressure, automated reading vi...	27 -0.256043	Impedance of leg (left) 0.0	27 0.165338	Peak expiratory flow (PEF) visit-9 2.2
28 0.314085	Diastolic blood pressure, automated reading vi...	28 -0.248760	Time spent outdoors in winter 0.0	28 -0.165011	Impedance of leg (left) 0.0
29 -0.311503	Impedance of arm (left) 0.0	29 0.247667	Arm fat percentage (right) 0.0	29 0.163760	Peak expiratory flow (PEF) visit-3 0.2

Figure 5.9: Top 30 canonical loadings for the first three sets of significant SM canonical variables in the CCA of FC and SM. From top left to bottom right are the first to the seventh canonical loadings respectively. Variables that are sign-flipped have a ‘-’ sign in front of their names.

sures apart from one, ‘time spent watching TV’. Only the 30th variable, ‘impedance in arm (left)’ has a different sign from the others. Interestingly, *fluid intelligence*, *qualification* (sign-flipped) and *education* appear on the top of the second set, and they are in line (having same sign) with *job involves heavy labour* which was sign-flipped, and in contrast with *time spent outdoor* and *jobs involves mainly walking and standing*. From the third set, we start to see physical measures been mixed with cognition, lifestyle & environment and mental health variables. Loadings for all seven significant SM canonical variables can be seen in Fig. B.9.

Based on NeuroSynth, the most correlated terms with each brain map are shown in Tab. 5.1. Fig. B.10 shows the positive and negative maps for the first seven CCA modes thresholded at 95 percentile. Apart from mode 3 whose positive and negative maps largely overlapped in the cerebellum and motor area, the positive and negative maps in all other modes show fairly distinct patterns. We also notice that the functional correlations of mode 3 and 4 in Tab. 5.1 are larger than other modes, especially Mode 1 and 7 are particularly weak.

	Positive map	Negative map
Mode 1	pain (0.174), secondary somatosensory (0.167), default mode (0.146)	sentence (0.173), comprehension (0.171), supplementary motor (0.142)
Mode 2	default (0.193), mind (0.148)	action (0.218), default (0.207)
Mode 3	motor (0.303), primary motor (0.283), sensorimotor (0.275)	somatosensory (0.268), pain (0.23), sensory & primary motor (0.216)
Mode 4	sentence (0.265), language (0.254), speech (0.246), comprehension (0.237), listening (0.237), somatosensory (0.235), pain (0.227)	speech (0.269), motor (0.26), auditory (0.255), somatosensory (0.252), sensorimotor (0.242)
Mode 5	premotor (0.208), task (0.184), working memory (0.149)	goal (0.142), demand (0.137)
Mode 6	somatosensory (0.199), motion (0.19), pain (0.184), perception (0.175)	motor (0.24), movements (0.238), sensorimotor (0.21)
Mode 7	resting (0.109)	retrieval (0.182), working memory (0.167)

Table 5.1: Functional interpretation for the positive and negative brain maps in the CCA between FC and SM (shown in Fig. B.10). The keywords are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.

Combining the results from both sides, in the first mode, we have physical measures positively correlated with the prefrontal area and negatively correlated with the cerebellum. They also have weak positive correlation with pain and negative correlation with sentence and comprehension related brain functions. However, it becomes more complicated to make interpretations once the SM modes start to have variables from different functional sub-domains.

CCA between IDP and SM. The SM canonical loadings are still heavily dominated by physical measures, especially for the first two sets (Fig. B.11). Notably in the first SM canonical loadings set, fluid intelligence is the only non-physical variable, and all the top 30 canonical loadings have the same sign. From the third set, the proportion of physical measures on the list starts to fade away. The last two significant sets are mainly lifestyle & environment, cognition and tobacco variables.

On the IDP side (Fig. B.12), the first set of canonical loadings are almost all grey matter volume related variables. In combination with the first SM loading

set, physical measures are most closely related to grey matter volume variables with the canonical correlation being 0.62 (Fig. 5.6). Grey matter volume variables still have a considerable proportion in the second IDP loading set, but the remaining four significant canonical loadings sets are mixed with different IDP variables. Because of this, it is very challenging to make firm conclusions about domination of individual variables.

Fig. B.11 and B.12 show the full lists of significant canonical loadings for SM and IDP respectively.

CCA between FC and IDP. The canonical loadings between FC and IDP are the hardest to interpret and visualise due to the large number of significant canonical pairs and high canonical correlations. Therefore, we focus on the first six pairs despite the number of significant canonical pair being 31. The first set of IDP canonical loadings are mainly grey matter volume variables. Unlike the loadings in the CCA between IDP and SM, grey matter variables here play a dominating role in the rest sets. This suggests that grey matter volume variables drive the canonical relationship between IDP and SM, whereas other IDP variables like the diffusion measures play important roles in the interplay between FC and IDP. Fig. B.13 displays top IDP loadings for the first six canonical variables.

Fig. 5.10 shows the positive (red) and negative (blue) maps for the first set of FC canonical loadings. In fact, positive and negatives maps for the first six sets canonical loadings (Fig. B.14) display extensive overlaps in prefrontal, parietal and precuneus areas. These overlaps cross different canonical modes. Tab. 5.2 gives the functional interpretation of the maps in Fig. B.14 by NeuroSynth decoding. The keywords ‘default’, ‘theory of mind’ and ‘retrieval’ repeatedly come up in the table, which also suggests the functions of the brain maps overlap significantly. However, as the mode increases, we observe more variations in positive/negative map functions, and less occurrence of ‘default mode’, where the second and fourth modes have considerable high correlations with the default mode network. When it comes to mode 6, semantic becomes the highest related function (highlighted in Tab. 5.2). The above evidence suggests that although the canonical structures between FC and IDP are strongly related (high canonical correlation and more significant canonical pairs), they appear to be very homogeneous from both side (at least for the first six sets), and are mainly driven by grey matter volume from the IDP side and default mode network from the FC side.

Finally, the 10-fold CV study gives 11, 3 and 28 as the best number of

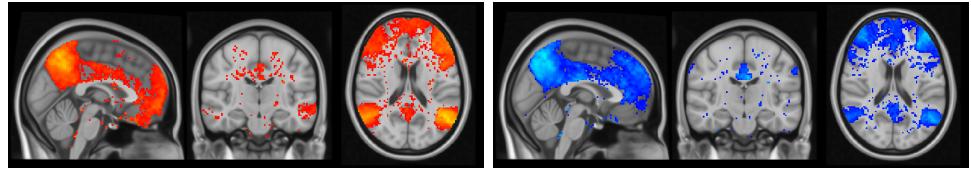


Figure 5.10: Canonical loaded maps for the first set significant FC canonical variables in the CCA with IDP. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).

canonical pairs in the prediction sense for FC and SM, IDP and SM, and FC and IDP respectively. The general trend is in line with permutation testing. However, they are following different validation basis, and therefore, it is hard to compare between the two.

5.4.3 Multi-view CCA on non-reduced data

The analysis in the previous section explored the underlying structure between the three modalities pair-wisely. Next, we conduct MCCA to investigate common structures that are shared by three modalities.

Fig. 5.11 shows the pairwise canonical correlations generated by one MCCA model. FC and IDP still have the strongest canonical correlations and IDP and SM have the weakest relationship. Similar to the pairwise scenarios, FC canonical variables explain the least variance in the original dataset, whilst SM and IDP canonical variables explain roughly similar amount of variance even though the original dimension of IDP is twice as high as the SM's.

Importantly, we notice that the canonical correlation between each pair of modalities is not monotonic anymore (left plot in Fig. 5.11). In traditional CCA (two-view version), we would always observe canonical correlations appearing in descending order. This violation in MCCA is explained by the mechanism of the multi-view setting and it is discussed in Section 2.3.2.

Permutation testing becomes more complicated to apply in the multi-view setting as a result of the non-monotonic behaviour of canonical correlations. We can no longer use the distribution of the first canonical correlation in the permuted data to define significant canonical pairs, since it may not be the largest correlation anymore (Section 2.9.2). Therefore, we came up with the following approach.

Instead of using the distribution of the first canonical correlation, we use the distribution of the canonical correlation with the largest mean value. Moreover, we

	Positive map	Negative map
Mode 1	theory of mind (0.259), default (0.257), retrieval (0.221), mental states (0.216)	default (0.23), working memory (0.216), retrieval (0.207)
Mode 2	default (0.304), theory of mind (0.292), mental states (0.227)	default (0.256), theory of mind (0.229), retrieval (0.193)
Mode 3	default (0.259), retrieval (0.216), theory of mind (0.145)	default (0.278), theory of mind (0.231), mental states (0.192)
Mode 4	default (0.306), theory of mind (0.19), retrieval (0.18)	default (0.241), working memory (0.203)
Mode 5	retrieval (0.181), working memory (0.171), task (0.163), default (0.158)	default (0.242), retrieval (0.193), theory of mind (0.188)
Mode 6	working memory (0.22), retrieval (0.199), task (0.185), default (0.164)	semantic (0.306), retrieval (0.264), sentence (0.245), language (0.232), comprehension (0.227)

Table 5.2: Functional interpretation for the positive and negative brain maps in the CCA between FC and IDP (shown in Fig. B.14). The keywords are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.

carry out two sets of permutation tests.

- Permutation test 1: this tests the pairwise significance. To be able to fairly test every pair of modalities, this procedure consists of two steps: the first step has only FC permuted, and SM and IDP stay the same. We use this experiment to test the significance of canonical pairs between FC and SM, and FC and IDP; the second step has SM permuted and the other two fixed to test the significant between SM and IDP.
- Permutation test 2: this permutation testing is to permute both FC and SM at the same time and keep IDP remain the same. We use this setting to test the significance of the sum of the correlations.

The top left plot in Fig. 5.12 shows that four pairs of FC and SM canonical variables survived the significance level (upper bound of the grey band). However, in the top right plot of Fig. 5.12 (test between SM and IDP), due to the non-monotonic property of the canonical correlations, the true canonical correlation starts to fluctuate around the borderline from the 7th pair. For the permutation testing between FC and IDP, only one (48th) of the first 50 canonical pairs falls

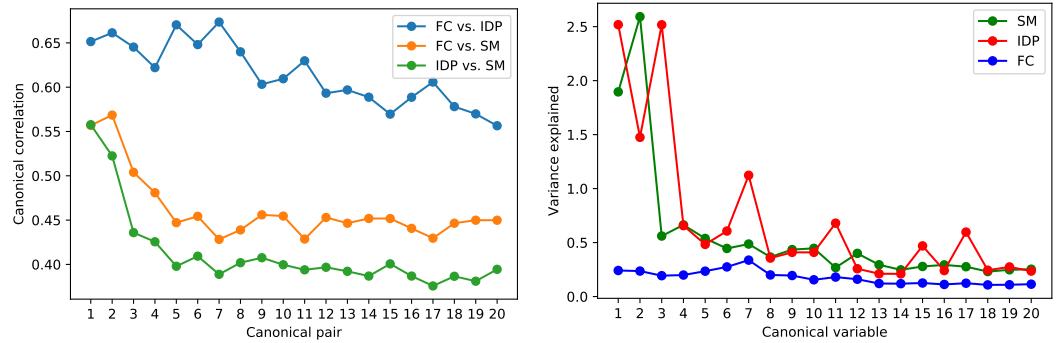


Figure 5.11: The left subplot shows the first 20 multi-view canonical correlations. The right subplot shows the corresponding variance explained in the original datasets.

under the significance level. Therefore, it is hard to draw conclusions for the second and third sets of permutation testing.

We then carried out Permutation test 2 to examine the significance of the sum of the correlations. Fig. 5.13 shows that the sum of the correlations is actually monotonic (blue line). It gives 18 sets of significant canonical variables.

The CV study gives 19 as the best number of canonical pairs to reconstruct the test set with the smallest error, which is similar to the permutation testing result on the sum of the correlations.

5.4.3.1 Canonical loadings

Due to the large number of the significant canonical correlations and the difficulty of interpretation over individual variables, we only present the results from the first four sets of the canonical loadings.

The first set of loadings of SM, IDP and FC (top row in Fig. 5.14, 5.15 and 5.16) are physical measures versus grey matter volumes versus activation in prefrontal area for both positive and negative maps. From the functional interpretation keywords table (Tab. 5.3), we also know that the activation areas in the brain are associated with working memory network, and negatively correlated with the default mode network. The second set of loadings for SM and IDP look very similar to the first set, whereas for the positive brain map, it mainly focuses on precuneus

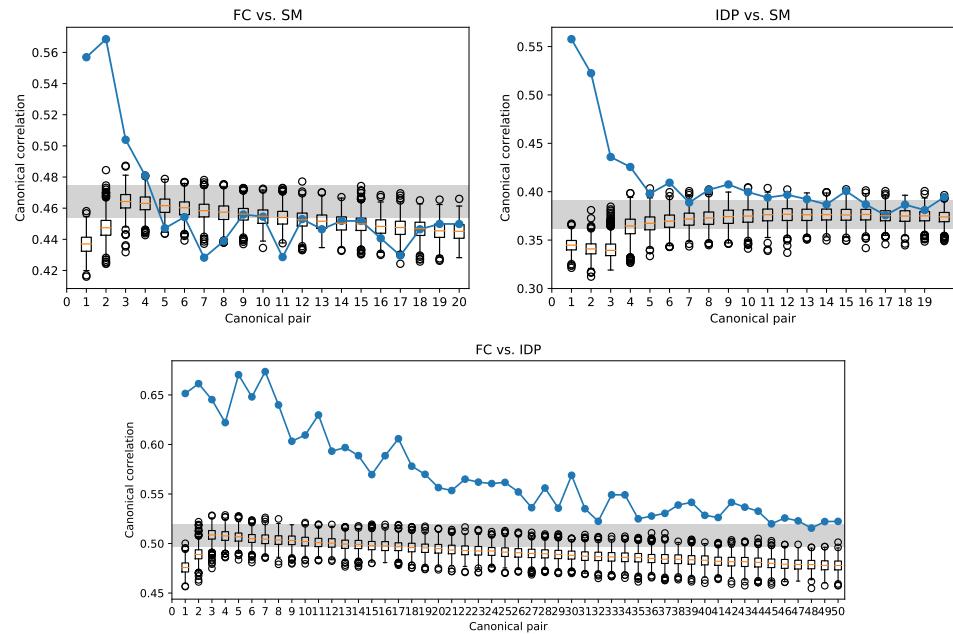


Figure 5.12: The first set of permutation testing in multi-view setting. Plots on the first row are obtained by permuting FC only. Bottom plot is obtained by permuting SM only. Blue lines are the true canonical correlations between FC and SM (top left), IDP and SM (top right) and FC and IDP (bottom). Box plots are the distributions of the canonical correlation between permuted data. Grey band is plotted from the 5th to the 95th percentile of the permuted distribution with the largest mean.

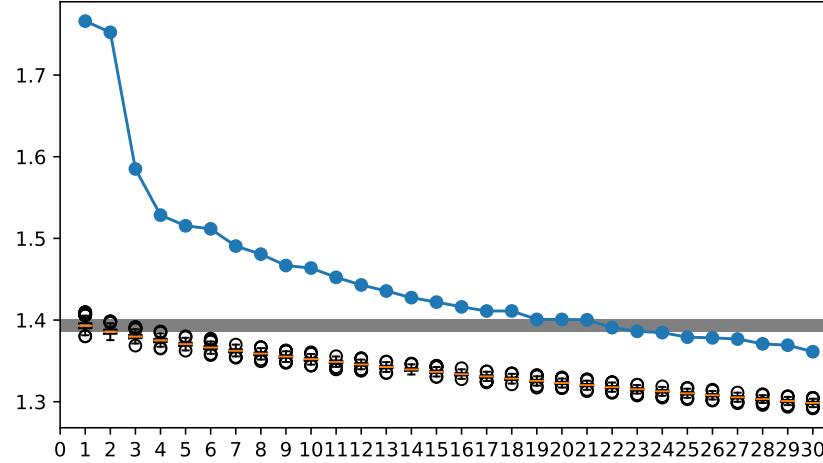


Figure 5.13: The second set of permutation testing, testing the significance of the sum of the correlations with FC and SM permuted at the same time. Blue line is the sum of the true canonical correlations between all three pairs of modalities. Box plots are the distributions of the sum of the canonical correlations between permuted data. Grey band is plotted from the 5th to the 95th percentile of the permuted distribution with the largest mean.

Loading	Name	Loading	Name	Loading	Name	Loading	Name
0 0.56234	Standing height:0.0	0 0.792040	Weight:visit-3:2.0	0 0.297388	Number of older siblings:2.0	0 -0.427817	Number of older siblings:2.0
1 0.490283	Sitting height:visit-3:2.0	1 0.749832	Weight:0.0	1 -0.227442	Pulse rate, automated reading:visit-5:2.0	1 0.280392	Weight:visit-3:2.0
2 0.469301	Sitting height:0.0	2 0.726539	Body mass index (BMI):visit-3:2.0	2 -0.211257	-Time to complete round:0.1	2 -0.277201	Number of full brothers:0.0
3 0.462772	Forced vital capacity (FVC):visit-8:2.1	3 0.713521	Whole body fat mass:0.0	3 -0.210940	Pulse rate, automated reading:visit-6:2.1	3 -0.276475	Number of full brothers:2.0
4 0.456811	Forced vital capacity (FVC):visit-8:2.1	4 0.712567	Leg fat mass (right):0.0	4 0.203554	-Time to complete round:2.2	4 0.276325	Standing height:0.0
5 0.440837	Forced vital capacity (FVC):0.0	5 0.708090	Arm fat mass (left):0.0	5 -0.194494	Pulse rate, automated reading:0.0	5 -0.263257	Number of full sisters:2.0
6 0.428948	Forced vital capacity (FVC):visit-2:0.1	6 0.705615	Arm fat mass (right):0.0	6 0.195250	Body mass index (BMI):visit-3:2.0	6 -0.255807	Number of full sisters:0.0
7 0.425064	Fluid intelligence score:visit-3:2.0	7 0.702595	Waist circumference:visit-3:2.0	7 -0.186255	Year ended full time education:0.0	7 0.249234	Waist circumference:visit-3:2.0
8 -0.415817	Leg fat percentage (right):0.0	8 0.684852	Trunk fat mass:0.0	8 -0.185307	Impedance of arm (right):0.0	8 0.247179	Hip circumference:visit-3:2.0
9 0.401886	Forced expiratory volume in 1-second (FEV1):0.0	9 0.684616	Body mass index (BMI):0.0	9 -0.180799	-Qualifications:0.0	9 0.236718	Leg fat-free mass (right):0.0
10 0.395792	Forced expiratory volume in 1-second (FEV1):visit-1:0.0	10 0.678466	Waist circumference:0.0	10 -0.179560	Pulse rate, automated reading:visit-2:0.1	10 0.234893	Whole body fat-free mass:0.0
11 0.391646	Forced vital capacity (FVC):visit-9:2.2	11 0.634619	Hip circumference:visit-3:2.0	11 0.175246	-Mean time to correctly identify matches:0.0	11 0.223522	Hand grip strength (left):visit-3:2.0
12 0.384551	Forced expiratory volume in 1-second (FEV1):0.0	12 0.603304	Body fat percentage:0.0	12 0.169346	Body mass index (BMI):0.0	12 0.223247	Weight:0.0
13 0.383780	Fluid intelligence score:0.0	13 0.594826	Hip circumference:0.0	13 0.169074	-Time to complete round:0.2	13 0.212920	Arm fat-free mass (right):0.0
14 0.382577	Forced vital capacity (FVC):visit-3:0.2	14 0.591382	Arm fat percentage (right):0.0	14 0.164832	-Time to complete round:0.0	14 0.204597	Hip circumference:0.0
15 0.382816	Forced expiratory volume in 1-second (FEV1):visit-1:0.0	15 0.591380	Leg fat percentage (right):0.0	15 0.164767	Use of sun/uv protection:0.0	15 0.201855	Forced expiratory volume in 1-second (FEV1):visit-1:0.0
16 -0.354967	Body mass index (BMI):visit-3:2.0	16 0.590713	Arm fat percentage (left):0.0	16 0.164601	Number of symbol digit matches attempted:0.0	16 0.196629	Hand grip strength (right):visit-3:2.0
17 -0.346812	Body fat percentage:0.0	17 0.588596	Arm fat-free mass (right):0.0	17 -0.163893	-Time to complete round:0.1	17 0.194544	Hand grip strength (right):0.0
18 -0.345636	Arm fat percentage (right):0.0	18 0.580835	Whole body fat-free mass:0.0	18 0.163370	-Diabetes diagnosed by doctor:2.0	18 0.173976	Forced expiratory volume in 1-second (FEV1):visit-1:0.0
19 -0.337212	Arm fat percentage (left):0.0	19 0.578946	Leg fat-free mass (right):0.0	19 0.161862	Number of full brothers:2.0	19 0.168319	Sitting height:0.0
20 0.333950	Forced expiratory volume in 1-second (FEV1):visit-1:0.0	20 0.567261	Trunk fat percentage:0.0	20 0.158115	Number of full brothers:0.0	20 0.166371	Trunk fat mass:0.0
21 0.330889	Year ended full time education:0.0	21 -0.383315	Impedance of whole body:0.0	21 -0.155954	Age completed full time education:0.0	21 0.166288	Sitting height:visit-3:2.0
22 -0.318509	Time spent watching television (TV):0.0	22 -0.364176	Impedance of arm (left):0.0	22 -0.155013	-Qualifications:2.0	22 0.165104	Forced vital capacity (FVC):visit-7:2.0
23 -0.317750	Body mass index (BMI):0.0	23 -0.352714	Impedance of arm (right):0.0	23 0.153875	Number of symbol digit matches made correctly:0.0	23 0.161218	Waist circumference:0.0
24 0.306103	-Duration to complete alphanumeric path (fall...)	24 -0.332321	Impedance of leg (left):0.0	24 -0.153857	Diastolic blood pressure, automated reading:vi...	24 -0.158580	Time spent outdoors in winter:2.0
25 0.305046	Forced expiratory volume in 1-second (FEV1):vi...	25 -0.330020	Impedance of leg (right):0.0	25 0.153091	-Mean time to correctly identify matches:vi...	25 0.157588	Peak expiratory flow (PEF):visit-8:2.1
26 -0.302340	Time spent watching television (TV):2.0	26 0.310199	Sitting height:visit-3:2.0	26 0.152397	-Number of self-reported non-cancer illnesses:...	26 0.157337	Forced expiratory volume in 1-second (FEV1):vi...
27 0.287138	Age completed full time education:2.0	27 0.306784	Sitting height:0.0	27 -0.151690	Impedance of arm (left):0.0	27 0.156683	Body mass index (BMI):visit-3:2.0
28 -0.286509	Trunk fat percentage:0.0	28 0.261758	Standing height:0.0	28 0.151242	Number of live births:0.0	28 0.153850	-Time to complete round:0.0
29 0.283528	Age completed full time education:0.0	29 0.264200	-Snoring:0.0	29 0.149329	Hand grip strength (left):visit-3:2.0	29 0.152249	Age completed full time education:0.0

Figure 5.14: Top 30 SM canonical loadings for the first four sets in multi-view CCA. From top left to bottom right are the first to the fourth set respectively.

Loading	Name	Loading	Name	Loading	Name	Loading	Name
0 -0.625298	Volumetric scaling from T1 head image to stand...	0 0.463991	Volume of peripheral cortical grey matter	0 0.418035	Volume of grey matter (normalised for head size)	0 -0.282118	Volume of grey matter in Cingulate Gyrus, ante...
1 0.604489	Volume of brain, grey+white	1 -0.446778	Volumetric scaling from T1 head image to stand...	1 0.401051	Volume of peripheral cortical grey matter (nor...	1 -0.269629	Volume of grey matter in Cingulate Gyrus, ante...
2 0.604179	Volume of grey matter	2 0.428408	Volume of grey matter in Precuneous Cortex (r...	2 -0.396015	Volume of ventricular cerebrospinal fluid (nor...	2 -0.216558	Volume of grey matter in Postcentral Gyrus (r...
3 0.576879	Volume of peripheral cortical grey matter	3 0.418830	Volume of grey matter in Precuneous Cortex (l...	3 -0.372338	Mean ISQVF in superior cerebellar peduncle on ...	3 -0.210780	Volume of grey matter in Middle Temporal Gyrus...
4 0.544894	Volume of white matter	4 0.418658	Volume of brain, grey+white	4 0.371668	Mean MD in body of corpus callosum on FA skele...	4 -0.202350	Volume of grey matter in Supramarginal Gyrus, ...
5 0.502559	Volume of thalamus (right)	5 0.404981	Volume of grey matter	5 -0.368929	Mean MD in superior cerebellar peduncle on FA ...	5 -0.199046	Volume of grey matter in Cingulate Gyrus, post...
6 0.482941	Volume of thalamus (left)	6 0.394891	Volume of grey matter in Subcallosal Cortex (l...	6 -0.363514	Volume of ventricular cerebrospinal fluid	6 -0.195700	Volume of grey matter in Postcentral Gyrus (l...
7 0.466265	Volume of grey matter in Insular Cortex (right)	7 0.391068	Volume of grey matter in Subcallosal Cortex (r...	7 0.359662	Volume of brain, grey+white, normalised for hea...	7 -0.195616	Volume of grey matter in Cuneal Context (left)
8 0.463756	Volume of grey matter in Frontal Pole (right)	8 0.389233	Volume of white matter	8 -0.354915	Mean MD in superior cerebellar peduncle on FA ...	8 -0.191322	Volume of grey matter in Central Opercular Cor...
9 0.458732	Volume of grey matter in Subcallosal Cortex (l...	9 0.376007	Volume of grey matter in Frontal Pole (right)	9 -0.347737	Mean ISQVF in superior cerebellar peduncle on ...	9 0.190233	Volume of grey matter in Vermis Vitta/Cerebellum
10 0.446322	Volume of grey matter in Frontal Pole (left)	10 0.371043	Volume of grey matter in Middle Frontal Gyrus ...	10 0.346277	Mean FA in anterior corona radiata on FA skele...	10 0.189221	Mean OD in splenium of corpus callosum on FA s...
11 0.445057	Volume of grey matter in Subcallosal Cortex (r...	11 0.369428	Volume of grey matter in Frontal Pole (left)	11 0.340845	Volume of grey matter in Occipital Pole (left)	11 -0.188648	Median T2star in hippocampus (right)
12 0.445053	Volume of grey matter in Insular Cortex (left)	12 0.368438	Volume of grey matter in Cingulate Gyrus, post...	12 -0.339112	Mean OD in body of corpus callosum on FA skeleton	12 -0.188071	Volume of grey matter in Middle Frontal Gyrus ...
13 0.426478	Volume of grey matter in Amygdala (right)	13 0.365493	Volume of grey matter in Cingulate Gyrus, post...	13 -0.338351	Mean ISQVF in fornix on FA skeleton	13 -0.184543	Median T2star in hippocampus (left)
14 0.425132	Volume of grey matter in Central Opercular Cor...	14 0.349959	Volume of grey matter in Middle Frontal Gyrus (...	14 -0.338403	Mean L3 in fornix on FA skeleton	14 -0.181360	Weighted-mean L1 in tract posterior thalamic r...
15 0.424314	Volume of grey matter in Temporal Fusiform Cor...	15 0.333784	Volume of grey matter in Supracalcarine Cortex...	15 0.331307	Mean FA in fornix on FA skeleton	15 -0.180702	Mean OD in splenium of corpus callosum on FA s...
16 0.416261	Volume of grey matter in Central Opercular Cor...	16 0.330787	Volume of grey matter in Paracingulate Gyrus (...	16 0.327547	Mean FA in anterior corona radiata on FA skele...	16 -0.180045	Volume of grey matter in Inferior Temporal Gyr...
17 0.412707	Volume of grey matter in Lingual Gyrus (right)	17 0.322977	Volume of grey matter in Lateral Occipital Cor...	17 -0.321778	Mean L1 in fornix on FA skeleton	17 -0.177061	Mean L1 in cingulum hippocampus on FA skeleton...
18 0.409452	Volume of grey matter in Temporal Fusiform Cor...	18 0.320547	Volume of grey matter in Central Opercular Cor...	18 -0.315844	Mean L1 in superior cerebellar peduncle on FA ...	18 -0.176469	Volume of grey matter in Inferior Temporal Gyr...
19 0.402699	Volume of brain stem + 4th ventricle	19 0.316168	Volume of grey matter in Frontal Orbital Corte...	19 -0.303773	Mean L3 in superior fronto-occipital fasciculu...	19 -0.175899	Volume of grey matter in Cingulate Gyrus, post...
20 0.402070	Volume of grey matter in Amygdala (left)	20 0.309143	Volume of grey matter in Supracalcarine Cortex...	20 -0.299423	Mean L1 in superior cerebellar peduncle on FA ...	20 -0.173055	Volume of white matter
21 0.399000	Volume of putamen (left)	21 0.303990	Volume of grey matter in Central Opercular Cor...	21 0.293588	Mean FA in superior fronto-occipital fasciculu...	21 -0.172363	Mean MO in superior corona radiata on FA skele...
22 0.398949	Volume of grey matter in Precuneous Cortex (l...	22 0.303038	Volume of grey matter in Insular Cortex (left)	22 -0.294294	Mean L2 in external capsule on FA skeleton (left)	22 -0.171629	Volume of grey matter in Supramarginal Gyrus, ...
23 0.397837	Volume of grey matter in Precuneous Cortex (left)	23 0.302134	Volume of grey matter in Cuneal Context (right)	23 0.292736	Volume of grey matter in Occipital Pole (right)	23 -0.167670	Volume of peripheral cortical grey matter
24 0.393366	Volume of grey matter in Frontal Orbital Corte...	24 0.296287	Volume of grey matter in Cingulate Gyrus, ante...	24 0.287982	Mean ISQVF in superior fronto-occipital fascic...	24 -0.167535	Volume of grey matter in Precuneous Cortex (r...
25 0.391860	Volume of putamen (right)	25 0.293388	Volume of grey matter in Paracingulate Gyrus (...	25 -0.282832	Mean L2 in superior cerebellar peduncle on FA ...	25 -0.167460	Volume of grey matter in Cuneal Context (right)
26 0.391319	Volume of grey matter in Cingulate Gyrus, post...	26 0.292768	Volume of grey matter in Temporal Fusiform Cor...	26 -0.279054	Weighted-mean L2 in tract anterior thalamic ra...	26 -0.165178	Mean MD in cingulum hippocampus on FA skeleton...
27 0.387844	Volume of grey matter in Paracingulate Gyrus (...	27 0.290107	Volume of grey matter in Occipital Pole (right)	27 -0.273872	Mean L2 in external capsule on FA skeleton (r...	27 -0.164133	Volume of grey matter in Supracalcarine Cortex...
28 0.386220	Volume of grey matter in Parahippocampal Gyrus...	28 0.287404	Volume of grey matter in Superior Frontal Gyru...	28 -0.278415	Mean L2 in superior fronto-occipital fasciculu...	28 -0.163793	Volume of brain, grey+white
29 0.385298	Volume of grey matter in Cingulate Gyrus, post...	29 0.286584	Volume of grey matter in Cuneal Context (left)	29 -0.277575	Mean OD in anterior corona radiata on FA skele...	29 0.161671	Volume of grey matter in V Cerebellum (left)

Figure 5.15: Top 30 IDP canonical loadings for the first four sets in multi-view CCA. From top left to bottom right are the first to the fourth set respectively.

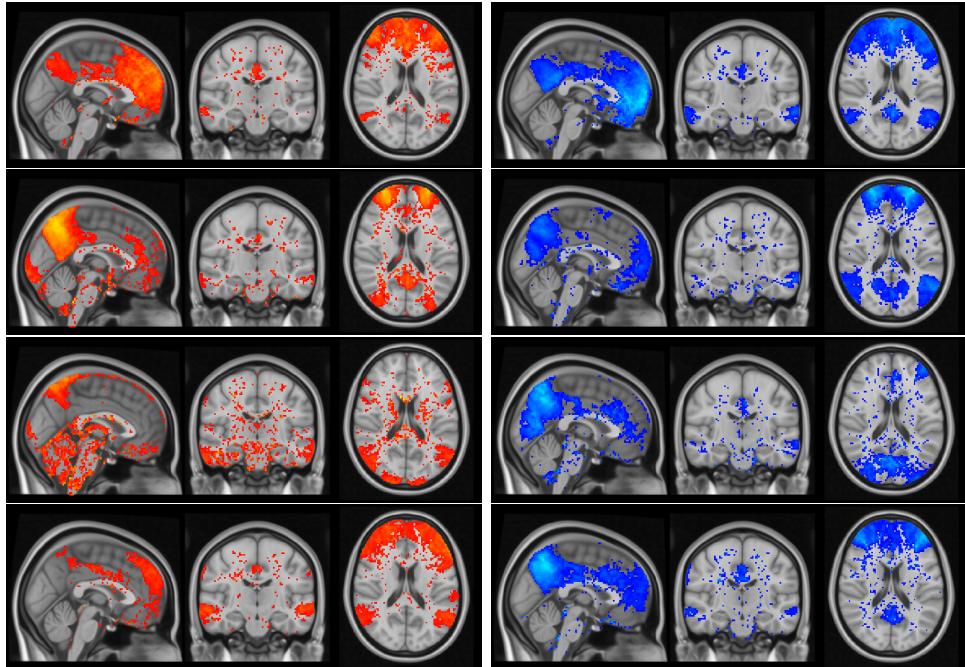


Figure 5.16: Canonical loaded maps for the first four significant FC canonical variables in the multi-view CCA. From top to bottom are the first to the fourth canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).

	Positive map	Negative map
Mode 1	working memory (0.206), task (0.194), default (0.177)	default (0.311), retrieval (0.232), mind (0.209)
Mode 2	default (0.237), retrieval (0.143)	default (0.198), theory of mind (0.184), mental states (0.174)
Mode 3	eye (0.126), spatial/location (0.114)	default (0.278), theory of mind (0.231), mental states (0.192)
Mode 4	sentences (0.239), comprehension (0.236), semantic (0.234), language (0.223)	default (0.219), retrieval (0.162), working memory (0.161)

Table 5.3: Functional interpretation for the positive and negative brain maps in the multi-view CCA (shown in Fig. 5.16). The keywords are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.

and default mode network; for the negative map, it is more focused on prefrontal but also activates in default mode. The third and fourth IDP and SM sets look harder to summarise over these individual variables. For FC, it presents more distinct patterns between positive and negative maps. Particularly, the positive map in mode 3 largely activates in cerebellum and brain stem, which is not shown in all other maps.

We have also noticed that some canonical loadings in the multi-view setting largely overlap with the pairwise results, especially the physical measures in SM, grey matter volume in IDP and default mode network in FC. It also seems for the first mode (first canonical loadings for all three modalities) in MCCA, the SM set of loadings is driven by the relationship with IDP rather than FC since the SM loadings are more similar to the ones in the pairwise CCA with IDP. Similarly, for FC, it is also driven by the relationship with IDP rather than SM.

Overall, the interpretation is not easier for three modalities over individual variable loadings compared with the pairwise results.

5.4.4 Pairwise CCA on SDR-reduced data

Recall that SDR reduces FC from 1485 to 57 dimensions on its 55 sub-domains (ICA regions); SM is reduced from 482 to 107 dimensions on the nine functional sub-domains; IDP is reduced by SDR from 869 to 205 dimensions on the ten functional sub-domains. We first applied pairwise CCA again to the three SDR-reduced modalities, denoted as SDR FC, SDR SM and SDR IDP.

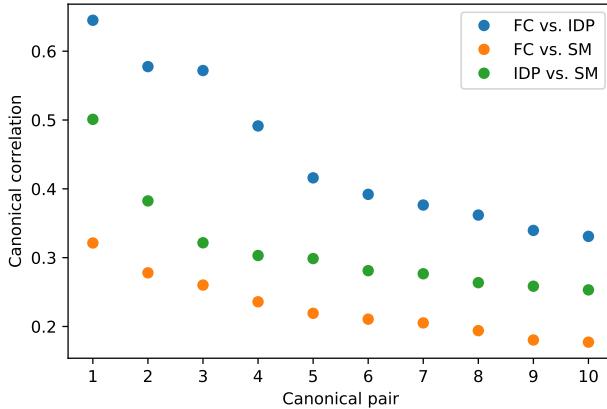


Figure 5.17: Canonical correlation for the first ten canonical pairs in pairwise CCA for SDR-reduced datasets.

Among all three combinations, canonical correlations between FC and IDP remain the strongest despite the fact that FC was reduced with the largest proportion. The canonical correlations between SM and IDP have overtaken the ones between FC and SM become the second strongest pair, which was not the case on the non-reduced data. This might be due to the dimensionalities they are reduced to, and higher dimensional data may produce stronger correlated canonical variables.

Permutation testing for the CCA between SDR FC and SDR SM gives seven significant canonical pairs (same as the non-reduced case), 8 for the CCA between SDR SM and SDR IDP (7 in the non-reduced case), and 19 for the SDR FC and SDR IDP (31 in the non-reduced case; Fig. 5.19). Fig. 5.18 shows the variance explained by these significant canonical variables in their original datasets. The pattern here is fairly similar to the non-reduced case, FC canonical variables explain least variance, and SM and IDP canonical variables explain roughly similar amount. Interestingly, we notice that FC canonical variables explain more variance in the SDR reduced case (left and right most subplots in Fig. 5.8 and 5.18); SM canonical variables explain less variance than the non-reduced case (left two subplots in Fig. 5.8 and 5.18); IDP canonical variables explain more variance in the CCA with SM (middle subplot in Fig. 5.8 and 5.18), and slightly less variance in the CCA with FC (right subplot in Fig. 5.8 and 5.18) compared with the non-reduced case. These observations are to our surprise considering we feed much lower dimensional data to CCA, especially for FC. We expected less variance been explained in general. This indicates that SDR reduces the noise level in the data and lead CCA to produce more explainable

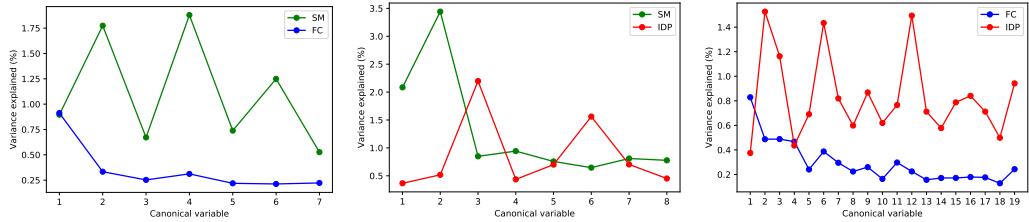


Figure 5.18: Variance explained by the significant SDR canonical variables in their original dataset. From left to right are the CCA between FC and SM (7 pairs), IDP and SM (8 pairs), IDP and FC (19 pairs).

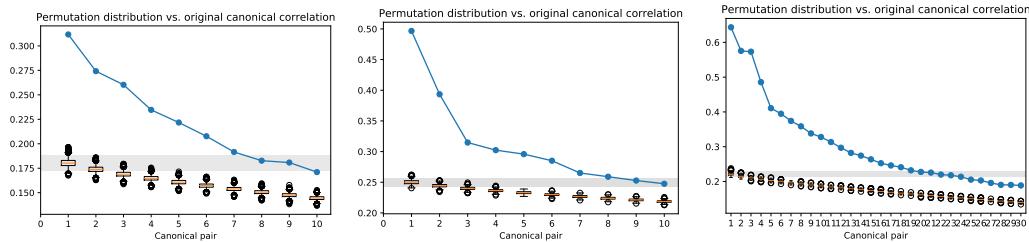


Figure 5.19: Permutation testing on the SDR reduced data for 1000 permutes. Observed canonical correlations (blue line) versus the distribution of canonical correlation between the permuted canonical pairs (box plot). The grey shaded area is the 5 to 95 percentile from the distribution of first permuted canonical pair (first box plot) which is used to define the significance of the canonical pairs (canonical correlation falls under the upper bound of this band is defined as insignificant). From left to right are the CCA between FC and SM, IDP and SM, and FC and IDP respectively.

canonical variables.

5.4.4.1 Canonical loadings

Previously in the non-reduced case, canonical loadings are correlations between the canonical variables and the CCA inputs, which are the observed data. In the SDR reduced case, the inputs of CCA are latent factors from all sub-domains. Therefore, canonical loadings here would represent the importance of the latent components, reflecting the contribution from each sub-domain.

CCA between SDR FC and SDR SM. Fig. 5.20 is an example figure showing the first three sets of significant SM canonical loadings. As shown in the figure, the loadings are not assigned to the observed variables anymore, instead they

are higher-level importance of the sub-domains. Therefore, we have better ideas of the constitution of each set of loadings. Unlike the non-reduced case, the top 30 loadings in the first set (first subplot in Fig. 5.9) are mainly physical measures. In the SDR reduced analysis, we are able to see a more interesting pattern with fewer loadings (first subplot in Fig. 5.20), a combination of ‘Physical Measures’, ‘Exercise & Work’ and ‘Lifestyle & Environment’. Moreover, we can summarise all (instead of the top 30) of the positive and negative loadings by mean squared values as shown in Fig. 5.21. It shows much clearer patterns of the contributions from the sub-domains. It is clear that ‘Physical Measures’ sub-domain plays an important role in almost all significant sets.

To further explore the details of the latent components in the sub-domains, we can observe the rotated loadings, i.e. loadings to construct the latent components during SDR. First seven rotated loadings for the ‘Physical Measures’ sub-domain are shown in Fig. 5.22 as an example. The latent component details of other sub-domains are in Appendix B.1. They are plotted based on the benchmark rotated loadings introduced in Section 5.3.1.2. For further convenience, we summarise the meaning of each latent component in all sub-domains and display them in Tab. 5.4.

By making use of Tab. 5.4, we are able to conclude that, for example, *Physical measure 5* is in contrast with *Physical measure 3* in the first set of canonical loadings (Fig. 5.20). It is in fact Spirometry measures against Body fat-free mass.

Table 5.4: Summary of SM sub-domains. The factors are orthogonally rotated principal components and ordered by R-squared values in the original sub-domain. Second column shows the factor names summarised from figures like Fig. 5.22.

Sub-domain Factors	Factor Summary
Mental Health 1	Depressed, disinterest, fed-up, restless, miserable and lonely feelings (imaging visit)
Mental Health 2	Depressed, disinterest, fed-up, restless, miserable and lonely feelings (initial visit)
Mental Health 3	Worrisome, Sensitivity, guilty and anxious feeling
Mental Health 4	Irritability and mood swing
Mental Health 5	Doctor visit for depression, anxiety or tension
Mental Health 6	Nervous feeling and tense
Mental Health 7	Life satisfaction
Mental Health 8	Able to confide, loneliness, family and friendship satisfaction
Mental Health 9	Risk taking and work satisfaction
Health & Medical History 1	Hospital recorded episodes
Health & Medical History 2	Medication, treatment, long-term illness
Health & Medical History 3	Overall health, falls, other serious condition against wear glasses/lenses
Health & Medical History 4	Self-report and diagnosed cancer

Continued on next page

Table 5.4 – continued from previous page

Sub-domain Factors	Factor Summary
Health & Medical History 5	Chest and lung problem
Health & Medical History 6	When operation took place (imaging visit)
Health & Medical History 7	When operation took place (initial visit)
Health & Medical History 8	Self-reported and other major operations
Health & Medical History 9	Hearing problem
Health & Medical History 10	When non-cancer illness diagnosed (imaging visit)
Health & Medical History 11	When non-cancer illness diagnosed (initial visit)
Health & Medical History 12	Age started wearing glasses/lenses
Alcohol Use 1	Intake frequency against drinker status
Alcohol Use 2	Intake versus 10 years ago and reason for reducing
Alcohol Use 3	Red wine intake
Alcohol Use 4	Alcohol taken with meals
Alcohol Use 5	Intake versus 10 years ago and reason for reducing (imaging visit)
Alcohol Use 6	Beer and cider intake
Alcohol Use 7	Champagne and white wine intake
Alcohol Use 8	Spirit intake
Alcohol Use 9	Fortified wine intake
Tobacco Use	Current and past smoking status
Cognition 1	Symbol digit substitution, trail making, pairs matching
Cognition 2	Fluid intelligence
Cognition 3	Time to complete touchscreen questionnaire
Cognition 4	Prospective memory
Cognition 5	Symbol digit substitution?
Lifestyle & Environment 1	Number in household
Lifestyle & Environment 2	Age of first sexual intercourse versus number of sexual partners
Lifestyle & Environment 3	Qualification
Lifestyle & Environment 4	Mother's and father's age at death
Lifestyle & Environment 5	Length of time at current address
Lifestyle & Environment 6	Father still alive
Lifestyle & Environment 7	Number of full brothers
Lifestyle & Environment 8	Number of full sisters
Lifestyle & Environment 9	Illnesses of father
Lifestyle & Environment 10	Father's age at death against mother still alive
Lifestyle & Environment 11	Home location 1
Lifestyle & Environment 12	Own or rent accommodation lived in against towns deprivation index
Lifestyle & Environment 13	Home location 2
Lifestyle & Environment 14	Illnesses of mother
Lifestyle & Environment 15	Time taken for blood phlebotomy, biometric and conclusion station, time taken for verbal Lifestyle & Environment & interview stage
Lifestyle & Environment 16	against private health care
Lifestyle & Environment 17	Frequency of solarium/sunlamp use
Lifestyle & Environment 18	against private health care
Lifestyle & Environment 19	Frequency of friend/family visits
Lifestyle & Environment 20	Use of sun/uv protection
Food & Drink 1	Bilateral oophorectomy,Ever used hormone-replacement therapy (HRT), had menopause
Food & Drink 2	Noisy workplace, loud music exposure
Food & Drink 3	Meat intake
	Veg and fruit intake
	Fish intake

Continued on next page

Table 5.4 – continued from previous page

Sub-domain Factors	Factor Summary
Food & Drink 4	Coffee against tea intake
Food & Drink 5	Water intake
Food & Drink 6	Bread intake
Food & Drink 7	Cereal and fruit intake
Food & Drink 8	Cheese intake
Food & Drink 9	Milk intake
Food & Drink 10	Salt added to food
Food & Drink 11	Hot drink temperature
Food & Drink 12	Variation in diet
Exercise & Work 1	Timer spend outdoors in summer and winter
Exercise & Work 2	Job involves heavy manual and walking/standing
Exercise & Work 3	Frequency of other exercises, number of days/week of vigorous physical activity
Exercise & Work 4	Frequency of walking for pleasure
Exercise & Work 5	Duration of vigorous activity and other exercise
Exercise & Work 6	Unpleasant work place
Exercise & Work 7	Mobile phone use
Exercise & Work 8	Chronotype
Exercise & Work 9	Distance to work and time spent driving
Exercise & Work 10	Light DIY
Exercise & Work 11	Daytime sleeping
Exercise & Work 12	Time spent watching TV
Exercise & Work 13	Age complete education
Exercise & Work 14	Sleep duration
Exercise & Work 15	Stair climbing
Exercise & Work 16	Not snore
Exercise & Work 17	Drive under speed limit against walking pace
Exercise & Work 18	Side of head for mobile phone use
Physical Measure 1	Body fat mass and BMI
Physical Measure 2	Spirometry
Physical Measure 3	Body fat-free mass
Physical Measure 4	Standing and sitting height
Physical Measure 5	Spirometry
Physical Measure 6	Spirometry
Physical Measure 7	Hand grip
Physical Measure 8	Blood pressure
Physical Measure 9	Blood pressure
Physical Measure 10	Pulse rate
Physical Measure 11	Skin colour
Physical Measure 12	Hair colour
Physical Measure 13	Weight
Physical Measure 14	Handedness
Physical Measure 15	Spirometry
Physical Measure 16	Facial ageing
Physical Measure 17	Spirometry
Physical Measure 18	Spirometry
Physical Measure 19	Spirometry

To interpret SDR FC canonical loading, we come up with similar summary statistics as in the non-reduced case and then map them onto brain volumes to

	Loading	Name		Loading	Name		Loading	Name
0	0.449165	Exercise&Work 6	0	-0.403427	Cognition 1	0	0.648665	Physical measure 6
1	-0.350545	Physical measure 5	1	-0.390396	Exercise&Work 2	1	-0.274279	LifeStyle&Environment 3
2	0.346866	Physical measure 3	2	0.355644	Physical measure 1	2	-0.251072	LifeStyle&Environment 4
3	-0.300155	LifeStyle&Environment 2	3	0.343014	Cognition 4	3	0.226149	Cognition 4
4	-0.296628	Exercise&Work 3	4	-0.292404	Physical measure 2	4	-0.219989	Exercise&Work 2
5	0.217453	Food&Drink 5	5	-0.290523	Food&Drink 4	5	-0.198834	Exercise&Work 1
6	-0.212531	Exercise&Work 19	6	0.287988	Alcohol 1	6	0.180154	Cognition 5
7	0.177649	Physical measure 1	7	-0.273570	Physical measure 6	7	0.175956	Food&Drink 1
8	0.175362	LifeStyle&Environment 3	8	-0.261749	LifeStyle&Environment 1	8	0.167416	Exercise&Work 11
9	0.174378	MentalHealth 2	9	0.256982	LifeStyle&Environment 5	9	0.161437	MentalHealth 4
10	-0.167564	Cognition 5	10	-0.235481	LifeStyle&Environment 3	10	-0.139063	Food&Drink 4
11	-0.156440	Alcohol 2	11	-0.224812	Physical measure 3	11	0.135887	MentalHealth 5
12	-0.154549	Food&Drink 7	12	-0.218299	LifeStyle&Environment 9	12	-0.135136	Food&Drink 3
13	-0.151986	Exercise&Work 10	13	-0.202280	LifeStyle&Environment 2	13	-0.134819	Health&MedicalHist 7
14	0.147594	LifeStyle&Environment 5	14	0.195669	Food&Drink 1	14	-0.133800	Exercise&Work 9
15	0.146861	Exercise&Work 8	15	0.192821	Exercise&Work 5	15	0.130591	Exercise&Work 8
16	-0.146765	Physical measure 2	16	-0.191041	Physical measure 11	16	0.124412	Physical measure 9
17	-0.145561	MentalHealth 4	17	-0.190060	Exercise&Work 3	17	0.117649	LifeStyle&Environment 7
18	0.139306	Food&Drink 12	18	-0.183205	Exercise&Work 7	18	0.116155	MentalHealth 3
19	0.136622	Exercise&Work 7	19	0.173966	Physical measure 17	19	-0.113972	Physical measure 11
20	-0.125979	MentalHealth 6	20	-0.153920	Food&Drink 10	20	0.113237	Health&MedicalHist 2

Figure 5.20: Top 20 canonical loadings for the first three significant SM canonical variables in the CCA of SDR FC and SDR SM. The number after the domain name indicates the n th latent component in the sub-domain. Canonical loadings for all seven significant canonical variables can be found in Fig. B.9.

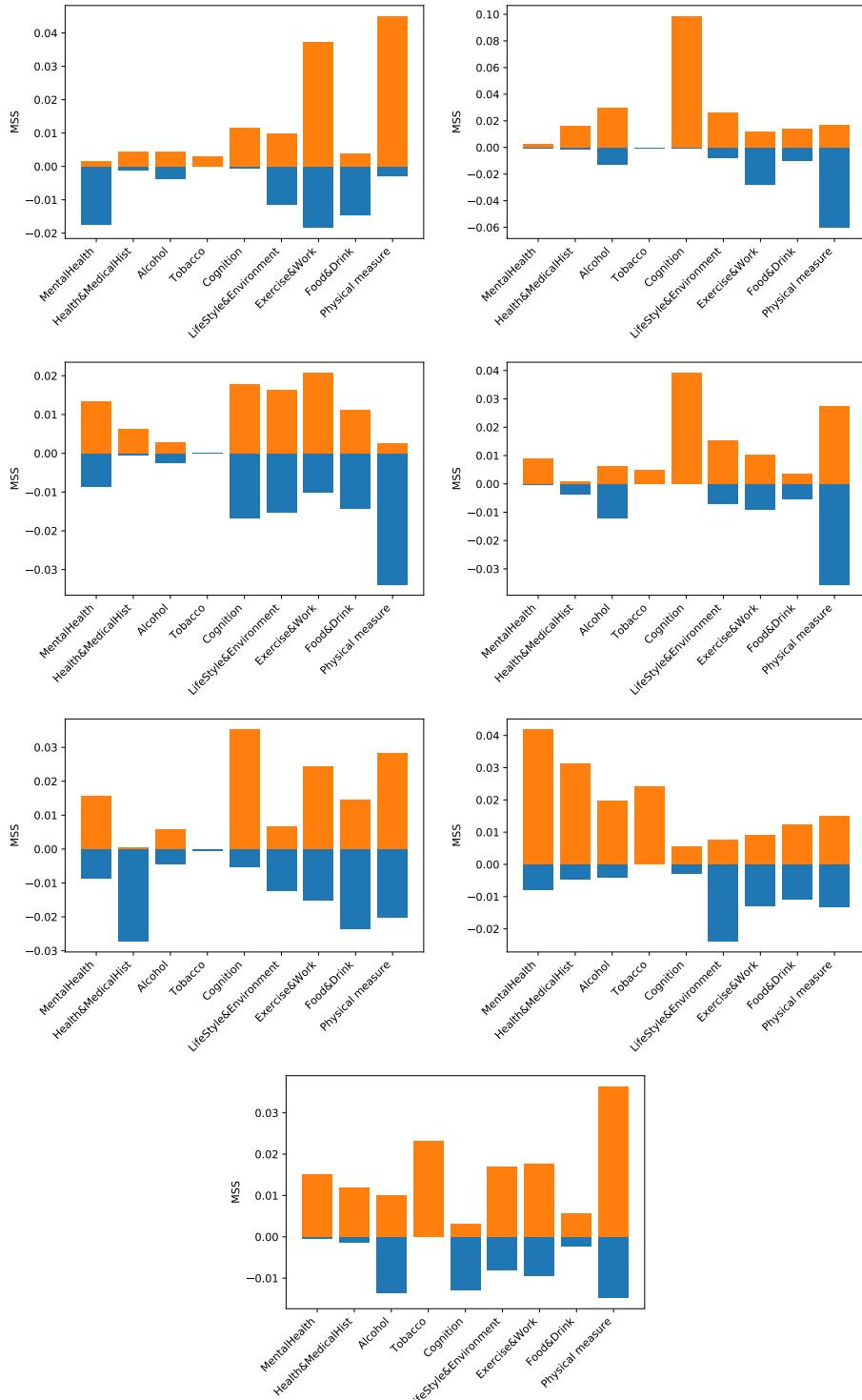


Figure 5.21: Mean squared SM loadings summarised from each sub-domain in the CCA of SDR FC and SDR SM. Blue bars are the mean squared positive loadings and orange bar are the mean squared negative loadings. From top left to bottom right are the first to the seventh set respectively.

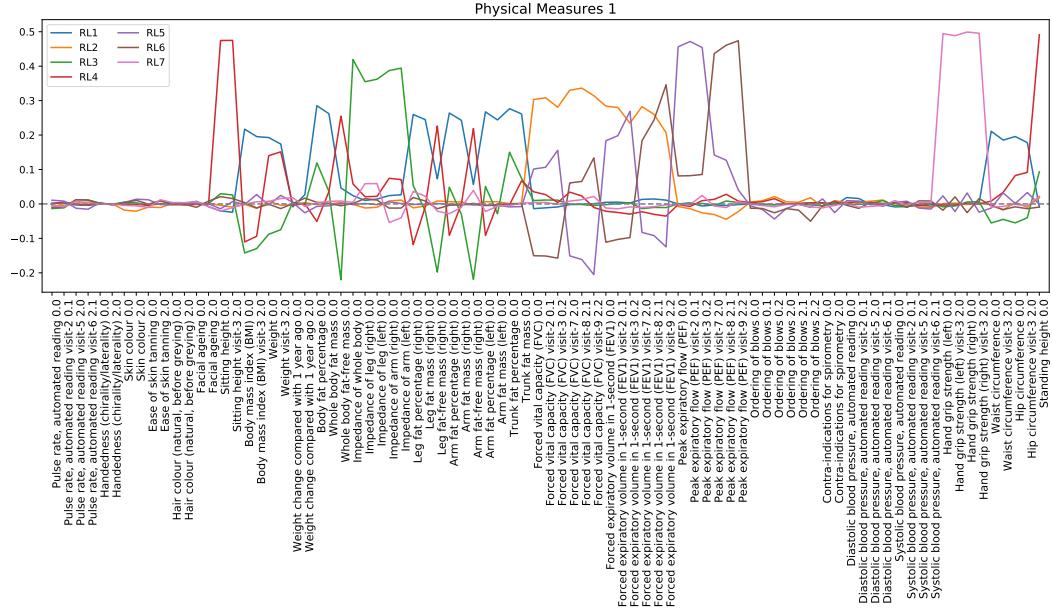


Figure 5.22: First seven latent factor (rotated) loadings from sub-domain Physical Measures. There are in total 20 such rotated loadings in the ‘Physical Measures’ sub-domain.

visualise. Instead of having a canonical loading for each pair of ICA regions, in SDR reduced case, we have one (two for ICA region 5 and 22) latent loading(s) for each ICA region. It provides a more straightforward way to load the brain volume maps. For the regions with two latent loadings (ICA 5 and 22), we take the average of maps.

Now, we observe very distinct patterns between the positive and negative maps in each mode (brain maps for all seven sets of significant canonical loadings are shown in Fig. B.16, with examples for the first three sets shown in Fig. 5.23).

Again, we used NeuroSynth decoding to assist the functional interpretation of all maps (Tab. 5.5). Combining results from both SM and FC sides, in mode 1, the physical measures such as healthy exercise habit and working environment are positively correlated with comprehension, language related areas in the brain; poor mental health and hard physical work are related to the action, sensory related brain regions. In mode 2, positive cognition functions are in line with the visual and motion areas; physical measures are related to the default mode network in the brain. From Fig. 5.21, we notice Tobacco plays a significant role in the 6th and 7th sets of SM loadings, and this may relate to dementia, working memory related brain

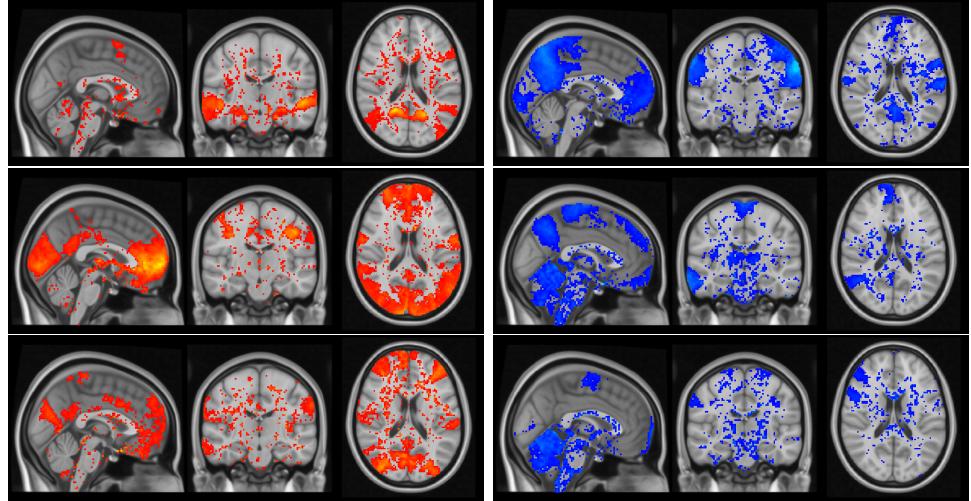


Figure 5.23: Brain volume maps for the first three significant SDR FC canonical loadings in the CCA with SDR SM. From top to bottom are the first to the third canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1). Maps for all seven sets significant canonical loadings can be found in Fig. B.15.

areas as shown in the last two rows of Tab. 5.5.

CCA between SDR IDP and SDR SM. Due to the length of the results, we show the full results in Appendix B.3.2, and present the first set of mean squared loadings in Fig. 5.24. The first canonical relationship is driven by ‘Physical Measures’ and ‘Cognition’ in SM and the volume sub-domains and ‘T2 & Bold’ in IDP. For the rest of the significant sets, SDR IDP loadings in general are more evenly distributed across sub-domains compared with the SDR SM loadings.

CCA between SDR FC and SDR IDP. There are 19 significant canonical pairs between SDR FC and SDR IDP by permutation testing. Due to the large number of results, we focus on the first eight pairs to match the previous two sets of CCAs. Brain maps and mean squared IDP loadings by sub-domain are presented in Appendix B.3.3. In the first eight sets of loadings, the volume sub-domains of IDP (‘Volume’ and ‘Cerebellum volume’) show dominating role, and Fig. 5.25 lists the first two sets as examples. In general, the contribution between sub-domains tend to be more even and weaker as the canonical correlation decreases (Fig. 5.21 and Fig. B.17 to B.19).

	Positive map	Negative map
Mode 1	comprehension (0.335), sentences (0.334), linguistic (0.297), language (0.279)	action (0.268), premotor (0.244), somatosensory (0.24), execution (0.22)
Mode 2	visual (0.261), motion (0.228)	default (0.198), resting (0.164), theory of mind (0.158)
Mode 3	spatial (0.185), default (0.15)	task (0.166), working memory (0.151), coordination (0.147)
Mode 4	visual (0.288), motion (0.205), perception (0.179)	demands (0.255), tasks (0.224), working memory (0.195), semantic (0.187)
Mode 5	episodic memory (0.191), retrieval (0.175)	speech production (0.164), intentions (0.141)
Mode 6	motion (0.217), social (0.163), pain (0.16), fear (0.155)	primary somatosensory (0.168), dorsal premotor (0.167), hand/finger movements (0.167)
Mode 7	dementia (0.2), reward (0.164), fear (0.14)	tasks (0.237), working memory (0.223), motion (0.21), demands (0.195)

Table 5.5: Functional interpretation for the positive and negative brain maps in the CCA between SDR FC and SDR SM (shown in Fig. B.16). The keywords are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.

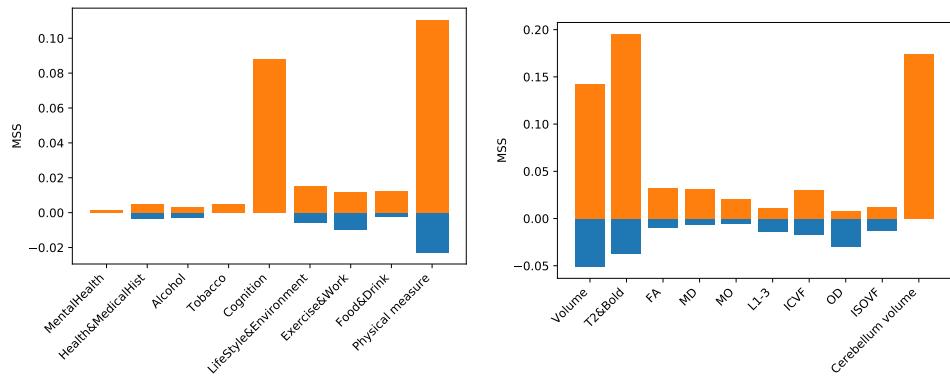


Figure 5.24: First set of Mean squared SDR SM loadings (left) and SDR IDP loadings (right) summarised from each sub-domain. Orange bars are the mean squared positive loadings and blue bars are the mean squared negative loadings. Loading for all significant sets for SDR SM and SDR IDP canonical variables can be found in Fig. B.17 and B.18 respectively.

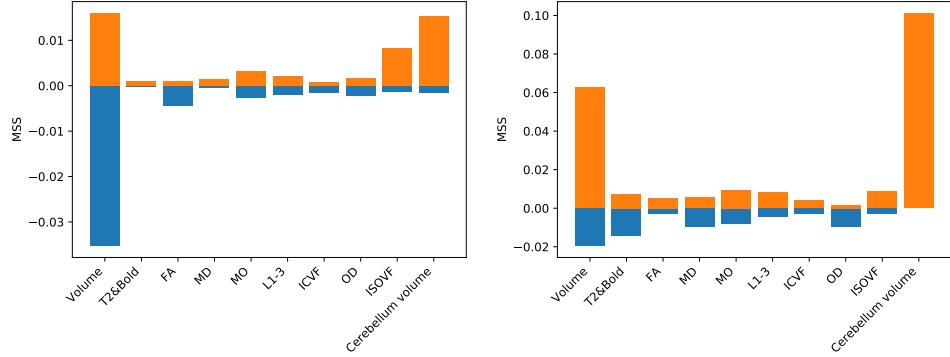


Figure 5.25: Mean squared IDP loadings for the first (left) and second (right) canonical variables summarised from each sub-domain in the CCA of SDR IDP and SDR FC. Orange bars are the positive loadings and blue bars are the negative loadings. The first eight sets of summarised loadings can be found in Fig. B.19.

	Positive map	Negative map
Mode 1	force (0.089)	dementia (0.175)
Mode 2	motor (0.203), tasks (0.154), execution (0.142), coordination (0.134)	early visual (0.098)
Mode 3	social (0.189), dementia (0.174), emotional (0.155)	motor (0.301), movement (0.266), execution (0.219)
Mode 4	auditory (0.343), speech (0.333), acoustic (0.308), listening (0.297), music (0.273)	coordination (0.204), motor (0.203), tasks (0.193), working memory (0.175)
Mode 5	premotor (0.234), motor (0.197), execution (0.164)	visual (0.1)
Mode 6	auditory cortex (0.112)	motion (0.199), vision (0.159), sensorimotor (0.154)
Mode 7	primary motor (0.105), motor cortex (0.099)	auditory (0.253), listening (0.235)
Mode 8	sentence (0.238), semantic (0.218), comprehension (0.215)	speech production (0.235), vocal (0.221), auditory (0.199)

Table 5.6: Functional interpretation for the positive and negative brain maps in the CCA between SDR FC and SDR IDP (shown in Fig. B.20). The keywords are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.

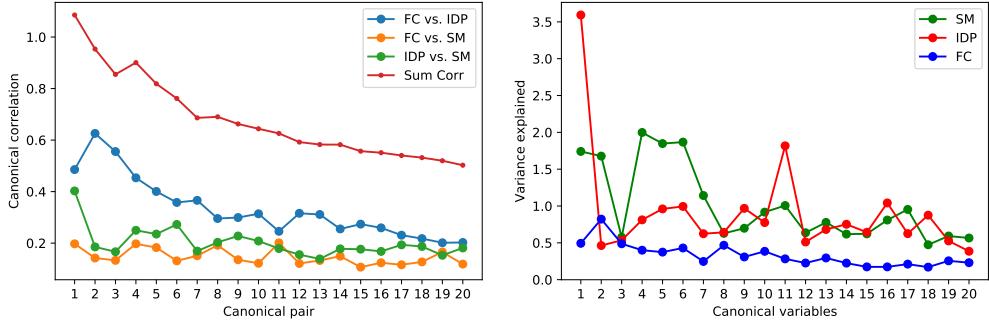


Figure 5.26: Left: top 20 canonical correlation between every pair of the modalities plotted together with the sum of the correlation (red line). Noticing here the sum of the correlation is not monotonic anymore. Right: corresponding variance explained for each modality.

Compared with IDP loadings in the CCA with SDR SM (Fig. B.18), IDP loadings in the CCA with SDR FC appear to be much more focused on ‘Volume’ and ‘Cerebellum volume’ sub-domains. It implicates that the IDP measures driving the canonical relationship with FC are more monotonic and volume based. Volume measures seem to be highly relevant with brain connectivity whereas other measures like diffusion, T2 & Bold are more involved with behavioural and demographic measures.

The related brain functions corresponding to the brain maps shown in Fig. B.20 appear to be a lot weaker as well (Tab. 5.6). In Tab. 5.6, we notice that the default mode network appear much less frequently, and we get more varied keywords compared with the non-reduced CCA case. However, they are all weak functions related to brain volumes.

5.4.5 Multi-view CCA for SDR reduced data

SDR has improved the interpretability from variable level to sub-domain level, making the contributions from sub-domains easier to visualise and summarise. Pairwise CCA helps us to understand the latent structures between pairs of views, however it does not advise the latent structures underpinning all views. In this section, we apply MCCA to further explore such latent structures on SDR reduced data.

Comparing with the canonical correlations in the non-reduced multi-view case (Fig. 5.11), Fig. 5.26 shows that correlations between SDR FC and SDR SM drop considerably (the first pair drops from 0.55 to 0.2), and become the weakest

related pair among the three combinations. This order of correlations is consistent with the pairwise SDR CCA. Notably here the sum of all pairwise correlations is not monotonic anymore (red line in Fig. 5.26), and this is possible as discussed in Section 5.4.3. The variance explained by the SDR FC canonical variables has increased which is line with the non-reduced multi-view case. However, it is hard to conclude for SM and IDP, since the first canonical variable for IDP explains more variance, for SM explains less, however, the rest of the variables may balance out the differences.

We implemented the two permutation tests as illustrated in 5.4.3. 16 non-consecutive pairs of SDR FC and SDR SM canonical variables appear to be significant, with the third pair being insignificant (top plot in Fig. 5.27). 21 canonical pairs for SDR IDP and SDR SM are significant, however, from the third pair, the correlation starts to fluctuate around the significance level (upper bound of the grey band in the bottom plot of Fig. 5.27). SDR FC and SDR IDP have 34 significant multi-view canonical pairs, and they are not consecutive pairs either.

Testing the significance of the sum of the correlation with Permutation test 2 (from 5.4.3) gives better result. Fig. 5.28 shows that there are 20 consecutive sets of significant canonical variables with no fluctuation around boarder line. Furthermore, CV study gives 12 pairs as the best number of canonical variables in the prediction sense.

5.4.5.1 Canonical loadings

Similar to the non-reduced multi-view case, we focus on the interpretation of the first eight sets of significant canonical loadings. All eight sets of loadings for SDR SM, SDR IDP and SDR FC are shown in Fig. B.21, B.22 and B.23 respectively. Tab. 5.7 shows the functional interpretation (by the help of NeuroSynth decoding) for the maps displayed in Fig. B.23. There are some interesting discoveries. For example, for the second mode (Fig. 5.29), it is ‘Physical Measures’ (top left plot in Fig. 5.29) correlated with cerebral volume measures (top right plot in Fig. 5.29), and motor related brain areas (bottom right plot in Fig. 5.29, and the first row in Tab. 5.7), which makes sense to be related from SM, IDP and FC perspectives. Another interesting example is mode 8 which is shown in Fig. 5.30. The SM side of this mode is positively loaded in ‘Alcohol’ and ‘Tobacco’ sub-domains (most variables are sign-flipped, therefore positive loadings mean less alcohol/tobacco intake), and they are positively correlated with motion, speech production related brain areas, and

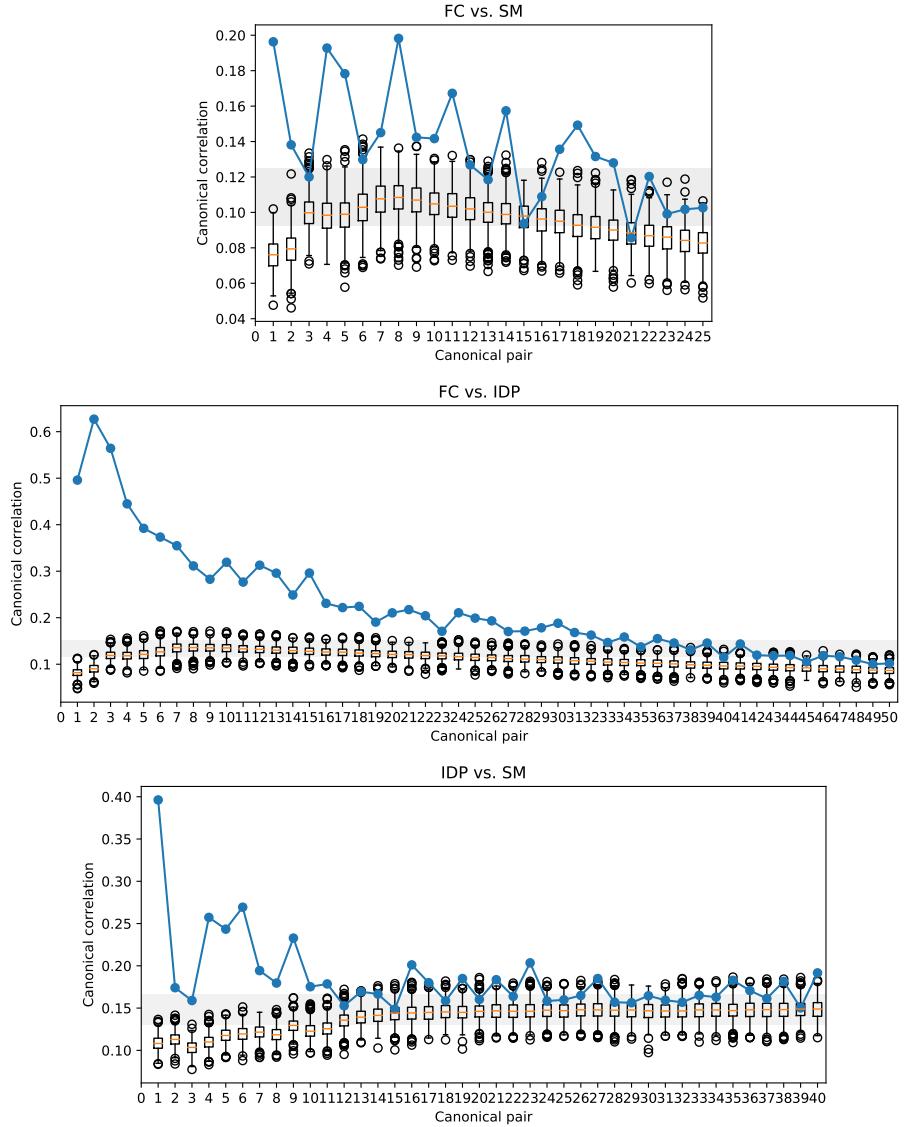


Figure 5.27: The first set of permutation testing: testing the significance of the canonical correlation between individual pairs. Plots on the first two plots are obtained by permuting FC only. Bottom plot is obtained by permuting SM only. Blue lines are the true canonical correlations between FC and SM (top), FC and IDP (middle), and IDP and SM (bottom) and. Box plots are the distributions of the canonical correlation between permuted data. Grey band is plotted from the 5th to the 95th percentile of the permuted distribution with the largest mean.

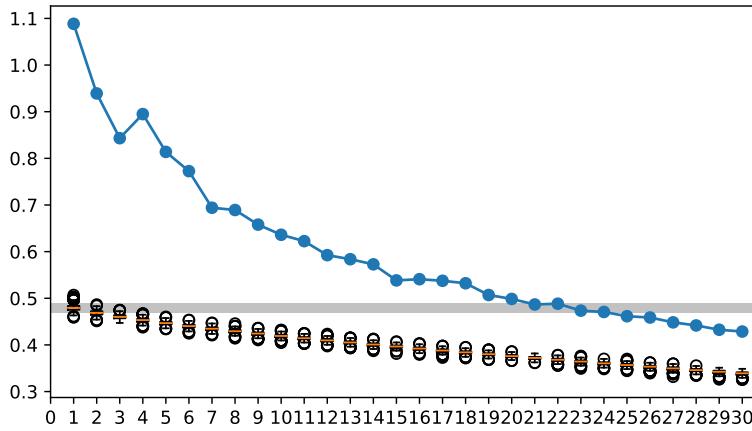


Figure 5.28: The second set of permutation testing on SDR reduced data, testing the significance of the sum of the correlations with FC and SM permuted at the same time. Blue line is the sum of the true canonical correlations between all three pairs of modalities. Box plots are the distributions of the sum of the canonical correlations between permuted data. Grey band is plotted from the 5th to the 95th percentile of the permuted distribution with the largest mean.

negatively correlated with most of the IDP sub-domains, especially ‘T2 & BOLD’ and diffusion sub-domains MD and ISOVF. This may imply the influence of alcohol and tobacco intake on brain functions.

In general, on the SM side, physical measures still dominate many of the top modes, so do the brain volume measures on the IDP side. The 4th to the 8th modes of IDP have fairly balanced contributions from all sub-domains whereas SM canonical loadings are more focused on particular sub-domains. On the FC side, many of the positive/negative maps show very weak brain functions (positive map for mode 1, both maps for mode 2, negative map for mode 6 in Tab. 5.7).

Not every mode gives a clear contrast between all three modalities. Due to the nature of multi-view CCA, each mode would have a dominating pair of modalities whose correlation is stronger than other pairs. We can see from the above results that at a lot of times, the picture presented by one of the modalities is not very specific. However with SDR, the patterns between all three modalities are much easier to summarise, despite the large number of significant modes.

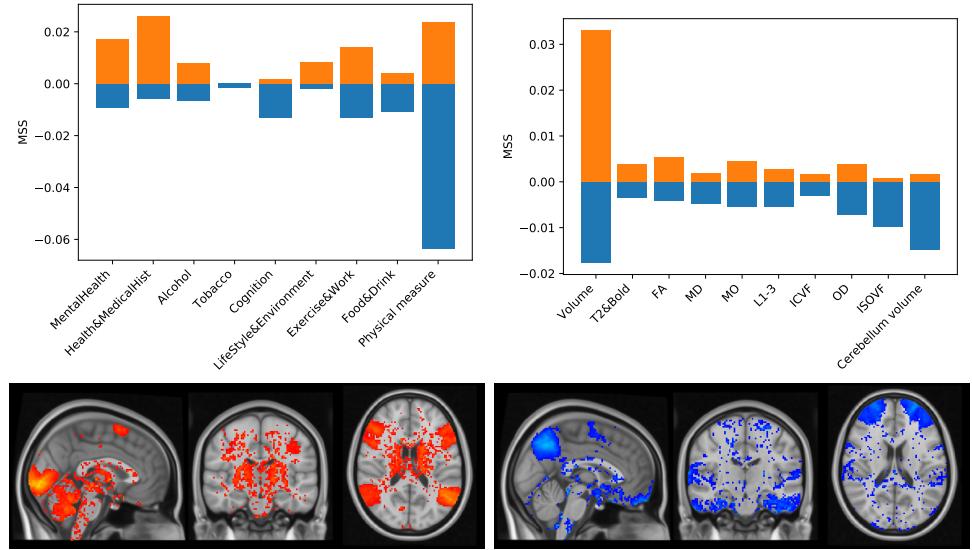


Figure 5.29: The second mode of multi-view SDR CCA. The top two plots are mean squared SM loadings (left) and IDP loadings (right) summarised from each sub-domain. Orange bars are the positive loadings and blue bars are the negative loadings. Bottom plots are positive brain map (left) and negative brain map (right) for the SDR FC loadings.

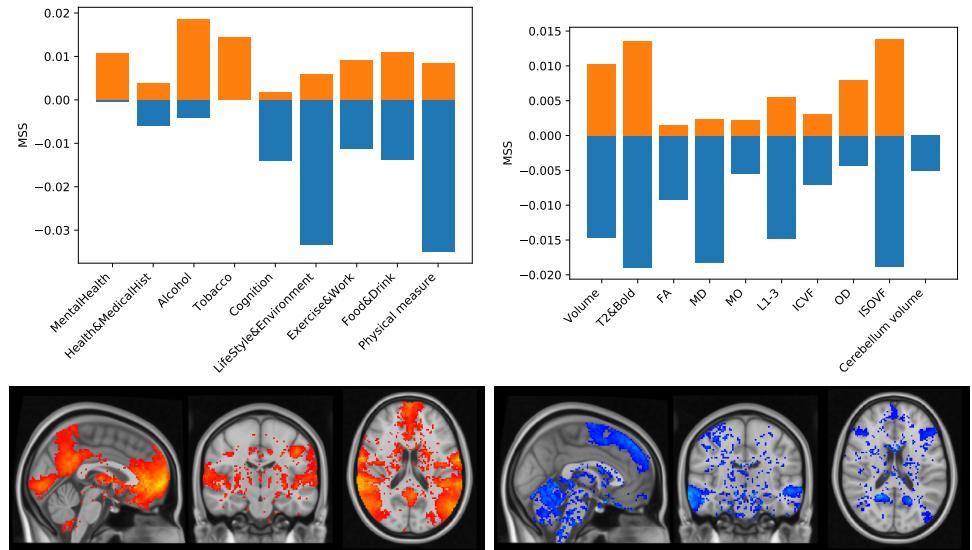


Figure 5.30: The eighth mode of multi-view SDR CCA. The top two plots are mean squared SM loadings (left) and IDP loadings (right) summarised from each sub-domain. Orange bars are the positive loadings and blue bars are the negative loadings. Bottom plots are positive brain map (left) and negative brain map (right) for the SDR FC loadings.

	Positive map	Negative map
Mode 1	dementia (0.091)	visual (0.255), motion (0.187)
Mode 2	primary visual (0.103)	dementia (0.155)
Mode 3	emotional (0.099)	motor (0.409), execution (0.307), sensorimotor (0.284), imagery (0.249)
Mode 4	coordination (0.139), movement (0.132)	listening (0.223), speech (0.223), comprehension (0.19)
Mode 5	auditory (0.295), sound (0.281), listening (0.247), musical (0.203)	premotor (0.177), tasks (0.159), movements (0.151)
Mode 6	speech (0.242), auditory (0.237), listening (0.216), speech perception (0.184)	dementia (0.08)
Mode 7	somatosensory (0.246), primary motor (0.14)	working memory (0.138), arithmetic (0.136)
Mode 8	motion (0.276), speech production (0.232), vocal (0.206), default (0.193)	semantic (0.186), comprehension (0.171), sentence (0.162), retrieval (0.148)

Table 5.7: Functional interpretation for the positive and negative brain maps in the multi-view CCA on SDR SM, SDR FC and SDR IDP (corresponding to the maps shown in Fig. B.23). The key words are obtained by NeuroSynth decoding (only function related keywords with are selected). In brackets are the correlations between the defined functional areas and the brain maps.

5.4.6 Stability study on SDR CCA

We now examine the stability of the results we found in SDR CCA with 10-fold CV procedure introduced in Section 5.3.1.2. Here we only investigate the stability of the latent factors since they are more practical and interpretable compared with the individual variables.

We are only interested in examining the stability of latent (rotated) components (RCs) with large loadings, since small loadings are less stable and likely to be caused by noise. Therefore, we come up with the following visualisation. For each CV fold, we take the top 30 canonical loadings ordered by absolute value and take the ones occurred more than three times out of the 10 folds. To be able to compare the differences across different canonical modes, we apply this procedure to all or the first few significant canonical modes. Finally, we take the union of all RCs appeared in every fold selected by the above steps and calculate their means and standard deviations of canonical loadings in CV.

In figures presented in this section, we define *very stable factors* as factors that occur more than seven out of 10 folds and label them with dark grey bars. The rest of the factors (that occur at least in three out of 10 folds) are labelled with light grey bars.

5.4.6.1 CCA between SDR FC and SDR SM

Stability of the SDR SM and SDR FC canonical loadings in pairwise CCA are shown in Fig. 5.31 and 5.32.

For the seven significant modes of SDR SM in Fig. 5.31, the factors get less and less stable as the mode increases: fewer factors appear for more than seven out of 10 folds (dark grey bars), and the standard deviations (std) of the canonical loadings get larger (blue bars). Physical measures show considerable occurrence and stability across all seven modes; tobacco fails to appear more than seven times in all modes. The behaviours of the first four modes look similar with very stable factors (dark grey bars with small std) loaded on most of the sub-domains. The rest of the modes have much fewer very stable factors. Especially the last mode, almost all factors appear as light grey bars, however, there are only five very stable factors out of the 107 total SDR SM factors. Moreover, the variances of canonical loadings are much larger and often cross the x-axis, which implies that the result of this mode is not stable and generalisable.

Similar observation can be found on the FC side (Fig. 5.32): as the mode

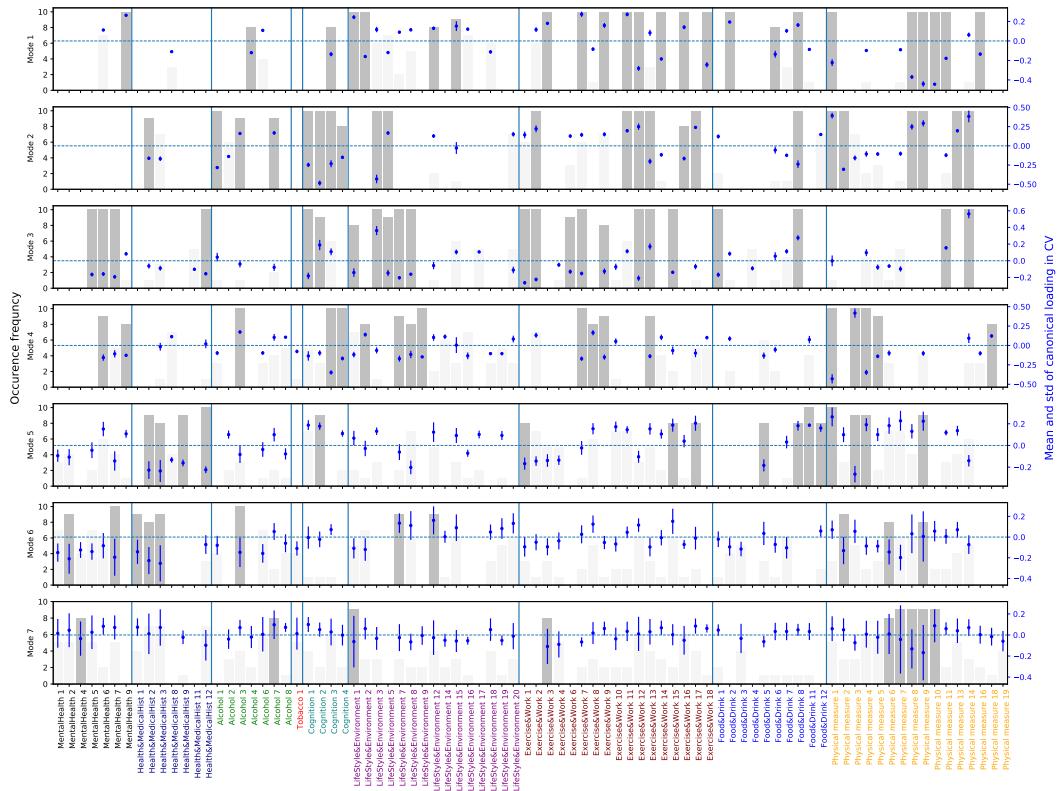


Figure 5.31: Stability of SDR SM canonical loadings for the seven significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than three times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than seven out of 10 times are shadowed with dark grey, the rest is shadowed by light grey.

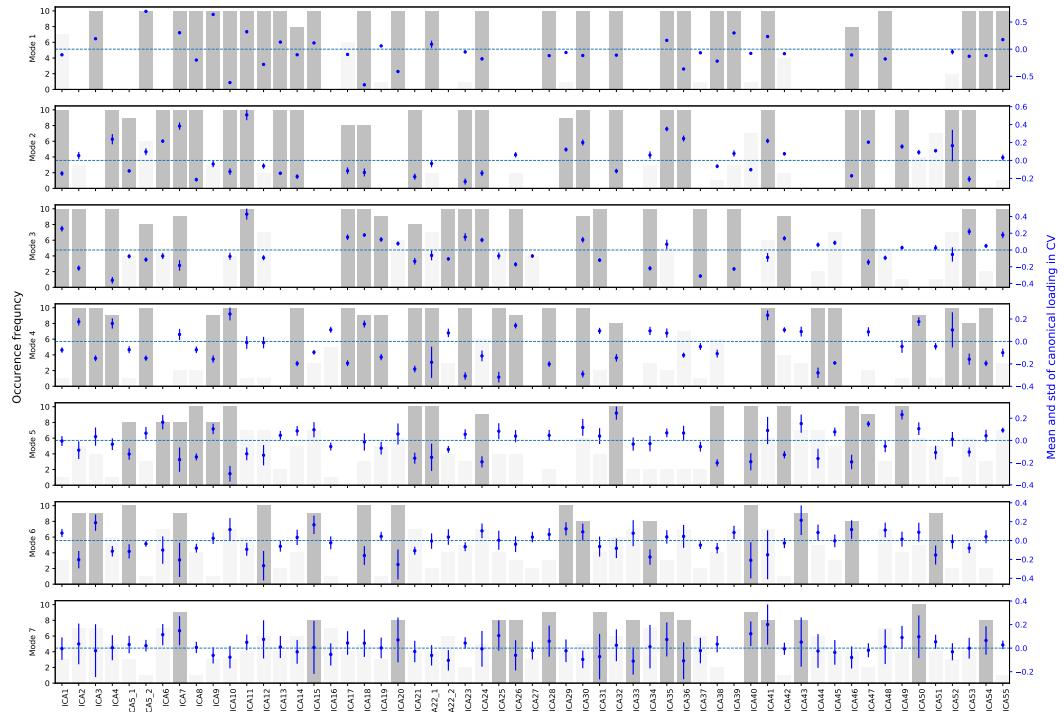


Figure 5.32: Stability of SDR FC canonical loadings for the seven significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than three times out of the 10 folds). The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than seven out of ten times are shadowed with dark grey, the rest is shadowed by light grey.

increases, the factors get less stable. Compared with the last three modes, the first four modes display higher stability and have more very stable factors. Since the ICA regions are not function based, we do not interpret their functionalities.

Although permutation testing gives seven significant canonical pairs in the CCA between SDR SM and SDR FC, the stability of the latent factors drop considerably after four modes. Therefore, we should always be cautious when interpreting results for all significant modes.

5.4.6.2 CCA between SDR IDP and SDR SM

Following similar rules, the stability of the latent factors in the CCA between SDR IDP and SDR SM decreases as the number of mode increases. Moreover, SDR IDP factors show worse stability compared with the SDR SM factors. In particular, we observe much larger std and fewer very stable factors from the 4th mode of SDR IDP. The stability of the SDR SM factors drop gradually and become noticeably less stable from the 5/6th mode. We attach similar figures as shown in the previous section in Fig. B.24 and B.25.

For the first three modes, there are relatively clear patterns presented, and they are consistent with the results shown in Section 5.4.4.1. Especially the first mode shows very stable pattern relating cerebral and cerebellum volume measures with the physical and cognitive measures.

5.4.6.3 CCA between SDR FC and SDR IDP

The latent factors in the CCA between SDR FC and SDR IDP display the strongest stability among all three sets of pairwise CCA. We observe a similar amount of very stable factors for SDR FC across the eight significant CCA modes we investigated (Fig. B.27). The 8th mode of SDR IDP has fewer very stable factors compared with the first seven modes, and all modes on the IDP side show similar patterns: mostly cerebral and cerebellum volume focused (Fig. B.26).

An interesting observation is that for both SDR FC and SDR IDP, mode 2 and 3 show the least stability among the first eight significant CCA modes by having significantly larger standard deviations on the canonical loadings. We do not interpret the SDR FC factors in CV since they are not interpretable if not mapped to brain volumes. However, the CV study shows that the results shown in Section 5.4.4.1 are very stable apart from mode 2 and 3.

5.4.6.4 Multi-view CCA

Based on the permutation testing on the sum of the canonical correlations in the multi-view setting, we get 20 significant canonical sets (Fig. 5.28). We visualise the stability of the first eight canonical sets to match the previous results.

Mode 3, 7 and 8 for SDR SM have significantly less stable factors compared with the rest of the first eight modes with higher variances and less very stable factors (Fig. 5.33). Physical measures still play an important role in every mode. Mode 1 is a strong Cognition and Physical measures mode, with dark grey bar absent in ‘Mental Health’ and ‘Alcohol’ sub-domains; mode 2 shifts the high stability to the ‘Mental Health’ and ‘Health & Medical History’ sub-domains; the patterns in mode 4, 5 and 6 look fairly similar, all missing very stable factors in ‘Mental Health’, and ‘Tobacco’. Mode 6 misses dark grey bars in ‘Alcohol’, and mode 5 is particularly stable in ‘Exercise & Work’ sub-domain. We notice that the only RC in ‘Tobacco’ sub-domain fails to be very stable in all 8 modes.

The patterns of stability across modes on IDP side (Fig. 5.34) are similar to the ones in the pairwise CCA with SDR FC (Fig. B.26). They are very focused on the two volume sub-domains. The variances of canonical loadings of the 7th and the 8th mode are large and with much fewer very stable factors. Mode 1 is still a strong volume mode and the only mode with total absence in L1-L3 and ISOVF sub-domain.

The 4th, 7th and 8th modes in Fig. 5.35, FC canonical loading, show weaker stability. Interestingly, the second mode is the most stable one among the first eight significant modes.

The above stability study suggests that in the multi-view setting, the relationship between SDR IDP and SDR FC dominates the canonical correlation among all three modalities, and this was shown in the one-off analysis as well (Section 5.4.5). With figures like 5.33 to 5.35, we have better ideas on the focus of each CCA mode.

5.4.6.5 Cross-validated canonical correlation

Another aspect to investigate in the stability study is the canonical correlations. We examine the strength of the cross-validated canonical correlations as shown in Fig. 5.2. Recall from Section 2.9 that cross-validated canonical correlations are calculated by applying the training canonical weights to the test data to construct

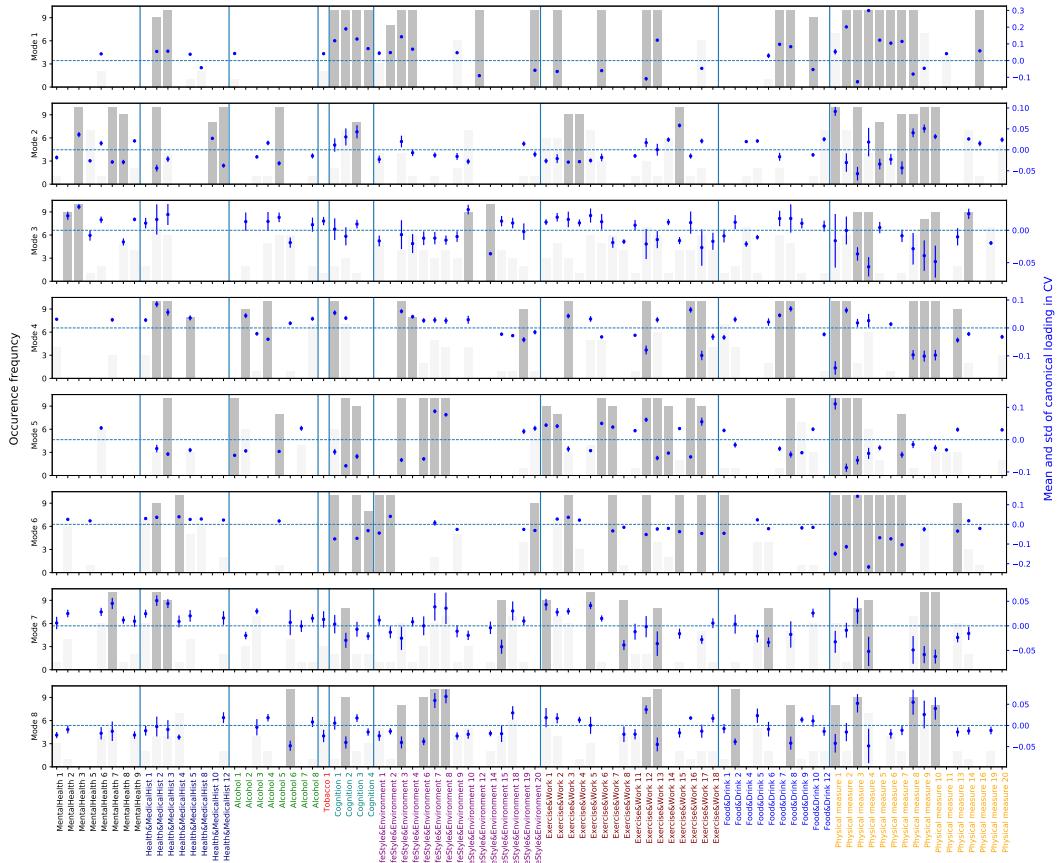


Figure 5.33: Stability of SDR SM canonical loadings in the multi-view CCA of SDR FC, SDR IDP and SDR SM for the first eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than three times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than seven out of ten times are shadowed with dark grey, the rest is shadowed by light grey.

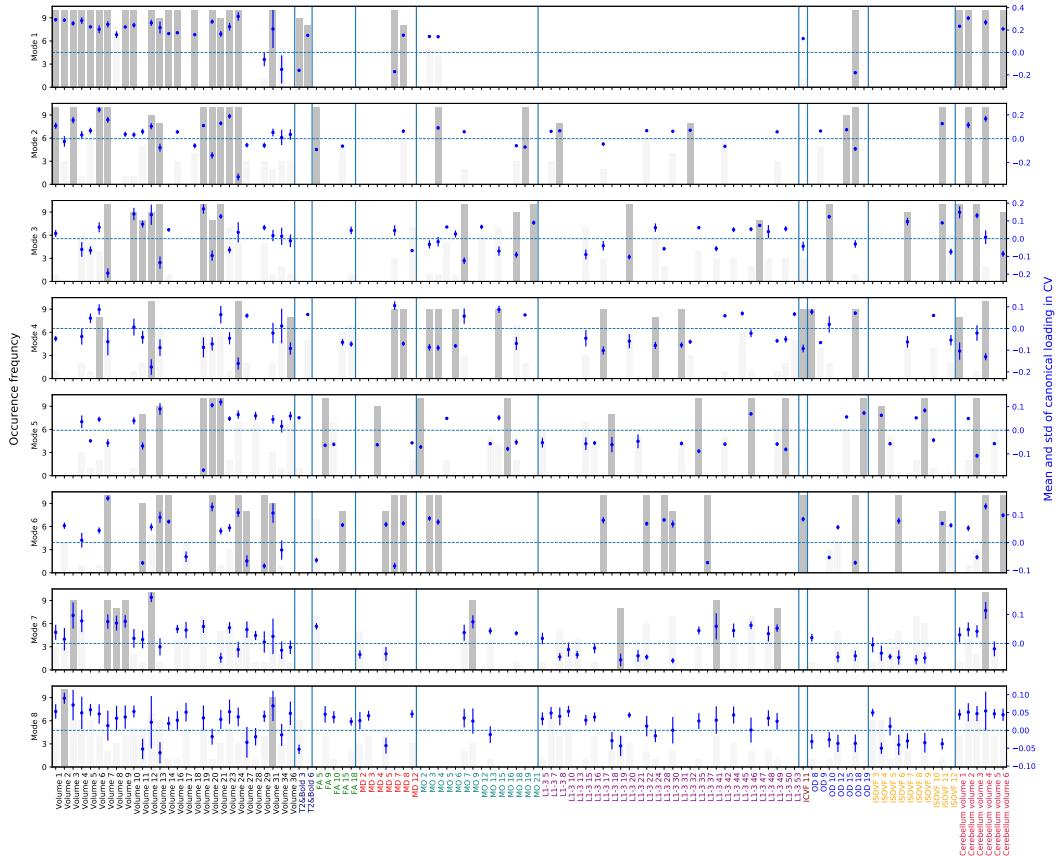


Figure 5.34: Stability of SDR IDP canonical loadings in the multi-view CCA of SDR FC, SDR IDP and SDR SM for the first eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than three times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than seven out of ten times are shadowed with dark grey, the rest is shadowed by light grey.

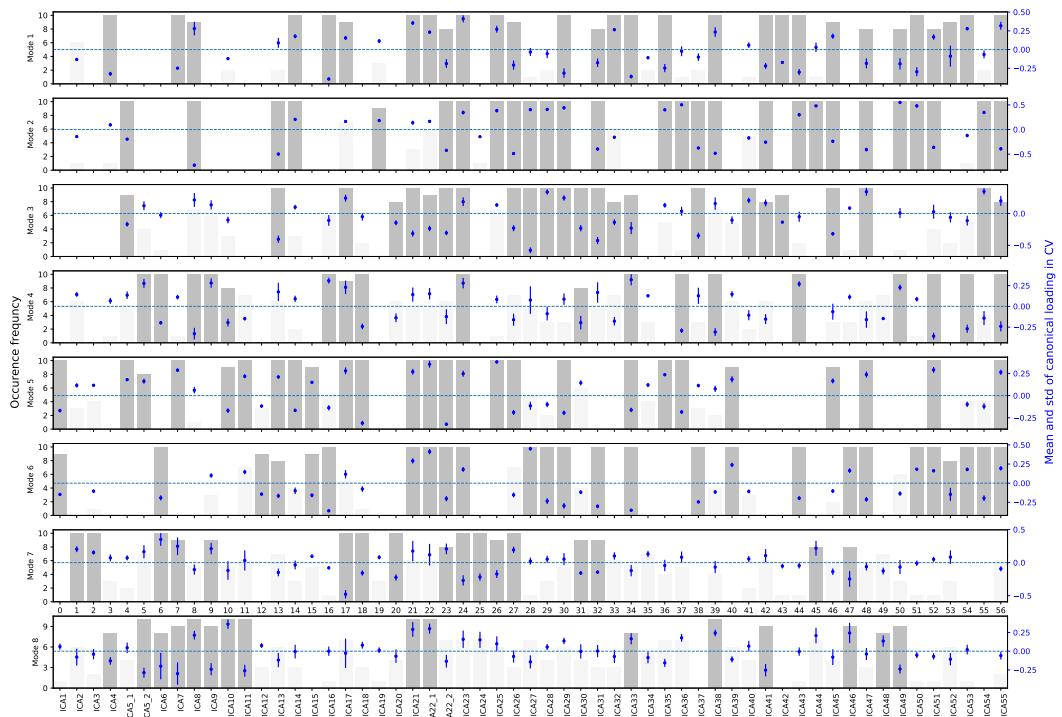


Figure 5.35: Stability of SDR FC canonical loadings in the multi-view CCA of SDR FC, SDR IDP and SDR SM for the first eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than three times out of the 10 folds). The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than seven out of ten times are shadowed with dark grey, the rest is shadowed by light grey.

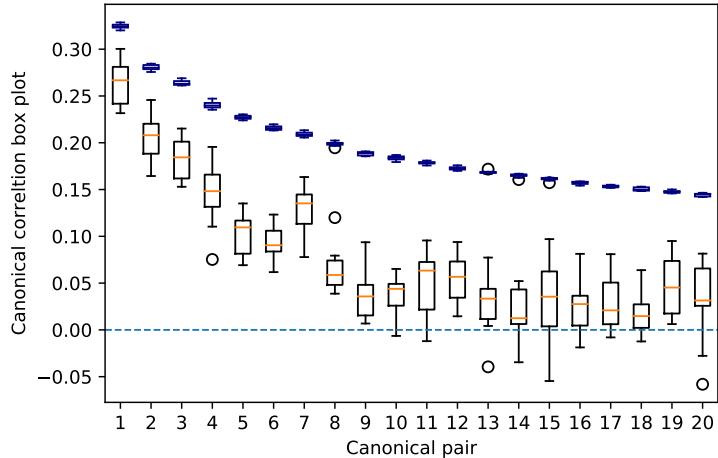


Figure 5.36: Distributions of the training set canonical correlations (blue box plot on the top) compared with the cross-validated canonical correlations (bottom box plot) in the CV study of CCA between SDR FC and SDR SM.

canonical variables, and then calculate the correlations between them. They show the reliability of generalising the canonical relations to unknown data. We compare the distribution of the canonical correlations in the training sets with the distribution of the cross-validated canonical correlations obtained from the 10 folds of CV study. The results for the four sets of CCA (FC vs. SM, IDP vs. SM, FC vs. IDP and multi-view CCA) are shown in Fig. 5.36, 5.37, 5.38 and 5.39. Most of the cross-validated correlations in Fig. 5.36 to 5.39 are significantly larger than 0. In particular, the distributions of the first few sets are comparable with the correlations in the training sets. The positions where the lower whiskers of the cross-validated canonical correlations (black box plot) crosses zero are roughly the same with the number of significant canonical pairs/sets permutation testing identifies. All of the above evidence shows that the results we discovered previously are stable and generalisable.

5.5 GFA Results

We have applied multi-view CCA to explore the latent structures shared by all three modalities. Next, we apply GFA to explore the latent structures shared by each pair of the modalities as well as the by all three modalities in one model (Section 2.5).

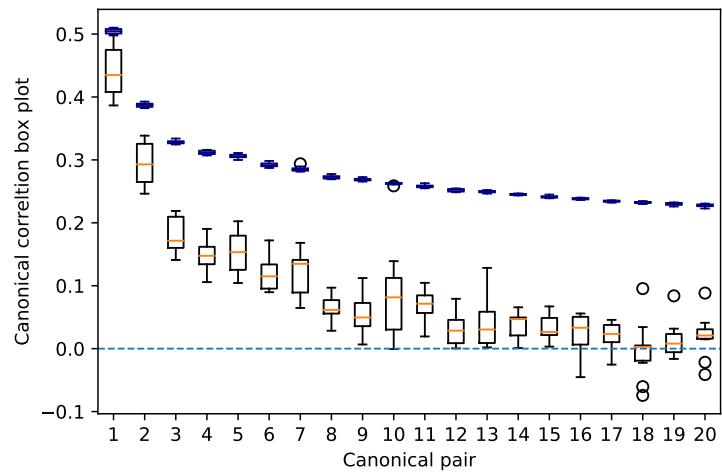


Figure 5.37: Distributions of the training set canonical correlations (blue box plot on the top) compared with the cross-validated canonical correlations (bottom box plot) in the CV study of CCA between SDR IDP and SDR SM.

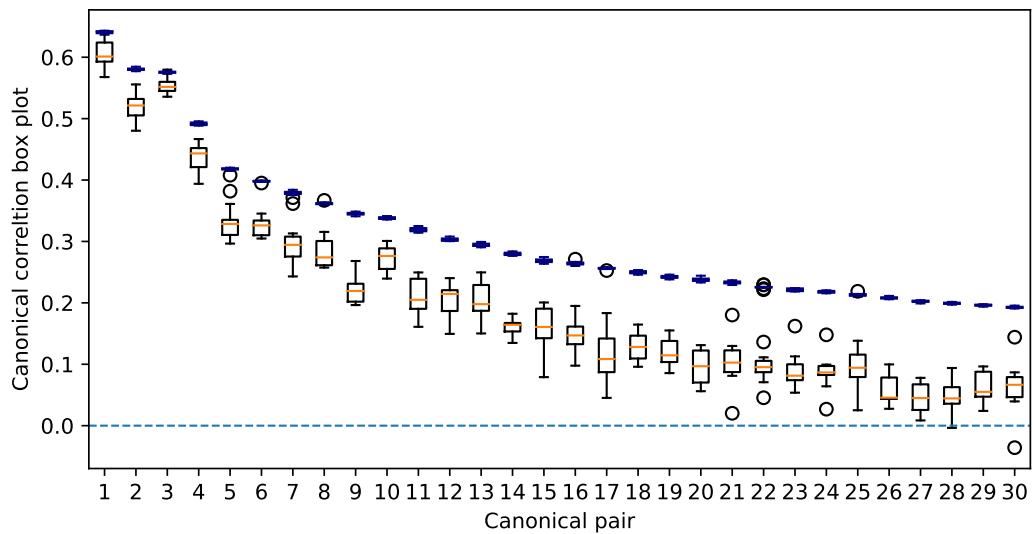


Figure 5.38: Distributions of the training set canonical correlations (blue box plot on the top) compared with the cross-validated canonical correlations (bottom box plot) in the CV study of CCA between SDR FC and SDR IDP.

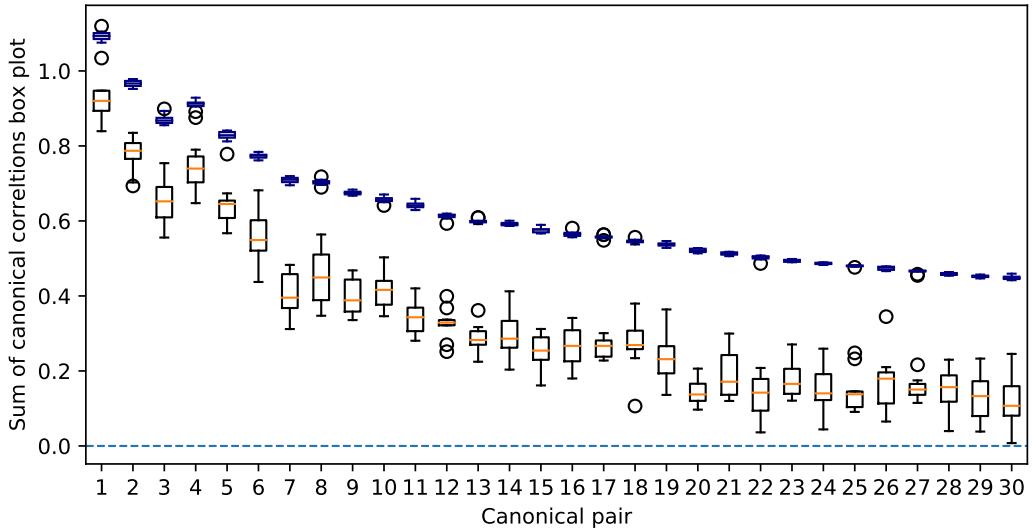


Figure 5.39: Distributions of the training set canonical correlations (blue box plot on the top) compared with the cross-validated canonical correlations (bottom box plot) in the CV study of multi-view CCA between SDR FC, SDR SM and SDR IDP.

We hope to compare the results of GFA with the pairwise and multi-view CCA results to gain more insights on the interplay between these modalities of data.

We attempted applying GFA to the non-reduced and PCA-reduced data and found that it is almost impossible to interpret the results. For the non-reduced case, due to the high dimensionality of the inputs (FC 1485 + SM 481+ IDP 869), the dimensionality of the latent space identified by GFA is very high (over 200). This leads to the loadings matrix (W in Eqn. (2.31)) being very dense, therefore, too difficult to interpret. Moreover, as introduced in Section 2.5, having the latent dimensionality (K in Section 2.5) this high makes the model too slow to be practicable. For PCA-reduced data, GFA identifies a reasonable number of latent factors, however, the loading matrix W are weights on the principal components which are linear combinations of the observed variables. It makes the weight matrix uninterpretable. Therefore, we show results on GFA applied to SDR-reduced SM, IDP and FC.

The input dimensionality of GFA in the SDR-reduced case is 369 (57 SDR FC + 107 SDR SM + 205 SDR IDP). We initialise K as 80 (after several adjustments as instructed in Section 2.5.1) and repeat the experiment for ten times. By selecting the model with the best performance (the one has the lowest lower bound $L(\Theta)$ in Eqn. (2.36)), we get 78 latent factors.

To interpret those factors, we visualise the loading matrix W which is a canonical loading equivalent measure in GFA. For each modality, FC, SM or IDP, GFA gives a $D \times k$ loading matrix, where D is the input dimension for the respective modality (57, 107 or 205) and k is the number of latent factors GFA identifies, which is 78 in this case. Fig. 5.40 shows the loading matrices for all 78 components and three modalities on the SDR latent factors. IDP is loaded on almost all components (Fig. 5.40c) and SM is loaded with the least components (Fig. 5.40b). We also notice that many of the components are modality-specific, i.e. single modality accounts for the whole component. Moreover, from the histograms on the top of the plots in Fig. 5.40, the magnitudes of the loadings are mostly very close to 0. Those components might be caused by noise and therefore, not stable. To focus on the components with significant loadings, we take components with mean absolute values larger than 0.05 across all modalities and show them in Fig. 5.41.

There are 14 components in Fig. 5.41. IDP is still the heaviest loaded modality, and SM is the least loaded modality. Component 2 loads on all three modalities, and is the most loaded component for both FC and SM (Fig. 5.41a and 5.41b). It has large negative loadings on the two volume sub-domains of IDP. It is also negatively loaded on some physical measures (especially Physical measure 4 and 2) and cognition ones. By checking the SDR factor summary table (Tab. 5.4), Physical measure 4 is ‘standing and sitting height’ and Physical measure 2 is a spirometry factor. It suggests that they are positively correlated (both having negative loadings) to the brain volume measures. Component 2 on the FC side (Fig. 5.41a) is loaded on a wide range of ICA regions, which is not surprising since volume can be an important factor to all brain regions. This result is consistent with the CCA finding. Component 5 is the heaviest loaded component for IDP (Fig. 5.41c). It shows contrasts between diffusion, L1-L3 factors and some of the fractional anisotropy ones, and within ICVF and OD sub-domains. This component does not have significant loadings on FC, and lightly loaded on some of the physical and cognition measures of SM.

SDR also enable us to visualise the loadings by sub-domains. However, due to the identifiable issue of factor analysis (signs of the loadings can be flipped without changing the components), we cannot take the mean/sum of all loadings within a sub-domain. We resolved this issue in CCA by fixing the signs of the latent factors at various stages, however in GFA this is not possible. Therefore, we are only able to view the absolute importance of the sub-domains. We first group the factors into their sub-domains (SM has 9 sub-domains; IDP has 10 and FC has 55 sub-

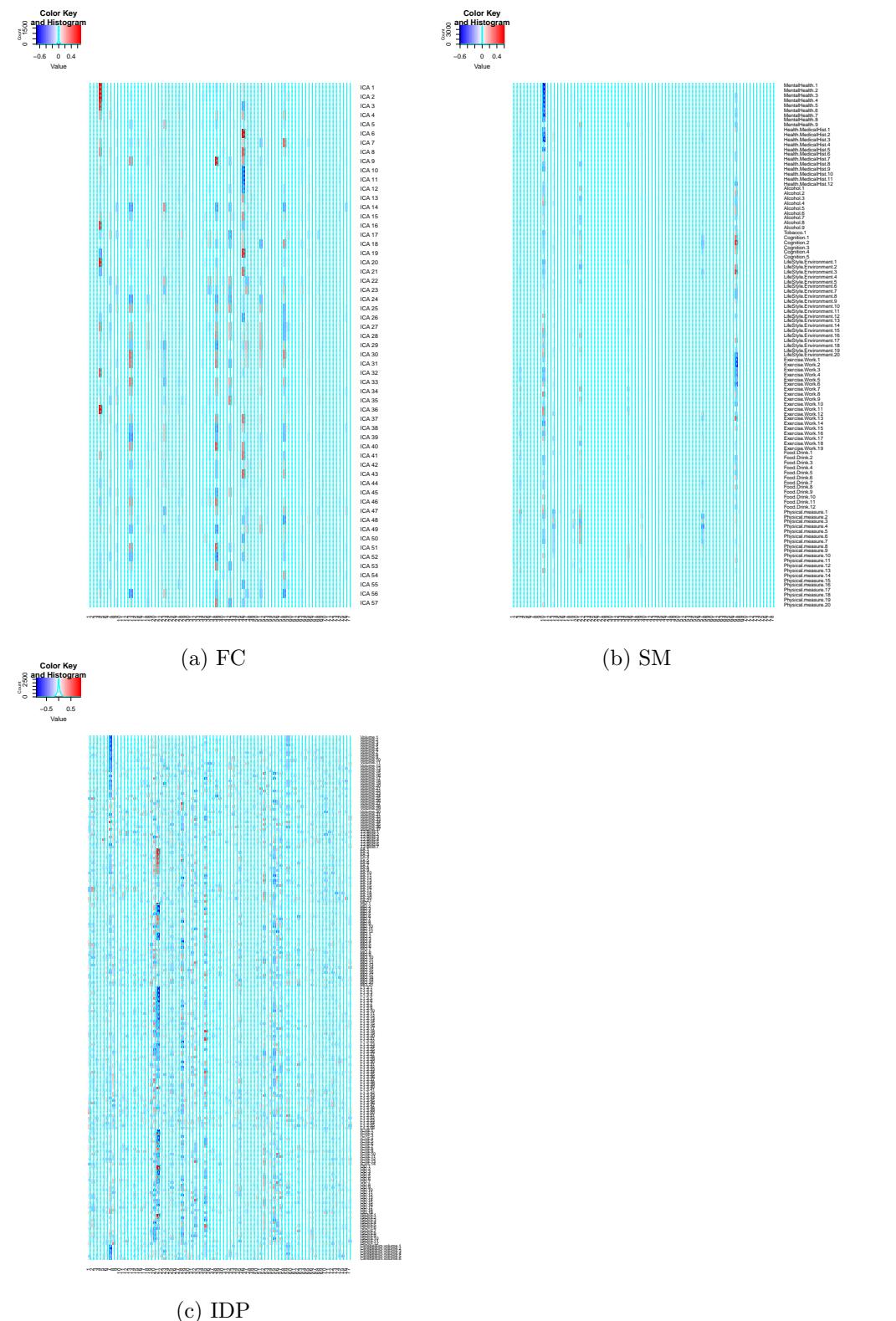


Figure 5.40: Loading matrices for all 78 GFA components.

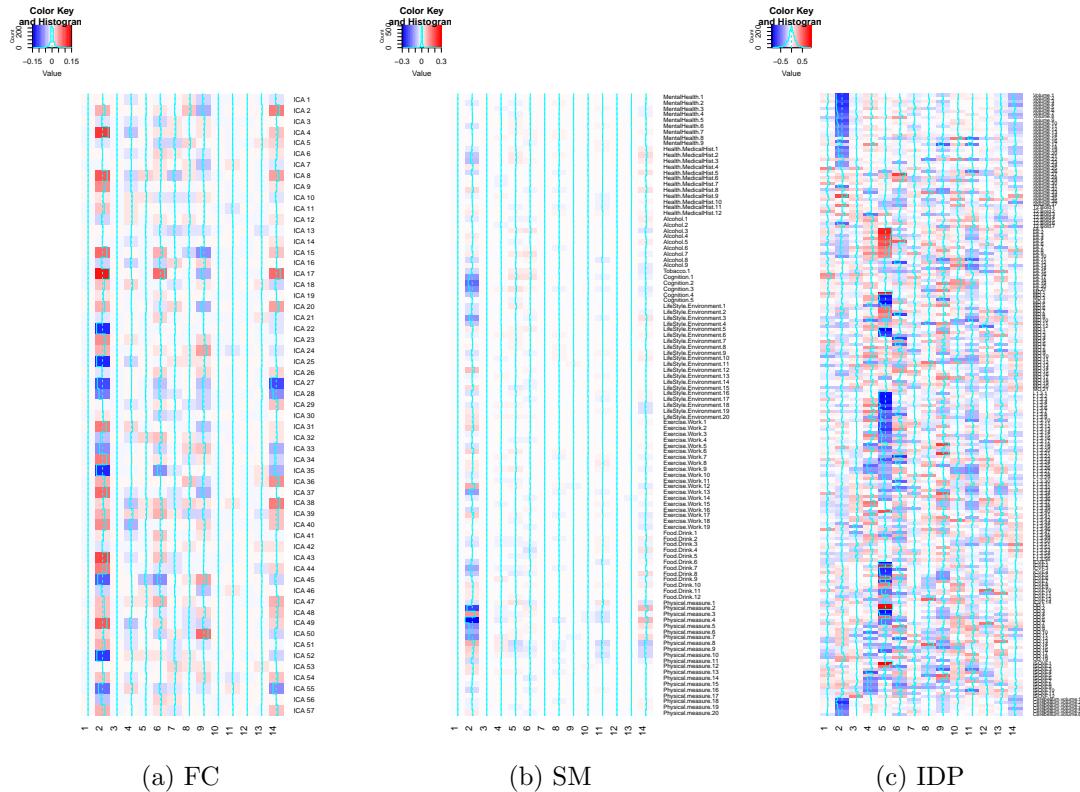


Figure 5.41: Significant GFA loadings (loadings with mean absolute values larger than 0.05) for the three modalities. Subplot (a), (b) and (c) are filtered from (a), (b) and (c) in Fig. 5.40 respectively.

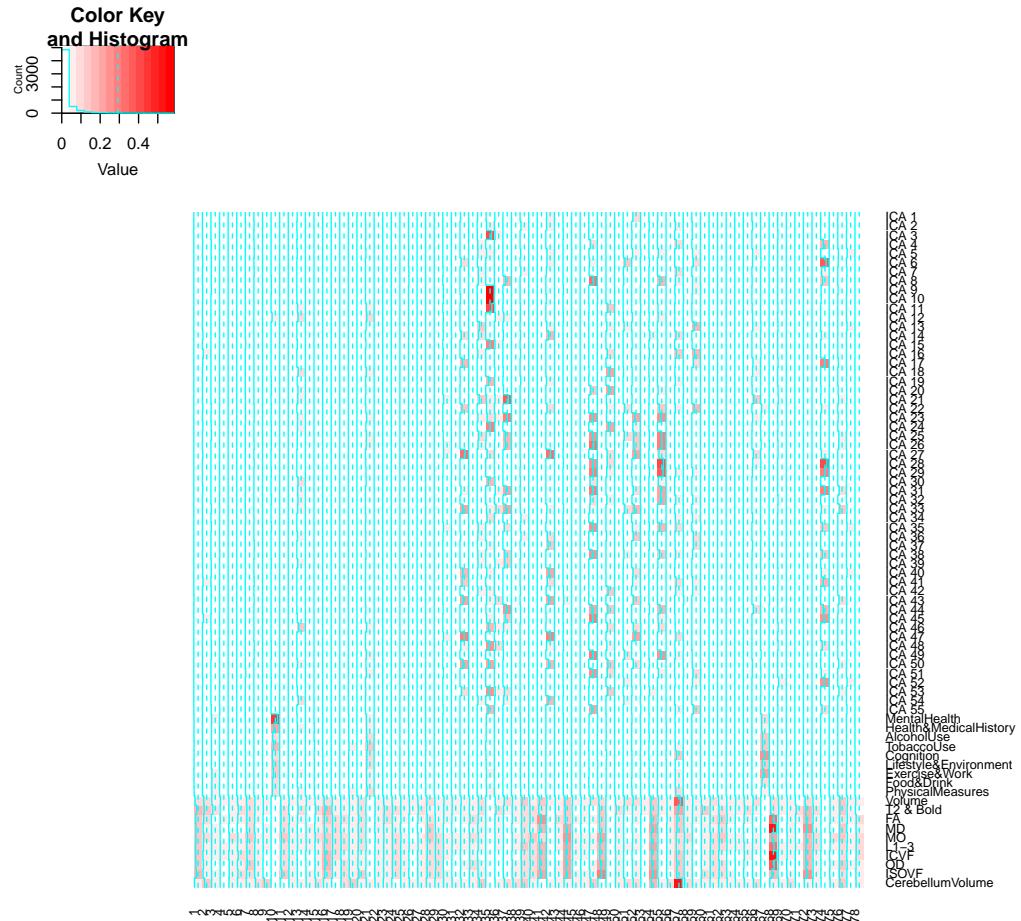


Figure 5.42: GFA loading matrix summarised by modality sub-domains for all GFA components. For a component in a sub-domain, the summarised loading is calculated by taking the mean of absolute loadings that are larger than 0.01.

domains). Within each sub-domain, we turn loadings with absolute value smaller than 0.01 to zeros to alleviate noise, and then take the mean of the absolute values of the non-zero loadings.

Fig. 5.42 shows the concatenation of the loadings for all three modalities summarised by sub-domains and for all 78 GFA components. It illustrates the distribution of the loadings across modalities and sub-domains. It is also clearer to see that many of the components only load on one modality. Fig. 5.42 is equivalent with the concatenation of all subplots in Fig. 5.40 and summarise the rows by sub-domains. To further explore the shared components, we take the sum of the absolute loadings for each component within each modality, and extract the components having the sum larger than 0.01 in more than one modalities.

There are 49 components shared by at least two modalities and are shown in Fig. 5.43. Most of them are shared between IDP and FC. Component 19 is shared between Physical Measures and Exercise & Work in SM and a few ICA regions in FC, with ICA 9, 10, 11 and 3 particularly loaded (the thumbnails of these ICA regions can be found Fig. 5.44). ICA 9 is temporal area and mostly related to language and sentence comprehension; the activated area in ICA 10 is related to episodic memory; ICA 11 is a default mode region and ICA 3 is a visual region. This finding coincides with the first and fifth modes in pairwise CCA on SDR-reduced SM and FC (Fig. B.15 and Tab. 5.5). One other interesting component is component 7. It has the largest loading on the ‘Mental Health’ sub-domain of SM and is lightly loaded on ICA region 12 which is a dorsolateral prefrontal area. However, based on NeuroSynth decoding, this region does not associate with any brain functions significantly.

In the end, we look at the components shared by all three modalities. There are 16 such components (Fig. 5.45). Many of them have dominating modalities. Only three of them are noticeably coloured across all modalities, component 4, 5 and 14. Component 14 is the one discussed above focusing on the IDP volume sub-domains, physical measure, cognition and a wide range of ICA regions. Component 5 is loaded on all SDR SM factors which are related to the volume and L1-L3 measures in IDP. Component 9 displays high loadings on the two volumes sub-domains of IDP and ICA regions 26, 31 and 49, which are frontal parietal, prefrontal and temporal area respectively.

5.5.0.1 Comparison between CCA and GFA

We applied pairwise CCA to investigate latent structures shared by each pair of the modalities, and multi-view CCA for the structures shared by all three modalities. GFA perform both tasks in one model. Moreover, GFA also identify factors that explain variance in only one of the modalities. These two models reveal similar results. For example, IDP and FC show closer relationship compared with the other two combinations. CCA demonstrates this point showing higher canonical correlations whereas GFA proves it by showing most of the latent factors are loaded on IDP and FC. Furthermore, the latent factors from GFA and CCA display similar patterns such as the positive correlation between the volume sub-domains of IDP and physical and cognitive measures in SM.

By applying CCA, we can perform more specific tasks such as pairwise or

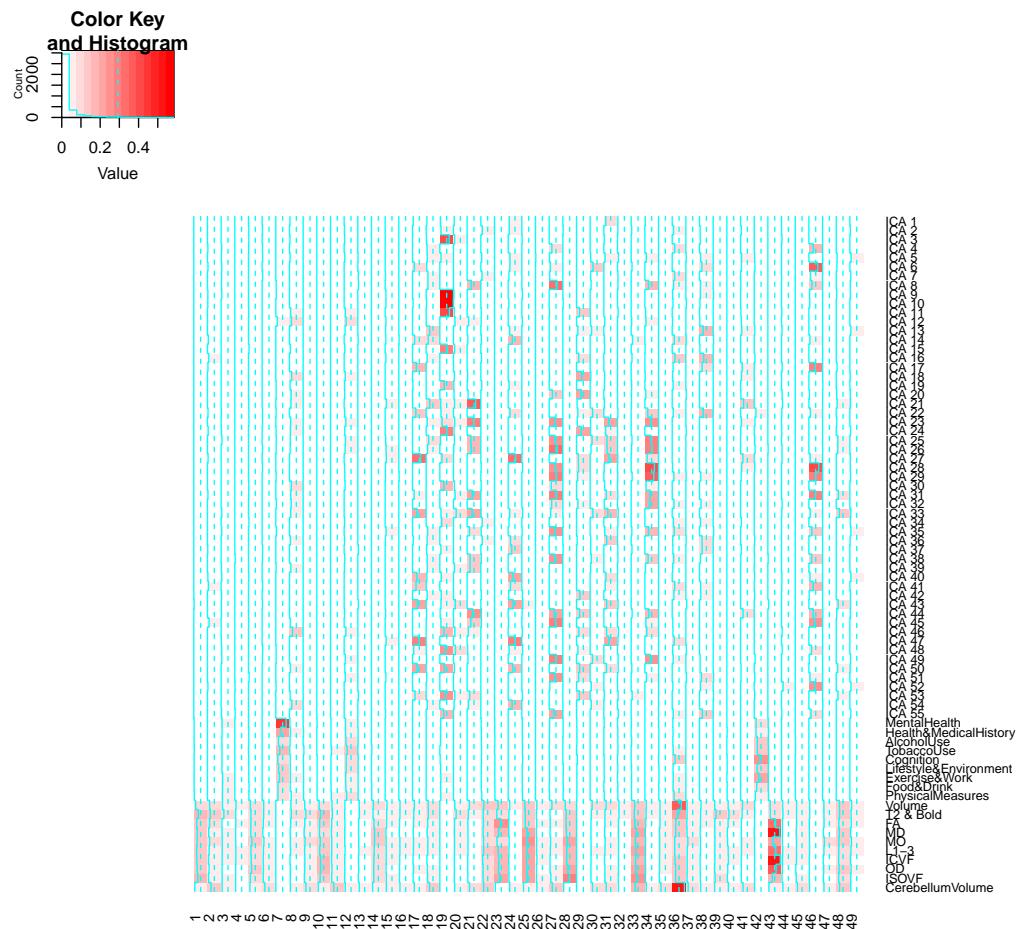


Figure 5.43: GFA loading matrix for components shared by at least two modalities. Loadings are summarised by sub-domains (same as shown in Fig. 5.42).

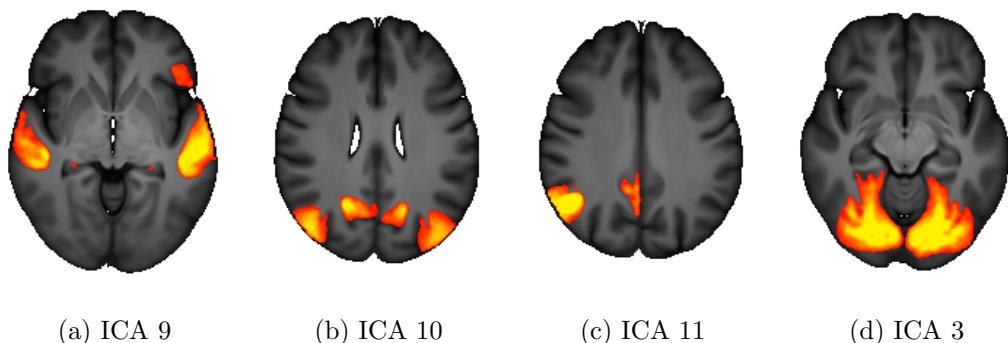


Figure 5.44: Thumbnails for ICA regions 9, 10, 11 and 3.

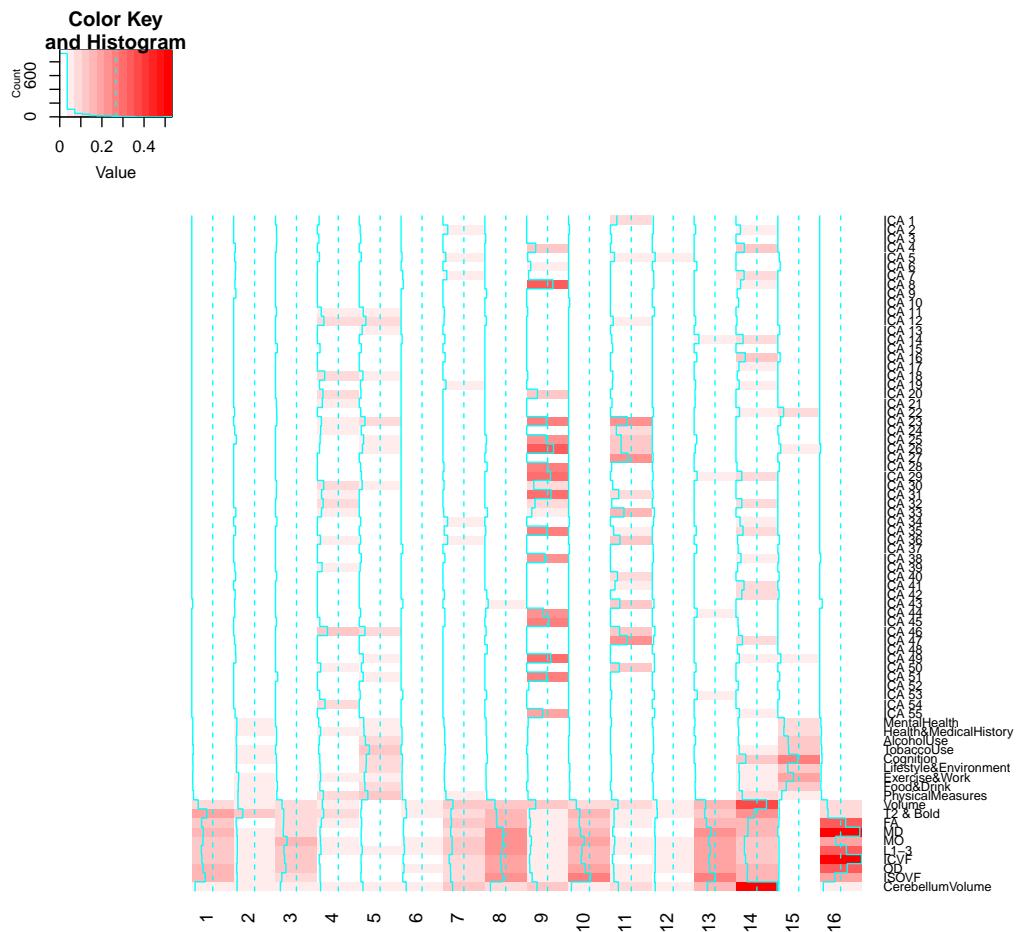


Figure 5.45: GFA loading matrix for components shared by all three modalities. Loadings are summarised by sub-domains (same as shown in Fig. 5.42).

multi-view. Therefore, we gain more detailed results in terms of each task. However this mechanism ignores the relationship and balance of the structures shared by two or three modalities. For example, from GFA we learnt that most of the latent structures are share by two modalities particularly IDP and FC. There are fewer and weaker components shared by all three modalities compared with the ones shared by only two of the modalities. Overall CCA and GFA provide insights on the latent structures from different perspectives: GFA focuses on a higher level interplay between modalities; CCA offers deeper understanding of specific relationship of interests.

5.6 Conclusion

In this project, we extended the study carried out on the Human Connectome Project (Chapter 4) to the UK Biobank project, a much larger health data project with more detailed health measures and data modalities. We studied the relationships between three data modalities, SM (subject measure), FC (functional connectivity) and IDP (image derived phenotype), which consist of 482, 1485 and 896 variables respectively after QC. We encountered several data challenges specific to SM including hierarchical missingness and heterogeneous data coding, which was solved by hierarchical missing data imputation and data re-coding prior to the analysis. We also applied sign-flipping to keep the meaning of the variables consistent.

Since the number of subjects in the UK Biobank dataset is much larger than the number of variables, it is not necessary to apply dimension reduction prior to CCA, which is the primary technique that was used to explore the shared patterns between modalities. Thus we first applied pairwise and multi-view CCA analysis without dimension reduction. From the pairwise CCA analysis on the non-reduced data, we found that IDP and FC show stronger relationship by having larger canonical correlations and more significant canonical pairs, which is not surprising since both modalities are brain imaging related and FC focuses on the functional aspect and IDP focuses on the structural aspect. Then comes the relationship between FC and SM which is only slightly stronger than SM and IDP. It implies that behavioural and demographics measures are more closely related to brain functions than structures. The MCCA gives the same order of relationship. We adapted permutation testing to the multi-view setting to test the significance of the sum of the canonical correlations, and obtained 18 significant sets of canonical variables. Although the canonical loadings on individual variables are difficult to summarise, we still gained some interesting insights. In conclusion, the pairwise CCA between

FC and SM, and between FC and IDP lead to very different brain activation maps. Comparing Tabs. 5.1 and 5.2 (and Figs. B.10 and B.14), areas in the brain that drive the correlation with SM are mainly motor, sensory, speech, and auditory related areas, however, the correlation with those areas are relatively weak. For areas with IDP, there are mainly default mode, retrieval, and theory of mind related areas. For the differences of SM canonical loadings in the CCA with FC and IDP, physical measures seem to dominate both, however different kinds of physical measures stood out in the two CCA studies. For FC, there are mainly body fat mass measures; for IDP there are height, weight and spirometry related measures in the first mode. Grey matter volumes in different parts of the brain from IDP side dominate the top modes in both of the CCA, with SM and FC. The canonical loadings obtained from multi-view CCA overlap with the ones in the pairwise analysis, but mainly dominated by the loadings appeared in the CCA between FC and IDP. Overall, for the non-reduced CCA case, due to the high number of variables, we identified more significant canonical modes compared with the HCP project. Moreover, without dimension reduction, canonical loadings become even more difficult to summarise. The latent structures shared between modalities focus on the details of the data. Therefore, we added dimension reduction SDR to alleviate this problem and improve the interpretability.

SDR reduces SM, FC and IDP to 107, 57 and 205 dimensions respectively. It is surprising to see the proportion of reduction for FC is the largest, leaving one or two component(s) for each FC sub-domain (ICA region). This may suggest that the noise level of FC data is high. After the dimension reduction, it is expected to see the canonical correlations drop. FC and IDP remain being the strongest correlated. However, the relationship between FC and SM becomes the weakest in the pairwise CCA analysis. We observe the same order in the multi-view CCA on SDR-reduced data and CCA on PCA-reduced data (see Appendix B.4.3), which implies that with higher dimensional data as inputs, CCA can generate more correlated canonical variables. With the help of SDR, we are able to visualise the constitution of each CCA mode by sub-domains and conclude the following interesting discoveries. Healthy exercise habit and working environment are positively correlated with comprehension and language related areas of the brain; unhealthy mental status and hard physical work are related to action and sensory related brain regions; tobacco usage affects working memory and is linked to dementia related brain area; physical and cognitive measures are most correlated with brain volumes between SM and IDP; brain volume measures drive the relationship with FC. Finally when considering all three SDR reduced datasets together, we found similar behaviour as in

the non-reduced datasets: IDP and FC dominate the latent patterns. Nevertheless, out of the eight significant CCA modes we examined, we found alcohol and tobacco intake negatively correlated with motion, speech production brain areas, and also largely related to the diffusion measures in IDP.

Cross-validation study showed that although the number of significant canonical variables might be high, the stability of them may decrease quickly. Most of the CCA modes we examined have descent stability and generalisability with very stable latent factors and significantly large loadings. Once more, IDP and FC present stronger stability among all combinations of modalities.

Although CCA improved our understanding of the latent structures shared by two or three modalities, it does not show the balance between the structures shared between two and three modalities, nor the amount of variance explained by single modality. We conducted GFA to compliment the CCA analysis, and with the assistance of SDR, we are able to interpret the GFA results. Some of the results overlap with CCA. However, the main conclusions we gained are that most of the variance in the datasets cannot be explained by the shared latent structures; SM contributes much less to the shared structure compared with FC and IDP, and IDP is loaded on almost all shared factors.

In this project, SDR proved its advantage in improving the interpretability of latent factor models. We came across more complicated identifiability issue of the latent factors such as the problem displayed in Fig. 5.3. As the analysis pipeline gets more complicated, one should always be cautious when interpreting the latent factors and cross-validating them since they may represent different components after multiple simulation or applying factor rotation. One inadequacy of the analysis is that the FC sub-domains are not functionally interpretable, which leads to the interpretation of the results on the FC side very challenging and left uninterpreted in some cases. Moreover, the data quality compared with the HCP datasets is poorer especially for FC. This is reflected by the low dimensionality after applying SDR and weak signal in the function interpretation by NeuroSynth decoding.

CHAPTER 6

Predicting Personality with Functional Connectivity

6.1 Introduction

Predicting personal traits using fMRI has been a popular research topic in recent years. [43], [104] and [111] have shown success in predicting fluid intelligence, individual behaviour and sustained attention respectively using functional brain connectome.

Personality is a collection of individual features that are formed during the development of an individual. It reflects peoples values, attitudes, habits etc., which are of vital importance for individual behaviours and social relationships. Personality can be measured by various tests and is usually broken into components called the Big Five, including Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism ([50], [35]). This project looks at the connections between the Big Five personality factors and brain imaging data, particularly resting-state fMRI (rfMRI), and tries to predict each of the five factors by identifying personality-related brain network. rfMRI is especially advantageous for this investigation as no task paradigms are needed, which avoids the possible biases incurred by tasks.

[43] came up with two summery statistics, *positive feature network strength* and *negative feature network strength*, that are derived from functional connectivity and served as predictive features. With these statistics as input of linear regression, [43] and [104] successfully predicted individual's fluid intelligence, sustained attention and used these statistics as powerful neuromarkers for identifying individuals. In this study, we replicate their method to predict personality and extend the method to include SVR (Section 2.10.2). We consider two independent datasets, one from Southwest University, Chongqing China and the other is HCP 900 release. We

also consider the effect of different brain parcellation schemes which was suggested being influential in prediction ([43]).

In the dataset from Southwest University, we significantly predict Extraversion and Conscientiousness by using linear regression. By applying SVR, we predict Extraversion significantly as well. However, the accuracy of all predictions are low. For the HCP dataset, we have more information on participants demographics, therefore predicting using linear regression model is applied to rfMRI directly, and to rfMRI with some nuisances removed such as age and gender. These parallel analysis give very different results as predicting using rfMRI directly gives significant predictions on most of the factors, however, with the nuisances removed from rfMRI, all the predictions become non-significant. In the end, we study the effects of all the nuisances removed from rfMRI on personality predicting.

Being able to predict personality has great value in developing personalised medication. Also, this research would offer insights on factors that affect personality. Especially, these factors may be originally on different scales, bringing them all together in predicting personality could enable us to investigate their contributions/information on the same scale.

6.2 Method

6.2.1 Data

The data is from Southwest University, Chongqing China, consisting of 131 healthy young subjects (mean age = 19.7) with 44 males and 67 females. The personality data is formed by the well-defined Big Five factors in psychology: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. For each of the subjects, scores for all the five factors are obtained. All resting-state fMRI data is collected in the Southwest University Centre for Brain Imaging, Chongqing, China, using a 3.0-T Siemens Trio MRI scanner. Each subject is required not to drink alcohol the day before the experiments, which is then confirmed right before the scanning by questionnaires.

6.2.1.1 Resting-state fMRI acquisition

In resting-state fMRI scanning, the subjects were instructed to rest without thinking about a particular topic, and not to fall asleep or close their eyes. The 8-min scan of 242 contiguous whole-brain resting-state functional images was obtained using gradient-echo planar imaging (EPI) sequences with the following parameters: slices

= 32, repetition time (TR)/echo time (TE) = 2000/30ms, flip angle = 90, field of view (FOV) = 220 mm 220 mm, and thickness/slice gap = 3/1 mm, voxel size $3.4 \times 3.4 \times 3$.

6.2.2 fMRI data pre-processing and functional connectivity matrix

All fMRI data was pre-processed by SPM8 and the Data Processing Assistant for Resting-State fMRI (DPARSF) ([144]). We applied both Power atlas ([99]) and AAL2 atlas ([103]) to parcellate the brain. The Power atlas consists of 264 brain regions and for AAL2, we used 94 regions excluding the cerebellum. Then each of the regions was represented as a time series and the Pearson's correlation between every pair of them was calculated to form the connectivity matrix. Therefore, for each subject, Power atlas gives the connectivity matrix of dimension 264×264 and the connectivity matrix for AAL2 has dimension 94×94 . For the sake of simplification, we will illustrate the analysis using Power atlas only. All the analysis was repeated to the AAL2 atlas.

6.2.3 Personality related functional brain network

We applied the analysis separately to each of the Big Five factors. We computed Pearson's correlation between each of the brain connectivities (links) and the personality scores. The personality related functional brain network was then formed by links with correlation p-value smaller than a pre-defined threshold. We first tried the threshold with 0.01. To compare the difference, we later tried with 0.05. The brain network is divided into two, *positive brain network*, whose links are positively correlated with the personality score, and the *negative brain network* whose links are negatively related. Once we have these networks, we need to test whether we could make predictions (with positive correlation between predicted and true personality scores) about personality from these positive and negative networks.

6.2.4 Prediction using personality related brain network

Due to the small number of subjects in the dataset from the Southwest University, we applied LOOCV to perform prediction in the following way: one subject was excluded at a time to serve as the test set and the rest of the 130 subjects formed the training set. The positive and negative brain networks were generated within the training sets, i.e. the brain networks selection is independent between different folds of CV. These networks were then fitted into linear regression (Section 2.10.1) and SVR (Section 2.10.2) for prediction. As described in [43], for each subject in

the training set, we calculated the positive/negative network strength by summing up the connectivity strength in the positive and negative networks respectively. We then used this network strength as an independent variable to predict the personality scores. A linear model was fitted between the network strengths and the personality scores in the training set to obtain the linear coefficients for both networks separately. The positive and negative network strengths for the test set were calculated, and the linear coefficients from the training set were applied to the test subject to acquire two predicted personality scores for the test subject, one from the positive network and one from the negative network. Finally we looped over all subjects to get two predicted values for each subject.

For SVR, we used Matlab toolbox *Libsvm* ([28]) with the following kernel functions: linear kernel, RBF and Sigmoid kernel ((2.80) with $d = 1$, (2.81) and (2.82) respectively in Section 2.10.2). We tried both ϵ -SVR and ν -SVR methods where the latter one gives control on the proportion of support vectors in the solution, and the former controls how much error there will be in the model (see Section 2.10.2).

The prediction power of a network for a personality factor is measured by correlating the predicted scores with the true scores. Both correlation and p-value of the correlation were considered. We took p-value < 0.01 as significant predictions. We only considered results where the predictions are positively correlated with the true scores.

6.2.5 Common network for all subjects

To quantify the extent to which different brain links contribute to predict each personality factor for both networks, we followed the methods suggested in [43] and [104], identifying *common positive/negative network*. They are generated by selecting links that pass the threshold for every subject. Although the predictors we use are summaries (positive/negative network strength) of the individual connectivity links, by peeking into their constitution, we may gain further insights on what links drive such summaries, therefore, cause significant predictions. For easier interpretation, the links in Power atlas were matched into the Brodmann areas (BA), which are defined purely by neuronal organisation and closely related to cortical functions ([25]).

Big Five Factors	Positive Network		Negative Network	
	Correlation	p-value	Correlation	p-value
Agreeableness	0.0192	0.8276	0.082	0.3519
Conscientiousness	0.1666	0.0572	0.2981	5.4440e-04*
Extraversion	0.1979	0.0234	0.3187	2.0684e-04*
Neuroticism	0.2010	0.0214	-0.0233	0.7914
Openness to Experience	-0.3909	3.9154e-06*	-0.0576	0.5132

Table 6.1: Prediction power on Big Five personality factors using Power atlas. Significance of individual links is 0.01. Values with * are significant.

6.3 Results on Southwest University dataset

6.3.1 Linear regression with Power atlas

Three of the Big Five factors turned out to have significant predictions for one of the networks: negative network of Conscientiousness, negative network of Extraversion and positive network of Openness to Experience (Tab. 6.1).

Among all the significant networks ($p\text{-value} < 0.01$), the predicted values of negative network of Conscientiousness and of Extraversion are positively correlated with the true Conscientiousness scores and Extraversion scores respectively. The most significant network in predicting personality is the positive network of Openness to Experience. However, the predictions are negatively correlated with the true scores (Fig. 6.1), and thus not sensible predictions.

We then examined the links in the common negative networks for Conscientiousness and Extraversion which have significant positive predictions.

For the common negative network of Conscientiousness, there are 74 links across all subjects (all links are attached in Appendix Tab. C.1). Brodmann area (BA) 40 comes up most frequently which is part of the parietal cortex and involved with language perception and processing. It is mostly connected to areas 7 (parietal cortex involved in spatial vision), 20 (inferior temporal visual cortex involved in object and face perception) and 13 (posterior orbitofrontal cortex involved in emotion) ([102]). Area 19 is involved as frequently as area 40 and area 19 is an area in the ventral visual stream. It is mostly connected with area 20 and 21, which are inferior and middle, temporal gyrus and related to perception.

There are 231 links in the common negative network of Extraversion (all links are attached in Appendix Tab. C.2). Except for the visual areas, BA 32 comes up very frequently which is part of the anterior cingulate cortex involved in emotion

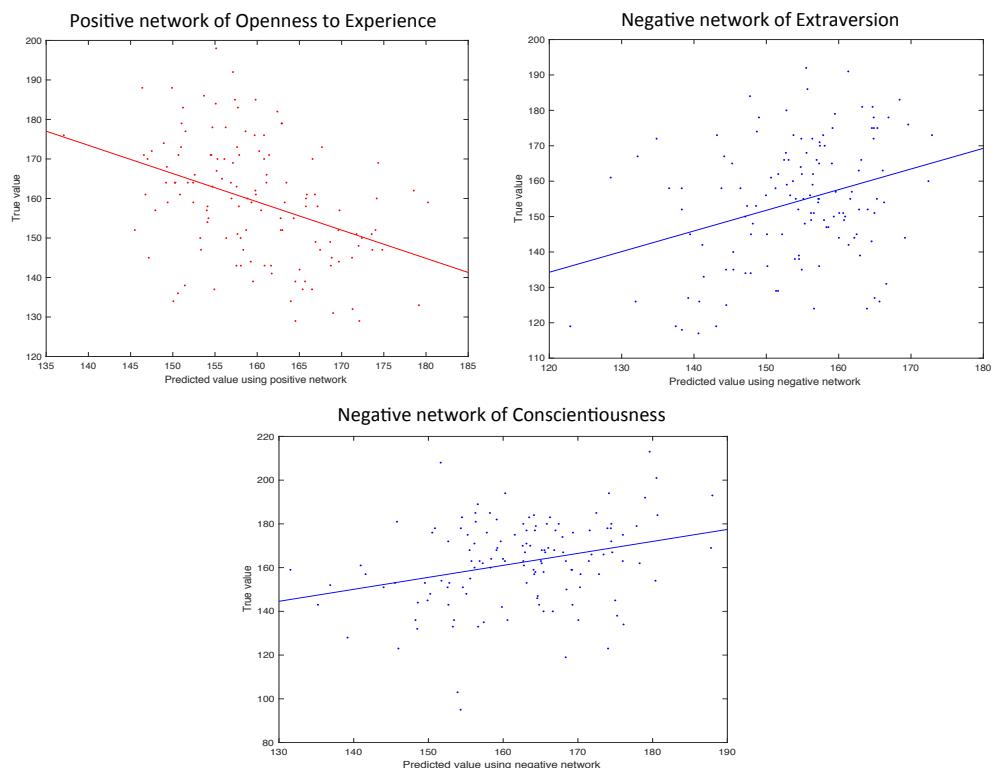


Figure 6.1: Predicted values versus true values for the positive network of Openness to Experience, negative network of Extraversion and negative network of Conscientiousness (from left to right, top to bottom).

([101], [102]). But area 32 is mostly linked to visual areas. BA 47/12 in the lateral orbitofrontal cortex is also an interesting emotion area, particularly punishment and non-reward related ([30], [101]), which comes up a few times in this network. It connects to areas that are involved in vision, auditory and speech related (e.g. BA 7, BA 22 and BA 19). Interestingly, it is also connected to BA 39, the angular gyrus, which is involved in language.

6.3.2 Linear regression with AAL2 atlas

Although the Power atlas parcellates the brain more finely (264 regions), it is difficult to make interpretations of those regions since there are not grouped based on functions. Therefore, we repeated our analysis using the AAL2 atlas for 94 regions excluding the cerebellum since the interpretations of the AAL2 areas are more widely understood. We first tried with the correlation threshold of 0.01. Among all five-personality factors, two had networks that could make significant predictions ($p\text{-value} < 0.01$). One was the negative network of Extraversion (Pearson's correlation $r = 0.2543$, $p\text{-value} = 0.0034$). Alternatively this network is positively correlated with introversion and the other one is the negative network of Openness to Experience ($r = -0.2660$, $p = 0.0021$). Since the negative network of Openness to Experience negatively predicts the personality which implies the model is not fitting, we will not consider its results.

In the common negative network of Extraversion, 76 links passed the threshold across all subjects. A considerable proportion of them link the occipital lobe (visual cortical areas) to the parietal, temporal, and central (somatosensory and motor) (Fig. 6.2, right; Fig. 6.3 Bottom).

Since only one set of predictions turned out to be successful for the single edge threshold of $p < 0.01$, we relaxed the threshold to 0.05. However, only one network appeared to be significant, negative network of Extraversion ($r = 0.3183$, $p = 2.1379e - 04$). This network consists of 133 links and is shown in the circular graph Fig. 6.4. This set of predictions is more positively related (having larger Pearson's correlation) to the true personality scores than using the threshold 0.01. When the threshold of the correlation between the connectivity and personality is 0.01, the correlation of predicted score and true score is 0.2543. When the threshold of the correlation between the connectivity and personality is 0.05, the correlation of predicted score and true score is 0.3183.

This could imply edges that can efficiently predict Extraversion are added to

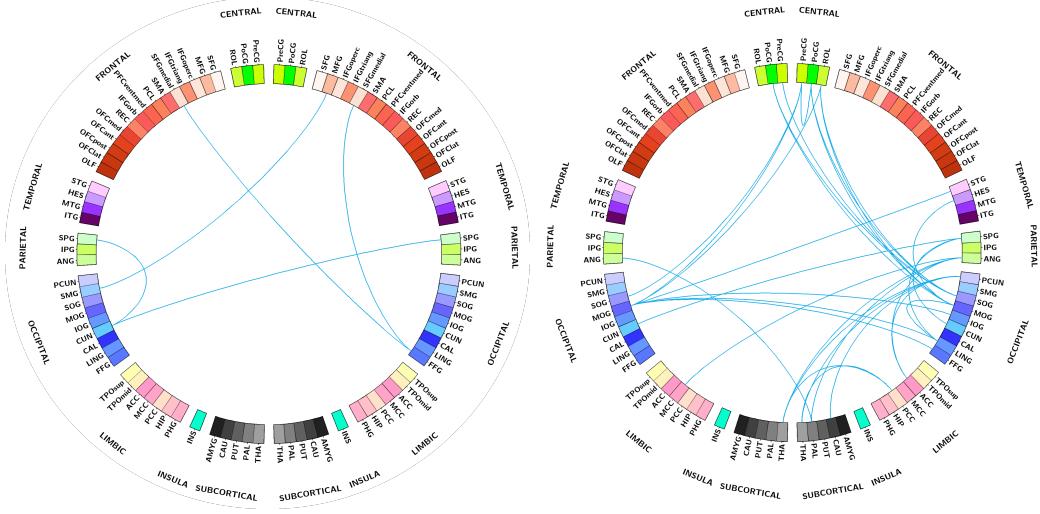


Figure 6.2: On the left is the circular graph of links in the common negative network of Openness to Experience; on the right is the circular graph of links in the common negative network of Extraversion. Links in both graphs were generated using threshold p-value smaller than 0.01.

the negative network. They include edges in the orbital frontal areas and singular areas which are of interests because these brain areas maybe related to personality ([101]). From Fig. 6.4 we can see that most of the links connect occipital lobes between two hemispheres, occipital to central and occipital to sub-cortical. However, due to the links in the negative network are selected being negatively correlated with Extraversion, therefore, the stronger these links are, the less extrovert a person can be.

To see the differences of using different thresholds, we then varied the threshold of just this network to 0.02 and found that it is still significant with $r = 0.2986$ and $p = 5.3232e - 04$. However, using threshold 0.02 gives more interpretable links than the threshold of 0.01. For example, anterior cingulate cortex (ACC), posterior cingulate cortex (PCC) and hippocampus (HIP) which are all related to Extraversion. In total, 78 links passed this threshold that shared by all subjects. The right graph in Fig. 6.5 shows that the right middle occipital and the left inferior occipital have the most links connected to. Most interestingly, we found links in anterior and medial orbital frontal gyrus, inferior and superior frontal gyrus which connected to Putamen, Caudate, temporal and anterior cingulate respectively (Fig. 6.5). Orbitofrontal cortex (OFCpos) with parahippocampal gyrus (PHG) link is negatively correlated with Extraversion, i.e positively correlated introversion.

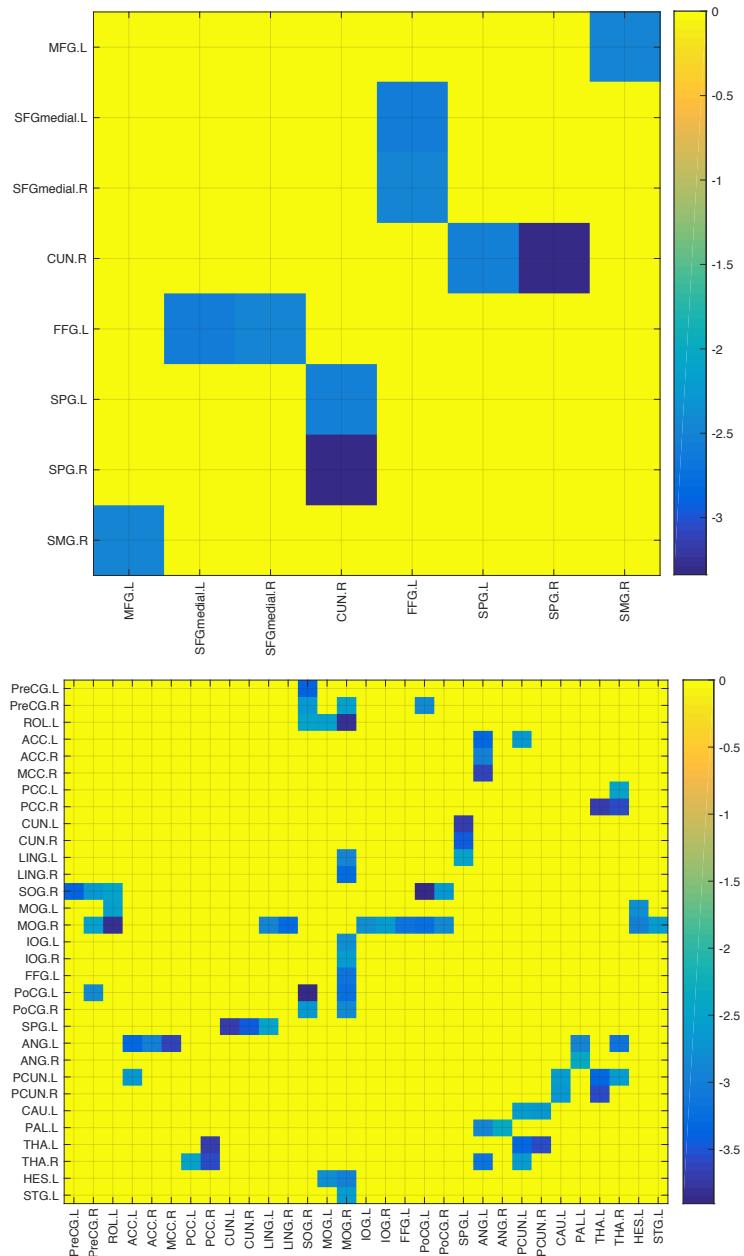


Figure 6.3: Matrix graph of links in the common negative network of Openness to Experience (top) and Extraversion (bottom). The colour stands for the significance level of the link in predicting respective personality factor, the logarithm of p-values of the correlation between each of the edges and personality score.

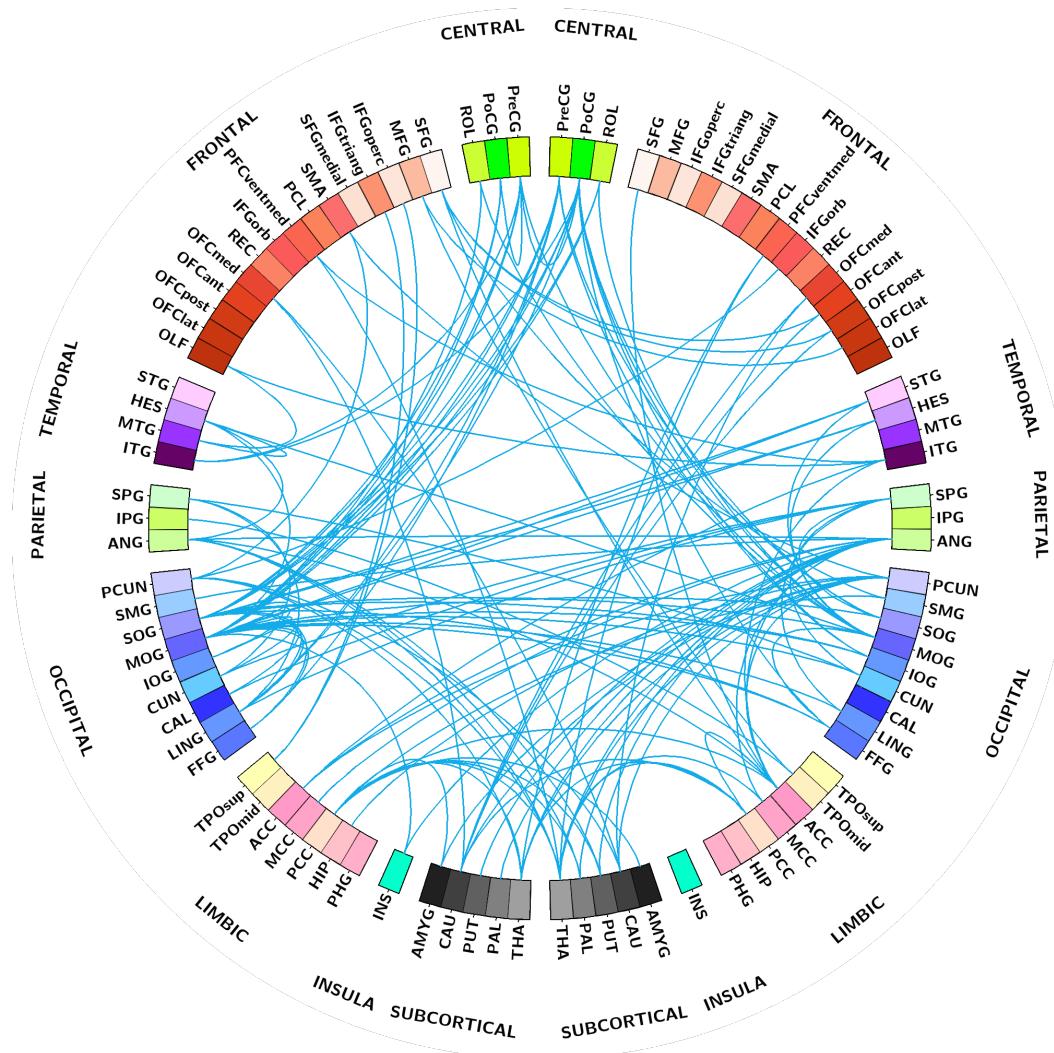


Figure 6.4: Circular graph of the negative network of Extraversion using threshold 0.05.

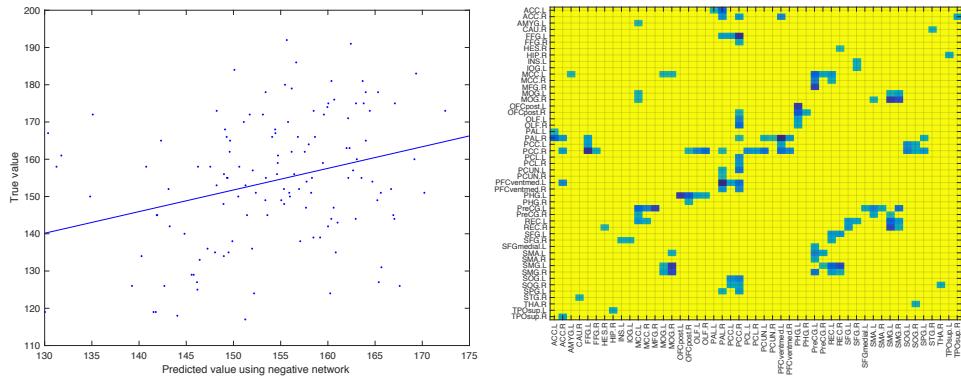


Figure 6.5: On the left is the predicted values versus true values for the negative network of Extraversion using threshold 0.02 with AAL2 atlas; On the right is the matrix graph showing the common links across all subjects in this network. The colour stands for the logarithm of p-values of the correlation between each of the edges and Extraversion score.

6.3.3 Comparison between Power and AAL2 parcellations

Using the Power atlas, we found three networks with significant results, however, only two of them gave positive correlation between the predicted and true scores. They are the negative networks for Extraversion and Conscientiousness. With AAL2, only one network gave positive significant result, the negative network of Extraversion. Although both parcellations did not do well in predicting personality factors, in general, the Power atlas with its greater number of areas (264 vs. 94) seems having slightly stronger prediction power. However, it is difficult to interpret the links in each of the networks with the Power atlas.

6.3.4 Analysis using Support Vector Regression

We used functional connectivity as the explanatory variables to predict Extraversion since this personality factor turned out being significant in the previous analysis. The AAL2 atlas was first applied because it is easier to interpret. To predict individual personality scores, we applied LOOCV. We trained the SVR model on $N - 1$ subjects with the positive and negative network strengths as inputs and predicted on the left out one and then loop over all of the subjects.

For ϵ -SVR, we first tried RBF kernel which is the default setting, and of the form (2.81) in Section 2.10.2. However, we got negative prediction correlation ($r = -0.6216, p = 10^{-15}$). From Fig. 6.6a we can see that not only there is a strong negative predicting power, but also the range of the predicted value is smaller than the true values, therefore we cannot accept this model. Since we have much more

Kernel	Correlation	p-value
Linear	0.1265	0.1499
RBF	-0.6216	10e-15
Sigmoid	bad performance	

Table 6.2: Summary statistics using AAL2 atlas for SVR modelling.

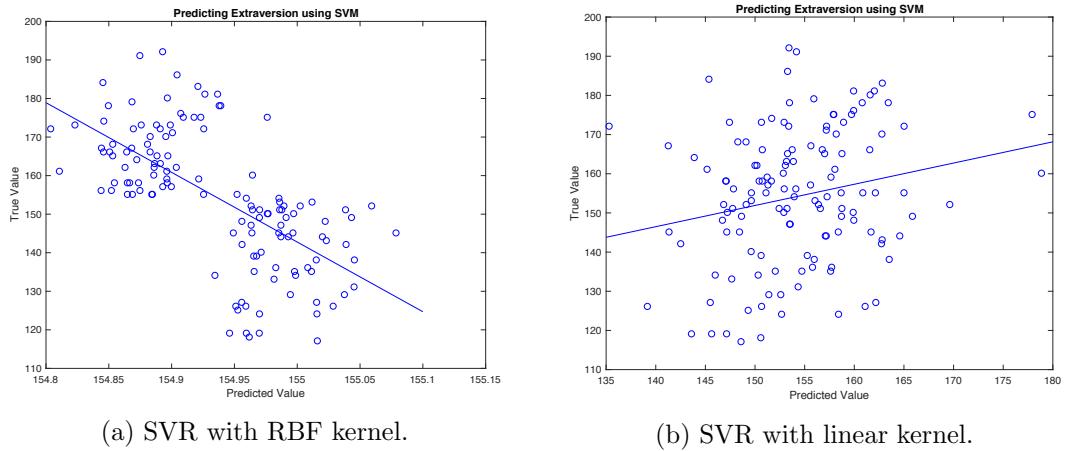


Figure 6.6: Predicted Extraversion score against true Extraversion score using Power atlas and Support Vector Regression (SVR) with RBF kernel (left) and linear kernel (right).

features than subjects, we then tried with the linear kernel which has form (2.80) with $d = 1$. However, it gave insignificant results ($r = 0.1265$, $p = 0.1499$). We have also tried tuning the parameters in the cost function (C in (2.74) and (2.75)) and optimisation conditions (γ in (2.81)) but they did not outperform the default setting. Summary see Tab. 6.2.

Because of the poor performance of AAL2 atlas, we switched to the Power atlas. Again we first tried RBF kernel and got similar results with the AAL2 atlas ($r = -0.645$, $p = 10^{-16}$). The linear kernel made an improvement in the prediction task and gave $r = 0.2131$ and $p = 0.0145$ (see Fig. 6.6b). The range of the predicted scores is more alike to the true range as we can see from Fig. 6.6b. Tab. 6.3 presents a summary comparison.

Moreover, we applied as well the Sigmoid kernel and ν -SVR. The results are neither insignificant nor negatively predicting and having too few predicted values, therefore not reported. The model with the best performance is the Power atlas with linear kernel in ϵ -SVR model.

Kernel	Correlation	p-value
Linear	0.2131	0.0148
RBF	-0.645	10e-16
Non-linear	bad performance	

Table 6.3: Summary statistics using Power atlas for SVR modelling.

In general, we find that our limited application of SVR has very poor performance, and it seems that linear regression outperforms SVR. However in most application cases, SVR performs at least as good as linear regression. Several reasons may be accounted for the poor performance of SVR in our case. First of all, we have only tried with one of the five factors of personality, and this factor is selected based on the performance of linear regression. It is possible that SVR may work better on other factors. Secondly, there might be other kernels that fit better with this dataset, or the kernel hyper-parameters can be better tuned. Lastly, the worse performance of SVR can be due to the overfitting of linear regression.

6.4 Personality Prediction on HCP data

To validate the results, we applied the same predicting analysis to the Human Connectome Project (HCP) data, with more subjects and potentially higher quality of the data. This dataset also has the Big Five personality tests scores. When this project commenced, we only had access to the HCP 900 release which consists of around 900 subjects. After removing the subjects with missing data, we have 813 subjects left.

The imaging data was rfMRI with ICA pre-processed (see Section 4.2.1 for more details) into 200 independent brain regions (same as the HCP project in Chapter 4), therefore there is no brain atlas involved for this dataset. Then a partial correlation connectivity matrix was calculated. Therefore, we have a 200×200 functional connectivity matrix for each subject. We then selected 11 variables as potential confounders. They are data release, gender, age, rfMRI motion, height, weight, blood pressure systolic, blood pressure diastolic, haemoglobin, intercranial volume and brain volume. We carried out the prediction separately using two design matrices: the original connectivity matrix and the de-confounded connectivity matrix (with the effects of the confounds removed from the target matrix; Section 2.8.4). The reason of making this distinction is that we were interested to see the effects of these confounders in predicting personality, i.e. whether the functional

connectivity or the confounders play a more important prediction role. We limit the HCP study to only replicate the method in [43], i.e. only apply linear regression to the HCP dataset.

We applied the same analysis as the one for the Southwest University dataset, but instead of using leave-one-out cross-validation, we applied split-half cross-validation for 10 different splits (a.k.a Monte Carlo CV) to increase the statistical power. This is because the sample size in the HCP data is much larger than the Southwest University dataset, therefore the training and test sets can have decent sizes without losing power when K -fold is applied (for more details see Section 2.9.1).

For both design matrices, the partial connectivity matrix was normalised (by subtracting the mean and dividing by the standard deviation). For the de-confounded matrix, the partial connectivity matrix went through de-confounding, and normalisation again.

A specific personality is then selected at a time as the response variable. All the subjects are grouped into 2 halves without breaking the family structures. Half of the subjects serve as the training set, the other half as the test set. The significant links are selected based on the correlation significance of 0.01. Again the links are divided into the positive network and negative network.

We train a linear regression model on the training set using the same feature selection method with the previous dataset, i.e. using the summarised positive or negative network strength as model input and the personality score for a specific factor as response. The coefficients obtained from the training set are then applied to the test set to get a predicted personality score for each of the subjects. The accuracy of the prediction is measured by calculating the correlation between true personality score and the predicted score. We have measured the difference between the actual score and the predicted score. The procedure is repeated for 10 different splits. We assess the results by examining the mean correlation between the predicted and true scores averaged over 10 splits, the mean error which is calculated as the percentage of the error (difference between predicted and true scores) in true score, and the number of significant simulations out of the 10 splits. A simulation is considered to be significant if the p-value of the correlation between predicted and true scores is smaller than 0.01.

6.4.1 Prediction using the original connectivity matrix

Tabs. 6.4 and 6.5 report the summary results of predicting all the five personality factors using the original (non-de-confounded) connectivity matrix. They show the

	Positive Network		
	Mean correlation (std)	# of sig. simulations	Mean error
Agreeableness	0.1519 (0.0391)	8	14.5%
Openness	0.1229 (0.0538)	5	21.64%
Conscientiousness	0.0904 (0.0404)	3	17.54%
Neuroticism	0.1062 (0.024)	1	72.48%
Extraversion	0.072 (0.055)	1	20.51%

Table 6.4: Prediction results on the five personality factors using for the positive network extracted from the non-de-confounded connectivity matrix. The second columns shows the mean correlation between predicted and true personality scores over 10 simulations with standard deviation in brackets; the third column shows the number of significant simulations out of the 10 splits; the last column is the deviation of the predicted score from the true score in percentage.

mean correlation for 10 simulations, and the number of significant simulations (with p-value less than 0.01) out the total 10 and mean error. All of them are calculated for both positive network and negative network. The mean error is defined in the following way: first of all, the absolute difference between predicted personality score and the true score is calculated for each subject in the test set and is denoted as d . Secondly, the proportion of d in the true score is computed ($d/\text{true score}$). Then we calculate the average of this proportion of all the subjects in the test set and this is the mean error for one simulation. In the end, the mean of all 10 simulations is reported in the table.

We can see that Agreeableness is the most predictable factor with 8 out of 10 and 9 out of 10 being significant simulations (with p-value less than 0.01) for positive and negative networks respectively. Besides, the mean error is the lowest among all five personalities. Openness and the negative network of Conscientiousness have similar performance. The total number of significant simulations for Openness (5) is higher than Conscientiousness (3), however, the mean error is higher as well. Both Neuroticism and Extraversion look pretty unpredictable by the non-de-confounded functional connectivity. Although the negative network of Neuroticism has 5 significant simulations, the mean correlation is low and the mean error is extremely high.

6.4.2 Prediction using the de-confounded connectivity matrix

We compare the results from the non-de-confounded connectivity matrix with predictions using de-confounded partial connectivity matrix. The same summary statis-

	Negative Network		
	Mean correlation (std)	# of sig. simulations	Mean error
Agreeableness	0.183 (0.040)	9	14.03%
Openness	0.1081 (0.0353)	5	21.88%
Conscientiousness	0.131 (0.035)	6	16.96%
Neuroticism	0.114 (0.0389)	5	70.23%
Extraversion	0.082 (0.041)	1	20.63%

Table 6.5: Prediction results on the five personality factors using for the negative network extracted from the non-de-confounded connectivity matrix. The second columns shows the mean correlation between predicted and true personality scores over 10 simulations with standard deviation in brackets; the third column shows the number of significant simulations out of the 10 splits; the last column is the deviation of the predicted score from the true score in percentage.

	Positive Network		
	Mean correlation (std)	# of sig. simulations	Mean error
Agreeableness	0.0544 (0.056)	1	15.15%
Openness	0.055 (0.052)	1	23.04%
Conscientiousness	-0.002 (0.061)	0	18.66%
Neuroticism	0.023 (0.050)	0	78.14%
Extraversion	0.077 (0.024)	0	20.96%

Table 6.6: Prediction results on the five personality factors using for the positive network extracted from the de-confounded connectivity matrix. The second columns shows the mean correlation between predicted and true personality scores over 10 simulations with standard deviation in brackets; the third column shows the number of significant simulations out of the 10 splits; the last column is the deviation of the predicted score from the true score in percentage.

tics are reported in Tab. 6.6 and 6.7. None of the predictions looks successful - most number of significant simulations being zero. Agreeableness is still the most predictable factor, however only one simulation in each network is significant ($p\text{-value} < 0.01$). The mean errors do not change significantly, but the mean correlations for Agreeableness, Openness, Conscientiousness and Neuroticism become much lower. Extraversion is unpredictable by either the positive or the negative network.

6.4.3 Effects of confounders

We studied the effect of all confounders in predicting personalities. We excluded one confounder at a time and used the rest to de-confound the connectivity matrix. Then we kept the rest of the analysis the same and saw the effect of missing

	Negative Network		
	Mean correlation (std)	# of sig. simulations	Mean error
Agreeableness	0.090 (0.035)	1	15.03%
Openness	0.046(0.030)	0	22.79%
Conscientiousness	0.037 (0.047)	0	18.20%
Neuroticism	0.027 (0.043)	0	76.52%
Extraversion	0.109 (0.032)	1	20.78%

Table 6.7: Prediction results on the five personality factors using for the negative network extracted from the de-confounded connectivity matrix. The second columns shows the mean correlation between predicted and true personality scores over 10 simulations with standard deviation in brackets; the third column shows the number of significant simulations out of the 10 splits; the last column is the deviation of the predicted score from the true score in percentage.

confounder in predicting Agreeableness since Agreeableness has shown of being the most predictable personality factor.

We found that none of the confounders played a critical role in affecting the predictions except the rfMRI motion, which increased the number of significant splits in the negative network significantly (see Tab. 6.6 and 6.7), but still not as high as in Tab. 6.4 and 6.5. Therefore, we can conclude that rfMRI motion has the highest impact on predicting Agreeableness using partial connectivity matrix. Each of the other confounders has some but not significant effect. However, with the removal of all confounders, the prediction of Agreeableness changed from fairly successful to almost unpredictable at all.

6.5 Conclusion

In this chapter, we tried to predict the Big Five personality factors using functional connectivity on two datasets, a dataset from Southwest University China and the HCP 900 dataset. For the dataset from Southwest University China, we applied linear regression model and SVR, and investigated the predictive power of functional connectivity derived from different parcellation schemes, in particular, the AAL2 atlas and the Power atlas. We found that SVR did not perform well with this analysis pipeline in predicting personality factors for both brain atlas and all kernel functions that were applied (RBF, linear and Sigmoid). This could be due to the choices of kernel functions and their hyper-parameters and the features that were fed into the model. With linear regression, the Power atlas appeared to have stronger predicting power than the AAL2 atlas. This may be due to the fact that Power atlas is more finely defined parcellation with considerably more regions than the AAL2 atlas (264

vs. 94). Specifically, Extraversion and Openness to Experience were significantly predicted by functional connectivity generated from both parcellation schemes with the Power atlas predicting one other factor, Conscientiousness. Interestingly, all of the successful cases were predicted by the negative networks of the functional connectivity.

HCP dataset was then used as an independent dataset to examine whether we can get similar results with the same model and feature selection method. We further extended the study to investigate the effect of confounders on the HCP data, where this information was missing from the previous dataset. Only linear regression was applied to this dataset due to the poor performance of SVR on the dataset from Southwest University. In this study, Agreeableness turned out to be the most predictable, and Extraversion is the least predictable. This is the total opposite case in the first dataset. Among the 11 confounders considered, rfMRI motion has the highest impact. Moreover, we found that with the effect of all 11 confounders removed from the functional connectivity, none of the personality factors can be predicted significantly. In this case, it is possible that the significant predictions were achieved by the confounders rather than the functional connectivity.

Results from the two datasets appear to be inconsistent, and we are aware that these two studies are not comparable due to the different parcellation schemes and the data pre-processing pipelines. However, we can see from these two studies that the prediction accuracy is affected by several factors such as the predictive models and confounders. To be able to conclude whether personality is predictable or not, more work needs to be done. For example, we used positive and negative network strengths derived from functional connectivity as predictors. They are very high level summaries of the functional connectivity, therefore, we lose a lot of detailed information. Other features or feature extraction methods can be explored such as using the significant individual links rather than the network strengths. Furthermore, more complicated models such as generalised linear models, neural networks are worth exploring, whereas in the latter one, much more data is needed.

CHAPTER 7

Conclusions

This thesis is inspired by the recent advances of MRI technology, computation power and machine learning methods and the interplay between brain and behaviours. With the help of publicly available health data projects such as the Human Connectome Project (HCP) and the UK Biobank project, large-scale neuroimaging and health-related non-imaging is now available. Therefore, researchers can finally bring neuroimaging studies to the population level.

In recent years, latent variable models have been widely applied to investigate the relationships between various brain and behavioural/demographics measures. Linking datasets from different scales and modalities often requires the discovery of some hidden spaces that are shared by the datasets. Latent variable models are effective tools to find the hidden spaces associated with the observed data. Moreover, latent variable models have extensive flexibility, identifying appropriate latent spaces for different objectives such as dimension reduction, feature selection and missing data imputation. In Chapter 2, we discussed the latent variable models applied in this thesis and their limitations and variations. In particular, Principal Component Analysis which is a popular method for dimension reduction, and Canonical Correlation Analysis which finds shared underlying structures between two and more sets of data. One of the main challenges we encountered during the application of latent variable models is the identifiability issue (discussed in Section 2.6). When the analysis pipeline gets complex, e.g. in the application of chained latent variable models, this issue is easily neglected and therefore, causes misinterpretations of the results or even wrong results. Therefore, we need to be particularly cautious when interpreting the signs of the latent variables/loadings, and apply sign-flipping or factor rotation when needed.

The projects carried out in this thesis were all based on real-world data. Real-world data, especially when collected over large scales, is often noisy and has various quality issues. Therefore, dealing with such data properly is a vital part

Conclusions

of the analysis. Moreover, due to the high complexity of the data and the analysis pipelines, studies performed on real-world data, particularly neuroimaging data, face the problem of being non-reproducible. Thus results validation and model assessment become more complicated and crucial. Chapter 2 also discussed commonly applied data processing techniques that solve issues like missing data, ill-conditioned measures and confounded data; it also introduced model validation and assessment methods which can be adapted to multi-step latent variable models.

One of the main contributions of this thesis is to improve the interpretability of the latent variable models. In Chapter 3, we introduced a dimension reduction technique named Supervised Dimension Reduction to help achieve this. The idea of SDR is to group variables into functional sub-domains using human prior knowledge on the data, and then reduce the dimension of the data by sub-domain. Moreover, the dimension of the sub-domains is estimated automatically by a two-way cross-validation algorithm. SDR alleviates the problem in PCA that the principal components are hard to interpret. In particular, when using principal components as inputs for other latent variable models such as CCA and Group Factor Analysis, the outputs become uninterpretable. With SDR, we were able to interpret the latent variables by the importance of the sub-domains of the original data. This makes it easier to summarise and visualise the results from latent variable models like CCA and GFA. To make the interpretation more consistent, we incorporated a sign-flipping step in SDR so that the signs of latent variables do not depend on the variable encoding.

In Chapter 4, we applied SDR and CCA to the HCP dataset to explore the relationship between brain measures, functional connectivity, and health-related subject measures. We also compared the performance of CCA using SDR and PCA reduced datasets as inputs. We found that their performances are comparable with the canonical correlation generated by SDR being slightly lower. However, we gained interpretability of the canonical loadings and variables. With the SDR CCA analysis, we discovered a canonical mode that is driven by the subjects who have good cognition and motor ability and do not smoke, but take drugs and drink, and have mental disorders. This set of behaviours is positively correlated with language, sentences, semantic related brain areas, and negatively correlated with pre-motor, motor and primary areas. We have also discovered two other CCA modes, both displaying high social-economic status and positive well-being patterns, with one focused on no drug use and the other dominated by no tobacco use. Furthermore, the cross-validation study proved that these results are stable.

Chapter 5 extended the SDR CCA analysis to the UK Biobank project

Conclusions

and investigated the relationships between three data modalities: subject measures (SM), functional connectivity (FC) and image-derived phenotypes (IDP). The sample size and the number of variables considered in this project were much larger than the HCP project. We carried out two sets of CCA analysis, without and with dimension reduction before CCA, and continued to use SDR as the dimension reduction technique. Without dimension reduction, the canonical variables had higher correlations than the dimension reduced case; however, the canonical loadings were complicated to summarise. After dimension reduction, the canonical correlations for all CCA combinations became weaker, especially for the CCA between FC and SM. We noticed that SDR reduced FC to only 57 dimensions, which is the lowest dimension among three reduced modalities. We found the same phenomenon in the HCP project, which is with lower-dimensional datasets as inputs for CCA, it tends to generate less correlated canonical variables. Although the number of variables for IDP is twice as many as the ones for SM in both non-reduced and reduced cases, the canonical variables for both modalities explain a similar amount of variance in their own original datasets. This implies that the canonical variables for IDP are more informative since we would expect the canonical variables to explain less variance for higher-dimensional data. In general, IDP and FC unsurprisingly have stronger correlations than other relationships. The physical measures in SM dominated the canonical relationship with both FC and IDP; the brain volume measures for IDP overshadowed other ones in both of the relationships with FC and SM. For FC, language and comprehension related brain areas stood out in the relationship with SM, and default mode network appeared most frequently in the CCA with IDP. In the multi-view setting, latent spaces were dominated by the relationship between IDP and FC. In the end, we applied group factor analysis (GFA) which identified latent structures shared by any subsets of the three modalities in one model. The results of GFA partly overlapped with CCA, with the extra insights on the proportion of contributions each modality made to the shared latent structures. We found that IDP basically loaded on every latent component GFA identified, and SM loaded on only a few latent components.

In Chapter 6, we focused on the relationship between functional connectivity (FC) with one specific subject measure, personality. We tried to use FC to predict the Big Five factors of personality. We carried out the analysis on two sets of data independently. One was from Southwest University in China and consisted of 131 subjects, and the other was the HCP 900 release. We found weak evidence that some personality factors are predictable from the two datasets: in the dataset from Southwest University China, Conscientiousness and Extraversion were predicted

Conclusions

significantly but with weak correlations between the true and predicted personality scores; in the HCP dataset, Agreeableness appeared to be the most predictable, however, with very weak correlations as well. We also investigated the effect of FC derived from different brain atlases in predicting and found that the prediction accuracy is brain atlas dependent. When the brain is parcellated into finer regions, the accuracy increases. Predictive models undoubtedly play an important role. Among the two predictive models we applied, linear regression had better performance than Support Vector Machines. On the HCP dataset, we further explored the effect of confounders in prediction. We selected several confounders including fMRI motion, age, gender, and discovered that after removing the effects of those confounders from FC, FC could barely predict any factors of personality. Therefore, the predictive power of FC towards personality remains questionable.

Latent variable models are powerful tools in data science and are widely applied to real-world data in different areas such as advertisement and engineering. For medical/health-related data, the interpretability of the models is essential. Being able to interpret the results offers better understanding to the target subjects or the mechanisms of the disease, so that researchers can generalise the results to a broader population, build personalised treatment or even create the cure of the disease. We spent a considerable amount of effort in this thesis to improve the interpretability of latent factor models. We applied techniques such as factor rotation and sign-flipping during the analysis, as well as introduced SDR to reduce the dimension. Dimension reduction is one of the primary purposes of applying latent variable models and is often a necessary step in the analysis pipeline when dealing with real-world big data. However, it may complicate the interpretation of the results, especially in the application of chained latent variable models. We carefully designed the analysis pipelines of latent variable models to make sure the variance of the pipeline is low, and the results are interpretable and stable. Nevertheless, we still faced many challenges. For example, like the CCA analysis presented in Chapter 5, when the data became very big, the sets of statistically significant results also became very large. With 500 subjects, permutation testing identified one significant CCA mode; for 1200 subjects, there were three significant modes; when there were 9000 subjects, this number increased to nearly ten. With these many sets of results, we struggled to summarise/visualise all of them. It shows that in the application of latent variable models to modern big or even mega data, model interpretation remains challenging. Besides, from the extensive study on the performance of chained latent variables models applied to neuroimaging and behavioural data, we discovered some model behaviour that is hard to explain. For example, in the HCP project, the amount of

the variance canonical variables explain in the observed datasets is non-monotonic with the input dimensions. Highly correlated canonical variables do not necessarily explain large amount of variance in the observed datasets. However, variance explained is also an important measure in CCA. Finding the relationship between canonical correlation and variance explained require further research.

The potential future work can also include further improvements of the interpretation of CCA results on the FC side. So far, we interpret the results by mapping FC canonical loadings onto brain volumes/surfaces via a rather complex procedure. It works well in terms of providing reasonable and insightful brain maps. However, due to the high dimensionality of the parcellation scheme we adopted, and the fact that the parcellation regions are not structurally/functionally defined, directly interpreting FC canonical loadings is very challenging. It is worth exploring structure/function-based parcellation schemes. Moreover, multi-view models such as MCCA and GFA are potent tools to unveil latent structures between data modalities. We explored FC, SM and IDP; however, other modalities such as structural connectivity and genetics would be very interesting to investigate. Although our prediction attempt on Personality was not very successful, we did see evidence of a connection between functional connectivity and behaviours from the latent variable model studies. Therefore, we still believe the possibility of predicting other individual traits using functional connectivity, and using latent variable models as the predictive models would be a natural next step.

Finally, neuroimaging, particularly functional connectivity, has attracted much interest, and has proved useful in understanding the brain and behaviours. We have shown that functional connectivity have close relationships with demographic, behavioural and brain structural measures. Nonetheless, we have also seen that it may not be the solution to every brain/behavioural related problem. There are various factors affecting the investigation of the question of interests such as the choice of models, confounders and data quality. In addition, studies like those exploring the links between brain and behaviours are highly interdisciplinary, and require expertise from multiple subjects. To better apply the models and interpret the results, interdisciplinary collaborations should be strengthened.

Bibliography

- [1] Hcp gallery. <http://www.humanconnectomeproject.org/gallery/>. Accessed: 2019-10-17.
- [2] A slice of axial mri at the level of the basal ganglia depicting changes in the fmri signal in red (an increase in blood oxygenation level) and blue (its decrease). <https://ru.wikipedia.org/wiki/file:FMRIscan.jpg>.
- [3] Uk biobank official website. <https://www.ukbiobank.ac.uk>. Accessed: 2019-07-13.
- [4] Abdi, H. (2003a). Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences*. Sage: Thousand Oaks, CA, pages 792–795.
- [5] Abdi, H. (2003b). Multivariate analysis. *Encyclopedia for Research Methods for the Social Sciences*. Thousand Oaks: Sage, pages 699–702.
- [6] Abdi, H. and Valentin, D. (2007). Multiple correspondence analysis. *Encyclopedia of Measurement and Statistics*, 2:651–66.
- [7] amoeba (<https://stats.stackexchange.com/users/28666/amoeba>). How to perform cross-validation for pca to determine the number of principal components? Cross Validated. <https://stats.stackexchange.com/q/115477> (version: 2018-02-28).
- [8] Anderson, T. W. (1962). An introduction to multivariate statistical analysis. Technical report, Wiley New York.
- [9] Ashburner, J. and Friston, K. J. (2007). Rigid body registration. *Statistical parametric mapping: The analysis of functional brain images*, pages 49–62.
- [10] Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49.
- [11] Bailey, D. L., Maisey, M. N., Townsend, D. W., and Valk, P. E. (2005). *Positron emission tomography*. Springer.

-
- [12] Barcikowski, R. S. and Stevens, J. P. (1975). A monte carlo study of the stability of canonical correlations, canonical weights and canonical variate-variable correlations. *Multivariate Behavioral Research*, 10(3):353–364.
 - [13] Beasley, T. M., Erickson, S., and Allison, D. B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics*, 39(5):580.
 - [14] Beaulieu-Jones, B. K., Lavage, D. R., Snyder, J. W., Moore, J. H., Pendergrass, S. A., and Bauer, C. R. (2017). Characterizing and managing missing structured data in electronic health records. *bioRxiv*, page 167858.
 - [15] Bela Ajtai, Eric Lindzen, J. C. M. Neuroimaging: Structural imaging: Magnetic resonance imaging, computed tomography. <https://clinicalgate.com/neuroimaging-structural-imaging-magnetic-resonance-imaging-computed-tomography/#f0060>. Accessed: 2019-10-17.
 - [16] Bennett, C. (2008). Aal brain atlas. "<http://prefrontal.org/blog/2008/05/brain-art-aal-patchwork/>".
 - [17] Bennett, C. M. and Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, 1191(1):133–155.
 - [18] Berry, W. D. (1993). *Understanding regression assumptions*, volume 92. Sage Publications.
 - [19] Bilenko, N. Y. and Gallant, J. L. (2016). Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in Neuroinformatics*, 10:49.
 - [20] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
 - [21] Bliss, C. I. et al. (1967). Statistics in biology. statistical methods for research in the natural sciences. *Statistics in Biology. Statistical Methods for Research in the Natural Sciences*.
 - [22] Blom, G. (1958). Statistical elements and transformed beta variables. *Wiley, New York. Boeschen, LE, Koss, MP, Figueiredo, AJ, & Coan, JA (2001). Experiential avoidance and post-traumatic stress disorder: A cognitive mediational model of rape recovery. Journal of Aggression, Maltreatment & Trauma*, 4(2):211–245.

BIBLIOGRAPHY

- [23] Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. *International Statistical Review/Revue Internationale de Statistique*, pages 291–319.
- [24] Bro, R., Kjeldahl, K., Smilde, A. K., and Kiers, H. A. L. (2008). Cross-validation of component models: A critical look at current methods. *Analytical and Bioanalytical Chemistry*, 390(5):1241–1251.
- [25] Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth.
- [26] Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186.
- [27] Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika*, 18(1):23–38.
- [28] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.
- [29] Cheng, W., Rolls, E. T., Gu, H., Zhang, J., and Feng, J. (2015). Autism: reduced connectivity between cortical areas involved in face expression, theory of mind, and the sense of self. *Brain*, 138(5):1382–1393.
- [30] Cheng, W., Rolls, E. T., Qiu, J., Liu, W., Tang, Y., Huang, C.-C., Wang, X., Zhang, J., Lin, W., Zheng, L., et al. (2016). Medial reward and lateral non-reward orbitofrontal cortex circuits change in opposite directions in depression. *Brain*, 139(12):3296–3309.
- [31] Cliff, N. and Krus, D. J. (1976). Interpretation of canonical analysis: Rotated vs. unrotated solutions. *Psychometrika*, 41(1):35–42.
- [32] Cole, J. H., Poudel, R. P., Tsagkrasoulis, D., Caan, M. W., Steves, C., Spector, T. D., and Montana, G. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124.
- [33] Corner, S. (2009). Choosing the right type of rotation in pca and efa. *JALT Testing & Evaluation SIG Newsletter*, 13(3):20–25.

-
- [34] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.
 - [35] Costa, P. T. and McCrae, R. R. (2008). The revised neo personality inventory (neo-pi-r). *The SAGE Handbook of Personality Theory and Assessment*, 2(2):179–198.
 - [36] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.
 - [37] Cristianini, N., Shawe-Taylor, J., et al. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
 - [38] Dattalo, P. V. (2014). A demonstration of canonical correlation analysis with orthogonal rotation to facilitate interpretation. *Manuscript*.
 - [39] Do Tromp (2009). Diffusion tensor imaging 101. <http://www.diffusion-imaging.com/2009/05/diffusion-tensor-imaging-101.html>. [Online; accessed July-2019].
 - [40] Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905.
 - [41] Ferreira, F., Rosa, M. J., Moutoussis, M., Dolan, R., Shawe-Taylor, J., Ashburner, J., and Mourao-Miranda, J. (2018). Sparse pls hyper-parameters optimisation for investigating brain-behaviour relationships. In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 1–4. IEEE.
 - [42] Filmer, D. and Pritchett, L. H. (2001). Estimating wealth effects without expenditure data or tears: an application to educational enrollments in states of india. *Demography*, 38(1):115–132.
 - [43] Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., and Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11):1664.
 - [44] Friman, O., Cedefamn, J., Lundberg, P., Borga, M., and Knutsson, H. (2001). Detection of neural activity in functional mri using canonical correlation analysis. *Magnetic Resonance in Medicine*, 45(2):323–330.

BIBLIOGRAPHY

- [45] Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Human Brain Mapping*, 2(1-2):56–78.
- [46] Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., and Frackowiak, R. S. (1995). Spatial registration and normalization of images. *Human Brain Mapping*, 3(3):165–189.
- [47] Gale, D. and Shapley, L. S. (2013). College admissions and the stability of marriage. *The American Mathematical Monthly*, 120(5):386–391.
- [48] Ge, T., Yeo, B. T., and Winkler, A. A brief overview of permutation testing with examples. <https://www.ohbmbrainmappingblog.com/blog/a-brief-overview-of-permutation-testing-with-examples>. Accessed: 2018-03-19.
- [49] Gittins, R. (2012). *Canonical analysis: a review with applications in ecology*, volume 12. Springer Science & Business Media.
- [50] Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48(1):26.
- [51] Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., Sochat, V. V., Nichols, T. E., Poldrack, R. A., Poline, J.-B., Yarkoni, T., and Margulies, D. S. (2015). Neurovault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, 9:8.
- [52] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.
- [53] Grellmann, C., Bitzer, S., Neumann, J., Westlye, L. T., Andreassen, O. A., Villringer, A., and Horstmann, A. (2015). Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of mri and genetic data. *NeuroImage*, 107:289–310.
- [54] Guadagnoli, E. and Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103(2):265.
- [55] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.
- [56] Hansen, P., Kringlebach, M., and Salmelin, R. (2010). *MEG: an introduction to methods*. Oxford university press.

-
- [57] Hardoon, D. R., Mourao-Miranda, J., Brammer, M., and Shawe-Taylor, J. (2007). Unsupervised analysis of fmri data using kernel canonical correlation. *NeuroImage*, 37(4):1250–1259.
 - [58] Harman, H. H. (1960). *Modern factor analysis*. University of Chicago Press.
 - [59] Harris, R. J. (1989). A canonical cautionary. *Multivariate Behavioral Research*, 24(1):17–39.
 - [60] Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, springer series in statistics.
 - [61] Hendrickson, A. E. and White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 17(1):65–70.
 - [62] Henson, R., Buechel, C., Josephs, O., and Friston, K. (1999). The slice-timing problem in event-related fmri. *NeuroImage*, 9:125.
 - [63] Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., and Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, 65(2):202–226.
 - [64] Hope, T. M., Seghier, M. L., Leff, A. P., and Price, C. J. (2013). Predicting outcome and recovery after stroke with lesions extracted from mri images. *NeuroImage: clinical*, 2:424–433.
 - [65] Horst, P. (1961). Relations among sets of measures. *Psychometrika*, 26(2):129–149.
 - [66] Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90.
 - [67] Hsieh, J. et al. (2009). Computed tomography: principles, design, artifacts, and recent advances. SPIE Bellingham, WA.
 - [68] Jennrich, R. I. and Sampson, P. (1966). Rotation for simple loadings. *Psychometrika*, 31(3):313–323.
 - [69] Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

BIBLIOGRAPHY

- [70] Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451.
- [71] Klamí, A., Virtanen, S., Leppäaho, E., and Kaski, S. (2014). Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2136–2147.
- [72] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.
- [73] Kolenikov, S., Angeles, G., et al. (2004). The use of discrete data in pca: theory, simulations, and applications to socioeconomic indices. *Chapel Hill: Carolina Population Center, University of North Carolina*, pages 1–59.
- [74] Kujala, J., Aho, T., and Elomaa, T. (2009). A walk from 2-norm svm to 1-norm svm. In *2009 Ninth IEEE International Conference on Data Mining*, pages 836–841. IEEE.
- [75] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- [76] Kumar, K., Chauvin, L., Toews, M., Colliot, O., and Desrosiers, C. (2017). Multi-modal brain fingerprinting: a manifold approximation based framework. *bioRxiv*, page 209726.
- [77] Laken, P. V. D. Simpson’s paradox: Two hr examples with r code. <https://paulvanderlaken.com/2017/09/27/simpsons-paradox-two-hr-examples-with-r-code/>. Accessed: 2019-09-17.
- [78] Lawley, D. N. and Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229.
- [79] Le Roux, B. and Rouanet, H. (2010). *Multiple correspondence analysis*, volume 163. Sage.
- [80] Lee, H. S. (2007). Canonical correlation analysis using small number of samples. *Communications in Statistics-Simulation and Computation*, 36(5):973–985.

-
- [81] Leung, D., Han, X., Mikkelsen, T., and Nabors, L. B. (2014). Role of mri in primary brain tumor evaluation. *Journal of the National Comprehensive Cancer Network*, 12(11):1561–1568.
 - [82] Li, Y.-O., Adali, T., Wang, W., and Calhoun, V. D. (2009). Joint blind source separation by multiset canonical correlation analysis. *IEEE Transactions on Signal Processing*, 57(10):3918–3929.
 - [83] Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F. C., and Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC bioinformatics*, 15(1):346.
 - [84] M. Udell, C. Horn, R. Z. and Boyd, S. (2016). Generalized low rank models. <https://github.com/powerscorinne/GLRM>.
 - [85] MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1):84.
 - [86] Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(Aug):2287–2322.
 - [87] McCarthy, P. (2019). funpack (version 1.3.1). <http://doi.org/10.5281/zenodo.3235753>, [Online; accessed 11/06/2019].
 - [88] McCullagh, P. (2018). *Generalized linear models*. Routledge.
 - [89] McRobbie, D. W., Moore, E. A., Graves, M. J., and Prince, M. R. (2017). *MRI from Picture to Proton*. Cambridge university press.
 - [90] Mihalik, A., Ferreira, F. S., Moutoussis, M., Ziegler, G., Adams, R. A., Rosa, M. J., Prabhu, G., de Oliveira, L., Pereira, M., Bullmore, E. T., et al. (2020). Multiple holdouts with stability: Improving the generalizability of machine learning analyses of brain–behavior relationships. *Biological Psychiatry*, 87(4):368–376.
 - [91] Mihalik, A., Ferreira, F. S., Rosa, M. J., Moutoussis, M., Ziegler, G., Monteiro, J. M., Portugal, L., Adams, R. A., Romero-Garcia, R., Vértes, P. E., et al. (2019). Brain-behaviour modes of covariation in healthy and clinically depressed young people. *Scientific Reports*, 9(1):1–11.
 - [92] Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiroopoulos, S. N., Andersson, J. L.,

BIBLIOGRAPHY

- et al. (2016). Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523.
- [93] Monteiro, J. M., Rao, A., Shawe-Taylor, J., Mourão-Miranda, J., Initiative, A. D., et al. (2016). A multiple hold-out framework for sparse partial least squares. *Journal of Neuroscience Methods*, 271:182–194.
- [94] Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). The alzheimer’s disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877.
- [95] Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.
- [96] Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15(1):1–25.
- [97] Nocedal, J. (1980). Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782.
- [98] Ogawa, S., Lee, T.-M., Kay, A. R., and Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24):9868–9872.
- [99] Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., et al. (2011). Functional network organization of the human brain. *Neuron*, 72(4):665–678.
- [100] Rasser, P. E., Johnston, P. J., Ward, P. B., and Thompson, P. M. (2004). A deformable brodmann area atlas. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, pages 400–403. IEEE.
- [101] Rolls, E. T. (2013). *Emotion and decision making explained*. Oxford University Press.
- [102] Rolls, E. T. (2016). *Cerebral cortex: principles of operation*. Oxford University Press.
- [103] Rolls, E. T., Joliot, M., and Tzourio-Mazoyer, N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *NeuroImage*, 122:1–5.

-
- [104] Rosenberg, M. D., Finn, E. S., Scheinost, D., Papademetris, X., Shen, X., Constable, R. T., and Chun, M. M. (2016). A neuromarker of sustained attention from whole-brain functional connectivity. *Nature Neuroscience*, 19(1):165.
 - [105] Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283.
 - [106] Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*, 52(3):1059–1069.
 - [107] Rupnik, J. and Shawe-Taylor, J. (2010). Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, pages 1–4.
 - [108] Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, 12(5):1207–1245.
 - [109] Schomer, D. L. and Da Silva, F. L. (2012). *Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins.
 - [110] Searle, S. R. and Gruber, M. H. (1971). *Linear models*, volume 12. Wiley Online Library.
 - [111] Shen, X., Finn, E. S., Scheinost, D., Rosenberg, M. D., Chun, M. M., Papademetris, X., and Constable, R. T. (2017). Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nature Protocols*, 12(3):506.
 - [112] Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.
 - [113] Sladky, R., Friston, K. J., Tröstl, J., Cunnington, R., Moser, E., and Windischberger, C. (2011). Slice-timing effects and their correction in functional mri. *NeuroImage*, 58(2):588–594.
 - [114] Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., et al. (2013). Resting-state fmri in the human connectome project. *NeuroImage*, 80:144–168.
 - [115] Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., and Miller, K. L.

BIBLIOGRAPHY

- (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, 18(11):1565.
- [116] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779.
- [117] Sui, J., Adali, T., Pearlson, G., Yang, H., Sponheim, S. R., White, T., and Calhoun, V. D. (2010). A cca+ ica based model for multi-task brain imaging data fusion and its application to schizophrenia. *NeuroImage*, 51(1):123–134.
- [118] Sui, J., He, H., Pearlson, G. D., Adali, T., Kiehl, K. A., Yu, Q., Clark, V. P., Castro, E., White, T., Mueller, B. A., et al. (2013). Three-way (n-way) fusion of brain imaging data based on mcca+ jica and its application to discriminating schizophrenia. *NeuroImage*, 66:119–132.
- [119] T BO, T. H., Dysvik, B., and Jonassen, I. (2004). Lsimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research*, 32(3):e34–e34.
- [120] Tan, P. N. (2018). *Introduction to data mining*. Pearson Education India.
- [121] Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Number 47. Sage.
- [122] Thompson, B. (2007). Factor analysis. *The Blackwell Encyclopedia of Sociology*.
- [123] Thorndike, R. M. and Weiss, D. J. (1973). A study of the stability of canonical correlations and canonical components. *Educational and Psychological Measurement*, 33(1):123–134.
- [124] Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of The Vectors of Mind*. University of Chicago Press.
- [125] Tibshirani, R., Wainwright, M., and Hastie, T. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.
- [126] Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences.

-
- [127] Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.
 - [128] Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 15(1):273–289.
 - [129] Udell, M., Horn, C., Zadeh, R., Boyd, S., et al. (2016). Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118.
 - [130] Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.
 - [131] Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: an overview. *NeuroImage*, 80:62–79.
 - [132] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer science & business media.
 - [133] Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179.
 - [134] Vidaurre, D., Smith, S. M., and Woolrich, M. W. (2017). Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832.
 - [135] Virtanen, S., Klami, A., Khan, S. A., and Kaski, S. (2011). Bayesian group factor analysis. *arXiv preprint arXiv:1110.3204*.
 - [136] Virtanen, S., Leppaaho, E., Klami, A., and Virtanen, M. S. (2013). R package ccagfa. <https://CRAN.R-project.org/package=CCAGFA>.
 - [137] Waerden, B. V. D. (1952). Order tests for the two-sample problem and their power. *Indagationes Mathematicae (Proceedings)*, 55:453–458.
 - [138] Wang, L., Zang, Y., He, Y., Liang, M., Zhang, X., Tian, L., Wu, T., Jiang, T., and Li, K. (2006). Changes in hippocampal connectivity in the early stages of alzheimer’s disease: evidence from resting state fmri. *NeuroImage*, 31(2):496–504.

BIBLIOGRAPHY

- [139] Westfall, J. and Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS One*, 11(3):e0152719.
- [140] Whitaker, K. J., Vértes, P. E., Romero-Garcia, R., Váša, F., Moutoussis, M., Prabhu, G., Weiskopf, N., Callaghan, M. F., Wagstyl, K., Rittman, T., et al. (2016). Adolescence is associated with genomically patterned consolidation of the hubs of the human brain connectome. *Proceedings of the National Academy of Sciences*, 113(32):9105–9110.
- [141] Wikipedia contributors (2019a). Alzheimer’s disease neuroimaging initiative — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Alzheimer%27s_Disease_Neuroimaging_Initiative&oldid=882873984. [Online; accessed 20-June-2019].
- [142] Wikipedia contributors (2019b). Human connectome project — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Human_Connectome_Project&oldid=897932949. [Online; accessed 20-June-2019].
- [143] Wikipedia contributors (2019c). Uk biobank — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=UK_Biobank&oldid=895103123. [Online; accessed 20-June-2019].
- [144] Yan, C. and Zang, Y. (2010). Dparsf: a matlab toolbox for “pipeline” data analysis of resting-state fmri. *Frontiers in Systems Neuroscience*, 4:13.
- [145] Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665.

APPENDIX A

Appendix for Human Connectome Project

A.1 Confounds for the HCP project

Confounds include: data release, data acquisition, gender, age (and age²), race white (binary), race black (binary), other race (binary), ethnicity, height (and height²), weight (and weight²), BMI, 3T fMRI Reconstruction Version, head motion, intra-cranial volume (cubed) and brain segmentation volume (cubed).

A.2 Full list of 234 SMS (HCP 1200 release)

Subject, Release, Acquisition, Gender, Age_in_Yrs, Race_white, Race_black, Race_other, Ethnicity, Height, Weight, BMI, Head_motion, fMRI_3T_ReconVrs, FS_IntraCranial_Vol, FS_BrainSeg_Vol, Handedness, SSAGA_Employ, SSAGA_Income, SSAGA_Educ, SSAGA_InSchool, SSAGA_Rlshp, SSAGA_MOBorn, -SSAGA_BMICat, -SSAGA_BMICatHeaviest, Hematocrit_1, Hematocrit_2, BPSystolic, BPDiastolic, ThyroidHormone, HbA1C, Menstrual_RegCycles, Menstrual_AgeBegan, Menstrual_CycleLength, Menstrual_DaysSinceLast, Menstrual_UsingBirthControl, -FamHist_Moth_Dep, -FamHist_Fath_Dep, -FamHist_Fath_DrgAlc, FamHist_Moth_None, FamHist_Fath_None, -ASR_Anxd_Raw, -ASR_Anxd_Pct, -ASR_Witd_Raw, -ASR_Witd_T, -ASR_Soma_Raw, -ASR_Soma_T, -ASR_Thot_Raw, -ASR_Thot_T, -ASR_Attn_Raw, -ASR_Attn_T, -ASR_Aggr_Raw, -ASR_Aggr_T, -ASR_Rule_Raw, -ASR_Rule_T, -ASR_Intr_Raw, -ASR_Intr_T, -ASR_Oth_Raw, -ASR_Crit_Raw, -ASR_Intn_Raw, -ASR_Intn_T, -ASR_Extn_Raw, -ASR_Extn_T, -ASR_TAO_Sum, -ASR_Totp_Raw, -ASR_Totp_T, -DSM_Depr_Raw, -DSM_Depr_T, -DSM_Anxi_Raw, -DSM_Anxi_T, -DSM_Somp_Raw, -DSM_Somp_T, -DSM_Avoid_Raw, -DSM_Avoid_T, -DSM_Adh_Raw, -DSM_Adh_T, -DSM_Inat_Raw, -DSM_Hype_Raw, -DSM_Antis_Raw, -DSM_Antis_T, -SSAGA_ChildhoodConduct, -SSAGA_PanicDisorder, -SSAGA_Agoraphobia, -SSAGA_Depressive_Ep, -SSAGA_Depressive_Sx, -EVA_Denom, Correction, -Noise_Comp, Odor_Unadj, Odor_AgeAdj, -PainIntens_RawScore, -PainInterf_Tscore, -Taste_Unadj, -Taste_AgeAdj, Mars_Log_Score, -Mars_Errs, Mars_Final, -THC, -SSAGA_Times_Used_Illicits, -SSAGA_Times_Used_Cocaine, -SSAGA_Times_Used_Hallucinogens, -SSAGA_Times_Used_Opiates, -SSAGA_Times_Used_Sedatives, -SSAGA_Times_Used_Stimulants, -SSAGA_Mj_Use, -SSAGA_Mj_Ab_Dep, SSAGA_Mj_Age_1st_Use, -SSAGA_Mj_Times_Used, -Total_Drinks_7days, -Num_Days_Drank_7days, -Avg_Weekday_Drinks_7days, -Avg_Weekend_Drinks_7days, -Total_Beer_Wine_Cooler_7days, -Avg_Weekday_Beer_Wine_Cooler_7days, -Avg_Weekend_Beer_Wine_Cooler_7days, -Total_Wine_7days, -Avg_Weekday_Wine_7days, -Avg_Weekend_Wine_7days, -Total_Hard_Liquor_7days, -Avg_Weekday_Hard_Liquor_7days, -Avg_Weekend_Hard_Liquor_7days, -SSAGA_Alc_D4_Dp_Sx, -SSAGA_Alc_D4_Ab_Dx, -SSAGA_Alc_D4_Ab_Sx, -SSAGA_Alc_D4_Dp_Dx, -SSAGA_Alc_12_Drinks_Per_Day, SSAGA_Alc_12_Frq, SSAGA_Alc_12_Frq_5plus, SSAGA_Alc_12_Frq_Drk, -SSAGA_Alc_Hvy_Drinks_Per_Day, SSAGA_Alc_Hvy_Frq, SSAGA_Alc_Hvy_Frq_5plus, SSAGA_Alc_Hvy_Frq_Drk, -SSAGA_Alc_Hvy_Max_Drinks, -Total_Any_Tobacco_7days, -Times_Used_Any_Tobacco_Today, -Num_Days_Used_Any_Tobacco_7days, -Avg_Weekday_Any_Tobacco_7days, -Avg_Weekend_Any_Tobacco_7days, -Total_Cigarettes_7days, -Avg_Weekday_Cigarettes_7days, -Avg_Weekend_Cigarettes_7days, -SSAGA_TB_Smoking_History, -SSAGA_TB_Still_Smoking, MMSE_Score, -PSQI_Score, -PSQI_SleepQuality1, -PSQI_SleepLatency, -PSQI_SleepQuality2, -PSQI_SleepDuration, -PSQI_SleepDisturbance, -PSQI_SleepMeds, -PSQI_DayDysfunction, PicSeq_Unadj, PicSeq_AgeAdj, CardSort_Unadj, CardSort_AgeAdj, Flanker_Unadj, Flanker_AgeAdj, PMAT24_A_CR, -PMAT24_A_SI, -PMAT24_A_RTCR, ReadEng_Unadj, ReadEng_AgeAdj, PicVocab_Unadj, PicVocab_AgeAdj, ProcSpeed_Unadj, ProcSpeed_AgeAdj, DDisc_SV_1mo_200, DDisc_SV_6mo_200, DDisc_SV_1yr_200, DDisc_SV_3yr_200, DDisc_SV_5yr_200, DDisc_SV_10yr_200, DDisc_SV_1mo_40K, DDisc_SV_6mo_40K, DDisc_SV_1yr_40K, DDisc_SV_3yr_40K, DDisc_SV_5yr_40K, DDisc_SV_10yr_40K, DDisc_AUC_200, DDisc_AUC_40K, VSPLIT_TC, -VSPLIT_CRTE,

Appendix for HCP

-VSPLIT_OFF, SCPT_TP, SCPT_TN, -SCPT_FP, -SCPT_FN, -SCPT_TPRT, SCPT_SEN, SCPT_SPEC, -SCPT_LRNR, IWRD_TOT, -IWRD_RTC, ListSort_Unadj, ListSort_AgeAdj, ER40_CR, -ER40_CRT, ER40ANG, ER40FEAR, ER40NOE, ER40SAD, -AngAffect_Unadj, -AngHostil_Unadj, -AngAggr_Unadj, -FearAffect_Unadj, -FearSomat_Unadj, -Sadness_Unadj, LifeSatisf_Unadj, MeanPurp_Unadj, PosAffect_Unadj, Friendship_Unadj, -Loneliness_Unadj, -PercHostil_Unadj, -PercReject_Unadj, EmotSupp_Unadj, InstruSupp_Unadj, -PercStress_Unadj, SelfEff_Unadj, Endurance_Unadj, Endurance_AgeAdj, GaitSpeed_Comp, Dexterity_Unadj, Dexterity_AgeAdj, Strength_Unadj, Strength_AgeAdj, NEOFAC_A, NEOFAC_O, NEOFAC_C, NEOFAC_N, NEOFAC_E.

Note: variable name with a ‘-’ in front suggests that it has been sign-flipped.

A.3 Sub-domain summary reports

The following figures show the sub-domain summary reports for all sub-domains. In each of the following figures of this section, the top left panel shows the eigen-spectrum (blue), cumulative eigen-spectrum (red) and null eigen-spectrum (green); middle left panel shows the cumulative variance explained by principal components (PCs) in cross-validation; bottom left panel is the summary table for panel B showing 3 benchmark percentages 50%, 70% and 90%; top right panel shows the principal loadings for optimal number of PCs; middle right panel shows the rotated loadings in D; bottom right panel shows the error curves calculated by Eqn.3.11 and Eqn.3.13. The naive way of calculating PRESS (dotted line) is monotonically decreasing, while the two-way CV method (red line) offers a minimum point.

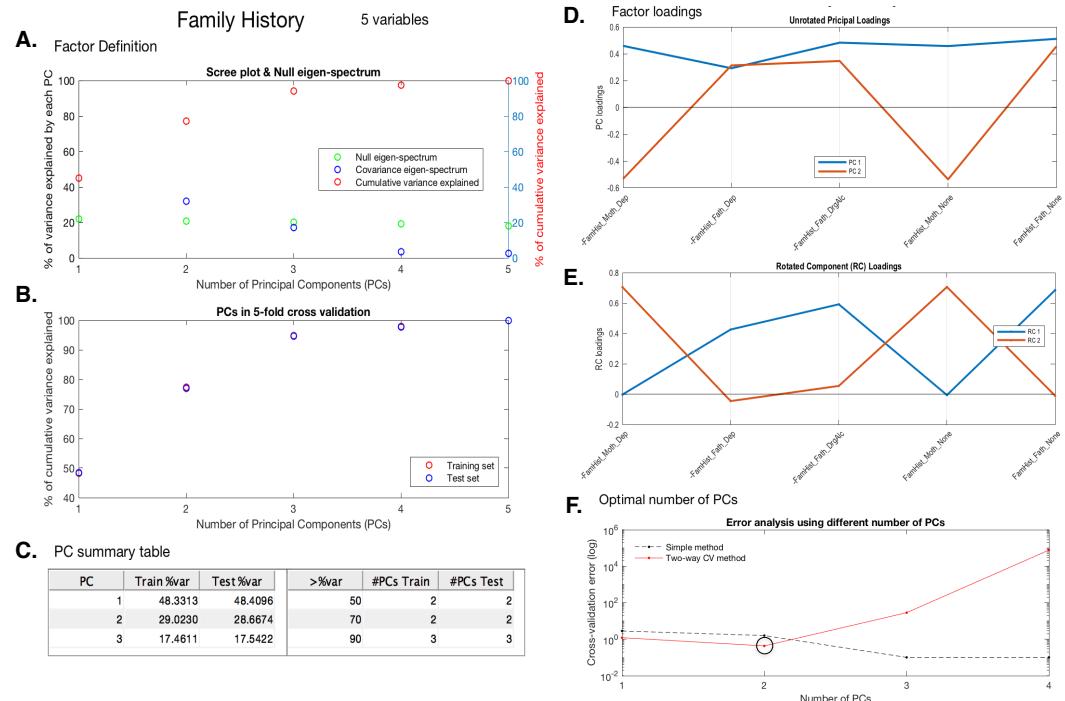


Figure A.1: Summary report of Family History.

Appendix for HCP

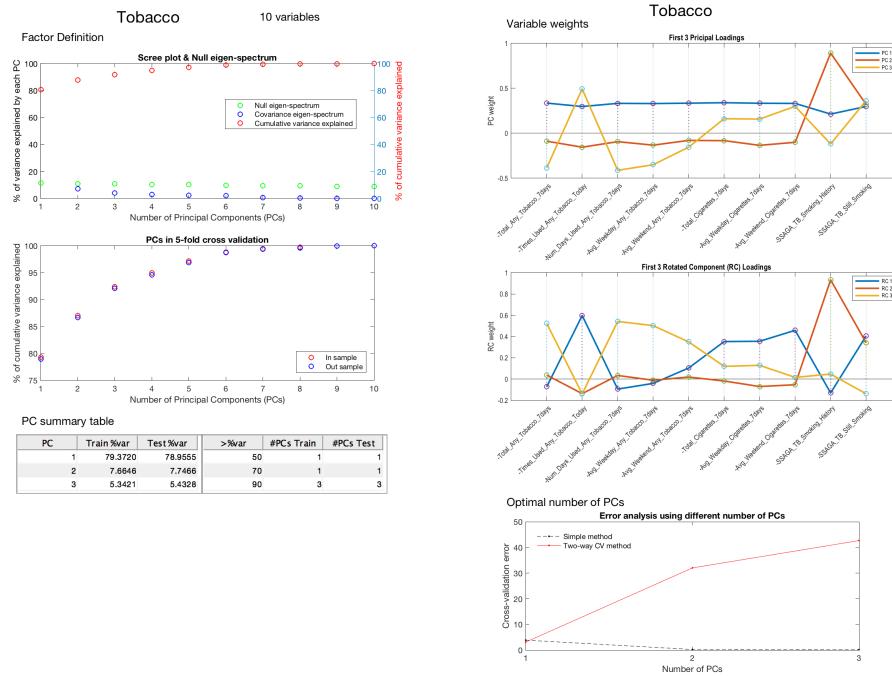


Figure A.2: Tobacco Use sub-domain summary report.

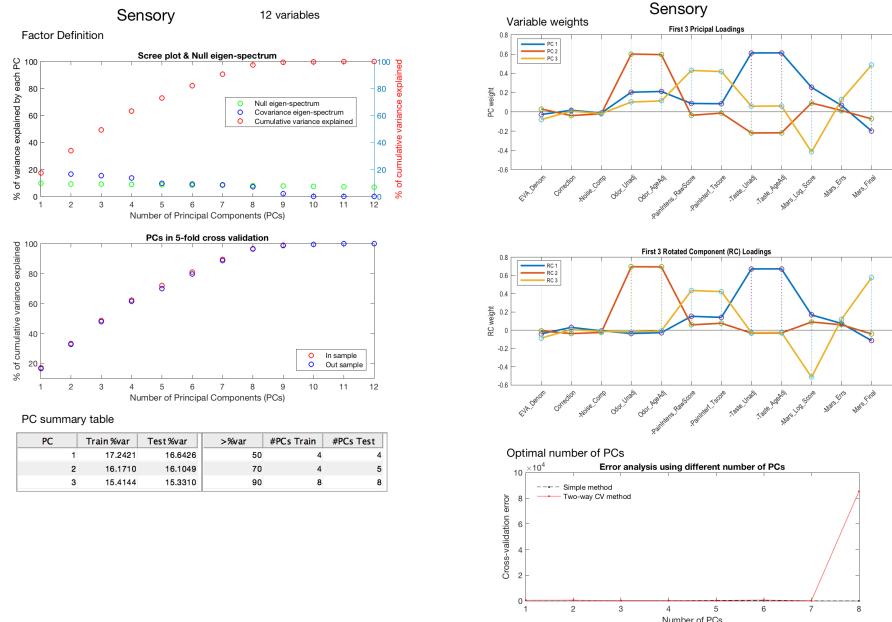


Figure A.3: Sensory sub-domain summary report.

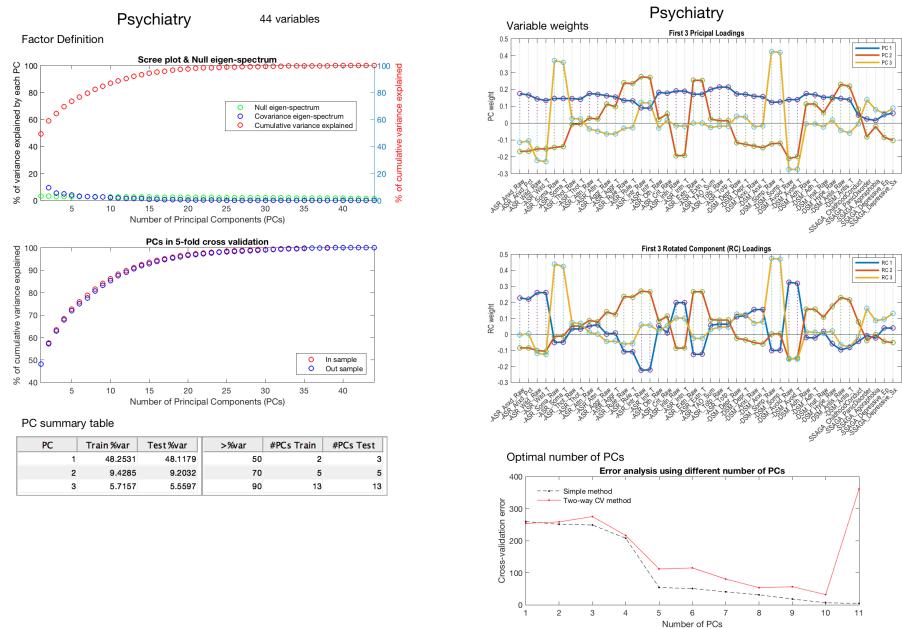


Figure A.4: Psychiatry sub-domain summary report.

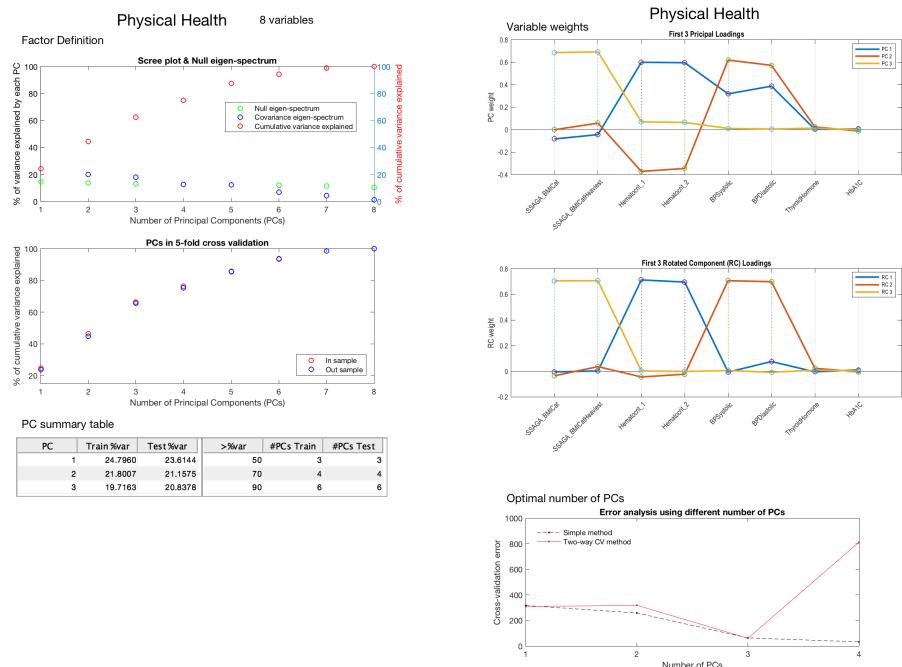


Figure A.5: Physical Health sub-domain summary report.

Appendix for HCP

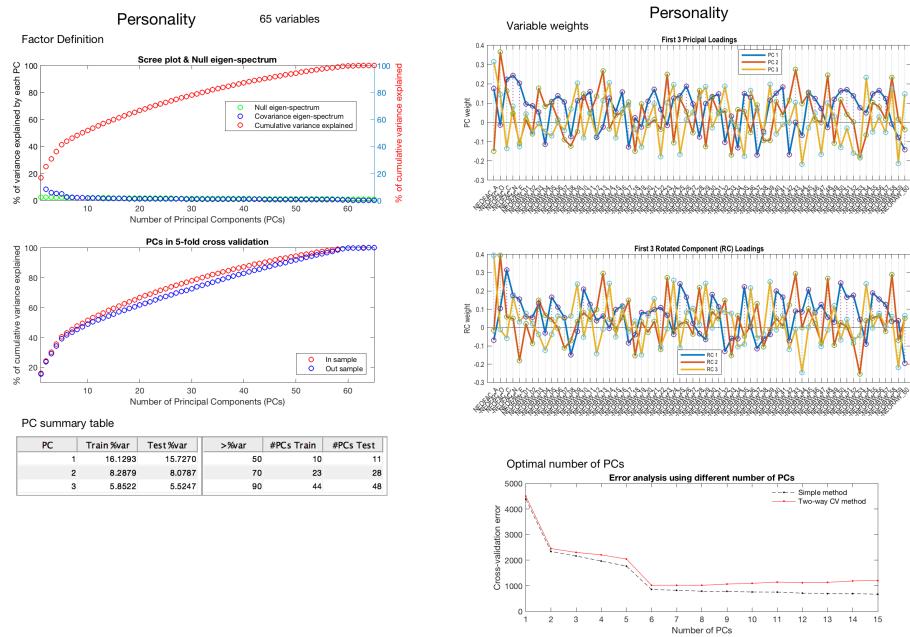


Figure A.6: Personality sub-domain summary report.

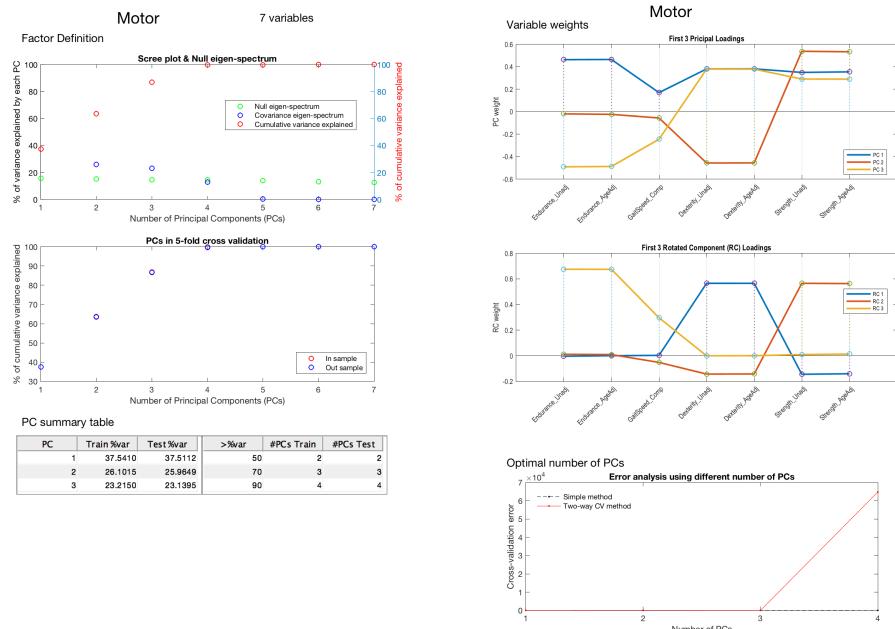


Figure A.7: Motor sub-domain summary report.

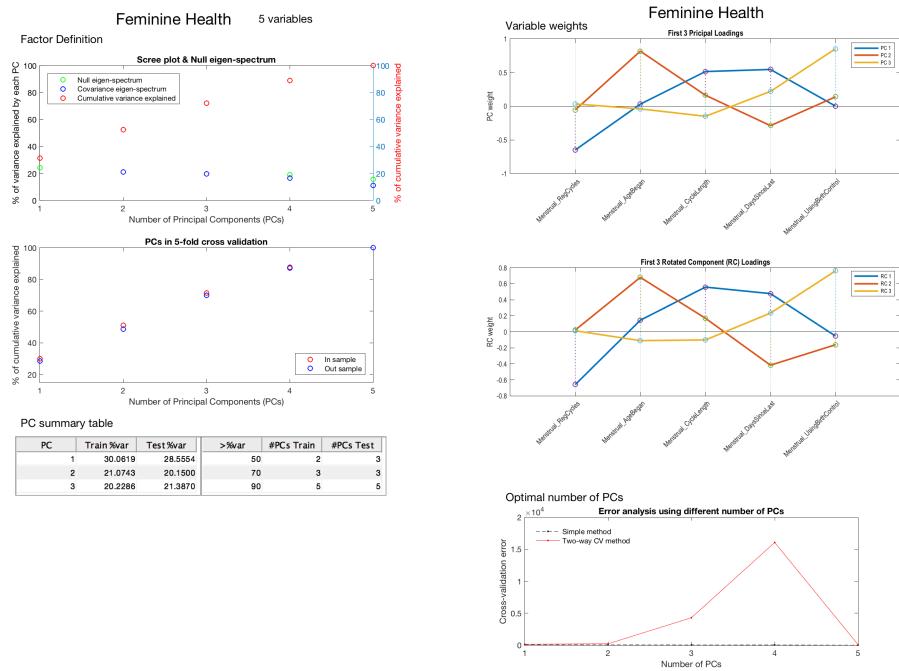


Figure A.8: Feminine Health sub-domain summary report.

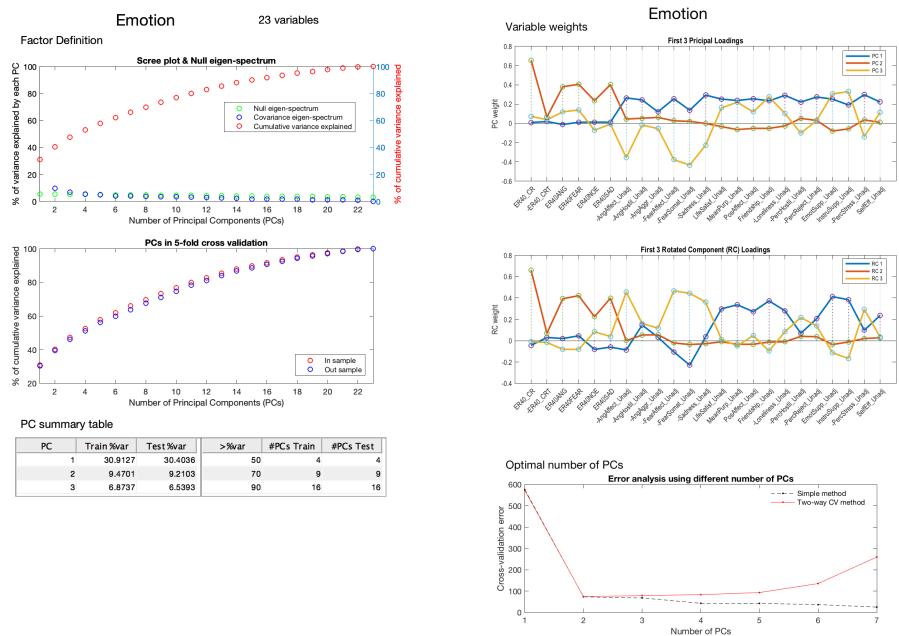


Figure A.9: Emotion sub-domain summary report.

Appendix for HCP

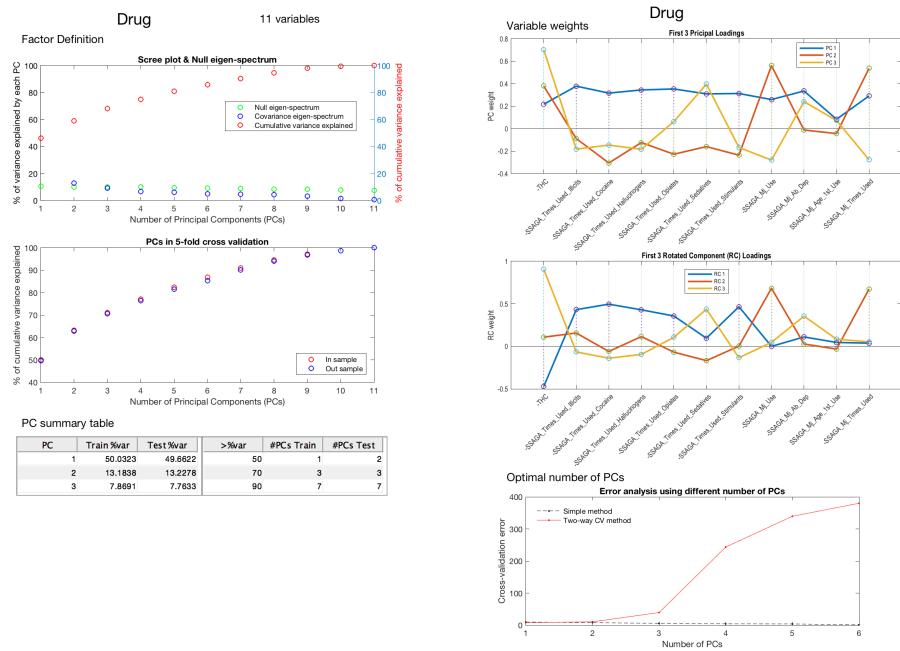


Figure A.10: Drug Use sub-domain summary report.

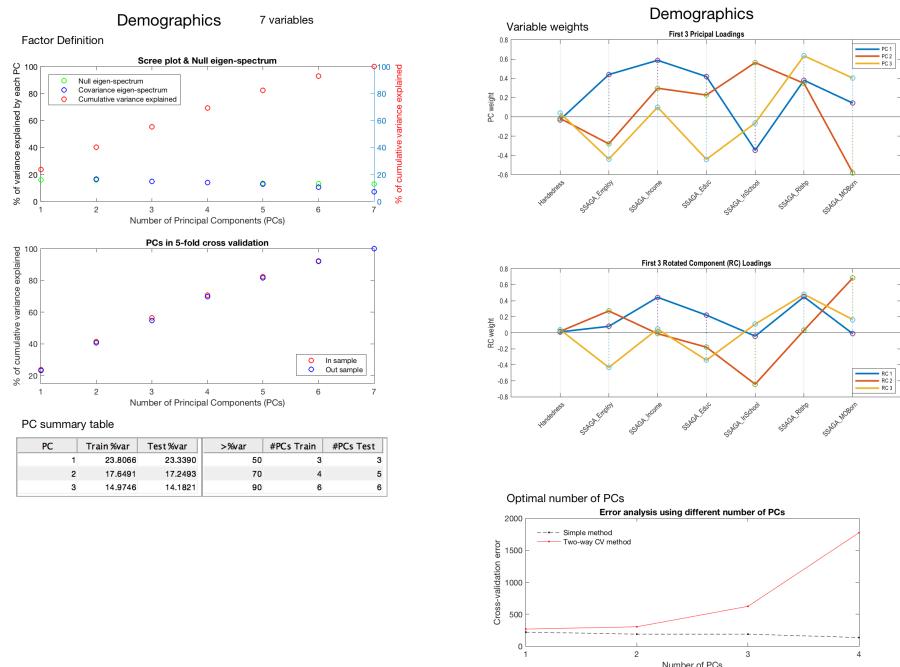


Figure A.11: Demographics and SES sub-domain summary report.

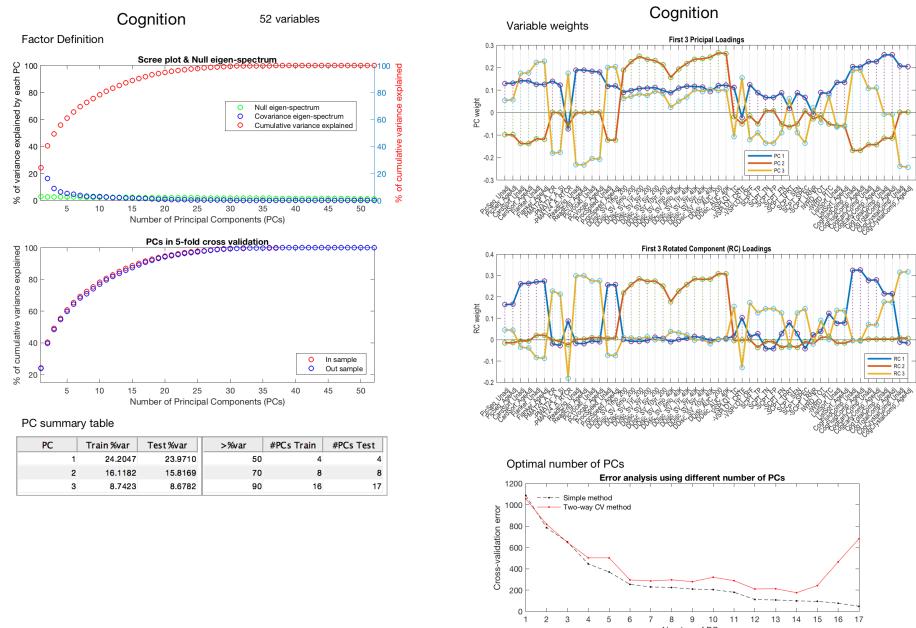


Figure A.12: Cognition sub-domain summary report.

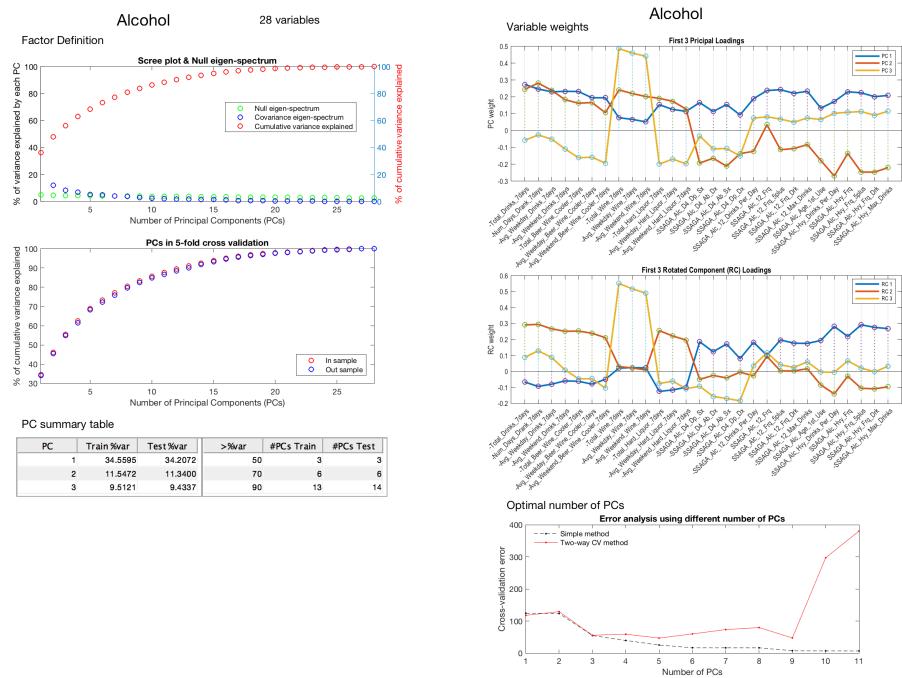


Figure A.13: Alcohol Use sub-domain summary report.

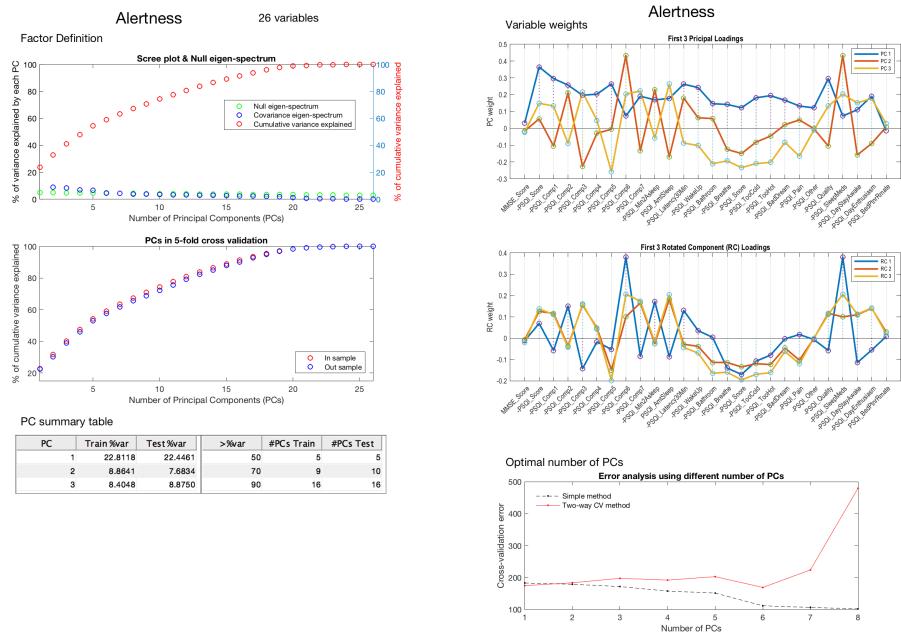


Figure A.14: Alertness sub-domain summary report.

A.4 Stability of SDR CCA canonical loading

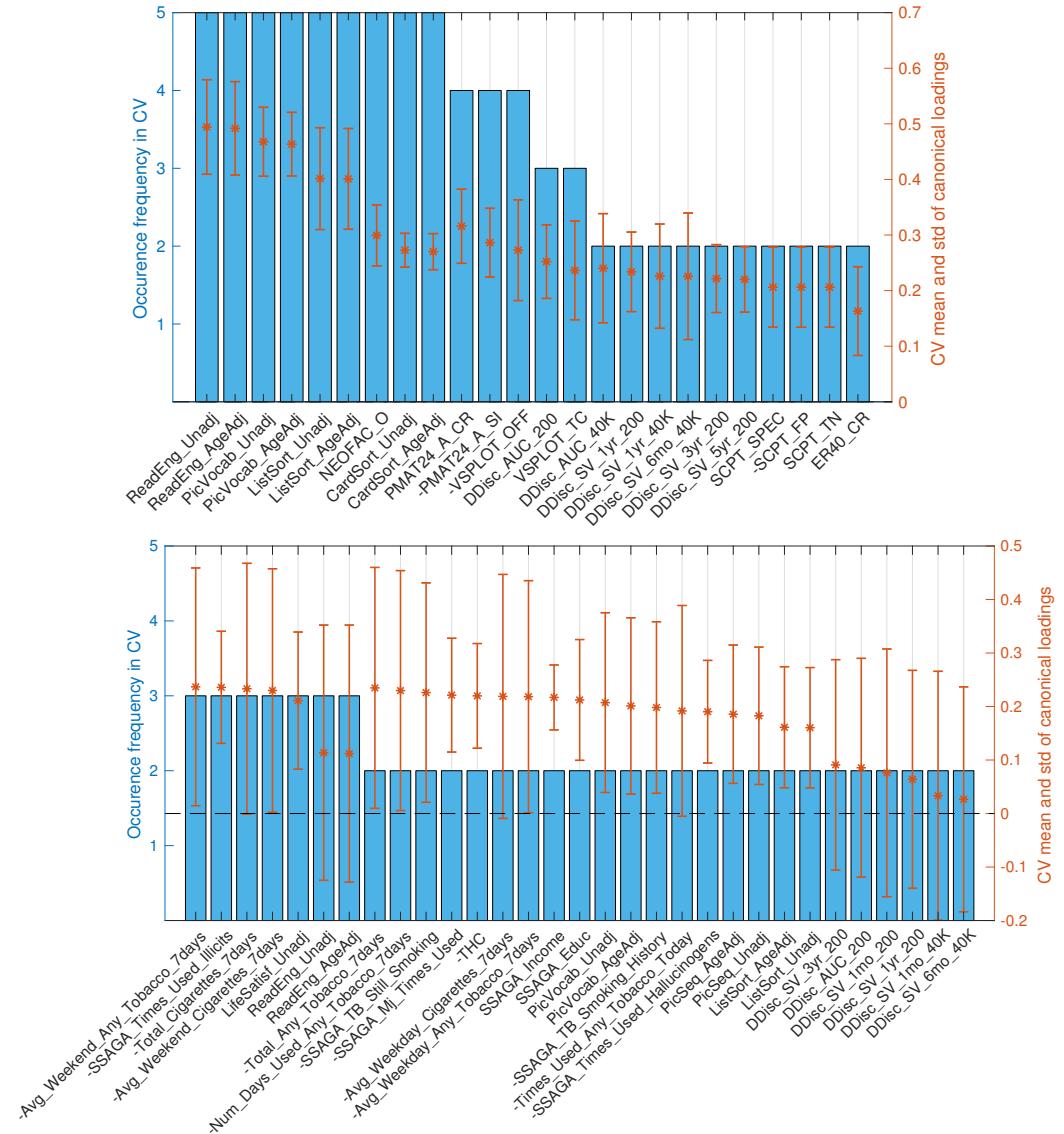


Figure A.15: Stability of SM canonical loadings on observed variables in SDR CCA. Bar plot shows the occurrence frequency in CV out of the 5 folds. Variables are chosen by selecting the top 20 mostly weighted ones in each fold. The ones appeared at least twice are shown above. Right axis shows the mean and the standard deviation over all occurred loadings. Top and bottom plots are the canonical loadings for the first and second canonical variables respectively. It is obvious that the second canonical loadings are less stable than the first set.

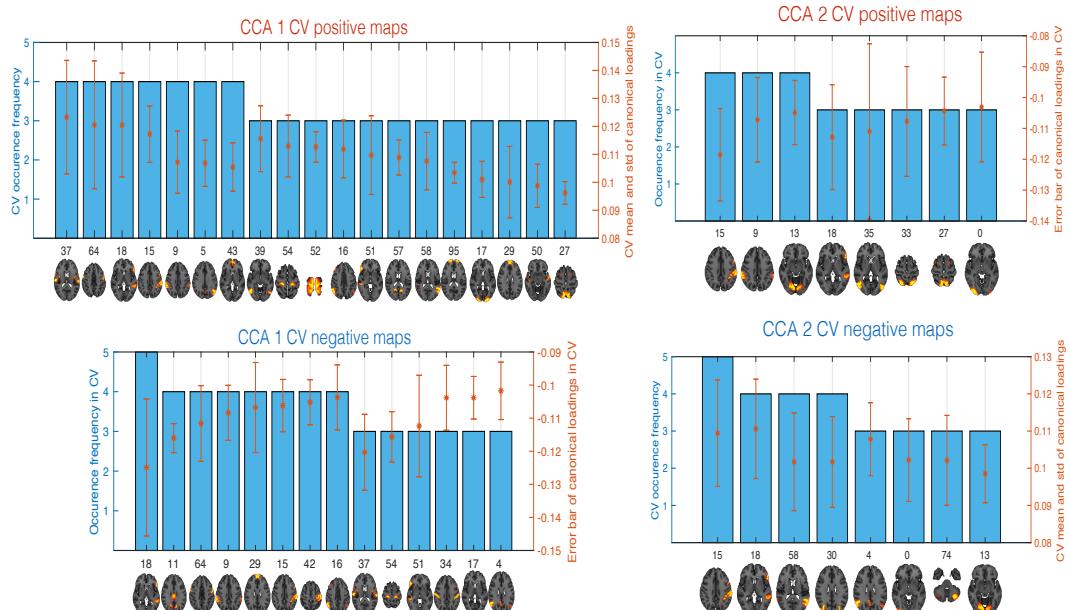


Figure A.16: Stability of BM canonical loadings on observed data in SDR CCA. Bar plot shows the occurrence frequency in CV out of the 5 folds. The positive (top plots) and negative (bottom plots) maps are chosen by first averaging the top 20 positive and negative canonical loadings within each region respectively; then select the top 20 nodes with the highest positive and negative mean loadings in each fold. The ones occurred at least three times are shown above. Right axis shows the mean and the standard deviation over all occurred loadings. Similar to SM canonical loadings, the first set shows better stability than the second set.

APPENDIX B

Appendix for UK Biobank Project

B.1 Rotated Loadings in SM Sub-domains

The following figures show the rotated principal loadings in the 9 SM sub-domain apart from sub-domain Physical Health which is displayed in Fig. 5.22.

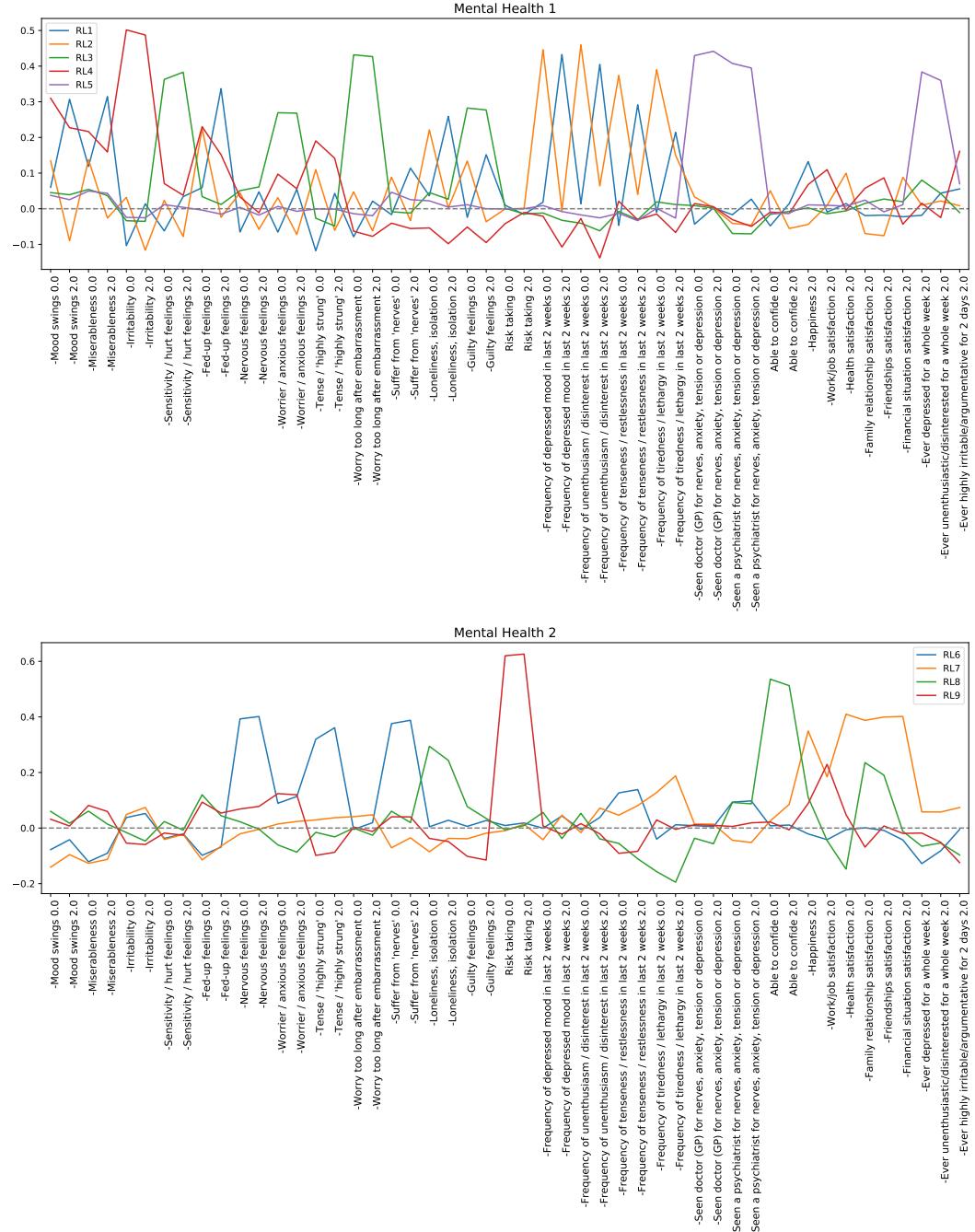


Figure B.1: Mental Health sub-domain rotated loadings.

Appendix for UK Biobank

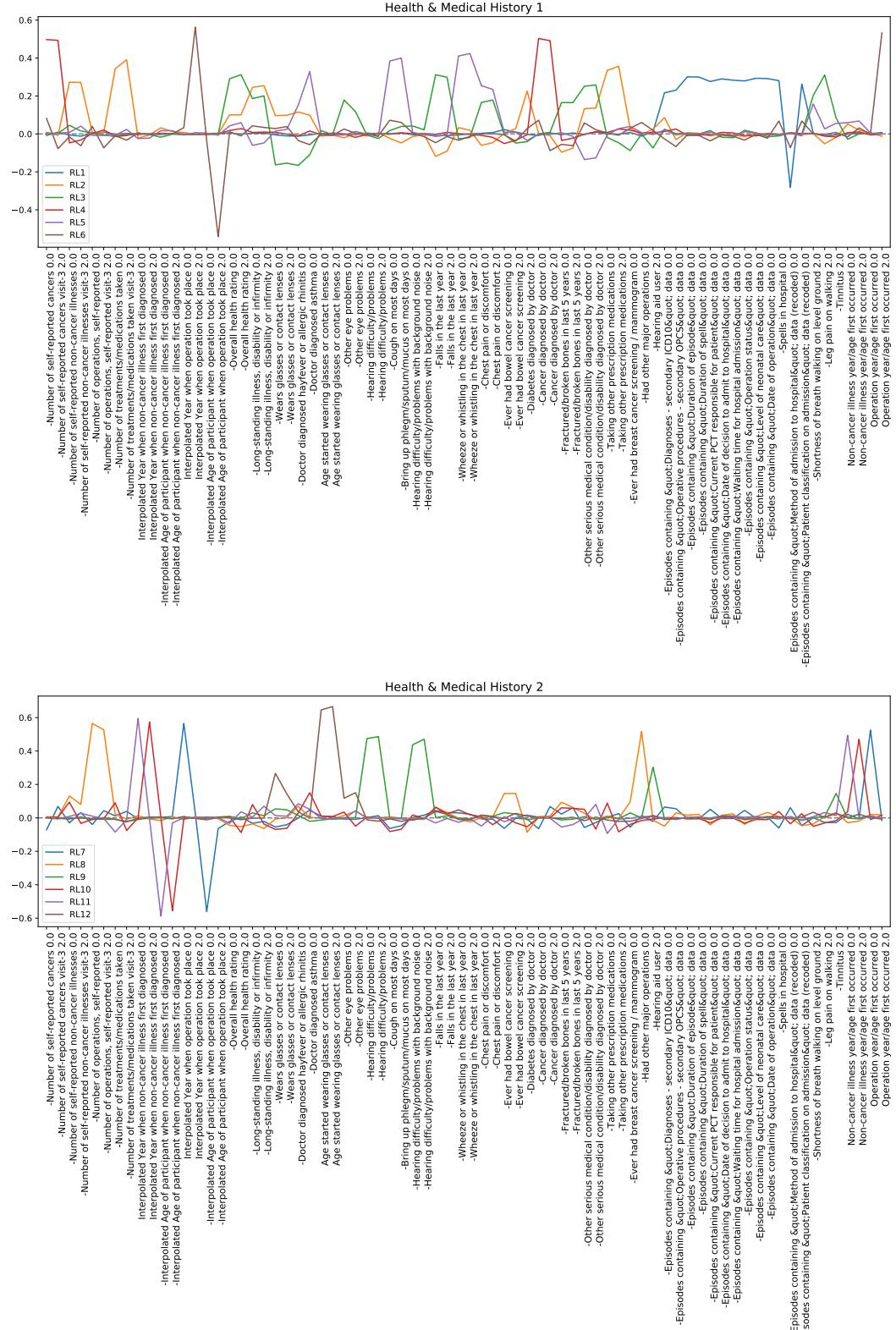


Figure B.2: Health & Medical History sub-domain rotated loadings.

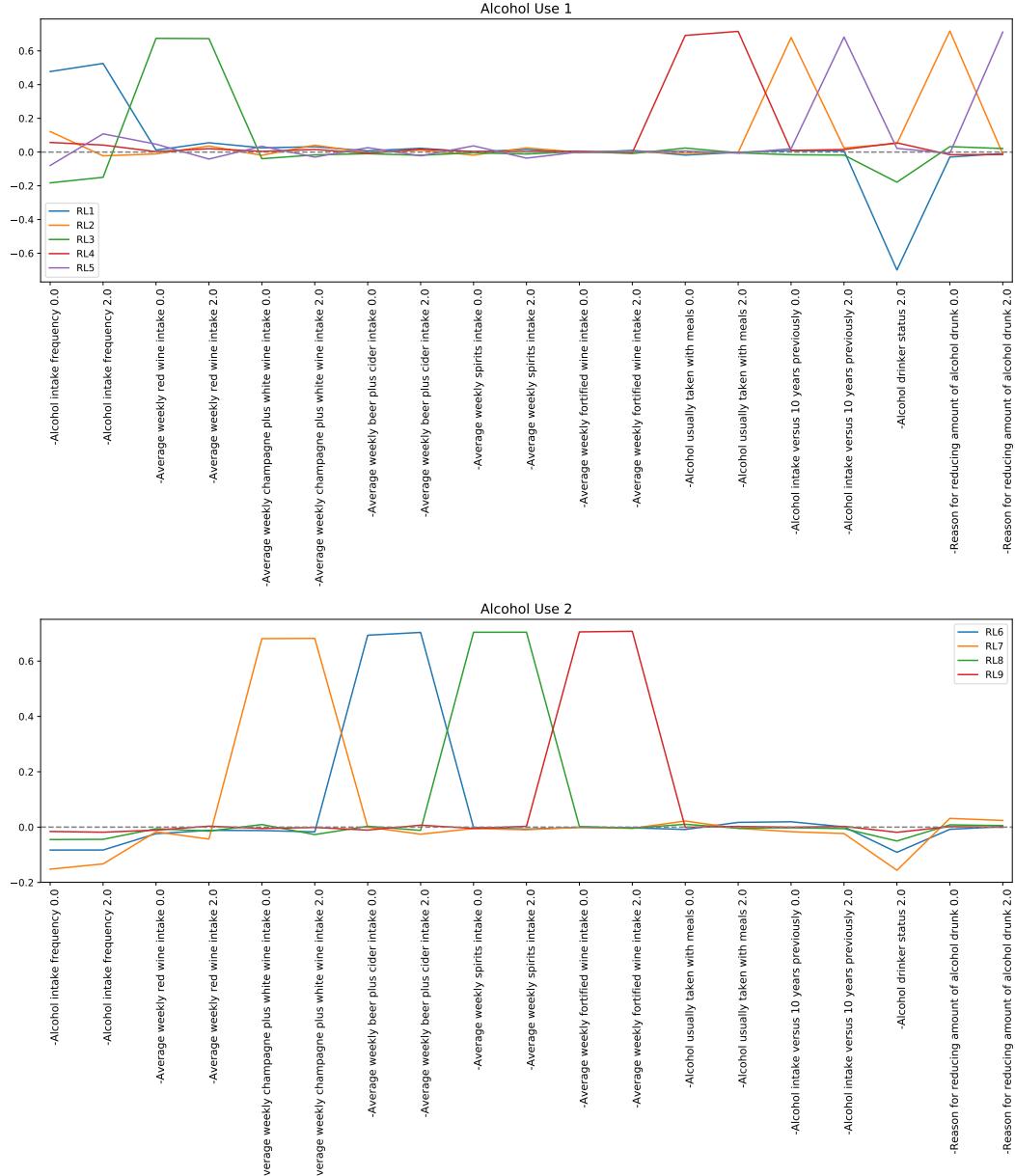


Figure B.3: Alcohol Use sub-domain rotated loadings.

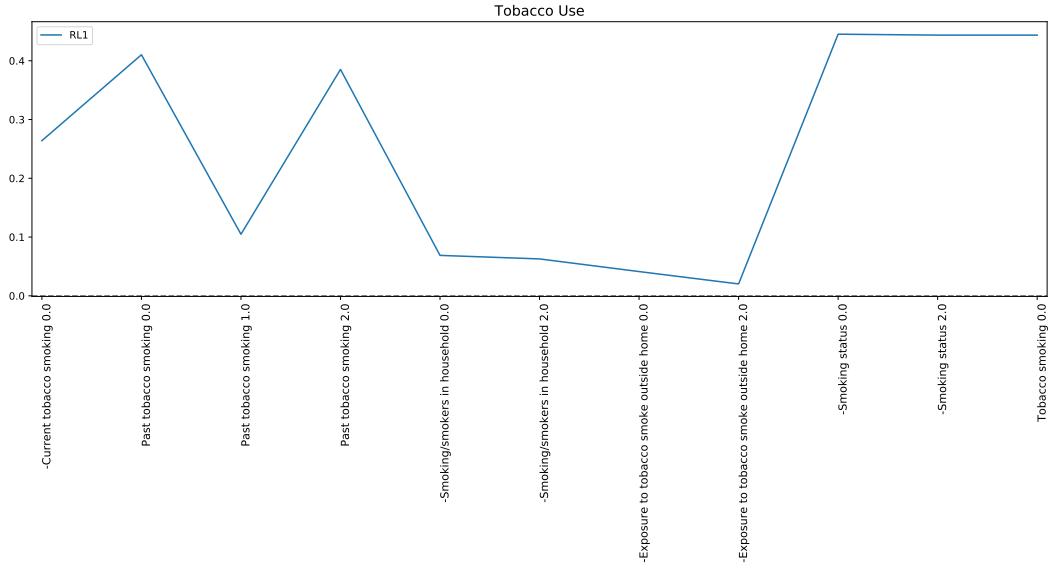


Figure B.4: Tobacco Use sub-domain rotated loadings.

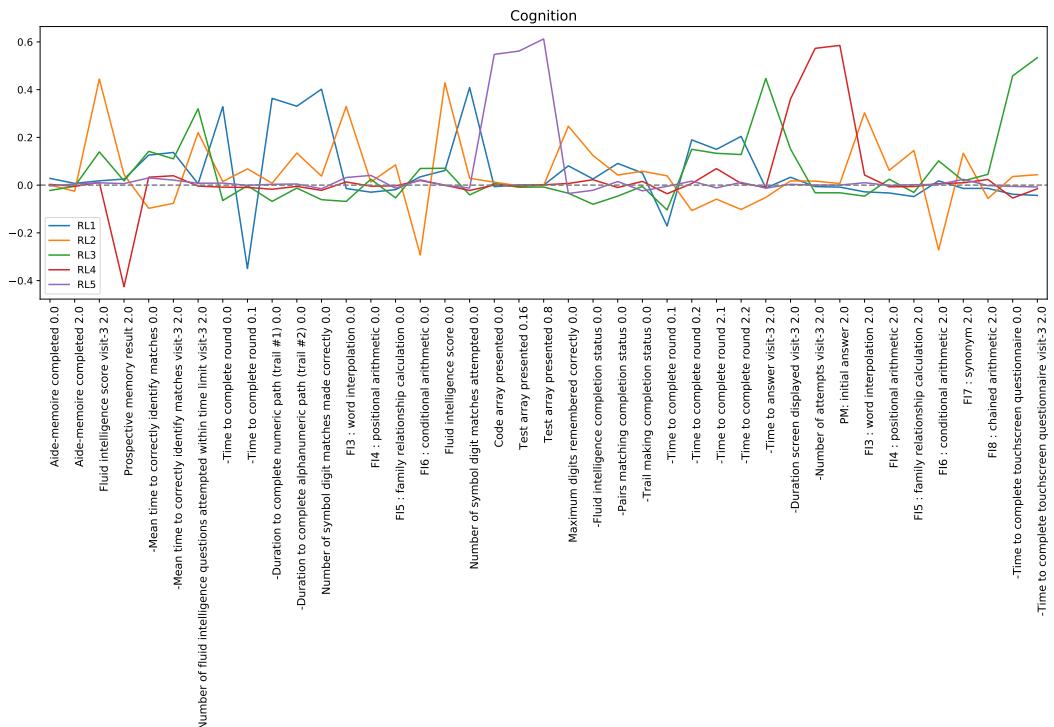


Figure B.5: Cognition sub-domain rotated loadings.

Appendix for UK Biobank B.1. ROTATED LOADINGS IN SM SUB-DOMAINS

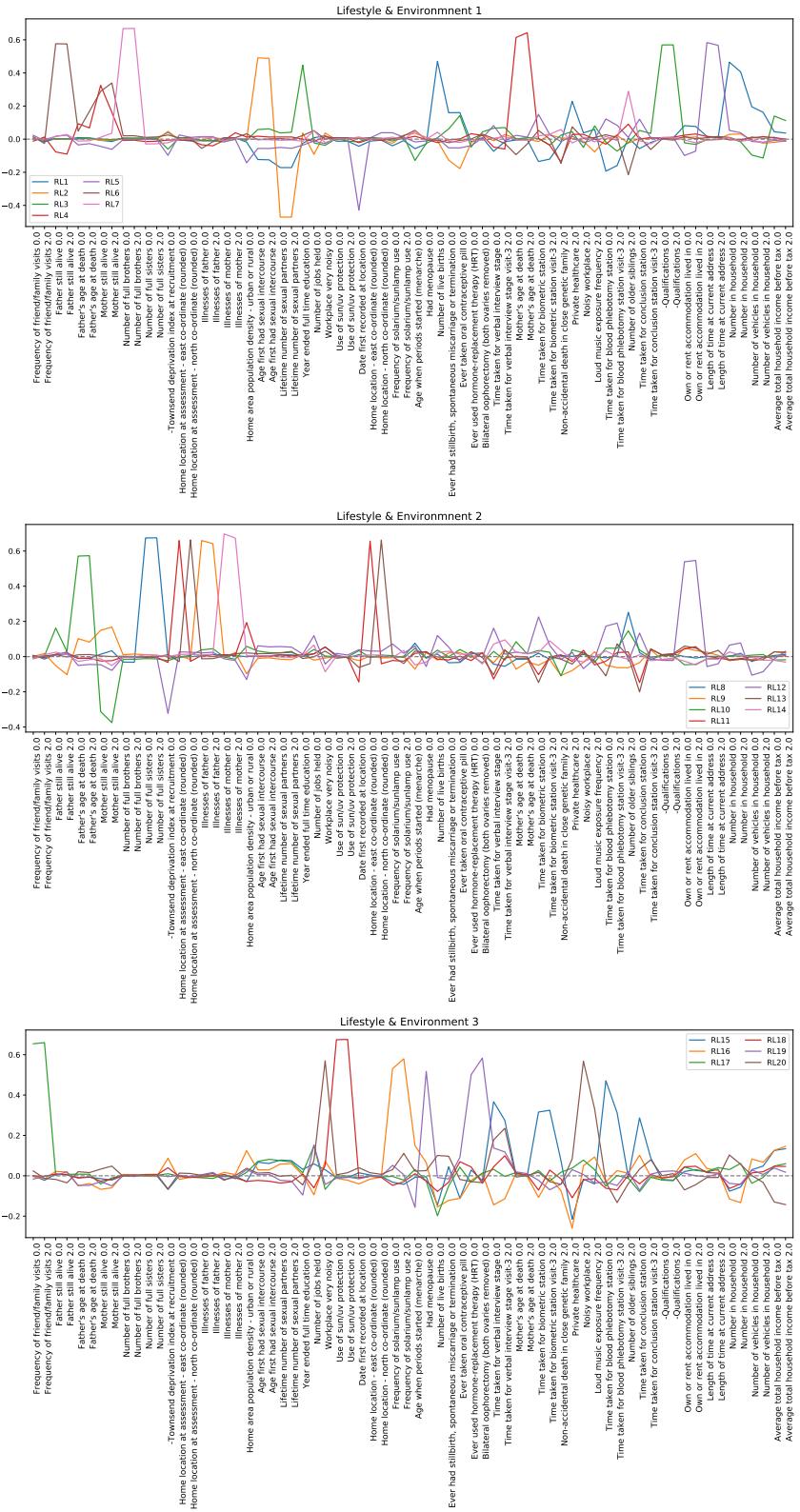


Figure B.6: Lifestyle & Environment sub-domain rotated loadings.

Appendix for UK Biobank

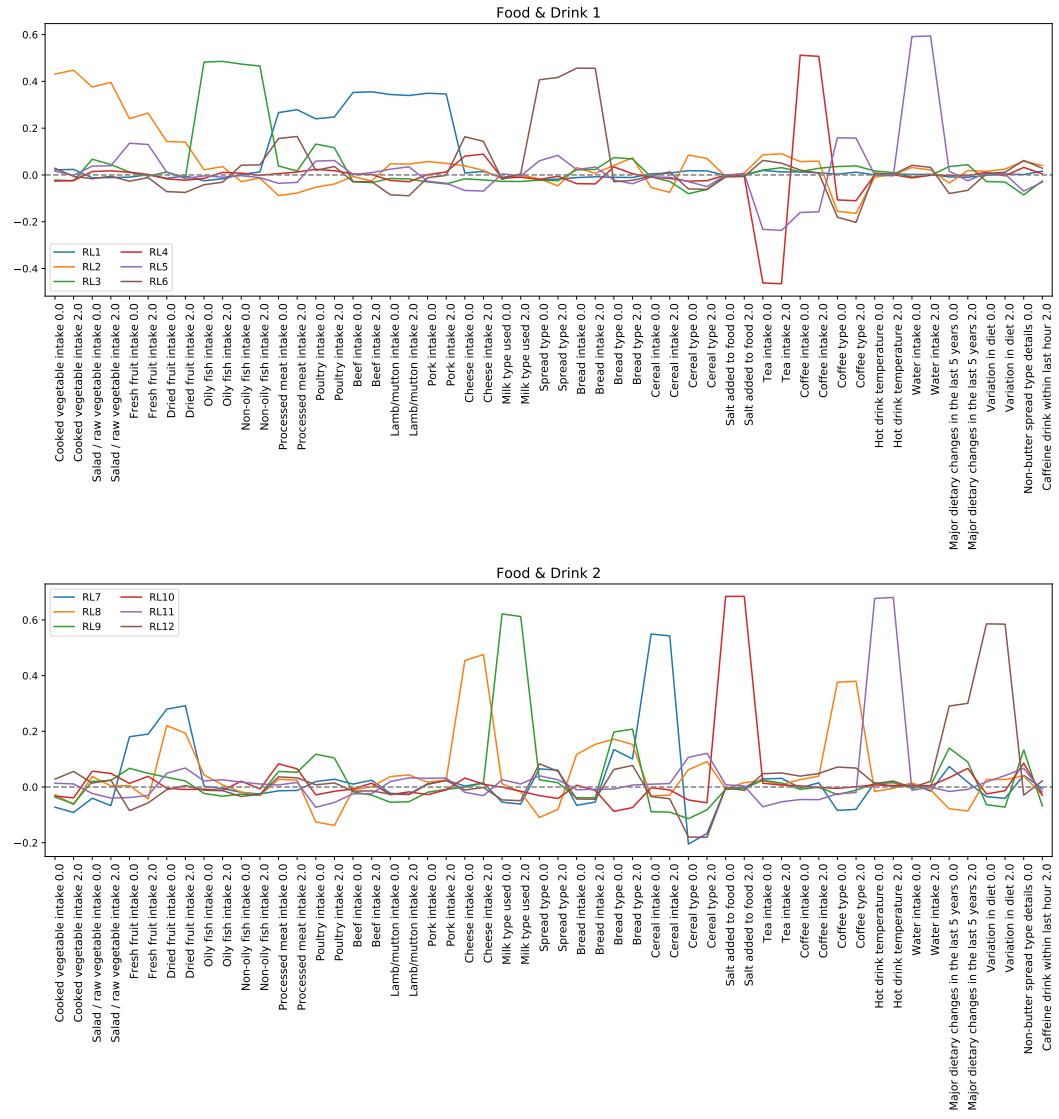


Figure B.7: Food & Drink sub-domain rotated loadings.

Appendix for UK Biobank B.1. ROTATED LOADINGS IN SM SUB-DOMAINS

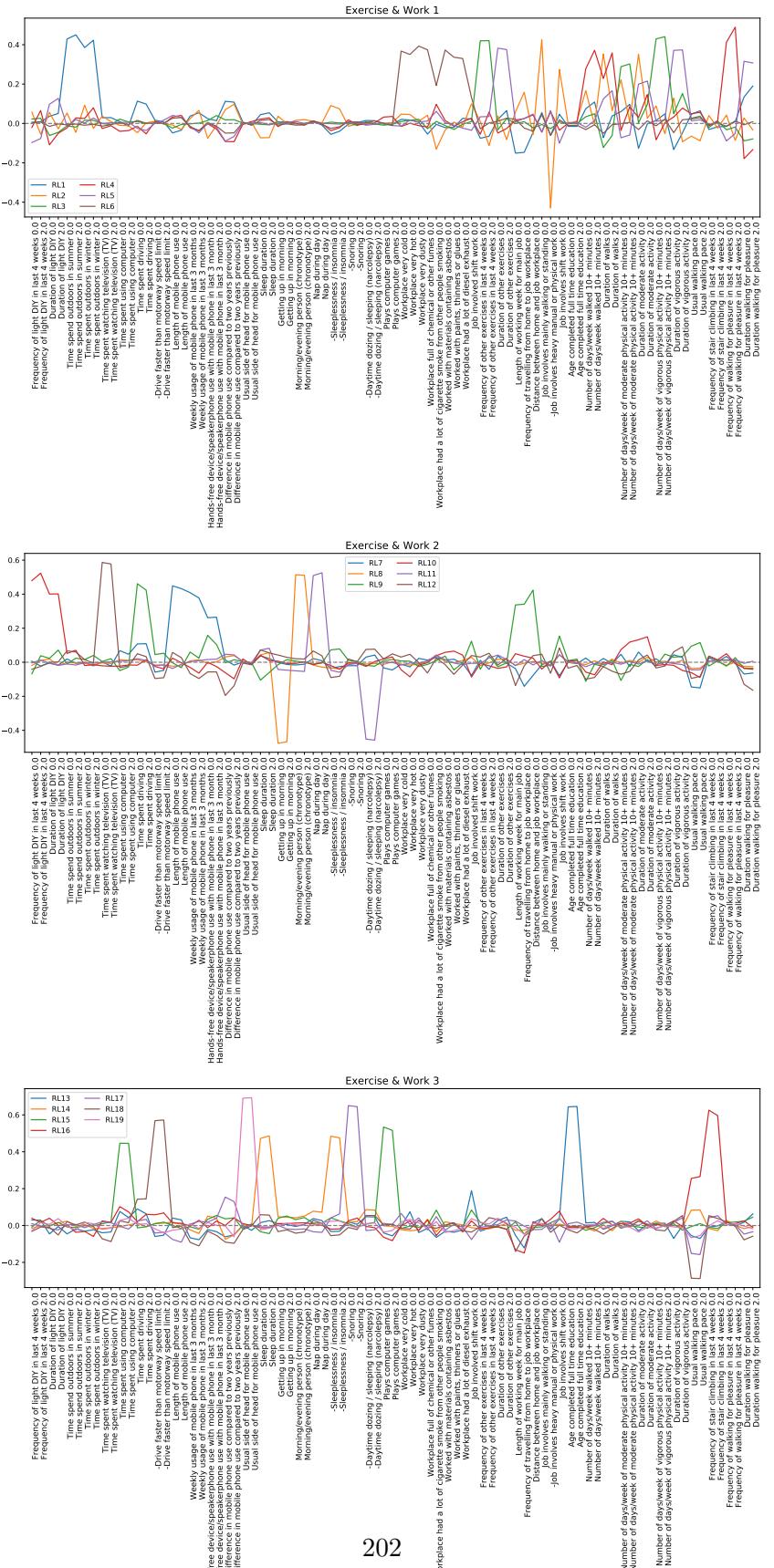


Figure B.8: Exercise & Work sub-domain rotated loadings.

B.2 CCA Results for Non-reduced Data

B.2.1 Canonical loadings between non-reduced SM and FC

Loading	Name	Loading	Name	Loading	Name	Loading	Name
0 0.734214	Body mass index (BMI) visit-3.2.0	0 0.406918	Fluid intelligence score visit-3.2.0	0 -0.252215	Pulse rate, automated reading 0.0	0 0.373445	Systolic blood pressure, automated reading vis...
1 0.670712	Body mass index (BMI) 0.0	1 0.376907	Weight visit-3.2.0	1 -0.234901	Pulse rate, automated reading visit-5 2.0	1 0.353149	Systolic blood pressure, automated reading vis...
2 0.667755	Leg fat mass (right) 0.0	2 0.375939	-Qualifications 0.0	2 -0.231920	Pulse rate, automated reading visit-2 0.1	2 0.316430	Diastolic blood pressure, automated reading vi...
3 0.662667	Waist circumference visit-3.2.0	3 0.370264	Fluid intelligence score 0.0	3 0.228062	Number of vehicles in household 2.0	3 0.306220	Systolic blood pressure, automated reading 0.0
4 0.658388	Leg fat percentage (right) 0.0	4 0.366745	Leg fat-free mass (right) 0.0	4 -0.225194	Pulse rate, automated reading visit-6 2.1	4 0.298669	Systolic blood pressure, automated reading vis...
5 0.637466	Whole body fat mass 0.0	5 0.365383	Weight 0.0	5 0.212439	Number of vehicles in household 0.0	5 0.298324	Diastolic blood pressure, automated reading vi...
6 0.632968	Weight visit-3.2.0	6 0.362716	Year ended full time education 0.0	6 -0.209059	Handedness (chirality/laterality) 0.0	6 0.255952	Forced vital capacity (FVC) visit-2 0.1
7 0.626562	Arm fat mass (left) 0.0	7 0.360578	-Qualifications 2.0	7 0.201202	Peak expiratory flow (PEF) visit-2 0.1	7 0.239272	Forced vital capacity (FVC) 0.0
8 0.625501	Arm fat mass (right) 0.0	8 0.358988	Whole body fat-free mass 0.0	8 -0.197953	Handedness (chirality/laterality) 2.0	8 0.237499	Forced vital capacity (FVC) visit-9 2.2
9 0.614080	Body fat percentage 0.0	9 0.359695	Waist circumference visit-3.2.0	9 0.196563	-Time to complete round 0.0	9 0.236156	Diastolic blood pressure, automated reading 0.0
10 0.609484	Wast circumference 0.0	10 0.352734	Waist circumference 0.0	10 -0.194716	-Time to complete round 0.1	10 0.233681	Forced expiratory volume in 1-second (FEV1) vi...
11 0.594852	Trunk fat mass 0.0	11 -0.324994	Time spent outdoors in summer 0.0	11 -0.187661	Impedance of arm (right) 0.0	11 0.231030	Forced expiratory volume in 1-second (FEV1) 0.0
12 0.590770	Arm fat percentage (left) 0.0	12 0.315371	Arm fat mass (right) 0.0	12 0.184256	Hand grip strength (right) 0.0	12 0.229427	Forced vital capacity (FVC) visit-8 2.1
13 0.587975	Arm fat percentage (right) 0.0	13 0.312250	Arm fat mass (left) 0.0	13 0.181378	Hand grip strength (left) visit-3 2.0	13 0.227404	Diastolic blood pressure, automated reading vi...
14 0.565113	Weight 0.0	14 -0.308995	Time spent outdoors in summer 2.0	14 0.181040	Peak expiratory flow (PEF) visit-7 2.0	14 0.214002	Forced vital capacity (FVC) visit-7 2.0
15 0.560300	Hip circumference visit-3.2.0	15 0.300668	Arm fat-free mass (right) 0.0	15 0.180443	-Trail making completion status 0.0	15 -0.205238	Lifetime number of sexual partners 2.0
16 0.555584	Trunk fat percentage 0.0	16 0.298401	Hip circumference 0.0	16 0.178027	-Duration screen displayed visit-3 2.0	16 0.200246	Forced vital capacity (FVC) visit-3 2.0
17 0.490437	Hip circumference 0.0	17 0.295339	Whole body fat mass 0.0	17 0.177898	Average total household income before tax 2.0	17 -0.195716	Lifetime number of sexual partners 0.0
18 0.367626	Arm fat-free mass (right) 0.0	18 -0.294967	Time spent outdoors in winter 2.0	18 0.176490	Hand grip strength (left) visit-3 2.0	18 -0.189522	Weekly usage of mobile phone in last 3 months 2.0
19 0.341210	Time spent watching television (TV) 0.0	19 0.293462	Body mass index (BMI) visit-3.2.0	19 -0.174416	Impedance of leg (right) 0.0	19 0.188384	Forced expiratory volume in 1-second (FEV1) vi...
20 0.339498	Diastolic blood pressure, automated reading 0.0	20 0.292969	Trunk fat mass 0.0	20 -0.173634	Impedance of whole body 0.0	20 0.181817	Forced expiratory volume in 1-second (FEV1) vi...
21 0.337385	Systolic blood pressure, automated reading vis...	21 0.283603	Hip circumference visit-3.2.0	21 0.173105	Number in household 2.0	21 0.175927	Forced expiratory volume in 1-second (FEV1) vi...
22 0.334840	Leg fat-free mass (right) 0.0	22 0.282685	Body mass index (BMI) 0.0	22 0.172054	-Nervous feelings 0.0	22 0.175045	Hand grip strength (right) visit-3 2.0
23 0.334266	Time spent watching television (TV) 0.0	23 0.281762	Leg fat mass (right) 0.0	23 -0.167952	-Time to complete round 0.1	23 -0.174100	Body mass index (BMI) 0.0
24 0.329736	Systolic blood pressure, automated reading vis...	24 -0.273620	Job involves mainly walking or standing 0.0	24 0.167318	-Worrier / anxious feelings 2.0	24 0.173850	Forced expiratory volume in 1-second (FEV1) vi...
25 0.325669	Whole body fat-free mass (right) 0.0	25 -0.260419	Impedance of leg (right) 0.0	25 0.166978	-Time to complete round 2.1	25 0.168602	Pulse rate, automated reading visit-5 2.0
26 0.323231	Diastolic blood pressure, automated reading vi...	26 0.256721	-Job involves heavy manual or physical work 0.0	26 0.166601	Peak expiratory flow (PEF) visit-8 2.1	26 -0.167158	Leg fat mass (right) 0.0
27 0.322341	Diastolic blood pressure, automated reading vi...	27 -0.256043	Impedance of leg (left) 0.0	27 0.165338	Peak expiratory flow (PEF) visit-9 2.2	27 -0.167010	Leg fat percentage (right) 0.0
28 0.314085	Diastolic blood pressure, automated reading vi...	28 -0.248768	Time spent outdoors in winter 0.0	28 -0.165011	Impedance of leg (left) 0.0	28 -0.163058	Weekly usage of mobile phone in last 3 months 0.0
29 -0.311503	Impedance of arm (left) 0.0	29 0.247667	Arm fat percentage (right) 0.0	29 0.165760	Peak expiratory flow (PEF) visit-3 2.0	29 0.161705	-Episodes containing "Diagnoses - secondary IC...
Loading	Name	Loading	Name	Loading	Name	Loading	Name
0 0.275709	Fluid intelligence score visit-3.2.0	0 -0.153895	Skin colour 0.0	0 0.274851	Standing height 0.0	0 0.274851	Skin colour 0.0
1 0.242532	Number of older siblings 2.0	1 0.152005	Forced expiratory volume in 1-second (FEV1) vi...	1 0.263558	Whole body fat-free mass 0.0	1 0.263558	Number of older siblings 2.0
2 -0.220930	Handedness (chirality/laterality) 2.0	2 0.148805	Forced vital capacity (FVC) visit-3 2.0	2 0.261789	Arm fat-free mass (right) 0.0	2 0.261789	Handedness (chirality/laterality) 2.0
3 0.205569	Number of fluid intelligence questions attempted 0.0	3 0.147333	-Fractured/broken bones in last 5 years 0.0	3 0.252256	Sitting height 0.0	3 0.252256	Number of fluid intelligence questions attempted 0.0
4 0.204167	Fluid intelligence score 0.0	4 0.142700	Length of working week for main job 0.0	4 0.251593	Sitting height visit-3 2.0	4 0.251593	Fluid intelligence score 0.0
5 -0.204134	-Time to complete round 0.1	5 0.139614	Mother still alive 2.0	5 0.235288	Leg fat-free mass (right) 0.0	5 0.235288	-Time to complete round 0.1
6 -0.202671	Handedness (chirality/laterality) 0.0	6 0.132413	Forced vital capacity (FVC) visit-8 2.1	6 0.182474	Weight 0.0	6 0.182474	Handedness (chirality/laterality) 0.0
7 0.197255	-Duration to complete alphanumeric path (trail) ...	7 0.131375	Forced expiratory volume in 1-second (FEV1) vi...	7 0.176656	-Time to complete touchscreen questionnaire vi...	7 0.176656	-Duration to complete alphanumeric path (trail) ...
8 0.193600	Pulse rate, automated reading visit-2 2.1	8 -0.124939	Hair colour (natural, before greying) 0.0	8 0.172712	Weight visit-3 2.0	8 0.172712	Pulse rate, automated reading visit-2 2.1
9 0.187894	Pulse rate, automated reading visit-2 2.0	9 0.121490	Forced vital capacity (FVC) visit-9 2.2	9 0.165454	-Duration to complete alphanumeric path (trail) ...	9 0.165454	Pulse rate, automated reading visit-2 2.0
10 -0.175573	Systolic blood pressure, automated reading vis...	10 0.116734	FIS - word interpretation 0.0	10 0.157236	Tobacco smoking 0.0	10 0.157236	Systolic blood pressure, automated reading vis...
11 0.175121	Number of symbol digit matches attempted 0.0	11 -0.116528	Skin colour 0.0	11 0.154283	Number of symbol digit matches attempted 0.0	11 0.154283	Number of symbol digit matches attempted 0.0
12 0.168754	-Time to complete round 0.1	12 0.116107	Oily fish intake 0.0	12 0.153668	-Current tobacco smoking 0.0	12 0.153668	-Time to complete round 0.1
13 0.168401	Number of symbol digit matches made correctly 0.0	13 0.115956	Forced vital capacity (FVC) visit-2 2.1	13 0.151802	Age completed full time education 2.0	13 0.151802	Number of symbol digit matches made correctly 0.0
14 0.166062	Forced vital capacity (FVC) visit-7 2.0	14 0.113781	Forced expiratory volume in 1-second (FEV1) vi...	14 0.151735	Number of symbol digit matches made correctly 0.0	14 0.151735	Forced vital capacity (FVC) visit-7 2.0
15 0.163644	Maximum digits remembered correctly 0.0	15 -0.112776	Ordering of blouse 2.2	15 -0.150182	-Time to complete round 0.1	15 -0.150182	Maximum digits remembered correctly 0.0
16 -0.160714	-Time to complete round 0.1	16 0.112130	Forced vital capacity (FVC) 0.0	16 0.142751	-Smoking status 2.0	16 0.142751	-Time to complete round 0.1
17 0.158915	Forced vital capacity (FVC) visit-8 2.1	17 0.111909	-Smoking status 2.1	17 0.141034	Past tobacco smoking 0.0	17 0.141034	Forced vital capacity (FVC) visit-8 2.1
18 0.156953	Forced expiratory volume in 1-second (FEV1) vi...	18 0.111529	Forced expiratory volume in 1-second (FEV1) vi...	18 0.138178	Forced expiratory volume in 1-second (FEV1) 0.0	18 0.138178	Forced expiratory volume in 1-second (FEV1) vi...
19 0.155772	Pulse rate, automated reading 0.0	19 -0.110621	-Alcohol drinker status 2.0	19 0.136371	Forced expiratory volume in 1-second (FEV1) vi...	19 0.136371	Pulse rate, automated reading 0.0
20 0.152951	-Duration screen displayed visit-3 2.0	20 0.110569	Nap during day 0.0	20 0.136507	-Time to complete touchscreen questionnaire 0.0	20 0.136507	-Duration screen displayed visit-3 2.0
21 0.152506	Pulse rate, automated reading visit-2 0.1	21 -0.107791	Hair colour (natural, before greying) 2.0	21 0.135827	Forced expiratory volume in 1-second (FEV1) vi...	21 0.135827	Pulse rate, automated reading visit-2 0.1
22 0.152414	-Time to complete round 0.2	22 0.104851	Forced vital capacity (FVC) visit-7 2.2	22 0.134723	-Time to complete round 2.2	22 0.134723	-Time to complete round 0.2
23 0.148084	Forced expiratory volume in 1-second (FEV1) vi...	23 -0.104268	-Ever unenthusiastic/disinterested for a whole...	23 0.134357	-Smoking status 0.0	23 0.134357	Forced expiratory volume in 1-second (FEV1) vi...
24 0.147246	Sitting height visit-3 2.0	24 0.103389	Sleep duration 2.0	24 0.134236	Past tobacco smoking 2.0	24 0.134236	Sitting height visit-3 2.0
25 -0.143474	Systolic blood pressure, automated reading vis...	25 -0.103017	-Cough on most days 0.0	25 -0.132197	Impedance of whole body 0.0	25 -0.132197	Systolic blood pressure, automated reading vis...
26 0.142427	-Duration to complete numeric path (trial #1) 0.0	26 0.101723	Standing height 0.0	26 -0.131114	Impedance of arm (right) 0.0	26 -0.131114	-Duration to complete numeric path (trial #1) 0.0
27 0.141751	Forced expiratory volume in 1-second (FEV1) vi...	27 -0.100678	Water intake 2.0	27 0.127081	Age completed full time education 0.0	27 0.127081	Forced expiratory volume in 1-second (FEV1) vi...
28 0.140100	Time spent watching television (TV) 0.0	28 -0.098068	-Tinnitus 2.0	28 -0.125364	Plays computer games 0.0	28 -0.125364	Time spent watching television (TV) 0.0
29 0.139619	Number of full brothers 0.0	29 0.097784	-Current tobacco smoking 0.0	29 -0.124642	Impedance of arm (left) 0.0	29 -0.124642	Number of full brothers 0.0

Figure B.9: Top 30 canonical loadings for the 7 significant SM canonical variables in the CCA of FC and SM. From top left to bottom right are the first to the seventh canonical loadings respectively. Variables that are sign-flipped have a '-' sign in front of their names.

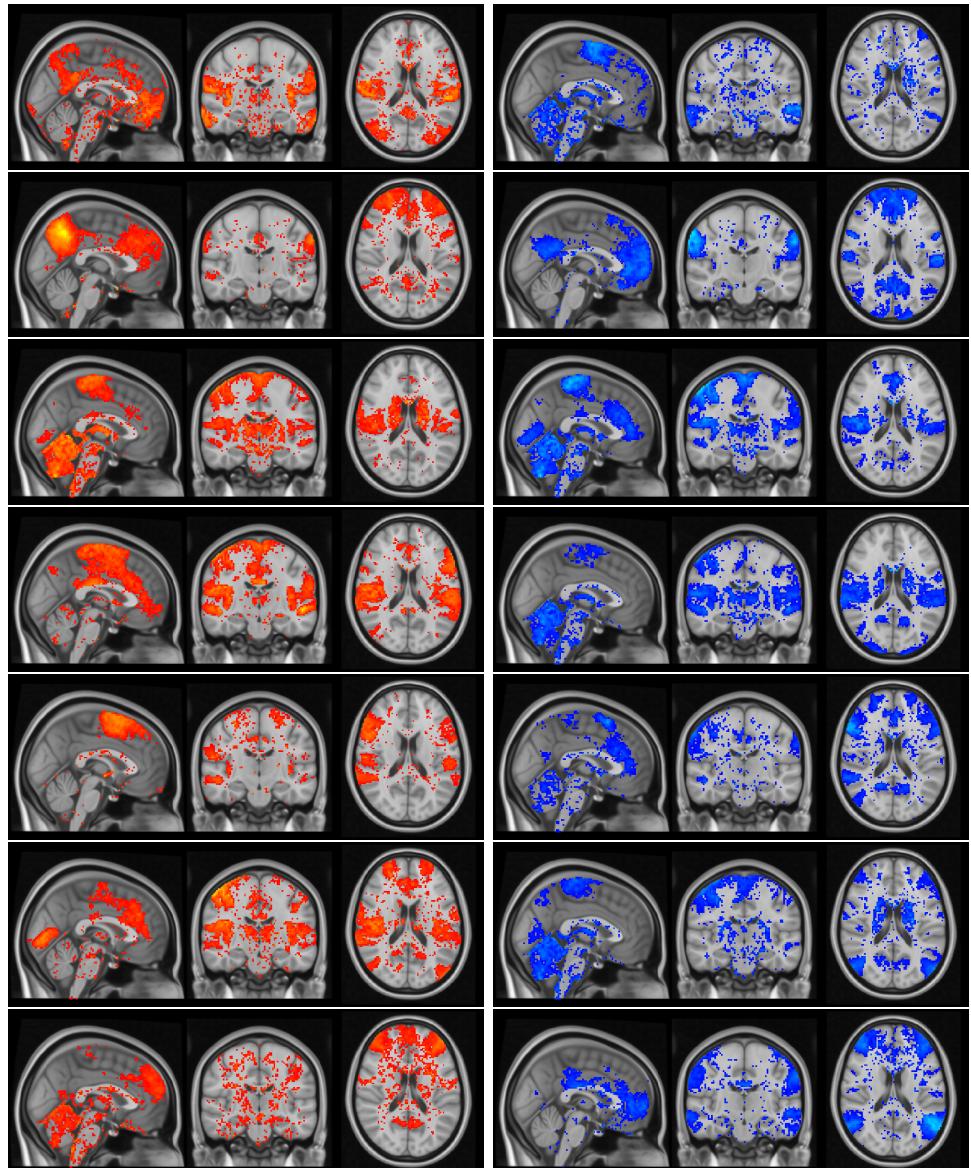


Figure B.10: Canonical loaded maps for the 7 significant FC canonical variables in the CCA with SM. From top to bottom are the first to the seventh canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).

B.2.2 Canonical loadings between non-reduced SM and IDP

Loading	Name	Loading	Name	Loading	Name
0 0.742223	Standing height 0.0	0 0.773736	Body mass index (BMI) visit-3 2.0	0 0.350361	Number of older siblings 2.0
1 0.701486	Sitting height visit-3 2.0	1 0.717088	Body mass index (BMI) 0.0	1 -0.216803	Systolic blood pressure, automated reading vis...
2 0.667639	Sitting height 0.0	2 0.694244	Leg fat mass (right) 0.0	2 0.202603	Number of full brothers 2.0
3 0.598161	Whole body fat-free mass 0.0	3 0.692562	Waist circumference visit-3 2.0	3 -0.199550	Impedance of arm (right) 0.0
4 0.548353	Arm fat-free mass (right) 0.0	4 0.690190	Arm fat mass (right) 0.0	4 -0.196312	Systolic blood pressure, automated reading vis...
5 0.535920	Leg fat-free mass (right) 0.0	5 0.689649	Whole body fat mass 0.0	5 0.195951	Number of full brothers 0.0
6 0.487776	Forced vital capacity (FVC) visit-7 2.0	6 0.688332	Arm fat mass (left) 0.0	6 -0.194578	Pulse rate, automated reading visit-5 2.0
7 0.474984	Forced vital capacity (FVC) visit-8 2.1	7 0.680275	Weight visit-3 2.0	7 -0.194388	Pulse rate, automated reading visit-6 2.1
8 0.460435	Forced vital capacity (FVC) 0.0	8 0.661406	Leg fat percentage (right) 0.0	8 0.184021	Number of full sisters 2.0
9 0.443206	Forced vital capacity (FVC) visit-2 0.1	9 0.658176	Waist circumference 0.0	9 0.179401	Number of full sisters 0.0
10 0.436841	Forced expiratory volume in 1-second (FEV1) ...	10 0.656309	Body fat percentage 0.0	10 -0.173414	Impedance of arm (left) 0.0
11 0.429373	Weight visit-3 2.0	11 0.655832	Trunk fat mass 0.0	11 0.164893	Number of live births 0.0
12 0.420790	Weight 0.0	12 0.649971	Arm fat percentage (right) 0.0	12 -0.163776	Impedance of whole body 0.0
13 0.418899	Forced expiratory volume in 1-second (FEV1) 0.0	13 0.643061	Arm fat percentage (left) 0.0	13 -0.162124	Systolic blood pressure, automated reading 0.0
14 0.414279	Forced vital capacity (FVC) visit-9 2.2	14 0.634179	Weight 0.0	14 -0.155215	Impedance of leg (right) 0.0
15 0.412398	Forced expiratory volume in 1-second (FEV1) v...	15 0.609134	Hip circumference visit-3 2.0	15 -0.153289	Systolic blood pressure, automated reading vis...
16 0.402414	Forced expiratory volume in 1-second (FEV1) v...	16 0.605914	Trunk fat percentage 0.0	16 -0.142316	F13 : word interpolation 0.0
17 0.385997	Forced vital capacity (FVC) visit-3 0.2	17 0.525538	Hip circumference 0.0	17 -0.139353	Age completed full time education 0.0
18 0.361028	Forced expiratory volume in 1-second (FEV1) v...	18 0.405920	Arm fat-free mass (right) 0.0	18 -0.133053	Pulse rate, automated reading 0.0
19 0.358954	Fluid intelligence score visit-3 2.0	19 0.401940	Leg fat-free mass (right) 0.0	19 -0.124589	Fluid intelligence score 0.0
20 0.351569	Forced expiratory volume in 1-second (FEV1) v...	20 0.376679	Whole body fat-free mass 0.0	20 -0.121945	Impedance of leg (left) 0.0
21 0.333435	Fluid intelligence score 0.0	21 -0.322438	Impedance of arm (left) 0.0	21 0.120001	Arm fat-free mass (right) 0.0
22 0.283704	Hand grip strength (left) visit-3 2.0	22 -0.320238	Impedance of whole body 0.0	22 -0.118964	Age first had sexual intercourse 2.0
23 0.277239	Hip circumference 0.0	23 -0.310658	Impedance of arm (right) 0.0	23 0.118280	Job involves mainly walking or standing 0.0
24 0.273542	-Duration to complete alphanumeric path (trail...	24 0.298114	-Snoring 0.0	24 -0.117296	Pulse rate, automated reading visit-2 0.1
25 0.269085	Hip circumference visit-3 2.0	25 -0.291811	-Number of treatments/medications taken vis...	25 0.111008	Time spent outdoors in winter 2.0
26 0.263833	Hand grip strength (right) visit-3 2.0	26 -0.287347	Usual walking pace 2.0	26 -0.108563	-Smoking/smokers in household 2.0
27 0.257783	Hand grip strength (left) 0.0	27 -0.277257	Impedance of leg (right) 0.0	27 -0.108287	Frequency of friend/family visits 0.0
28 0.254732	Hand grip strength (right) 0.0	28 -0.274443	Impedance of leg (left) 0.0	28 -0.107268	Fluid intelligence score visit-3 2.0
29 0.242556	Peak expiratory flow (PEF) 0.0	29 0.265432	Time spent watching television (TV) 2.0	29 0.100290	Nap during day 2.0
Loading	Name	Loading	Name	Loading	Name
0 0.259253	Number of older siblings 2.0	0 -0.163941	Number of full brothers 0.0	0 -0.196698	Number of older siblings 2.0
1 0.233019	Systolic blood pressure, automated reading vis...	1 -0.160241	Number of full brothers 2.0	1 0.192053	-Duration to complete numeric path (trial #1) 0.0
2 0.217541	Systolic blood pressure, automated reading vis...	2 0.156577	Pulse rate, automated reading visit-5 2.0	2 0.186782	F13 : word interpolation 0.0
3 0.212379	Sitting height visit-3 2.0	3 0.149230	Pulse rate, automated reading visit-6 2.1	3 0.176990	Number of symbol digit matches attempted 0.0
4 0.209827	Pulse rate, automated reading visit-6 2.1	4 0.146019	-Time to complete round 0.1	4 0.175776	Number of symbol digit matches made correctly 0.0
5 0.198027	Pulse rate, automated reading visit-5 2.0	5 -0.145498	-Mood swings 2.0	5 0.152664	Fluid intelligence score 0.0
6 0.196176	Sitting height 0.0	6 -0.140020	-Number of self-reported non-cancer illnesses ...	6 0.144895	-Duration to complete alphanumeric path (trail...
7 0.176812	Standing height 0.0	7 -0.138858	-Diabetes diagnosed by doctor 2.0	7 0.144231	Fluid intelligence score visit-3 2.0
8 0.174603	Number of full brothers 0.0	8 -0.137751	-Number of treatments/medications taken vis...	8 0.142699	Past tobacco smoking 0.0
9 0.174518	Systolic blood pressure, automated reading vis...	9 -0.136362	-Current tobacco smoking 0.0	9 0.142050	-Other serious medical condition/disability di...
10 0.174176	Number of full brothers 2.0	10 -0.134202	-Duration to complete alphanumeric path (trail...	10 0.138913	-Smoking status 2.0
11 -0.171339	Age completed full time education 2.0	11 -0.132316	Able to confide 2.0	11 0.138268	Time taken for blood phlebotomy station visit...
12 0.168133	Time spent watching television (TV) 0.0	12 -0.129349	-Smoking status 0.0	12 0.135905	Hand grip strength (right) visit-3 2.0
13 -0.167257	Year ended full time education 0.0	13 -0.126894	Number of symbol digit matches made correctly 0.0	13 0.135149	Tobacco smoking 0.0
14 0.167172	Systolic blood pressure, automated reading 0.0	14 -0.125593	Tobacco smoking 0.0	14 0.130535	Age completed full time education 0.0
15 -0.162633	Age completed full time education 0.0	15 0.124507	Time taken for biometric station visit-3 2.0	15 0.130250	Hand grip strength (left) visit-3 2.0
16 0.150447	Pulse rate, automated reading visit-2 0.1	16 0.124058	Pulse rate, automated reading visit-2 0.1	16 0.127315	-Time to complete touchscreen questionnaire 0.0
17 0.150315	Impedance of leg (left) 0.0	17 -0.123625	Number of symbol digit matches attempted 0.0	17 0.127250	Past tobacco smoking 2.0
18 0.147724	Impedance of leg (right) 0.0	18 -0.123613	Duration of vigorous activity 2.0	18 -0.126841	-Time to complete round 0.1
19 0.145571	Pulse rate, automated reading 0.0	19 -0.121186	Number of live births 0.0	19 0.125239	-Exposure to tobacco smoke outside home 2.0
20 -0.144646	-Qualifications 2.0	20 -0.120841	-Smoking status 2.0	20 0.123040	-Smoking status 0.0
21 0.141075	Impedance of whole body 0.0	21 -0.117424	-Irritability 2.0	21 0.121769	-Average weekly spirits intake 2.0
22 0.135098	Number of full sisters 0.0	22 0.117185	Frequency of friend/family visits 0.0	22 0.120723	Major dietary changes in the last 5 years 2.0
23 -0.133965	-Qualifications 0.0	23 -0.116656	-Number of treatments/medications taken 0.0	23 -0.120285	Sitting height 0.0
24 0.133677	Time spent watching television (TV) 2.0	24 -0.116388	-Duration screen displayed visit-3 2.0	24 0.119539	Age completed full time education 2.0
25 -0.130180	-Average weekly beer plus cider intake 0.0	25 -0.115317	F14 : positional arithmetic 0.0	25 -0.119355	Number of full brothers 2.0
26 -0.127883	Age first had sexual intercourse 2.0	26 -0.114909	-Time to complete round 0.0	26 -0.117597	Number of full sisters 2.0
27 0.127650	Number of full sisters 2.0	27 -0.114028	-Seen a psychiatrist for nerves, anxiety, tens...	27 0.116763	-Time to complete round 0.0
28 -0.126703	Age first had sexual intercourse 0.0	28 0.111662	-Alcohol intake versus 10 years previously 0.0	28 -0.115985	Number of full sisters 0.0
29 -0.125454	Time spent using computer 0.0	29 0.110863	Standing height 0.0	29 -0.114843	Number of full brothers 0.0

Figure B.11: Top 30 canonical loadings for the 6 significant SM canonical variables in the CCA of IDP and SM. From top left to bottom right are the first to the sixth canonical loadings respectively. Variables that are sign-flipped have a ‘-’ sign in front of their names.

Loading	Name	Loading	Name	Loading	Name
0 -0.606769	Volumetric scaling from T1 head image to stand...	0 -0.343047	Volume of grey matter in Vermis Cerebellum (right)	0 0.164779	Mean MO in anterior corona radiata on FA skele...
1 0.547882	Volume of brain, grey+white	1 -0.295969	Volume of grey matter in Vllb Cerebellum (right)	1 0.150357	Volume of grey matter in Parahippocampal Gyrus...
2 0.543187	Volume of grey matter	2 -0.289646	Volume of grey matter in Villa Cerebellum (left)	2 -0.146043	Mean OD in uncinate fasciculus on FA skeleton ...
3 0.526676	Volume of peripheral cortical grey matter	3 -0.287899	Median T2star in hippocampus (left)	3 0.140508	Volume of grey matter in Temporal Fusiform Cor...
4 0.497269	Volume of white matter	4 -0.273847	Volume of grey matter in Vllb Cerebellum (left)	4 -0.137846	Mean ISOVF in anterior corona radiata on FA sk...
5 0.482325	Volume of thalamus (right)	5 -0.271427	Volume of grey matter in IX Cerebellum (right)	5 0.137616	Mean FA in uncinate fasciculus on FA skeleton ...
6 0.460416	Volume of thalamus (left)	6 -0.269673	Volume of grey matter in IX Cerebellum (left)	6 -0.134533	Mean L2 in anterior corona radiata on FA ske...
7 0.449897	Volume of grey matter in Subcallosal Cortex (...)	7 -0.262112	Median T2star in hippocampus (right)	7 -0.132805	Mean ISOVF in anterior corona radiata on FA sk...
8 0.429820	Volume of grey matter in Subcallosal Cortex (...)	8 -0.253389	Volume of grey matter in Vllb Cerebellum (right)	8 0.132370	Volume of grey matter in Temporal Fusiform Cor...
9 0.415507	Volume of grey matter in Frontal Pole (right)	9 -0.234599	Volume of grey matter in Vllb Cerebellum (left)	9 -0.130959	Mean L2 in anterior corona radiata on FA ske...
10 0.412847	Volume of grey matter in Frontal Pole (left)	10 -0.218789	Volume of grey matter in Vermis IX Cerebellum	10 -0.125285	Weighted-mean L1 in tract superior longitudina...
11 0.389036	Volume of putamen (left)	11 -0.217942	Mean L3 in pontine crossing tract on FA skeleton	11 0.124800	Volume of peripheral cortical grey matter (nor...
12 0.388799	Volume of grey matter in Temporal Fusiform Cor...	12 -0.216575	Mean MD in pontine crossing tract on FA skeleton	12 -0.120350	Weighted-mean ISOVF in tract superior longitud...
13 0.386617	Volume of brain stem + 4th ventricle	13 -0.215168	Volume of grey matter in Vermis Vllb Cerebellum	13 -0.120119	Weighted-mean ISOVF in tract superior thalamic...
14 0.375325	Volume of grey matter in Temporal Fusiform Cor...	14 -0.197152	Mean L2 in pontine crossing tract on FA skeleton	14 -0.119648	Weighted-mean MD in tract superior thalamic ra...
15 0.375139	Volume of grey matter in Parahippocampal Gyrus...	15 -0.195988	Mean ISOVF in pontine crossing tract on FA skele...	15 -0.119608	Mean ISOVF in superior corona radiata on FA sk...
16 0.371975	Volume of grey matter in Insular Cortex (right)	16 -0.194592	Mean L1 in pontine crossing tract on FA skeleton	16 -0.118818	Weighted-mean L2 in tract anterior thalamic ra...
17 0.371646	Volume of grey matter in Frontal Orbital Corte...	17 0.193077	Mean ICVF in cerebral peduncle on FA skeleton ...	17 -0.117825	Weighted-mean ISOVF in tract anterior thalamic...
18 0.367857	Volume of grey matter in Lingual Gyrus (left)	18 -0.184271	Volume of grey matter in Crus II Cerebellum (l...	18 -0.117309	Weighted-mean L1 in tract superior thalamic ra...
19 0.365403	Volume of grey matter in Temporal Pole (right)	19 -0.181645	Volume of grey matter in Crus II Cerebellum (r...	19 -0.117275	Mean L3 in superior cerebellar peduncle on FA ...
20 0.361255	Volume of grey matter in Amygdala (right)	20 -0.177791	Volume of grey matter in Vermis Vllb Cerebellum	20 -0.110710	Mean ISOVF in posterior corona radiata on FA s...
21 0.360575	Volume of grey matter in Precuneous Cortex (left)	21 -0.176719	Mean L1 in inferior cerebellar peduncle on FA ...	21 -0.115830	Weighted-mean ISOVF in tract inferior fronto-o...
22 0.359981	Volume of grey matter in Precuneous Cortex (...)	22 -0.175643	Total volume of white matter hyperintensities ...	22 0.115663	Volume of grey matter in Temporal Pole (left)
23 0.359077	Volume of grey matter in Insular Cortex (left)	23 -0.169011	Mean L1 in medial lemniscus on FA skeleton (r...	23 0.114908	Mean MO in anterior corona radiata on FA skele...
24 0.358783	Volume of grey matter in Central Opercular Cor...	24 -0.168091	Mean L1 in superior cerebellar peduncle on FA ...	24 -0.113881	Mean ISOVF in corticospinal tract on FA skelet...
25 0.356923	Volume of grey matter in Central Opercular Cor...	25 -0.166054	Mean L1 in superior cerebellar peduncle on FA ...	25 0.113254	Volume of grey matter in Superior Temporal Gyr...
26 0.354662	Volume of grey matter in Temporal Fusiform Cor...	26 0.159677	Mean ICVF in medial lemniscus on FA skeleton (...)	26 -0.113132	Weighted-mean ISOVF in tract superior thalamic...
27 0.354227	Volume of grey matter in Hippocampus (left)	27 0.156989	Mean OD in superior cerebellar peduncle on FA ...	27 -0.113048	Weighted-mean L1 in tract anterior thalamic ra...
28 0.354197	Volume of grey matter in Lingual Gyrus (right)	28 0.156452	Mean OD in superior cerebellar peduncle on FA ...	28 0.111890	Median T2star in amygdala (right)
29 0.353966	Volume of grey matter in Temporal Fusiform Cor...	29 0.156126	Mean ICVF in cingulum hippocampus on FA skelet...	29 -0.111423	Mean L2 in corticospinal tract on FA skeleton ...
Loading	Name	Loading	Name	Loading	Name
0 0.202370	Mean ICVF in pontine crossing tract on FA skele...	0 0.215932	Mean ISOVF in superior cerebellar peduncle on ...	0 0.147704	Volume of brain, grey+white, normalised for he...
1 0.184019	Mean ISOVF in corticospinal tract on FA skelet...	1 0.213732	Mean MD in superior cerebellar peduncle on FA ...	1 0.142323	Volume of grey matter (normalised for head size)
2 0.176214	Volume of grey matter in Vllb Cerebellum (left)	2 0.195184	Mean MD in superior cerebellar peduncle on FA ...	2 0.137639	90th percentile of z-statistic (in group-defin...
3 0.174382	Volume of grey matter in Brain-Stem	3 -0.191986	Mean MO in external capsule on FA skeleton (left)	3 0.132592	Median z-statistic (in group-defined mask) for...
4 0.171970	Volume of grey matter in Vllb Cerebellum (right)	4 -0.191166	Volume of grey matter (normalised for head size)	4 -0.131216	Mean OD in cingulum cingulate gyrus on FA skele...
5 0.158594	Mean ICVF in corticospinal tract on FA skeleto...	5 0.184344	Mean L2 in superior cerebellar peduncle on FA ...	5 0.129422	Volume of grey matter in Frontal Orbital Corte...
6 0.158140	Mean ICVF in corticospinal tract on FA skeleto...	6 0.181798	Mean L2 in external capsule on FA skeleton (left)	6 0.128216	Volume of grey matter in Parahippocampal Gyrus...
7 0.154322	Mean L1 in corticospinal tract on FA skeleton (...)	7 -0.181678	Volume of peripheral cortical grey matter (nor...	7 0.119442	Volume of grey matter in Amygdala (right)
8 -0.149378	Weighted-mean OD in tract medial lemniscus (left)	8 -0.179987	Mean L2 in sagittal stratum on FA skeleton (left)	8 0.119400	Volume of peripheral cortical grey matter (nor...
9 0.148410	Mean ISOVF in cingulum cingulate gyrus on FA s...	9 0.177595	Mean L2 in superior cerebellar peduncle on FA ...	9 0.118295	Volume of grey matter in Vermis Vllb Cerebellum
10 0.148307	Volume of grey matter in IX Cerebellum (left)	10 0.175565	Mean ISOVF in superior cerebellar peduncle on ...	10 -0.117654	Mean OD in superior longitudinal fasciculus on...
11 0.145775	Mean ISOVF in pontine crossing tract on FA skele...	11 0.173893	Mean L3 in superior cerebellar peduncle on FA ...	11 0.117444	Volume of grey matter in Frontal Orbital Corte...
12 0.143371	Weighted-mean ISOVF in tract cingulate gyrus p...	12 -0.171890	Volume of grey matter	12 0.116733	Median z-statistic (in group-defined amygdala ...)
13 0.141536	Mean ICVF in medial lemniscus on FA skeleton (...)	13 -0.171673	Mean L3 in superior cerebellar peduncle on FA ...	13 0.114067	Median BOLD effect (in group-defined mask) for...
14 0.140043	Mean FA in anterior limb of internal capsule o...	14 0.170469	Weighted-mean L2 in tract inferior longitudina...	14 0.114020	Volume of grey matter in Precuneous Cortex (r...
15 0.139322	Mean ICVF in medial lemniscus on FA skeleton (...)	15 -0.168226	Volume of peripheral cortical grey matter	15 0.111187	Volume of grey matter in Parahippocampal Gyrus...
16 0.138344	Volume of grey matter in Vermis X Cerebellum	16 0.163947	Total volume of white matter hyperintensities ...	16 0.105524	Mean FA in cingulum hippocampus on FA skeleton...
17 0.136800	Weighted-mean ISOVF in tract forceps minor	17 0.161960	Mean L1 in superior cerebellar peduncle on FA ...	17 0.104473	90th percentile of BOLD effect (in group-defin...
18 0.136544	Mean ISOVF in corticospinal tract on FA skelet...	18 0.160804	Weighted-mean L2 in tract inferior longitudina...	18 -0.103131	Mean ISOVF in superior cerebellar peduncle on ...
19 0.136226	Volume of grey matter in IX Cerebellum (right)	19 0.160720	Mean L3 in sagittal stratum on FA skeleton (left)	19 0.102523	Volume of hippocampus (left)
20 -0.135734	Mean OD in anterior limb of internal capsule o...	20 0.159812	Weighted-mean MD in tract inferior longitudina...	20 -0.102463	Mean L2 in superior cerebellar peduncle on FA ...
21 0.131922	Weighted-mean L1 in tract cingulate gyrus part...	21 0.159656	Mean MD in sagittal stratum on FA skeleton (left)	21 0.100310	Mean MO in retrolicular part of internal ca...
22 0.130842	Mean L1 in cingulum cingulate gyrus on FA skele...	22 -0.157503	Volume of grey matter in insular Cortex (right)	22 0.100078	Volume of grey matter in Caudate (left)
23 0.129473	Mean FA in posterior limb of internal capsule ...	23 -0.157442	Volume of grey matter in Insular Cortex (left)	23 0.099728	Median BOLD effect (in group-defined amygdala ...)
24 0.129405	Mean ISOVF in medial lemniscus on FA skeleton ...	24 -0.157076	Mean MO in external capsule on FA skeleton (ri...	24 0.099651	Volume of thalamus (right)
25 0.128844	Weighted-mean L1 in tract corticospinal tract ...	25 -0.157041	Mean FA in sagittal stratum on FA skeleton (left)	25 0.099597	Volume of caudate (right)
26 0.127853	Mean MD in corticospinal tract on FA skeleton ...	26 -0.154412	Volume of brain, grey+white, normalised for he...	26 0.098952	Weighted-mean MO in tract acoustic radiation (...
27 0.127698	Mean L1 in body of corpus callosum on FA skeleton	27 0.154267	Weighted-mean MD in tract inferior longitudina...	27 0.098929	Volume of grey matter in Inferior Temporal Gyr...
28 -0.127675	Mean OD in superior fronto-occipital fasciculu...	28 0.153804	Mean L2 in external capsule on FA skeleton (ri...	28 0.098939	Volume of white matter (normalised for head size)
29 0.126329	Weighted-mean FA in tract anterior thalamic ra...	29 0.153114	Mean L1 in tapetum on FA skeleton (left)	29 -0.098347	Mean OD in retrolicular part of internal ca...

Figure B.12: Top 30 canonical loadings for the 6 significant IDP canonical variables in the CCA of IDP and SM. From top left to bottom right are the first to the sixth canonical loadings respectively.

B.2.3 Canonical loadings between non-reduced FC and IDP

Loading	Name	Loading	Name	Loading	Name
0	0.491263	Volume of grey matter in Angular Gyrus (right)	0	0.390489	Volume of grey matter in Lateral Occipital Cor...
1	0.484269	Volume of grey matter in Cingulate Gyrus, post...	1	-0.367429	Volume of grey matter in Supramarginal Gyrus, ...
2	0.479931	Volume of peripheral cortical grey matter	2	-0.355967	Volume of grey matter in Supramarginal Gyrus, ...
3	0.477623	Volume of grey matter in Cingulate Gyrus, post...	3	0.347914	Volume of grey matter in Lateral Occipital Cor...
4	0.471205	Volume of grey matter in Angular Gyrus (left)	4	0.342060	Volume of grey matter in Middle Frontal Gyrus ...
5	0.467662	Volume of brain, grey+white	5	0.341450	Volume of grey matter in Middle Frontal Gyrus ...
6	-0.457713	Volumetric scaling from T1 head image to stand...	6	0.324877	Volume of grey matter in Planum Polare (right)
7	0.457514	Volume of grey matter in Precuneous Cortex (left)	7	0.318651	Volume of grey matter in Cingulate Gyrus, ante...
8	0.455760	Volume of grey matter	8	0.317122	Volume of grey matter in Cingulate Gyrus, ante...
9	0.451250	Volume of grey matter in Precuneous Cortex (ri...	9	0.308473	Volume of grey matter in Insular Cortex (left)
10	0.432530	Volume of grey matter in Cingulate Gyrus, ante...	10	0.297152	Volume of grey matter in Planum Polare (left)
11	0.432308	Volume of white matter	11	0.296052	Volume of grey matter in Insular Cortex (right)
12	0.428937	Volume of grey matter in Middle Temporal Gyrus...	12	0.293107	Volume of brain, grey+white
13	0.410220	Volume of grey matter in Cingulate Gyrus, ante...	13	0.292848	Volume of grey matter in Superior Frontal Gyrus...
14	0.400721	Volume of grey matter in Supramarginal Gyrus, ...	14	0.285902	Volume of grey matter
15	0.398367	Volume of grey matter in Supramarginal Gyrus, ...	15	0.282900	Volume of grey matter in Superior Temporal Gyrus...
16	0.396149	Volume of grey matter in Subcallosal Cortex (f...	16	0.282642	Volume of peripheral cortical grey matter
17	0.392839	Volume of grey matter in Planum Polare (left)	17	0.272189	Volume of grey matter in Frontal Pole (right)
18	0.387274	Volume of grey matter in Subcallosal Cortex (f...	18	0.271572	Volume of grey matter in Superior Frontal Gyrus...
19	0.379765	Volume of grey matter in Middle Frontal Gyrus ...	19	-0.271480	Volumetric scaling from T1 head image to stand...
20	0.376977	Volume of grey matter in Supracalcarine Cortex...	20	0.269599	Volume of white matter
21	0.373981	Volume of grey matter in Central Opercular Cor...	21	0.258610	Volume of grey matter in Frontal Pole (left)
22	0.369426	Volume of grey matter in Middle Temporal Gyrus...	22	0.254457	Volume of grey matter in Subcallosal Cortex (f...
23	0.369380	Volume of grey matter in Postcentral Gyrus (ri...	23	0.250259	Volume of grey matter in Paracingulate Gyrus (...)
24	0.367900	Volume of grey matter in Supracalcarine Cortex...	24	0.249236	Volume of grey matter in Lateral Occipital Cor...
25	0.362907	Volume of grey matter in Postcentral Gyrus (left)	25	0.245473	Volume of grey matter in Subcallosal Cortex (f...
26	0.360824	Volume of grey matter in Supramarginal Gyrus, ...	26	0.244039	Volume of grey matter in Superior Temporal Gyrus...
27	0.359941	Volume of grey matter in Middle Temporal Gyrus...	27	0.237119	Volume of grey matter in Supracalcarine Cortex...
28	0.359661	Volume of grey matter in Middle Temporal Gyrus...	28	0.229011	Volume of grey matter in Paracingulate Gyrus (...)
29	0.358048	Volume of grey matter in Planum Polare (right)	29	0.214467	Volume of grey matter in Cuneal Cortex (right)
Loading	Name	Loading	Name	Loading	Name
0	0.368234	Volume of grey matter in Lateral Occipital Cor...	0	0.349690	Volume of grey matter in Lateral Occipital Cor...
1	0.330745	Volume of grey matter in Intracalcarine Cortex...	1	-0.260704	Volume of grey matter in Supramarginal Gyrus, ...
2	0.322662	Volume of grey matter in Intracalcarine Cortex...	2	-0.255242	Volume of grey matter in Angular Gyrus (right)
3	0.312750	Volume of grey matter in Lateral Occipital Cor...	3	0.241122	Volume of grey matter in Supramarginal Gyrus, ...
4	0.267472	Volume of grey matter in Occipital Pole (left)	4	-0.236774	Volume of grey matter in Intracalcarine Cortex...
5	0.250673	Mean MO in body of corpus callosum on FA skeleton	5	-0.235117	Volume of grey matter in Lateral Occipital Cor...
6	0.248481	Volume of grey matter in Occipital Pole (right)	6	-0.210907	Volume of grey matter in Intracalcarine Cortex...
7	-0.231300	Volume of grey matter in Superior Frontal Gyrus...	7	0.210723	Volume of grey matter in Parietal Operculum Co...
8	0.229588	Mean FA in superior fronto-occipital fasciculu...	8	0.189972	Volume of grey matter in Precuneous Cortex (left)
9	-0.228477	Weighted-mean MD in tract forceps major	9	0.183562	Volume of grey matter in Postcentral Gyrus (left)
10	0.224197	Volume of grey matter in Ventral Striatum (left)	10	-0.183510	Volume of grey matter in Occipital Pole (left)
11	-0.220823	Weighted-mean L3 in tract forceps major	11	0.180554	Volume of grey matter in Parietal Operculum Co...
12	0.215075	Volume of grey matter in Paracingulate Gyrus (...)	12	0.179269	Weighted-mean MD in tract forceps major
13	-0.213153	Mean OD in superior fronto-occipital fasciculu...	13	0.174646	Weighted-mean L3 in tract forceps major
14	-0.209273	Weighted-mean L1 in tract forceps major	14	0.174350	Volume of grey matter in Supramarginal Gyrus, ...
15	0.207156	Volume of grey matter in Ventral Striatum (right)	15	0.173632	Weighted-mean L2 in tract forceps major
16	0.205538	Mean FA in superior fronto-occipital fasciculu...	16	0.172385	Volume of grey matter in Planum Temporale (right)
17	0.202583	Mean FA in fornix on FA skeleton	17	0.162818	Weighted-mean ISOVF in tract forceps major
18	0.200101	Weighted-mean MO in tract cingulate gyrus plus...	18	0.147854	Volume of grey matter in Superior Parietal Lobu...
19	-0.199474	Mean MD in superior cerebellar peduncle on FA...	19	0.146875	Volume of grey matter in Temporal Operculum Fo...
20	-0.199324	Mean OD in body of corpus callosum on FA skeleton	20	0.141735	Weighted-mean L1 in tract forceps major
21	-0.199172	Mean OD in superior fronto-occipital fasciculu...	21	0.139620	Volume of grey matter in Precuneous Cortex (ri...
22	0.199028	Volume of grey matter in Frontal Pole (left)	22	-0.139020	Weighted-mean FA in tract forceps major
23	-0.189876	Mean ISOVF in superior cerebellar peduncle on ...	23	0.129514	Volume of grey matter in Heschl's Gyrus (inclus...
24	0.197978	Volume of grey matter in Middle Frontal Gyrus ...	24	0.127619	Volume of grey matter in Planum Temporale (left)
25	-0.197967	Mean ISOVF in fornix on FA skeleton	25	-0.127029	Weighted-mean L2 in tract superior longitudina...
26	0.1966529	Mean MO in external capsule on FA skeleton (ri...	26	0.126015	Mean MO in superior longitudinal fasciculus on...
27	0.195175	Volume of grey matter in Middle Frontal Gyrus ...	27	0.123961	Volume of grey matter in Inferior Frontal Gyrus...
28	-0.193330	Weighted-mean L2 in tract forceps major	28	-0.119432	Weighted-mean OD in tract superior longitudina...
29	-0.193232	Volume of grey matter in Superior Frontal Gyrus...	29	-0.118579	Mean OD in superior longitudinal fasciculus on...
Loading	Name	Loading	Name	Loading	Name
0	0.452710	Volume of grey matter in Superior Frontal Gyrus...	1	0.419583	Volume of grey matter in Superior Frontal Gyrus...
2	0.294985	Volume of grey matter	3	0.293936	Volume of grey matter in Insular Cortex (right)
4	0.279960	Volume of grey matter in Central Opercular Cor...	5	0.277583	Volume of grey matter in Central Opercular Cor...
6	0.272507	Volume of grey matter in Villa Cerebellum (left)	7	0.271903	Volume of peripheral cortical grey matter
8	0.269105	Volume of grey matter in Insular Cortex (left)	9	0.262193	Volume of brain, grey+white
10	0.260778	Volume of grey matter in Postcentral Gyrus (left)	11	0.258819	Volume of grey matter in Villa Cerebellum (right)
12	0.257116	Volume of grey matter in Juxtapositional Lobul...	13	-0.256369	Volumetric scaling from T1 head image to stand...
14	0.252851	Volume of grey matter in Juxtapositional Lobul...	15	0.248120	Volume of grey matter in Temporal Fusiform Cor...
16	0.247591	Volume of thalamus (right)	17	-0.246332	Mean OD in body of corpus callosum on FA skeleton
18	0.244297	Volume of grey matter in IX Cerebellum (right)	19	0.243392	Volume of grey matter in Vermis Vilb Cerebellum
20	0.242571	Volume of grey matter in Vilb Cerebellum (left)	21	0.241072	Volume of grey matter in Lingual Gyrus (right)
22	0.240037	Volume of thalamus (left)	23	0.238965	Volume of grey matter in Vilb Cerebellum (right)
24	0.236920	Volume of grey matter in Vermis VI Cerebellum	25	0.232883	Volume of grey matter in VI Cerebellum (left)
26	0.228737	Mean MO in body of corpus callosum on FA skeleton	27	0.228484	Volume of grey matter in Frontal Pole (left)
28	0.228285	Volume of grey matter in IX Cerebellum (left)	29	0.228251	Volume of grey matter in Amygdala (right)

Figure B.13: Top 30 canonical loadings for the first 6 significant IDP canonical variables in the CCA of IDP and FC. From top left to bottom right are the first to the sixth canonical loadings respectively.

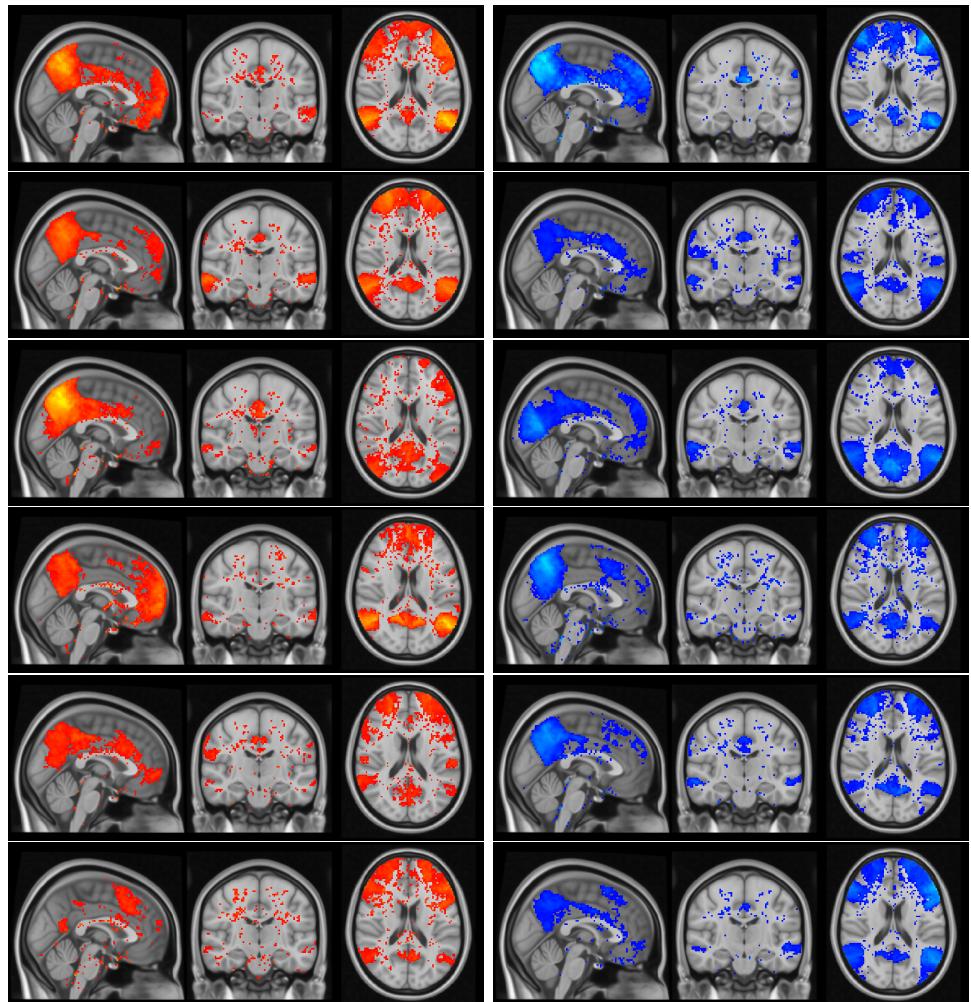


Figure B.14: Canonical loaded maps for the first 6 significant FC canonical variables in the CCA with IDP. From top to bottom are the first to the sixth canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).

B.3 CCA Results for SDR-reduced Data

B.3.1 Canonical loadings between SDR SM and SDR FC

Loading	Name	Loading	Name	Loading	Name	Loading	Name
0 0.449165	Exercise&Work 6	0 -0.403427	Cognition 1	0 0.648665	Physical measure 6	0 0.588900	Physical measure 1
1 -0.350545	Physical measure 5	1 -0.390396	Exercise&Work 2	1 -0.274279	LifeStyle&Environment 3	1 0.306917	Cognition 1
2 0.346866	Physical measure 3	2 0.355644	Physical measure 1	2 -0.251072	LifeStyle&Environment 4	2 0.244613	MentalHealth 4
3 -0.300155	LifeStyle&Environment 2	3 0.343014	Cognition 4	3 0.226149	Cognition 4	3 0.226846	Physical measure 2
4 -0.296628	Exercise&Work 3	4 -0.292404	Physical measure 2	4 -0.219989	Exercise&Work 2	4 -0.224139	Exercise&Work 8
5 0.217453	Food&Drink 5	5 -0.290523	Food&Drink 4	5 -0.198834	Exercise&Work 1	5 0.202837	Physical measure 14
6 -0.212531	Exercise&Work 19	6 0.287988	Alcohol 1	6 0.180154	Cognition 5	6 0.191767	LifeStyle&Environment 2
7 0.177649	Physical measure 1	7 -0.273570	Physical measure 6	7 0.175956	Food&Drink 1	7 0.180089	Exercise&Work 3
8 0.175362	LifeStyle&Environment 3	8 -0.261749	LifeStyle&Environment 1	8 0.167416	Exercise&Work 11	8 0.178479	LifeStyle&Environment 12
9 0.174378	MentalHealth 2	9 0.256982	LifeStyle&Environment 5	9 0.161437	MentalHealth 4	9 -0.176889	MentalHealth 2
10 -0.167564	Cognition 5	10 -0.235481	LifeStyle&Environment 3	10 -0.139063	Food&Drink 4	10 -0.170895	Exercise&Work 7
11 -0.156440	Alcohol 2	11 -0.224812	Physical measure 3	11 0.135887	MentalHealth 5	11 0.156145	Food&Drink 11
12 -0.154549	Food&Drink 7	12 -0.218299	LifeStyle&Environment 9	12 -0.135136	Food&Drink 3	12 0.152653	Health&MedicalHist 11
13 -0.151986	Exercise&Work 10	13 -0.202280	LifeStyle&Environment 2	13 -0.134819	Health&MedicalHist 7	13 -0.152574	Cognition 5
14 0.147594	LifeStyle&Environment 5	14 0.195669	Food&Drink 1	14 -0.133800	Exercise&Work 9	14 -0.148324	Alcohol 3
15 0.146861	Exercise&Work 8	15 0.192821	Exercise&Work 5	15 0.130591	Exercise&Work 8	15 0.143356	Physical measure 17
16 -0.146765	Physical measure 2	16 -0.191041	Physical measure 11	16 0.124412	Physical measure 9	16 0.132552	LifeStyle&Environment 9
17 -0.145561	MentalHealth 4	17 -0.190060	Exercise&Work 3	17 0.117649	LifeStyle&Environment 7	17 0.130095	Physical measure 8
18 0.139306	Food&Drink 12	18 -0.183205	Exercise&Work 7	18 0.116155	MentalHealth 3	18 -0.125440	Exercise&Work 2
19 0.136622	Exercise&Work 7	19 0.173966	Physical measure 17	19 -0.113972	Physical measure 11	19 0.124542	Exercise&Work 1
20 -0.125979	MentalHealth 6	20 -0.153920	Food&Drink 10	20 0.113237	Health&MedicalHist 2		
Loading	Name	Loading	Name	Loading	Name	Loading	Name
0 -0.280200	Health&MedicalHist 2	0 -0.360058	Alcohol 7	0 0.250038	Physical measure 6		
1 0.273074	Food&Drink 9	1 -0.305462	Health&MedicalHist 1	1 -0.222310	Alcohol 2		
2 0.260745	Cognition 5	2 -0.300489	LifeStyle&Environment 1	2 0.208340	Exercise&Work 10		
3 -0.241394	Exercise&Work 4	3 -0.281168	Food&Drink 4	3 0.201678	Exercise&Work 12		
4 0.219321	Exercise&Work 19	4 -0.275644	Physical measure 2	4 -0.191779	LifeStyle&Environment 16		
5 0.217054	Alcohol 4	5 -0.261123	MentalHealth 1	5 -0.188507	Exercise&Work 6		
6 0.212023	Exercise&Work 1	6 -0.235565	Exercise&Work 10	6 0.188206	Physical measure 1		
7 -0.210924	Exercise&Work 12	7 0.203047	Exercise&Work 1	7 -0.187763	Food&Drink 8		
8 0.192931	Health&MedicalHist 7	8 -0.196012	Tobacco 1	8 -0.185652	Exercise&Work 13		
9 -0.177307	LifeStyle&Environment 18	9 -0.190174	Cognition 5	9 -0.176629	Cognition 5		
10 -0.176331	LifeStyle&Environment 9	10 -0.187125	Health&MedicalHist 2	10 0.174812	Physical measure 7		
11 0.175825	MentalHealth 9	11 -0.182559	Exercise&Work 2	11 -0.169791	Exercise&Work 18		
12 -0.169124	Exercise&Work 3	12 -0.180558	Alcohol 3	12 0.165762	Exercise&Work 17		
13 0.168798	LifeStyle&Environment 3	13 -0.179622	Alcohol 9	13 0.151810	Exercise&Work 9		
14 -0.165858	Food&Drink 12	14 -0.178600	MentalHealth 2	14 -0.150438	Alcohol 9		
15 -0.164397	Exercise&Work 8	15 0.171280	LifeStyle&Environment 8	15 -0.140326	Cognition 1		
16 -0.163743	MentalHealth 2	16 -0.170044	Exercise&Work 16	16 -0.140086	Physical measure 11		
17 0.163332	Food&Drink 8	17 -0.166226	Food&Drink 9	17 -0.138216	MentalHealth 3		
18 -0.151782	Physical measure 8	18 0.155098	MentalHealth 9	18 -0.137944	LifeStyle&Environment 17		
19 -0.149799	LifeStyle&Environment 6	19 -0.149412	MentalHealth 5	19 0.133411	LifeStyle&Environment 15		

Figure B.15: Top 20 canonical loadings for the 7 significant SM canonical variables in the CCA of SDR FC and SDR SM. From top left to bottom right are the first to the seventh canonical loadings respectively. The number after the domain name means the nth latent component in the sub-domain.

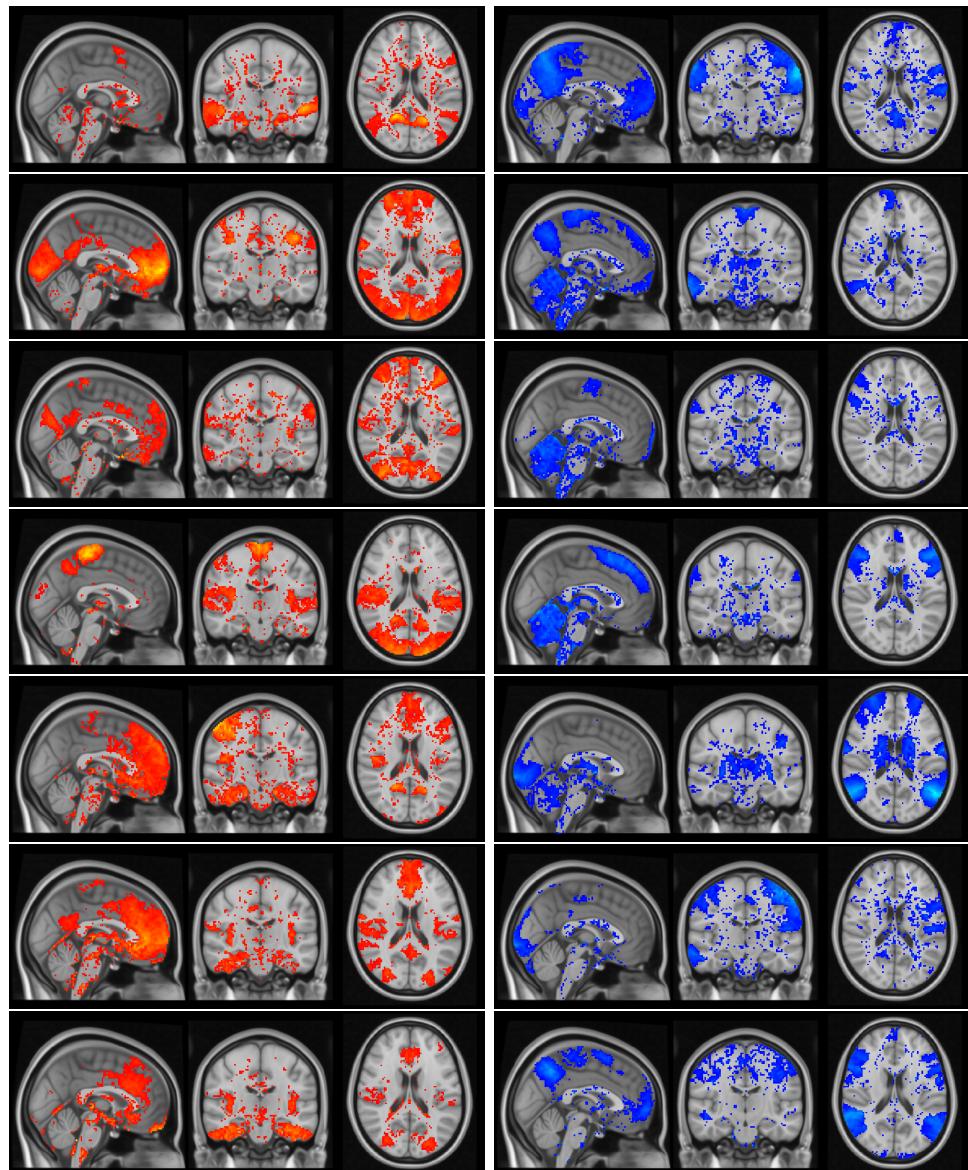


Figure B.16: Canonical loaded maps for the 7 significant SDR FC canonical variables in the CCA with SDR SM. From top to bottom are the first to the seventh canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).

B.3.2 Canonical loadings between SDR SM and SDR IDP

From Fig. B.17 and B.18 we can see clearly the contrast between each pair of SM and IDP loadings. Compared with the SM loadings, IDP loadings look more evenly distributed over its sub-domains for the 8 sets of significant canonical variables except the first set being largely focused on the cerebellum volume and the cerebral volume. Each set of the SM loadings seem to have a/some dominating domain(s) apart from the last set. For example, the first set if dominated by physical measure and cognition, and the second set is only dominated by physical measures. Moreover, the fourth set shows the contrast between cognition against tobacco.

Combining results from both sides, the first mode presents positive correlation between Physical measures, Cognition and brain volumes and T2 & Bold measures. The second mode is mainly physical measures and T2 & Bold measures. The third mode is more complicated since the contributions are more spread. However, overall many diffusion measures like MD and ICVF and even cerebral volumes are positively related to physical measures. The fourth mode is an interesting one. It shows the relationship that not taking tobacco and alcohol is positively related to larger brain volumes.

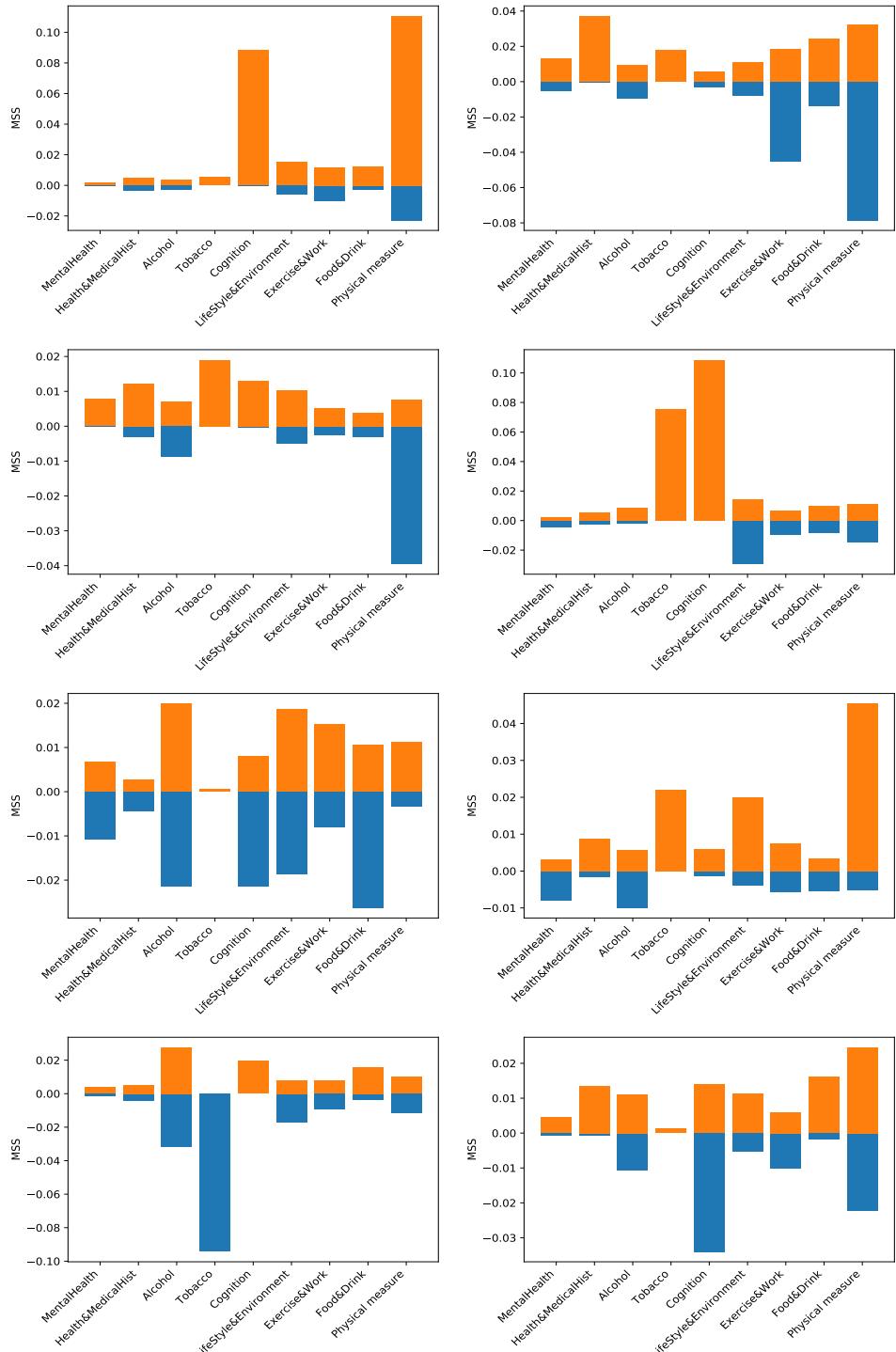


Figure B.17: Mean sum of squared SM loadings summarised from each sub-domain in the CCA of SDR IDP and SDR SM. Blue bars are the mean sum of squared positive loadings and orange bar are the mean sum of squared negative loadings. From top left to bottom right are the first to the eighth set respectively.

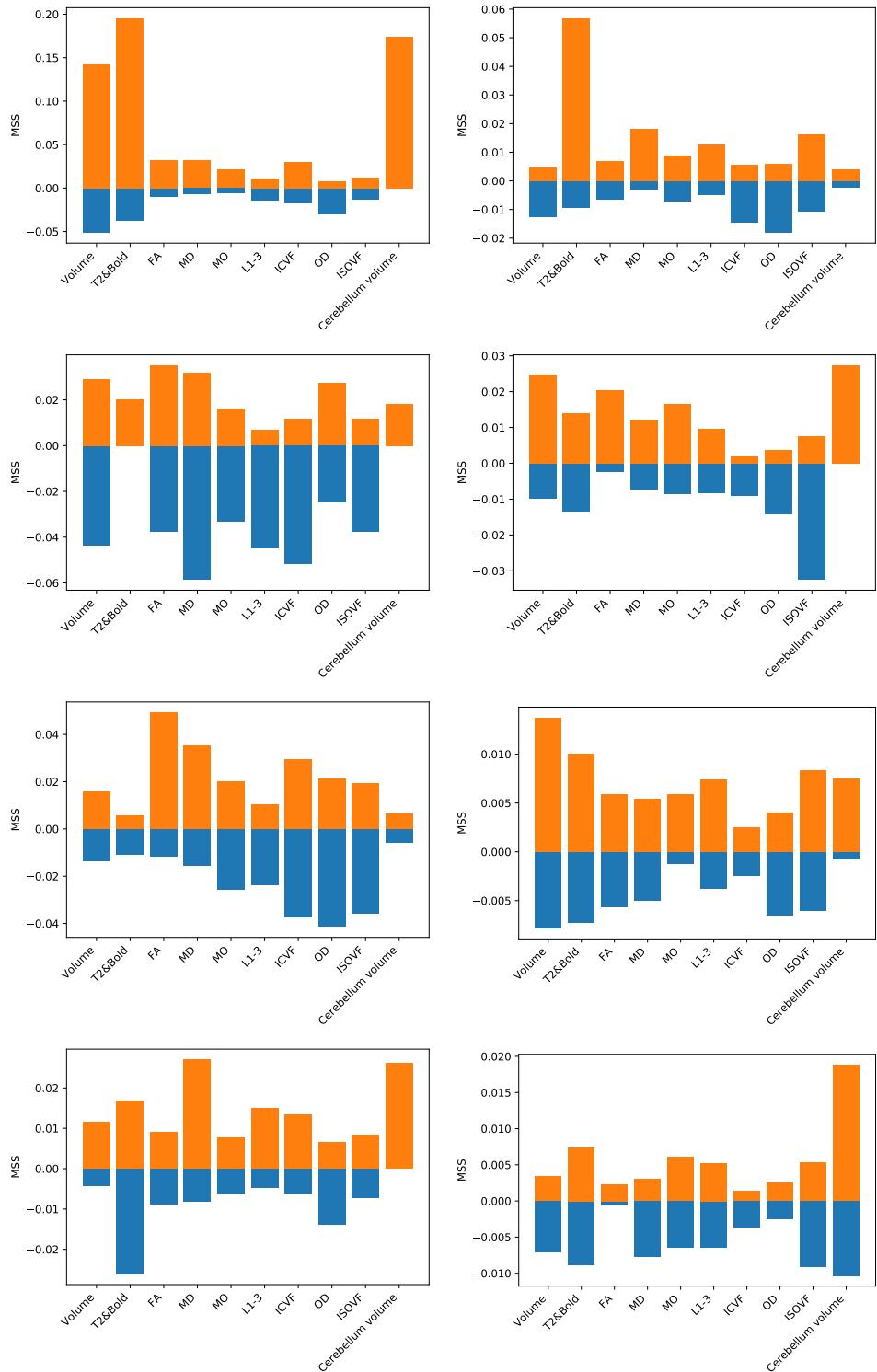


Figure B.18: Mean sum of squared IDP loadings summarised from each sub-domain in the CCA of SDR IDP and SDR SM. Blue bars are the mean sum of squared positive loadings and orange bar are the mean sum of squared negative loadings. From top left to bottom right are the first to the eighth set respectively.

B.3.3 Canonical loadings between SDR FC and SDR IDP

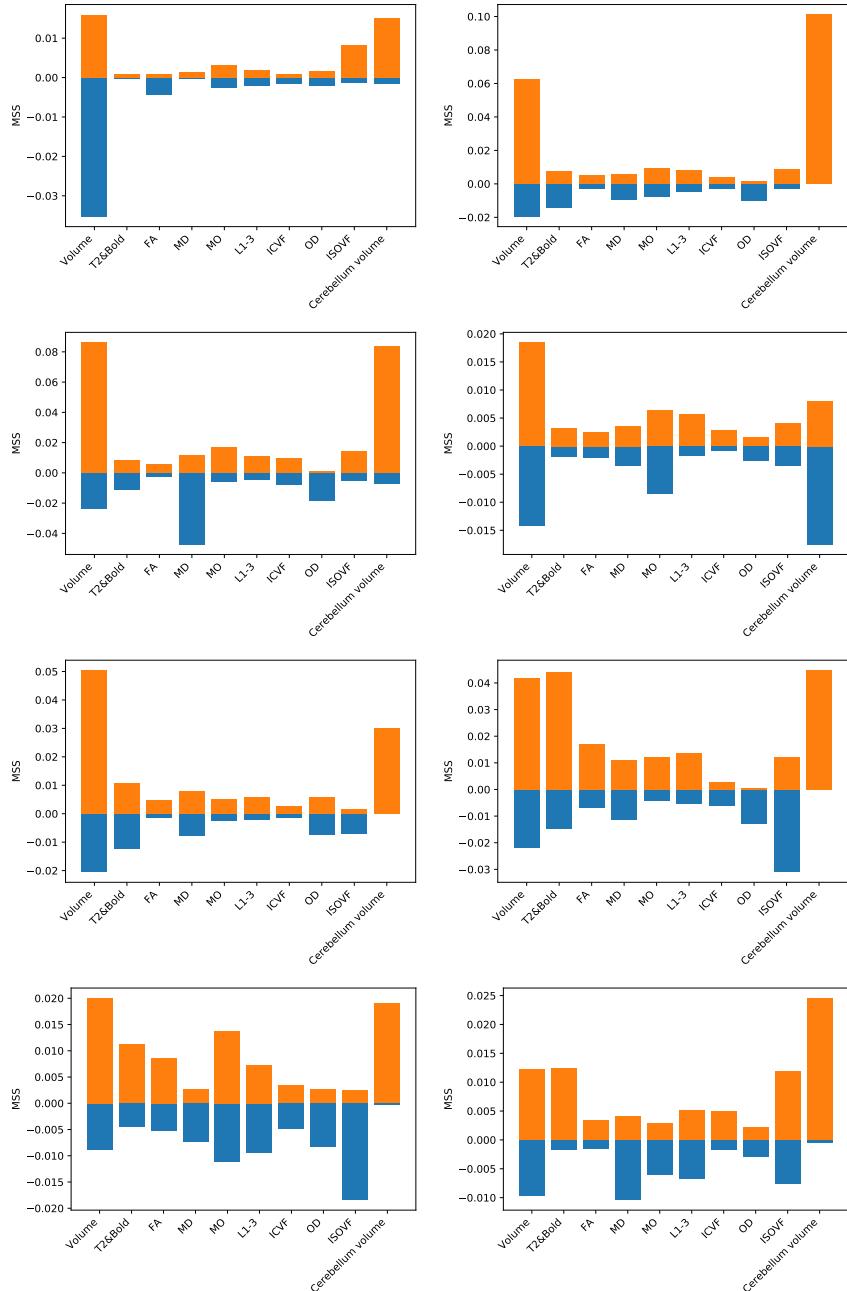


Figure B.19: Mean squared value of IDP loadings summarised from each sub-domain in the CCA of SDR IDP and SDR FC. Blue bars are the mean sum of squared positive loadings and orange bar are the mean sum of squared negative loadings. From top left to bottom right are the first to the eighth set respectively.

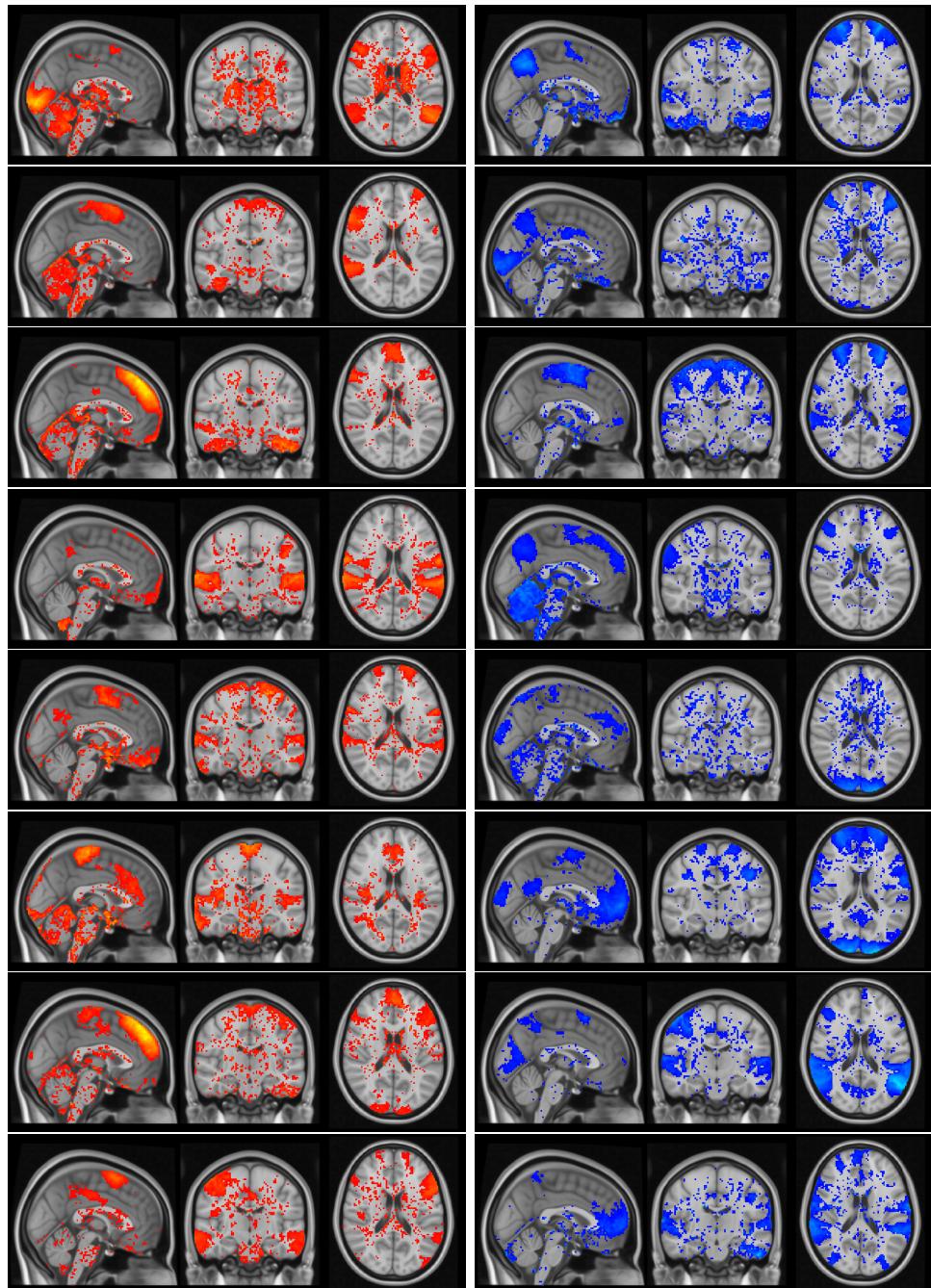


Figure B.20: Canonical loaded maps for the first 8 significant SDR FC canonical variables in the CCA with SDR IDP. From top to bottom are the first to the eighth canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).

B.3.4 Canonical loadings for multi-view SDR CCA

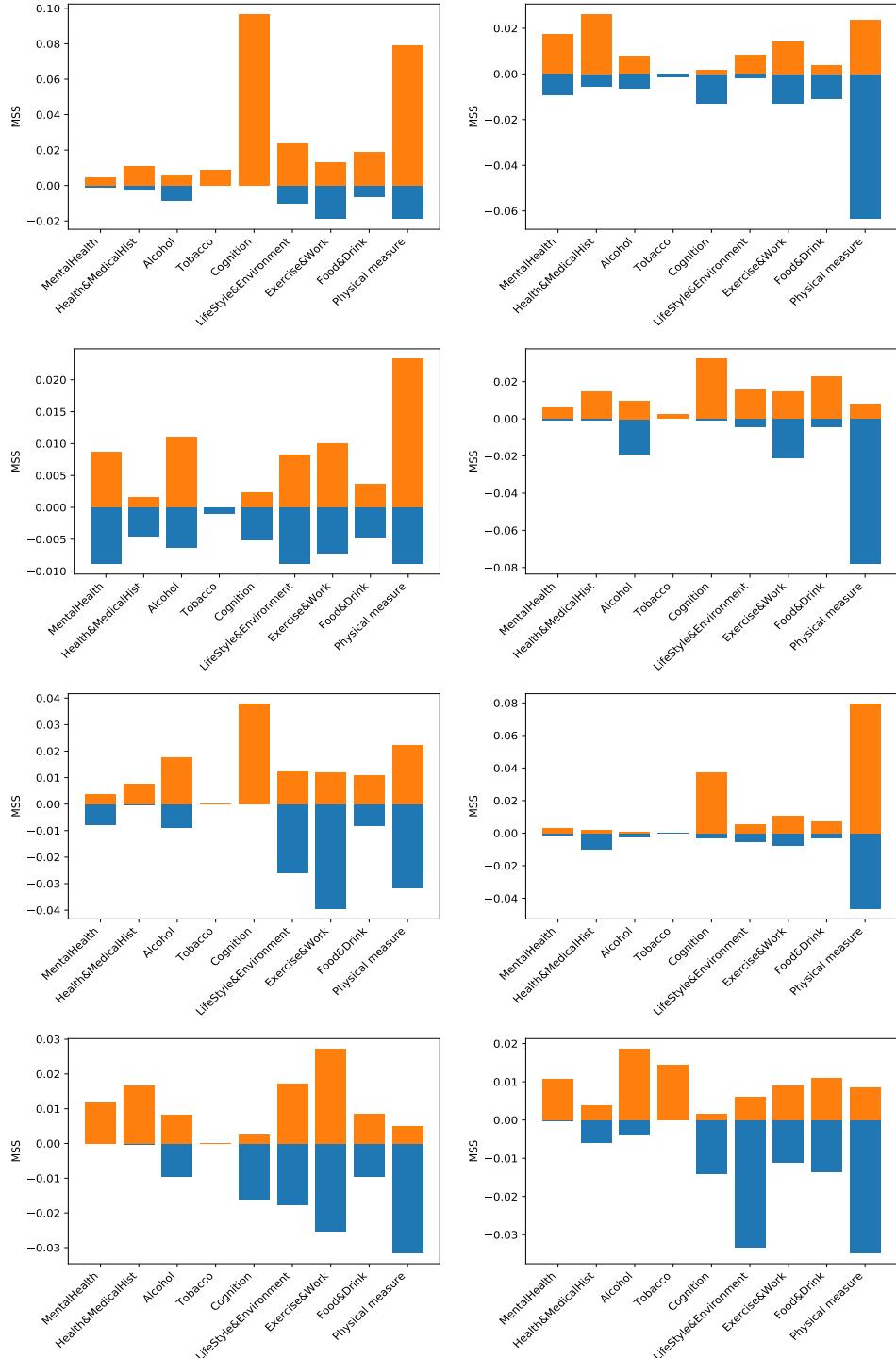


Figure B.21: Mean sum of squared SM loadings summarised from each sub-domain in the SDR multi-view CCA. Blue bars are the mean sum of squared positive loadings and orange bar are the mean sum of squared negative loadings. From top left to bottom right are the first to the eighth set respectively.

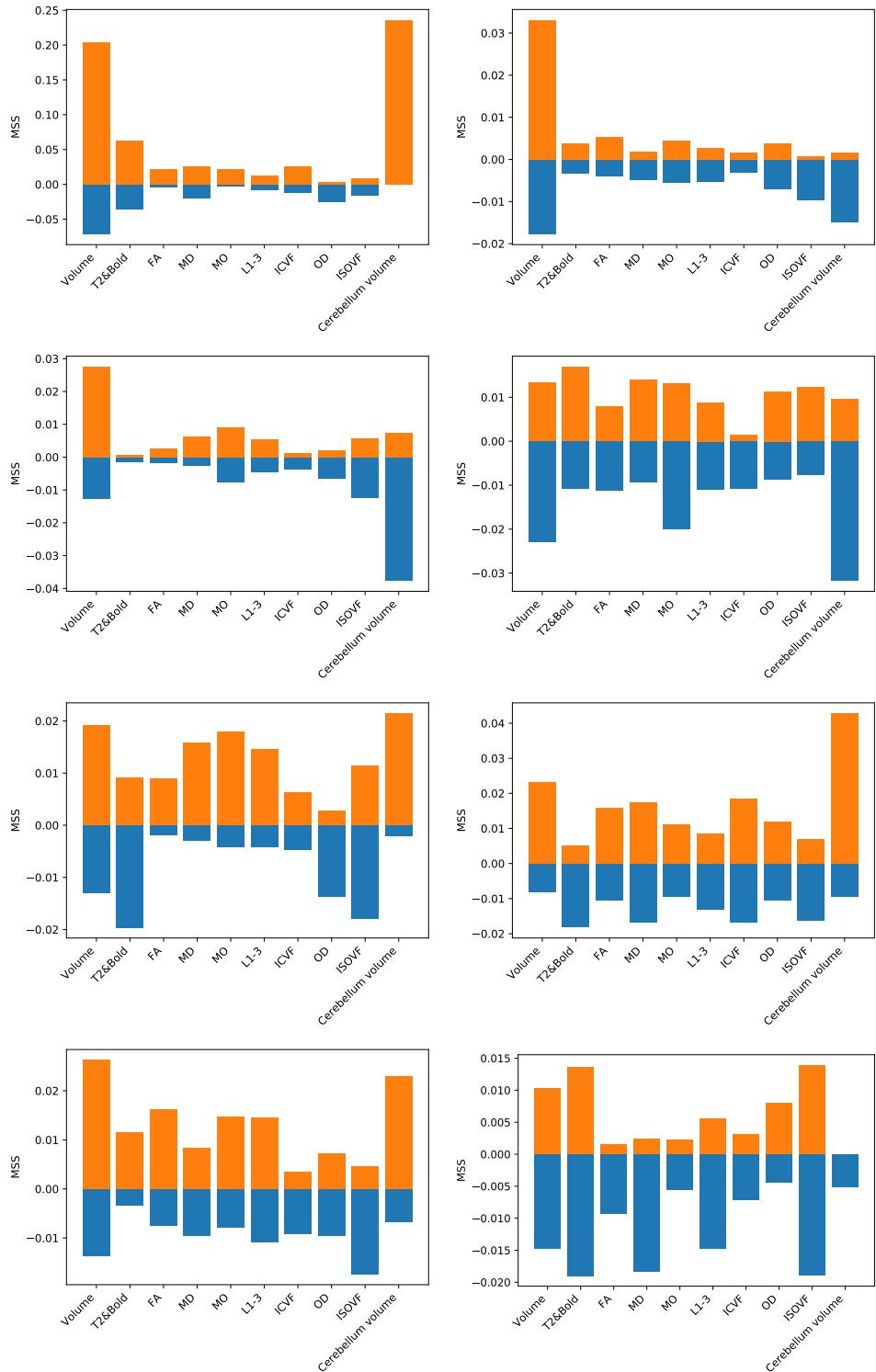


Figure B.22: Mean sum of squared IDP loadings summarised from each sub-domain in the SDR multi-view CCA. Blue bars are the mean sum of squared positive loadings and orange bar are the mean sum of squared negative loadings. From top left to bottom right are the first to the eighth set respectively.

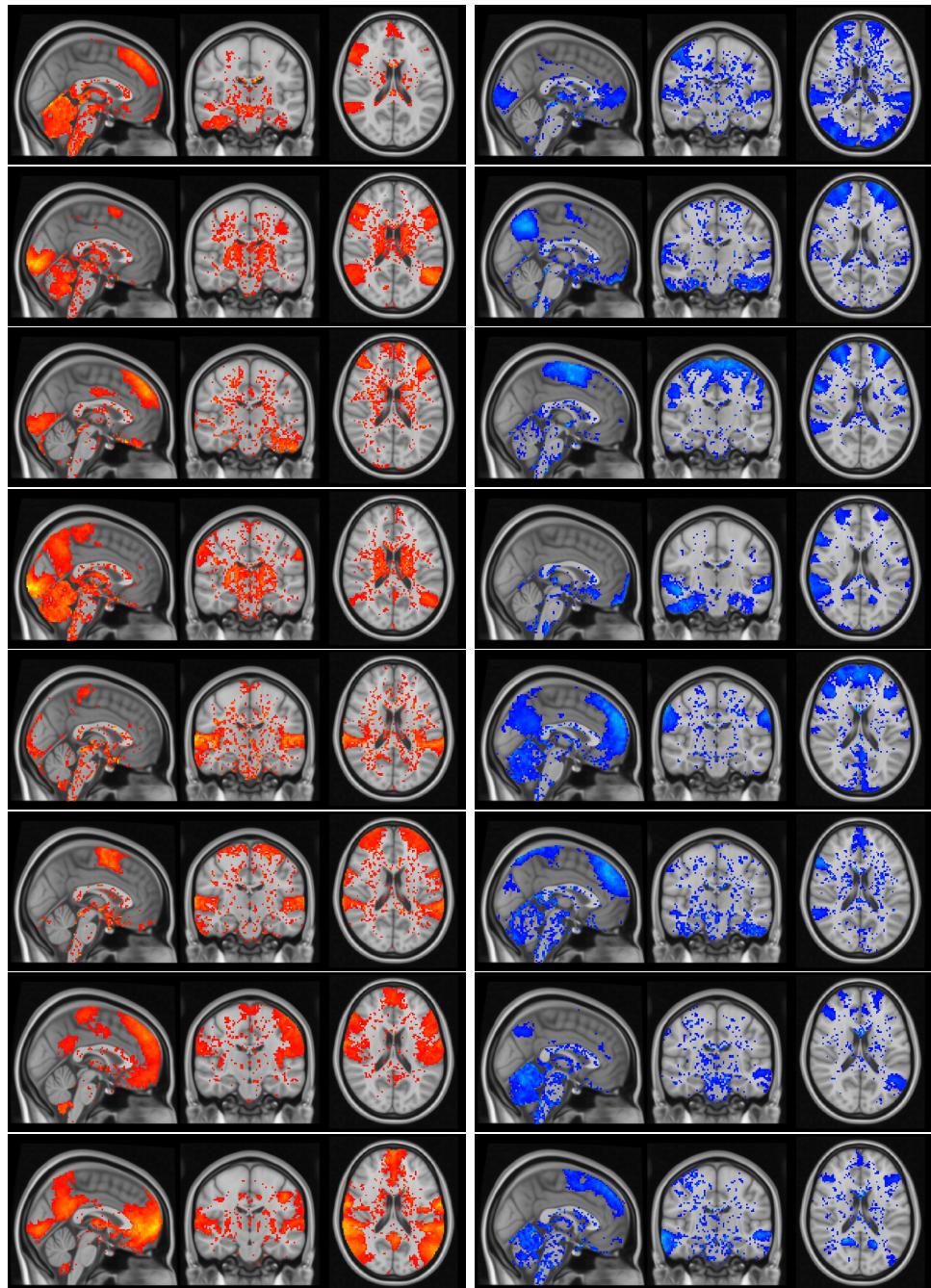


Figure B.23: Canonical loaded maps for the first 8 significant SDR FC canonical variables in the multi-view CCA. From top to bottom are the first to the eighth canonical maps. Red maps are the positive maps, and blue maps are the negative maps (the generation of the brain maps is introduced in Section 5.4.2.1).

B.4 Stability of SDR CCA

B.4.1 CCA between SDR IDP and SDR SM

Mode 1 in Fig. B.24 looks densely loaded on Cognition and Physical measures. Interestingly, it is totally absent in Mental Health, and has only one light grey bar in Alcohol. The second mode in Fig. B.24 is a mode of Health & Medical History, Physical measures and Exercise & Work. It is almost totally absent in Cognition. The third mode appear to be mostly concentrated on Physical measures, with no dark grey bars in Food & Drink and Tobacco, and the other domains are sparsely loaded. The fourth mode show strong pattern on Cognition and Tobacco with high means and small standard deviations of the canonical loadings. The other sub-domains look fairly sparse compared with the first three modes. The rest of the modes have much less dark grey bars with considerably high standard deviations of the canonical loadings. One notable observation is that the seventh mode is the only other mode highly stable in Tobacco sub-domain, and also has a noticeable proportion of highly stable factors in Cognition and Alcohol sub-domains.

On the IDP side, in general, it shows much lower stability compared with the SM side. From Fig. B.25 we can see that from the fourth to the last significant mode, there are very few dark grey bars and the standard deviations of the canonical loading are very high, many of the ones in the last three modes crosses 0. However for the first three modes, there are relatively clear patterns displayed for each of them. Especially the first mode is a strong Cerebral and Cerebellum Volume mode. The second mode has sizeable loadings in T2 & Bold, L1-L3 and ISOVF. The third mode is more focused on L1-L3 and MO sub-domains.

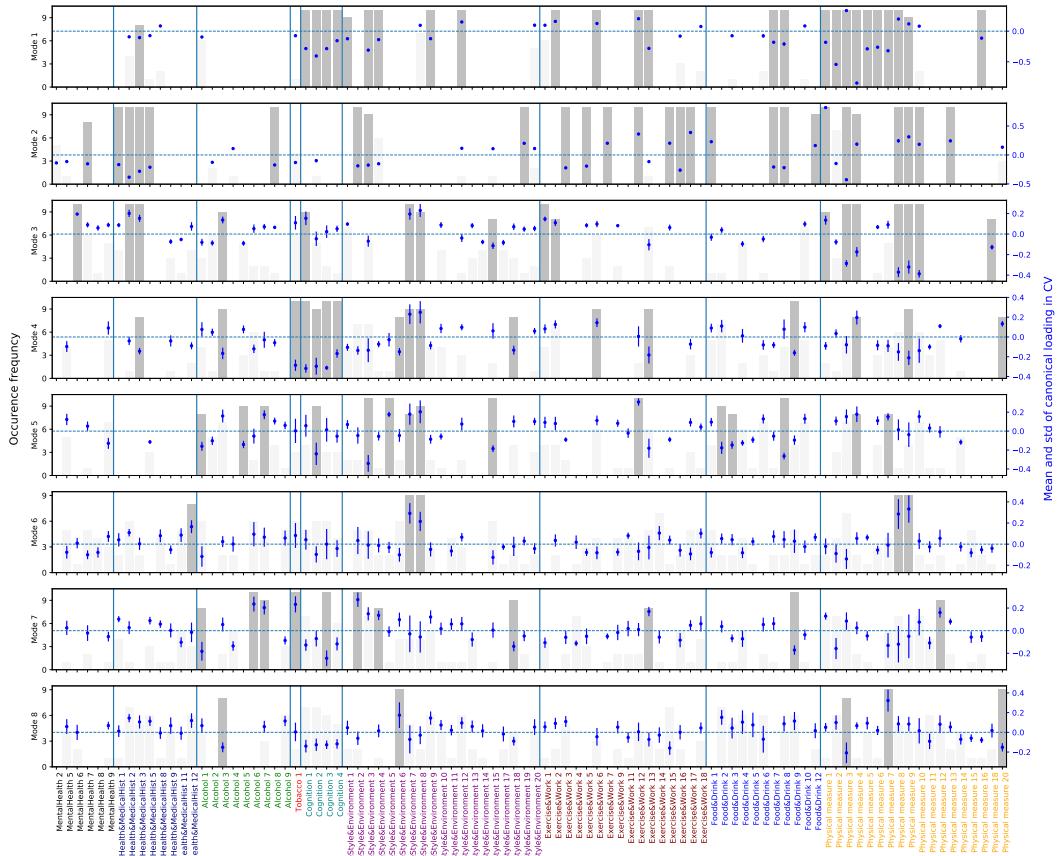


Figure B.24: Stability of SDR SM canonical loadings in the CCA of SDR SM and SDR IDP for the eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than 3 times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than 7 out of 10 times are shadowed with dark grey, the rest is shadowed by light grey.

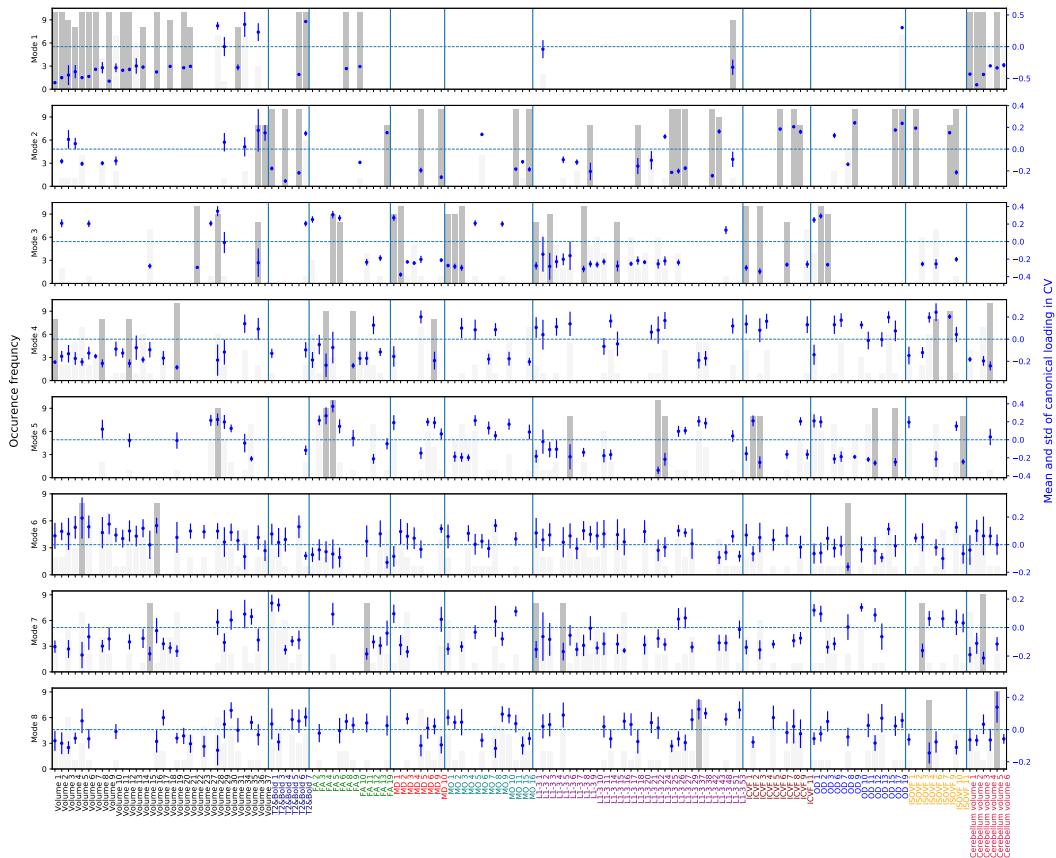


Figure B.25: Stability of SDR IDP canonical loadings in the CCA of SDR SM and SDR IDP for the eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than 3 times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than 7 out of 10 times are shadowed with dark grey, the rest is shadowed by light grey.

B.4.2 CCA between SDR FC and SDR IDP

Due to the high significance number for the CCA between SDR FC and SDR IDP, we only show results for the first 8 mode. Fig. B.26 and B.27 display stronger stability compared with Fig. 5.31 and 5.32 and Fig. B.24 and B.25. This is mainly reflected by having smaller standard deviations on the canonical loadings, and no significant dark grey bar decrease as mode increases. All modes in Fig. B.26 (IDP side) show roughly the same pattern, mostly Cerebral and Cerebellum volume focused. An interesting observation here is that mode 2 and 3 have significantly larger standard deviations on the canonical loadings than other modes. Moreover, apart from the last mode, all the other modes have approximately similar amount of dark grey bars. The fifth mode has the most concentration on Cerebral volume sub-domain.

Similar on the FC side (Fig. B.27), mode 2 and 3 present much weaker stability compared with other modes. Again, it is hard to interpret FC here without mapping them on to the brain. However the point of this set of result is to assure that the results shown in Fig. B.20 are stable.

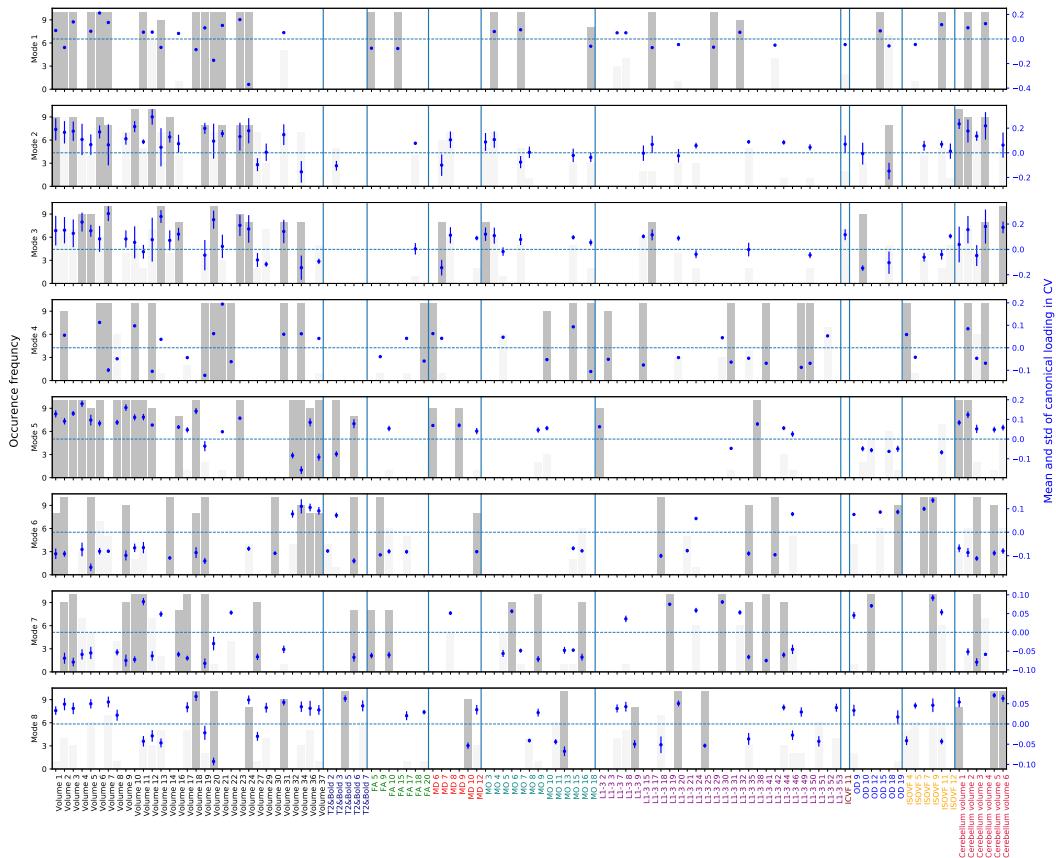


Figure B.26: Stability of SDR IDP canonical loadings in the CCA of SDR FC and SDR IDP for the eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than 3 times out of the 10 folds). Sub-domains are differentiated by different tick label colours and augmented by blue vertical lines. The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than 7 out of 10 times are shadowed with dark grey, the rest is shadowed by light grey.

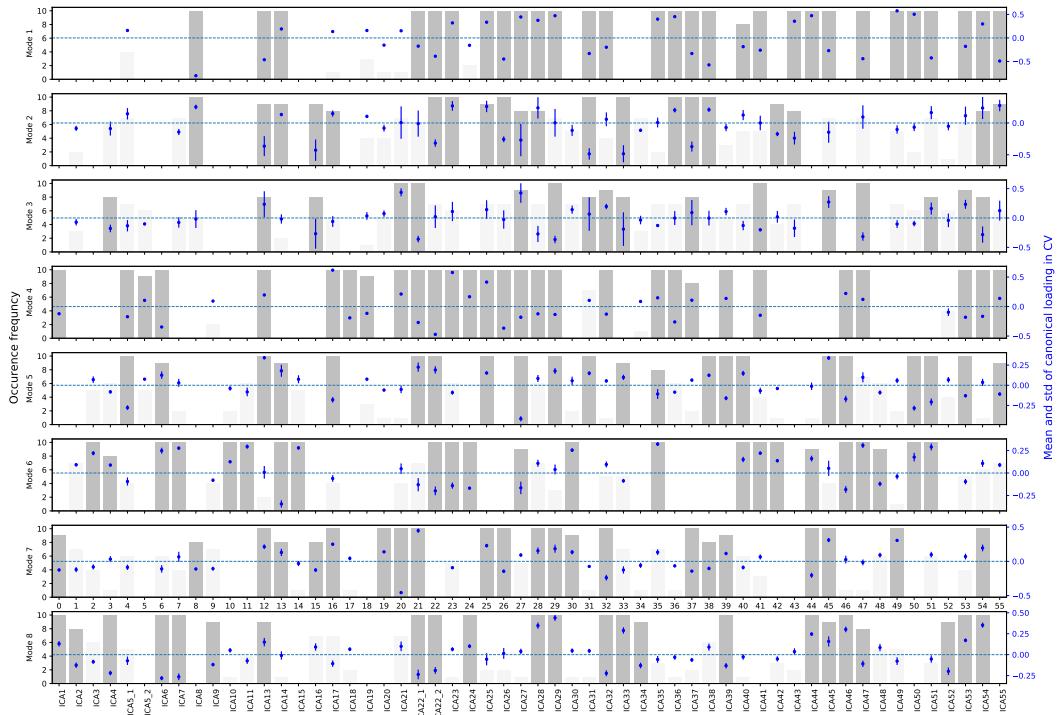


Figure B.27: Stability of SDR FC canonical loadings in the CCA of SDR FC and SDR IDP for the eight significant canonical modes. SDR factors shown here are the union of the top factors across all modes (top factors in each mode are defined by appearing in the top 30 loadings for more than 3 times out of the 10 folds). The left axis show the occurrence frequency out of 10 folds; the right axis shows the mean (blue dots) and std (blue bars) of the canonical loadings. Factors appeared more than 7 out of 10 times are shadowed with dark grey, the rest is shadowed by light grey.

B.4.3 Pairwise and multi-view CCA on PCA-reduced data

To be able to compare the performance of SDR, we applied PCA to reduce the data to the same dimensionalities as the SDR datasets, i.e. reduce FC to 57, SM to 107 and IDP to 205.

B.4.3.1 Pairwise CCA

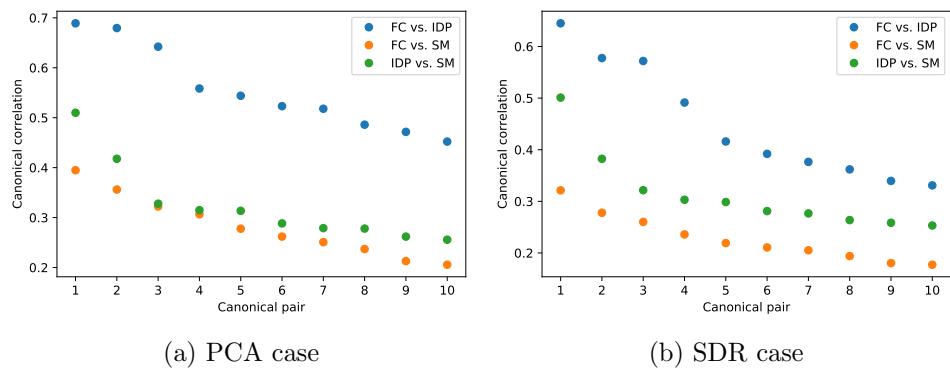


Figure B.28: Comparison between SDR and PCA on canonical correlations in pairwise CCA analysis.

PCA reduced canonical variables (Fig. B.28a) have slightly higher correlations compared with the SDR reduced case (Fig. B.28b). More specifically, correlations between FC and SM drop most, however, only with the decrease being less than 0.1, and correlations between FC and IDP are roughly the same.

Permutation testing on the PCA reduced data gives 12 significant canonical pairs between FC and SM, 9 pairs between IDP and SM, and 28 pairs between FC and IDP (Fig. B.29). These numbers are all higher than the SDR case (Fig. 5.19).

Due to different numbers of significant pairs are detected in the PCA and SDR cases, we show the variance explained by the first 10 canonical variables for both cases. Fig. B.30 illustrates that the SDR FC canonical variables explain more variance in the original data than the PCA case for both CCAs that FC was involved. SM canonical variables look very similar in the CCA between IDP and SM in both cases, and slightly lower for the SDR ones in the CCA with FC. SDR and PCA IDP canonical variables seem to explain roughly the same amount of variance in the CCA with FC, however for the CCA with SM, SDR IDP canonical variables explain less variance than the PCA ones. The general performances between PCA and SDR look comparable in terms of variance explained.

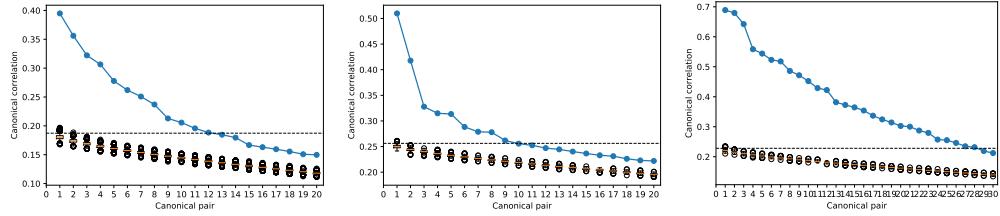


Figure B.29: Permutation testing on the PCA reduced data for 1000 permutes. True canonical correlations (blue line) versus the distribution of canonical correlation between the permuted canonical pairs (box plot). The dotted line is the 95 percentile of the distribution of the first permuted canonical pair (first box plot), which is used to define the significance of the canonical pairs (canonical correlation falls under the this line is defined as insignificant). From left to right are the CCA between FC and SM, IDP and SM, and FC and IDP respectively.

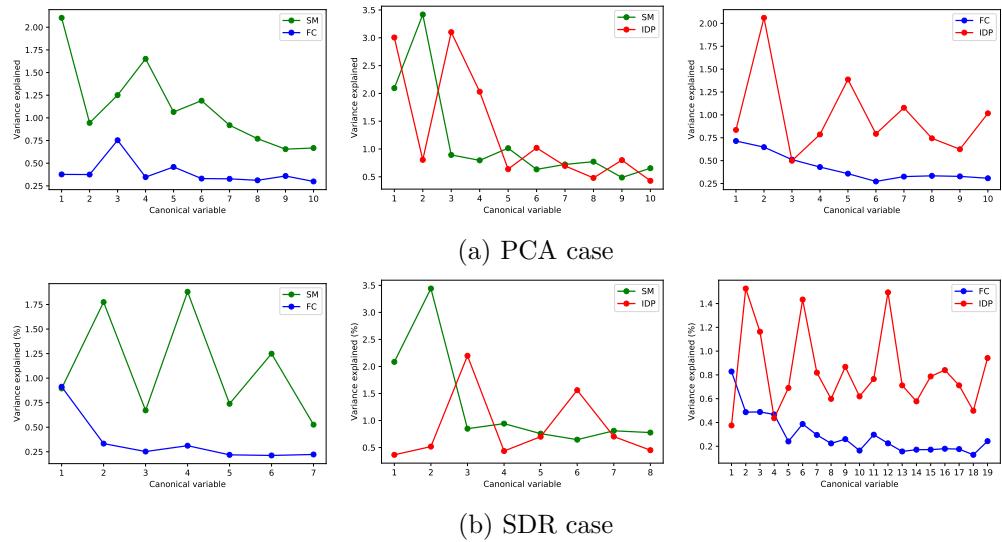


Figure B.30: Comparison between SDR and PCA on variance explained by the first 10 canonical variables in pairwise CCA analysis.

B.4.3.2 Canonical loadings

We are not going to interpret canonical loadings in the PCA case simply because they are not interpretable. As discussed earlier, principal components are linear combinations of the observed variables from the whole data space. Therefore, loadings on them would be even more abstract than the non-reduced case.

B.4.3.3 Multi-view CCA on PCA reduced data

Fig. B.31a shows the pairwise and sum canonical correlations in the multi-view setting for PCA. We observe the same non-monotonic behaviour as in the SDR case. The performances of canonical correlations for PCA and SDR in the multi-view setting are roughly comparable as well.

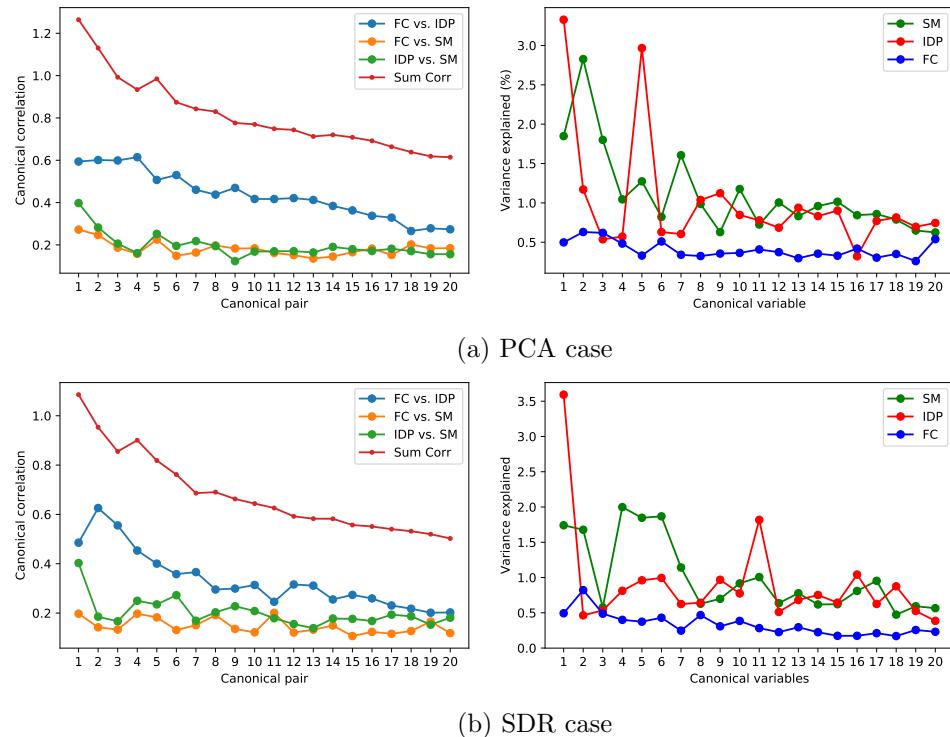


Figure B.31: Comparison of multi-view CCA between SDR and PCA on canonical correlation and variance explained.

We only implemented Permutation test 2 as introduced in 5.4.3, which is to test the significance of the sum of the canonical correlations. The sum of the correlations is more monotonic so that permutation testing gives unambiguous results. Fig. B.32 shows that there are 29 significant canonical sets in the multi-view

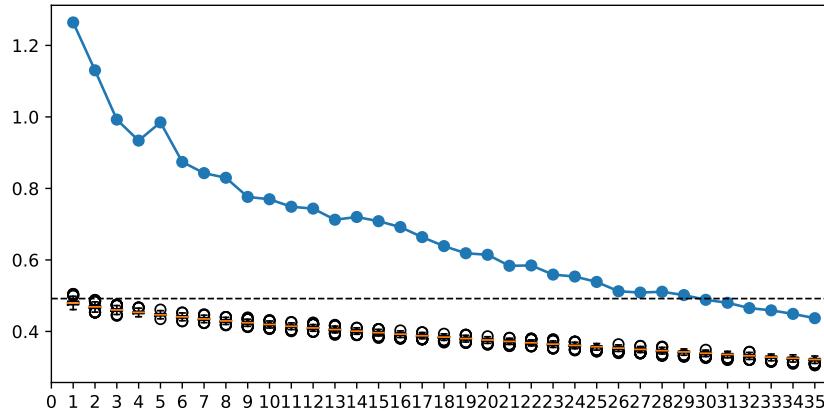


Figure B.32: Permutation testing on the sum of the canonical correlations for the multi-view PCA CCA.

PCA CCA. CV study gives 16 as the best number of canonical sets in the prediction setting.

B.4.4 Comparing SDR with PCA

From the above comparisons, we can see that in general, CCA on PCA reduced data would generate more correlated latent factors. We observed the same performance on the HCP data in Chapter 4. It can be explained by the nature of the methods. PCA reduces the data into an orthogonal latent space. When such space servers as the inputs for CCA, it offers more freedom to generate related latent factors compared with the same dimensional non-orthogonal space. However, SDR sacrifices the global orthogonality in the reduced space to represent the sub-domains. The orthogonality becomes local - factors are orthogonal within each domain. In terms of the variance explained by the canonical variables, both methods have their own wins and loses. It is hard to explain the behaviour of explained variance since the objective of CCA is to maximise between sets correlations. Furthermore, permutation testing and CV study always detect more significant/best canonical variables PCA. It may imply that the PCA canonical variables capture more structures in the data. However, these structures may not be informative and make the results harder to summarise. Moreover, as discussed earlier, the canonical loadings based on PCA factors are not interpretable at all.

APPENDIX C

Appendix for Predicting Personality Project

C.1 Significant network links

Table C.1: All significant links in the common network that is positively related to Openness to Experience using Power atlas. A link is the connection between the strength of the link and the Openness to Experience score. The links are ordered by R-value. * indicates no matching Brodmann area found by the MNI coordinates.

Brodmann Area 1	Brodmann Area 2	R-value	p-value
BA 22	BA 19	0.3376	8.02E-05
BA 40	BA 7	0.298	5.46E-04
BA 21	BA 39	0.2969	5.74E-04
BA 31	BA 8	0.2942	6.48E-04
BA 19	BA 40	0.2896	7.94E-04
*	BA 6	0.2874	8.72E-04
BA 6	*	0.2793	1.24E-03
BA 39	BA 21	0.2751	1.47E-03
BA 10	BA 2	0.2691	1.88E-03
BA 40	BA 20	0.2686	1.92E-03
BA 7	BA 8	0.2668	2.07E-03
BA 10	BA 23	0.2668	2.07E-03
BA 21	BA 19	0.2665	2.10E-03
BA 6	BA 47	0.2659	2.15E-03
BA 13	BA 40	0.2639	2.32E-03
Putamen	*	0.2636	2.35E-03
BA 4	BA 10	0.2629	2.41E-03
BA 7	BA 8	0.2625	2.45E-03
BA 19	BA 20	0.2619	2.51E-03
BA 19	BA 21	0.2616	2.55E-03
BA 47	BA 17	0.2597	2.74E-03
*	BA 40	0.2596	2.75E-03
BA 39	BA 19	0.259	2.81E-03
BA 6	BA 23	0.2585	2.87E-03
BA 24	BA 37	0.2579	2.94E-03

BA 40	BA 7	0.2561	3.15E-03
BA 10	BA 46	0.2559	3.17E-03
BA 13	BA 40	0.2557	3.21E-03
BA 24	BA 3	0.2541	3.41E-03
BA 7	BA 42	0.253	3.55E-03
BA 31	*	0.2523	3.65E-03
BA 21	BA 19	0.2522	3.66E-03
BA 39	BA 21	0.2519	3.70E-03
BA 37	BA 37	0.2517	3.73E-03
BA 10	BA 21	0.2517	3.73E-03
BA 42	*	0.2474	4.39E-03
*	*	0.245	4.80E-03
BA 19	BA 40	0.2896	7.94E-04
BA 46	BA 10	0.2559	3.17E-03
BA 40	BA 20	0.2686	1.92E-03
*	BA 31	0.2523	3.65E-03
*	BA 6	0.2793	1.24E-03
*	*	0.245	4.80E-03
*	BA 40	0.2596	2.75E-03
BA 10	BA 46	0.2559	3.17E-03
BA 7	BA 8	0.2668	2.07E-03
BA 19	BA 21	0.2522	3.66E-03
BA 19	BA 21	0.2665	2.10E-03
BA 19	BA 22	0.3376	8.02E-05
BA 19	BA 20	0.2619	2.51E-03
BA 40	*	0.2596	2.75E-03
BA 10	BA 2	0.2691	1.88E-03
BA 10	BA 4	0.2629	2.41E-03
BA 8	BA 7	0.2668	2.07E-03
BA 8	BA 7	0.2625	2.45E-03
BA 47	BA 17	0.2597	2.74E-03
BA 23	BA 10	0.2668	2.07E-03
BA 23	BA 6	0.2585	2.87E-03
Putamen	*	0.2636	2.35E-03
BA 40	BA 19	0.2896	7.94E-04
BA 40	BA 7	0.2561	3.15E-03
BA 47	BA 6	0.2659	2.15E-03
*	Putamen	0.2636	2.35E-03
BA 20	BA 40	0.2686	1.92E-03
BA 20	BA 19	0.2619	2.51E-03
BA 37	BA 24	0.2579	2.94E-03
BA 37		0.2517	3.73E-03
BA 40	BA 13	0.2639	2.32E-03
BA 40	BA 13	0.2557	3.21E-03
BA 7	BA 8	0.2625	2.45E-03
BA 19	BA 21	0.2616	2.55E-03
BA 7	BA 42	0.253	3.55E-03
BA 7	BA 40	0.298	5.46E-04
BA 7	BA 40	0.2561	3.15E-03

Appendix for Predicting Personality

Table C.2: All significant links in the common negative network of Extraversion using Power atlas. A link is the connection between Brodmann Area 1 to Brodmann Area 2. R-value is the Pearson's correlation between the strength of the link and the Extraversion score. The links are ordered by R-value. * indicates no matching Brodmann area found by the MNI coordinates.

Brodmann Area 1	Brodmann Area 2	R-value	p-value
BA 19	*	-0.4076	1.35E-06
BA 2	BA 19	-0.4032	1.80E-06
BA 6	BA 19	-0.3802	7.51E-06
BA 3	BA 7	-0.3588	2.57E-05
BA 39	BA 19	-0.3504	4.09E-05
BA 3	*	-0.35	4.17E-05
BA 37	BA 4	-0.3464	5.05E-05
BA 31	*	-0.3445	5.60E-05
BA 39	BA 19	-0.342	6.40E-05
BA 40	BA 32	-0.3414	6.60E-05
BA 31	Medial Dorsal Nucleus	-0.3398	7.18E-05
*	*	-0.3397	7.20E-05
BA 19	BA 41	-0.3382	7.81E-05
BA 19	*	-0.3367	8.42E-05
BA 7	*	-0.3362	8.62E-05
BA 19	Putamen	-0.3359	8.77E-05
BA 40	BA 32	-0.3355	8.94E-05
BA 19	Putamen	-0.3329	1.02E-04
BA 39	BA 32	-0.3323	1.05E-04
BA 3	BA 37	-0.3313	1.11E-04
*	BA 7	-0.3302	1.17E-04
BA 4	BA 19	-0.33	1.19E-04
BA 19	*	-0.3265	1.41E-04
Putamen	BA 39	-0.3255	1.49E-04
BA 4	BA 19	-0.3226	1.72E-04
BA 4	BA 6	-0.3178	2.17E-04
*	BA 30	-0.3176	2.19E-04
BA 37	BA 6	-0.3173	2.21E-04
BA 19	*	-0.317	2.25E-04
BA 2	BA 19	-0.3167	2.29E-04
BA 19	BA 32	-0.3162	2.34E-04
BA 19	*	-0.3156	2.41E-04
BA 7	Medial Dorsal Nucleus	-0.3143	2.56E-04
BA 43	BA 19	-0.314	2.59E-04
BA 39	*	-0.3128	2.75E-04
BA 31	*	-0.3127	2.76E-04
BA 39	BA 19	-0.3122	2.83E-04
BA 17	*	-0.3119	2.87E-04
BA 6	BA 19	-0.3116	2.91E-04
BA 20	BA 32	-0.3114	2.94E-04
*	BA 19	-0.3065	3.71E-04
BA 19	Putamen	-0.3064	3.73E-04
Putamen	BA 39	-0.3063	3.73E-04
BA 4	*	-0.3055	3.88E-04
*	BA 19	-0.3052	3.94E-04

BA 39	*	-0.3047	4.03E-04
*	BA 37	-0.3044	4.08E-04
BA 19	BA 3	-0.3025	4.46E-04
BA 6	BA 7	-0.3013	4.70E-04
BA 6	BA 19	-0.2989	5.23E-04
BA 17	*	-0.2987	5.30E-04
BA 39	BA 18	-0.2981	5.44E-04
BA 32	BA 20	-0.2964	5.87E-04
BA 47	BA 7	-0.2963	5.89E-04
BA 39	BA 40	-0.2961	5.96E-04
BA 6	BA 21	-0.296	5.99E-04
BA 4	BA 40	-0.2941	6.51E-04
BA 6	BA 19	-0.2936	6.65E-04
*	BA 7	-0.293	6.85E-04
BA 19	BA 19	-0.2927	6.93E-04
BA 6	BA 32	-0.2912	7.41E-04
*	BA 19	-0.291	7.46E-04
BA 19	BA 39	-0.2906	7.58E-04
*	*	-0.29	7.79E-04
BA 2	BA 18	-0.29	7.80E-04
BA 47	BA 22	-0.2899	7.84E-04
BA 18	BA 6	-0.2894	8.02E-04
*	BA 19	-0.289	8.15E-04
BA 4	BA 19	-0.2882	8.43E-04
BA 6	BA 39	-0.2882	8.44E-04
BA 6	*	-0.2877	8.64E-04
BA 3	BA 6	-0.287	8.90E-04
BA 20	BA 32	-0.2861	9.25E-04
BA 39	BA 18	-0.2849	9.71E-04
BA 31	Putamen	-0.2849	9.72E-04
BA 7	BA 18	-0.2848	9.79E-04
BA 19	BA 6	-0.2846	9.87E-04
BA 4	BA 7	-0.2845	9.89E-04
BA 18	*	-0.2838	1.02E-03
*	BA 7	-0.2836	1.03E-03
BA 4	BA 6	-0.283	1.06E-03
*	BA 4	-0.2829	1.06E-03
BA 18	BA 37	-0.2828	1.06E-03
BA 39	BA 32	-0.2827	1.07E-03
BA 4	BA 39	-0.2827	1.07E-03
BA 2	*	-0.2824	1.08E-03
BA 19	BA 6	-0.2821	1.10E-03
BA 19	*	-0.2819	1.11E-03
Ventral Anterior Nucleus	BA 19	-0.2818	1.11E-03
BA 7	*	-0.2818	1.11E-03
BA 19	BA 19	-0.2816	1.12E-03
BA 19	BA 39	-0.2816	1.12E-03
BA 4	BA 6	-0.2808	1.16E-03
BA 39	Medial Dorsal Nucleus	-0.2805	1.18E-03
BA 4	BA 4	-0.28	1.20E-03
BA 47	BA 19	-0.2795	1.23E-03
BA 4	*	-0.2792	1.24E-03
BA 6	BA 19	-0.279	1.25E-03

Appendix for Predicting Personality

BA 19	*	-0.2788	1.26E-03
BA 40	Putamen	-0.2771	1.35E-03
BA 19	*	-0.2771	1.36E-03
*	BA 31	-0.2766	1.38E-03
BA 17	*	-0.2766	1.38E-03
BA 19	BA 32	-0.2765	1.39E-03
BA 6	BA 6	-0.2763	1.40E-03
BA 8	BA 40	-0.276	1.42E-03
BA 7	BA 4	-0.2756	1.44E-03
BA 19	BA 6	-0.2751	1.47E-03
BA 3	*	-0.2734	1.58E-03
BA 7	BA 19	-0.2733	1.59E-03
BA 39	BA 32	-0.2733	1.59E-03
BA 4	BA 39	-0.273	1.61E-03
BA 6	BA 39	-0.2726	1.63E-03
BA 32	BA 40	-0.2725	1.64E-03
BA 19	BA 19	-0.2723	1.65E-03
BA 23	*	-0.2723	1.65E-03
*	BA 3	-0.2723	1.66E-03
*	BA 17	-0.2722	1.66E-03
BA 39	*	-0.2721	1.67E-03
BA 31	Medial Dorsal Nucleus	-0.2716	1.70E-03
BA 7	BA 19	-0.2713	1.72E-03
BA 7	BA 8	-0.2713	1.72E-03
BA 3	BA 19	-0.2709	1.75E-03
BA 32	BA 40	-0.2707	1.76E-03
BA 37	BA 19	-0.2706	1.77E-03
*	*	-0.2706	1.78E-03
BA 21	BA 6	-0.2705	1.78E-03
BA 4	BA 7	-0.2705	1.78E-03
*	*	-0.2702	1.80E-03
BA 32	BA 39	-0.2701	1.81E-03
BA 31	BA 7	-0.2701	1.81E-03
BA 39	BA 17	-0.2699	1.82E-03
BA 39	BA 23	-0.2696	1.85E-03
BA 7	*	-0.2695	1.86E-03
BA 7	BA 19	-0.2693	1.87E-03
BA 4	BA 7	-0.2691	1.89E-03
BA 6	BA 32	-0.269	1.89E-03
*	*	-0.2689	1.90E-03
BA 17	*	-0.2685	1.93E-03
BA 4	BA 19	-0.2684	1.94E-03
BA 6	BA 39	-0.268	1.97E-03
BA 19	BA 2	-0.268	1.97E-03
BA 19	BA 2	-0.2677	1.99E-03
BA 32	BA 39	-0.2677	1.99E-03
BA 21	Ventral Anterior Nucleus	-0.2676	2.01E-03
BA 19	BA 7	-0.2672	2.03E-03
BA 20	Medial Dorsal Nucleus	-0.2662	2.12E-03
BA 19	*	-0.2662	2.12E-03
*	BA 32	-0.2656	2.17E-03
BA 7	BA 32	-0.2654	2.19E-03
BA 6	BA 18	-0.2653	2.20E-03

BA 19	BA 4	-0.2653	2.20E-03
BA 19	BA 7	-0.2649	2.23E-03
BA 2	*	-0.2649	2.23E-03
Putamen	BA 39	-0.2645	2.27E-03
BA 18	BA 7	-0.2644	2.28E-03
Putamen	BA 39	-0.2643	2.28E-03
BA 41	BA 17	-0.2641	2.31E-03
BA 6	BA 18	-0.264	2.31E-03
BA 6	BA 39	-0.2638	2.34E-03
BA 19	*	-0.2636	2.35E-03
BA 31	*	-0.2632	2.39E-03
BA 19	BA 6	-0.2629	2.41E-03
BA 19	BA 7	-0.2629	2.42E-03
*	BA 7	-0.2622	2.49E-03
BA 31	BA 6	-0.262	2.50E-03
BA 19	Putamen	-0.2616	2.54E-03
BA 39	*	-0.2614	2.56E-03
BA 3	BA 19	-0.2612	2.58E-03
BA 19	BA 6	-0.261	2.60E-03
BA 4	BA 6	-0.2609	2.62E-03
BA 4	BA 4	-0.2609	2.62E-03
BA 2	BA 39	-0.2608	2.62E-03
BA 17	BA 7	-0.2601	2.70E-03
BA 19	BA 2	-0.26	2.71E-03
BA 19	*	-0.2599	2.72E-03
*	*	-0.2596	2.76E-03
BA 32	*	-0.2595	2.76E-03
BA 7	BA 21	-0.2595	2.76E-03
BA 3	BA 7	-0.2592	2.80E-03
BA 8	Medial Dorsal Nucleus	-0.2591	2.80E-03
BA 39	Putamen	-0.259	2.82E-03
BA 4	BA 19	-0.2589	2.82E-03
BA 19	BA 32	-0.2586	2.86E-03
BA 6	BA 32	-0.2582	2.90E-03
BA 7	BA 6	-0.258	2.92E-03
BA 31	*	-0.2575	2.98E-03
BA 8	BA 37	-0.2574	2.99E-03
*	BA 18	-0.2574	3.00E-03
BA 6	BA 19	-0.257	3.05E-03
BA 39	BA 47	-0.2569	3.05E-03
BA 40	*	-0.2569	3.06E-03
BA 6	BA 6	-0.2568	3.07E-03
BA 19	BA 32	-0.2568	3.07E-03
*	BA 19	-0.2567	3.08E-03
BA 39	BA 19	-0.2565	3.10E-03
BA 18	BA 7	-0.2565	3.10E-03
BA 9	BA 21	-0.2564	3.12E-03
BA 7	BA 18	-0.2563	3.13E-03
BA 4	BA 37	-0.256	3.16E-03
BA 19	BA 32	-0.2557	3.20E-03
*	BA 40	-0.2552	3.27E-03
BA 4	BA 19	-0.255	3.29E-03
BA 4	BA 6	-0.2547	3.33E-03

Appendix for Predicting Personality

BA 19	BA 32	-0.2545	3.35E-03
*	BA 19	-0.2544	3.37E-03
*	*	-0.2538	3.45E-03
BA 4	BA 37	-0.2537	3.46E-03
*	BA 23	-0.2537	3.46E-03
BA 3	BA 6	-0.2535	3.48E-03
BA 3	BA 19	-0.2533	3.50E-03
*	*	-0.2533	3.51E-03
BA 10	BA 9	-0.2526	3.61E-03
BA 4	BA 19	-0.2525	3.62E-03
BA 32	BA 39	-0.2523	3.64E-03
BA 2	BA 3	-0.2523	3.64E-03
BA 19	BA 32	-0.2518	3.72E-03
BA 3	BA 19	-0.2518	3.72E-03
BA 2	BA 18	-0.2514	3.77E-03
BA 6	BA 9	-0.251	3.83E-03
BA 19	Putamen	-0.2509	3.84E-03
BA 4	BA 19	-0.2507	3.87E-03
BA 17	*	-0.2507	3.88E-03
BA 31	BA 47	-0.2507	3.88E-03
BA 19	BA 4	-0.2492	4.11E-03
BA 19	*	-0.2488	4.16E-03
BA 39	Putamen	-0.2483	4.25E-03
Putamen	BA 19	-0.2481	4.28E-03
*	BA 19	-0.2479	4.31E-03
BA 19	BA 4	-0.2475	4.38E-03
BA 19	BA 18	-0.245	4.79E-03