Supervised Dimension Reduction Canonical Correlation Analysis of UK Biobank

Alan Turing Zhangdaihong Liu^{1,2}; Kirstie Whitaker^{2,3}; Stephen Smith⁴; Thomas Nichols⁵ ¹MathSys CDT, University of Warwick; ²Alan Turing Institute; ³University of Cambridge; ⁴FMRIB, University of Oxford; ⁵Big Data Institute, University of Oxford



Introduction

UK Biobank data offers us great opportunities for discovering relationships between different types or 'modalities' of health data, especially between neuroimaging and nonimaging data. In this work, we consider three such modalities, functional and structural imaging, and behavioral variables. We apply a multivariate method, canonical correlation

analysis (CCA), and explore relationships between modalities with and without dimension reduction on the input data. We use a supervised dimension reduction method based on sub-domains of each modality. We present results on 9301 UK Biobank subjects, and compare the merits of the two different ways of applying CCA.

Data	Results		
 We considered three modalities in the data: Subject measure (SM): Includes behavioural and demographics variables consisting 698 variables 	We applied the above analysis to every pair of the three modalities. Here, we present results on the CCA of SM and FC.	0.665 0.6666 0.666 0.666 0.666 0.666 0.666 0.666 0.6666 0.666 0.666 0.666 0.66	

- ieniugiaphilus vanabies, consisting 050 vanabies from 8 sub-domains.
- Functional connectivity (FC): Derived from rfMRI, \bullet using same 55-node ICA parcellation as in Miller et al. (2016), which gives 1485 edges for each subject. Here, the nodes define 55 sub-domains.
- Structural imaging derived phenotypes (**sIDP**): Includes structural and diffusion MRI, with 828 variables in 10 sub-domains.

Data pre-processing

The

Institute

For all three modalities, variables with more than 50% missing values were dropped. For highly correlated pairs of variables (r>0.99), the one with higher missingness was dropped. SM was normalised by rank-based inverse normalisation and missing values were imputed by soft impute (Mazumder et al. 2010); FC and sIDP were standardized and missing values in IDP were then filled with mean. Finally, all three modalities were de-confounded with age, sex, scan date, head size, rfMRI head motion and tfMRI head motion.

Methods

Non-reduced CCA

Figure 1 top shows the canonical correlations (blue) for the first 10 canonical components and the variance explained (orange) by each of them in the original dataset. Permutation testing gave 3 pairs of significant canonical variables. Figure 2 shows the top SM canonical loadings for the first 3 canonical variables.

SDR CCA

Table 1 shows the number of dimensions reduced by SDR in each SM sub-domain. The SM input of SDR CCA is the concatenation of these numbers of PCs from each sub-domain.

Mental health	Health & Medical History	Alcohol	Tobacco	
9/49	18/122	7/19	3/14	
Cognition	General lifestyle& Environment	Exercise & work	Food & drink	
28/271	19/76	23/90	13/57	

Table 1. Dimensionality of 8 SM sub-domain that was reduced by SDR. The denominators show the total number of variables in each sub-domain.

Canonical correlations and variance explained for the first 10 canonical variables are shown in the bottom subplot in figure 1. Compared with the non-reduced CCA, canonical correlations have decreased significantly. However, the variance explained by each of the canonical variables has increased. Same permutation testing gave 6 pairs of significant canonical variables. Figure 3 shows the mean squares of positive and negative loadings suggest the role of each sub-domain for the first 3 canonical loadings ; e.g. "Exercise & work" dominates the first mode while "General Lifestyle & Environment" dominates the second.



Canonical correlation analysis (CCA)

CCA finds relationship between two sets of variables, finding modes of maximal correlation between pair of datasets.

Supervised dimension reduction (SDR)

We apply SDR in each sub-domain, aiding in the interpretation of the discovered CCA modes. SDR consists of a Principal Components (PC) dimensionreduction, where a two-way cross-validation (CV) method is used to automatically estimate the dimensionality in each sub-domain. The final input to CCA is the concatenated PC from each of the subdomains.

We compare CCA computed in 2 ways: Without any dimension reduction ("non-reduced CCA") and with SDR as described ("SDR CCA").

Measures of assessment

We assess the performance of CCA by looking at the canonical correlations, canonical loadings and the variance explained by the canonical variables. The number of significant canonical variables was determined (1000 permutation bv testing permutations).

Canonical variable

Figure 1. Canonical correlations and variance explained by the first 10 canonical variables in the original SM and FC datasets for non-reduced CCA (top) and SDR CCA (bottom).

0.4	Year ended full time education Fluid intelligence score Cheese intake Age completed full time education Maximum digits remembered correctly FI3 : word interpolation (Fluid intelligence) Average total household income before tax Age first had sexual intercourse Average total household income before tax	0.24	Time to complete round (Pairs matching) Duration to complete alphanumeric path (Trail making) Duration screen displayed (Prospective memory) Interval taken in alphanumeric path (Trail making) Number of incorrect matches in round (Pairs matching) Duration to entering value (Symbol digit substitution) Duration to complete numeric path (Trail making)	0.19	Time to complete touchscreen questionnaire Number of live births Number of vehicles in household Time to complete touchscreen questionnaire visit-3 Variation in diet Alcohol intake frequency Number in household Job involves mainly walking or standing Hands-free device/speakerphone use with mobile phone in last 3 month Number of vehicles in household Time spend outdoors in summer Cereal intake
).2).4	Time spent outdoors in winter Job involves mainly walking or standing Job involves heavy manual or physical work Time spend outdoors in summer Time spent watching television (TV) Qualifications (higher value means lower level)	-0.15	Number of fluid intelligence questions attempted within time limit Number of symbol digit matches made correctly Number of symbol digit matches attempted Fluid intelligence score Plays computer games (higher value mean worse degree)	0.11 -0.11 -0.12	Daytime dozing / sleeping (narcolepsy) Weekly usage of mobile phone in last 3 months Qualifications Past tobacco smoking Happiness Smoking status Maximum digits remembered correctly Frequency of other exercises in last 4 weeks

Figure 2. Top SM canonical loadings from the non-reduced CCA of SM and FC for the significant canonical variables. Left to right are the first to the third canonical loadings.



Conclusions

From non-reduced CCA, we can produce highly correlated latent variables, which however explain less variance. With SDR CCA, we sacrifice canonical correlation to produce canonical variables which explain more variance and more interpretable. By applying SDR, we are able to track contributions of each sub-domain. Moreover, SDR allows us to produce canonical loadings in two ways (as shown in figure 3), which indicate the relationship between canonical variable with observed data and with CCA inputs respectively. In contrast, when PCA is used en masse domain-specific canonical loadings (Fig. 3, lighter bars) cannot be computed for the dimensionreduced dataset.

Figure 3. Mean squares (MS) of SM canonical loadings in each sub-domain. Dark blue and red bars are amount of variance the canonical variables explain in the observed SM dataset; light blue and orange bars show the amount of variance explained in the CCA input (SDR reduced SM). Bars above zero are the MS for all variables with positive loadings in each sub-domain; Bars blow zero are MS for all variables with negative loadings. We present them on the negative axis to show contrast with the positive loadings. In general the right sets of bars are larger than the left ones due to the lower dimensions of the CCA input compared with the observed dataset.

Acknowledgement

Z.Liu would like to acknowledge the generous support of the Guarantors of Brain, Mathsys CDT at The University of Warwick and the Alan Turing Institute in attending this conference, and the Chinese Scholarship Council for their support of her PhD research.

References

1. Miller, Karla L., et al. Nature neuroscience 19.11 (2016): 1523. 2. Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani. Journal of machine learning research11.Aug (2010): 2287-2322. 3. Smith, Stephen M., et al. Nature neuroscience 18.11 (2015): 1565.