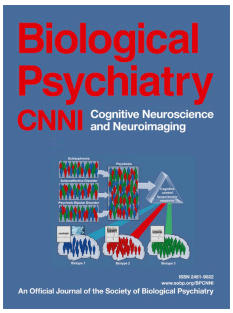


Canonical correlation analysis for identifying biotypes of depression

Agoston Mihalik, Rick A. Adams, Quentin Huys

PII: S2451-9022(20)30032-X

DOI: https://doi.org/10.1016/j.bpsc.2020.02.002

Reference: BPSC 556

To appear in: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*

Received Date: 3 February 2020

Accepted Date: 4 February 2020

Please cite this article as: Mihalik A., Adams R.A. & Huys Q., Canonical correlation analysis for identifying biotypes of depression, *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (2020), doi: https://doi.org/10.1016/j.bpsc.2020.02.002.

Canonical correlation analysis for identifying biotypes of depression

Agoston Mihalik[1,2,*], Rick A. Adams[1,2,*], Quentin Huys[2,3]

[1] *Centre for Medical Image Computing, Department of Computer Science, University College London, United Kingdom*

[2] *Max Planck-UCL Centre for Computational Psychiatry and Ageing Research, University College London, United Kingdom*

[3] *Division of Psychiatry, University College London, United Kingdom*

* *These authors contributed equally to this work*

Corresponding author: Agoston Mihalik MD, PhD (90 High Holborn, London WC1V 6LJ, Department of Computer Science, UCL, UK; +44 7552235333; a.mihalik@ucl.ac.uk)

An old question is challenging novel methods: do diagnoses such as schizophrenia (1) and depression (2) contain subgroups of patients? And might such subgroups be identifiable by neurobiological means, and have differential responses to therapies? If so, then long-awaited biomarkers for psychiatric diagnosis and therapeutics might be found.

These powerful multivariate methods can detect patterns of common variation in symptom and brain variables. One example is canonical correlation analysis (CCA). CCA finds linear combinations of variables within two types of datasets (e.g. biological and clinical measures) that maximally correlate with each other, termed 'canonical variates'. Such linear combinations of variables can be thought of as 'factors' in each of the datasets which relate to each other. However, brain data can be very high-dimensional, making CCA prone to overfitting (i.e. finding spurious associations that exist by chance in a given sample and thus don't generalize). Rigorous methodological steps such as regularization and validation on held-out data are therefore necessary to avoid this.

Much can be learned about these issues from a landmark study applying CCA to clinical and resting state functional magnetic resonance imaging (rs-fMRI) data from participants with depression (2) and subsequent discussions between those authors and another group (3–6). Here we summarise these exchanges:

Drysdale et al (2) applied CCA to a large rs-fMRI dataset of treatment-resistant major depression (Figure 1A). As subjects' full connectivity profiles were too high-dimensional, they preselected a small number of connections that correlated (p<0.005) with one or more items of the Hamilton Depression Rating Scale (HAMD). These connections and the HAMD items were analysed using CCA, yielding two canonical variates relating two sets of connections broadly to anxiety and anhedonia, respectively. A clustering algorithm then grouped the subjects into four 'biotypes': groups with high/low anxiety and high/low anhedonia canonical variate scores. The biotypes generalized in clinically meaningful ways

('clinical validation'): one was associated with successful response to repetitive transcranial magnetic stimulation (rTMS) therapy. Second, patients with generalized anxiety disorder tended to fall into three of the biotypes, and when they did they also had depressive symptoms. Third, however, patients with schizophrenia did not fall into any of the biotypes.

Dinga et al (3) attempted to replicate these findings in a separate sample of 187 subjects with a history of depression or anxiety disorder, but using more rigorous versions of the same methods (Figure 1B). In particular, they used permutation testing of the entire feature selection and CCA procedure to assess the significance of the canonical variates, and cross-validation to assess their generalizability. They also tested the null hypothesis that 'no clusters' were present in the data, and assessed the stability of both the canonical loadings (i.e. univariate correlations between variables and canonical variates) and clusters by repeating the whole process with single subjects omitted. They found that the canonical correlations – though very high ($\geq$0.97) – were not higher than expected by chance and were very low in out-of-sample data ($<|0.1|$). The null hypothesis of 'no clusters' could not be rejected, and the clusters were very unstable. They did not assess whether Drysdale et al's canonical variates could predict symptom scores from rs-fMRI data in their sample.

In response, Grosenick et al (4) repeated their feature selection and CCA analysis on their previous 220 participants, adding L2 regularisation (see below) to reduce overfitting and cross-validation to assess stability. Regularisation made an enormous difference to the stability of the canonical variates, increasing out-of-sample canonical correlations from medians of around 0 to 0.85 (for ~176 rs-fMRI features). However, as Dinga et al also point out (5), the response by Grosenick et al. does not clarify whether the original CCA's canonical variates or clustering results can be reproduced ('statistical validation'). Commendably, this is work in progress (6).

Statistical validation is extremely important. A strength of the original paper was the multiple clinical validations. However, these were either not fully independent (in the case of rTMS response), or mainly relating to diagnoses rather than the biotypes. Indeed, if there are underlying relationships between prefrontal cortical connectivity in depression and rTMS response, and between symptoms in GAD and in depression, then even unstable clusters may have distinct relationships to rTMS response or symptoms. Hence validation of any specific set of biotypes must be statistical in the first instance. *Clinical* validation, however, will make biotypes practically useful.

Grosenick et al. also suggest that an important factor in Dinga et al's failure to find a significant canonical variate in their own analysis may have been the heterogeneity and lower severity of psychopathology in their sample. They did not test this suggestion using CCA in their own mildly-unwell sample, however (4).

We believe there are some key lessons to learn from this illuminating exchange:

A conceptual point of importance is that when using clustering methods, the null hypothesis of 'no clusters' is often (wrongly) disregarded. Nevertheless, clusters *per se* aren't necessarily important. Defining key axes of variation (i.e. using a dimensional rather than categorical approach) could be just as (or more) useful for assigning treatments or predicting outcomes. Indeed, one could even test whether categorical or dimensional approaches have better predictive power.

In addition, some might conclude from the strengths of the univariate brain-symptom associations in severely versus mildly unwell participants in (4) that future analyses ought to be conducted in more homogeneous samples with more severe illness. However, we would not necessarily recommend this approach. First, such samples are harder to obtain and therefore (usually) smaller. Second, only mixed samples can demonstrate what variance is unique to a disorder (or its subgroups), and what is transdiagnostic. Third, we speculate that

the failure to find any significant canonical variates in (3) is more due to the methods used than the sample's milder, more heterogeneous illnesses. The lack of regularisation causes huge overfitting (the median of the null distribution of canonical correlations is around 0.99), making it very hard for any genuine associations to achieve statistical significance. With regularisation, a depression-related canonical variate can be found in a mildly unwell sample (7).

To conclude, we discuss three aspects of the CCA methodology in detail, and make some recommendations (Figure 1C).

First, out-of-sample evaluation is crucial, given multivariate methods' strong tendency to overfit high-dimensional data. To perform statistical inference on how the fitted CCA model generalizes to unseen data, independent test data are needed. Critically, when a validation set is used to select the optimal regularization parameter (see below), then three divisions of the data are required: training, validation and test data (7). We also recommend performing statistical evaluation on out-of-sample correlations (i.e., test canonical correlations) as it assesses the generalization of the CCA model explicitly.

Second, we prefer regularization to feature selection. Standard CCA is limited to cases when the number of examples (e.g. number of subjects) exceeds the number of variables (e.g. connectivity features). One can therefore employ feature selection – using univariate tests or principal component analysis (PCA) – to reduce the number of variables (as in (2)). However, if the number of selected features is not sufficiently reduced, CCA is still prone to overfitting, as shown in (4). Moreover, if the top-ranked variables in one dataset are highly intercorrelated, these feature selection techniques will favour their inclusion in the CCA at the expense of lower-ranked variables that can potentially account for more shared variance across the datasets.

Regularization avoids the need for feature selection, reduces overfitting, and can bring additional benefits to the CCA model. For instance, L1 regularization results in automatic feature selection (i.e. sparse solutions) which facilitates the interpretation of the results; L2 regularization makes CCA more stable (8). Elastic net regularization, combining L1 and L2, has the advantages of both, which explains its overwhelming popularity in CCA applications, e.g. (7,9). Interestingly, Partial Least Squares (PLS) can be viewed as a special case of CCA with maximal L2 regularisation, which maximises the covariance (rather than correlation) between the datasets (8). Regularized CCA/PLS models can be further improved by using stability selection (9) or a stability criterion (7), which promotes the inclusion of variables in the model that appear consistently across different subsamples of the data.

Third, standard (or regularized) CCA is limited to extracting linear combinations of variables. Non-linear extensions, such as kernel CCA with non-linear kernels, could explore more complex relationships between the datasets. A promising future direction involves the combination of more than two types of datasets, e.g. functional and structural brain data and behavior. This was recently demonstrated in (9) and may enable a more complete description of latent neurobiological (and other) factors. Probabilistic approaches, such as Group Factor Analysis (10), can decompose two datasets into both shared and unique variances, and are also better equipped to deal with missing data (further references to these novel methods are in (7)).

These are highly complex methods, and novel ones are constantly being developed. Best practices are changing year on year. Critical exchanges such as these are invaluable in advancing the field and building expertise, and as such should be welcomed.

Acknowledgements and disclosures

Agoston Mihalik is supported by the Wellcome Trust under grant number WT102845/Z/13/Z. Rick A. Adams is supported by an MRC Skills Development Fellowship (MR/S007806/1). The other authors report no biomedical financial interests or potential conflicts of interest.

Figure legends

Figure 1. Schematic illustration and comparison of the canonical correlation analysis (CCA) pipelines used in Drysdale et al (2), Dinga et al (3) and our suggestion for future studies. NB, we have omitted the clustering procedures and their statistical evaluation for clarity. **(A)** Drysdale et al trained a CCA model on a selected set of rs-fMRI features (178 out of >33000 features that correlated (p<0.005) with one or more HAMD items) and 17 HAMD items. They used a parametric test (orange box) to assess the statistical significance of the results, which did not take into account the previous feature selection step. They did not perform out-of-sample evaluation of the CCA model, i.e., calculating canonical correlations on test data using the trained CCA model from the previous step. **(B)** Dinga et al copied Drysdale et al's feature selection method but used a permutation test (orange box) on the (in-sample) canonical correlation of the trained CCA model to assess the statistical significance of the results, taking into account the previous feature selection step. In addition, they calculated test (out-of-sample) canonical correlations using the trained CCA model, moreover, they used cross-validation including all previous steps (green box) to assess the robustness of the results. **(C)** We suggest using regularization rather than feature selection: the regularization parameter can be selected in an inner cross-validation loop (dashed green box). A CCA model can then be trained on the brain and behavioral features using the best regularization parameter. To assess the statistical significance of the results, we recommend a permutation test (orange box) on the test canonical correlation, including re-training the CCA model.

Optionally, the permutation test can extended to include the feature selection step as well, however, it greatly increases the computational costs, and is unnecessary if the test canonical correlation is included in the permutation test. Finally, the whole procedure can be embedded in a cross-validation (green box) to assess the robustness of the results.

References

1. Clementz BA, Sweeney JA, Hamm JP, Ivleva EI, Ethridge LE, Pearlson GD, *et al.* (2016): Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *Am J Psychiatry* 173: 373–384.

2. Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, *et al.* (2017): Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med* 23: 28–38.

3. Dinga R, Schmaal L, Penninx BWJH, van Tol MJ, Veltman DJ, van Velzen L, *et al.* (2019): Evaluating the evidence for biotypes of depression: Methodological replication and extension of. *NeuroImage Clin* 22: 101796.

4. Grosenick L, Shi TC, Gunning FM, Dubin MJ, Downar J, Liston C (2019): Functional and Optogenetic Approaches to Discovering Stable Subtype-Specific Circuit Mechanisms in Depression. *Biol Psychiatry Cogn Neurosci Neuroimaging* 4: 554–566.

5. Dinga R, Schmaal L, Marquand AF (2020): A Closer Look at Depression Biotypes: Correspondence Relating to Grosenick et al. (2019). *Biol Psychiatry Cogn Neurosci Neuroimaging*. https://doi.org/10.1016/j.bpsc.2019.09.011

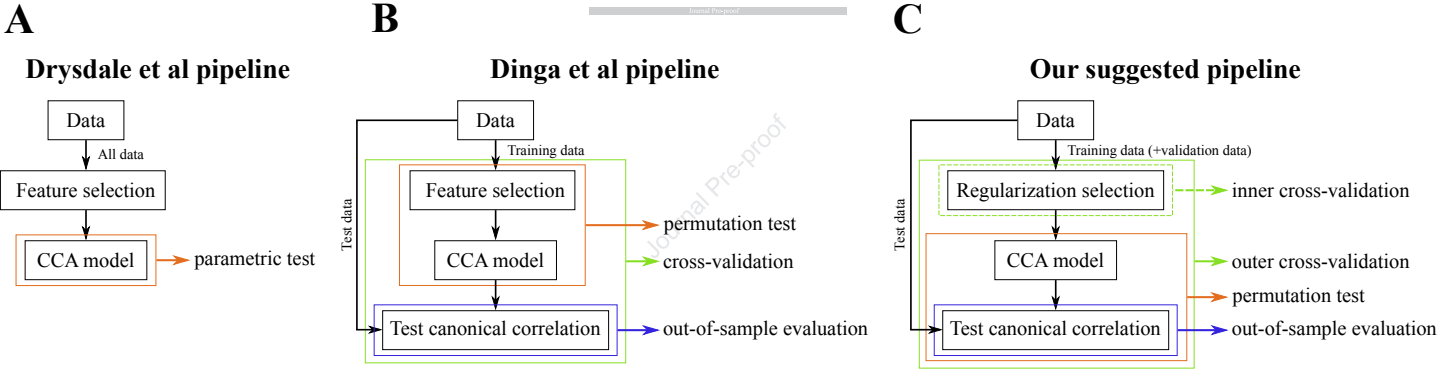
6. Grosenick L, Liston C (2020): Reply to: A Closer Look at Depression Biotypes: Correspondence Relating to Grosenick et al. (2019). *Biol Psychiatry Cogn Neurosci Neuroimaging*. https://doi.org/10.1016/j.bpsc.2019.11.002

7. Mihalik A, Ferreira FS, Moutoussis M, Ziegler G, Adams RA, Rosa MJ, *et al.* (2020): Multiple Holdouts With Stability: Improving the Generalizability of Machine Learning Analyses of Brain–Behavior Relationships. *Biol Psychiatry* 87: 368–376.

8. Shawe-Taylor J, Cristianini N (2004): *Kernel Methods for Pattern Analysis*. USA: Cambridge University Press.

9. Ing A, Sämann PG, Chu C, Tay N, Biondo F, Robert G, *et al.* (2019): Identification of neurobehavioural symptom groups based on shared brain mechanisms. *Nat Hum Behav* 3: 1306–1318.

10. Klami A, Virtanen S, Leppäaho E, Kaski S (2015): Group Factor Analysis. *IEEE Trans Neural Netw Learn Syst* 26: 2136–2147.


A
Drysdale et al pipeline
Data
All data
Feature selection
CCA model
parametric test
B
Dinga et al pipeline
Data
Training data
Test data
Feature selection
CCA model
Test canonical correlation
permutation test
cross-validation
out-of-sample evaluation
C
Our suggested pipeline
Data
Training data (+validation data)
Test data
Regularization selection
CCA model
Test canonical correlation
inner cross-validation
outer cross-validation
permutation test
out-of-sample evaluation