Evaluating a Novel Methodology for Multivariate Analysis of Brain-Behavior Relationships Across Datasets

Nina Bernick
advised by Dr. John Murray and Dr. Markus Helmer
Murray Lab, Yale School of Medicine

May 6, 2022

A Senior Thesis in the Department of Applied Mathematics

Abstract

Canonical Correlation Analysis (CCA) is a widely used method for discovering and analyzing multivariate associations between datasets. CCA is often used in computational neuroscience to relate neural and behavioral data, but the limited sample size of datasets in the field pose considerable issues for the method. When applied to high-dimensional datasets with limited sample sizes, as is common in practical applications, CCA can be highly unstable and overestimate the association strength between datasets. Additionally, CCA is unable to harness the power of multiple datasets if the feature sets are distinct. To address these issues at small sample sizes, an extension of CCA called Combination Canonical Correlation Analysis (C3A) is proposed to incorporates an additional dataset(s) with distinct behavioral features. By relying on underlying common modes of association between the datasets, this can effectively increase the sample size of the primary dataset to improve reliability of the analysis. Exploratory testing of the performance of C3A was conducted using generated synthetic data sets that model empirical data trends. Results are visualized over several parameters of the generated data, such as the number of samples per feature, true correlations, and similarity of neural weight vectors between datasets. To attempt to further improve C3A, an adaptation that considers within the optimization the number of samples per feature in each dataset is proposed and tested. Both C3A and its adaptation show significant improvement in predicting association strength and weighting vectors over CCA for small sample sizes, showing promise for usefulness in analyzing brain-behavior associations.

Contents

1 Introduction

Within the study of brain-behavior relationships, major knowledge gaps persist regarding how neural variations map onto clinical features across psychiatric disease. Psychiatric diagnoses of patients are made using categorical groupings of behavioral symptoms according to the Diagnostic and Statistical Manual of Mental Disorders (DSM), but it has been shown that these diagnoses map poorly onto neural perturbations [1, 2]. A significant open question in neuropsychiatry is whether there exists a better scale for combining behavioral scores that more effectively explains variation in affected neural circuits.

One area of inquiry within this problem space focuses on identifying latent associations between individual variation in behavioral features with neural features for a set of subjects, irrespective of DSM diagnosis [3, 4]. Large-scale publicly available datasets such as the Human Connectome Project (HCP)[6], Connectomes Related to Human Disease (CRHD), and the UK Biobank [7] provide promising resources for discovering such associations through statistical analysis on high-quality and clinically variable neural and behavioral data.

A common approach for mapping multivariate associations is Canonical Correlation Analysis (CCA) [5, 10], which will be discussed in greater depth in the following section. Despite its popularity, CCA has been shown to face significant challenges in stability for small sample sizes which are characteristic of neural-behavioral datasets [8]. CCA is also unable to harness the power of multiple datasets if their features are distinct – if we have one dataset of neural and behavioral measures for a group of individuals, and a second dataset of identical neural and distinct behavioral measures for a second group, CCA would not be able to consider both these datasets in the analysis. Because of the high cost and time necessary to collect neural-behavioral data, most studies are of small-to-medium size and individually lack sufficient sample size to produce stable and reproducible results given the number of features present in the data [8, 9]. Additionally, separate studies often collect different behavioral measures. Harnessing the power of multiple datasets presents an opportunity to conduct more robust and reliable analyses, but to date there is a significant lack of quantitative frameworks that would allow researchers to combine data across multiple studies and datasets.

This paper presents a potential framework for utilizing multiple datasets to discover and analyze latent associations between behavioral features and neural variation. Building off of CCA and extending it to multiple datasets, we propose C3A, a method that simultaneously performs CCA on each dataset while constraining the weight vectors to be similar. This method has the potential to improve accuracy and stability of multivariate association analyses by incorporating supplemental, nonoverlapping datasets that effectively increase the sample size of the analysis. Since the regime of small sample sizes is where CCA fails to provide reliable results and where many neuro-behavioral datasets fall, a method that improves accuracy in that domain has the potential to improve multivariate analyses that are most often used in computational neuroscience.

2 Canonical Correlation Analysis

2.1 Mathematical Definition

Canonical Correlation Analysis (CCA) [5] is one of several common methods used to discover associations between two multivariate data sets [10]. Each dataset is composed of multiple variables measured on a single set of individuals. Note that our definition of “multiple” datasets depends on context: the “two” datasets here can be considered as partitions of the same dataset, since they are measured over the same group of individuals; this is separate from the above discussion that considers the case of two datasets on different populations, each with neural and behavioral features. Given two datasets X and Y, CCA defines linearly weighted composites or ‘scores’ of features in each dataset and chooses the set of weights that maximizes the correlation between scores of the two datasets. Predicted association strength is an important output of the method, as are the weight vectors, which reveal how each feature in the dataset contributes to the association. Many other metrics can be calculated, but we focus on these two in the report.

To provide a strong background on the mathematical process of CCA, the following derivation is provided from [13]. Let X and Y be zero-meaned datasets of sizes $n \times p$ and $n \times q$, respectively. The rows of X and Y represent observations from a population of size n and the columns denote the variable or feature vectors for each subject. Consider the following transformations:

$$\begin{aligned} z_a &= w_a X \\ z_b &= w_b Y \end{aligned} \tag{1}$$

where $w_a \in R^p$ and $w_b \in R^q$ are called the canonical weight vectors and $z_a, z_b \in R^n$ are called the canonical variates or scores. The goal of CCA is to find the set of weights w_a and w_b such that the cosine of the angle θ between z_a and z_b is maximized, subject to the constraints that z_a and z_b are unit norm vectors and $\theta \in [0, \pi/2]$. This can be written as

$$\cos\theta = \max_{z_a, z_b \in R^n} \langle z_a, z_b \rangle \tag{2}$$

$$||z_a|| = 1 \quad ||z_b|| = 1 \tag{3}$$

The cosine of the angle θ is also called the canonical correlation. There exist $min(p, q)$ pairs of (z_a^i, z_b^i), each z_a^i orthogonal to all $z_a^j \ \forall j \neq i$ and z_b^i orthogonal to all $z_b^j \ \forall j \neq i$. Most often only the first canonical correlation, which is the largest value of $\cos\theta$ and therefore the minimum θ is considered.

Because the correlation (or angle between) z_a and z_b does not change with their scaling, we can constrain w_a and w_b such that z_a and z_b have unit variance:

$$\begin{aligned} z_a^T z_a &= w_a^T X^T X w_a = w_a^T S_{XX} w_a = 1 \\ z_b^T z_b &= w_b^T Y^T Y w_b = w_b^T S_{YY} w_b = 1 \end{aligned} \tag{4}$$

where S_{XX} and S_{YY} are the sample variance matrices for X and Y respectively. Since X and Y have zero mean, the covariance between z_a and z_b is

$$z_a^T z_b = w_a^T X^T Y w_b = w_a^T S_{XY} w_b \tag{5}$$

where S_{XY} is the sample covariance matrix between the variable column vectors of X and Y. Using (4) and (5), we can rewrite the problem of canonical correlation in (2) as:

$$\begin{aligned} \cos\theta &= \max_{z_a, z_b \in R^n} w_a^T S_{xy} w_b \\ ||z_a|| = \sqrt{w_a^T S_{XX} w_a} \quad ||z_b|| &= \sqrt{w_b^T S_{yy} w_b} \end{aligned} \tag{6}$$

Expressing the constraints instead as squares, we get $w_a^T S_{XX} w_a = 1$ and $w_b^T S_{yy} w_b = 1$. This optimization can be solved by using Lagrange multipliers to enforce the constraints, differentiating with respect to w_a and w_b, and solving the resulting generalized eigenvalue problem. It can also be solved using Singular Value Decomposition (SVD).

2.2 Limitations of CCA

Although CCA can be a useful tool for analyzing multivariate associations, it can suffer from several limitations when applied to typical neural-behavioral datasets.

First, CCA performs poorly in terms of stability and accuracy at low sample sizes [8, 10, 11]. This is exacerbated for neuroimaging studies because the ground-truth effects are often quite small, requiring even more samples/feature to produce stable results [10]. Stability can be defined as the ability to reliably estimate elements of the CCA solution across varying samples from the same population, and accuracy is the ability of CCA to discover the true properties and associations of the dataset. Helmer et. al. discuss these drawbacks in great detail [8], finding that the stability of CCA solutions requires more samples than are usually present in published neuroimaging studies. Considering two large example datasets, the HCP and UK Biobank datasets, they found that while CCA was able to produce stable results on UK Biobank data (about 20,000 subjects and 100 principal components after PCA), association strengths and weight vectors for HCP data (about 1200 subjects and 100 principal components after PCA) were unstable (Fig. 1). At low sample sizes, CCA tends to report association strengths that are too high as a result of overfitting the data, and in fact reported association strength in the analyzed studies could actually be predicted using only the number of samples per feature (used instead of samples to account for widely varying dimensionality of feature spaces). Most reliable studies reported correlations around or below 0.3; higher reported associations may indicate instability and overfitting [8]. This sample size dependence is one form of instability. Thus, CCA suffers from both instability and inaccuracy for low samples per feature.

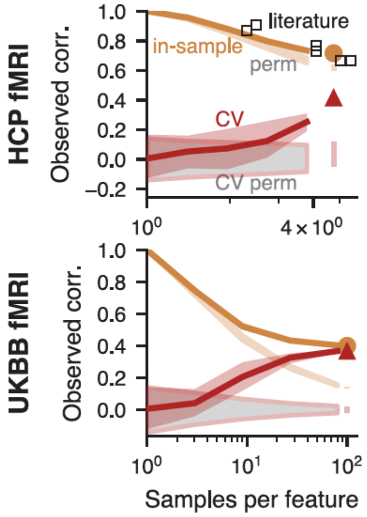


Figure 1: Observed correlation converges to stable values for UKBB dataset but not for HCP dataset. Reprinted from [8] with permission.

Second, CCA is unable to combine multiple datasets in the same analysis if their feature spaces differ. Because there exist many small-to-medium size neural-behavioral datasets measuring slightly different features, it would be advantageous to be able to combine studies together to effectively increase the number of samples per feature. Consider a case where we have datasets 1 and 2, each containing X and Y data, and while the X features are consistent between the data sets, the Y features differ. Thus we cannot simply stack X_1 on X_2 and Y_1 on Y_2, since the features do not match up.

There exists, therefore, a significant need to find improvements over classical CCA to produce more stable and accurate results for typical neural-behavioral data and to extend it to multi-dataset scenarios.

3 Combination Canonical Correlation Analysis (C3A)

3.1 Approach

To address the previously discussed shortcomings of CCA as applied to neural-behavioral data, we propose and evaluate a novel methodology for discovering multivariate relationships between neural and behavioral data from multiple populations or studies. The aim of Combination Canonical Correlation Analysis (C3A) is to combine multiple datasets with comparable neuroimaging data and distinct behavioral data to effectively increase the sample size of the analysis and mitigate CCA's issues of instability. C3A considers two datasets, each with comparable neuroimaging features and potentially nonoverlapping behavioral features, and simultaneously performs CCA on each dataset while enforcing that the weight vectors for the two neuroimaging feature sets are similar. This method relies on a crucial assumption that even though behavioral features may differ between datasets, related or common underlying modes of neural-behavioral association are present in the datasets and encoded by different behavioral features [9]. For example, conditions like depression may be measured using different behavioral scales. This means that even though behavioral features in the second dataset may be distinct from those in the first, the second dataset can still contribute valuable information to improve the stability and accuracy of neural-behavioral association analyses performed on the first dataset. C3A draws on similar ideas of dataset combination presented in [9, 12], but couples the analyses using a weight penalty term and avoids assigning a mathematical order of priority to the datasets. In this paper, we consider $n = 2$ datasets, although C3A is designed to be compatible with arbitrarily many datasets.

3.2 Mathematical Definition

Consider two datasets, each broken up into neural data X and behavioral data Y. The optimization equation that C3A solves is:

$$f = min_{w_{\vec{x}1},\vec{w}_{x2},\vec{w}_{y_a},\vec{w}_{y_b}} \Bigg(-\vec{w}_{x1}^T S_{xy_a} \vec{w}_{y_a} - \vec{w}_{x2}^T S_{xy_b} \vec{w}_{y_b} + \\ \sum_{i \in \{x_1, x_2, y_a, y_b\}} \frac{1}{2} \lambda_i \left(-\vec{w}_i^T S_{ii} \vec{w}_i - 1 \right) + \rho r \left(\vec{w}_{x1}, \vec{w}_{x2} \right) \Bigg) \quad (7)$$

where x_1 and x_2 refer to neuroimaging data in the two datasets, which share the same feature set. y_a is the behavioral feature set for dataset 1 and y_b is the behavioral feature set for dataset 2. Each $\vec{w}$ is a weight vector estimated by the method and S denotes a covariance matrix, e.g. S_{xy_a} is the covariance matrix between the stacked X neural data and the behavioral data y_a from the first dataset. The first two terms in this optimization thus represent the two different canonical correlation terms that we seek to maximize. The third term is a summation of Lagrange multiplier terms that enforce a scaling of weight vectors such that the first two terms represent correlations, not just covariances (e.g. enforce that the product of each weight vector times its corresponding dataset has unit variance). The fourth term enforces similarity of X weight vectors by imposing a penalty that scales with dissimilarity. Here, $r(\cdot)$ is a function of w_{x1} and w_{x2}. In our analyses, we use $r = \frac{1}{2}||w_{x_1} - w_{x_2}||^2$ and $\rho = 0.5$. Larger values of ρ enforce more similar weights.

It is assumed that PCA has already been performed on the datasets such that they have zero mean and the PCs scores' variances (which are the eigenvalues of S_{xx} and S_{yy}) decay with a power-law dependence.

The above optimization problem has been written as the minimization of negative canonical correlations to be compatible with the solver in `scipy.optimize.differential_evolution`. It could be written equivalently as the maximization of positive canonical correlations.

3.3 Iterative Solution

Note that the above optimization of (7) has dimension $4+p_x+p_x+p_{y_a}+p_{y_b}$, where the 4 comes from the four λ_i and p refers to the number of features in the corresponding data set. With some manipulation, we can reduce this to a 4-dimensional optimization over λ_i. The iterative solver at each step guesses values for λ_i, solves for w vectors using λs, then evaluates the loss function in (7) to inform the next iteration.

First, take the partial derivative of (7) with respect to each weight vector (each equation here has been multiplied by -1):

$$\frac{\partial f}{\partial w_{x_1}} = w_{y_a}^T S_{xy_a}^T + \lambda_{x_1} w_{x_1}^T S_{xx} - \rho \frac{\partial r}{\partial w_{x_1}} = 0 \tag{8}$$

$$\frac{\partial f}{\partial w_{x_2}} = w_{y_b}^T S_{xy_b}^T + \lambda_{x_2} w_{x_2}^T S_{xx} - \rho \frac{\partial r}{\partial w_{x_2}} = 0 \tag{9}$$

$$\frac{\partial f}{\partial w_{y_a}} = w_{x_1}^T S_{xy_a}^T + \lambda_{y_a} w_{y_a}^T S_{y_a y_a} = 0 \tag{10}$$

$$\frac{\partial f}{\partial w_{y_b}} = w_{x_2}^T S_{xy_b}^T + \lambda_{y_b} w_{y_b}^T S_{y_b y_b} = 0 \tag{11}$$

Substituting (10) into the transpose of (8) and simplifying, we get

$$S_{xy_a}\left[-\frac{1}{\lambda_{y_a}} S_{y_a y_a}^{-1} S_{xy_a}^T w_{x_1}\right] + \lambda_{x_1} S_{xx} w_{x_1} = \rho \frac{\partial r}{\partial w_{x_1}}^T \tag{12}$$

$$-\left(\frac{1}{\lambda_{y_a}} S_{xy_a} S_{y_a y_a}^{-1} S_{xy_a}^T - \lambda_{x_1} S_{xx}\right) w_{x_1} = \frac{\partial r}{\partial w_{x_1}}^T \tag{13}$$

Defining M_a as $\frac{1}{\lambda_{y_a}} S_{xy_a} S_{y_a y_a}^{-1} S_{xy_a}^T - \lambda_{x_1} S_{xx}$, (12) can be written as:

$$-M_a w_{x_1} = \frac{\partial r}{\partial w_{x_1}}^T \tag{14}$$

Analogously manipulating (9) and (11):

$$-M_b w_{x_2} = \frac{\partial r}{\partial w_{x_2}}^T \tag{15}$$

where $M_b = \frac{1}{\lambda_{y_b}} S_{xy_b} S_{y_b y_b}^{-1} S_{xy_b}^T - \lambda_{x_2} S_{xx}$.
From this point, the solution depends on the choice of the function $r(w_{x_1}, w_{x_2})$. Assume that $r = \frac{1}{2}||w_{x_1} - w_{x_2}||^2$. Then $\frac{\partial r}{\partial w_{x_1}} = w_{x_1}^T - w_{x_2}^T$ and $\frac{\partial r}{\partial w_{x_2}} = w_{x_2}^T - w_{x_1}^T$. Substituting these identities into (13) and (14), the following system of equations results:

$$\rho w_{x_2} = (M_a + \rho I) w_{x_1} \tag{16}$$

$$\rho w_{x_1} = (M_b + \rho I) w_{x_2} \tag{17}$$

This can be manipulated to form an eigenvalue problem which can be solved for w_{x_2}:

$$\rho w_{x_2} = (M_a + M_b)^{-1} M_a M_b w_{x_2} \tag{18}$$

Once w_{x_2} has been solved for, it can be substituted into (17) to solve for w_1.

This effectively reduces the problem such that, given covariance matrices S, which are determined at the start of the algorithm, and λ_1, λ_2, λ_3, λ_4, which are iteratively solved for in the differential evolution function, each w vector can be solved for. Within each iteration of the differential evolution solver, the w vectors are calculated using the λs as follows:

1. Solve (18) for w_{x_2}.
2. Using w_{x_2} and (17), solve for w_{x_1}.
3. Using w_{x_1} and (10), solve for w_{y_a}.
4. Using w_{x_2} and (11), solve for w_{y_b}.

Then the w vectors and λs are substituted into the loss function (7).

4 Method

In the following sections, an exploratory analysis of the accuracy of C3A compared to CCA is conducted using synthetic generated data over a wide range of data parameters. Metrics of X weight vector error and association strength are visualized over different parameter sets to provide insight into uses and limitations of the new method.

4.1 Generative Model

In order to test C3A and CCA over a robust set of parameters to inspect how each parameter impacts performance, we used generated synthetic datasets for which the ground truth is known. The GEMMR framework developed by Helmer et. al. [8] allows us to generate data with varying parameters such as dimensionality of datasets, number of samples per feature, association strength, similarity of X weight vectors (for C3A) and within-set variance spectra that are based on empirical trends observed in neural-behavioral data. These datasets can then be analyzed to inspect how parameter values impact stability and accuracy to show how C3A performs under different data conditions. For example, how does C3A compare to CCA under varying samples/feature, or low association strength? Relying on existing datasets such as HCP or UK Biobank constrains the parameter sets that can be analyzed by what is available in the data, but GEMMR can create datasets of arbitrary parameter combinations. Results of C3A and CCA can then easily be compared to the known ground truth to see how well each method discovers latent associations and weight vectors. A detailed explanation of how GEMMR generates data is available in the appendix of [8].

4.2 Running Analyses

Because C3A is solved iteratively rather than deterministically, the computational time can be expensive for high dimensional data and low true correlations. The analyses were therefore run on a high performance computing cluster and parameter combinations were broken up into approximately 150 jobs per method analyzed. Each job was run on a set of parameter combinations and took between several hours to one day to complete. For each possible combination of parameters, a dataset was generated and the performance of CCA and C3A were analyzed on that dataset. This process of generating a dataset and running CCA and C3A was repeated 5 times for each combination of parameters and results were averaged over all repetitions. Data from each job was then merged into a composite dataset spanning the entire parameter space.

4.3 Error Metrics

Although GEMMR outputs many error metrics after running each statistical analysis, the two primary metrics used in this report to quantify C3A performance are X weight errors and association strength. These two metrics are particularly useful because the primary results from CCA used to interpret associations are association strength and weight vectors. Association strength predicted by C3A and CCA can be compared easily to the ground truth value passed into the GEMMR model that is used to generate the datasets. X and Y weight vectors are predicted by C3A and CCA and are not parameters that are passed into GEMMR, so in order to evaluate predicted X weight vectors, we compare them to the predicted X and Y weight vector of a CCA analysis run on a dataset with effectively infinite sample size. Practically, this means that the sample covariance matrices used by the CCA analysis are actually the true population covariance matrices. Because CCA only deals with one dataset split into X and Y features and therefore predicts one X and Y weight vector, we can compare X and Y weight vectors for only the primary dataset considered by C3A to the CCA ground truth solution. Weight error is calculated as $1 - |\cos\theta|$, where θ is the angle between the true weight vector and the predicted weight vector. Generally, X weight error showed much larger variation with the parameters and so is considered in these results instead of Y weight error.

5 Standard C3A Results

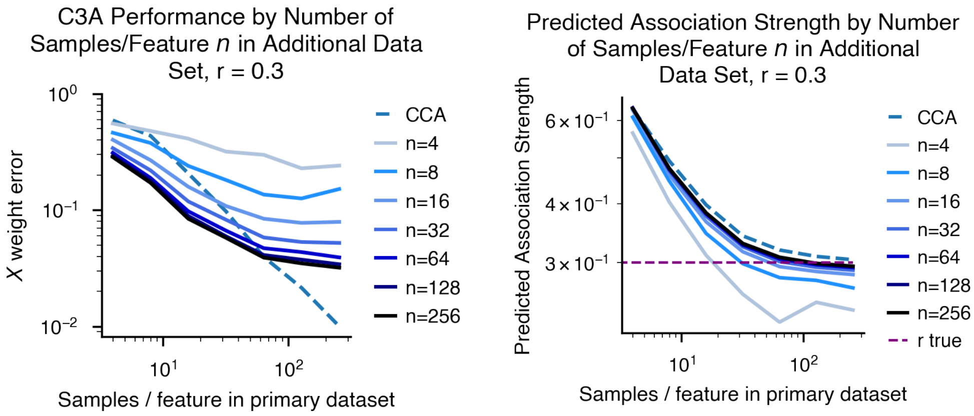
Results for the most important parameters of generated data are analyzed in the following sections. C3A and CCA performance depended most strongly on samples/feature, true association strength, and similarity of datasets (C3A only).

Sample Size Dependence CCA stability and accuracy exhibits a strong dependence on sample size, so we predicted C3A would also have a strong sample size dependence as an extension of CCA. Note that the more comprehensive samples per feature is used throughout this report rather than number of samples as a way of eliminating feature set size dependence. Samples per feature is defined as the number of individual samples (rows in the data sets) divided by the sum of the number of neural and behavioral features. The ability of C3A to accurately predict X weights and association strength strongly depended on both the number of samples per feature in the primary dataset and the number of samples per feature in the supplemental dataset.

At low sample sizes for the primary dataset, C3A generally outperforms CCA by incorporating data from the secondary dataset to increase the amount of information included in the analysis and effectively decrease the variance in data (Fig. 2). At larger sample sizes for the primary dataset, C3A performance plateaus and CCA outperforms C3A at predicting X weight because of the bias introduced by the second dataset (Fig. 2a). In this region, the secondary dataset functions as a distractor to the method because the two datasets are not perfectly aligned, and therefore information from the secondary dataset begins to detract from the analysis of the primary dataset. Relevant data from the secondary dataset proves useful at low sample sizes of the primary dataset, but when the primary dataset has enough samples to produce accurate results, the irrelevant data from the secondary dataset outweighs the benefit of adding the relevant data from the secondary dataset.

Considering the samples/feature of the secondary dataset, C3A performance increases with increasing samples up to a point, then plateaus after about n = 128 samples/feature (Fig. 2). For a small secondary dataset of only $n = 4 - 8$ samples per feature, C3A provides virtually no gains, but for larger secondary datasets C3A improves performance in the region of low samples/feature of the primary dataset. Practically, this means that incorporating a secondary dataset of medium to large size will generally improve prediction of X weight vectors if the number of samples/feature for the primary dataset is small.

Association Strength Dependence Both X weight error and predicted association strength depend on true association strength between neural and behavioral features. Accuracy of predicted association strength depends much more heavily on true correlation



(a)

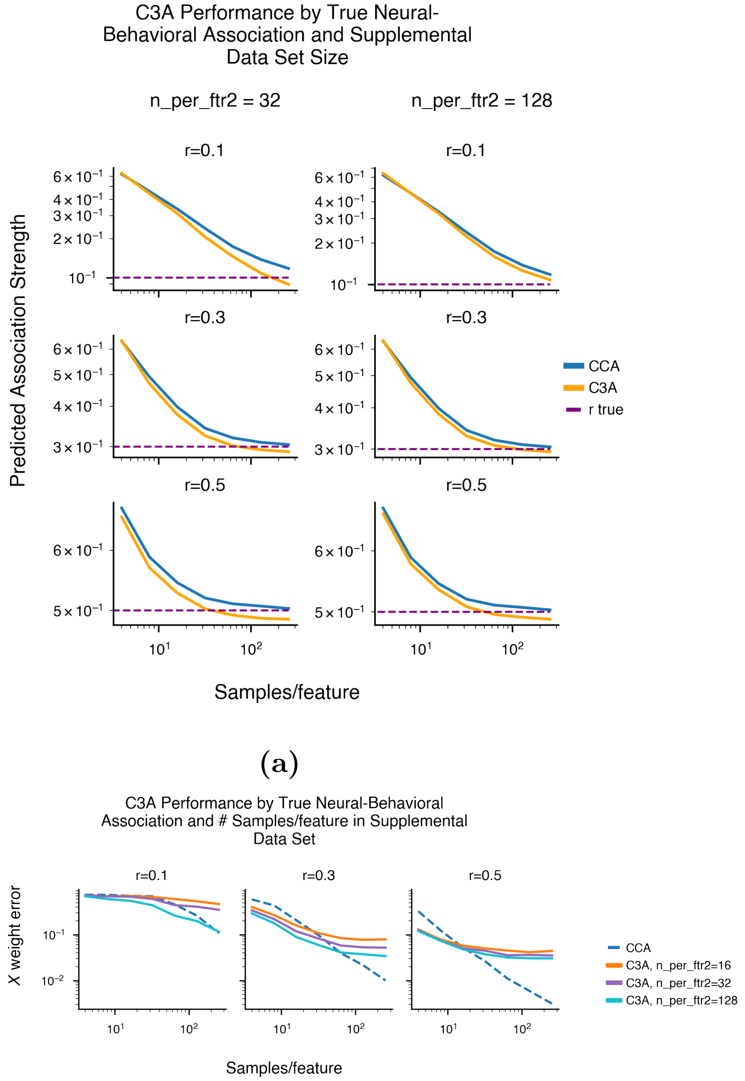
(b)

Figure 2: Error in predicted X weight vectors and predicted association strength by samples/feature of secondary dataset. Error decreases with increasing samples/feature in both the primary and secondary datasets. C3A improvements begin to level off around 64 samples/feature for the secondary dataset. X weight error plateaus for C3A for large samples/feature in the primary dataset.

than on samples/feature of the secondary dataset (Fig. 3a). As discussed above, predicted association strength has a strong dependence on the number of samples/feature in the primary dataset. Thus, although association strength predicted by C3A decreases more quickly towards the true value for all true association strengths, at high sample sizes for the primary dataset C3A underestimates true association strength. As correlation increases, this behavior of underestimating the true association at large samples/feature for the primary dataset becomes more pronounced and the regime for which C3A is more accurate than CCA is smaller. In the region of low samples/feature for the primary dataset, however, C3A converges much more quickly towards the true association strength. Thus, C3A shows promise in predicting association strengths if the true association strength is small, as it often is in the literature, and for low samples/feature in the primary dataset.

Similarly, prediction of X weight vectors by C3A is strongly dependent on true association strength (Fig. 3b). For lower true correlations, C3A predicts X weight vectors more accurately than CCA for a larger range of samples/feature for the primary data set and is therefore more useful for smaller primary datasets. Both methods are less accurate for lower true association strengths. Again, the true association strength parameter influences results more than the number of samples/feature in the secondary dataset.

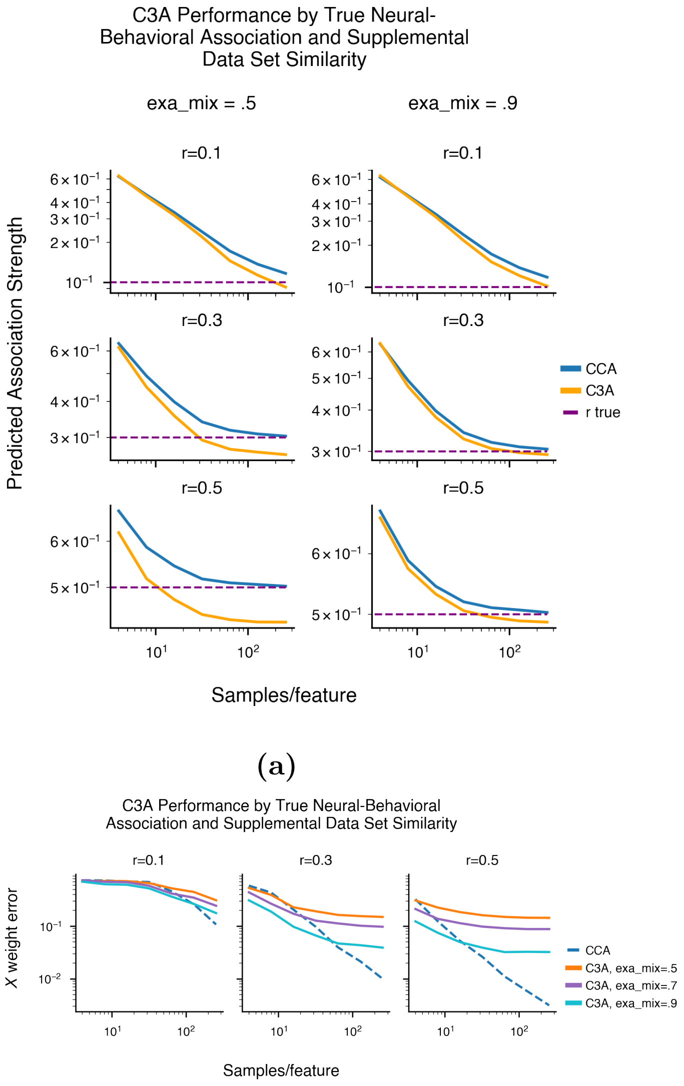
Dependence on Dataset Similarity As intuition would predict, the similarity between the two datasets used by C3A significantly impacts results. This measure is quantified by the parameter `exa_mix`, which is the similarity between the true X weight vectors for the two datasets. Although less intuitive than the previous two parameters discussed,



(a)

(b)

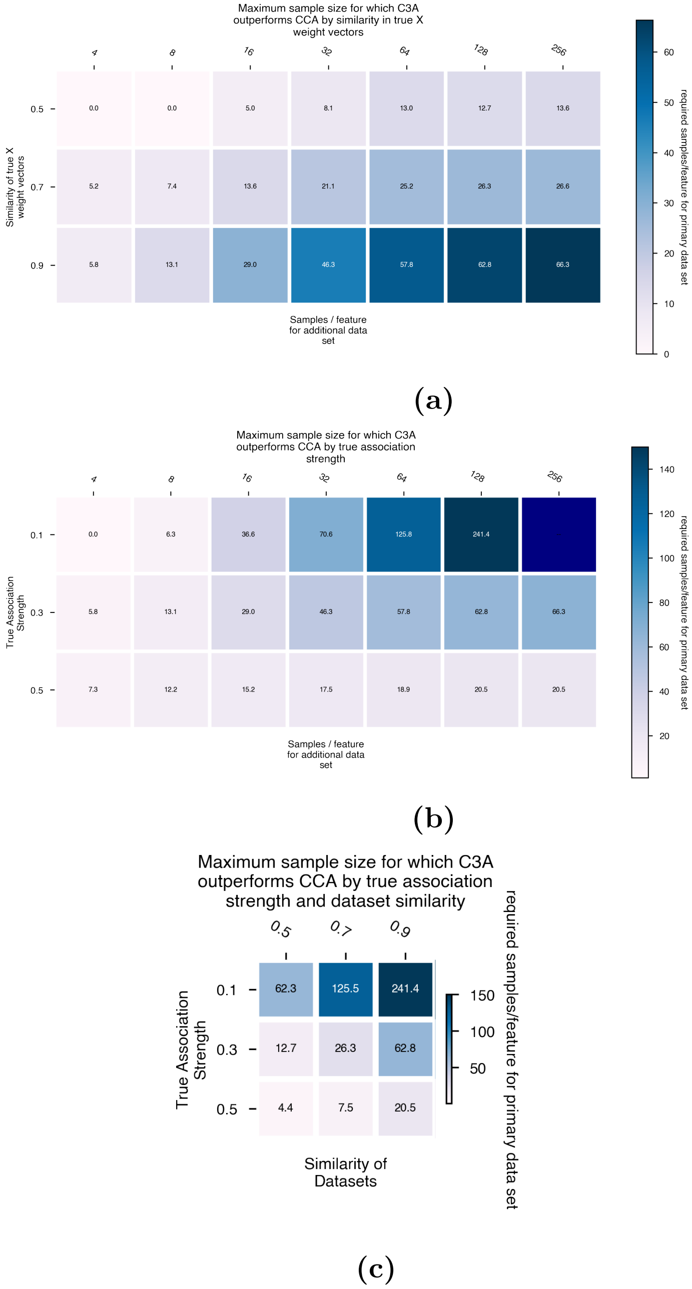
Figure 3: Predicted association strength and X weight error for varying true association strength and supplemental dataset size. C3A performance in accurately predicting association strength and X weight vectors depends heavily on true association strength. C3A provides gains in predicting X weight vectors for $r = .3, .5$ for low samples/feature in the primary dataset but underpredicts association strength for $r = .5$. CCA performs well under high correlations, making C3A less useful.



(a)

(b)

Figure 4: Predicted association strength and X weight error by true association and dataset similarity. For lower similarity between datasets 1 & 2, C3A underpredicts true association strengths more dramatically. C3A provides gains in predicting X weight vectors in regions of low samples/feature for dataset 1 for $exa\_mix = .7, .9$



(a)

(b)

(c)

Figure 5: C3A outperforms CCA for regimes of varying size. For each combination of parameters, the intersection point of the X weight error curves for CCA and C3A was linearly interpolated over the number of samples/feature in the primary dataset (see Fig. 2 for an example of the curves examined). The maximum number of samples/feature for the primary dataset for which C3A had a lower X weight error than CCA is plotted in the heatmaps. The first plot uses r = .3, the second uses exa_mix=.9, the third uses n = 128 samples/feature for the secondary dataset.

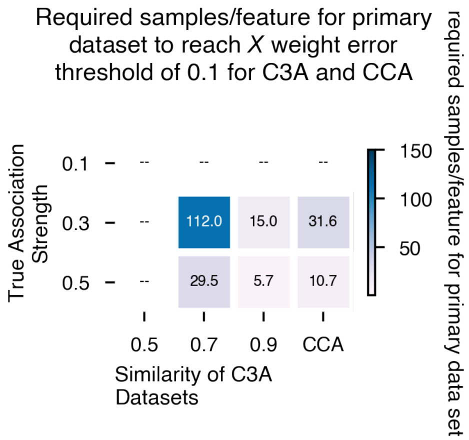


Figure 6: Required samples/feature to reach error threshold depends on algorithm parameters. Because X weight error decreases monotonically with samples/feature of the primary dataset, it is appropriate to find the minimum value at which X weight error drops below an acceptable threshold. This minimum value was calculated by linearly interpolating X weight error over samples/feature. Missing values mean the error threshold was never reached for that parameter combination. The plot uses $n = 128$ samples/feature for the secondary dataset.

similarity of X weight vectors helps to describe the general applicability of both the neural and behavioral data of the second dataset to the first dataset. Lower values of exa_mix amplify the behavior that C3A exhibits of underpredicting association strength; larger values of exa_mix produce better predictions that are more comparable to C3A (Fig. 4a). C3A with larger values of exa_mix predicts X weight vectors more accurately, and the regime for which C3A performs better than CCA is larger (Fig. 4b). C3A still showed improvements over CCA for exa_mix= .7, but the gains were much smaller than exa_mix = .9 (Fig. 4b).

Dependence on Other Parameters Less important parameters for the performance of CCA and C3A were dimensionality of feature spaces and variance spectra of PCA-reduced datasets. Both CCA and C3A were less accurate for larger dimensionality of feature sets, holding samples/feature constant. The relative improvement of C3A over CCA was mostly consistent over different dimensionalities. X weight error was virtually identical for different decay functions of PC components or variance spectra.

Intersections of Error Curves An exploratory analysis of the intersections of error curves was conducted by plotting the maximum samples/feature of the primary dataset for which C3A outperforms CCA as a function of two parameters. This maximum samples/feature was calculated by linearly interpolating the intersection point of CCA and C3A X weight error curves and is plotted in Fig. 5; missing values on the heatmap indicate that C3A always outperformed CCA for that parameter combination. Darkness of colors in the heatmap corresponds to this maximum sample size, thus darker colors indicate situations in which C3A is more useful. As expected based on individual parameter analyses, the combination of lower true association strength and increasing number of samples/feature for the secondary dataset produces a larger regime in which C3A outperforms CCA, with association strength having a much larger effect than samples/feature (Fig. 5a). The same is true for high `exa_mix` and high samples/feature in the secondary dataset (Fig. 5b). For association strengths of 0.1, the regime of C3A outperforming CCA is just as large even for `exa_mix`$= .5$ as `exa_mix`$= .9$ and $r = 0.3$ (Fig. 5c). Association strength has a larger impact on the size of this regime than does `exa_mix`.

Required Samples to Reach Error Threshold Fig. 6 plots the required number of samples/feature for the primary dataset to attain an X weight error less than 0.1. Because X weight error is monotonically decreasing, unlike predicted association strength, it is appropriate to define a minimum number of samples required. For true associations strengths of 0.1 or `exa_mix` of 0.5, the error threshold is never reached. C3A reaches this error threshold faster than CCA for `exa_mix` of 0.9 only.

6 Adaptation: Weighted C3A

6.1 Approach

The results for standard C3A show that it can provide gains in accuracy of predicting weights and associations, especially for small samples/feature in the primary dataset. After a certain number of samples/feature in the primary dataset, however, the secondary dataset incorporated by C3A functions as a distraction rather than a helpful addition. We then considered potential adaptations of our optimization function that could either improve accuracy of C3A in the region where it performs better than CCA, or extend that region to a larger number of samples/feature so that C3A is applicable in a wider variety of situations. Many adaptations of CCA exist that seek to optimize different functions of correlation and covariance [10, 12]. In order to maintain the relationship discussed in section 3.3 that allows reduction of a high-dimensional optimization down to 4 dimensions, we kept the adaptation of C3A mathematically simple so that we could use the same solution framework. The error for weights and associations was strongly related to number of samples/feature, so an adaptation is proposed that weights the canonical correlation terms in the optimization function based on the number of samples/feature in each dataset. The new optimization problem can be written as:

$$f = min_{\vec{w_{x1}},\vec{w}_{x2},\vec{w}_{ya},\vec{w}_{y_b}} \Bigg(-c_1 \vec{w}_{x1}^T S_{xy_a} \vec{w}_{y_a} - c_2 \vec{w}_{x2}^T S_{xy_b} \vec{w}_{y_b} + \\ \sum_{i \in \{x_1, x_2, y_a, y_b\}} \frac{1}{2} \lambda_i \left(-\vec{w}_i^T S_{ii} \vec{w}_i - 1 \right) + \rho r \left(\vec{w}_{x1}, \vec{w}_{x2} \right) \Bigg) \quad (19)$$

Where c_1 and c_2 are constants calculated from the number of samples/feature in each dataset. A number of different weighting functions were tested that apply a function to each dataset's samples/feature before scaling the constants around 1 in order to maintain similar ratios between the correlation terms and the weight similarity penalty term. Then an optional postprocessing function, either squaring or taking the square root, was applied to c_1 and c_2 before beginning to solve the optimization problem.

$$c_1 = \frac{2f(n_1)}{f(n_1) + f(n_2)} \qquad c_2 = \frac{2f(n_2)}{f(n_1) + f(n_2)} \tag{20}$$

where n_1 and n_2 are the number of samples/feature in datasets 1 and 2, respectively. The relevant M matrices can be redefined thus as:

$$\begin{aligned} M_a &= \frac{1}{\lambda_{y_a}} c_1^2 S_{xy_a} S_{y_a y_a}^{-1} S_{xy_a}^T - \lambda_{x_1} S_{xx} \\ M_b &= \frac{1}{\lambda_{y_b}} c_2^2 S_{xy_b} S_{y_b y_b}^{-1} S_{xy_b}^T - \lambda_{x_2} S_{xx} \end{aligned} \tag{21}$$

6.2 Results and Summary of Recommendations

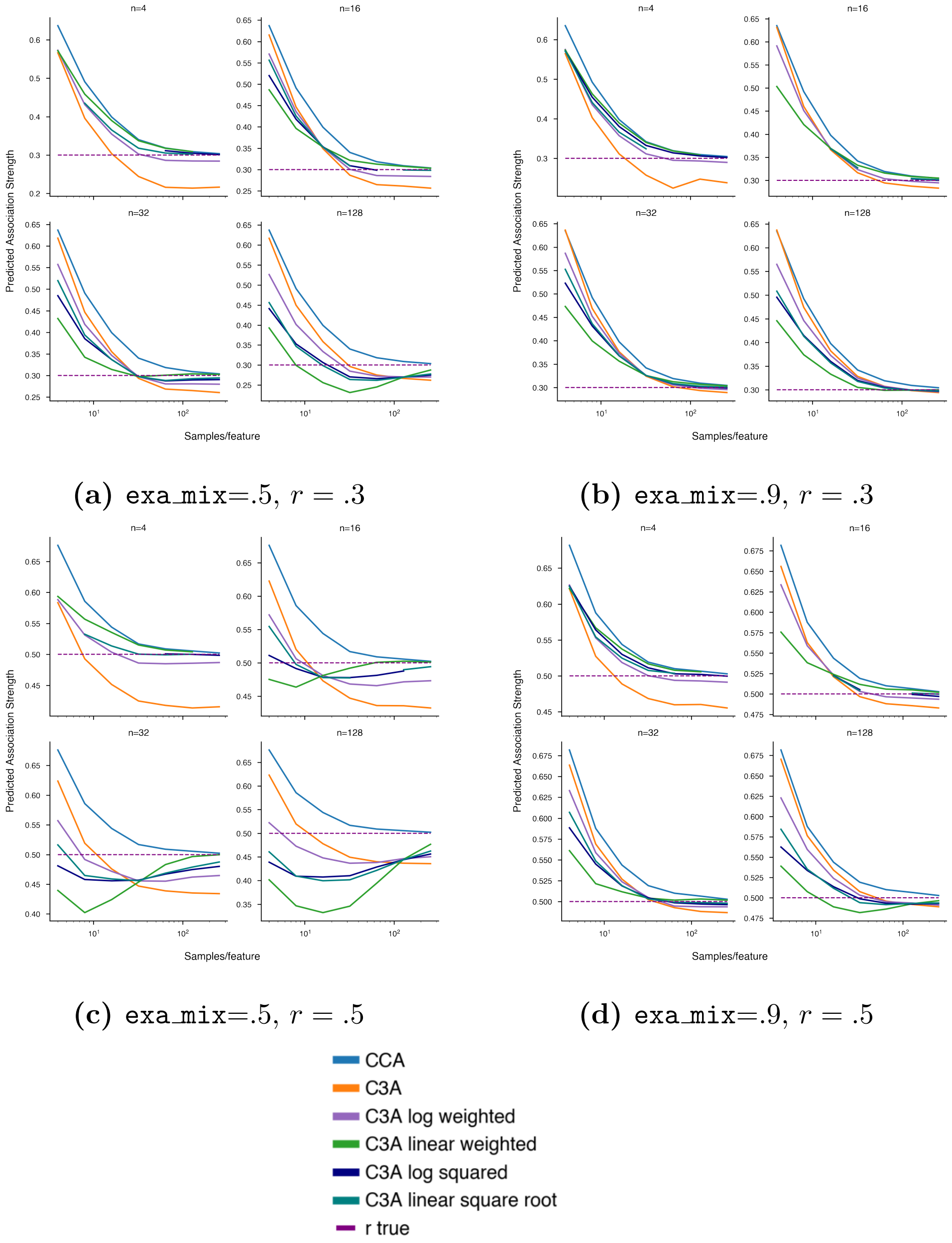
Four different weighting algorithms are considered to calculate c_1, c_2 as in (20)

1. linear weighting: $f(n) = n$
2. logarithmic weighting: $f(n) = log(n)$
3. linear-square root: $f(n) = n$, then take the square root of the resulting c_1, c_2
4. log-squared: $f(n) = log(n)$, then square the resulting c_1, c_2

Linear-square root and log-squared were tested because it was observed that linear weighting produced results dramatically different from standard C3A and log weighting produced results quite similar to standard C3A (Fig. 7, 8), and so the weighting constants were pushed closer together and farther apart, respectively. We predicted that some kind of logarithmic weighting would produce the best results because visualization of standard C3A results showed that the relationship between error and samples/feature was roughly logarithmic.

Weighted C3A produced significantly different results than standard C3A for both predicted association strength and X weight error. Weighting C3A did not significantly change the size of the region of samples/feature in the primary dataset for which C3A outperforms CCA at predicting X weight vectors (Fig. 8) but it slightly decreased the required samples/feature in the primary dataset to reach an error threshold of 0.1 if the datasets had high similarity (Fig. 9). Significant gains were observed in reducing the underprediction of true association strength discussed in section 5 (Fig. 7). Modest gains were observed in prediction of X weight vectors for some parameter combinations for the region of very low samples/feature in the primary dataset where C3A is useful (Fig. 8). Weighting generally improved prediction of X weight vectors for large samples/feature for the primary dataset, but this is not a region in which C3A would be useful.

Weighted C3A Predicted Association Strength by True Association Strength and Dataset Similarity



(a) `exa_mix`=.5, $r = .3$

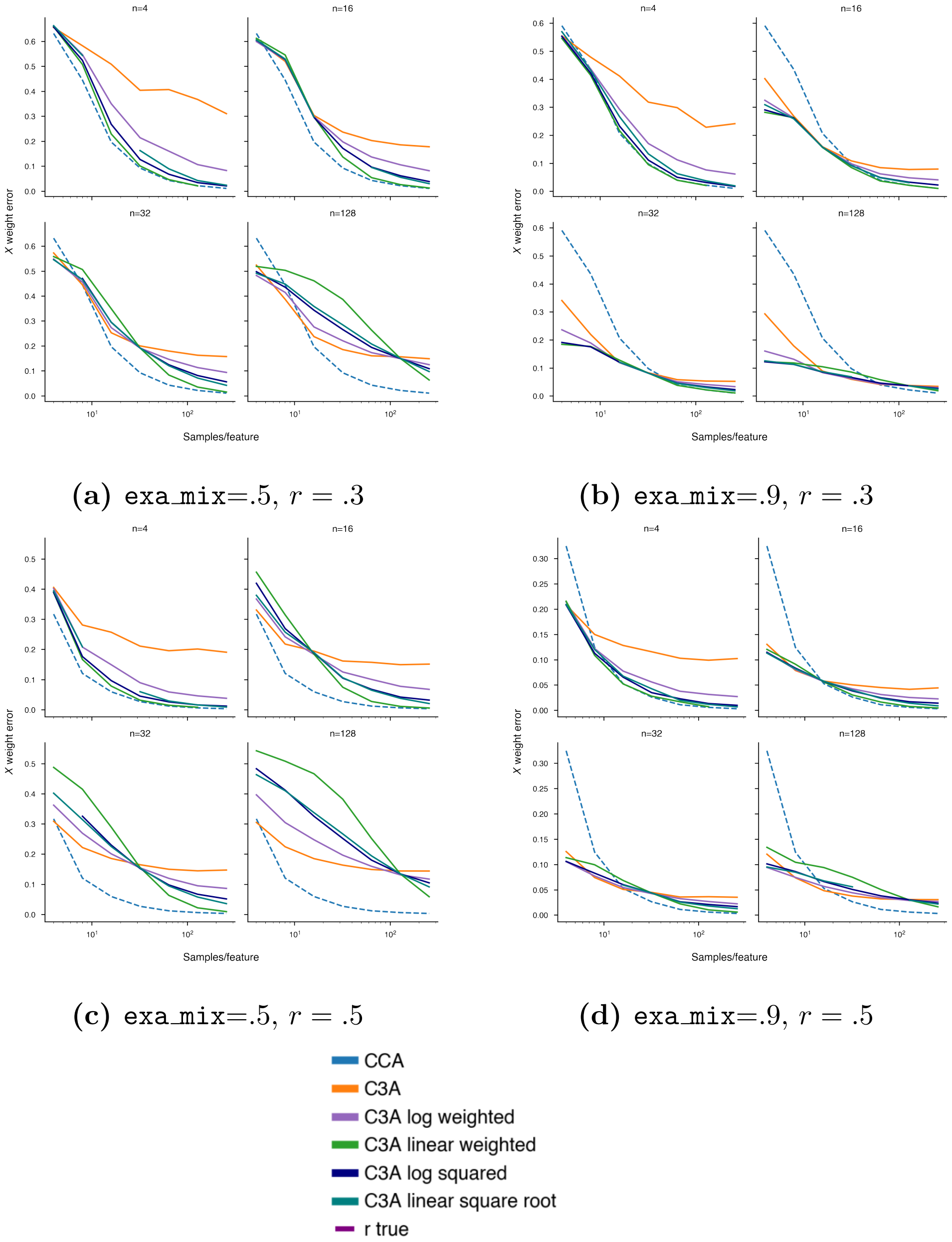
(b) `exa_mix`=.9, $r = .3$

(c) `exa_mix`=.5, $r = .5$

(d) `exa_mix`=.9, $r = .5$

Figure 7: Predicted association strength for weighted C3A variants. Weighting C3A generally improves prediction of association strength in low samples/feature regimes. Figures are plotted with varying number of samples/feature of the secondary dataset n from 4 - 128.

Weighted C3A X Weight Error by True Association Strength and Dataset Similarity



(a) exa_mix=.5, $r = .3$

(b) exa_mix=.9, $r = .3$

(c) exa_mix=.5, $r = .5$

(d) exa_mix=.9, $r = .5$

Figure 8: X weight error for weighted C3A variants. Weighting C3A can improve prediction of X weight vectors for sufficiently high levels of similarity of the two datasets and sufficiently low true association values. The region in which weighting Figures are plotted with varying number of samples/feature of the secondary dataset n from 4 - 128.

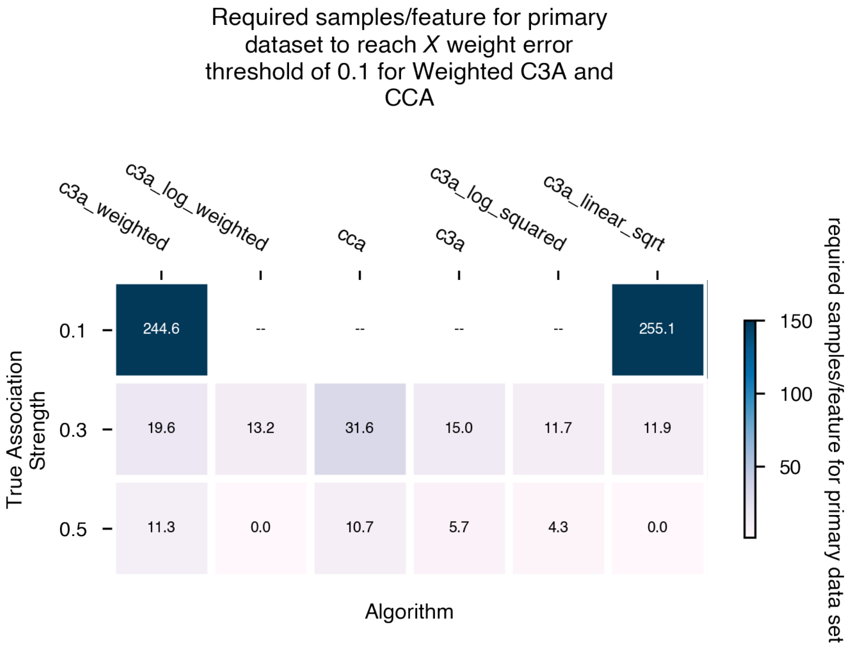


Figure 9: Weighting C3A slightly decreases the minimum samples/feature required to reach error threshold. As in figure 6, X weight error was linearly interpolated over samples/feature for the primary dataset to find the minimum number of samples/feature required to reach an error threshold of 0.1. Missing values indicate the error threshold was never reached. For this figure, `exa_mix`$= .9$.

Several parameter combinations of true correlation, dataset similarity, and supplemental dataset samples/feature are considered below to make recommendations of algorithm consideration for each use case depending on which metric is of greatest importance. Recommendations are based on a small number of samples/feature for the primary dataset, as this is where CCA performs poorly. Trying out different methods is also advised because the ranking of algorithm performance depends on which metric is being considered. A slightly simplified set of parameters is visualized here, using $r = .3, .5$, `exa_mix` $= .5, .9$, and number of samples/feature in the secondary dataset $n = 4, 16, 32, 128$.

Low true correlation, high dataset similarity (Fig. 7b, 8b) This is the ideal case for C3A to be used instead of CCA, and weighting C3A improves predicted association strength accuracy for all sample size regimes. Weighting decreases X weight error for low samples/feature for the primary dataset and does not have a significant impact for medium sized primary datasets. Linear weighting performs best at predicting association strength, while log-squared weighting performed best at predicting X weight vectors.

Low true correlation, low dataset similarity (Fig. 7a, 8a) Weighting improves C3A accuracy of predicted association strength for all but very large samples/feature in the secondary dataset, in which case underpredicting association strength happens more dramatically and for a lower number of samples/feature in the primary dataset. X weight error is reduced by weighting C3A for only very small primary datasets and medium to large secondary datasets. Linear weighting performs best at predicting association strength except for very large secondary datasets, in which case log-squared or linear-square root should be used. If samples/feature for the primary dataset is extremely small (< 5), log weighting is best for predicting X weight error, else CCA is best.

High true correlation, high dataset similarity (Fig. 7d, 8d) Weighting improves C3A accuracy of predicted association strength consistently, with linear weighting performing best at predicting association strength except for very large secondary datasets, in which case log-squared or linear-square root should be used. Weighting C3A has little effect on X weight error for all but very small (< 5) samples/feature for the primary dataset, so standard C3A should be used in the region determined by figure 5.

High true correlation, low dataset similarity (Fig. 7c, 8c) This is the regime under which standard C3A performs the worst with respect to CCA. Here, weighted C3A dramatically underpredicts true association strength, especially for large secondary dataset sizes. For predicting association strength at small sample sizes (< 10) standard C3A can be used, but for all other cases and for predicting X weight error, standard CCA performs best.

7 Limitations and Future Directions

Given the scale of this project and the computational time needed to produce results, it was necessary to make a number of assumptions. First, we assumed that the data was normally distributed, which is not always true in practice. Datasets 1 and 2 were assumed to all have the same number of features for neural (X) and behavioral (Y) data and the same $X - Y_1$ and $X - Y_2$ association. It is assumed that PCA has been applied to the data, which in practice introduces additional uncertainty based on the choice of number of PCs. This uncertainty can be reduced by using data-driven approaches to appropriately choose the number of PCs based on data characteristics like between-set correlation [**?**].

Because this analysis was performed on synthetic data from a generative model, it is strongly recommended to test C3A on empirical data. CCA and PLS have been shown to behave similarly in empirical data and generated data [8], but additional assumptions that C3A introduces must be investigated thoroughly by comparing empirical and generative results. The most crucial assumption is that there exist underlying modes of neural-behavioral association between the two datasets, as this is what allows the addition of a secondary dataset to improve results on the primary dataset. Use of large and robust secondary datasets like the UK Biobank and the HCP improve the likelihood of the existence of these common modes because their scope helps to capture more clinical variability [9]. Nevertheless, this assumption should be rigorously investigated on empirical data.

This project focused on generating visual analyses of the accuracy C3A and CCA in predicting true associations and weight vectors for the data. Robust statistical analyses to find statistical power of the methods and confidence intervals and p-values for the numerical results are recommended to be performed in future investigations to validate visual results. Because of the computational expense required to permute data and perform C3A on it hundreds of times, this aspect of analysis was not feasible for the time frame of this project.

Because CCA – and by extension C3A – is restricted to discovering linear relationships within the data, it is possible there are nonlinear relationships that better describe associations between neural and behavioral data. Nonlinear methods, however, are more susceptible to overfitting, and so the low sample size region observed in neuroscience literature would likely produce instability and inaccuracy in results for nonlinear methods. Within the class of linear methods, there exist several related methods that propose to reduce shortcomings of classical CCA. Regularized or sparse CCA uses a penalty to encourage weight vectors to shrink to smaller values, and has the potential to reduce required sample size in some situations [8]. Applying regularization to C3A could further reduce sample sizes beyond what has been shown in the results above. Other functions of correlation such as squaring could be applied to the canonical correlation terms in equation (7) as proposed in [12]; this would require adaptation of the C3A solver.

8 Conclusion

Canonical correlation analysis has been shown to perform poorly at discovering multivariate associations between datasets at low sample sizes. This problem is especially prominent in analyses of neural-behavioral associations, which usually contain less than 10 samples per feature [8]. We proposed a novel methodology for improving CCA for low sample sizes by incorporating supplemental data to effectively increase the sample size. Given common underlying latent modes of association between datasets, C3A shows promise at improving accuracy of predicted association strength and weight vectors, especially for low samples/feature of the primary dataset. Weighting the C3A optimization by samples/feature of the dataset shows additional gains over C3A in several situations. Accuracy of C3A depends heavily on parameters of the data, namely samples/feature, true association strength, and dataset similarity; C3A exhibits the most prominent gains over CCA when true correlation is low, dataset similarity is relatively high, and the secondary dataset has a substantial number of samples/feature. Characteristics of the datasets should be considered when deciding whether C3A should be used instead of classical CCA. This analysis provides an initial justification for the value of C3A as demonstrated on generated synthetic data through visualization of accuracy metrics. C3A holds promise for improving multivariate analyses of neural-behavioral association by introducing the possibility of combining multiple datasets with non-overlapping behavioral features. This presents great potential for utilizing the multitude of small to medium sized datasets in neuroscience in combination with large, robust datasets like the HCP and the UK Biobank.

References

[1] Gillihan SJ, Parens E. Should we expect ”neural signatures” for DSM diagnoses? *The Journal of Clinical Psychiatry.* **72**, 1383–1389 (2011). doi: 10.4088/JCP.10r06332gre.

[2] Phillips KA, First MB, Pincus HA. Advancing DSM: Dilemmas in Psychiatric Diagnosis. *American Psychiatric Pub* **2008**

[3] 1.Ji, J. L. *et al.* Mapping brain-behavior space relationships along the psychosis spectrum. *eLife* **10**, e66968.

[4] 1.Insel, T. R. & Cuthbert, B. N. Brain disorders? Precisely. *Science* **348**, 499–500 (2015).

[5] Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28**, 321-377 (1936). URL https://www.jstor.org/stable/2333955.

[6] Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: An overview. *NeuroImage* **80**, 62-79 (2013).
URL http://www.sciencedirect.com/science/article/pii/S1053811913005351.

[7] Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience* **19**, 1523-1536 (2016).
URL https://www.nature.com/articles/nn.4393.

[8] Helmer, M. *et al.* On stability of Canonical Correlation Analysis and Partial Least Squares with application to brain-behavior associations. 2020.08.25.265546 (2021) doi:10.1101/2020.08.25.265546.

[9] He, T. *et al.* Meta-matching: a simple framework to translate phenotypic predictive models from big to small data. 2020.08.10.245373 (2020) doi:10.1101/2020.08.10.245373.

[10] Wang, H.-T. *et al.* Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. *NeuroImage* **216**, 116745 (2020).

[11] Thorndike, R. M. & Weiss, D. J. A study of the stability of canonical correlations and canonical components. *Educational and Psychological Measurement* **33**, 123-134 (1973).

[12] Tenenhaus, M., Tenenhaus, A. & Groenen, P. J. F. Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods. *Psychometrika* **82**, 737–777 (2017).

[13] Uurtio, V. *et al.* A tutorial on canonical correlation methods. *ACM Computing Surveys (CSUR)*, **50**(6), 1-33 (2017).

[14] Song, Y. *et al.* Canonical correlation analysis of high-dimensional data with very small sample support. *Signal Processing* **128**, 449–458 (2016).