Open access • Posted Content • DOI:10.1101/2020.08.25.265546

On stability of Canonical Correlation Analysis and Partial Least Squares with application to brain-behavior associations — Source link

Markus Helmer, Shaun Warrington, Ali-Reza Mohammadi-Nejad, Ali-Reza Mohammadi-Nejad ...+8 more authors

Institutions: Yale University, University of Nottingham, National Institute for Health Research, John Radcliffe Hospital

Published on: 25 Aug 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Partial least squares regression, Overfitting, Canonical correlation and Principal component analysis

On stability of Canonical Correlation Analysis and Partial Least Squares with application to brain-behavior associations

Markus Helmer[1], Shaun Warrington[2], Ali-Reza Mohammadi-Nejad[2,3], Jie Lisa Ji[1,4], Amber Howell[1,4], Benjamin Rosand[5], Alan Anticevic[1,4,6] Stamatios N. Sotiropoulos[2,3,7,*] John D. Murray[1,4,5,*]

1 Department of Psychiatry, Yale School of of Medicine, New Haven, CT 06511

2 Sir Peter Mansfield Imaging Centre, School of Medicine, University of Nottingham, Nottingham, NG7 2UH, United Kingdom

3 National Institute for Health Research (NIHR) Nottingham Biomedical Research Ctr, Queens Medical Ctr, Nottingham, United Kingdom

4 Interdepartmental Neuroscience Program, Yale University School of Medicine, New Haven, CT 06511, USA

5 Department of Physics, Yale University, New Haven, CT 06511, USA

6 Department of Psychology, Yale University, New Haven, CT 06511, USA

7 FMRIB, Wellcome Centre for Integrative Neuroimaging, Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, University of Oxford, Oxford, OX3 9DU, United Kingdom

*** john.murray@yale.edu (JDM), stamatios.sotiropoulos@nottingham.ac.uk (SNS)**

Abstract

Associations between datasets, each comprising many features, can be discovered through multivariate methods like Canonical Correlation Analysis (CCA) or Partial Least Squares (PLS). Application of CCA/PLS to high-dimensional datasets raises critical questions about reliability and interpretability. To study this, we developed a generative modeling framework to simulate synthetic datasets, parameterized by dimensionality, variance structure, and association strength. We found that CCA/PLS associations could be highly inaccurate when the number of samples per feature is relatively small. For PLS, profiles of feature weights exhibit detrimental bias toward leading principal component axes. We confirmed these trends in state-of-the-art neuroimaging datasets, Human Connectome Project (n≈1000) and UK Biobank (n=20000), finding that only the latter comprised sufficient samples for stable estimates. Analysis of the neuroimaging literature using CCA to map brain-behavior relationships revealed also that the commonly employed sample sizes yield unstable CCA solutions. Finally, we provide a calculator of dataset properties required for CCA/PLS stability. Collectively, we characterize how to limit detrimental effects of overfitting on CCA/PLS stability, and provide recommendations for future studies.

Introduction

Discovery of associations between datasets is a topic of growing importance across scientific disciplines in analysis of data comprising a large number of samples across high-dimensional sets of features. For instance, large initiatives in human neuroimaging collect, across thousands of subjects, rich multivariate neural measures as one dataset and psychometric and demographic measures as another linked dataset [1, 2]. A major goal is to determine, in a data-driven way, the dominant latent patterns of association linking individual variation in behavioral features to variation in neural features [3, 4].

A widely employed approach to map such multivariate associations is to define linearly weighted composites of features in both datasets (e.g., neural and psychometric) and to choose the sets of weights—which correspond to axes of variation—to maximize the association strength (Fig. 1A). The resulting profiles of weights for each dataset can be examined for how the features form the association. If the association strength is measured by the correlation coefficient, the method is called *canonical correlation analysis* (CCA) [5], whereas if covariance is used the method is called *partial least squares* (PLS, or PLS correlation, see Discussion) [6, 7]. CCA and PLS are commonly employed across scientific fields, including genomics [8], and neuroimaging [3, 9].

Although the utility of CCA and PLS is well established, a number of open challenges exist regarding their stability in characteristic regimes of dataset properties. Stability implies that elements of CCA/PLS solutions, such as association strength and weight profiles, are reliably estimated across different independent sample sets from the same population, despite inherent variability in the data. Instability or overfitting can occur if an insufficient amount of data is available to properly constrain the model. Manifestations of instability and overfitting in CCA/PLS include inflated association strengths [10–12], cross-validated

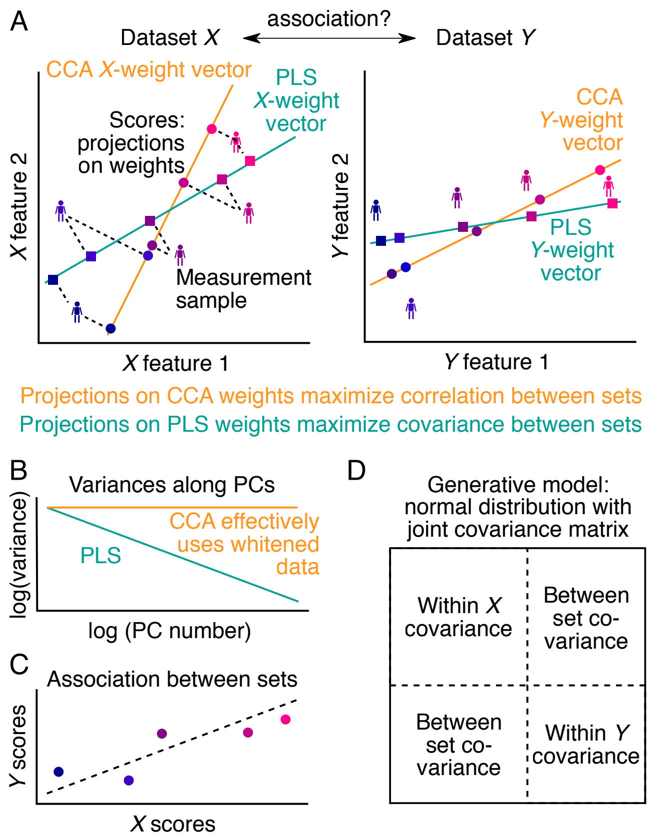

Figure 1. Overview of CCA, PLS and the generative model used to investigate their properties. A) Two multivariate datasets, X and Y, are projected separately onto respective weight vectors, resulting in univariate scores for each dataset. The weight vectors are chosen such that the correlation (for CCA) or covariance (for PLS) between X and Y scores is maximized. **B)** In the principal component coordinate system, the variance structure within each dataset can be summarized by its principal component spectrum. For simplicity, we assume that these spectra can be modeled as power-laws. CCA, uncovering correlations, disregards the variance structure and can be seen as effectively using whitened data (cf. Methods). **C)** The association between sets is encoded in the association strength of X and Y scores. **D)** Datasets X and Y are jointly modeled as a multivariate normal distribution. The within-set variance structure (**B**) corresponds to the blocks on the diagonal, and the associations between datasets (**C**) are encoded in the off-diagonal blocks.

association strengths that are markedly lower than in-sample estimates [13], or feature profiles that vary from study to study [10, 13–16]. Stability of models is essential for their replicability, generalizability, and interpretability. Therefore, it is important to assess how stability of CCA/PLS solutions depends on dataset properties.

Instability of CCA/PLS solutions is in principle a known issue [4, 14]. Prior studies using a small number of specific datasets or Monte-Carlo simulations have suggested to use between 10 and 70 samples per feature in order to obtain stable models [11, 15, 17]. However, it remains unclear how the various elements of CCA/PLS solutions (including association strengths, weights, and statistical power) differentially depend on dataset properties and sampling error, nor how CCA and PLS as distinct methods may exhibit differential robustness across data regimes. To our knowledge, no framework exists to systematically quantify errors in CCA/PLS results, depending on the numbers of samples and features, the assumed latent (between-set) correlation and the variance structure in the data, for both CCA and PLS.

To investigate these issues, we developed a generative statistical model to simulate synthetic datasets with known latent axes of association. Sampling from the generative model allowed quantification of deviations between estimated and true CCA or PLS solutions. We found that stability of CCA/PLS solutions requires more samples than are commonly used in published neuroimaging studies. With too few samples, estimated association strengths were too high, and estimated weights could be unreliable for interpretation. CCA and PLS differed in their dependences and robustness, in part due to PLS exhibiting a detrimental bias of weights toward principal axes. We analyzed two large state-of-the-art neuroimaging-psychometric datasets, the Human Connectome Project [1] and the UK Biobank [2], which followed similar trends as our model. These model and empirical findings, in conjunction with a meta-analysis of estimated stability in the brain-behavior CCA literature, suggest that typical CCA/PLS studies in neuroimaging are prone to instability. Finally, we applied the generative model to develop algorithms and a software package for calculation of estimation errors and required sample sizes for CCA/PLS. We end with 10 practical recommendations for application and interpretation of CCA and PLS in future studies (see also Tab. 1).

Results

A generative model for cross-dataset multivariate associations. To analyze sampling properties of CCA and PLS, we need to generate synthetic datasets of stochastic samples with known properties and with known correlation structure across two multivariate datasets. We therefore developed a generative statistical modeling framework that satisfies these requirements, which we refer to as GEMMR (Generative Modeling of Multivariate Relationships). GEMMR is central to all that follows as it allows us to design and generate synthetic datasets, investigate the dependence of CCA/PLS sampling errors on dataset size and assumed covariances, estimate weight errors in CCAs reported in the literature, and calculate sample sizes required to bound estimation errors.

To describe GEMMR, first note that data for CCA and PLS consist of two datasets, given as data matrices X and Y, with respectively p_x and p_y features (columns) and an equal number n of samples (rows). We assume a principal component analysis (PCA) has been applied separately to each dataset so that, without loss of information, the columns of X and Y are principal component (PC) scores. The PC scores' variances, which are also the eigenvalues of the within-set covariance matrices, S_{XX} and S_{YY}, are modeled to decay with a power-law dependence (Fig. 1B) for PLS, as empirical variance spectra often follow approximate power-laws (for examples, see Fig. S1A-J). For CCA, which optimizes correlations instead of covariances, the two datasets are effectively whitened internally during the analysis (see Methods).

Between-set associations between X and Y (Fig. 1C) are summarized in the cross-covariance matrix S_{XY}. By performing a singular value decomposition of S_{XY} a solution for CCA and PLS can be obtained (after whitening for CCA, see Methods) with the singular values giving the association strengths and the singular vectors encoding the weight vectors for the latent between-set association modes. Conversely, given association strengths and weight vectors for between-set association modes (i.e., the solution to CCA or PLS), the corresponding cross-covariance matrix can be assembled making use of the same singular value decomposition, where different weight normalizations reflect the distinction between CCA and PLS (see Methods and Fig. S4). The joint covariance matrix for X and Y is then composed from the within- and between-set covariances (Fig. 1D) and the normal distribution associated with this joint covariance matrix constitutes our generative model for CCA and PLS.

In the following we systematically vary the parameters on which the generative model depends and investigate their downstream effects on the stability of CCA and PLS solutions. Specifically, we vary the number of features (keeping the same number of features for both datasets for simplicity), the assumed between-set correlation, the power-laws describing the within-set variances, and the number of samples drawn. Weight vectors are chosen randomly and constrained such that the ensuing X and Y scores explain at least half as much variance as an average principal component in their respective sets. For simplicity, we restrict our present analyses to a single between-set association mode. Of note, in all of the manuscript, "number of features" denotes the total number across both X and Y, i.e., $p_x + p_y$. Also of note, the terminology used for CCA properties (e. g. weights, scores) is not uniform across the literature. CCA/PLS "scores" (as described above) could also be called "variates", "weights" (as described above) could also be called "vectors" or "saliences", and "loadings" (as described below) could also be called "parameters" or "structure coefficients". For CCA, the correlation between the score vectors, i. e. the "between-set correlations" or "inter-set correlations" are also called "canonical correlations" [4, 6, 9, 18, 19].

Sample-size dependence of estimation error. Using randomly sampled surrogate datasets from our generative model, we characterized the estimation error in multiple elements of CCA/PLS solutions. First, we asked whether a significant association can robustly be detected, quantified by statistical power. To that end we calculate the association strength in each synthetic dataset as well as in 1000 permutations of sample labels, and calculate the probability that association strengths are stronger in permuted datasets, giving a p-value. We repeat this process, and estimate statistical power as the probability that the p-value is below $\alpha = 0.05$ across 100 synthetic datasets drawn from the same normal distribution with given covariance matrix. For a sufficient number of samples that depends on the other parameter values statistical power eventually becomes 1 (Fig. 2A-B). Note that here we use "samples per feature" as an effective sample-size measurement to account for the fact that datasets in practice can have widely varying dimensionalities (see also Fig. S5). A typical value in the brain-behavior CCA/PLS literature is about 5 samples per feature (Fig. S6A), which is also marked in Fig. 2.

Second, we evaluated the association strength (Fig. 2C-D). While the observed association strength converges to its true value for sufficiently large sample sizes, it consistently overestimates the true value and decreases monotonically with sample size. Moreover, for very small sample sizes, observed association strengths are very similarly high, independent of the true between-set correlation (Fig. S8O-P). Thus as above, a sufficient sample size, depending on other parameters of the covariance matrix, is needed to bound the error in the association strength. We also compared in-sample estimates for the association strength to cross-validated estimates. We found that cross-validated estimates underestimate the true value (Fig. S9A-B) to a similar degree as in-sample estimates overestimate it (Fig. S9C-D). Interestingly, the average of in-sample and cross-validated association strength was a better estimator than either of the two alone in our simulations (Fig. S9E-F). Finally, bootstrapped association strengths overestimated, on average, slightly more than in-sample estimates (Fig. S9G-H).

Third, CCA/PLS solutions provide weights that encode the nature of the association in each dataset. We quantify the corresponding estimation error as the cosine distance between the true and estimated weights, separately for X and Y and taking the greater of the two. As the sign of weights is ambiguous in CCA and PLS it is chosen to obtain a positive between-set correlation between observed and true weight. We found that weight error decreases monotonically with sample size (Fig. 2E-F). Bootstrapped weight errors were again, on average, slightly larger than in-sample estimates (Fig. S9I-L), while the variability of individual weight elements across repeated datasets can be well approximated through bootstrapping (Fig. S9M-N).

Fourth, CCA/PLS solutions provide scores which represent a latent value assigned to each sample (e.g., subject). Applying true and estimated weights to common test data to obtain test scores, score error is quantified as $1 -$ Spearman correlation between true and estimated scores. It also decreased with sample size (Fig. 2G-H).

Finally, some studies report loadings, i. e. the correlations between original data features and CCA/PLS scores (Fig. S10A-B). In practice, original data features are generally different from principal component scores, but as the relation between these two data representations cannot be constrained, we calculate all loadings here with respect to principal component scores. Moreover, to compare loadings across repeated datasets we calculate loadings for a common test set, as for CCA/PLS scores. The loading error is then obtained as ($1 -$ Pearson correlation) between test loadings and true loadings. Like other error metrics, it decayed with sample size (Fig. 2I-J). Interestingly, convergence for PLS is somewhat worse than for CCA across all metrics assessed in Fig. 2.

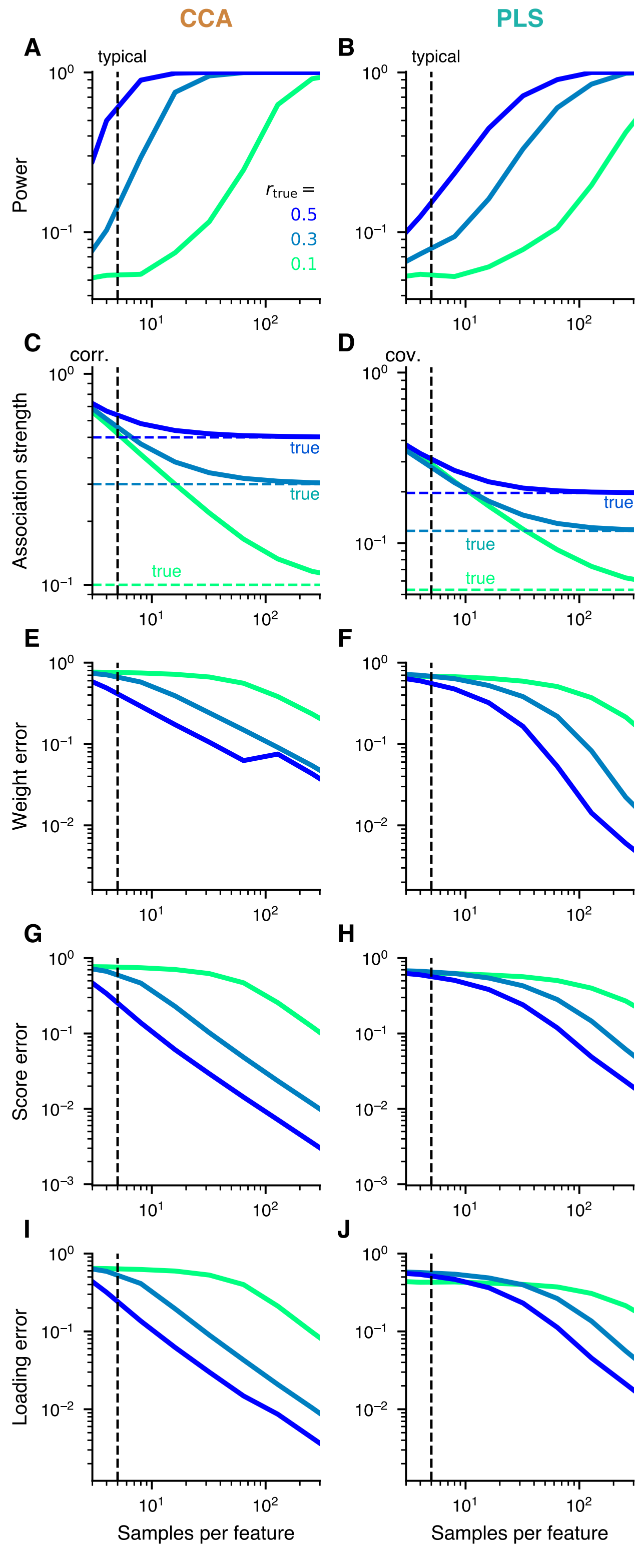

Figure 2. Sample-size dependence of CCA and PLS. (Caption follows)

Figure 2. Sample-size dependence of CCA and PLS. For sufficiently large sample sizes, statistical power to detect a non-zero correlation converges to 1 **(A, B)**, between-set covariances approach their assumed true value **(C, D)**, and weight **(E, F)**, score **(G, H)**, and loading **(I, J)** errors become close to 0. Left and right columns show results for CCA and PLS, respectively. For all metrics, convergence depends on the true between-set correlation r_{true} and is slower if r_{true} is low. Note in **C, D)** that estimated between-set association strengths overestimate the true values. The true value in **C)** is the indicated correlation, whereas in **D)** it is given by the indicated correlation multiplied by the standard deviations of X and Y scores which depend on the specific weight vectors. The dashed vertical line at 5 samples per feature represents a typically used value (Fig. S6A).

Weight error and stability. Fig. 2 quantifies the effect of sampling error on various aspects of the model in terms of summary statistics. We next focus on the error and stability of the weights, due to their centrality in CCA/PLS analysis in describing how features carry between-set association. First we illustrate how weight vectors are affected when typically used sample-to-feature ratios are used. For this illustration we set up a joint covariance matrix with a true between-set correlation of 0.3 and assuming 100 features per dataset, and then generated synthetic datasets with either 5 or 50 samples per feature. Using 5 samples per feature, estimated CCA weights varied so strongly that the true weight were not discernable in the confidence intervals (Fig. 3A). In contrast, with 50 samples per feature the true weights became more resolved. For PLS, the confidence interval for weights estimated with 5 or 50 samples per feature did not even align with the true weights (Fig. 3B) indicating that even more samples than for CCA should be used.

We next assessed weight stability, i.e., the consistency of estimated weights across independent sample datasets. We quantified weight stability as the cosine-similarity between weights obtained from two independently drawn datasets and averaged across pairs of datasets. When the datasets consisted of only few samples, the average weight stability was close to 0 for CCA and eventually converged to 1 (i. e. perfect similarity) with more samples (Fig. 3E). PLS exhibited striking differences from CCA: mean weight stability had a relatively high value even at low sample sizes where weight error is very high (Figs. 3F, 2F), with high variability across datasets.

Finally, to show the dependence of weight error on the assumed true between-set correlation and the number of features we estimated the number of samples required to obtain less than 10% weight error (Fig. 3C-D). The required sample size is higher for increasing number of features, and lower for increasing true between-set correlation. We also observe that, by this metric, required sample sizes can be much larger than typically used sample sizes in CCA/PLS studies.

Weight PC1 similarity in PLS. Figs. 3A-B,E-F and 2E-F show that at low sample sizes, PLS weights exhibit, on average, high error but also reasonably high stability. This combination suggests a systematic bias in PLS weights toward a different axis than the true latent axis of association. To gain further intuition of this phenomenon, we first consider the case of both datasets comprising 2 features each, so that weight vectors are 2-dimensional unit vectors lying on a circle. Setting $r_{\mathrm{true}} = 0.3$, we drew synthetic datasets from the normal distribution and performed CCA or PLS on these. When 50 samples per feature were used, all resulting weight vectors scattered tightly around the true weight vectors (Fig. 3G-H). With only 5 samples per feature, which is typical in CCA/PLS studies (Fig. S6A), the distribution was much wider. For CCA the circular histogram peaked around the true value. In contrast, for PLS the peak was shifted towards the first principal component axis when 5 samples per feature were used.

Next, we investigated how this weight bias toward the first principal component axis in PLS manifests more generally. We first considered an illustrative data regime (64 features/dataset, $r_{\text{true}} = 0.3$). We quantified the PC similarity as the cosine similarity between estimated weight vectors and a principal component axis. Compared to CCA, PC similarity to the dominant principal comonents was strong for PLS, even with a large number of samples (Fig. 3I-J), and more so for a small number of samples. Note also, that the average PC similarity in permuted datasets was similar to that in unpermuted datasets, for both CCA and PLS. Finally, these observations also held for datasets with differing number of features and true correlations. For PLS the weight vectors are biased toward the first principal component axis, compared to CCA, and more strongly than random weight vectors, particularly when few samples per feature were used to estimate them (Fig. 3L).

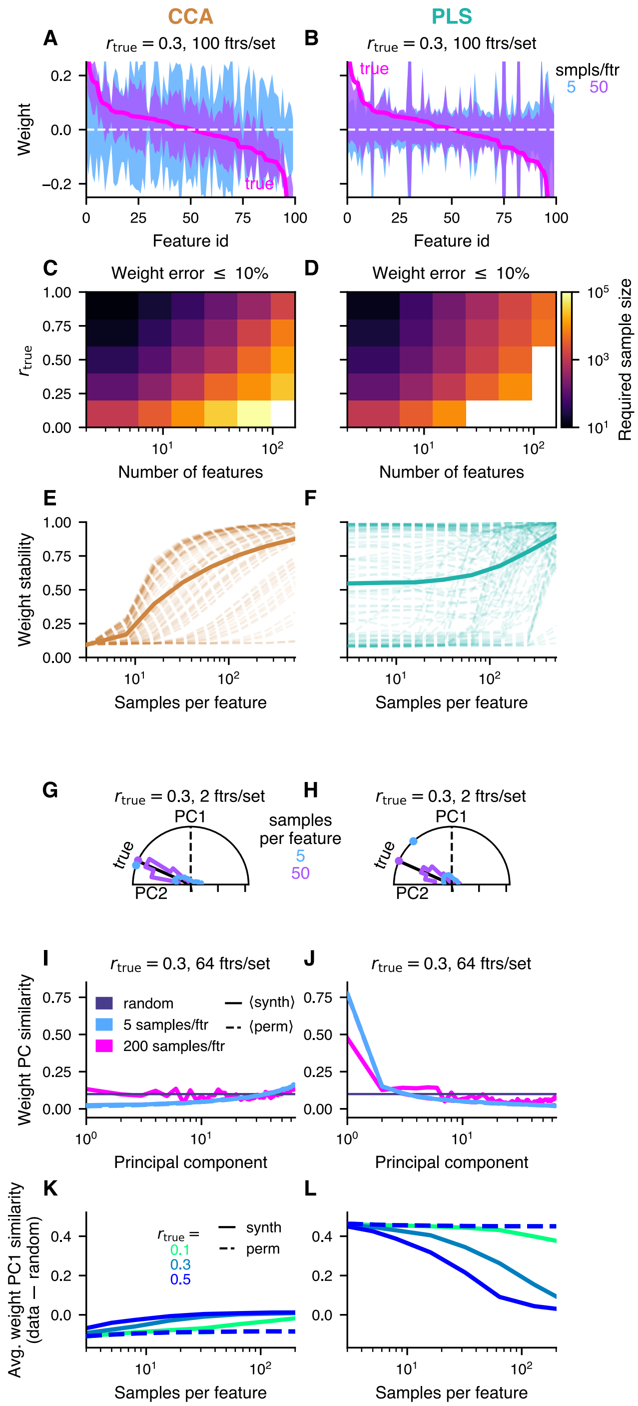

Figure 3. Large number of samples required to obtain good weight estimates. (Caption follows)

Figure 3. Large number of samples required to obtain good weight estimates. A, B) Realistic example where the true betwen-set correlation was set to r_{true} =0.3. The area plots show 95 % confidence intervals. **A)** For CCA, good weight estimates could be obtained with 50, but not 5, samples per feature. **B)** For PLS even more samples were necessary. **C-D)** Sample sizes required to obtain less than 10% weight errors. **E-F)** Weight stability, i. e. the average cosine-similarity between weights across pairs of repetitions, increases towards 1 (identical weights) with more samples. For PLS, weight stability can be high, even with few samples. The true between-set correlation was set to r_{true}=0.3. **G-H)** Example situation assuming a true between-set correlation of r_{true}=0.3 between datasets and 2 features each for both X and Y datasets. In this 2-dimensional setting weight vectors, scaled to unit length, lie on a circle. Synthetic datasets were generated repeatedly. 5 samples per feature gave good estimates in many cases but notably all possible weight vectors occurred frequently. 50 sampes per feature resulted in consistently better estimates. Dots near border of semi-circles indicate directional means of distributions. **I-J)** Another example with 64 features per dataset and a between-set correlation r_{true}=0.3. PLS weights have a strong PC1 similarity (cosine-similarity with first principal component). **K-L)** PC1 similarity was stronger for PLS (**L**) than for CCA (**K**) also for datasets with varying number of features and true between-set correlations r_{true}. Shown is relative PC1 similarity across synthetic datasets with varying number of features, relative to the expected PC1 similarity of a randomly chosen vector with dimension matched to each synthetic dataset.

Comparison of loadings and weights. In addition to weights, loadings provide a measure of importance for each variable. It has previously been argued that loadings provide provide a better basis for interpretation of linear models than weights, as weights are more susceptible to noise [**?**]. Moreover, mixed results have been reported regarding differential stability of weights and loadings [**?**, 18]. Here, we have used simulations to elucidate these issues. We generated data with GEMMR assuming 32 features / set and a true between-set correlation of 0.3. Here (unlike in the rest of the manuscript), we also worked in coordinate systems that were randomly rotated from the principal-component coordinate system, to mimic each variable's original coordinate system. This matters, as the coordinate system affects loadings.

We first investigated stability. Stability was measured as pairwise cosine-similarity of weight vectors, and pairwise pearson-correlation of loading vectors, respectively, obtained from a CCA / PLS of 25 independent data samples. This was then repeated for 25 different data scenarios, each of which had different true weight / loading vectors and different relative orientation to the principal component axes. For CCA (Fig. 4A) loadings were slightly more stable than weights. At the same time, loading error decreased as loading stability increased. For PLS (Fig. 4B) weight and in particular loading stability was high already for small sample sizes, where the loading error was high as well. This indicates that, for PLS, stability of loadings is not a good indicator for accuracy of loadings.

Stability depended on the within-set variance spectrum (Fig. 4C,D). For both CCA and PLS, the steeper the powerlaw describing this spectrum (i. e. if there exist strongly dominating PCs) stability was higher compared to shallower powerlaws. For PLS, we observed a similar effect for weights.

Finally, we evaluated the pattern of weights and loadings in terms of their similarity to principal component axes (Fig. 4E, F). CCA weights for small sample sizes overlapped more with low-variance principal component axes. On the other hand, CCA loadings, as well as PLS loadings and weights resembled more dominant principal component axes. Thus, the within-set variance can have a strong influence on these CCA / PLS properties. For a specific example of this effect in real data see Fig. S7. Of note, strong similarity of weights / loadings with dominant PC-axes might complicate interpretation of between-set effects based on weights or loadings of, say, modality X, as these weights or loadings might appear similar even for disparate input datasets Y.

Empirical brain-behavior data. Do these phenomena observed in synthetic data from our generative modeling framework also hold in empirical data? We focused on two state-of-the-art population neuroimaging datasets: Human Connectome Project (HCP) [1] and UK Biobank (UKB) [2]. Both datasets

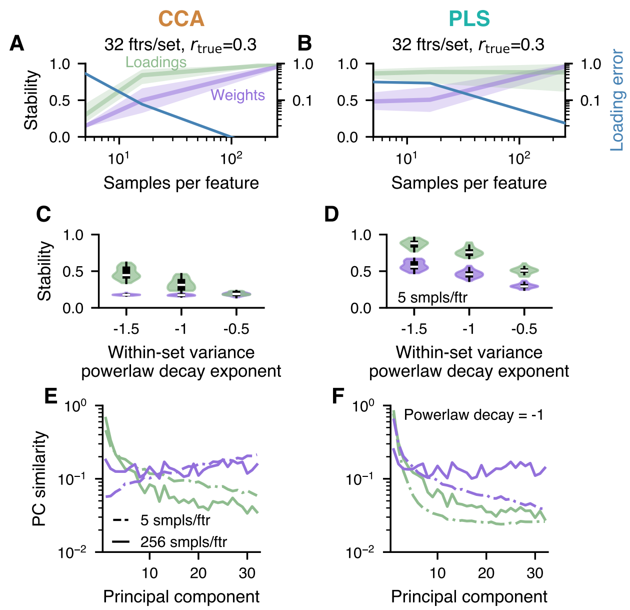

Figure 4. Stability and PC similarity of weights and loadings. A CCA loadings are slightly more stable than weights. **B** Similarly for PLS. Also, PLS loadings were extremely stable despite large loading error. **C-D** The steeper the within-set variance spectrum the more stable were loadings (and weights for PLS). **E** CCA loadings resembled dominant principal component axes while CCA weights for small sample sizes resembled more low-variance PC-axes. **F** PLS weights and loadings resembled PC-axes.

provide multi-modal neuroimaging data along with a wide range of behavioral and demographic measures, and both have been used in prior studies using CCA to map brain-behavior relationships [2, 3, 20–24]. HCP, comprising around 1200 subjects, is one of the lager neuroimaging datasets available and is of exceptional quality. We analyzed two neuroimaging modalities in the HCP dataset, resting-state functional MRI (fMRI) (in 948 subjects) and diffusion MRI (dMRI) (in 1020 subjects, shown in Fig. S2A-D). UKB is a population-level study and, to our knowledge, the largest available neuroimaging dataset. We analyzed fMRI features from 20000 UKB subjects. HCP and UKB thereby provide two independent testbeds, across neuroimaging modalities and with large numbers of subjects, to investigate error and stability of CCA/PLS in brain-behavior data.

After modality-specific preprocessing (see Methods), both datasets in each of the three analyses were deconfounded and reduced to 100 principal components (see Methods and Fig. S1K), in agreement with prior CCA studies of HCP data [3, 20–24]. Functional connectivity features were extracted from fMRI data and structural connectivity features were extracted from dMRI. Note that, as only a limited number of samples were available in these empirical datasets, we cannot use increasingly more samples to determine how CCA or PLS converge with sample size (as we did with synthetic data above). Instead, we repeatedly formed two non-overlapping sets of subjects of the available data, varying their sizes from 202 up to 50 % of the available number of subjects.

We found that the first mode of association was statistically significant for all three sets of data and for both CCA and PLS (p-values from a permutation test were A) 0.001, C) 0.004, I) 0.001, K) 0.001). Association strengths decreased with increasing size of the subsamples, but clearly converged only for the UKB data. Cross-validated association strengths estimates increased with subsample size and, for UKB, converged to the same value as the in-sample size. Fig. 5A overlays reported CCA results from other publications that used 100 features per set in HCP data, which further confirms the decreasing trend of association strength as a function of sample size.

Weight stabilities (i. e., the cosine-similarities between the two estimated weight vectors of a pair of data matrices with non-overlapping subjects) for the HCP datasets remained low and at intermediate values for CCA and PLS, respectively. In contrast, in the UKB dataset weight stabilities reached values close to 1 (perfect similarity). Moreover, for all datasets weight PC1 similarity was close to 0 for CCA but markedly larger for PLS weights. We also investigated stability (i. e. the Pearson correlation between the two estimated loading vectors of a pair of data matrices with non-overlapping subjects) and PC1 similarity for loadings in the HCP-fMRI and UKB dataset. Here, loadings were calculated either for principal components (i. e. the correlations between PC-scores and CCA/PLS scores) or for original variables (i. e. the correlations between original data variables and CCA/PLS scores). Both PC-loadings and original-variable loadings show a similar behavior as weights, with loadings being slightly more similar to PC1 than weights.

All these results were in agreement with analyses of synthetic data discussed above (Figs. 2-4). Altogether, we emphasize the overall similarity between CCA analyses of different data modalities and features (first and second row in Fig. 5) and data of similar nature from different sources (first and third row in Fig. 5). This suggests that sampling error is a major determinant in CCA and PLS outcomes and this is valid across imaging modalities and for independent data sources. Note also that stable CCA and PLS results with a large number of considered features can be obtained with sample sizes that become available with UKB-level datasets.

We also considered reducing the data to different numbers of principal components than 100. Fig. S2E-H shows a re-analysis of HCP data in which a smaller number of principal components was selected according to an optimization procedure [25]. Moreover, using UKB data, we separately varied the number of retained neuroimaging and behavioral principal components from 1 to 100 and calculated in-sample and cross-validated association strengths for CCA and PLS (Fig. S3). For both methods, we found that the obtained association strength rose strongly when retaining an increasing number of behavioral PCs, but only up to about 10. Retaining more than 10 behavioral PCs lead to more marginal increases in association strength. The situation for neuroimaging PCs differed between the methods, however. For CCA, retaining more neuroimaging PCs improved the association strength up to about 20-40 before plateauing. For PLS, on the other hand, the top PCs ($\approx$ 5-10) were enough for the association strength to plateau. This is in line with the described PC similarity for PLS. Altogether, these findings suggest that, for both methods, the between-set association is encoded in the top few behavioral PCs, and this can be exploited with dimensionality reduction methods before using CCA / PLS.

Samples per feature alone predicts published CCA strengths. We next examined stability and association strengths in CCA analyses of empirical datasets more generally. To that end we performed an analysis of the published literature using CCA with neuroimaging data to map brain-behavior relationships. From 100 CCAs that were reported in 31 publications (see Methods), we extracting the number of samples, number of features, and association strengths. As the within-set variance spectrum is not typically reported, but would be required to assess PLS results (as described above), we did not perform such an analysis for PLS.

Most studies used less than 10 samples per feature (Fig. 6A and S6A). Overlaying reported canonical correlations as a function of samples per feature on top of predictions from our generative model shows that most published CCAs we compiled are compatible with a range of true correlations, from about 0.5 down to 0 (Fig. 6A). Interestingly, despite the fact that these studies investigated different questions using different datasets and modalities, the reported canonical correlation could be well predicted simply by the number of samples per feature alone ($R^2 = 0.83$).

We next asked whether weight errors can be estimated from published CCAs. As these are unknown in principle, we estimated them using our generative modeling framework. We did this by (i) generating synthetic datasets of the same size as a given empirical dataset, and sweeping through assumed true correlations between 0 and 1 (ii) selecting those synthetic datasets for which the estimated canonical correlation matches the empirically observed one, and (iii) using the weight errors in these matched synthetic datasets as a proxy for the weight error in the empirical dataset (Fig. S6C). This resulted in a distribution of weight errors across the matching synthetic datasets for each published CCA study that we considered. The mean of these distributions are overlaid in color in Fig. 6A and the range of the distributions is shown in Fig. 6B. The mean weight error falls off roughly with the distance to the correlation-vs-samples/feature curve for

[h!]

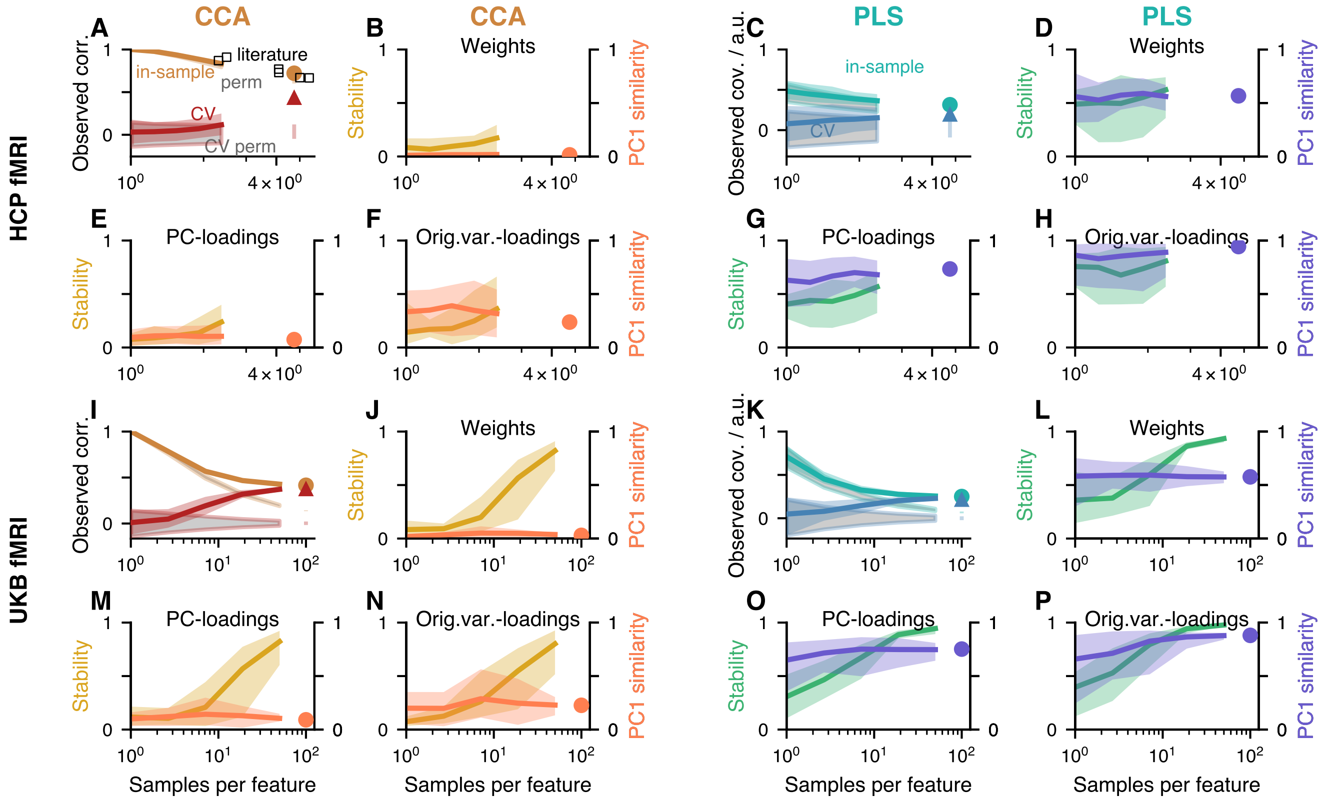

Figure 5. CCA and PLS analysis of empirical population neuroimaging datasets. For all datasets and for both CCA and PLS a significant mode of association was detected. Association strengths monotonically decreased with size of the subsamples (orange in column 1, green in column 3). Association strengths for permuted data are shown in grey (with orange and green outlines in columns 1 and 3, respectively). Deviations of the orange and green curves from the grey curves occur for sufficient sample sizes and correspond to significant p-values. Note how the curves clearly flatten for UKB but not for HCP data where the number of available subjects is much lower. Circle indicates the estimated value using all available data and the vertical bar in the same color below it denotes the corresponding 95 % confidence interval obtained from permuted data. In **A)** we also overlaid reported canonical correlations from other studies that used HCP data reduced to 100 principal components. Cross-validated association strengths shown in red (column 1) and blue (column 3), cross-validated estimation strengths of permuted datasets in grey with red and blue outlines in columns 1 and 3, respectively. Triangle indicates the cross-validated association strength using all data and the vertical bar in the same color below it denotes the corresponding 95 % confidence interval obtained from permuted data. Cross-validated association strengths were always lower than in-sample estimates and increased with sample size. For UKB (but not yet for HCP) cross-validated association strengths converged to the same value as the in-sample estimate. Weight stabilities increased with sample size for UKB and slightly for the PLS analyses of HCP datasets, while they remained low for the CCA analyses of HCP datasets. PC1 weight similarity was low for CCA but high for PLS. Both PC-loadings and original-variable-loadings show a similar pattern as weights, with loadings being slightly more similar to PC1 than weights. All analyses were performed with repeatedly subsampled data of varying sizes (x-axis). For each subsample size and repetition, we created two non-overlapping sets of subjects and calculated stability of weights / loadings using these non-overlapping pairs.

permuted data (see also Fig. S6B). Altogether, these analyses suggest that many published CCA studies might have unstable feature weights due to an insufficient sample size.

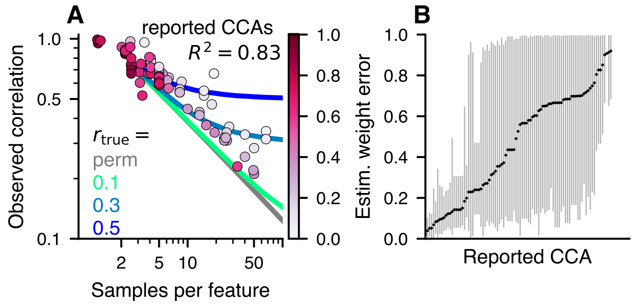

Figure 6. CCAs reported in the population neuroimaging literature might often be unstable. A) Canonical correlations and the number of samples per features are extracted from the literature and overlaid on predictions from the generative model for various between-set correlations r_{true}. Many studies employed a small number of samples per feature (cf. also Fig. S6A) and reported a large canonical correlation. These studies fall in the top-left corner of the plot, where predictions from the generative model for $r_{\text{true}} < 0.5$ and also the null-data (having no between-set correlation, resulting from permuted datasets) are indistinguishable (see also Fig. S8O-P). In fact, the reported canonical correlation can be predicted from the used number of samples per feature alone using linear regression ($R^2 = 0.83$). We also estimated the weight error (encoded in the colorbar) for each reported CCA (details are illustrated in Fig. S6C). The farther away a CCA lies from the predictions for permuted data the lower the mean-estimated weight error (cf. Fig. S6B). **B)** The distribution of estimated weight errors for each reported CCA is shown along the y-axis. For many studies weight errors could be quite large, suggesting that conclusions drawn from interpreting weights might not be robust.

Benefit of cross-loadings in PLS. Given the instability associated with estimated weight vectors, we investigated whether other measures provide better feature profiles. Specifically, we compared loadings and cross-loadings. Cross-loadings are the correlations across samples between CCA/PLS scores of one dataset with the original data features of the other dataset (unlike loadings, which are the correlations between CCA/PLS scores and original features of the same dataset). In CCA, they are collinear (see Methods and Fig. S10C) and to obtain estimates that have at most 10 % loading or cross-loading error required about the same number of samples (Fig. S10E). For PLS, on the other hand, true loadings and cross-loadings were, albeit not collinear still very similar (Fig. S10D), but cross-loadings could be estimated to within 10 % error with about 20 % to 50 % less samples as loadings in our simulations (Fig. S10F).

Calculator for required sample size. In both synthetic and empirical datasets we have seen that sample size plays a critical role to guarantee stability and interpretability of CCA and PLS, and that many existing applications may suffer from a lack of samples. How many samples are required, given particular dataset properties? We answer this question with the help of *GEMMR*, our generative modeling framework described above. Specifically, we suggest to base the decision on a combination of criteria, by bounding statistical power as well as relative error in association strength, weight error, score error and loading error at the same time. Requiring at least 90 % power and admitting at most 10 % error for the other metrics, we determined the corresponding sample sizes in synthetic datasets by interpolating the curves in Fig. 2 (see Fig. S11A and Methods). The results are shown in Fig. 7 (see also Fig. S8A-L). Assuming, for example, that the decay constants of the variance spectra satisfy $a_x + a_y = -2$, several hundreds to thousands of samples are necessary to achieve the indicated power and error bounds when the true correlation is 0.3 (Fig. 7A). More generally, the required sample size per feature as a function of the true correlation roughly follows a power-law dependence, with a strong increase in required sample size when the true correlation is low (Fig. 7B). Interestingly, PLS generally needs more samples than CCA (see also Fig. S12). As mentioned above, accurate estimates of the association strength alone (as opposed to power, association strength, weight, score and loading error at the same time) could be obtained in our simulations with fewer samples: by averaging the in-sample with a cross-validated estimate (Fig. S9E-F). Moreover, accurate estimates of a PLS feature profile required fewer samples when assessed as cross-loadings (Fig. S10F). We also evaluated required sample sizes for sparse CCA with our analysis framework (Fig. S8M-N) but note that an in-depth analysis of sparse CCA is beyond the scope of this manuscript.

Given the complexity and computational expense to generate and analyze enough synthetic datasets to obtain sample size estimates in the way described above, we finally asked whether we could formulate a concise, easy-to-use description of the relationship between model parameters and required sample size. To that end, we fitted a linear model to the logarithm of the required sample size, using logarithms of total number of features and true correlation as predictors (Figs. S11). We additionally included a predictor for the decay constant of the within-set variance spectrum, $|a_x + a_y|$. Using split-half predictions to validate the model, we find good predictive power for CCA and PLS (Fig. S11C-D).

Discussion

We characterized CCA and PLS through a parameterized generative modeling framework. CCA and PLS require a sufficient number of samples to work as intended and the required sample size depends on the number of features in the data, the assumed true correlation, and the principal component variance spectrum for each dataset.

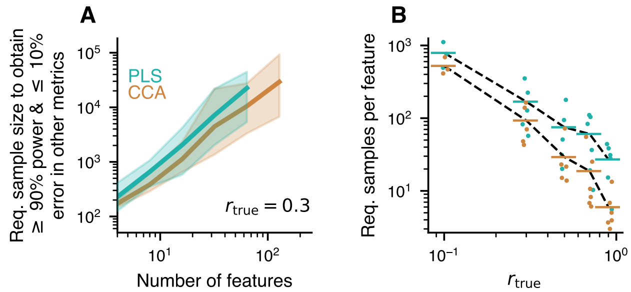

Figure 7. Required sample sizes. Sample sizes to obtain at least 90 % power as well as at most 10 % error for the association strength, weight, scores and loadings. Shown estimates are constrained by the within-set variance spectrum (here $a_x + a_y = -2$, cf. Fig. S8E-L for other values). **A)** Assuming a true between-set correlation of $r_{\text{true}} = 0.3$ 100s to several 1000s of samples are required to reach target power and error levels. See Fig. S8A-D for other values of r_{true}. Shaded areas show 95 % confidence intervals. **B)** The required number of samples divided by the total number of features in X and Y scales with r_{true}. For $r_{\text{true}} = 0.3$ about 50 samples per feature are necessary to reach target power and error levels in CCA, which is much more than typically used (cf. Fig. S6A). Generally, more features are necessary for PLS than CCA (see also Fig. S12), and if the true correlation is smaller. Every point for a given r_{true} represents a different number of features and is slightly jittered for visibility. Values for a given p_X are only shown here if simulations were available for both CCA and PLS.

Generative model for CCA and PLS. At least for CCA, the distribution of canonical correlations has been reported to be intractable, even for normally distributed data [26]. Thus, a generative model is an attractive alternative to investigate sampling properties. Our generative model for CCA and PLS made it possible to investigate all aspects of a solution, beyond just the canonical correlations, at the cost of higher computational expenses. For example, the generative model can be used to systematically explore parameter dependencies, to assess stability, to calculate required sample sizes in new studies, and to estimate weight stability in previously published studies. While this generative model was developed for CCA and PLS, it can also be used to investigate related methods like sparse variants [27]. More broadly, extending the generative model to other domains, like associations with gene expression maps where spatial autocorrelations between samples have to be taken into account, could be a promising future direction.

Pitfalls in CCA and PLS. Association strengths can be overestimated and, at least for CCA when the number of samples per feature as well as the true correlation are low, observed canonical correlations can be compatible with of true correlations, down to zero (Fig. S8O-P). Estimated weight vectors do not need to resemble the true weights when the number of samples is low and can overfit, i. e. vary strongly between datasets sampled from the same population(Fig. 3), affecting significantly their interpretability and generalizability. Furthermore, PLS weights also show a consistent similarity to the first principal component axis (Fig. 3G-L). As a consequence, similarity of weights obtained for two datasets drawn from the same population is necessary but not sufficient to infer replicability. The PC1 similarity also existed for null data. Therefore, estimated weights that strongly resemble the first principal component axis need not indicate an association, but could instead indicate the absence of an association, or insufficient sample size. Importantly, we have shown that the same pitfalls also appear in empirical data.

Differences between CCA and PLS. First and foremost, CCA and PLS have different objectives: while CCA finds weighted composites with the stronfgest possible correlation between datasets, PLS maximizes their covariance. When features do not have a natural commensurate scale, CCA can be attractive due to its scale invariance (see Fig. 1 and Methods). In situations where both analyses make sense, PLS comes with the additional complication that estimated weights show a consistent similarity towards the first principal component axis. Moreover, our analyses suggest that the required number of samples for PLS is usually higher than for CCA, except maybe when the true PLS weights overlap strongly with the first principal component axis (see Fig. S12). Based on these arguments, CCA might often be preferable to PLS.

Sample size calculator for CCA and PLS. Previous literature, based on small numbers of specific datasets or Monte-Carlo simulations, has suggested using between 10 and 70 samples per feature for CCA [11, 15, 17]. Beyond that, our calculator is able to suggest sample sizes for the given characteristics of a dataset, and can do so for both CCA and PLS. As an example, consider the UKB data in Fig. 5. Both in-sample and cross-validated CCA association strengths converge to about 0.5. Fig. 7B then suggests to use about 20 samples per feature, i. e. 4000 samples, to obtain at least 90 % power and at most 10 % error in other metrics. This is compatible with Fig. 5J: at 4000 subjects weight stability is about 0.8 (note that weight stability measures similarity of weights between different repetitions of the dataset; we expect the similarity of a weight vector to the true weight vector—which is the measure going into the sample size calculation—to be slightly higher on average). Our calculator is made available as an open-source Python package named *GEMMR* (Generative Modeling of Multivariate Relationships).

Brain-behavior associations. CCA and PLS have become popular methods to reveal associations between neuroimaging and behavioral measures [2, 3, 9, 13, 21–24, 28]. The main interest in these applications lies in interpreting weights or loadings to understand the profiles of neural and behavioral features carrying the brain-behavior relationship. We have shown, however, that stability and interpretability of weights or loadings are contingent on a sufficient sample size which, in turn, depends on the true between-set correlation.

How strong are true between-set correlations? While this depends on the data at hand, and is in principle unknown *a priori*, our analyses provide estimates in the case of brain-behavior associations. First, we saw in UKB data that both in-sample and cross-validated canonical correlations converged to a value of around 0.5. As the included behavioral measures comprised a wide range of categories (cognitive, physical, life-style measures and early life factors) this canonical correlation is probably more on the upper end, such that brain-behavior associations probing more specialized modes are likely lower. Second, we saw in a literature analysis of brain-behavior CCAs that reported canonical correlations as a function of sample-to-feature ratios largely follow the trends predicted by our generative model, despite different datasets investigated in each study. We also saw that few studies which had 10-20 samples per feature reported canonical correlations around 0.5-0.7, while most studies with substantially more than 10 samples per feature appeared to be compatible only with values ≤ 0.3. In this way, we conclude that true canonical correlations in brain-behavior applications are probably not greater than 0.3 in many cases.

Assuming a true between-set correlation of 0.3, our generative model implies that about 50 samples per feature are required at minimum to obtain stability in CCA results. We have shown that many published brain-behavior CCAs do not meet this criterion. Moreover, in HCP data we saw clear signs that the available sample size was too small to obtain stable solutions—despite that the HCP data comprised around 1000 subjects which is one of the largest and highest-quality neuroimaging datasets available to date. On the other hand, with UKB data, where we used 20000 subjects, CCA and PLS results appeared to have converged. As the resources required to collect datasets of this size go well beyond what is available to typical research groups, this observation supports the accruement of datasets that are shared widely [29].

Generalizability. Small sample and effect sizes have been identified as challenges for neuroimaging that impact replicability and generalizability [30, 31]. Here, we have considered stability of CCA/PLS analyses and found that observed association strengths decrease with used sample-per-feature ratio. Similarly, a decrease in reported effect size with increasing sample size has been reported in meta-analyses of various classification tasks of neuroimaging measures [32]. These sample-size dependences of the observed effect sizes are an indication of instability.

A judicious choice of sample size, together with an estimate of the effect size, are thus advisable at the planning stage of an experiment or CCA/PLS analysis. Our generative modeling framework provide estimates for both. Beyond that, non-biological factors—such as batch or site effects [33] or flexibility in the data processing pipeline [34]—certainly contribute to unstable outcomes and could be addressed in extensions of the generative model. External validation with separate datasets is also necessary to establish generalizability of findings beyond the dataset under investigation.

Limitations and future directions. For tractability it was necessary to make a number of assumptions in our study. Except for Fig. 6 it was assumed that both datasets had an equal number of features (but see Fig. S5 where we used different number of features for the two datasets). We also assumed that data were normally distributed, which is often not true in practice. For example, cognitive scores are commonly recorded on an ordinal scale. To address that, we used empirical datasets and found similar sample-size dependencies as in synthetic datasets. In an investigation of the stability of CCA for non-normal data varying kurtosis had minimal effects [17]. We then assumed the existence of a single cross-modality axis of association, but in practice several ones might be present. Moreover, we assumed that data are described in a principal component (PC) basis. In practice, however, PCs and the number of PCs need to be estimated, too. This introduces an additional uncertainty, although, presumably, of lesser influence than the inherent sampling error in CCA and PLS. We therefore expect that a dataset whose features have been rotated into a new coordinate system by an orthogonal transformation matrix to have the same sample size requirements as the untransformed dataset. Furthermore, we used "samples per feature" as an effective sample-size parameter to account for the fact that datasets in practice have very different dimensionalities. This is in line with previous studies [32, 35]. Here, Fig. S5 show that power and error metrics for CCA are parameterized well in terms of "samples per feature", whereas for PLS it is only approximate. Nonetheless, as "samples per feature" is arguably most straightforward to interpret, we presented results in terms of "samples per feature" for both CCA and PLS.

Several related methods have been proposed to potentially circumvent shortcomings of standard CCA and PLS (see [36] for a recent review). Regularized or sparse CCA or PLS [27] have been designed to mitigate the problem of small sample sizes. They modify the modeling objective by introducing a penalty for the elements of the weight vectors, encouraging them to "shrink" to smaller values. This modification has the goal to obtain more accurate predictions, but will also bias the solutions away from their true values. (We assume that, in general, the true weight vectors are non-sparse.) Conceptually, thus, these variants follow more a "predictive" rather than "inferential" modeling goal [37, 38]. Our analysis pipeline evaluated with a commonly used sparse CCA method [27] suggested that in some situations–namely, high dimensionalities and low true correlations—fewer samples were required than for CCA to obtain the same bounds on evaluation metrics (Fig. S8M-N). Nonetheless, although sparse CCA can in principle be used with fewer samples than features, these required sample sizes for sparse CCA were still many times the number of features: when $r_{\text{true}} = 0.3$, for example, 35–50 (depending on the number of features) samples per feature were required. We note, however, that a complete characterization of sparse CCA or PLS methods was beyond the scope of this manuscript. PLS has ben compared to sparse CCA in a setting with more features than samples and it has been concluded that the former (latter) performs better when having fewer (more) than about 500 features per sample [39]. We note that sparse methods are also often used in classification tasks, where they have been observed to provide better prediction but less stable weights [40, 41], which indicates a trade-off between prediction and inference [40]. Correspondingly, it has been suggested to consider weight stability as a criterion in sparsity parameter selection [40, 42].

Moreover, whereas CCA and PLS are restricted to discovering linear relationships between two datasets, there exist non-linear extensions, such as kernel [43], deep [44] or nonparametric [45] CCA, as well as extensions to multiple datasets [46]. Due to their increased expressivity, and therefore capacity to overfit, we expect them to require even larger sample sizes. For classification, kernel and deep-learning methods have been compared to linear methods, using neuroimaging-derived features as input [47]. Accuracy was found similar for kernel, deep-learning and linear methods and also had a similar dependence on sample size, using up to 8000 subjects.

There exist several variants of PLS [6, 7]. Here, we have used one that is also sometimes referred to as PLS correlation (PLSC) or PLS-SVD. This variant treats both datasets X and Y symmetrically and is thus conceptually similar to CCA: both PLSC/PLS-SVD and CCA strive to optimize an association strength between weighted averages of their two original feature sets. In contrast, PLS regression approaches treat the two datasets X and Y asymmetrically, aiming to find the best possible prediction of one dataset's scores from the other dataset's scores. As prediction methods, PLS regression approaches are thus conceptually different from PLSC/PLS-SVD [37, 38] in that the focus lies on obtaining accurate out-of-sample scores, potentially even at the cost of less accurate weights. While evaluation of PLS regression with our pipeline remains an interesting direction for future work, we note that prediction approaches in general might still require considerable number of samples per feature [48].

The number of features in the datasets was an important determinant for stability. Thus, methods for dimensionality reduction hold great promise. On the one hand, there are data-driven methods that, for example, select the number of principal components in a way that takes the between-set correlation into account [25]. Applying this method to HCP data we saw that the reduced number of features the method suggests leads to slightly better convergence (Fig. S2E-H). On the other hand, previous knowledge could be used to preselect the features hypothesized to be most relevant for the question at hand.

Recommendations. We end with 10 recommendations for using CCA or PLS in practice (summarized in Tab. 1).

1. Sample size and the number of features in the datasets are crucial determinants for stability. Therefore, any form of dimensionality reduction as a preprocessing step can be useful, as long as it preserves the features that carry the between-set association. PCA is a popular choice and can be combined with a consideration for the between-set correlation [25].

2. Significance tests used with CCA and PLS usually test the null hypothesis that the between-set association strength is 0. This is a different problem than estimating the strength or the nature of the association [49, 50]. For CCA we find that the number of samples required to obtain 90% power at significance level $\alpha = 0.05$ is lower than to obtain stable association strengths or weights, whereas for PLS the numbers are about commensurate with required sample sizes for other metrics (Fig. S8C-D). As significant results can also be obtained even when power is low, detecting a significant mode of association with either CCA or PLS does not in general indicate that association strengths or weights are stable.

3. CCA and PLS overestimate the association strength for small sample sizes, and we found that cross-validated estimators underestimate it. Interestingly, the average of the in-sample and the cross-validated association strength was a much better estimator in our simulations.

4. The main interest of CCA/PLS studies is often the nature of the between-set association, which is encoded in the weight vectors, loadings and cross-loadings. Every CCA and PLS will provide weights, loadings and cross-loadings, but they may be inaccurate or unstable if an insufficient number of samples was used for estimation. In our PLS simulations, cross-loadings required less samples than weights and loadings to obtain an error of at most 10%.

5. PLS weights that strongly resemble the first principal component axis can indicate that either no association exist or that an insufficient number of samples was used.

6. As a side effect of this similarity of PLS weights towards the first principal component axis, PLS weights can appear stable across different sample sets, although they are inaccurate.

7. Performing CCA or PLS on subsamples of the data can indicate stability, if very similar results are obtained for varying number of samples used, and compared to using all data.

8. Bootstrapped estimates were useful in our simulations for assessing the variability or precision of elements of the weight vectors. Estimates were, however, not accurate: they were as biased as in-sample estimates, i. e. they overestimated association strengths, and both association strength and weight error had a similar sample-size dependence as in-sample estimates.
9. For CCA and PLS analyses in the literature it can be difficult to deduce what datasets precisely were used. We recommend to always explicitly state the used sample size, number of features in both datasets, and obtained association strength. Moreover, the within-set principal component variances are required and are thus useful to report.
10. CCA or PLS requires a sufficient number of samples for reliability. Sample sizes can be calculated using *GEMMR*, the accompanying software package. An assumed but unknown value for the true between-set correlation is needed for the calculation. Our literature survey suggests that between-set correlations are probably not greater than 0.3 in many cases.

Conclusion. We have presented a parameterized generative modeling framework for CCA and PLS. It allows analysis of the stability of CCA and PLS estimates, prospectively and retrospectively. Exploiting this generative model, we have seen that a number of pitfalls exist for using CCA and PLS. In particular, we caution against interpreting CCA and PLS models when the available sample size is low. We have also shown that CCA and PLS in empirical data behave similar to the predictions of the generative model. Sufficient sample sizes depending on characteristics of the data are suggested and can be calculated with the accompanying software package. Altogether, our analyses provide guidelines for using CCA and PLS in practice.

Materials and Methods

Experimental Design.

The goal of this work was to determine requirements for stability of CCA and PLS solutions, both in simulated and empirical data. To do so, we first developed a generative model that allowed us to generate synthetic data with known CCA/PLS solutions. This made it possible to systematically study deviations of estimated from true solutions. Second, we used large state-of-the-art neuroimaging datasets with associated behavioral measurements to confirm the trends that we saw in synthetic data. Specifically, we used data from the Human Connectome Project (HCP) ($n \approx 1000$) and UK Biobank (UKB) ($n = 20000$). Third, we analyzed published CCA results of brain-behavior relationships to investigate sample-size dependence of CCA results in the literature.

Human Connectome Project (HCP) dataset

fMRI data. We used resting-state fMRI (rs-fMRI) from 951 subjects from the HCP 1200-subject data release (03/01/2017) [1]. The rs-fMRI data were preprocessed in accordance with the HCP Minimal Preprocessing Pipeline (MPP). The details of the HCP preprocessing can be found elsewhere [51, 52]. Following the HCP MPP, BOLD time-series were denoised using ICA-FIX [53, 54] and registered across subjects using surface-based multimodal inter-subject registration (MSMAll) [55]. Additionally, global signal, ventricle signal, white matter signal, and subject motion and their first-order temporal derivatives were regressed out [56].

The rs-fMRI time-series of each subject comprised of 2 (69 subjects), 3 (12 subjects), or 4 (870 subjects) sessions. Each rest session was recorded for 15 minutes with a repetition time (TR) of 0.72 s. We removed the first 100 time points from each of the BOLD sessions to mitigate any baseline offsets or signal intensity variation. We subtracted the mean from each session and then concatenated all rest sessions for each subject into a single time-series. Voxel-wise time series were parcellated to obtain region-wise time series using the "RelatedValidation210" atlas from the S1200 release of the HCP [57]. Functional connectivity was then computed as the Fisher-z-transformed Pearson correlation between all pairs of parcels. 3 subjects were excluded (see section below), resulting in a total of 948 subjects with 64620 connectivity features each.

dMRI data. Diffusion MRI (dMRI) data and structural connectivity patterns were obtained as described in [58, 59]. In brief, 41 major white matter (WM) bundles were reconstructed from preprocessed HCP diffusion MRI data [60] using FSL's XTRACT toolbox [59]. The resultant tracts were vectorised and concatenated, giving a WM voxels by tracts matrix. Further, a structural connectivity matrix was computed using FSL's `probtrackx` [61, 62], by seeding cortex/white-grey matter boundary (WGB) vertices and counting visitations to the whole white matter, resulting in a WGB $\times$ WM matrix. Connectivity "blueprints" were then obtained by multiplying the latter with the former matrix. This matrix was parcellated (along rows) into 68 regions with the Desikan-Killany atlas [63] giving a final set of $68 \times 41 = 2788$ connectivity features for each of the 1020 HCP subjects.

Behavioral measures. The same list of 158 behavioral and demographic data items as in [3] was used.

Confounds. We used the following items as confounds: Weight, Height, BPSystolic, BPDiastolic, HbA1C, the third cube of FS_BrainSeg_Vol, the third cube of FS_IntraCanial_Vol, the average of the absolute as well as the relative value of the root mean square of the head motion, squares of all of the above, and an indicator variable for whether an earlier of later software version was used for MRI preprocessing. Head motion and software version were only included in the analysis of fMRI vs behavioral data, not in the analysis of dMRI vs behavioral data. Confounds were inverse-normal-transformed. Subsequently, missing values were set to 0. 3% and 5% of confound values were missing in the fMRI vs. behvior, and dMRI vs behavior analysis, respectively. All resulting confounds were z-scores once more.

UK Biobank (UKB) dataset

fMRI data. We utilized pre-processed resting-state fMRI data [64] from 20000 subjects, available from the UK Biobank Imaging study [2].

In brief, EPI unwarping, distortion and motion correction, intensity normalization and high-pass temporal filtering were applied to each subject's functional data using FSL's Melodic [65], data were registered to standard space (MNI), and structured artefacts are removed using ICA and FSL's FIX [53, 54, 65]. A set of resting-state networks were identified common across the cohort using a subset of subjects (≈ 4000 subjects) [64]. This was achieved by extracting the top 1200 components from a group-PCA [66] and a subsequent spatial ICA with 100 resting-state networks [65, 67]. Visual inspection revealed 55 non-artifactual ICA components. Next, these 55 group-ICA networks were dual regressed onto each subject's data to derive representative timeseries for each of the ICA components. Following the regression of the artifactual nodes for all other nodes and the subsequent removal of the artifactual nodes, the timeseries were used to compute partial correlation parcellated connectomes with a dimensionality of 55 × 55. The connectomes were z-score transformed and the upper triangle vectorized to give 1485 functional connectivity features per subject, for each of the 20000 subjects.

Behavioral measures. The UK Biobank contains a wide range of subject measures [68], including physical measures (e.g., weight, height), food and drink, cognitive phenotypes, lifestyle, early life factors and sociodemographics. We hand-selected a subset of 389 cognitive, lifestyle and physical measures, as well as early life factors. For categorical items, we replaced negative values with 0, as in [2]. Such negative values encode mostly "Do not know" / "Prefer not to answer". Measures with multiple visits were then averaged across visits, reducing the number of measures 226. We then performed a check for measures that had missing values in more than 50% of subjects and also for measures that had identical values in at least 90% of subjects; no measures were removed through these checks. We then performed a redundancy check. Specifically, if the correlation between any two measures was > 0.98, one of the two items was randomly chosen and dropped. This procedure further removed 2 measures, resulting in a final set of 224 behavioral measures, available for each of the 20000 subjects.

Confounds. We used the following items as confounds: acquisition protocol phase (due to slight changes in acquisition protocols over time), scaling of T1 image to MNI atlas, brain volume normalized for head size (sum of grey matter and white matter), fMRI head motion, fMRI signal-to-noise ratio, age, sex. In addition, similar to [2] we used the squares of all non-categorical items (i. e. T1 to MNI scaling, brain volume, fMRI head motion, fMRI signal-to-noise ratio and age), as well as age × sex and age^2× sex. Altogether these were 14 confounds. 6% of values were missing and set to 0. All resulting confounds were z-scores across subjects.

Preprocessing of empirical data for CCA and PLS

We prepared data for CCA following, for the most part, the pipeline in [3].

Deconfounding. Deconfounding of a matrix X with a matrix of confounds C was performed by subtracting linear predictions, i.e.

$$X_{\text{deconfounded}} = X - C\beta \tag{1}$$

where

$$\beta = C^{+}X = \left(C^{\mathsf{T}}C\right)^{-1} C^{\mathsf{T}}X \tag{2}$$

The confounds used were specific to each dataset and mentioned in the previous section.

Neuroimaging data. Neuroimaging measures, were, on the one hand, z-scored. On the other hand, normalized values were used as additional features: normalization was performed by calculating features' absolute value of the mean across subjects and, in case this mean was above 0.1 (otherwise this feature was not used in normalized form), the original values of the feature were divided by this mean, and the resulting values were z-scored across subjects.

The resulting data matrix was de-confounded (as described in the previous above), decomposed into principle components via a singular value decomposition, and the left singular vectors, multiplied by their respective singular values were used as data matrix X in the subsequent CCA or PLS analysis.

Behavioral and demographic data. The list of used behavioral items were specific to each dataset and mentioned in the previous sections. Given this list, separately for each item, a rank-based inverse normal transformation [69] was applied and the result z-scored. For both of these steps subjects with missing values were disregarded. Next, a subjects $\times$ subjects covariance matrix across variables was computed, considering for each pair of subjects only those variables that were present for both subjects. The nearest positive definite matrix of this covariance matrix was computed using the function `cov_nearest` from the Python `statsmodels` package [70]. This procedure has the advantage that subjects can be used without the need to impute missing values. An eigenvalue decomposition of the resulting covariance matrix was performed where the eigenvectors, scaled to have standard deviation 1, are principal component scores. They are then scaled by the square-roots of their respective eigenvalues (so that their variances correspond to the eigenvalues) and used as matrix Y in the subsequent CCA or PLS analysis.

Generating synthetic data for CCA and PLS

We analyzed properties of CCA and PLS with simulated datasets from a multivariate generative model. These datasets are be drawn from a normal distribution with mean 0 and a covariance matrix Σ that encodes assumed relationships in the data. To specify Σ we need to specify relationships of features within X, i. e. the covariance matrix $\Sigma_{XX} \in \mathbb{R}^{p_x \times p_x}$, relationships of features within Y, i. e. the covariance matrix $\Sigma_{YY} \in \mathbb{R}^{p_y \times p_y}$, and relationships between features in X on the one side and Y on the other side, i.e. the matrix $\Sigma_{XY} \in \mathbb{R}^{p_x \times p_y}$. Together, these three covariance matrices form the joint covariance matrix (Fig. 1D)

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY}^{\mathsf{T}} & \Sigma_{YY} \end{pmatrix} \in \mathbb{R}^{p_x+p_y \times p_x+p_y} \tag{3}$$

for X and Y and this allows us to generate synthetic datasets by sampling from the associated normal distribution $\mathcal{N}(0, \Sigma)$.

The covariance matrices Σ_{XX} and Σ_{YY}

Given a data matrix X, the features can be re-expressed in a different coordinate system through multiplication by an orthogonal matrix O: $\tilde{X} = XO$. No information is lost in this process, as it can be reversed: $X = \tilde{X}O^{\mathsf{T}}$. Therefore, we are free to make a convenient choice. We select the principal component coordinate system as in this case the covariance matrix becomes diagonal, i.e. $\Sigma_{XX} = \mathrm{diag}(\vec{\sigma}_{XX})$. Analogously, for Y we choose the principal component coordinate system such that $\Sigma_{YY} = \mathrm{diag}(\vec{\sigma}_{YY})$.

For modeling, to obtain a concise description of $\vec{\sigma}_{XX}$ and $\vec{\sigma}_{YY}$ we assume a power-law such that $\sigma_{XX,i} = c_{XX} i^{-a_{XX}}$ and $\sigma_{YY,i} = c_{YY} i^{-a_{YY}}$ with decay constants a_{XX} and a_{YY} (Fig. 1B). Unless a match to a specific dataset is sought, the scaling factors c_{XX} and c_{YY} can be set to 1 as they would only rescale all results without affecting conclusions.

The cross-covariance matrix Σ_{XY}

The between-set covariance matrix Σ_{XY} encodes relationships between the datasets X and Y. One such relationship is completely specified if we are given the weights of the variables in each dataset, $\vec{w}_X$ and $\vec{w}_Y$, and the association strength of the resulting weighted composite scores.

For PLS, the relation between the between-set covariance matrix, the weight vectors and association strengths is given by

$$\Sigma_{XY} = W_X \mathrm{diag}(\vec{\sigma}_{XY}) W_Y^{\mathsf{T}} \qquad \text{(for PLS)} \tag{4}$$

where the m columns of W_X and W_Y contain the weight vectors for the m between-set modes, $W_X^{\mathsf{T}} W_X = \mathbb{1}_m$, $W_Y^{\mathsf{T}} W_Y = \mathbb{1}_m$ and $\vec{\sigma}_{XY}$ are the covariances of the composite scores. Arguably, correlations are more accessible to intuition though and we therefore re-express $\vec{\sigma}_{XY}$ in terms of the assumed true (canonical) correlations. For each mode with weights $\vec{w}_X$ and $\vec{w}_Y$ and covariance σ_{XY} we have

$$\sigma_{XY} = r_{\text{true}} \sqrt{\mathrm{var}\left(X\vec{w}_X\right) \mathrm{var}\left(Y\vec{w}_Y\right)} \tag{5}$$

where $\mathrm{var}\,(X\vec{w}_X) = \vec{w}_X^\mathsf{T}\Sigma_{XX}\vec{w}_Y$ and $\mathrm{var}\,(Y\vec{w}_Y) = \vec{w}_Y^\mathsf{T}\Sigma_{YY}\vec{w}_{XY}$ are, respectively, the variances along the X and Y composite scores.

For CCA, we have to consider the singular value decomposition of $\Sigma_{XY}^{\mathrm{CCA}} = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$:

$$\begin{aligned}\Sigma_{XY} &= \Sigma_{XX}^{1/2}\Sigma_{XY}^{\mathrm{CCA}}\Sigma_{YY}^{1/2} \\ &= \Sigma_{XX}^{1/2}\left(U\mathrm{diag}(\vec{\sigma}_{XY})V^\mathsf{T}\right)\Sigma_{YY}^{1/2} \\ &= \Sigma_{XX}^{1/2}\left(\Sigma_{XX}^{1/2}W_X\right)\mathrm{diag}(\vec{\sigma}_{XY})\left(\Sigma_{YY}^{1/2}W_Y\right)^\mathsf{T}\Sigma_{YY}^{1/2} \qquad (6)\end{aligned}$$

where we have used (??) and (??). Here, $\vec{\sigma}_{XY}$ are directly the assumed true correlations and, by construction, the weights matrices W_X and W_Y (with m columns, one for each mode) are constrained to satisfy the normalization $\mathbb{1}_m = U^\mathsf{T}U = (\Sigma_{XX}^{1/2}W_X)^\mathsf{T}\Sigma_{XX}^{1/2}W_X$ and analogously for W_Y. If $m = 1$ and given W_X, W_Y (i. e. they have a single column) the normalization can be obtained by scaling $\Sigma_{XX}^{1/2}W_X$ to unit-length, and analogously for Y.

Thus, in summary, to specify Σ_{XY} we select the number m of between-set association modes, for each of them the association strength in form of the assumed true correlation, and sets of weight vectors $\vec{w}_{X,i}$ and $\vec{w}_{Y,i}$ (for $1 \le i \le m$). The weight vectors for each set need to be orthonormal ($W_X^\mathsf{T}W_X = W_Y^\mathsf{T}W_Y = \mathbb{1}_m$) for PLS, while for CCA they need to satisfy $W_X^\mathsf{T}\Sigma_{XX}W_X = W_Y^\mathsf{T}\Sigma_{YY}W_Y = \mathbb{1}_m$.

Choice of weight vectors

We impose two constraints on possible weight vectors:

1. We aim to obtain association modes that explain a "large" amount of variance in the data, otherwise the resulting scores could be strongly affected by noise. The decision is based on the explained variance of only the first mode and we require that it is greater than 1/2 of the average explained variance of a principal component in the dataset, i.e. we require that

$$\vec{w}_X^\mathsf{T}\Sigma_{XX}\vec{w}_X > \frac{1}{2}\frac{\mathrm{tr}\,\Sigma_{XX}}{p_X} \qquad (7)$$

   and analogously for Y.

2. The weight vectors impact the joint covariance matrix Σ (via (3), (4) and (6)). Therefore, we require that the chosen weights result in a proper, i. e. positive definite, covariance matrix Σ.

To increase chances of finding weights that satisfy the first constraint, we compose them as a linear combination of a high-variance subspace element, and another component from the low-variance subspace. The high-variance subspace is defined as the vector space spanned by the first q_X and q_Y (for datasets X and Y, respectively) components where q_X and q_Y are chosen to explain 90% of their respective within-set variances. Having chosen (see below) any unit vectors of the low- and high-variance subspaces, $\vec{w}_{\text{lo}}$ and $\vec{w}_{\text{hi}}$, they are combined as

$$\vec{w} = c\vec{w}_{\text{hi}} + \sqrt{1-c^2}\vec{w}_{\text{lo}} \tag{8}$$

so that $\|\vec{w}\| = 1$. Here, c is a uniform random number between 0 and 1 (but see also below). If the resulting weight vectors do not satisfy the imposed constraints, new values for $\vec{w}_{\text{lo}}$, $\vec{w}_{\text{hi}}$ and c are drawn. Note that, in case the number of between-set association modes m is greater than 1, only the first one is used to test the constraint (7), but weight vectors for the remaining modes are composed in the same way as just described.

Weight vector components of the low-variance subspace are found by multiplication of its basis vectors $U_{\text{lo}} \in \mathbb{R}^{p \times p-q}$ with a rotation matrix R_{lo}

$$W_{\text{lo}} = U_{\text{lo}} R_{\text{lo}} \tag{9}$$

where the first m columns of W_{lo} are used as the low-variance subspace components of the m between-set association modes. If $q_X \geq m > p_X - q_X$ (and analogously for Y) the dimensionality of the low-variance subspace is not large enough to get a component for all m modes in this way, so that only for the first m modes a low-variance subspace component will be used.

The rotation matrix R_{lo} is found as the Q-factor of a QR-decomposition of a $p_X - q_X \times p_X - q_X$ (analogously for Y) matrix with elements drawn from a standard normal distribution.

Weight vector components of the high-variance subspace are selected in the following way (see Fig. S4). First, 10000 attempts are made to find them in the same way as the low-variance component, i.e. as the first m columns of

$$W_{\text{hi}} = U_{\text{hi}} R_{\text{hi}} \tag{10}$$

where the columns of U_{hi} are the basis vectors for the high-variance subspace, and R_{hi} is found as the Q-factor of a QR-decomposition of a $q_X \times q_X$ (analogously for Y) matrix with elements drawn from a standard normal distribution. In case this fails (i. e. if one of the two constraints is not satisfied for all 10000 attempts), another 10000 attempts are made in which the coefficient c is not chosen randomly between 0 and 1, but the lower bound is increased stepwise from 0.5 to 1 to make it more likely that the first constraint is satisfied.

If this also fails (which tends to happen for large ground truth correlations r_{true} and large dimensionalities p_X and p_Y), and if $m = 1$, a differential evolution algorithm [71] is used to maximize the minimum eigenvalue of Σ, in order to encourage the second constraint to be satisfied. Specifically, q_X coefficients $\vec{c}_X$ and q_Y coefficients $\vec{c}_Y$ are optimized such that the weights $\vec{w}_X = U_{X,\text{hi}}\vec{c}_X$ and $\vec{w}_Y = U_{Y,\text{hi}}\vec{c}_Y$ satisfy the constraints. As soon as the minimum eigenvalue of a resulting Σ matrix is above 10^{-5} the optimization is stopped. 10000 attempts are made to add a low-variance component to the optimized high-variance component in this way, and if unsuccessful, another 10000 attempts are made in which the coefficient c is not chosen randomly between 0 and 1, but the lower bound is increased stepwise from 0.5 to 1.

If this also fails, and if $m = 1$, the high-variance components of the weight vectors are chosen as the first principal component axes as a fallback approach. To see why this works, recall that we have assumed to work in the principal component coordinate system so that $\vec{w}_{X,\text{hi},1} = (1, 0, \dots, 0)^\top$, $\vec{w}_{Y,\text{hi},1} = (1, 0, \dots, 0)^\top$ and Σ_{XX} as well as Σ_{YY} are diagonal. In addition, we assume that the principal component variances are normalized such that the highest (i. e. the top-left entry in Σ_{XX} and Σ_{YY}) is 1. We are seeking weight vectors that result in a positive definite covariance matrix Σ and Σ is positive definite if and only if both Σ_{YY} and the Schur complement of Σ, i. e. $\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}\Sigma_{XY}^\top$, are positive definite. Σ_{YY} is positive definite by construction. The between-set covariance matrix here is $\Sigma_{XY} = \sigma_{XY,1}\vec{w}_{X,\text{hi},1}\vec{w}_{Y,\text{hi},1}^\top$. For CCA, $\sigma_{XY,1}$ is the canonical correlation $r_{\text{true}} < 1$. For PLS, $\sigma_{XY,1} = r_{\text{true}}\sqrt{\operatorname{var} X\vec{w}_X \operatorname{var} Y\vec{w}_Y}$, which, with the specific choices of $\Sigma_{XX}, \Sigma_{YY}, \vec{w}_X$ and $\vec{w}_Y$ just described, also simplifies to $\sigma_{XY,1} = r_{\text{true}}$. Thus, $\Sigma_{XY}\Sigma_{YY}\Sigma_{XY}^\top = r_{\text{true}}^2(1, 0, \dots, 0)^\top(1, 0, \dots, 0)$ and consequently the diagonal entries of $\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}\Sigma_{XY}^\top$ are all greater than 0. That shows that Σ is positive definite if the weights are chosen as the first principal component axes. To not end up with the pure principal component axes in all cases, we add a low-variance subspace component as before, i. e. we make 10000 attempts to add a low-variance component with weight c chosen uniformly at random between 0 and 1, and, if unsuccessful, another 10000 attempts in which the lower bound for c is increased stepwise from 0.5 to 1.

Summary

Thus, to generate simulated data for CCA and PLS, we vary the assumed between-set correlation strengths $\vec{\rho}_{XY}$, setting them to select levels, while choosing random weights W_X and W_Y. The columns of the weight matrices W_X and W_Y must be mutually orthonormal for PLS, while for CCA they must satisfy $W_X^\top \Sigma_{XX} W_X = W_Y^\top \Sigma_{YY} W_Y = \mathbb{1}_m$. In addition, we assume that the weight vectors are contained within a subspace of, respectively, q_X and q_Y dominant principal components.

Performed simulations

For Figs. 2, 3C-D, the colored curves in Fig. 6A, Figs. S10E-F, 7, S8A-D, and the left 3 columns of Fig. S5, we ran simulations for $m = 1$ between-set association mode assuming true correlations of 0.1, 0.3, 0.5, 0.7 and 0.9, used dimensionalities $p_X = p_Y$ of 2, 4, 8, 16, 32, and 64 as well as 25 different covariance matrices. $a_X + a_Y$ was fixed at -2. 100 synthetic datasets were drawn from each instantiated normal distribution. Where not specified otherwise, null distributions were computed with 1000 permutations. Due to computational expense, some simulations did not finish and are reported as blank spaces in heatmaps.

Similar parameters were used for other figures, except for the following deviations.

For Fig. 3A-B p_X was 100, $r_{\text{true}} = 0.3$, $a_X = a_Y = -1$ and we used 1 covariance matrix for CCA and PLS.

For Fig. 3E-F p_X was 100, $r_{\text{true}} = 0.3$ and we used 100 different covariance matrices.

For Fig. 3G-H, p_X was 2, $r_{\text{true}} = 0.3$, $a_X = a_Y = -1$ and we used 10000 different covariance matrices for CCA and PLS.

For Fig. 3I-L, we used 2, 4, 8, 16, 32 and 64 for p_X, 0.1, 0.3 and 0.5 for r_{true}, 10 different covariance matrices for CCA and PLS, and 10 permutations. A subset of these, namely $p_X = 64$ and $r_{\text{true}} = 0.3$ was used for Fig. 3I-J.

For Fig. 6, we varied r_{true} from 0 to 0.99 in steps of 0.01 for each combination of p_X and p_Y for which we have a study in our database of reported CCAs, assumed $a_X = a_Y = 0$, and generated 1 covariance matrix for each r_{true}.

For the right 3 columns in Fig. S5 $p_X + p_Y$ was fixed at 64 and for p_X we used 2, 4, 8, 16, 32.

In Fig. S8O-P, for p_X we used 4, 8, 16, 32, 64, we generated 10 different covariance matrices for both CCA and PLS and varied r_{true} from 0 to 0.99 in steps 0.01.

For Fig. S9A-F we used 2, 4, 8, 16 and 32 for p_X, and 10 different covariance matrices for both CCA and PLS.

For Fig. S9G-N we used 2, 4, 8, 16, 32 and 64 for p_X, 5 different covariance matrices for both CCA and PLS, 100 bootstrap iterations and did not run simulations for $r_{\text{true}} = 0.1$.

For Fig. S8E-L, Fig. S11, and Fig. S12 we used 75 different covariance matrices. For each instantiated joint covariance matrix, $a_X + a_Y$ was chosen uniformly at random between -3 and 0 and a_X was set to a random fraction of the sum, drawn uniformly between 0 and 1.

In Fig. S8M-N we used 0.3, 0.5, 0.7 and 0.9 for r_{true}, 4, 8, 16, 32 and 64 for p_X, 6 different covariance matrices and 100 permutations.

Meta-analysis of prior literature

A PubMed search was conducted on December 23, 2019 using the query `("Journal Article"[Publication Type]) AND (fmri[MeSH Terms] AND brain[MeSH Terms]) AND ("canonical correlation analysis")` with filters requiring full text availability and studies in humans. In addition, studies known to the authors were considered. CCA results were included in the meta-analysis if they related a neuroimaging derived measures (e. g. structural or functional MRI, ...) to behavioral or demographic measures (e. g. questionnaires, clinical assessments ...) across subjects, if they reported the number of subjects and the number of features of the data entering the CCA analysis, and if they reported the observed canonical correlation. This resulted in 100 CCA analyses reported in 31 publications , which are summarized in SI Dataset 1.

The *gemmr* software package

We provide an open-source Python package, called *gemmr*, that implements the generative modeling framework presented in this paper `https://github.com/murraylab/gemmr`. Among other functionality, it provides estimators for CCA, PLS and sparse CCA; it can generate synthetic datasets for use with CCA and PLS using the algorithm laid out above; it provides convenience functions to perform sweeps of the parameters on which the generative model depends; it calculates required sample sizes to bound power and other error metrics as described above. For a full description, we refer to the package's documentation.

Statistical Analysis

Evaluation of sampling error

We use five metrics to evaluate the effects of sampling error on CCA and PLS analyses.

Statistical power. Power measures the capability to detect an existing association. It is calculated when the true correlation is greater than 0 as the probability across 100 repeated draws of synthetic datasets from the same normal distribution that the observed association strength (i. e. correlation for CCA, covariance for PLS) of a dataset is statistically significant. Significance is declared if the p-value is below $\alpha = 0.05$. The

p-value is evaluated as the probability that association strengths are greater in the null-distribution of association strengths. The corresponding null-distribution is obtained from performing CCA or PLS on 1000 datasets where the rows of Y were permuted randomly. Power is bounded between 0 and 1 and, unlike for the other metrics (see below), higher values are better.

Relative error in between-set covariance. The relative error of the between-set association strength is calculated as

$$\Delta r = \frac{\hat{r} - r}{r} \tag{11}$$

where r is the true between-set association strength and $\hat{r}$ is its estimate in a given dataset.

Weight error. Weight error Δw is calculated as 1 - absolute value of cosine similarity between observed ($\hat{\vec{w}}$) and true ($\vec{w}$) weights, separately for data sets X and Y, and the greater of the two errors is taken:

$$\Delta w = \max_{s \in \{X,Y\}} \left(1 - |\operatorname{cossim}(\hat{\vec{w}}_s, \vec{w}_s)|\right) \tag{12}$$

where

$$\operatorname{cossim}(\hat{\vec{w}}_s, \vec{w}_s) = \frac{\hat{\vec{w}}_s \cdot \vec{w}_s}{\|\hat{\vec{w}}_s\| \|\vec{w}_s\|}. \tag{13}$$

The absolute value of the cosine-similarity is used due to the sign ambiguity of CCA and PLS.

This error metric is bounded between 0 and 1 and measures the cosine of the angle between the two unit vectors $\hat{\vec{w}}_s$ and $\vec{w}_s$.

Score error. Score error Δt is calculated as 1 − absolute value of Spearman correlation between observed and true scores. The absolute value of the correlation is used due to the sign ambiguity of CCA and PLS. As for weights, the maximum over datasets X and Y is selected:

$$\Delta t = \max_{s \in X,Y} \left(1 - |\operatorname*{rankcorr}_i \left(\hat{t}_{s,i}^{(\text{test})}, t_{si}^{(\text{test})}\right)|\right) \tag{14}$$

Each element of the score vector represents a sample (subject). Thus, to be able to compute the correlation between estimated ($\hat{\vec{t}}$) and true ($\vec{t}$) score vectors, corresponding elements must represent the same sample, despite the fact that in each repetition new data matrices are drawn in which the samples have completely different identities. To overcome this problem and to obtain scores, which are comparable across repetitions (denoted $\hat{\vec{t}}^{(\text{test})}$ and $\vec{t}^{(\text{test})}$), each time a set of data matrices is drawn from a given distribution $\mathcal{N}(0, \Sigma)$ and a CCA or PLS model is estimated, the resulting model (i. e. the resulting weight vectors) is also applied to a "test" set of data matrices, $X^{(\text{test})}$ and $Y^{(\text{test})}$ (of the same size as X and Y) obtained from $\mathcal{N}(0, \Sigma)$ and common across repeated dataset draws.

The score error metric Δt is bounded between 0 and 1 and reflects the idea that samples (subjects) might be selected on the basis of how extreme they score and that the ordering of samples (subjects) is more important than the somewhat abstract value of their scores.

Loading error. Loading error $\Delta\ell$ is calculated as (1 − absolute value of Pearson correlation) between observed and true loadings. The absolute value of the correlation is used due to the sign ambiguity of CCA and PLS. As for weights, the maximum over datasets X and Y is selected:

$$\Delta\ell = \max_{s \in X,Y} \left(1 - |\operatorname*{corr}_i \left(\hat{\ell}_{s,i}^{(\text{test})}, \ell_{s,i}^{(\text{test})}\right)|\right) \tag{15}$$

True loadings are calculated with (**??**) (replacing the sample covariance matrix in the formula with its population value). Estimated loadings are obtained by correlating data matrices with score vectors ((**??**)). Thus, the same problem as for scores occurs: the elements of estimated and true loadings must represent the same sample. Therefore, we calculate loading errors with loadings obtained from test data ($X^{(\text{test})}$ and $Y^{(\text{test})}$) and test scores ($\hat{\vec{t}}^{(\text{test})}$ and $\vec{t}^{(\text{test})}$) that were also used to calculate score errors.

The loading error metric $\Delta\ell$ is bounded between 0 and 1 and reflects the idea that loadings measure the contribution of original data variables to the between-set association mode uncovered by CCA and PLS.

Loadings are calculated by correlating scores with data matrices. Of note, all synthetic data matrices in this study are based in the principal component coordinate system. In practice, however, this is not generally the case. Nonetheless, as the transformation between principal component and original coordinate system cannot be constrained, we here do not consider this effect.

Weight similarity to principal component axes

The directional means μ in Figs. 4A-B are obtained via

$$R = \frac{1}{n_\alpha} \sum_j^{n_\alpha} \mathrm{e}^{2i\alpha_j} \tag{16}$$

as $\mu = \arg(R)/2$.

To interpret the distribution of cosine similarities between weights and the first principal component axis we compare this distribution to a reference, namely to the distribution of cosine-similarities between a random n-dimensional unit vector and an arbitrary other unit vector $\vec{e}$. This distribution f is given by:

$$f_n(x) = \frac{\mathrm{d}\, P(X \leq x)}{\mathrm{d}\, x} \tag{17}$$

where P denotes the cumulative distribution function for the probability that a random unit-vector has cosine-similarity with $\vec{e}$ (or, equivalently, projection onto $\vec{e}$) $\leq x$. For $-1 \leq x \leq 0$, P can be expressed in terms of the surface area $A_n(h)$ of the n-dimensional hyperspherical cap of radius 1 and height h (i. e. $x - h = -1$)

$$P(X \leq x) = \frac{A_n(h)}{A_n(2)} \tag{18}$$

where $A_n(2)$ is the complete surface area of the hypersphere and

$$A_n(h) = \frac{1}{2} A_n(2) I\left(h(2-h); \frac{n-1}{2}, \frac{1}{2}\right) \tag{19}$$

and I is the regularized incomplete beta function. Thus,

$$f_n(x) = \frac{1}{2}\frac{\mathrm{d}\, I}{\mathrm{d}\, x}\left((x+1)(1-x); \frac{n-1}{2}, \frac{1}{2}\right) \tag{20}$$

$$= \frac{1}{2}\frac{1}{B(\frac{n-1}{2}, \frac{1}{2})}(1-x^2)^{\frac{n-3}{2}}\left(x^2\right)^{-1/2}(-2x) \tag{21}$$

$$= \frac{1}{B(\frac{n-1}{2}, \frac{1}{2})}(1-x^2)^{\frac{n-3}{2}} \tag{22}$$

where B is a beta function and

$$f_n(2\tilde{x}-1) \propto (2-2\tilde{x})^{\frac{n-1}{2}-1}(2\tilde{x})^{\frac{n-1}{2}-1} \tag{23}$$

$$\propto f_\beta\left(\tilde{x}; \frac{n-1}{2}, \frac{n-1}{2}\right) \tag{24}$$

where f_β is the probability density function for the beta distribution. Hence, $2\tilde{X} - 1$ with 795
$\tilde{X} \sim Beta(\frac{n-1}{2}, \frac{n-1}{2})$ is a random variable representing the cosine similarity between 2 random vectors (or, 796
equivalently, the projection of a random unit-vector onto another). 797

CCA/PLS analysis of empirical data

Permutation-based p-values in Fig. 5 and S2 were calculated as the probability that the CCA or PLS association strength of permuted datasets was at least as high as in the original, unpermuted data. Specifically, to obtain the p-value, rows of the behavioral data matrix were permuted and each resulting permuted data matrix together with the unpermuted neuroimaging data matrix were subjected to the same analysis as the original, unpermuted data, in order to obtain a null-distribution of between-set associations. 1000 permutations were used.

Due to familial relationships between HCP subjects they are not exchangeable so that not all possible permutations of subjects are appropriate [72]. To account for that, in the analysis of HCP fMRI vs behavioral data, we have calculated the permutation-based p-value as well as the confidence interval for the whole-data (but not the subsampled data) analysis using only permutations that respect familial relationships. Allowed permutations were calculated using the functions `hpc2blocks` and `palm_quickperms` with default options as described in `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/PALM/ExchangeabilityBlocks` (accessed May 18, 2020). No permutation indices were returned for 3 subjects that were therefore excluded from the functional connectivity vs behavior analysis.

Subsampled analyses (Fig. 5) were performed for 5 logarithmically spaced subsample-sizes between 202 and 50% of the total subject number. For each subsample size 100 pairs of non-overlapping data matrices were used.

Cross-validated analyses were performed with 5-fold cross-validation.

Principal component spectrum decay constants

The decay constant of a principal component spectrum (Fig. S1A-J) was estimated as the slope of a linear regression (including an intercept term)å of log(explained variance of a principal component) on log(principal component number). For each dataset in Fig. S1A-J we included as many principal components into the linear regression as necessary to explain either 30% or 90% of the variance.

Determination of required sample size

As all evaluation metrics change approximately monotonically with sample per feature, we fit splines of degree 3 to interpolate and to determine the number of samples per feature that approximately results in a given target level for the evaluation metric. For power (higher values are better) we target 0.9, for all the other metrics (lower values are better) we target 0.1. Before fitting the splines, all samples-per-feature are log-transformed and metrics are averaged across repeated datasets from the same covariance matrix. Sometimes the evaluation metrics show non-monotonic behavior and in case the cubic spline results in multiple roots we filter those for which the spline fluctuates strongly in the vicinity of the root (suggesting noise), and select the smallest remaining root $\tilde{n}$ for which the interpolated metric remains within the allowed error margin for all simulated $n > \tilde{n}$, or discard the synthetic dataset if all roots are filtered out. In case a metric falls within the allowed error margin for all simulated n (i. e. even the smallest simulated n_0) we pick n_0.

We suggest, in particular, a *combined* criterion to determine an appropriate sample size. This is obtained by first calculating sample-per-feature sizes with the interpolation procedure just described separately for the metrics power, relative error of association strength, weight error, score error and loading error. Then, for each parameter set, the maximum is taken across these five metrics.

Sample-size calculator for CCA and PLS

Estimating an appropriate sample size via the approach described in the previous section is computationally expensive as multiple potentially large datasets have to be generated and analyzed. To abbreviate this process (see also Fig. S11A) we do use the approach from the previous section to obtain sample-size estimates for $r_{\text{true}} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $p_x \in \{2, 4, 8, 16, 32, 64, 128\}$, $p_y = p_x$, and $a_x + a_y \sim \mathcal{U}(-3, 0)$, $a_x = c(a_x + a_y)$, and $c \sim \mathcal{U}(0, 1)$, where $\mathcal{U}$ denotes a uniform distribution. We then fit a linear model to the logarithms of the sample size, with predictors $\log(r_{\text{true}})$, $\log(p_x + p_y)$, $|a_x + a_y|$, and including an intercept term.

We tested the predictions of linear model using a split-half approach (Fig. S11B-F), i. e. we refitted the model using either only sample-size estimates for $r_{\text{true}} \in \{0.1, 0.3\}$ and half the values for $r_{\text{true}} = 0.5$, or the other half of the data, and tested the resulting refitted model on the remaining data in each case.

Table 1. Considerations and recommendations for using CCA and PLS in practice.

#	Keyword	Recommendation
1.	Importance of sample size and number of features	Sample size and the number of features in the dataset are of critical importance for the stability of CCA and PLS. Dimensionality reduction (e. g. PCA) is a useful preprocessing step, as long as it does not remove components correlated between sets. Methods for selecting number of components that take into account the correlation between sets have ben proposed, e. g. [25].
2.	Significance testing	A significant non-zero association does not necessarily indicate that estimated weights are reliable.
3.	Association strength error	In-sample estimates for association strengths are too high, cross-validated estimates too low, their average tended to be better.
4.	Weights & loadings	Weights and loadings estimated with too few samples are unreliable. For PLS, estimation of cross-loadings required fewer samples than loadings.
5.	PC1 similarity	In PLS, weights can appear consistently similar to the first principal component axis.
6.	Deceptive weight stability	For PLS, weights can appear stable, scattering around the first principal component axis, and converge to their true values only for very large sample sizes.
7.	Subsampling	Subsampling can be used to check stability of estimated association strengths in empirical data: similar results for varying subsample sizes indicate stability.
8.	Bootstrap	Bootstrapped estimates were useful to assess the variability of weights, but not for obtaining accurate estimates of association strengths or weights.
9.	Reporting	Number of samples, number of features (after dimensionality reduction) and obtained association strength should be reported. The within-set variance spectrum is useful as well.
10.	Required sample size	Generally, we recommend at least 50 samples per feature for CCA, more for PLS (depending on the variance spectrum). The accompanying Python package can be used to calculate recommended sample sizes for given dataset characteristics.

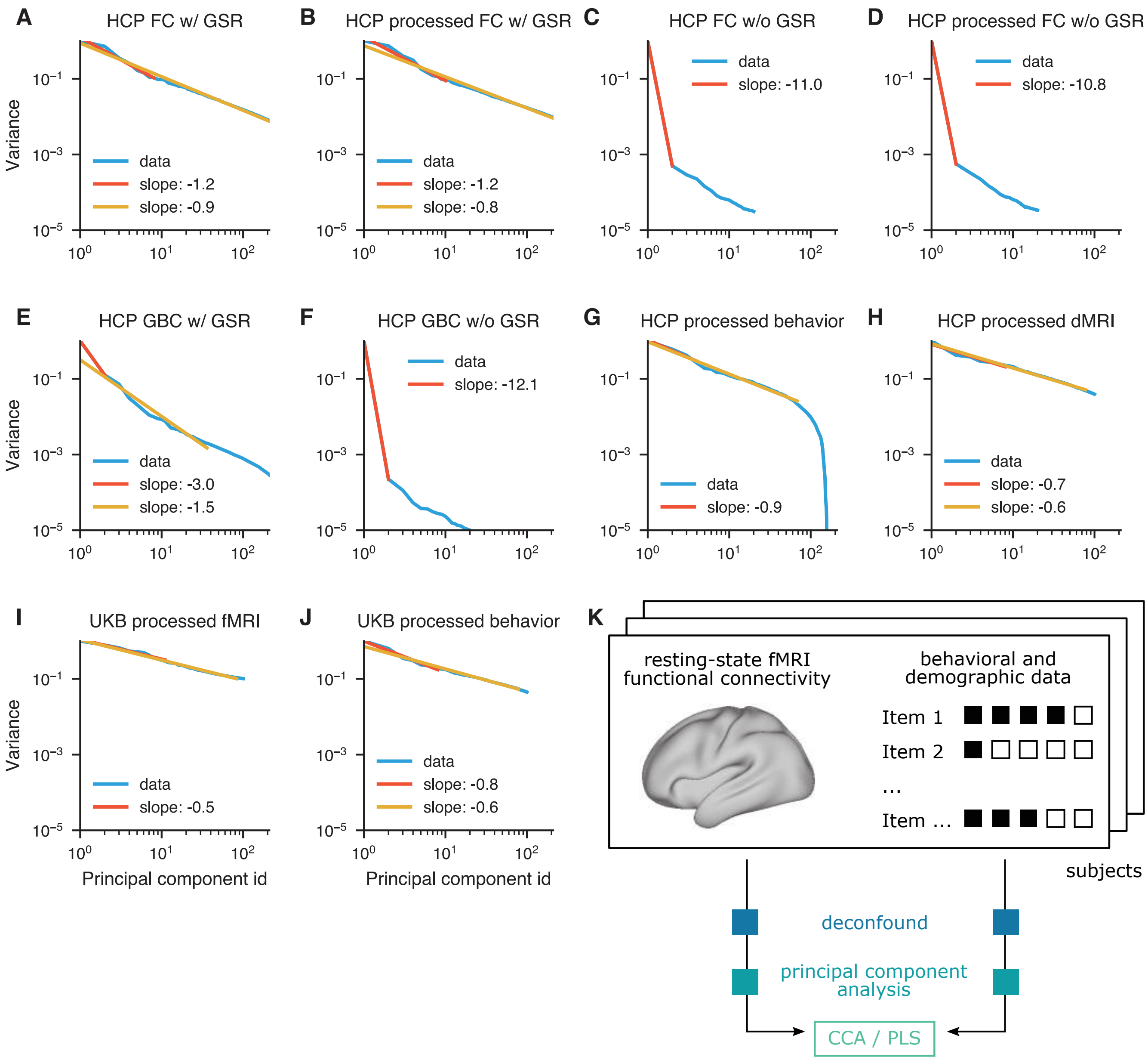

Figure S1. Supplementary analyses of empirical data. (Caption follows)

Figure S1. Supplementary results related to analysis of empirical data. A-J) Decay constants of principal component spectra in empirical data. Decay constants are estimated as the slope in a linear regression for the logarithm of the explained variance on the logarithm of the associated principal component number. We include enough components into the linear regression as necessary to explain either 30 % (red) or 90 % (yellow) of the variance. Where the two resulting slopes coincide only one is shown. Shown are decay constants for the following data matrices: **A)** HCP functional connectivity and **B)** HCP functional connectivity after preprocessing for CCA / PLS (as described in subsection), both based on 951 subjects. **C)** HCP functional connectivity for 877 subjects where global signal was not regressed out (cf. subsection) and **D)** HCP functional connectivity of 877 subjects where global signal was not regressed out after preprocessing for CCA / PLS. **E)** HCP global brain connectivity (GBC), i. e. the sum across rows of the parcel $\times$ parcel functional connectivity matrix (951 subjects) and **F)** HCP GBC where global signal was not regressed out (877 subjects). **G)** HCP behavioral data of 951 subjects after preprocessing for CCA / PLS **H)** HCP diffusion MRI of 1020 subjects after preprocessing for CCA / PLS. **I)** UK Biobank fMRI of 20000 subjects after preprocessing for CCA / PLS, **J)** UK Biobank behavioral measures of 20000 subjects after preprocessing for CCA / PLS. **K)** HCP data analysis workflow. Resting-state functional connectivity data and behavioral and demographic data from corresponding subjects were separately deconfounded, reduced to 100 principal components and then analyzed with CCA and PLS.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.25.265546; this version posted July 15, 2021. The copyright holder for this preprint (which

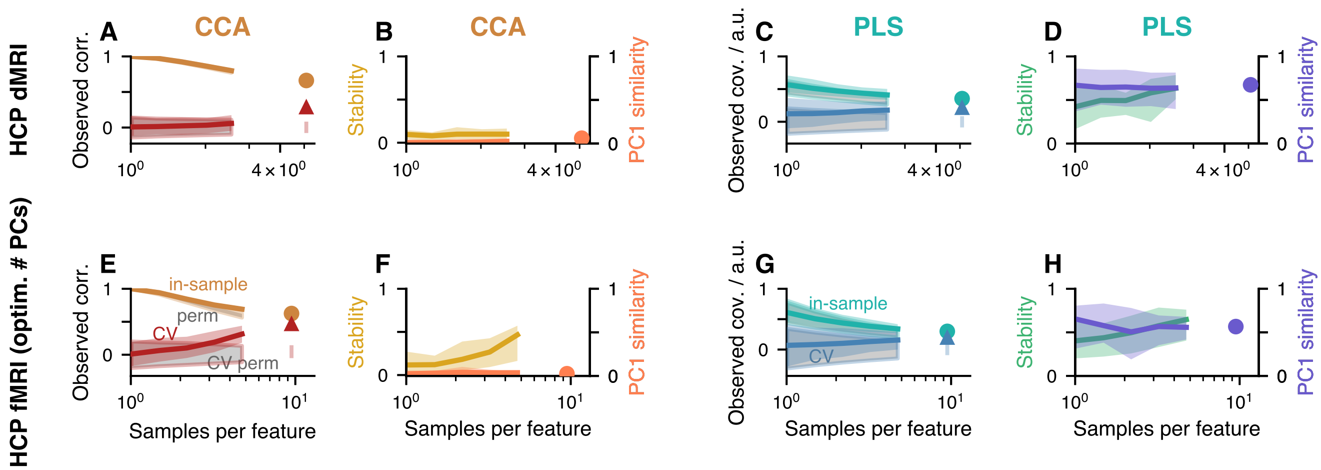

Figure S2. Additional CCA and PLS analyses of HCP data. Layout is similar to first row in Fig. 5. **A-D** HCP dMRI data was related to behavioral and demographic data. Overall, CCA and PLS behave similarly using dMRI compared to fMRI data (Fig. 5A-D). p-values in **A** and **C** were 0.001 and 0.001, respectively. **E-H)** Re-analysis of HCP fMRI vs behavior data with optimized number of principal components. Format is identical to Fig. 5. The only difference is the number of principal components retained for analysis: whereas in Fig. 5 100 principal components were used for both datasets, in agreement with previous studies of HCP data [3, 20–24], here we chose the number of principal component with the "max-min detector" from [25]. As the algorithm provided multiple values for the optimal number of components p_X (neuroimaging data) and p_Y (behavioral and demographic data), we selected here the pair that minimized $p_X + p_Y$. The optimized values were $p_X = 68$ and $p_Y = 32$, along with 13 between-set modes (we only consider the first one here). p-values for CCA and PLS were, respectively, 0.001 and 0.004. While the results are very similar to Fig. 5, (i) the observed correlations in **E)** appear to have stabilized more and are lower than in Fig. 5A, (ii) in-sample and cross-validated association strengths are more similar here in panels **A)** and **C)** than in Fig. 5, and (iii) weight similarities in **B)** and **D)** are higher than in Fig. 5. Altogether results seem to have converged more with the same sample size. This demonstrates the potential benefit of dimensionality reduction for CCA and PLS.

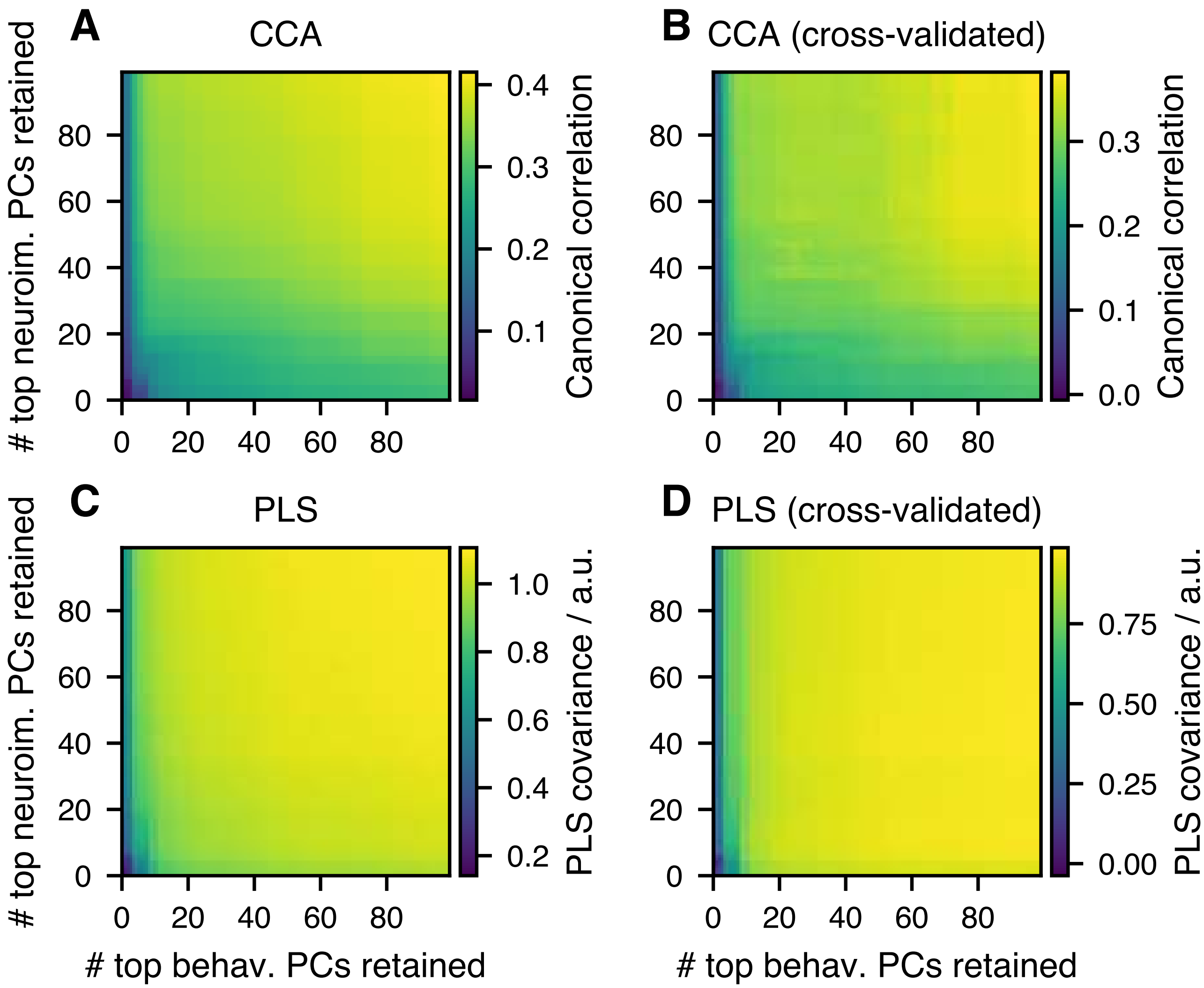

Figure S3. CCA and PLS association strength in UKB depending on retained number of principal components. **A)** In-sample and **B)** cross-validated association strength for CCA, measured as between-set correlation. **C)** In-sample and **D)** cross-validated association strength for PLS, measured as between-set covariance.

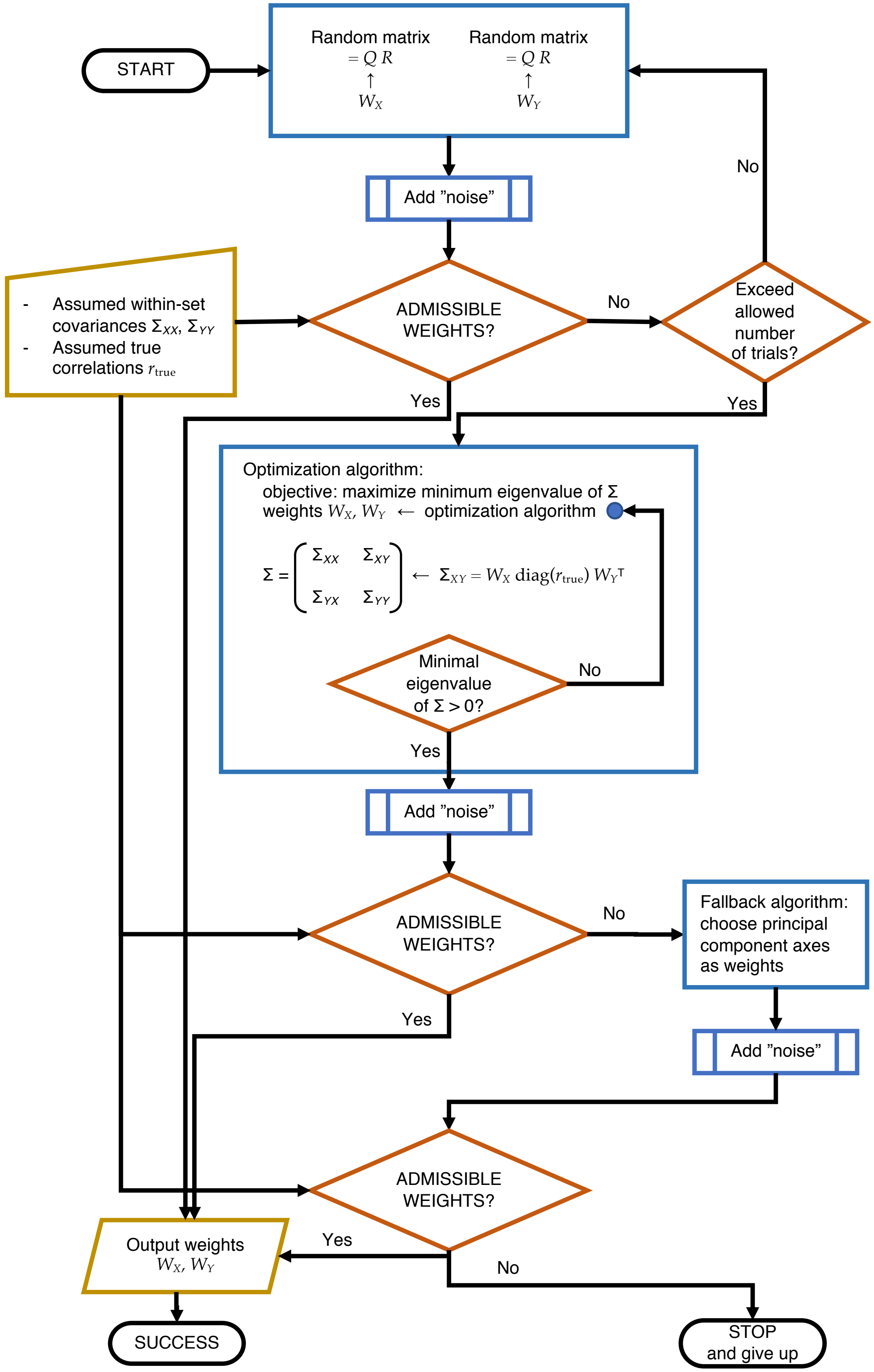

Figure S4. Algorithm for choosing weight vectors. (Caption follows)

Figure S4. Algorithm for choosing weight vectors. The flowchart illustrates the main logic of the algorithm. We require weight vectors (i) to be orthonormal within each set, (ii) to result in scores that explain at least a given fraction of variance, and (iii) to result in a proper, i. e. positive definite, joint covariance matrix Σ. Orthonormality is imposed directly when candidate weight vectors are proposed, and if the other two conditions are satisfied we say the weights are emphadmissible. In the first stage of the algorithm random weight vectors are generated as the Q factor of a QR-factorization of a matrix whose elements are drawn independently from a standard normal distribution. If this fails, an optimization algorithm is used to find weight vectors resulting in a positive definite matrix Σ. If this also fails the, the first principal component is used as first part of the weight vectors. In all three cases, after having found weight vectors in one of these ways, a component from the low-variance subspace is added, referred to in the flowchart as "noise".

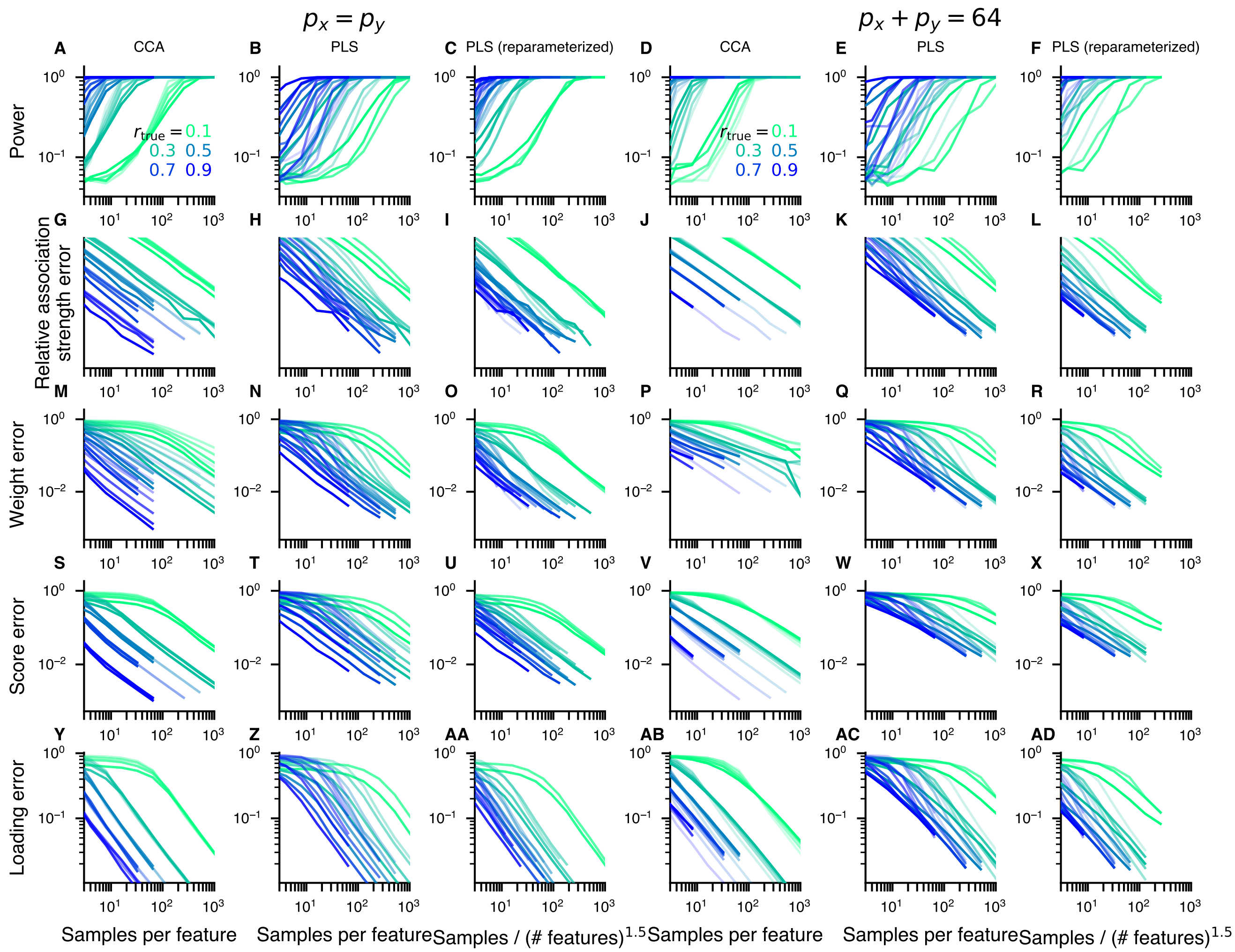

Figure S5. Samples per feature is a key effective parameter. (Caption follows)

Figure S5. Samples per feature is a key effective parameter. Throughout the manuscript we have presented results in terms of the parameter "samples per feature". Here, we demonstrate that this is, approximately, a good parameterization. Color hue represents true between-set correlation r_{true}, saturated colors are used for $p_X = 2$, and fainter colors for higher p_X. We fixed $p_X = p_Y$ in the left 3 columns, whereas we fixed $p_X + p_Y = 64$ (and thus had $p_X \neq p_Y$) in the right 3 columns. In CCA (first column), for a given r_{true}, power and error metric curves for various number of features are very similar when parameterized as "samples per feature". In PLS (second column), the same tendency can be observed, albeit the overlap between curves of the same hue (i. e. with same r_{true} but different number of features) is worse. When "samples / (number of features)$^{1.5}$" is used instead (third column), the curves overlap more. The same trends can be seen in the right 3 columns, where $p_X \neq p_Y$.

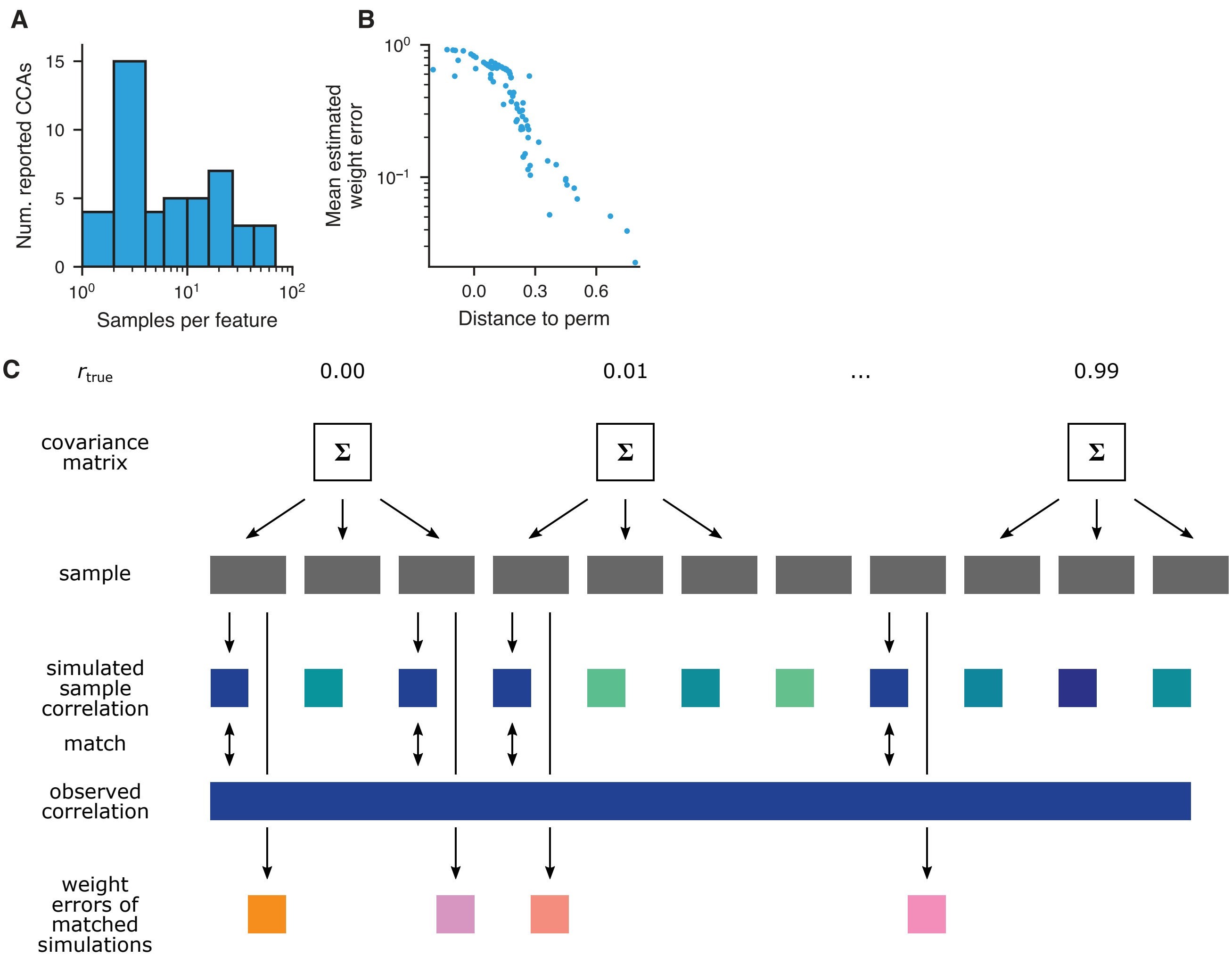

Figure S6. Supplementary results related to analysis of prior literature. (Caption follows)

Figure S6. Supplementary results related to analysis of prior literature. A) Typical number of samples per feature in brain-behavior CCAs. Studies using CCA to analyze brain-behavior relationships often used less than 5 samples per feature. Note that we here considered the number of features that entered into the CCA analysis, which, after preprocessing, can be considerably less than the "raw" number of features. **B)** Distance from null in *subjects-per-feature vs observed correlation* plot predicts weight error. A linear model was fit to the simulated, permuted data shown in Fig. 6A and for each reported CCA the orthogonal distance to the fit-line was measured and is shown here on the x-axis, with positive values indicating deviations towards the top-right corner of Fig. 6A. The mean estimated weight error for the reported CCAs is the smaller the farther away from the permuted data the CCA lies in the top-right part of the plot. **C)** Schematic for estimating weight errors for published CCA results. For each CCA from the literature in our database, synthetic data for CCA is generated with matching number of samples and features. Separate datasets are generated for assumed ground-truth between-set correlations r_{true} varying between 0 and 0.99. In each generated dataset the canonical correlation is estimated and if it is close to the value in the reported CCA, the weight error for the synthetic dataset is recorded. The distribution of recorded weight errors across assumed ground-truth correlations and repetitions of the whole process is shown in Fig. 6B and its mean in Fig. 6A.

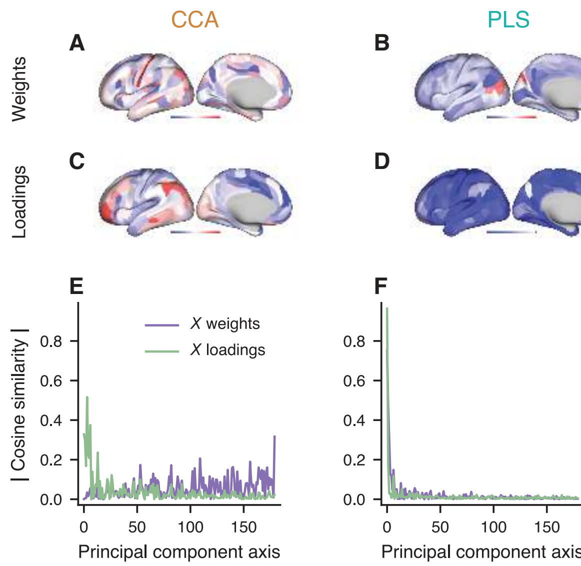

Figure S7. Weights vs loadings in real data. Using 180 fMRI-GBC features and the 5 dominant behavioral principal components as input to CCA / PLS we here illustrate GBC weights and loadings. **A** CCA weights, **B** PLS weights, **C** CCA loadings, and **D** PLS loadings. Note the relative noisiness of CCA weights. **E-F** shows a decomposition of weights and loadings into principal components, illustrating that CCA weights overlap more with low-variance PC-axes, while CCA loadings, as well as PLS weights and loadings overlap more with dominant PC-axes.

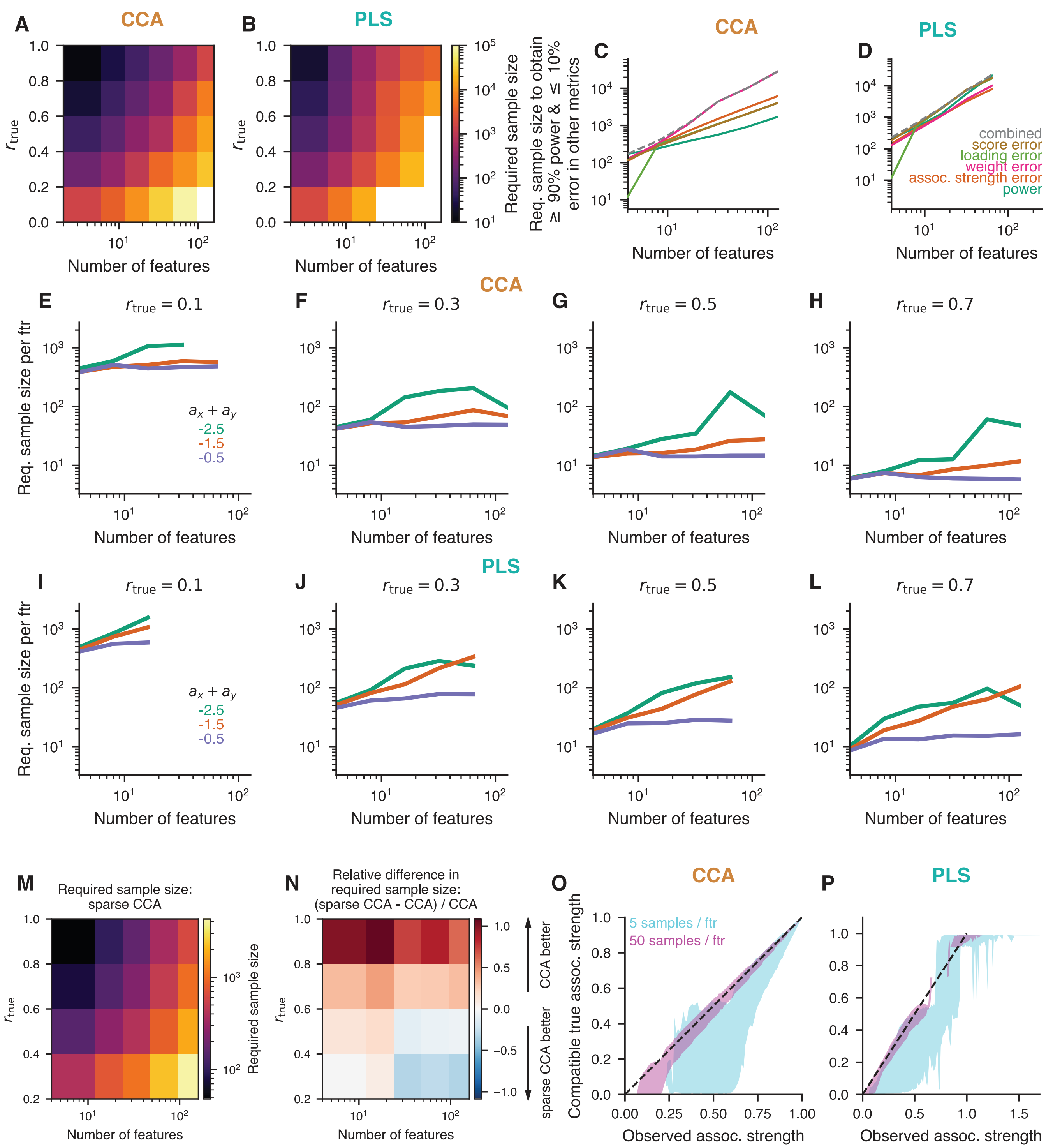

Figure S8. Parameter dependencies of required sample sizes. (Caption follows)

Figure S8. Parameter dependencies of required sample sizes. A-B) Required sample sizes based on the combined criterion increase with number of features and for low true between-set correlations r_{true}. Due to computational expense values for some parameter sets were not available (white). **C-D)** Scaling of sample-size dependence on number of features, shown here for $r_{\mathrm{true}} = 0.3$, for different metrics. **E-H)** Sample-size dependence of CCA and **I-L)** PLS on within-set variances. Simulated parameter sets were averaged across subsets having indicated values for the between-set correlation r_{true} and for $a_x + a_y$ (the sum of within-set power-law decay constants) ± 0.5. The closer $a_x + a_y$ was to 0 (i. e. the "whiter" the data) the fewer samples were required. **M-N)** Required sample size for sparse CCA. We determined required sample sizes with our analysis pipeline, for the sparse CCA variant *PMD* [27]. Due to the computational expense we ran only 6 repetitions per cell, 5 and 4, respectively, for the 2 right-most cells on the bottom. **A)** Required sample sizes increased with the number of features and with decreasing between-set correlation r_{true}. Layout is analogous to Fig. S8A-B. **B)** When the number of features was large and the true correlation r_{true} low, sparse CCA required somewhat less samples than CCA. For large r_{true}, in particular, we found the opposite. **O-P)** A wide range of true association strengths is compatible with a given observed association strength. Synthetic datasets were generated where the true correlation was varied from 0 to 0.99 in steps of 0.01 and analyzed with **O)** CCA, **P)** PLS. We investigated 4, 8, 16, 32, 64 and 128 features per set, set up 10 different covariance matrices with differing true weight vectors for each number of features and true correlation, and drew 100 repeated datasets from each corresponding normal distribution. For every CCA and PLS we recorded the observed association and binned them in bins with width 0.01. The plots show 95 % confidence intervals of the true association strength that were associated with a given observed association strength. Notably, apart from the very strongest observed association strengths which indicate an almost equally strong true correlation, compatible true association strengths can be markedly lower, down to essentially 0, when the number of used samples per feature is low.

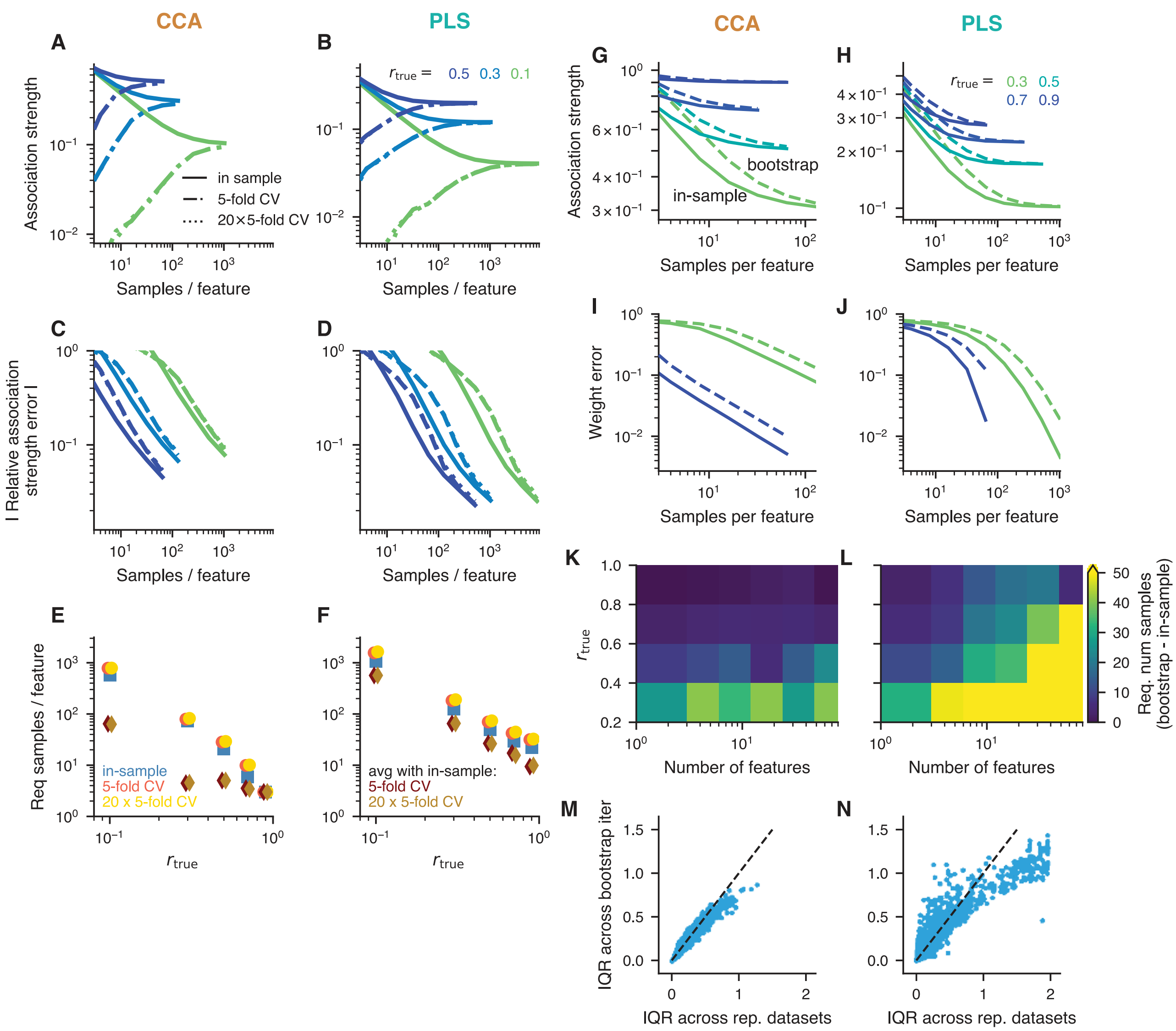

Figure S9. Cross-validated and bootstrapped estimation of association strength. (Caption follows)

Figure S9. Cross-validated and bootstrapped estimation of association strength. In contrast to in-sample estimates, cross-validated estimates (left 2 columns) of between-set association strengths underestimate the true value r_{true}. We tested two different cross-validation strategies here with very similar results (curves overlap): 5-fold cross-validation (dash-dotted line) and a strategy where the data were randomly split 20 times into 80 % train and 20 % test ("20×5-fold CV", dotted line). **C-D)** The absolute value of the relative estimation error is similar for in-sample and cross-validated estimates. **E-F)** Using the average of the in-sample and cross-validated estimates results in a better estimate than either of those, so that less samples are required to reach a target error level (here: 10 %). Bootstrapping (right 2 columns) affects CCA (3rd column) and PLS (4th column) in a similar manner. **G-H)** Bootstrapped association strengths averaged across 100 bootstrap iterations and repeated draws from a given normal distribution (dashed lines) are somewhat worse than estimates obtained from the full samples (solid lines) averaged across repetitions. Likewise, **I-J)** average weight errors and **K-L)** the number of samples required to obtain less than 10 % weight error are somewhat worse when estimated by bootstrapping. **M-N)** On the other hand, the variability of the bootstrap estimates, assessed as the interquartile range (IQR) across bootstrap iterations (and averaged across repetitions) of elements of the estimated weight vectors, match the IQR across repetitions. For each combination of the true between-set correlation $r_{\mathrm{true}} \in \{0.3, 0.5, 0.7, 0.9\}$, $p_x \in \{2, 4, 8, 16, 32, 64\}$ ($p_y = p_x$) and 5 different covariance matrices (with different true weight vectors), the scatter-plots show one dot for each element of the weight vector.

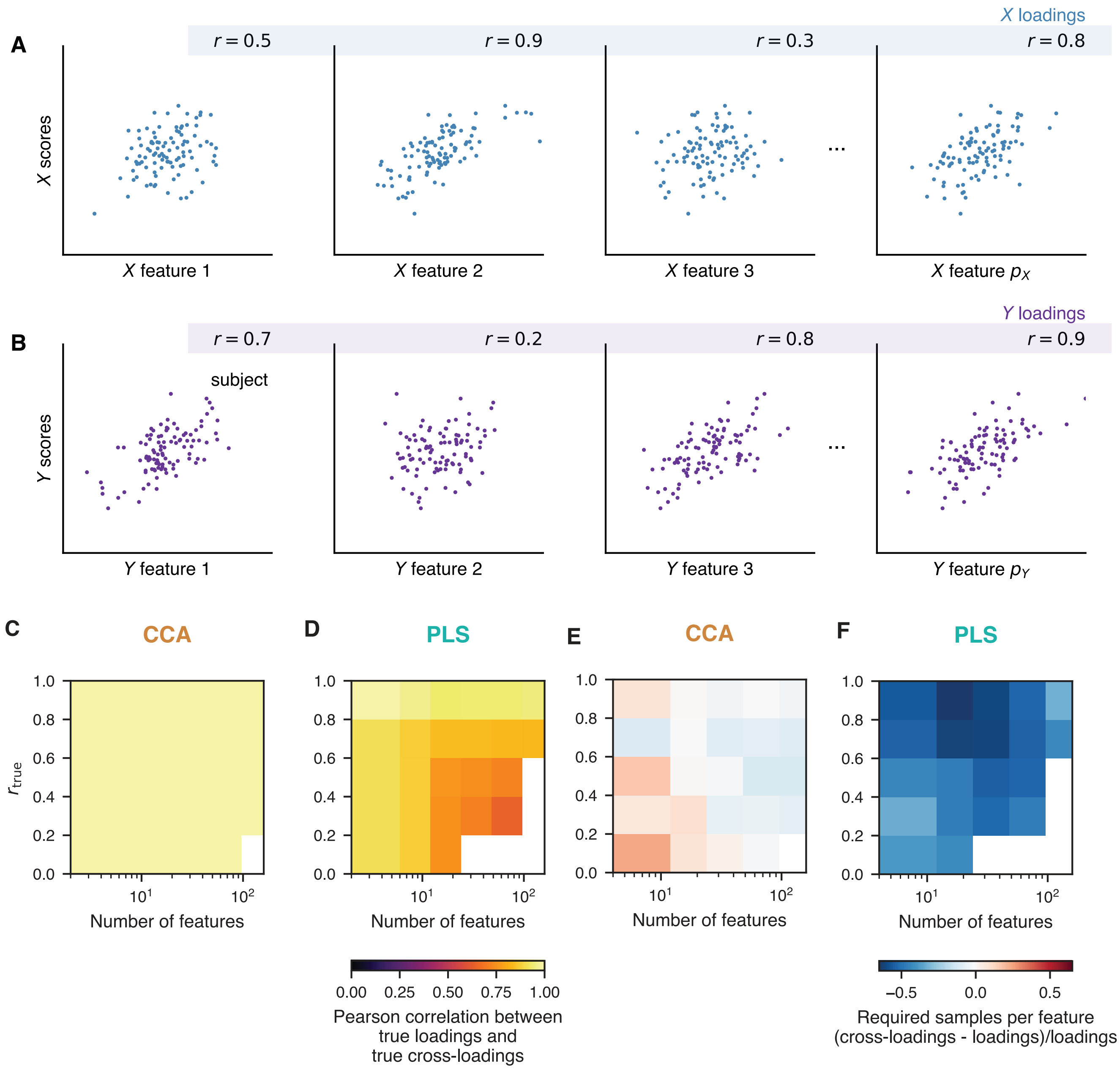

Figure S10. Loadings and cross-loadings. (Caption follows)

Figure S10. Loadings and cross-loadings. A-B) Loadings are defined as Pearson correlations across subjects of a feature with the CCA/PLS scores. The loadings vector contains these correlations for all variables. Apart from the illustrated loadings, *cross-loadings* in which scores of one set are correlated with the original features of the other set can also be computed. **C-D)**True loadings and cross-loadings were calculated with equations (**??**) and (**??**), respectively. **C)** In CCA, true loadings and true cross-loadings were collinear (as predicted by eq. (**??**)). **D)** For PLS, they were strongly correlated. The shown correlations were averaged across 25 covariance matrices with different true weight vectors. $a_x + a_y$ was constrained to -2. **E-F)** For PLS cross-loadings provide more stable estimates of feature profiles than loadings. Samples-per-feature required to obtain less than 10% error in either loadings or cross-loadings are compared. Shown here is their relative difference, i. e. the required sample-per-features for cross-loadings minus for loadings, divided by the required samples-per-feature for loadings. **E)** Relative differences were small for CCA. **F)** However, for PLS less samples were required with cross-loadings than with loadings to obtain the same error level. r_{true} indicates the true between-set correlations used in each respective simulation.

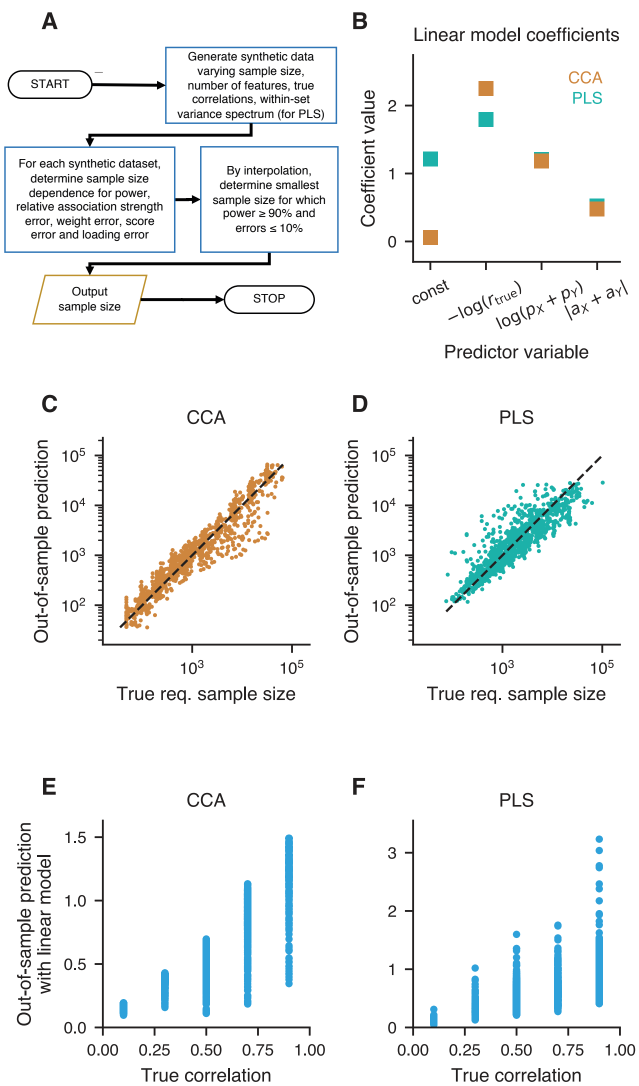

Figure S11. Sample-size calculator. (Caption follows)

Figure S11. Sample-size calculator. A) Algorithm for sample-size calculation. Sample sizes can, in principle, be calculated directly with GEMMR, as shown in Fig. 7. However, this is computationally expensive. To quickly obtain sample-size estimates, we developed the algorithm illustrated here. **B-G)** Especially for low assumed ground-truth correlations and a high number of features it is computationally expensive to estimate the required number of samples by generating synthetic datasets and searching the sample size such that error bounds are satisfied. To abbreviate this process we pre-calculate required sample sizes using the generative model approach for certain parameter values, fit a linear model to $\log(n_{\mathrm{required}})$ and then use it to quickly interpolate for parameter values not in the pre-calculated database. Predictors for the linear model are $-\log(r_{\mathrm{true}})$, $\log(p_x + p_y)$ and, for PLS only, $|a_x + a_y|$, where r_{true} indicates the true between-set correlation, p_x and p_y are the number of features in datasets X and Y, respectively, and a_x and a_y are the power-law decay constants for the within-set principal component spectrum, respectively. Shown here are linear model estimates for the required sample size based on the combined criterion, i. e. the sample sizes required to obtain 90 % power and at most 10 % error for the between-set association strength, weight, score and loading error. **B)** Linear model coefficients for CCA and PLS. **C-D)** The pre-calculated database was split in half where one half corresponded to true between-set correlations of $r_{\mathrm{true}} = 0.1$ and 0.3, the other to $r_{\mathrm{true}} = 0.7$ and 0.9 and entries for $r_{\mathrm{true}} = 0.5$ were divided between the two halves. The linear model was re-estimated separately for each half, and used to predict the other half. We obtained good predictions for CCA (**C**)) and PLS **D)**. **E, F)** Solving the linear model for r_{true}, we aim to predict correlations. We train the model using either simulation outcomes for $r_{\mathrm{true}} \in \{0.1, 0.3\}\}$, or $r_{\mathrm{true}} \in \{0.7, 0.9\}\}$ and testing the predictions on the remaining r_{true}s. **E)** Good predictions can be obtained in this way for CCA, **F)** but not for PLS.

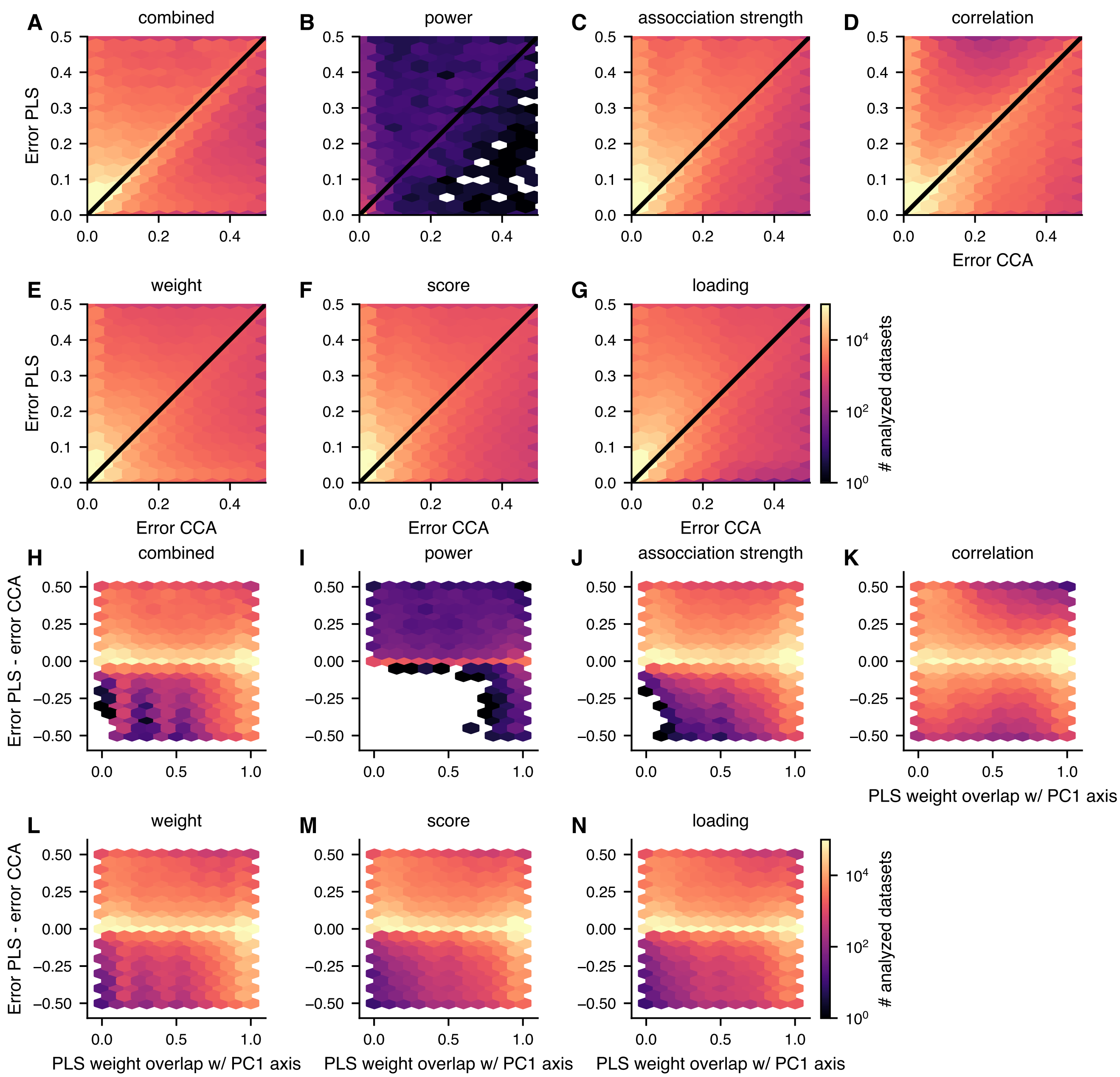

Figure S12. Required sample size for CCA vs PLS. (Caption follows)

Figure S12. Required sample size for CCA vs PLS. We instantiated joint covariance matrices (assuming 1 between-set association mode), drew samples from the associated normal distributions, and analyzed the resulting datasets with both CCA and PLS. The CCA and PLS estimations were then compared, respectively, to the true CCA and PLS solutions, which were derived from the joint covariance matrices. Panels **A)-G** show for various error metrics how resulting deviations from the truth compare between CCA and PLS. PLS errors for a given dataset tend to be larger than CCA errors in many, but not all, datasets. **H-N)** For various error metrics, when PLS has a smaller error than CCA, this tends to happen preferentially when the true PLS weight overlaps strongly with the PC1 axis. Datasets were included in these analyses if the CCA or PLS error were below 0.5.

References

1. 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`.

2. 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`.

3. Smith, S. M. *et al.* A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience* **18**, 1565–1567 (2015). URL `http://www.nature.com/neuro/journal/v18/n11/abs/nn.4125.html`.

4. Wang, H.-T. *et al.* Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. *NeuroImage* **216**, 116745 (2020). URL `https://linkinghub.elsevier.com/retrieve/pii/S1053811920302329`.

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

6. Rosipal, R. & Krämer, N. Overview and Recent Advances in Partial Least Squares. In Saunders, C., Grobelnik, M., Gunn, S. & Shawe-Taylor, J. (eds.) *Subspace, Latent Structure and Feature Selection*, Lecture Notes in Computer Science, 34–51 (Springer Berlin Heidelberg, 2006).

7. Abdi, H. & Williams, L. J. Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression. In Reisfeld, B. & Mayeno, A. N. (eds.) *Computational Toxicology*, vol. 930, 549–579 (Humana Press, Totowa, NJ, 2013). URL `http://link.springer.com/10.1007/978-1-62703-059-5_23`.

8. Le Floch, E. *et al.* Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage* **63**, 11–24 (2012). URL `http://www.sciencedirect.com/science/article/pii/S1053811912006775`.

9. Kebets, V. *et al.* Somatosensory-Motor Dysconnectivity Spans Multiple Transdiagnostic Dimensions of Psychopathology. *Biological Psychiatry* (2019). URL `http://www.sciencedirect.com/science/article/pii/S0006322319314751`.

10. Weinberg, S. L. & Darlington, R. B. Canonical Analysis when Number of Variables is Large Relative to Sample Size. *Journal of Educational Statistics* **1**, 313–332 (1976). URL `https://doi.org/10.3102/10769986001004313`.

11. Thompson, B. Finding a Correction for the Sampling Error in Multivariate Measures of Relationship: A Monte Carlo Study. *Educational and Psychological Measurement* **50**, 15–31 (1990). URL `https://doi.org/10.1177/0013164490501003`.

12. Lee, H.-S. Canonical Correlation Analysis Using Small Number of Samples. *Communications in Statistics - Simulation and Computation* **36**, 973–985 (2007). URL `https://doi.org/10.1080/03610910701539443`.

13. Dinga, R. *et al.* Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). *NeuroImage: Clinical* 101796 (2019). URL `http://www.sciencedirect.com/science/article/pii/S2213158219301469`.

14. 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).

15. Barcikowski, R. S. & Stevens, J. P. A Monte Carlo Study of the Stability of Canonical Correlations, Canonical Weights and Canonical Variate-Variable Correlations. *Multivariate Behavioral Research* **10**, 353–364 (1975). URL `https://doi.org/10.1207/s15327906mbr1003_8`.

16. Strand, K. H. & Kossman, S. Further Inquiry into the Stabilities of Standardized and Structure Coefficients in Canonical and Discriminant Analyses (New Orleans, 2000). URL https://eric.ed.gov/?id=ED572339.

17. Leach, L. & Henson, R. Bias and Precision of the Squared Canonical Correlation Coefficient Under Nonnormal Data Condition. *Journal of Modern Applied Statistical Methods* **13** (2014). URL https://digitalcommons.wayne.edu/jmasm/vol13/iss1/8.

18. Thorndike, R. M. 9 - Canonical Correlation Analysis. In Tinsley, H. E. A. & Brown, S. D. (eds.) *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, 237–263 (Academic Press, San Diego, 2000). URL http://www.sciencedirect.com/science/article/pii/B9780126913606500100.

19. Uurtio, V. *et al.* A Tutorial on Canonical Correlation Methods. *ACM Computing Surveys (CSUR)* **50**, 95:1–95:33 (2017). URL https://doi.org/10.1145/3136624.

20. Rahim, M., Thirion, B., Bzdok, D., Buvat, I. & Varoquaux, G. Joint prediction of multiple scores captures better individual traits from brain images. *NeuroImage* **158**, 145–154 (2017). URL http://www.sciencedirect.com/science/article/pii/S1053811917305438.

21. Bijsterbosch, J. D. *et al.* The relationship between spatial configuration and functional connectivity of brain regions. *eLife* **7**, e32992 (2018). URL https://elifesciences.org/articles/32992.

22. Bijsterbosch, J. D., Beckmann, C. F., Woolrich, M. W., Smith, S. M. & Harrison, S. J. The relationship between spatial configuration and functional connectivity of brain regions revisited. *eLife* **8**, e44890 (2019). URL https://elifesciences.org/articles/44890.

23. Li, J. *et al.* Topography and behavioral relevance of the global signal in the human brain. *Scientific Reports* **9**, 1–10 (2019). URL https://www.nature.com/articles/s41598-019-50750-8.

24. Han, F., Gu, Y., Brown, G. L., Zhang, X. & Liu, X. Neuroimaging contrast across the cortical hierarchy is the feature maximally linked to behavior and demographics. *NeuroImage* **215**, 116853 (2020). URL http://www.sciencedirect.com/science/article/pii/S1053811920303396.

25. Song, Y., Schreier, P. J., Ramírez, D. & Hasija, T. Canonical correlation analysis of high-dimensional data with very small sample support. *Signal Processing* **128**, 449–458 (2016). URL http://www.sciencedirect.com/science/article/pii/S0165168416300834.

26. Winkler, A. M., Renaud, O., Smith, S. M. & Nichols, T. E. Permutation inference for canonical correlation analysis. *NeuroImage* **220**, 117065 (2020). URL http://www.sciencedirect.com/science/article/pii/S1053811920305516.

27. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534 (2009). URL https://academic.oup.com/biostatistics/article/10/3/515/293026.

28. Drysdale, A. T. *et al.* Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine* **23**, 28–38 (2017). URL https://www.nature.com/articles/nm.4246.

29. Eickhoff, S., Nichols, T. E., Van Horn, J. D. & Turner, J. A. Sharing the wealth: Neuroimaging data repositories. *NeuroImage* **124**, 1065–1068 (2016). URL http://www.sciencedirect.com/science/article/pii/S1053811915010101.

30. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365–376 (2013). URL http://www.nature.com/nrn/journal/v14/n5/abs/nrn3475.html.

31. Poldrack, R. A. *et al.* Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience* **18**, 115–126 (2017). URL https://www.nature.com/articles/nrn.2016.167.

32. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **180**, 68–77 (2018). URL http://www.sciencedirect.com/science/article/pii/S1053811917305311.

33. Chen, J. *et al.* Exploration of scanning effects in multi-site structural MRI studies. *Journal of Neuroscience Methods* **230**, 37–50 (2014). URL http://www.sciencedirect.com/science/article/pii/S0165027014001393.

34. Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 1–7 (2020). URL https://www.nature.com/articles/s41586-020-2314-9. Publisher: Nature Publishing Group.

35. Bzdok, D., Engemann, D. & Thirion, B. Inference and Prediction Diverge in Biomedicine. *Patterns* 100119 (2020). URL http://www.sciencedirect.com/science/article/pii/S2666389920301604.

36. Zhuang, X., Yang, Z. & Cordes, D. A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping* hbm.25090 (2020). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25090.

37. Shmueli, G. To Explain or to Predict? *Statistical Science* **25**, 289–310 (2010). URL https://projecteuclid.org/euclid.ss/1294167961.

38. Bzdok, D. & Ioannidis, J. P. A. Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends in Neurosciences* **42**, 251–262 (2019). URL http://www.sciencedirect.com/science/article/pii/S0166223619300074.

39. Grellmann, C. *et al.* Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of MRI and genetic data. *NeuroImage* **107**, 289–310 (2015). URL http://www.sciencedirect.com/science/article/pii/S1053811914010179.

40. Rasmussen, P. M., Hansen, L. K., Madsen, K. H., Churchill, N. W. & Strother, S. C. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition* **45**, 2085–2100 (2012). URL http://www.sciencedirect.com/science/article/pii/S0031320311003906.

41. Varoquaux, G. *et al.* Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage* **145**, 166–179 (2017). URL http://www.sciencedirect.com/science/article/pii/S105381191630595X.

42. Mihalik, A. *et al.* Multiple Holdouts With Stability: Improving the Generalizability of Machine Learning Analyses of Brain–Behavior Relationships. *Biological Psychiatry* **87**, 368–376 (2020). URL http://www.sciencedirect.com/science/article/pii/S0006322319319183.

43. Akaho, S. A kernel method for canonical correlation analysis. *arXiv preprint* (2006). URL https://arxiv.org/abs/cs/0609071.

44. Andrew, G., Arora, R., Bilmes, J. & Livescu, K. Deep Canonical Correlation Analysis. In *Proceedings of the 30th International Conference on Ma- chine Learning*, vol. 28, 9 (Atlanta, Georgia, USA, 2013).

45. Michaeli, T., Wang, W. & Livescu, K. Nonparametric Canonical Correlation Analysis. *arXiv:1511.04839 [cs, stat]* (2016). URL http://arxiv.org/abs/1511.04839. ArXiv: 1511.04839.

46. Kettenring, J. R. Canonical analysis of several sets of variables. *Biometrika* **58**, 433–451 (1971). URL https://academic.oup.com/biomet/article/58/3/433/233349. Publisher: Oxford Academic.

47. Schulz, M.-A. *et al.* Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications* **11**, 1–15 (2020). URL https://www.nature.com/articles/s41467-020-18037-z. Number: 1 Publisher: Nature Publishing Group.

48. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. *Statistics in Medicine* **38**, 1262–1275 (2019). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7993. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7993.

49. Maxwell, S. E., Kelley, K. & Rausch, J. R. Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology* **59**, 537–563 (2008). URL http://www.annualreviews.org/doi/10.1146/annurev.psych.59.103006.093735.

50. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a World Beyond "p < 0.05". *The American Statistician* **73**, 1–19 (2019). URL https://doi.org/10.1080/00031305.2019.1583913. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2019.1583913.

51. Human Connectome Project. 1200 Subjects Data Release Reference. Tech. Rep. (2017). URL http://www.humanconnectome.org/documentation/S1200/.

52. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80**, 105–124 (2013). URL http://www.sciencedirect.com/science/article/pii/S1053811913005053.

53. Salimi-Khorshidi, G. *et al.* Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage* **90**, 449–468 (2014). URL http://www.sciencedirect.com/science/article/pii/S1053811913011956.

54. Griffanti, L. *et al.* ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage* **95**, 232–247 (2014). URL http://www.sciencedirect.com/science/article/pii/S1053811914001815.

55. Robinson, E. C. *et al.* MSM: A new flexible framework for Multimodal Surface Matching. *NeuroImage* **100**, 414–426 (2014). URL http://www.sciencedirect.com/science/article/pii/S1053811914004546.

56. Power, J. D. *et al.* Ridding fMRI data of motion-related influences: Removal of signals with distinct spatial and physical bases in multiecho data. *Proceedings of the National Academy of Sciences* **115**, E2105–E2114 (2018). URL http://www.pnas.org/lookup/doi/10.1073/pnas.1720985115.

57. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016). URL http://www.nature.com/nature/journal/v536/n7615/full/nature18933.html.

58. Mars, R. B. *et al.* Whole brain comparative anatomy using connectivity blueprints. *eLife* **7**, e35237 (2018). URL https://elifesciences.org/articles/35237.

59. Warrington, S. *et al.* XTRACT - Standardised protocols for automated tractography in the human and macaque brain. *NeuroImage* 116923 (2020). URL http://www.sciencedirect.com/science/article/pii/S1053811920304092.

60. Sotiropoulos, S. N. *et al.* Advances in diffusion MRI acquisition and processing in the Human Connectome Project. *NeuroImage* **80**, 125–143 (2013). URL http://www.sciencedirect.com/science/article/pii/S105381191300551X.

61. Behrens, T. E. J., Berg, H. J., Jbabdi, S., Rushworth, M. F. S. & Woolrich, M. W. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage* **34**, 144–155 (2007). URL http://www.sciencedirect.com/science/article/pii/S1053811906009360.

62. Hernandez-Fernandez, M. *et al.* Using GPUs to accelerate computational diffusion MRI: From microstructure estimation to tractography and connectomes. *NeuroImage* **188**, 598–615 (2019). URL http://www.sciencedirect.com/science/article/pii/S1053811918321591.

63. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980 (2006). URL http://www.sciencedirect.com/science/article/pii/S1053811906000437.

64. Alfaro-Almagro, F. *et al.* Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage* **166**, 400–424 (2018). URL http://www.sciencedirect.com/science/article/pii/S1053811917308613.

65. Beckmann, C. F. & Smith, S. M. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging* **23**, 137–152 (2004).

66. Smith, S. M., Hyvärinen, A., Varoquaux, G., Miller, K. L. & Beckmann, C. F. Group-PCA for very large fMRI datasets. *NeuroImage* **101**, 738–749 (2014). URL http://www.sciencedirect.com/science/article/pii/S105381191400634X.

67. Hyvärinen, A. & Oja, E. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation* **9**, 1483–1492 (1997). URL https://www.mitpressjournals.org/doi/10.1162/neco.1997.9.7.1483. Publisher: MIT Press.

68. Sudlow, C. *et al.* 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**, e1001779 (2015). URL https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001779. Publisher: Public Library of Science.

69. Beasley, T. M., Erickson, S. & Allison, D. B. Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited? *Behavior Genetics* **39**, 580 (2009). URL https://doi.org/10.1007/s10519-009-9281-0.

70. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In *9th Python in Science Conference* (2010).

71. Storn, R. & Price, K. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization* **11**, 341–359 (1997). URL https://doi.org/10.1023/A:1008202821328.

72. Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E. & Smith, S. M. Multi-level block permutation. *NeuroImage* **123**, 253–268 (2015). URL http://www.sciencedirect.com/science/article/pii/S105381191500508X.

Acknowledgments

Funding

This research was supported by NIH grants R01MH112746 (J.D.M.), R01MH108590 (A.A.), R01MH112189 (A.A.), U01MH121766 (A.A.), and P50AA012870 (A.A.); Wellcome Trust grant 217266/Z/19/Z (S.S.); a SFARI Pilot Award (J.D.M., A.A.); DFG research fellowship HE 8166/1-1 (M.H.), Medical Research Council PhD Studentship UK MR/N013913/1 (S.W.), NIHR Nottingham Biomedical Research Centre (A.M.). Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Data were also provided by the UK Biobank under Project 43822. In part, computations were performed using the University of Nottingham's Augusta HPC service and the Precision Imaging Beacon Cluster.

Author contributions

Conceptualization: MH, SW, AA, SNS, JDM. Methodology: MH, JDM. Software: MH. Formal analysis: MH, SW, AM, BR. Resources: AA, SNS, JDM. Data Curation: AM, JLJ, AH. Writing - Original Draft: MH, JDM. Writing - Review & Editing: All authors. Visualization: MH. Supervision: JDM. Project administration: JDM. Funding acquisition: AA, SNS, JDM.

Competing interests

J.L.J, A.A. and J.D.M. have received consulting fees from BlackThorn Therapeutics. A.A. has served on the Advisory Board of BlackThorn Therapeutics.

Data availability

Human Connectome Project and UK Biobank datasets cannot be made publicly available due to data use agreements. Human Connectome Project and UK Biobank are available for researchers to apply for data access. The outcomes of synthetic datasets that were analyzed with CCA or PLS are available from `https://osf.io/8expj/`.

Code availability

Our open-source Python software package, gemmr, will be freely available at `https://github.com/murraylab/gemmr`. Jupyter notebooks detailing the analyses and generation of figures presented in the manuscript will be made available as part of the package documentation.