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

Markus Helmer[a], Shaun Warrington[b], Ali-Reza Mohammadi-Nejad[b,c], Jie Lisa Ji[a,d], Amber Howell[a,d], Benjamin Rosand[e], Alan Anticevic[a,d,f], Stamatios N. Sotiropoulos[b,c,g,1], and John D. Murray[a,d,e,1]

[a]Department of Psychiatry, Yale School of of Medicine, New Haven, CT 06511; [b]Sir Peter Mansfield Imaging Centre, School of Medicine, University of Nottingham, Nottingham, NG7 2UH, United Kingdom; [c]National Institute for Health Research (NIHR) Nottingham Biomedical Research Ctr, Queens Medical Ctr, Nottingham, United Kingdom; [d]Interdepartmental Neuroscience Program, Yale University School of Medicine, New Haven, CT 06511, USA; [e]Department of Physics, Yale University, New Haven, CT 06511, USA; [f]Department of Psychology, Yale University, New Haven, CT 06511, USA; [g]FMRIB, Wellcome Centre for Integrative Neuroimaging, Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, University of Oxford, Oxford, OX3 9DU, United Kingdom

This manuscript was compiled on August 24, 2020

Associations between high-dimensional datasets, each comprising many features, can be discovered through multivariate statistical methods, like Canonical Correlation Analysis (CCA) or Partial Least Squares (PLS). CCA and PLS are widely used methods which reveal which features carry the association. Despite the longevity and popularity of CCA/PLS approaches, their application to high-dimensional datasets raises critical questions about the reliability of CCA/PLS solutions. In particular, overfitting can produce solutions that are not stable across datasets, which severely hinders their interpretability and generalizability. To study these issues, we developed a generative model to simulate synthetic datasets with multivariate associations, parameterized by feature dimensionality, data variance structure, and assumed latent association strength. We found that resulting CCA/PLS associations could be highly inaccurate when the number of samples per feature is relatively small. For PLS, the profiles of feature weights exhibit detrimental bias toward leading principal component axes. We confirmed these model trends in state-of-the-art datasets containing neuroimaging and behavioral measurements in large numbers of subjects, namely the Human Connectome Project ($n \approx 1000$) and UK Biobank ($n = 20000$), where we found that only the latter comprised enough samples to obtain stable estimates. Analysis of the neuroimaging literature using CCA to map brain-behavior relationships revealed that the commonly employed sample sizes yield unstable CCA solutions. Our generative modeling framework provides a calculator of dataset properties required for stable estimates. Collectively, our study characterizes dataset properties needed to limit the potentially detrimental effects of overfitting on stability of CCA/PLS solutions, and provides practical recommendations for future studies.

Data fusion | Multivariate associations | Canonical Correlation Analysis | Partial Least Squares | Stability | Brain-behavior associations

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 (1–3). 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 (4–6).

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) (7), whereas if covariance is used the method is called *partial least squares* (PLS) (5, 8, 9). CCA and PLS are commonly employed across scientific fields, including behavioral sciences (10), biology (11, 12), biomedical engineering (13), chemistry (14), environmental sciences (15), genomics (16), and neuroimaging (4, 17–19).

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 instabil-

Significance Statement

Scientific studies often begin with an observed association between different types of measures. When datasets comprise large numbers of features, multivariate approaches such as canonical correlation analysis (CCA) and partial least squares (PLS) are often used. These methods can reveal the profiles of features that carry the optimal association. We developed a generative model to simulate data, and characterized how obtained feature profiles can be unstable, which hinders interpretability and generalizability, unless a sufficient number of samples is available to estimate them. We determine sufficient sample sizes, depending on properties of datasets. We also show that these issues arise in neuroimaging studies of brain-behavior relationships. We provide practical guidelines and computational tools for future CCA and PLS studies.

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.

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.

[1]To whom correspondence should be addressed. E-mail: john.murray@yale.edu or stamatios.sotiropoulos@nottingham.ac.uk

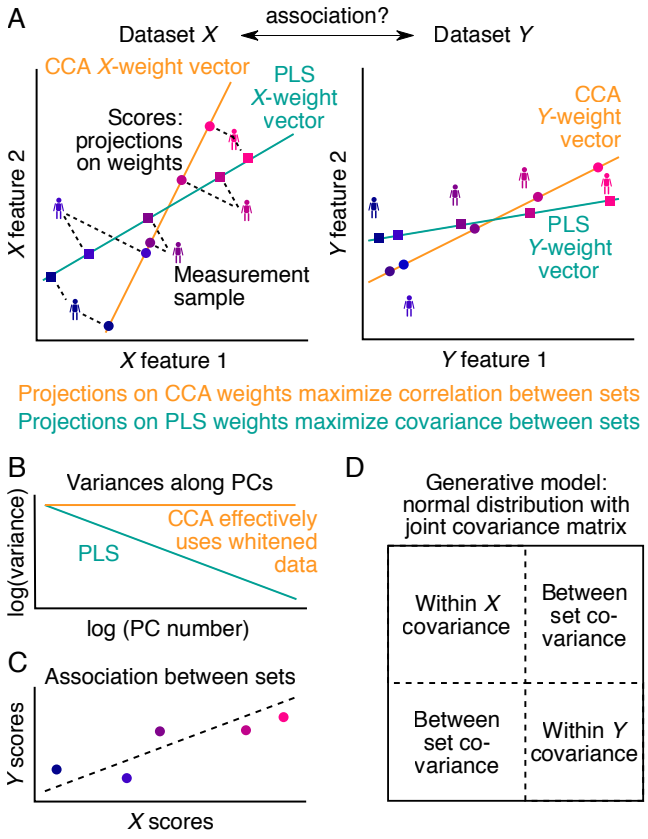

Fig. 1. Overview of CCA, PLS and the generative model used to investigate their properties. A) Two multivariate datasets, X and Y, are projected separately onto respective weight vectors, resulting in univariate scores for each dataset. The weight vectors are chosen such that the correlation (for CCA) 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.

ity and overfitting in CCA/PLS include inflated association strengths (20–22), cross-validated association strengths that are markably lower than in-sample estimates (23), or feature profiles that vary from study to study (20, 23–26). 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 (6, 24). 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 (21, 25, 27). 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 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 (2) and the UK Biobank (3), 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. S1).

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 satisfying 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 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. S1). For CCA, which optimizes correlations instead of covariances, the two datasets are effectively whitened during the analysis (see Methods) and we can therefore assume that all scores' variances are 1.

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 (see Methods and Fig. S2). 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 (for PLS), 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., $p_x + p_y$.

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 (Figs. S3-S4). A typical value in the brain-behavior CCA/PLS literature is about 5 samples per feature (Fig. S5A), 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 correlation (Fig. S6). 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. S7A-B) to a similar degree as in-sample estimates overestimate it (Fig. S7C-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. S7E-F). Finally, bootstrapped association strengths overestimated, on average, slightly more than in-sample estimates (Fig. S8A-B).

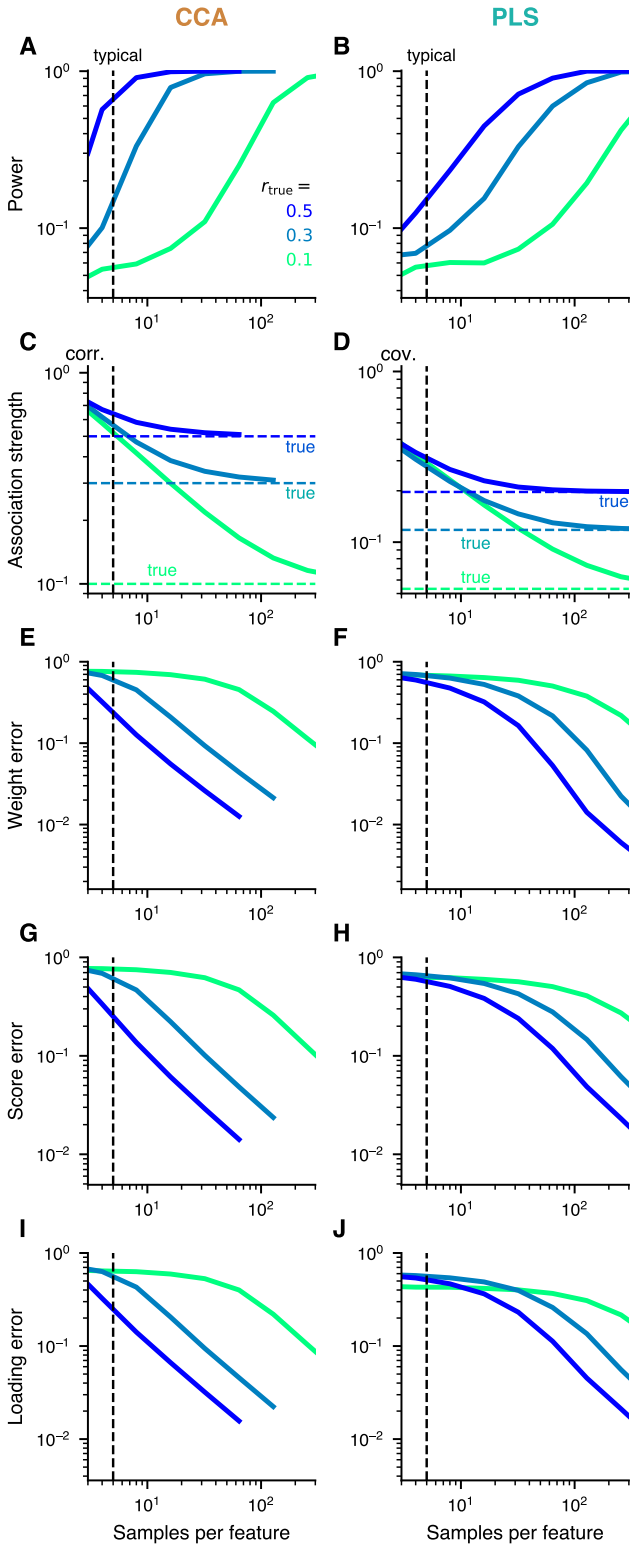

Fig. 2. Sample size dependence of CCA and PLS. For sufficiently large sample sizes, statistical power to detect a non-zero correlation converges to 1 **(A, B)**, between-set covariances approach their assumed true value **(C, D)**, and weight **(E, F)**, score **(G, H)**, and loading **(I, J)** errors 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. S5A).

Third, CCA and 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 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. S8C-F), while the variability of individual weight elements across repeated datasets can be well approximated through bootstrapping (Fig. S8G-H).

Fourth, CCA or 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. S9). 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.

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. 3C). PLS exhibited striking differences from CCA: mean weight stability had a relatively high value even at low sample sample sizes where weight error is very high (Figs. 3D, 2F), with high variability across datasets.

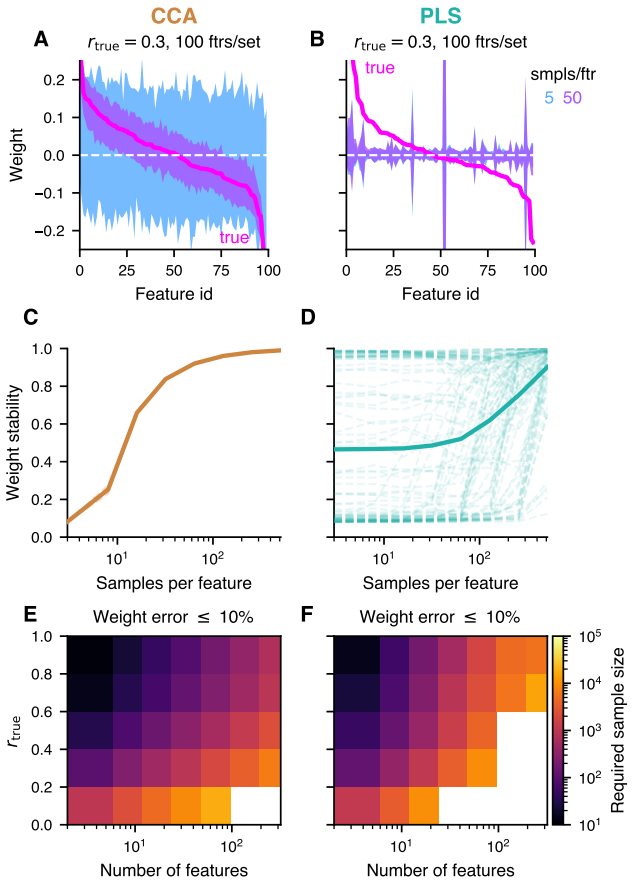

Fig. 3. A large number of samples is required to obtain good weight estimates. **A)** With $r_{true} = 0.3$ and 100 features per dataset, CCA leads to large uncertainty in weights when 5 samples per feature are used. With 50 samples per feature, on the other hand, a much better estimate is possible. The 100 dimensions of the weight vectors are shown along the x-axis, ordered according to the elements of the true weight vector. **B)** For PLS even more samples are necessary. Note that the confidence intervals for 5 and 50 samples per feature overlap almost entirely. **C)** Weight stability, i. e. the average cosine similarity between weights across pairs of repetitions, increases from very low values to 1 (identical weights) with more samples. 10 different covariance matrices were simulated and their individual similarity curves overlap with the mean curve (solid). **D)** For PLS, 100 different covariance matrices (faint dashed curves) have more variable similarity-curves than for CCA, but all eventually converge to 1, as does their mean (solid). **E-F)** Sample sizes required to obtain less than 10% weight errors are shown for CCA and PLS, depending on the assumed true correlation r_{true} and the total number of features in the data. Unless r_{true} is high, 100s to 1000s of samples are required, and more for PLS than for CCA.

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. 3E-F). The required sample size is higher for increasing number of features, and lower for increasing true between-set correlation. More samples were required for PLS than for CCA. 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 bias in PLS. Figs. 3 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 systemic 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 com-

prising 2 features each, so that weight vectors are 2-dimensional unit vectors lying on a circle. Setting $r_{\text{true}} = 0.3$, 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. 4A-B). With only 5 samples per feature, which is typical in CCA/PLS studies (Fig. S5A), 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 in PLS manifests more generally. We first considered an illustrative data regime (64 features/dataset, $r_{\text{true}} = 0.3$). We quantified the PC bias as the cosine similarity between estimated weight vectors and a principal component axis. Compared to CCA, for PLS there was a strong bias toward the dominant PCs, even with a large number of samples (Fig. 4C,D). Note also, that the average PC bias 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. 4F).

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) (2) and UK Biobank (UKBB) (3). Both datasets 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 (3, 4, 28–32). 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). UKBB is a population-level study and, to our knowledge, the largest available neuroimaging dataset. We analyzed fMRI features from 20000 UKBB subjects. HCP and UKBB 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. S10), in agreement with prior CCA studies of HCP data (4, 28–32) (see Fig. S11 for a re-analysis of HCP functional connectivity vs. behavior in which a smaller number of principal components was selected according to an optimization procedure (33)). 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 subsampled the available data to varying sizes from 202 up to 80 % of the available number of samples.

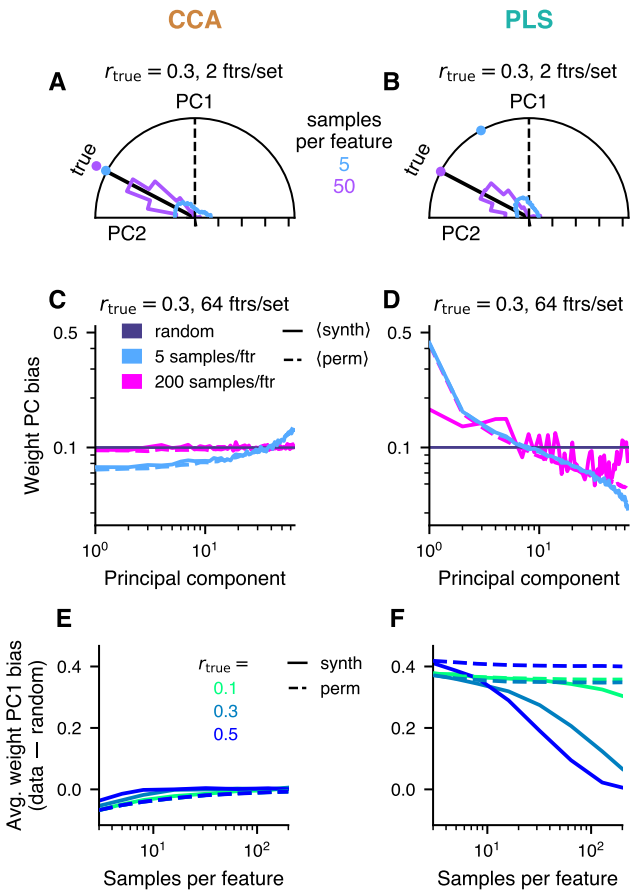

Fig. 4. PLS weights are biased toward first principal component axis, even in the absence of an association. A-B) For illustration, we used a joint covariance matrix assuming a true correlation of 0.3 between datasets and 2 features each for both X and Y datasets. In this 2-dimensional setting weight vectors, scaled to unit length, lie on a circle. Samples of indicated sizes were generated repeatedly from the model. The histogram of estimated weights (for X) as a function of the angle on the circle is shown. 50 samples per feature resulted in good estimates of the weight vectors. 5 samples per feature gave good estimates in many cases but notably all possible weight vectors occurred frequently. The estimation error was worse in PLS where also the mode of the distribution deviated from the true value and was shifted towards the first principal component axis when 5 samples per feature were used. Dots near the border of the semi-circles indicate directional means of the distributions. **C-D)** Another example with 64 features per dataset. Here, we define PC bias (on the y-axis) as the cosine similarity between a weight vector and a principal component axis. Compared to CCA (**C**), for PLS (**D**) there was a strong bias towards the dominant PC axes. **E-F)** The bias towards the first principal component (y-axis) was stronger for PLS (**F**) than for CCA (**E**) also for datasets with varying number of features and true correlations. Shown is the relative PC1 bias across synthetic datasets with varying number of features, relative to the expected PC1 bias of a randomly chosen vector with dimension matched to each synthetic dataset.

We found that the first mode of association was statistically significant for all three sets of data and for both CCA and PLS. Association strengths decreased with increasing size of the subsamples, but clearly converged only for the UKBB data. Cross-validated association strengths estimates increased with subsample size and, for UKBB, 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 weights estimated for different subsamples of the same size) increased with sample size but

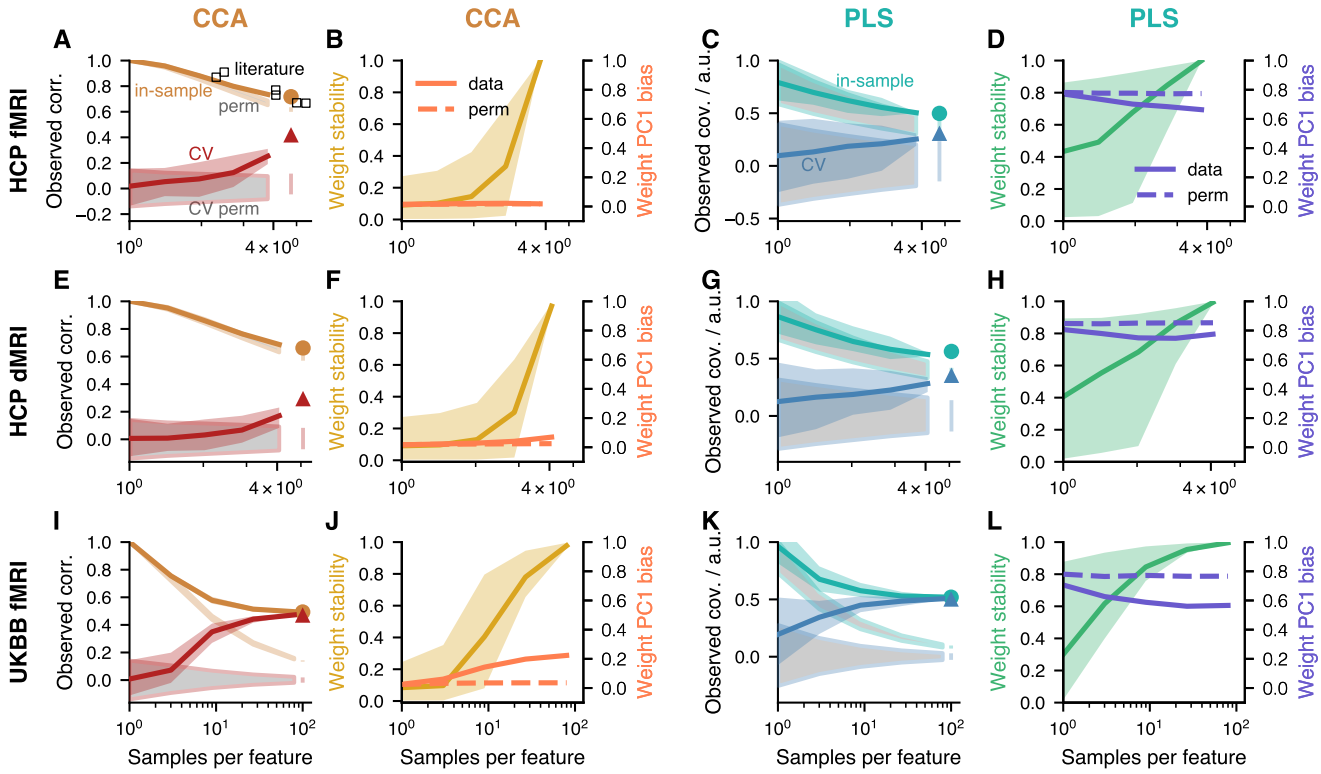

Fig. 5. CCA and PLS analysis of empirical population neuroimaging datasets. We employed CCA (left 2 columns) and PLS (right 2 columns) to investigate multivariate associations between **A-D)** Functional connectivity derived from resting-state functional MRI (fMRI) and behavior from the Human Connectome Project (HCP), **E-H)** Structural connectivity derived from diffusion MRI (dMRI)and behavior in HCP data, and **I-L)** fMRI-derived resting-state functional connectivity and behavior in data from UK Biobank (UKBB). All datasets were reduced to 100 principal components before CCA or PLS (see Methods for preprocessing details). Note the overall agreement of results between different types of data (first and second row) and between data of similar nature from different sources and with different sample size (first and third row): i) For all datasets and for both CCA and PLS a significant mode of association was detected. p-values were 0.001 for all 3 CCAs and 0.006, 0.001, 0.001 for PLS (top to bottom). ii) We subsampled the available data to varying degree and estimated the association strength. Association strengths monotonically decreased with sample size (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 UKBB but not for HCP data where the number of available subjects is much lower. The 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. iii) Cross-validated association strengths are shown in red (column 1) and blue (column 3). They start deviating from values obtained in permuted datasets (grey with red and blue outlines in columns 1 and 3, respectively) at around the same sample size as in-sample estimates do from their permuted datasets. The 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 UKBB (but not yet for HCP) cross-validated association strengths converged to the same value as the in-sample estimate. iv) Weight stability (column 2 and 4) reached values of around 0.8 - 0.9 in subsamples using 80 % of the HCP data, whereas with 80 % of UKBB data weight stability was essentially 1 (i. e. identical weights for different subsamples of the same size). Weight stability was defined as pairwise cosine similarity between weight estimates from repeated subsamples of same size. v). Weight PC1 bias was close to 0 (i. e. no overlap with the first principal component axis) for CCA weights and CCA weights estimated from permuted data (column 2). In contrast, weight PC1 bias for PLS was substantially higher. PC1 bias was defined as the maximum across the 2 datasets of the absolute value of the cosine similarity between weights and the first principal component axis.

reached values close to 1 (perfect similarity) only for UKBB data. Moreover, PC1 bias was close to 0 for CCA but markably larger for PLS 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 UKBB-level datasets.

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 S5A). 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,

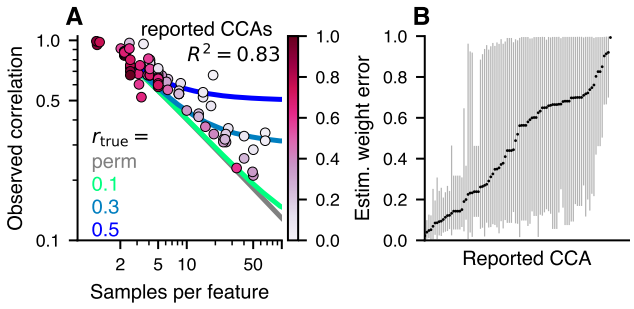

Fig. 6. CCAs reported in the literature might often be unstable. A) Canonical correlations and the number of samples per features are extracted from the literature and overlaid on predictions from the generative model. Many studies employed a small number of samples per feature (cf. also Fig. S5A) and 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. S6A). 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. S12). The farther away a CCA lies from the predictions for permuted data the lower the mean-estimated weight error (cf. Fig. S5B). **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.

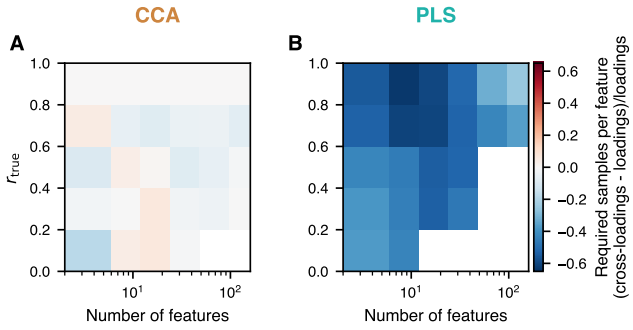

Fig. 7. For PLS cross-loadings provide more stable estimates of feature profiles than loadings. Samples-per-feature required to obtain less than 10% error in either loadings or cross-loadings are compared. Shown here is their relative difference, i. e. the required sample-per-features for cross-loadings minus for loadings, divided by the required samples-per-feature for loadings. **A)** Relative differences were small for CCA. **B)** However, for PLS less samples were required with cross-loadings than with loadings to obtain the same error level.

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. S12). 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 permuted data (see also Fig. S5B). Altogether, these analyses suggest that many published CCA studies might have unstable feature weights due to an insufficient sample size.

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. S13A) and to obtain estimates that have at most 10 % loading or cross-loading error required about the same number of samples (Fig. 7A). For PLS, on the other hand, true loadings and cross-loadings were, albeit not collinear still very similar (Fig. S13B), but cross-loadings could be estimated to within 10 % error with about 20 % to 50 % less samples as loadings in our simulations (Fig. 7B).

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. S14 and Methods). The results are shown in Fig. 8 (see also Figs. S15-S16). Assuming, for example, that the decay constants of the variance spectra satisfy $a_x + a_y = -2$ for PLS, several hundreds to thousands of samples are necessary to achieve the indicated power and error bounds when the true correlation is 0.3 (Fig. 8A). More generally, the required sample size per feature as a function of the true correlation roughy follows a power-law dependence, with a strong increase in required sample size when the true correlation is low (Fig. 8B). Interestingly, PLS generally needs more samples than CCA. 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. S7E-F). Moreover, accurate estimates of a PLS feature profile required fewer samples when assessed as cross-loadings (Fig. 7B).

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

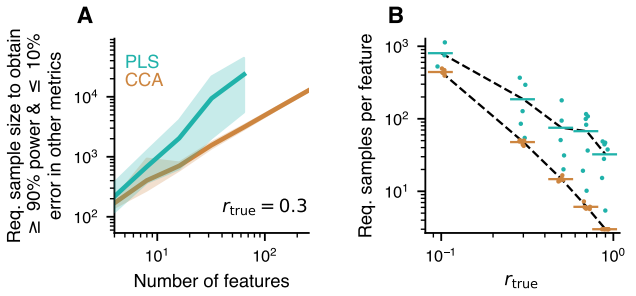

Fig. 8. Required sample sizes. Sample sizes to obtain at least 90 % power as well as at most 10 % error for the association strength, weight, scores and loadings. Shown PLS estimates are constrained by the within-set variance spectrum (here $a_x + a_y = -2$, cf. Fig. S16 for other values). **A)** Assuming 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. S15A-B for other values of r_{true}. **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. S5A). More features are necessary for PLS than CCA, 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.

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. S17). For PLS, 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 very good predictive power for CCA (Fig. S17B), while it is somewhat worse for PLS (Fig. S17C). When we added an additional predictor to the PLS model, measuring the fraction of explained variance along the weight vectors in both datasets, predictions improved notably (Fig. S17D), showing that the linear model approach is also suitable for PLS in principle. As the explained variance along the true weight vectors is unobservable in practice, though, we propose to use the linear model without the explained-variance-predictor.

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 (for PLS) the principal component variance spectrum for each dataset.

Generative model for CCA and PLS. At least for CCA, the distribution of sample canonical correlations has been reported to be intractable, even for normally distributed data (34). 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 (35, 36).

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 a wide range of true correlations, down to zero (Fig. S6). 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 sampled datasets (Fig. 3), affecting significantly their interpretability and generalizability. Furthermore, PLS weights are also biased away from the true value toward the first principal component axis (Fig. 4). As a consequence, similarity of weights from two different samples of the population is necessary but not sufficient to infer replicability. The PC1 bias 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 strongest 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 are biased towards the first principal component axis. Moreover, our analyses suggest that the required number of samples for PLS depends on the within-set principal component variance spectrum and is generally higher than for CCA. 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 (21, 25, 27). 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 UKBB data in Fig. 5. Both in-sample and cross-validated CCA association strengths converge to about 0.5. Fig. 8B 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 (3, 4, 17, 18, 23, 29–32, 37). 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 realtionship. 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 UKBB 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 UKBB data, where we used 20000 subjects, CCA and PLS results appeared to have converged. As the resources required to collect samples of this size go well beyond what is available to typical research groups, this observation supports the accruement of datasets that are shared widely (38, 39).

Generalizability Small sample and effect sizes have been identified as challenges for neuroimaging that impact replicability and generalizability (40, 41). 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 (42). 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 (43–46), scan duration (47), flexibility in the data processing pipeline (48, 49)—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. S4 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 (27). We then assumed the existence of a single cross-modality axis of association, but in practice several ones might be present. In that latter case, theoretical considerations suggest that even larger sample sizes are needed (50, 51). 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. 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. Figs. S3-S4 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 (19) for a recent review). Regularized or sparse CCA or PLS (35, 36) 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 (52, 53). Our analysis pipeline evaluated with a commonly used sparse CCA method (35) 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. S18). 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 (54). 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 (55, 56), which indicates a trade-off between prediction and inference (55). Correspondingly, it has been suggested to consider weight stability as a criterion in sparsity parameter selection (55, 57, 58).

Moreover, whereas CCA and PLS are restricted to discovering linear relationships between two datasets, there exist non-linear extensions, such as kernel (59, 60), deep (61) or non-

parametric (62) CCA, as well as extensions to multiple datasets (63). 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 (64). 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. Finally, we note that a relative of PLS, PLS regression, treats the two datasets asymetrically, deriving scores from one dataset to predict the other (5, 8, 9).

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 (33). Applying this method to HCP data we saw that the reduced number of features the method suggests leads to slightly better convergence (Fig. S11). On the other hand, previous knowledge could be used to preselect the features hypothesized to be most relevant for the question at hand (65–67).

Recommendations. We end with 10 recommendations for using CCA or PLS in practice (summarized in Tab. S1).

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 (33).
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 (68, 69). 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. S15C-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 bias 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, as we have argued above, to assess a PLS analysis 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. Assuming a true correlation of 0.3 results in a rule of thumb that CCA requires about 50 samples per feature. The number for PLS is higher and also depends on the within-set variance spectrum.

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

Materials and methods are summarized in the SI appendix.

ACKNOWLEDGMENTS. 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.

1. Biswal BB, et al. (2010) Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences* 107(10):4734–4739. Publisher: National Academy of Sciences Section: Biological Sciences.
2. Van Essen DC, et al. (2013) The WU-Minn Human Connectome Project: An overview. *NeuroImage* 80:62–79.
3. Miller KL, et al. (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience* 19(11):1523–1536.
4. Smith SM, et al. (2015) A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience* 18(11):1565–1567.
5. Krishnan A, Williams LJ, McIntosh AR, Abdi H (2011) Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage* 56(2):455–475.
6. Wang HT, et al. (2020) Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. *NeuroImage* 216:116745.
7. Hotelling H (1936) Relations Between Two Sets of Variates. *Biometrika* 28(3/4):321–377.
8. Rosipal R, Krämer N (2006) Overview and Recent Advances in Partial Least Squares in *Subspace, Latent Structure and Feature Selection*, Lecture Notes in Computer Science, eds. Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J. (Springer Berlin Heidelberg), pp. 34–51.
9. Abdi H, Williams LJ (2013) Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression in *Computational Toxicology*, eds. Reisfeld B, Mayeno AN. (Humana Press, Totowa, NJ) Vol. 930, pp. 549–579.
10. Sherry A, Henson RK (2005) Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer. *Journal of Personality Assessment* 84(1):37–48.
11. Reyment RA, Bookstein FL, Mckenzie KG, Majoran S (1988) Ecophenotypic variation in *Mutilus pumilus* (Ostracoda) from Australia, studied by canonical variate analysis and tensor biometrics. *Journal of Micropalaeontology* 7(1):11–20. Publisher: Copernicus GmbH.
12. Tabachnick RE, Bookstein FL (1990) The Structure of Individual Variation in Miocene Globorotalia. *Evolution* 44(2):416–434. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1558-5646.1990.tb05209.x.
13. Bin G, Gao X, Yan Z, Hong B, Gao S (2009) An online multi-channel SSVEP-based brain–computer interface using a canonical correlation analysis method. *Journal of Neural Engineering* 6(4):046002.
14. Mazerolles G, Devaux MF, Dufour E, Qannari EM, Courcoux P (2002) Chemometric methods for the coupling of spectroscopic techniques and for the extraction of the relevant information contained in the spectral data tables. *Chemometrics and Intelligent Laboratory Systems* 63(1):57–68.
15. Statheropoulos M, Vassiliadis N, Pappa A (1998) Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric Environment* 32(6):1087–1095.
16. Le Floch E, et al. (2012) Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage* 63(1):11–24.
17. Ziegler G, Dahnke R, Winkler AD, Gaser C (2013) Partial least squares correlation of multivariate cognitive abilities and local brain structure in children and adolescents. *NeuroImage* 82:284–294.
18. Kebets V, et al. (2019) Somatosensory-Motor Dysconnectivity Spans Multiple Transdiagnostic Dimensions of Psychopathology. *Biological Psychiatry*.
19. Zhuang X, Yang Z, Cordes D (2020) A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping* p. hbm.25090.
20. Weinberg SL, Darlington RB (1976) Canonical Analysis when Number of Variables is Large Relative to Sample Size. *Journal of Educational Statistics* 1(4):313–332.
21. Thompson B (1990) Finding a Correction for the Sampling Error in Multivariate Measures of Relationship: A Monte Carlo Study. *Educational and Psychological Measurement* 50(1):15–31.
22. Lee HS (2007) Canonical Correlation Analysis Using Small Number of Samples. *Communications in Statistics - Simulation and Computation* 36(5):973–985.
23. Dinga R, et al. (2019) Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). *NeuroImage: Clinical* p. 101796.
24. Thorndike RM, Weiss DJ (1973) A study of the stability of canonical correlations and canonical components. *Educational and Psychological Measurement* 33(1):123–134.
25. Barcikowski RS, Stevens JP (1975) A Monte Carlo Study of the Stability of Canonical Correlations, Canonical Weights and Canonical Variate-Variable Correlations. *Multivariate Behavioral Research* 10(3):353–364.
26. Strand KH, Kossman S (2000) *Further Inquiry into the Stabilities of Standardized and Structure Coefficients in Canonical and Discriminant Analyses*.
27. Leach L, Henson R (2014) Bias and Precision of the Squared Canonical Correlation Coefficient Under Nonnormal Data Condition. *Journal of Modern Applied Statistical Methods* 13(1).
28. Rahim M, Thirion B, Bzdok D, Buvat I, Varoquaux G (2017) Joint prediction of multiple scores captures better individual traits from brain images. *NeuroImage* 158:145–154.
29. Bijsterbosch JD, et al. (2018) The relationship between spatial configuration and functional connectivity of brain regions. *eLife* 7:e32992.
30. Bijsterbosch JD, Beckmann CF, Woolrich MW, Smith SM, Harrison SJ (2019) The relationship between spatial configuration and functional connectivity of brain regions revisited. *eLife* 8:e44890.
31. Li J, et al. (2019) Topography and behavioral relevance of the global signal in the human brain. *Scientific Reports* 9(1):1–10.
32. Han F, Gu Y, Brown GL, Zhang X, Liu X (2020) Neuroimaging contrast across the cortical hierarchy is the feature maximally linked to behavior and demographics. *NeuroImage* 215:116853.
33. Song Y, Schreier PJ, Ramírez D, Hasija T (2016) Canonical correlation analysis of high-dimensional data with very small sample support. *Signal Processing* 128:449–458.
34. Winkler AM, Renaud O, Smith SM, Nichols TE (2020) Permutation Inference for Canonical Correlation Analysis. *NeuroImage* p. 117065.
35. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10(3):515–534.
36. Tenenhaus A, Tenenhaus M (2011) Regularized Generalized Canonical Correlation Analysis. *Psychometrika* 76(2):257.
37. Drysdale AT, et al. (2017) Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine* 23(1):28–38.
38. Eickhoff S, Nichols TE, Van Horn JD, Turner JA (2016) Sharing the wealth: Neuroimaging data repositories. *NeuroImage* 124:1065–1068.
39. Nichols TE, et al. (2017) Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience* 20:299–303.
40. Button KS, et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14(5):365–376.
41. Poldrack RA, et al. (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience* 18(2):115–126.
42. Varoquaux G (2018) Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* 180:68–77.
43. Leek JT, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* 11(10):733–739. Number: 10 Publisher: Nature Publishing Group.
44. Chen J, et al. (2014) Exploration of scanning effects in multi-site structural MRI studies. *Journal of Neuroscience Methods* 230:37–50.
45. Shinohara RT, et al. (2017) Volumetric Analysis from a Harmonized Multisite Brain MRI Study of a Single Subject with Multiple Sclerosis. *American Journal of Neuroradiology* 38(8):1501–1509. Publisher: American Journal of Neuroradiology Section: ADULT BRAIN.
46. Garcia-Dias R, et al. (2020) Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. *NeuroImage* p. 117127.
47. Noble S, et al. (2017) Multisite reliability of MR-based functional connectivity. *NeuroImage* 146:959–970.
48. Carp J (2012) On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Frontiers in Neuroscience* 6. Publisher: Frontiers.
49. Botvinik-Nezer R, et al. (2020) Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* pp. 1–7. Publisher: Nature Publishing Group.
50. Lawley DN (1959) Tests of Significance in Canonical Analysis. *Biometrika* 46(1/2):59–66.
51. Loukas A (2017) How close are the eigenvectors of the sample and actual covariance matrices? in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. (JMLR.org), Vol. 70, pp. 2228–2237.
52. Shmueli G (2010) To Explain or to Predict? *Statistical Science* 25(3):289–310.
53. Bzdok D, Ioannidis JPA (2019) Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends in Neurosciences* 42(4):251–262.
54. Grellmann C, et al. (2015) Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of MRI and genetic data. *NeuroImage* 107:289–310.
55. Rasmussen PM, Hansen LK, Madsen KH, Churchill NW, Strother SC (2012) Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition* 45(6):2085–2100.
56. Varoquaux G, et al. (2017) Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage* 145:166–179.
57. Baldassarre L, Pontil M, Mourão-Miranda J (2017) Sparsity Is Better with Stability: Combining Accuracy and Stability for Model Selection in Brain Decoding. *Frontiers in Neuroscience* 11. Publisher: Frontiers.
58. Mihalik A, et al. (2020) Multiple Holdouts With Stability: Improving the Generalizability of Machine Learning Analyses of Brain–Behavior Relationships. *Biological Psychiatry* 87(4):368–376.
59. Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation* 16(12):2639–2664.
60. Akaho S (2006) A kernel method for canonical correlation analysis. *arXiv preprint*.
61. Andrew G, Arora R, Bilmes J, Livescu K (2013) Deep Canonical Correlation Analysis in *Proceedings of the 30th International Conference on Ma- chine Learning*. (Atlanta, Georgia, USA), Vol. 28, p. 9.
62. Michaeli T, Wang W, Livescu K (2016) Nonparametric Canonical Correlation Analysis. *arXiv:1511.04839 [cs, stat]*. arXiv: 1511.04839.
63. Kettenring JR (1971) Canonical analysis of several sets of variables. *Biometrika* 58(3):433–451. Publisher: Oxford Academic.
64. Schulz MA, et al. (2019) Deep learning for brains?: Different linear and nonlinear scaling in UK Biobank brain images vs. machine-learning datasets. *bioRxiv* p. 757054. Publisher: Cold Spring Harbor Laboratory Section: New Results.
65. Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD (2011) Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods* 8(8):665–670.
66. Chu C, Hsu AL, Chou KH, Bandettini P, Lin C (2012) Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage* 60(1):59–70.
67. Hong SJ, et al. (2020) Toward Neurosubtypes in Autism. *Biological Psychiatry* 88(1):111–128.
68. Maxwell SE, Kelley K, Rausch JR (2008) Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology* 59(1):537–563.
69. Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a World Beyond "p < 0.05". *The American Statistician* 73(sup1):1–19. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2019.1583913.

Supporting Information Text

1. CCA and PLS

We assume that we have two datasets in the form of data matrices X and Y, both of which have n rows representing samples, and, respectively, p_X and p_Y columns representing measured features (or variables). Throughout we also assume that all columns of X and Y have mean 0. If both datasets consisted of only a single variable, we could measure their association by calculating their covariance or correlation. On the other hand, if one or both consist of more than one variable, pairwise between-set associations can be obtained but the possibly huge number of pairs results in a loss of statistical sensitivity and a difficulty to concisely interpret a potentially large number of significant associations (1). To circumvent these problems, canonical correlation analysis (CCA) and partial least squares (PLS) estimate associations between weighted composites of the original data variables and find those weights that maximize the association strength.

A. Terminology. Given a data matrix, e. g. X, composite variables or *scores* $\vec{t}_X$ (a vector of the same size as the number of samples, n) are formed by projection of X onto a *weight* vector $\vec{w}_X$ (of same size as the number of variables in X, p_X), see Fig. 1A:

$$\vec{t}_X = X\vec{w}_X. \quad [1]$$

Loadings $\vec{\ell}_{XX}$ (of same size as the number of variables in X, p_X) characterize these composite variables by measuring their similarities with each of the original data variables in X (Fig. S9)

$$(\ell_{XX})_j = \underset{i}{\mathrm{corr}}(X_{ij}, t_{X,i}) = (\vec{x}_j)_z \cdot \left(\vec{t}_X\right)_z \quad [2]$$

where corr means Pearson correlation, $\vec{x}_j$ is the j-th column of X, and the subscript z represents z-scoring across samples (i. e. subtraction of the mean and subsequent division by the standard deviation across samples). The complete loading vector is then

$$\begin{aligned}\vec{\ell}_{XX} &= X_z^{\mathsf{T}} \left(\vec{t}_X\right)_z /(n-1) \\ &= \mathrm{diag}(S_{XX})^{-1/2} S_{XX}\vec{w}_X / \sqrt{\vec{w}_X^{\mathsf{T}} S_{XX}\vec{w}_X}\end{aligned} \quad [3]$$

where S_{XX} is the sample covariance matrix for X. Similarly, cross-loadings can be defined as

$$\begin{aligned}\vec{\ell}_{XY} &= X_z^{\mathsf{T}} \left(\vec{t}_Y\right)_z /(n-1) \\ &= \mathrm{diag}(S_{XX})^{-1/2} S_{XY}\vec{w}_Y / \sqrt{\vec{w}_Y^{\mathsf{T}} S_{YY}\vec{w}_Y}\end{aligned} \quad [4]$$

where S_{YY} and S_{XY} are, respectively, the sample covariance matrix for Y and the sample cross-covariance matrix between X and Y.

B. Partial Least Squares. Partial Least Squares (PLS) finds the maximal covariance achievable between weighted linear combinations of features from two data matrices X and Y (2):

$$w_X, w_Y = \underset{\|\tilde{w}_X\|=1, \|\tilde{w}_Y\|=1}{\arg\max} \mathrm{cov}\left(X\tilde{w}_X, Y\tilde{w}_Y\right) \quad [5]$$

The solution is based on the between-set covariance matrix Σ_{XY} which can be estimated from data via its sampled version $S_{XY} = \frac{1}{n-1} X^{\mathsf{T}} Y$. Performing a singular value decomposition yields

$$\Sigma_{XY} = U\mathrm{diag}(\vec{\sigma}_{XY})V^{\mathsf{T}} \quad [6]$$

such that the optimal weights are given by the first columns of U and V, and the maximal covariance

$$\max_{\|\tilde{w}_X\|=1, \|\tilde{w}_Y\|=1} \mathrm{cov}\left(X\tilde{w}_X, Y\tilde{w}_Y\right) \quad [7]$$

by the first singular value $\sigma_{XY,1}$ (3, 4).

Multiple modes of association can be estimated in this way: beyond only the first column, every pair of corresponding columns in U and V provides another mode such that $\mathrm{cov}(X\vec{u}_i, Y\vec{v}_i)$ (for $1 \leq i \leq \min(p_x, p_y)$) is maximal given that the covariance of lower modes (those with indices $< i$) has already been accounted for. There are a number of different algorithms for PLS that differ conceptually in how these higher modes are estimated (2, 3). The one presented above (sometimes called "partial least squares correlation" or PLS-SVD) was chosen for its similarity to canonical correlation analysis (see below). Another notable PLS algorithm is "PLS regression" which, in contrast to the above flavor, is asymmetrical in its handling of X and Y in that it estimates weighted composites (scores) for X and re-uses these as predictors for Y (2).

C. Canonical Correlation Analysis. Canonical Correlation Analysis (CCA) (5), as a multivariate extension of Pearson's correlation, finds maximal correlations between weighted linear combinations of variables from X and Y:

$$\vec{w}_X, \vec{w}_Y = \underset{\tilde{\vec{w}}_X, \tilde{\vec{w}}_Y}{\arg\max}\, \mathrm{corr}\left(X\tilde{\vec{w}}_X, Y\tilde{\vec{w}}_Y\right) \quad [8]$$

Note that $\mathrm{corr}\left(X\tilde{\vec{w}}_X, Y\tilde{\vec{w}}_Y\right)$ is independent of the scaling of $\tilde{\vec{w}}_X$ and $\tilde{\vec{w}}_Y$. I. e. if $\vec{w}_X$ and $\vec{w}_Y$ are solutions of Eq. (8), then $c_X\vec{w}_X$ and $c_Y\vec{w}_Y$, where $c_X \in \mathbb{R}$ and $c_Y \in \mathbb{R}$, are also solutions.

Also note that, as for PLS, several modes of association can be obtained with this framework by successively discounting the variance that has been explained by lower-order modes.

The maximal correlation in Eq. (8) is often called "canonical".

The further analysis is then based on the "whitened" between-set covariance matrix

$$\Sigma_{XY}^{(\mathrm{CCA})} = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2} \quad [9]$$

(6, 7). A singular value decomposition of $\Sigma_{XY}^{(\mathrm{CCA})}$ is performed, yielding

$$\Sigma_{XY}^{(\mathrm{CCA})} = U\mathrm{diag}(\vec{\sigma}_{XY})V^{\mathsf{T}} \quad [10]$$

and the singular values $\vec{\sigma}_{XY}$ turn out to be the canonical correlations from Eq. (8), i. e. the maximal achievable correlations between a weighted linear combination of variables in X on the one hand, and a weighted linear combination of variables in Y on the other hand. The corresponding weights are given by

$$W_X = \Sigma_{XX}^{-1/2}U \quad [11]$$

$$W_Y = \Sigma_{YY}^{-1/2}V. \quad [12]$$

The use of the "whitened" between-set covariance matrix in CCA leads to an invariance property between datasets that we will exploit later. To see this, let X_w, Y_w be whitened data matrices, i. e. $X_\mathrm{w} = X\Sigma_{XX}^{-1/2}$ and $Y_\mathrm{w} = Y\Sigma_{YY}^{-1/2}$ such that $\Sigma_{X_\mathrm{w}X_\mathrm{w}} = \mathbb{1}$, $\Sigma_{Y_\mathrm{w}Y_\mathrm{w}} = \mathbb{1}$. Then,

$$\begin{aligned}
\Sigma_{X_\mathrm{w}Y_\mathrm{w}}^{(\mathrm{CCA})} &= \Sigma_{X_\mathrm{w}X_\mathrm{w}}^{-1/2}\Sigma_{X_\mathrm{w}Y_\mathrm{w}}\Sigma_{Y_\mathrm{w}Y_\mathrm{w}}^{-1/2} && [13]\\
&= \Sigma_{X_\mathrm{w}Y_\mathrm{w}} && [14]\\
&= \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2} && [15]\\
&= \Sigma_{XY}^{(\mathrm{CCA})} && [16]
\end{aligned}$$

which is the same as for the original (non-whitened data). Consequently, canonical correlations for the original and whitened data are the same, given by the singular values of $\Sigma_{XY}^{(\mathrm{CCA})}$, canonical weights for the whitened data are directly its singular vectors and canonical weights for the original (non-whitened) data differ only by a matrix $\Sigma_{XX}^{-1/2}$ and $\Sigma_{YY}^{-1/2}$ for X and Y, respectively (Eq. (11)-Eq. (12)).

It can be shown that the invariance property is even more general (6). Let $O^{(X)} \in \mathbb{R}^{p_x \times p_x}$ and $O^{(Y)} \in \mathbb{R}^{p_y \times p_y}$ be non-singular and $\vec{d}^{(X)} \in \mathbb{R}^{p_x}$ and $\vec{d}^{(Y)} \in \mathbb{R}^{p_y}$ be arbitrary vectors. Then $\tilde{X} = O^{(X)}X + d^{(X)}$ and $\tilde{Y} = O^{(Y)}Y + d^{(Y)}$ have the same canonical correlations as X and Y, and the canonical vectors are related by

$$\tilde{\vec{w}}_X = \left(O^{(X)}\right)^{-1}\vec{w}_X \quad [17]$$

$$\tilde{\vec{w}}_Y = \left(O^{(Y)}\right)^{-1}\vec{w}_Y \quad [18]$$

Thus, in particular, z-scored data $X_z = \mathrm{diag}(S_{XX})^{-1/2}X$ and $Y_z = \mathrm{diag}(S_{YY})^{-1/2}Y$ as well as whitened data X_w and Y_w have the same canonical correlations as the original data X and Y.

In CCA, X- and Y-weights are related by (8)

$$w_X = \Sigma_{XX}^{-1}\Sigma_{XY}\vec{w}_Y/\sigma_{XY} \quad [19]$$

$$w_Y = \Sigma_{YY}^{-1}\Sigma_{YX}\vec{w}_X/\sigma_{XY} \quad [20]$$

$$[21]$$

Replacing sample with population covariance matrices in Eq. (3) and Eq. (4), we thus also see that loadings and cross-loadings are collinear

$$\begin{aligned}
\vec{\ell}_{XY} &= \mathrm{diag}(\Sigma_{XX})^{-1/2}\Sigma_{XY}\vec{w}_Y/\sqrt{\vec{w}_Y^{\mathsf{T}}\Sigma_{YY}\vec{w}_Y}\\
&= \mathrm{diag}(\Sigma_{XX})^{-1/2}\Sigma_{XX}\Sigma_{XX}^{-1}\Sigma_{XY}\vec{w}_Y/\sqrt{\vec{w}_Y^{\mathsf{T}}\Sigma_{YY}\vec{w}_Y}\\
&\propto \mathrm{diag}(\Sigma_{XX})^{-1/2}\Sigma_{XX}\vec{w}_X\\
&\propto \vec{\ell}_{XX} && [22]
\end{aligned}$$

D. Overestimation of association strength. Let Σ_{XY} be a population cross-covariance matrix with singular value decomposition

$$\Sigma_{XY} = U\mathrm{diag}(\vec{\sigma}_{XY})V^\mathsf{T} \quad [23]$$

and let $\vec{u}_1$, $\vec{v}_1$ and σ_1 be, respectively, the first columns of U, V and the first entry in $\vec{\sigma}_{XY}$. In PLS, $\vec{u}_1$, and $\vec{v}_1$ are the weight vectors of the first mode and σ_1 is the corresponding association strength. For CCA, as noted above, whitening the data leaves canonical correlations unchanged, so that we assume now data are white when performing CCA. Then, $\Sigma_{XY}^{(\mathrm{CCA})} = \Sigma_{XY}$ and the weights and association strength of the first mode are also given by $\vec{u}_1$, $\vec{v}_1$ and σ_1. In both cases, we have $\sigma_1 = \vec{u}_1^\mathsf{T}\Sigma_{XY}\vec{v}_1$.

The sample covariance matrix $S_{XY} = \frac{1}{n-1}X^\mathsf{T}Y$ is an unbiased estimator for Σ_{XY}, i. e. $\mathrm{E}[S_{XY}] = \Sigma_{XY}$. Therefore,

$$\mathrm{E}[\vec{u}_1^\mathsf{T}S_{XY}\vec{v}_1] = \vec{u}_1^\mathsf{T}\,\mathrm{E}[S_{XY}]\vec{v}_1 = \vec{u}_1^\mathsf{T}\Sigma_{XY}\vec{v}_1 = \sigma_1 \quad [24]$$

i. e. if the true (but unknown) weights were applied to a given dataset (between-set covariance matrix) the association strength of the resulting scores would, on average, match the true association strength. However, by definition, CCA and PLS select those weight vectors that maximize the association strength between resulting scores. If $\hat{\vec{u}}_1$ and $\hat{\vec{v}}_1$ are those optimal weights for a given dataset, then

$$\hat{\vec{u}}_1^\mathsf{T}S_{XY}\hat{\vec{v}}_1 \geq \vec{u}_1^\mathsf{T}S_{XY}\vec{v}_1 \quad [25]$$

and consequently also

$$\mathrm{E}[\hat{\vec{u}}_1^\mathsf{T}S_{XY}\hat{\vec{v}}_1] \geq \mathrm{E}[\vec{u}_1^\mathsf{T}S_{XY}\vec{v}_1] = \sigma_1 \quad [26]$$

i. e. the association strength is overestimated.

E. Sparse CCA. Multiple sparse CCA and PLS methods exist (9–12). Here, we use *penalized matrix decomposition* (PMD) (11), which has found widespread application, see e.g. (13–18). Briefly, the PMD algorithm repeats the following steps until convergence (11)

- $\vec{u} \leftarrow \arg\max_{\vec{u}} \vec{u}^\mathsf{T}X^\mathsf{T}Y\vec{v}$ subject to $\|\vec{u}\|_1 \leq c_1$ and $\|\vec{u}\|_2 \leq 1$
- $\vec{v} \leftarrow \arg\max_{\vec{v}} \vec{u}^\mathsf{T}X^\mathsf{T}Y\vec{v}$ subject to $\|\vec{v}\|_1 \leq c_2$ and $\|\vec{v}\|_2 \leq 1$

to maximize $\vec{u}^\mathsf{T}X^\mathsf{T}Y\vec{v}$. If $X^\mathsf{T}X \approx \mathbb{1}$ and $\|\vec{u}\|_2 = 1$, then $1 = \|\vec{u}\|_2 \approx \|X\vec{u}\|_2$ and analogously for Y. Consequently, $\vec{u}^\mathsf{T}X^\mathsf{T}Y\vec{v} \approx \vec{u}^\mathsf{T}X^\mathsf{T}Y\vec{v}/\sqrt{\|X\vec{u}\|_2\|Y\vec{v}\|_2} = \mathrm{corr}(X\vec{u}, Y\vec{v})$. Note that the approximation $X^\mathsf{T}X \approx \mathbb{1}$ (together with $Y^\mathsf{T}Y \approx \mathbb{1}$) makes this sparse "CCA" variant identical to sparse PLS (18, 19).

E.1. Implementation and sparsity parameter selection. We implemented a Python wrapper for the R-package PMA (20) which we used with default parameters. Sparsity parameters were estimated separately for each dataset subjected to sparse CCA via 5-fold cross-validation (11, 21): for X and Y we used 5 different candidate sparsity parameters (0.2, 0.4, 0.6, 0.8 and 1 where smaller values mean more sparsity and 1 corresponds to no sparsity) for a total of 25 parameter pairs. For each candidate parameter pair sparse CCA was estimated with 80 % of the data, the resulting weights applied to the remaining 20 % of the data to obtain test scores, the Pearson correlation calculated between the test scores and averaged across the 5 folds. The pair of sparsity parameters for which the test-correlation averaged across folds was maximal, was then selected and sparse CCA re-estimated on the whole data with these parameters.

2. Generating synthetic data for CCA and PLS

We will analyze properties of CCA and PLS with the help of simulated datasets. These datasets will be drawn from a normal distribution with mean 0 and a covariance matrix Σ that will encode 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)

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY}^\mathsf{T} & \Sigma_{YY} \end{pmatrix} \in \mathbb{R}^{p_x+p_y \times p_x+p_y} \quad [27]$$

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

A. 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.

B. 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

$$\Sigma_{XY} = W_X \mathrm{diag}(\vec{\sigma}_{XY}) W_Y^\mathsf{T} \qquad \text{(for PLS)} \tag{28}$$

where $W_X^\mathsf{T} W_X = \mathbb{1}_{p_x}$, $W_Y^\mathsf{T} W_Y = \mathbb{1}_{p_y}$ 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

$$\sigma_{XY} = r_\mathrm{true} \sqrt{\mathrm{var}\left(X\vec{w}_X\right) \mathrm{var}\left(Y\vec{w}_Y\right)} \tag{29}$$

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

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

$$\begin{aligned}
\Sigma_{XY} &= \Sigma_{XX}^{1/2} \Sigma_{XY}^\mathrm{CCA} \Sigma_{YY}^{1/2} \\
&= \Sigma_{XX}^{1/2} \left(U \mathrm{diag}(\vec{\sigma}_{XY}) V^\mathsf{T} \right) \Sigma_{YY}^{1/2} \\
&= \Sigma_{XX}^{1/2} \left(\Sigma_{XX}^{1/2} W_X \right) \mathrm{diag}(\vec{\sigma}_{XY}) \left(\Sigma_{YY}^{1/2} W_Y \right)^\mathsf{T} \Sigma_{YY}^{1/2}
\end{aligned} \tag{30}$$

where we have used Eq. (11) and Eq. (12). Here, $\vec{\sigma}_{XY}$ are directly the assumed true correlations and, by construction, the weights W_X and W_Y are constrained to satisfy $\mathbb{1} = U^\mathsf{T} U = (\Sigma_{XX}^{1/2} W_X)^\mathsf{T} \Sigma_{XX}^{1/2} W_X$ and analogously for W_Y. As mentioned above, we can exploit the property that pre-whitened data result in the same matrix Σ_{XY}^CCA. In the following, thus, assume that we have data X and Y with $\Sigma_{XX} = \mathbb{1}_{p_x}$ and $\Sigma_{YY} = \mathbb{1}_{p_y}$. But then,

$$\Sigma_{XY} = W_X \mathrm{diag}(\vec{\sigma}_{XY}) W_Y^\mathsf{T} \qquad \text{(for CCA)} \tag{31}$$

and $\mathbb{1} = U^\mathsf{T} U = W_X^\mathsf{T} W_X$, as well as $\mathbb{1} = W_Y^\mathsf{T} W_Y$. This is identical to the result for PLS, except that for CCA the assumption that the data are white is implicit.

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 a set of weight vectors $\vec{w}_{X,i}$ and $\vec{w}_{Y,i}$ (for $1 \leq i \leq m$). The weight vectors for each set need to be orthonormal ($W_X^\mathsf{T} W_X = W_Y^\mathsf{T} W_Y = \mathbb{1}_m$), and, for CCA, both X and Y need to be white, i. e. $\Sigma_{XX} = \mathbb{1}_{p_X}$ and $\Sigma_{YY} = \mathbb{1}_{p_Y}$.

C. 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

$$\vec{w}_X^\mathsf{T} \Sigma_{XX} \vec{w}_X > \frac{1}{2} \frac{\mathrm{tr}\, \Sigma_{XX}}{p_X} \tag{32}$$

   and analogously for Y.

2. The weight vectors impact the joint covariance matrix Σ (via Eq. (27), Eq. (28) and Eq. (31)). 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, $\vec{w}_\mathrm{lo}$ and $\vec{w}_\mathrm{hi}$, they are combined as

$$\vec{w} = c\vec{w}_\mathrm{hi} + \sqrt{1 - c^2}\vec{w}_\mathrm{lo} \tag{33}$$

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 $\vec{w}_\mathrm{lo}$, $\vec{w}_\mathrm{hi}$ 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 Eq. (32), 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 $U_{\mathrm{lo}} \in \mathbb{R}^{p \times p-q}$ with a rotation matrix R_{lo}

$$W_{\mathrm{lo}} = U_{\mathrm{lo}} R_{\mathrm{lo}} \quad [34]$$

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. S2). 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

$$W_{\mathrm{hi}} = U_{\mathrm{hi}} R_{\mathrm{hi}} \quad [35]$$

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 (22) 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 $\vec{w}_X = U_{X,\mathrm{hi}}\vec{c}_X$ and $\vec{w}_Y = U_{Y,\mathrm{hi}}\vec{c}_Y$ 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 $\vec{w}_{X,\mathrm{hi},1} = (1, 0, \ldots, 0)^{\mathsf{T}}$, $\vec{w}_{Y,\mathrm{hi},1} = (1, 0, \ldots, 0)^{\mathsf{T}}$ 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. $\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}\Sigma_{XY}^{\mathsf{T}}$, are positive definite. Σ_{YY} is positive definite by construction. The between-set covariance matrix here is $\Sigma_{XY} = \sigma_{XY,1}\vec{w}_{X,\mathrm{hi},1}\vec{w}_{Y,\mathrm{hi},1}^{\mathsf{T}}$. For CCA, $\sigma_{XY,1}$ is the canonical correlation $r_{\mathrm{true}} < 1$. For PLS, $\sigma_{XY,1} = r_{\mathrm{true}}\sqrt{\mathrm{var}\, X\vec{w}_X \,\mathrm{var}\, Y\vec{w}_Y}$, which, with the specific choices of $\Sigma_{XX}, \Sigma_{YY}, \vec{w}_X$ and $\vec{w}_Y$ just described, also simplifies to $\sigma_{XY,1} = r_{\mathrm{true}}$. Thus, $\Sigma_{XY}\Sigma_{YY}\Sigma_{XY}^{\mathsf{T}} = r_{\mathrm{true}}^2(1, 0, \ldots, 0)^{\mathsf{T}}(1, 0, \ldots, 0)$ and consequently the diagonal entries of $\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}\Sigma_{XY}^{\mathsf{T}}$ 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.

D. 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. For CCA, as outlined in the previous section, we can use pre-whitened data for which $\Sigma_{XX} = \mathbb{1}$ and $\Sigma_{YY} = \mathbb{1}$, and as a result, the cross-covariance matrix Σ_{XY} has the same form as for PLS. The columns of the weight matrices W_X and W_Y must be mutually orthonormal, and, in addition, we assume that they are contained within a subspace of, respectively, q_X and q_Y dominant principal components, that is $W_X = U_{XX}^{(q_X)}R_{XX}$ and $W_Y = U_{YY}^{(q_Y)}R_{YY}$, where $U_{XX}^{(q_X)} \in \mathbb{R}^{p_X \times q_X}$ is the matrix of the first q_X columns of U_{XX}, $R_{XX} \in \mathbb{R}^{q_X \times q_X}$ is unitary, and analogously for $U_{YY}^{(q_Y)}$ and R^{q_Y}.

E. Performed simulations. For Figs. 2, 3E-F, the colored curves in Fig. 6A, Figs. 7, 8, S15, the CCA results in Fig. S17 and Fig. S3, 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, 64, 128 as well as 25 different covariance matrices. $a_X + a_Y$ was fixed at 0 for CCA and -2 for PLS. 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, $r_{\mathrm{true}} = 0.3$ and we used 1 covariance matrix for CCA and PLS.

For Fig. 3C-D p_X was 100, $r_{\mathrm{true}} = 0.3$ and we used 10 and 100 different covariance matrices for CCA and PLS, respectively.

For Fig. 4A-B, p_X was 2, $r_{\mathrm{true}} = 0.3$ and we used 10000 different covariance matrices for CCA and PLS.

For Fig. 4C-D and 4G-H, 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 $r_{\mathrm{true}} = 0.3$ was used for Fig. 4E-F.

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, and 1 covariance matrix for each r_{true}.

For Fig. S4 $p_X + p_Y$ was fixed at 64 and for p_X we used 2, 4, 8, 16, 32.

In Fig. S6, for p_X we used 4, 8, 16, 32, 64 and, only for CCA, also 128, we used 10 different covariance matrices for both CCA and PLS and varied r_{true} from 0 to 0.99 in steps 0.01.

For Fig. S7 we used 2, 4, 8, 16 and 32 for p_X, and 10 different covariance matrices for both CCA and PLS.

For Fig. S8 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 the PLS results in Fig. S16 and Fig. S17 we used 50 different covariance matrices for $r_{\text{true}} = 0.1$, 0.9, as well as for $r_{\text{true}} = 0.7$ in combination with $p_X = 128$, 25 for $r_{\text{true}} = 0.5$ in combination with $p_X = 64$, and 75 for all other combinations of p_X and r_{true} for which the computational expense was not too high. 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. S18 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.

3. 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

$$\Delta r = \frac{\hat{r} - r}{r} \qquad [36]$$

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

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:

$$\Delta w = \max_{s \in \{X,Y\}} \left(1 - |\operatorname{cossim}(\hat{\vec{w}}_s, \vec{w}_s)|\right) \qquad [37]$$

where

$$\operatorname{cossim}(\hat{\vec{w}}_s, \vec{w}_s) = \frac{\hat{\vec{w}}_s \cdot \vec{w}_s}{\|\hat{\vec{w}}_s\| \|\vec{w}_s\|}. \qquad [38]$$

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:

$$\Delta t = \max_{s \in X,Y} \left(1 - |\operatorname*{rankcorr}_i \left(\hat{t}_{s,i}^{(\text{test})}, t_{si}^{(\text{test})}\right)|\right) \qquad [39]$$

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:

$$\Delta\ell = \max_{s\in X,Y}\left(1 - |\operatorname*{corr}_i\left(\hat{\ell}_{s,i}^{(\text{test})}, \ell_{s,i}^{(\text{test})}\right)|\right) \quad [40]$$

True loadings are calculated with Eq. (3) (replacing the sample covariance matrix in the formula with its population value). Estimated loadings are obtained by correlating data matrices with score vectors (Eq. (2)). 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.

4. Weight similarity to principal component axes

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

$$R = \frac{1}{n_\alpha}\sum_j^{n_\alpha} \mathrm{e}^{2i\alpha_j} \quad [41]$$

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 (23)

$$f_n(x) = \frac{\mathrm{d}\,P(X \le x)}{\mathrm{d}\,x} \quad [42]$$

where P denotes the cumulative distribution function for the probability that a random unit-vector has cosine similarity with $\vec{e}$ (or projection onto $\vec{e}$) $\le x$. For $-1 \le x \le 0$, 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. $x - h = -1$)

$$P(X \le x) = \frac{A_n(h)}{A_n(2)} \quad [43]$$

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

$$A_n(h) = \frac{1}{2}A_n(2)I\left(h(2-h); \frac{n-1}{2}, \frac{1}{2}\right) \quad [44]$$

and I is the regularized incomplete beta function. Thus,

$$\begin{aligned} f_n(x) &= \frac{1}{2}\frac{\mathrm{d}\,I}{\mathrm{d}\,x}\left((x+1)(1-x); \frac{n-1}{2}, \frac{1}{2}\right) && [45]\\ &= \frac{1}{2}\frac{1}{B(\frac{n-1}{2}, \frac{1}{2})}(1-x^2)^{\frac{n-3}{2}}\left(x^2\right)^{-1/2}(-2x) && [46]\\ &= \frac{1}{B(\frac{n-1}{2}, \frac{1}{2})}(1-x^2)^{\frac{n-3}{2}} && [47] \end{aligned}$$

where B is a beta function and

$$\begin{aligned} f_n(2\tilde{x}-1) &\propto (2-2\tilde{x})^{\frac{n-1}{2}-1}(2\tilde{x})^{\frac{n-1}{2}-1} && [48]\\ &\propto f_\beta\left(\tilde{x}; \frac{n-1}{2}, \frac{n-1}{2}\right) && [49] \end{aligned}$$

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

5. Analysis of empirical data

We demonstrate CCA and PLS analysis in empirical data using data from the Human Connectome Project (HCP) (24) and UK Biobank (25).

A. Human Connectome Project data.

A.1. fMRI data. We used resting-state fMRI (rs-fMRI) from 951 subjects from the Human Connectome Project (HCP) 1200-subject data release (03/01/2017) (24). 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 (26, 27). Following the HCP MPP, BOLD time-series were denoised using ICA-FIX (28, 29) and registered across subjects using surface-based multimodal inter-subject registration (MSMAll) (30). Additionally, global signal, ventricle signal, white matter signal, and subject motion and their first-order temporal derivatives were regressed out (31).

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 Human Connectome Project (32). Functional connectivity was then computed as the Fisher-z-transformed Pearson correlation between all pairs of parcels.

3 subjects were excluded (see section D below), resulting in a total of 948 subjects with 100 connectivity features each.

A.2. dMRI data. Diffusion MRI (dMRI) data and structural connectivity patterns were obtained as described in (33, 34). In brief, 41 major white matter (WM) bundles were reconstructed from preprocessed HCP diffusion MRI data (35) using FSL's XTRACT toolbox (34). 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` (36, 37), 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 (38) giving a final set of $68 \times 41 = 2788$ connectivity features for each of the 1020 HCP subjects.

A.3. Behavioral measures. The same list of 158 behavioral and demographic data items as in (39) was used.

A.4. Confounders. 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. Missing values were set to 0. All resulting confounds were z-scores across subjects.

B. UK Biobank data.

B.1. fMRI data. We utilised pre-processed resting-state fMRI data (40) from 20,000 subjects, available from the UK Biobank Imaging study (25).

In brief, EPI unwarping, distortion and motion correction, intensity normalisation and highpass temporal filtering were applied to each subject's functional data using FSL's Melodic (41), data were registered to standard space (MNI), and structured artefacts are removed using ICA and FSL's FIX (28, 29, 41).

A set of resting-state networks were identified common across the cohort using a subset of subjects ($\approx$ 4000 subjects) (40). This was achieved by extracting the top 1200 components from a group-PCA (42) and a subsequent spatial ICA with 100 resting-state networks (41, 43). Visual inspection revealed 55 non-artefactual ICA components. Next, these 55 group-ICA networks were dual regressed onto each subjects' data to define grey matter nodes. The average timeseries of each of the nodes were used to compute partial correlation parcellated connectomes with a dimensionality of 55 x 55. The connectomes were z-score transformed and the upper triangle vectorised to give 1485 functional connectivity features per subject, for each of the 20,000 subjects.

B.2. Behavioural measures. The UK Biobank contains a wide range of subject measures (44), including physical measures (e.g. weight, height), food and drink, cognitive phenotypes, lifestyle, early life factors and sociodemographics.

We hand-picked a subset of 3895 cognitive, lifestyle and physical measures, as well as early life factors. For categorical items, we replaced negative values with 0, as in (25). Such negative values encode mostly "Do not know"/"Prefer not to answer". We then removed measures that had missing values in more than 50% of subjects (for instance measures that reflected subsequent visits, which were not available for many subjects that only had one visit). We also removed measures that had identical values in at least 90% of subjects, leaving 633 non-imaging measures. 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 62 measures (mostly physical measures, also some less informative sections of tests), resulting in a final set of 571 behavioural measures, available for each of the 20,000 subjects.

B.3. 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 (25) 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 $\times$ sex and age$^2\times$ sex. Altogether these were 14 confounds.

Finally, we imputed 0 for missing values and z-scored all items.

C. Preprocessing for CCA and PLS. We prepared data for CCA following, for the most part, the pipeline in (39).

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

$$X_{\text{deconfounded}} = X - C\beta \quad [50]$$

where

$$\beta = C^{+}X = \left(C^{\mathsf{T}}C\right)^{-1}C^{\mathsf{T}}X \quad [51]$$

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 (45) 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 (46). 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.

D. CCA/PLS analysis. Permutation-based p-values in Fig. 5 and S11 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 (47). 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. 5A-D) were performed for 5 logarithmically spaced subsample-sizes between 202 and 80% of the total subject number. For each subsample size 100 subsampled data matrices were used.

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

E. Principal component spectrum decay constants. The decay constant of a principal component spectrum (Fig. S1) 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. S1 we included as many principal components into the linear regression as necessary to explain either 30% or 90% of the variance.

6. 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 (39, 48–77), which are summarized in SI Dataset 1.

7. 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 (e. g. due to numerical errors) 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.

8. 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. S14) 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 (for PLS) $a_x + a_y \sim \mathcal{U}(-3, 0)$, $a_x = c(a_x + a_y)$, 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 $\log(r_{\text{true}})$, $\log(p_x + p_y)$, (for PLS) $|a_x + a_y|$, and including an intercept term.

We tested the predictions of linear model using a split-half approach (Fig. S17), 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.

As for PLS (unlike CCA) the direction of the weight vectors relative to the principal component axes results in varying amounts of explained variance, we also tested an alternative linear model for PLS, in which $\log(v_x v_y)$ was included as additional predictor, where v_x and v_y denote, respectively, the explained variance ratio for the X and Y weight vector, i. e. the variance of the scores divided by the trace of the corresponding within-set covariance matrix. Note that, as the true weights are unknown in practice, this additional predictor is inaccessible in practice, and the alternative linear model only serves to gauge how much of the uncertainty in the linear model is due to this unobservable component.

9. The *gemmr* software package

We provide an open-source Python package, called *gemmr*, that implements the generative modeling framework presented in this paper. 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.

10. Code and data availability

Our open-source Python software package, *gemmr*, is freely available at https://github.com/murraylab/gemmr. It has dependencies on `scikit-learn` (78), `statsmodels` (46), `xarray` (79), `pandas` (80), `scipy` (81) and `numpy` (82) among others. Jupyter notebooks detailing the analyses and generation of figures presented in the manuscript are made available as part of the package documentation. The outcomes of synthetic datasets that were analyzed with CCA or PLS are available from https://osf.io/8expj/.

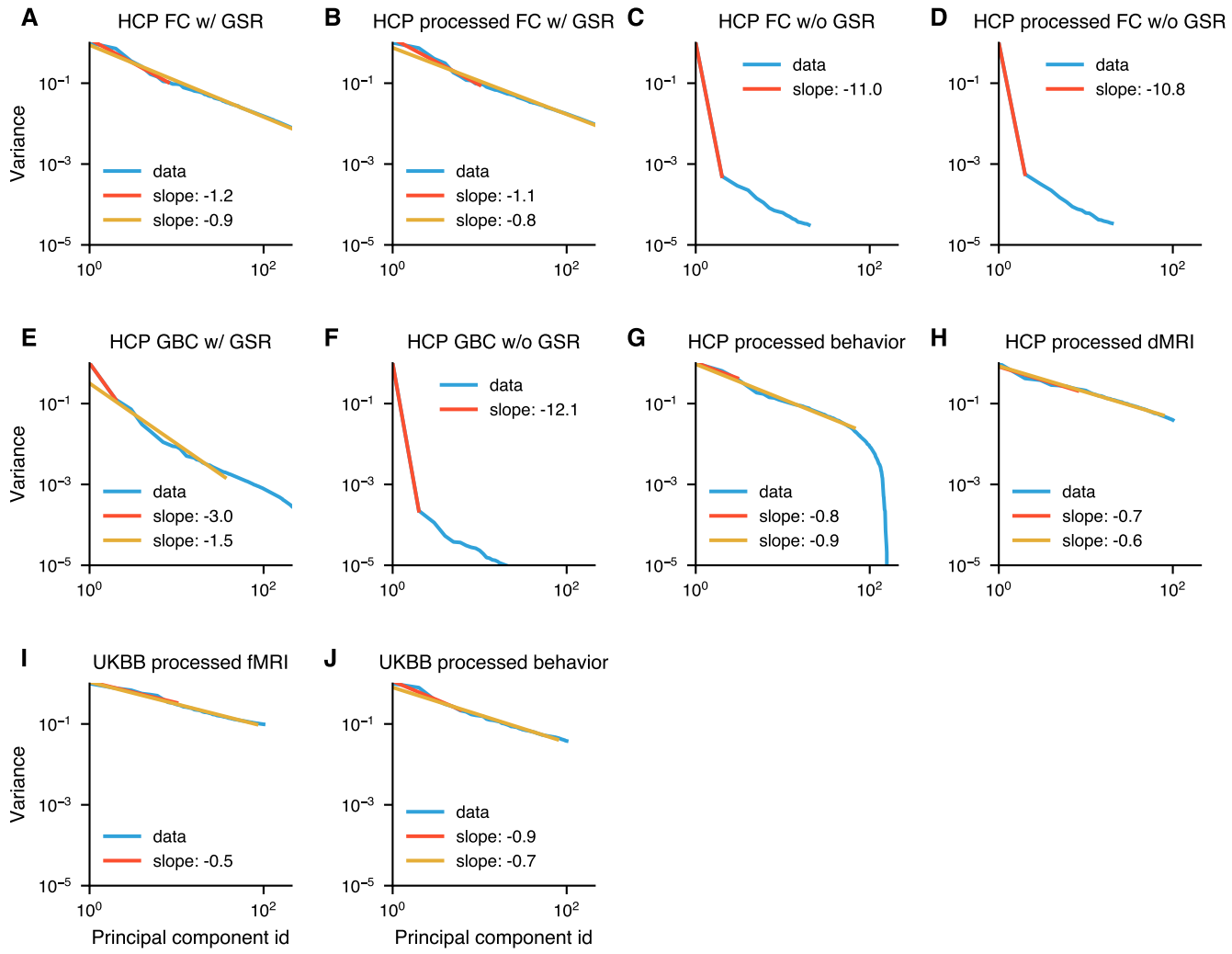

Fig. S1. Decay constants of principal component spectra in empirical data. Decay constants are estimated as the slope in a linear regression for the logarithm of the explained variance on the logarithm of the associated principal component number. We include enough components into the linear regression 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 C), both based on 951 subjects. **C)** HCP functional connectivity for 877 subjects where global signal was not regressed out (cf. subsection A.1) 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.

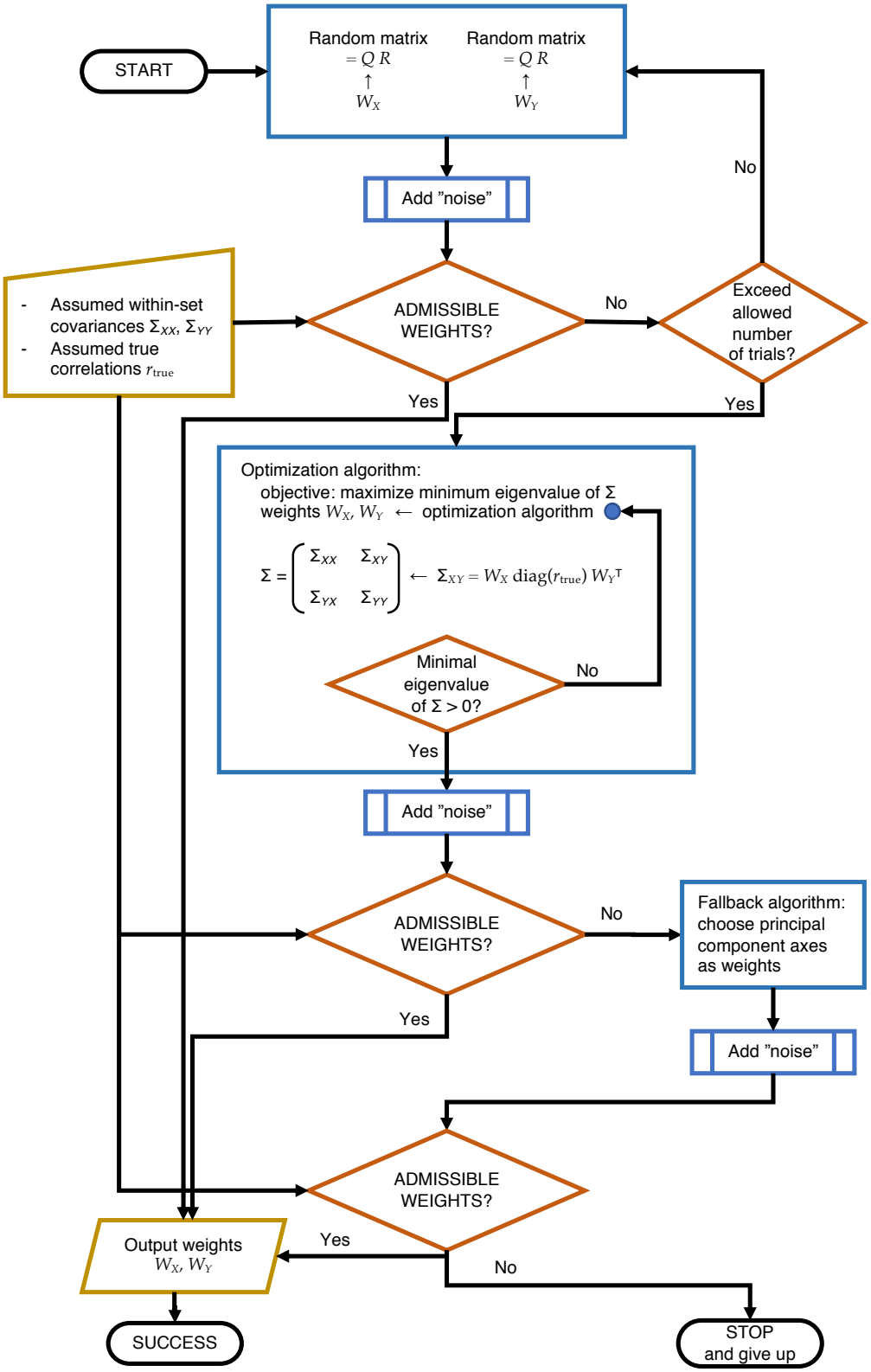

Fig. S2. Algorithm for choosing weight vectors. The flowchart illustrates the main logic of the algorithm. We require weight vectors (i) to be orthonormal within each set, (ii) to result in scores that explain at least a given fraction of variance, and (iii) to result in a proper, i. e. positive definite, 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".

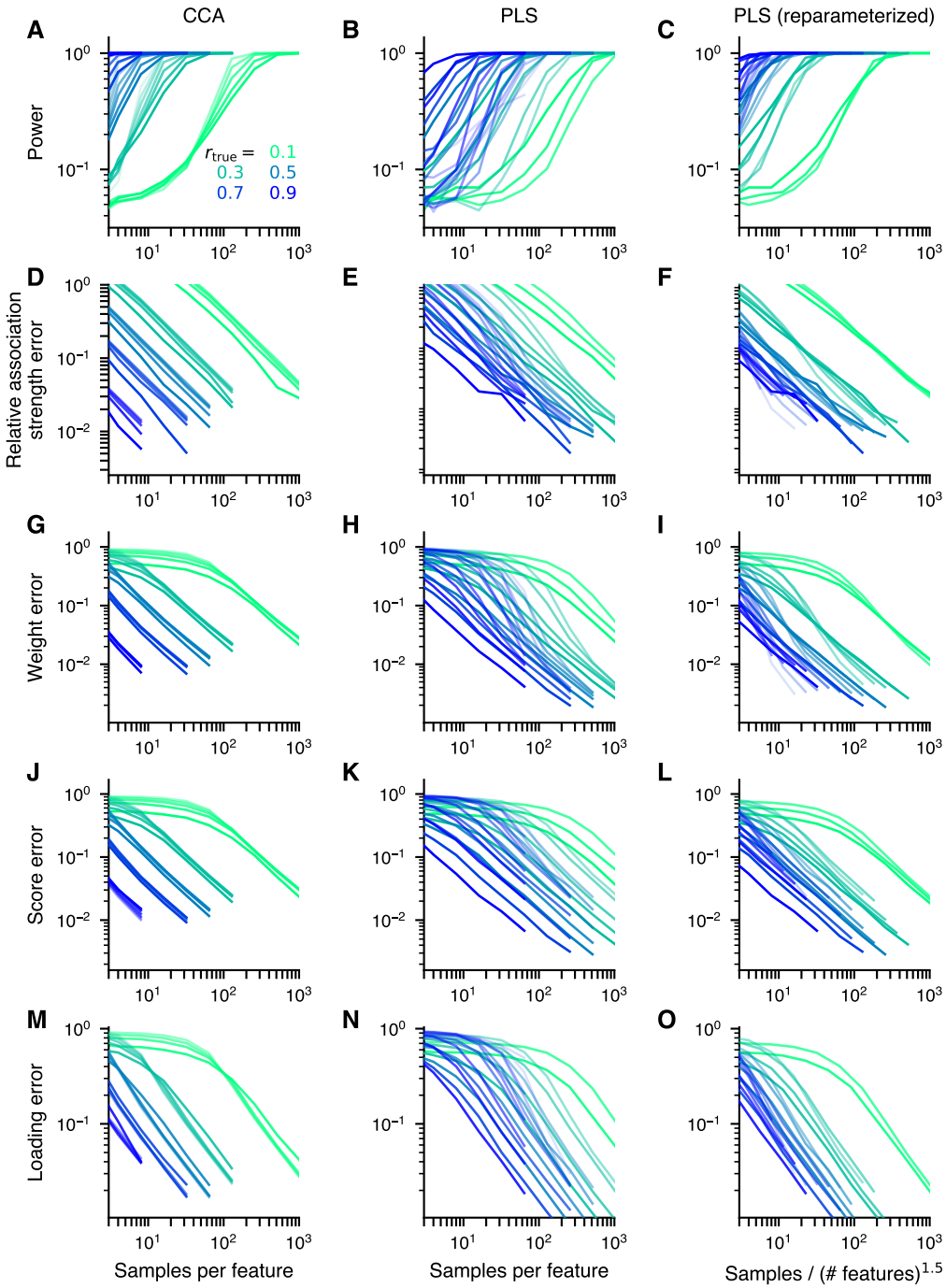

Fig. S3. Samples per feature is a key effective parameter (I). Throughout the manuscript we have presented results in terms of the parameter "samples per feature". Here we demonstrate that this is a suitable parameterization for CCA, while only approximately so for PLS. Color hue represents true correlation r_{true}, saturated colors are used for $p_X = p_Y = 2$, and fainter colors for higher p_X (and we used $p_Y = p_X$). In CCA (left 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 (middle 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 (right column), the curves overlap more. See also Fig. S4. Of note, the downstream effect of the “samples per feature” parameterization can be seen in Fig. 8B, where each dot represents a particular number of features: the dots for a given r_{true} do not scatter appreciatively for CCA but do for PLS. Likewise, $\log(p_x + p_y)$ was used as a predictor in the linear model used for prediction of $\log(n)$, and the corresponding coefficient was around 1 (indicating $n \sim p$) for CCA, but above 1 for PLS (Fig. S17A).

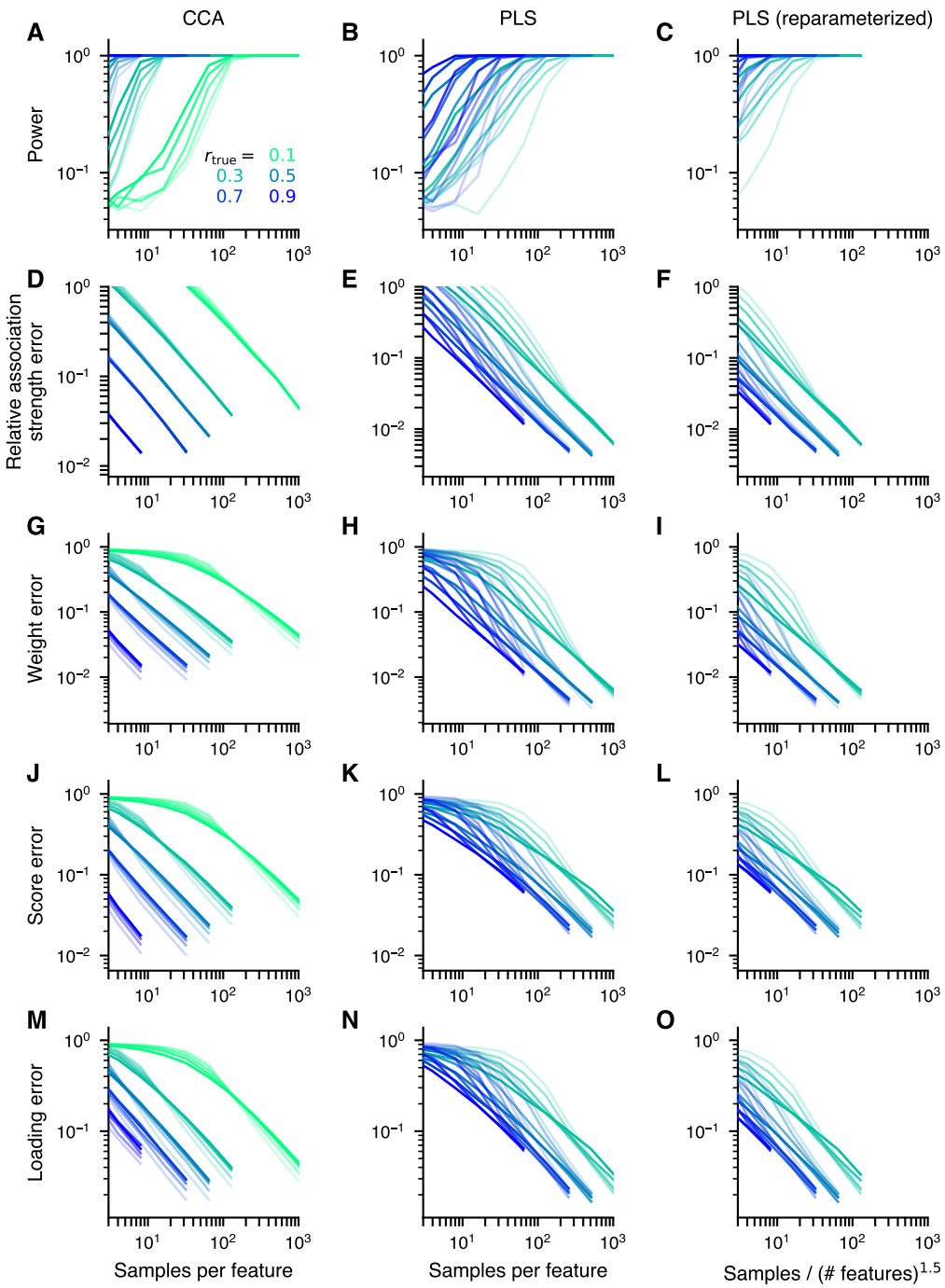

Fig. S4. Samples per feature is a key effective parameter (II). Same format as Fig. S3. Unlike there, here we used $p_X \neq p_Y$ and $p_X + p_Y$ was fixed at 64. Also unlike in Fig. S3, the difference between the middle and right column is smaller, suggesting that "samples per feature" is a good parameterization for PLS when the total number of features is fixed, but p_X and p_Y are not necessarily identical. Note that for PLS $r_{\text{true}} = 0.1$ is not shown due to computational expense.

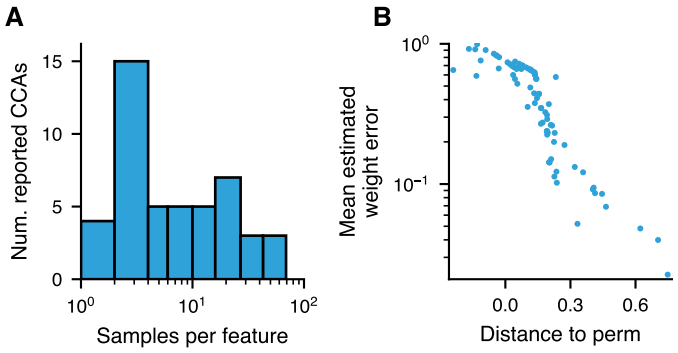

Fig. S5. Supplementary results related to meta-analysis. **A)** Typical number of samples per feature in brain-behavior CCAs. Studies using CCA to analyze brain-behavior relationships often used less than 5 samples per feature. **B)** Distance from null in *subjects-per-feature vs observed correlation* 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.

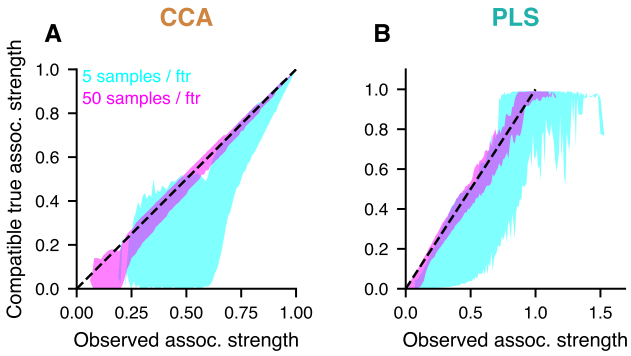

Fig. S6. A wide range of true association strengths is compatible with a given observed association strength. Synthetic datasets were generated where the true correlation was varied from 0 to 0.99 in steps of 0.01 and analyzed with **A)** CCA, **B)** PLS. We investigated 4, 8, 16, 32, 64 and 128 features 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.

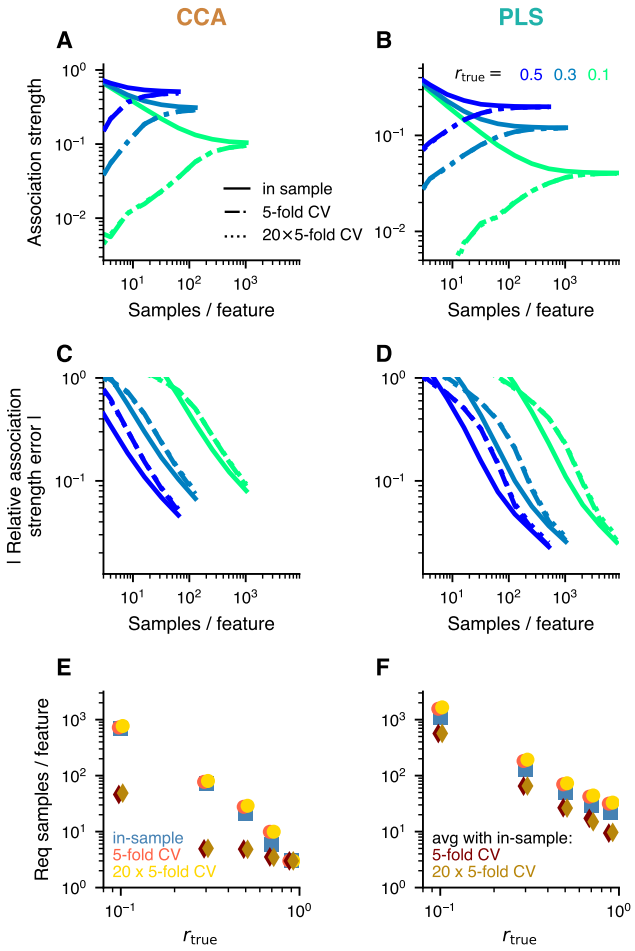

Fig. S7. Cross-validated estimation of association strength. In contrast to in-sample estimates, cross-validated estimates of between-set association strengths underestimate the true value. We tested two different cross-validation strategies here with very similar results (curves overlap): 5-fold cross-validation (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 %).

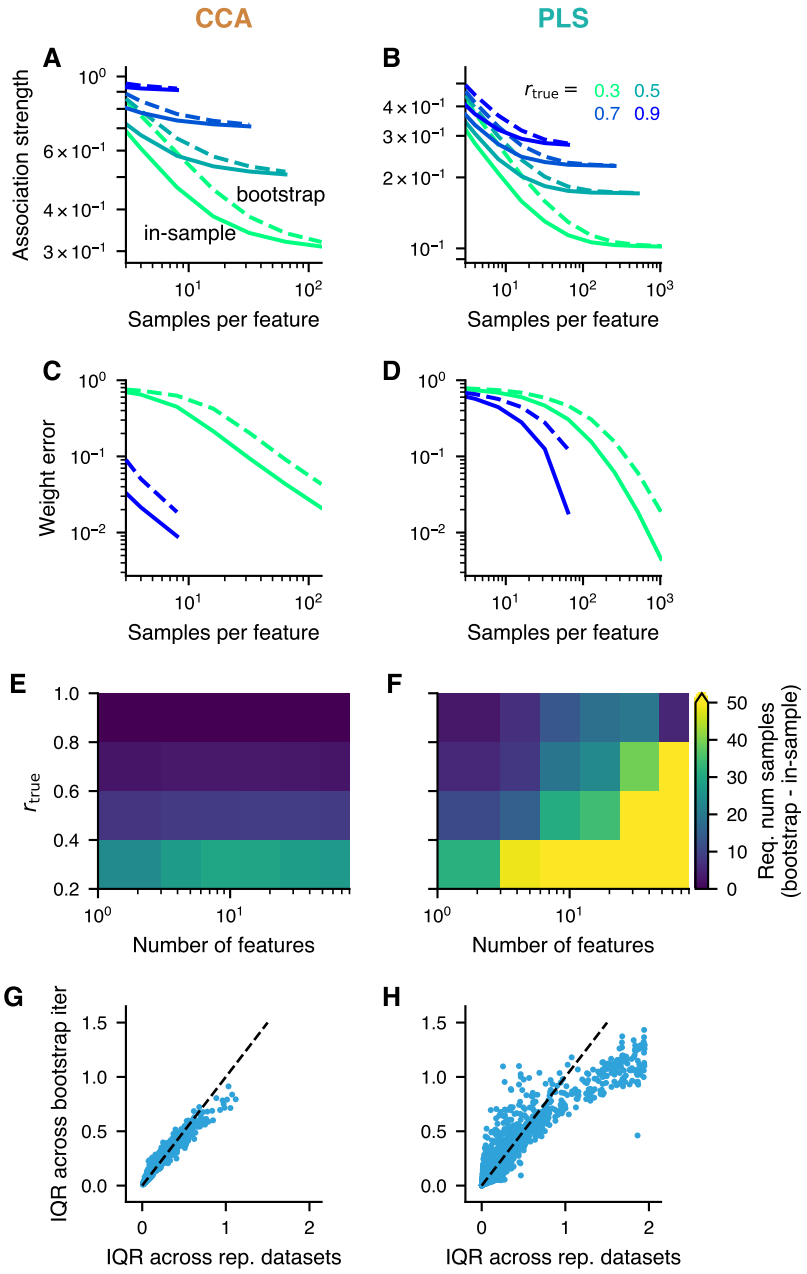

Fig. S8. Bootstrapped estimates are biased but recover variability. Bootstrapping affects CCA (left column) and PLS (right column) in a similar manner. **A, B)** Bootstrapped association strengths averaged across 100 bootstrap iterations and repeated draws from a given normal distribution (dashed lines) are somewhat worse than estimates obtained from the full samples (solid lines) averaged across repetitions. Likewise, **C, D)** average weight errors and **E, F)** the number of samples required to obtain less than 10 % weight error are somewhat worse when estimated by bootstrapping. **G, H)** 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 $r_{\text{true}} \in \{0.3, 0.5, 0.7, 0.9\}$, $p_x \in \{2, 4, 8, 16, 32, 64\}$ ($p_y = p_x$) and 5 different covariance matrices (with different true weight vectors), the scatter-plots show one dot for each element of the weight vector.

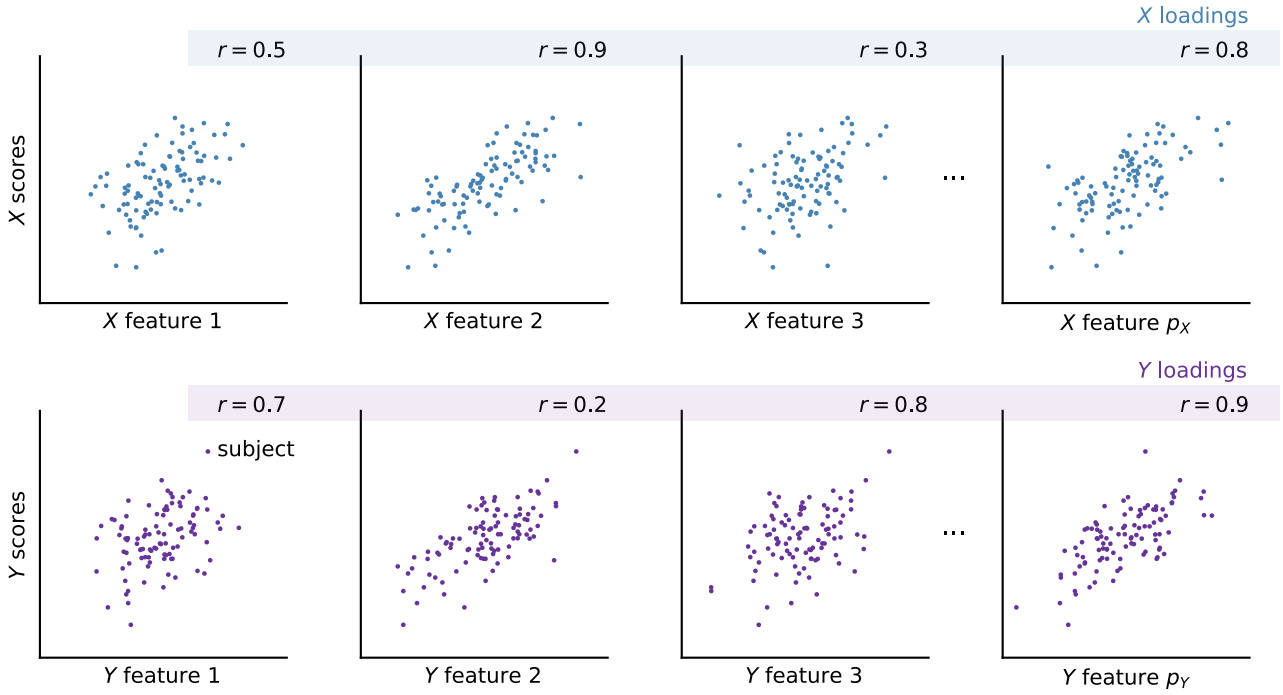

Fig. S9. Illustration of loadings. Loadings are defined as Pearson correlations across subjects of a feature with the CCA/PLS scores. The loadings vector contains these correlations for all variables. Apart from the illustrated loadings, *cross-loadings* in which scores of one set are correlated with the original features of the other set can also be computed.

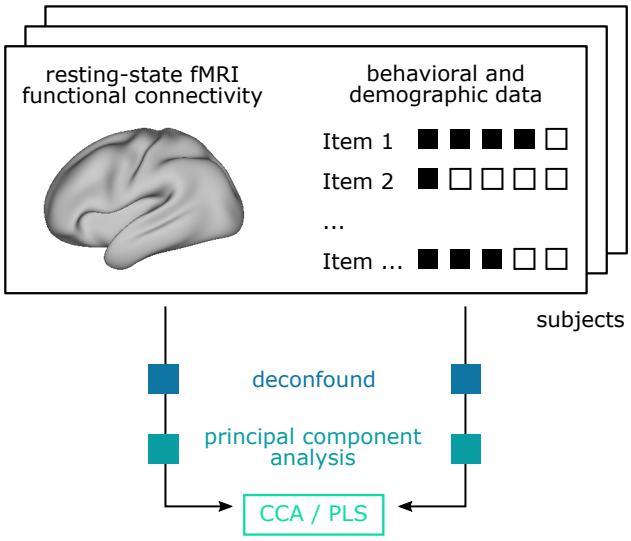

Fig. S10. HCP data analysis workflow. Resting-state functional connectivity data and behavioral and demographic data from corresponding subjects were separately deconfounded, reduced to 100 principal components and then analyzed with CCA and PLS.

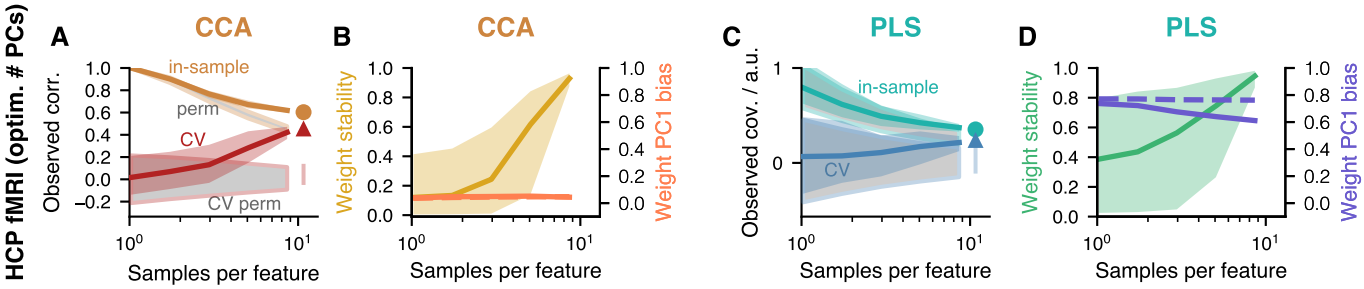

Fig. S11. Re-analysis of HCP fMRI vs behavior data with optimized number of principal components. Format is identical to Fig. 5. The only difference is the number of principal components retained for analysis: whereas in Fig. 5 100 principal components were used for both datasets, in agreement with previous studies of HCP data (39, 55, 60, 65, 72, 83), here we chose the number of principal component with the "max-min detector" from (84). 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 $p_X = 55$ and $p_Y = 33$, along with 12 between-set modes (we only consider the first one here). p-values for CCA and PLS were, respectively, 0.001 and 0.007. While the results are very similar to Fig. 5, (i) the observed correlations in **A)** 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.

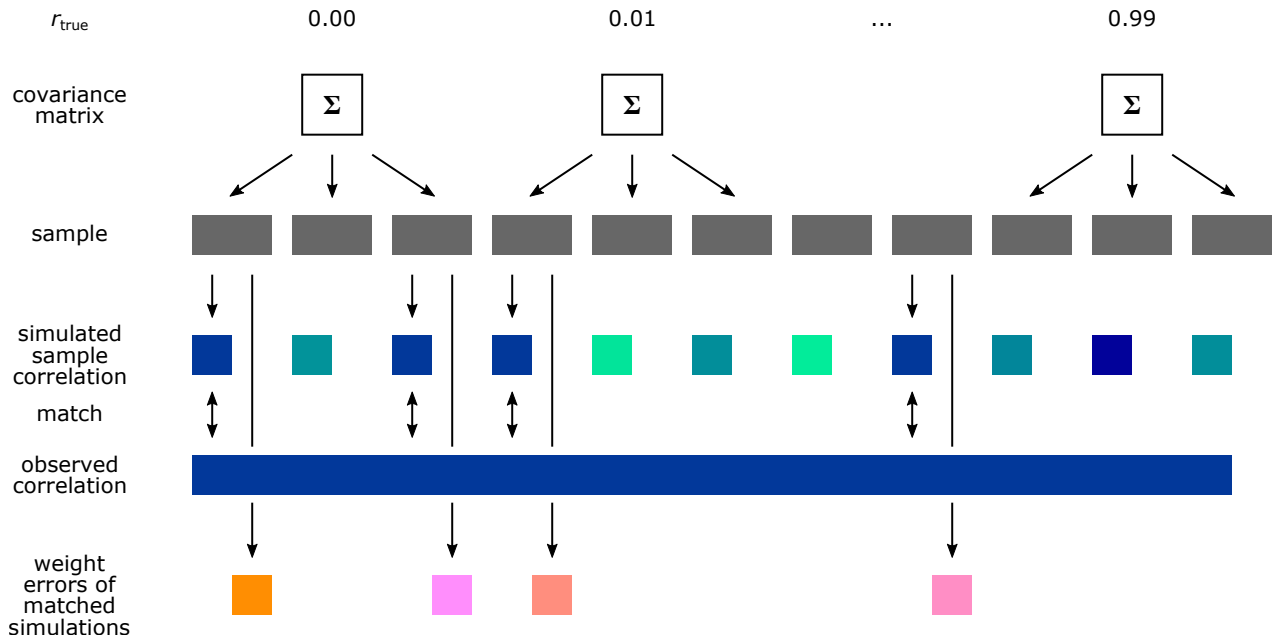

Fig. S12. Schematic for estimating weight errors for published CCA results. For each CCA from the literature in our database, synthetic data for CCA is generated with matching number of samples and features. Separate datasets are generated for assumed ground-truth correlations varying between 0 and 0.99. 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.

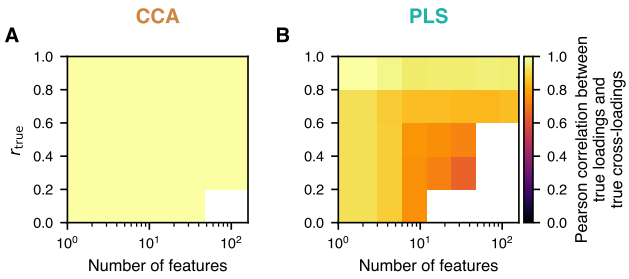

Fig. S13. True loadings and true cross-loadings are similar. True loadings and cross-loadings were calculated with $\mathrm{Eq.}\ (3)$ and $\mathrm{Eq.}\ (4)$, respectively. **A)** In CCA, true loadings and true cross-loadings were collinear (as predicted by $\mathrm{Eq.}\ (22)$). **B)** For PLS, they were strongly correlated. The shown correlations were averaged across 25 covariance matrices with different true weight vectors. Moreover, for PLS, $a_x + a_y$ was constrained to -2.

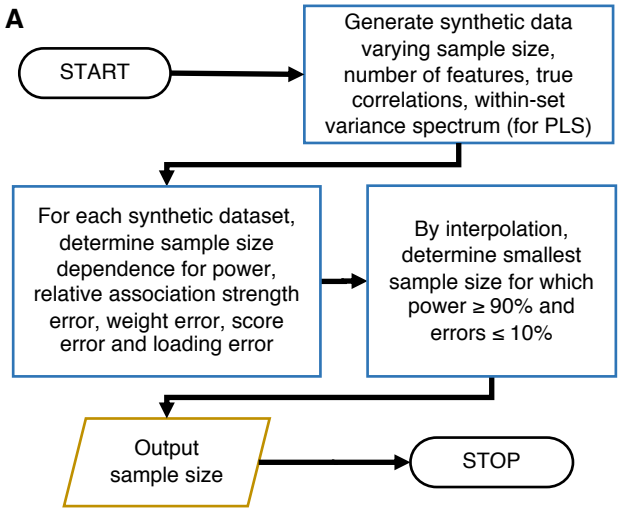

Fig. S14. Algorithm for sample size calculation. Sample sizes can, in principle, be calculated directly with GEMMR, as shown in Fig. 8. However, this is computationally expensive. To quickly obtain sample size estimates, we developed the algorithm illustrated here.

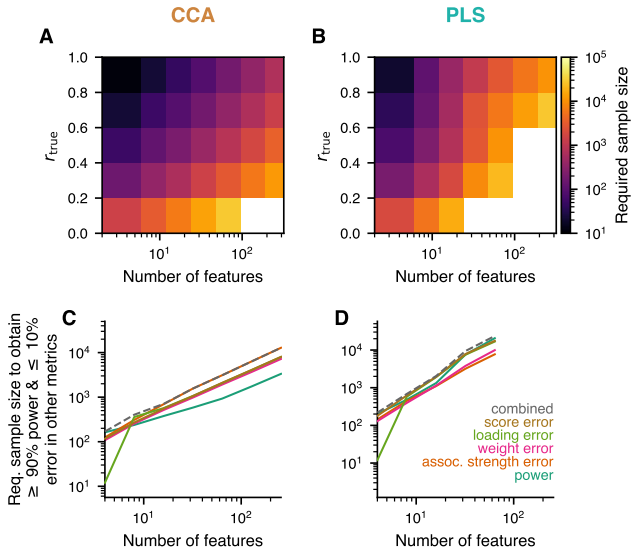

Fig. S15. Sample size dependence on ground truth correlation r_{true}, number of features and metric. A-B) Required sample sizes based on the combined criterion increase with number of features and for low true correlations. Due to computational expense values for some parameter sets were not available (white). **C-D)** Sample size dependence on number of features, shown here for $r_{\mathrm{true}} = 0.3$, scale similarly for all metrics, albeit with slight offsets.

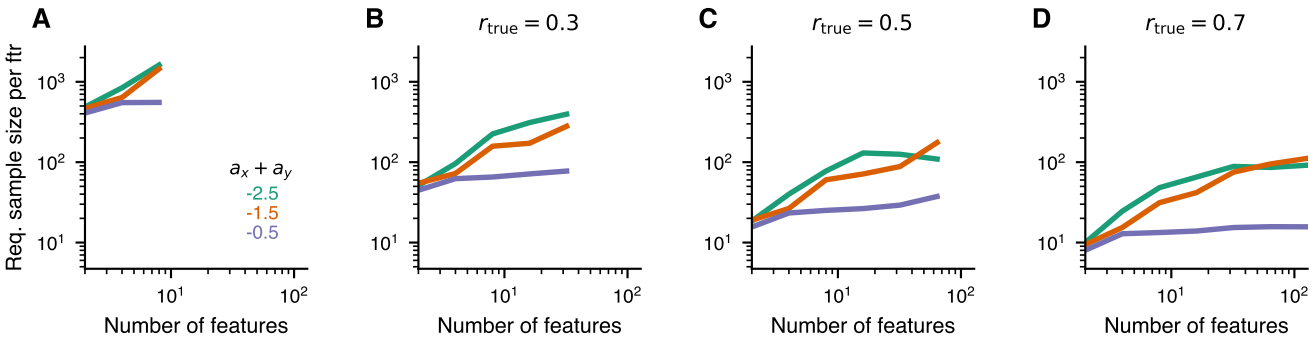

Fig. S16. Sample size dependence of PLS on within-set variances. Simulated parameter sets were averaged across subsets having indicated values for $a_x + a_y$ (the sum of within-set power-law decay constants) ± 0.5. The closer $a_x + a_y$ was to 0 (i. e. the "whiter" the data) the fewer samples were required.

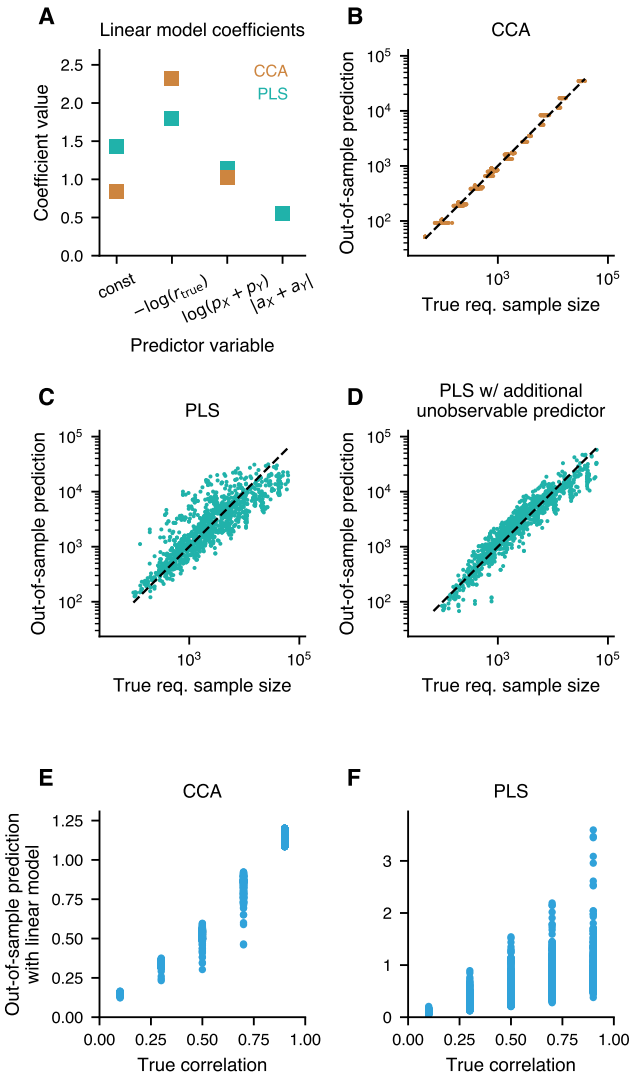

Fig. S17. Sample size calculator. Especially for low assumed ground-truth correlations and a high number of features it is computationally expensive to estimate the required number of samples by generating synthetic datasets and searching the sample size such that error bounds are satisfied. To abbreviate this process we pre-calculate required sample sizes using the generative model approach for certain parameter values, fit a linear model to $\log(n_{\text{required}})$ and then use it to quickly interpolate for parameter values not in the pre-calculated database. Predictors for the linear model are $-\log(r_{\text{true}})$, $\log(p_x + p_y)$ and, for PLS only, $|a_x + a_y|$, where 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. 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. **A)** Linear model coefficients for CCA and PLS. **B-D)** The pre-calculated database was split in half where one half corresponded to $r_{\text{true}} = 0.1$ and 0.3, the other to $r_{\text{true}} = 0.7$ and 0.9 and entries for $r_{\text{true}} = 0.5$ were divided between the two halves. The linear model was re-estimated separately for each half, and used to predict the other half. Predictions are **B)** good for CCA and **C)** somewhat worse for PLS. **D)** In contrast to CCA which effectively uses whitened data, PLS weights, even for a given ground-truth correlation, can differ by their direction relative to the principal component axes and thus by how much within-set variance the corresponding scores explain. In practice this is unobservable as it requires the knowledge of the true weight vectors. Using synthetic data where we know the true weight vectors we re-estimated the linear model for PLS with $\log(v_x v_y)$ as additional predictor, where v indicates the explained variance ratio for the weight vector, i. e. the variance of the scores divided by the trace of the corresponding within-set covariance matrix. This results in much better predictions and indicates that much of the unexplained variance in **C)** is due to unobservables. **E, F)** Solving the linear model for r_{true}, we aim to predict correlations. We train the model using either simulation outcomes for $r_{\text{true}} \in \{0.1, 0.3\}$, or $r_{\text{true}} \in \{0.7, 0.9\}$ and testing the predictions on the remaining r_{true}s. **E)** Good predictions can be obtained in that way for CCA, **F)** but not for PLS.

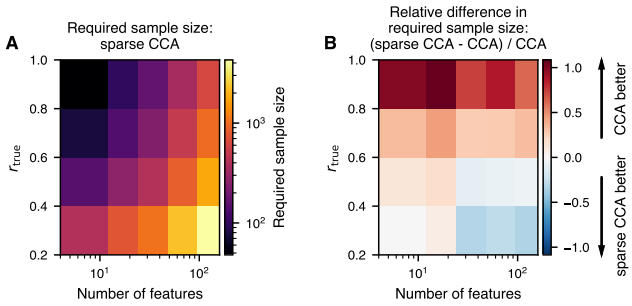

Fig. S18. Sparse CCA. We determined required sample sizes with our analysis pipeline, for the sparse CCA variant *PMD* (11). Due to the computational expense we ran only 6 repetitions per cell, 5 and 4, respectively, for the 2 right-most cells on the bottom. **A)** Required sample sizes increased with the number of features and with decreasing r_{true}. Layout is analogous to Fig. S15A-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.

Table S1. 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. (84).
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 bias	In PLS, weights can be biased towards the first principal component.
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. For PLS, 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.

References

1. Smith SM, Nichols TE (2018) Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron* 97(2):263–268.
2. Rosipal R, Krämer N (2006) Overview and Recent Advances in Partial Least Squares in *Subspace, Latent Structure and Feature Selection*, Lecture Notes in Computer Science, eds. Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J. (Springer Berlin Heidelberg), pp. 34–51.
3. Wegelin JA (2000) A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. *University of Washington, Department of Statistics, Tech. Rep.*
4. Abdi H, Williams LJ (2013) Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression in *Computational Toxicology*, eds. Reisfeld B, Mayeno AN. (Humana Press, Totowa, NJ) Vol. 930, pp. 549–579.
5. Hotelling H (1936) Relations Between Two Sets of Variates. *Biometrika* 28(3/4):321–377.
6. Mardia KV, Kent JT, Bibby JM (1979) *Multivariate analysis*. (Academic Press), 10 edition.
7. Härdle WK, Simar L (2019) *Applied Multivariate Statistical Analysis*. (Springer International Publishing, Cham).
8. Uurtio V, et al. (2017) A Tutorial on Canonical Correlation Methods. *ACM Computing Surveys (CSUR)* 50(6):95:1–95:33.
9. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P (2008) A Sparse PLS for Variable Selection when Integrating Omics Data. *Statistical Applications in Genetics and Molecular Biology* 7(1).
10. Parkhomenko E, Tritchler D, Beyene J (2009) Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. *Statistical Applications in Genetics and Molecular Biology* 8(1):1–34.
11. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10(3):515–534.
12. Tenenhaus A, et al. (2014) Variable selection for generalized canonical correlation analysis. *Biostatistics* 15(3):569–583.
13. Avants BB, Cook PA, Ungar L, Gee JC, Grossman M (2010) Dementia induces correlated reductions in white matter integrity and cortical thickness: A multivariate neuroimaging study with sparse canonical correlation analysis. *NeuroImage* 50(3):1004–1016.
14. Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y (2012) Relating drug–protein interaction network with drug side effects. *Bioinformatics* 28(18):i522–i528. Publisher: Oxford Academic.
15. Yahata N, et al. (2016) A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nature Communications* 7(1):1–12. Number: 1 Publisher: Nature Publishing Group.
16. Xia CH, et al. (2018) Linked dimensions of psychopathology and connectivity in functional brain networks. *Nature Communications* 9(1):3003.
17. Stuart T, et al. (2019) Comprehensive Integration of Single-Cell Data. *Cell* 177(7):1888–1902.e21.
18. Mihalik A, et al. (2020) Multiple Holdouts With Stability: Improving the Generalizability of Machine Learning Analyses of Brain–Behavior Relationships. *Biological Psychiatry* 87(4):368–376.
19. Zhuang X, Yang Z, Cordes D (2020) A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping* p. hbm.25090.
20. Witten D, Tibshirani R (2020) PMA: Penalized Multivariate Analysis.
21. Le Floch E, et al. (2012) Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage* 63(1):11–24.
22. Storn R, Price K (1997) Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization* 11(4):341–359.
23. Mathematics Stack Exchange (2018) x-coordinate distribution on the n-sphere. Library Catalog: math.stackexchange.com.
24. Van Essen DC, et al. (2013) The WU-Minn Human Connectome Project: An overview. *NeuroImage* 80:62–79.
25. Miller KL, et al. (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience* 19(11):1523–1536.
26. Human Connectome Project (2017) 1200 Subjects Data Release Reference, Technical report.
27. Glasser MF, et al. (2013) The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* 80:105–124.
28. Salimi-Khorshidi G, et al. (2014) Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage* 90:449–468.
29. Griffanti L, et al. (2014) ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage* 95:232–247.
30. Robinson EC, et al. (2014) MSM: A new flexible framework for Multimodal Surface Matching. *NeuroImage* 100:414–426.
31. Power JD, et al. (2018) 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(9):E2105–E2114.
32. Glasser MF, et al. (2016) A multi-modal parcellation of human cerebral cortex. *Nature* 536(7615):171–178.
33. Mars RB, et al. (2018) Whole brain comparative anatomy using connectivity blueprints. *eLife* 7:e35237.
34. Warrington S, et al. (2020) XTRACT - Standardised protocols for automated tractography in the human and macaque brain. *NeuroImage* p. 116923.
35. Sotiropoulos SN, et al. (2013) Advances in diffusion MRI acquisition and processing in the Human Connectome Project. *NeuroImage* 80:125–143.
36. Behrens TEJ, Berg HJ, Jbabdi S, Rushworth MFS, Woolrich MW (2007) Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage* 34(1):144–155.
37. Hernandez-Fernandez M, et al. (2019) Using GPUs to accelerate computational diffusion MRI: From microstructure

514 estimation to tractography and connectomes. *NeuroImage* 188:598–615.
515 38. Desikan RS, et al. (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into
516 gyral based regions of interest. *NeuroImage* 31(3):968–980.
517 39. Smith SM, et al. (2015) A positive-negative mode of population covariation links brain connectivity, demographics and
518 behavior. *Nature Neuroscience* 18(11):1565–1567.
519 40. Alfaro-Almagro F, et al. (2018) Image processing and Quality Control for the first 10,000 brain imaging datasets from UK
520 Biobank. *NeuroImage* 166:400–424.
521 41. Beckmann CF, Smith SM (2004) Probabilistic independent component analysis for functional magnetic resonance imaging.
522 *IEEE Transactions on Medical Imaging* 23(2):137–152.
523 42. Smith SM, Hyvärinen A, Varoquaux G, Miller KL, Beckmann CF (2014) Group-PCA for very large fMRI datasets.
524 *NeuroImage* 101:738–749.
525 43. Hyvärinen A, Oja E (1997) A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*
526 9(7):1483–1492. Publisher: MIT Press.
527 44. Sudlow C, et al. (2015) UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex
528 Diseases of Middle and Old Age. *PLOS Medicine* 12(3):e1001779. Publisher: Public Library of Science.
529 45. Beasley TM, Erickson S, Allison DB (2009) Rank-Based Inverse Normal Transformations are Increasingly Used, But are
530 They Merited? *Behavior Genetics* 39(5):580.
531 46. Seabold S, Perktold J (2010) Statsmodels: Econometric and Statistical Modeling with Python in *9th Python in Science*
532 *Conference*.
533 47. Winkler AM, Webster MA, Vidaurre D, Nichols TE, Smith SM (2015) Multi-level block permutation. *NeuroImage*
534 123:253–268.
535 48. Akisaki T, et al. (2006) Cognitive dysfunction associates with white matter hyperintensities and subcortical atrophy
536 on magnetic resonance imaging of the elderly diabetes mellitus Japanese elderly diabetes intervention trial (J-EDIT).
537 *Diabetes/Metabolism Research and Reviews* 22(5):376–384.
538 49. Tsvetanov KA, et al. (2016) Extrinsic and Intrinsic Brain Network Connectivity Maintains Cognition across the Lifespan
539 Despite Accelerated Decay of Regional Brain Activation. *Journal of Neuroscience* 36(11):3115–3126.
540 50. Ball G, et al. (2017) Multimodal image analysis of clinical influences on preterm brain development. *Annals of Neurology*
541 82(2):233–246.
542 51. Drysdale AT, et al. (2017) Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature*
543 *Medicine* 23(1):28–38.
544 52. Marquand AF, Haak KV, Beckmann CF (2017) Functional corticostriatal connection topographies predict goal-directed
545 behaviour in humans. *Nature Human Behaviour* 1(8):1–9.
546 53. Perry A, et al. (2017) The independent influences of age and education on functional brain networks and cognition in
547 healthy older adults. *Human Brain Mapping* 38(10):5094–5114.
548 54. Rahim M, Thirion B, Varoquaux G (2017) Population-Shrinkage of Covariance to Estimate Better Brain Functional
549 Connectivity in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, Lecture Notes in Computer
550 Science, eds. Descoteaux M, et al. (Springer International Publishing, Cham), pp. 460–468.
551 55. Bijsterbosch JD, et al. (2018) The relationship between spatial configuration and functional connectivity of brain regions.
552 *eLife* 7:e32992.
553 56. Lin HY, et al. (2018) Brain–behavior patterns define a dimensional biotype in medication-naïve adults with attention-deficit
554 hyperactivity disorder. *Psychological Medicine* 48(14):2399–2408.
555 57. Lin SJ, Baumeister TR, Garg S, McKeown MJ (2018) Cognitive Profiles and Hub Vulnerability in Parkinson's Disease.
556 *Frontiers in Neurology* 9.
557 58. Rodrigue AL, et al. (2018) Multivariate Relationships Between Cognition and Brain Anatomy Across the Psychosis
558 Spectrum. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3(12):992–1002.
559 59. Becker R, Hervais-Adelman A (2019) Resolving the connectome - Spectrally-specific functional connectivity networks and
560 their distinct contributions to behaviour. *bioRxiv* p. 700278.
561 60. Bijsterbosch JD, Beckmann CF, Woolrich MW, Smith SM, Harrison SJ (2019) The relationship between spatial configuration
562 and functional connectivity of brain regions revisited. *eLife* 8:e44890.
563 61. Dinga R, et al. (2019) Evaluating the evidence for biotypes of depression: Methodological replication and extension of
564 Drysdale et al. (2017). *NeuroImage: Clinical* p. 101796.
565 62. Manza P, Shokri-Kojori E, Volkow ND (2019) Reduced Segregation Between Cognitive and Emotional Processes in
566 Cannabis Dependence. *Cerebral Cortex*.
567 63. Menon V, et al. (2019) Quantitative modeling links in vivo microstructural and macrofunctional organization of human
568 and macaque insular cortex, and predicts cognitive control abilities. *bioRxiv* p. 662601.
569 64. Mihalik A, et al. (2019) Brain-behaviour modes of covariation in healthy and clinically depressed young people. *Scientific*
570 *Reports* 9(1):1–11.
571 65. Li J, et al. (2019) Topography and behavioral relevance of the global signal in the human brain. *Scientific Reports*
572 9(1):1–10.
573 66. Pusil S, Dimitriadis SI, López ME, Pereda E, Maestú F (2019) Aberrant MEG multi-frequency phase temporal synchro-
574 nization predicts conversion from mild cognitive impairment-to-Alzheimer's disease. *NeuroImage: Clinical* 24:101972.

67. Rodriguez C, et al. (2019) Structural Correlates of Personality Dimensions in Healthy Aging and MCI. *Frontiers in Psychology* 9.
68. Supekar K, Cai W, Krishnadas R, Palaniyappan L, Menon V (2019) Dysregulated Brain Dynamics in a Triple-Network Saliency Model of Schizophrenia and Its Relation to Psychosis. *Biological Psychiatry* 85(1):60–69.
69. Yu M, et al. (2019) Childhood trauma history is linked to abnormal brain connectivity in major depression. *Proceedings of the National Academy of Sciences* 116(17):8582–8590.
70. Zarnani K, et al. (2019) Discovering markers of healthy aging: a prospective study in a Danish male birth cohort. *Aging* 11(16).
71. Alnæs D, Kaufmann T, Marquand AF, Smith SM, Westlye LT (2020) Patterns of sociocognitive stratification and perinatal risk in the child brain. *Proceedings of the National Academy of Sciences*. Publisher: National Academy of Sciences Section: Biological Sciences.
72. Han F, Gu Y, Brown GL, Zhang X, Liu X (2020) Neuroimaging contrast across the cortical hierarchy is the feature maximally linked to behavior and demographics. *NeuroImage* 215:116853.
73. Hudgens-Haney ME, et al. (2020) Cognitive Impairment and Diminished Neural Responses Constitute a Biomarker Signature of Negative Symptoms in Psychosis. *Schizophrenia Bulletin*.
74. Jolly AE, Scott GT, Sharp DJ, Hampshire AH (2020) Distinct patterns of structural damage underlie working memory and reasoning deficits after traumatic brain injury. *Brain* 143(4):1158–1176. Publisher: Oxford Academic.
75. Mason NL, et al. (2020) Me, myself, bye: regional alterations in glutamate and the experience of ego dissolution with psilocybin. *Neuropsychopharmacology* pp. 1–9. Publisher: Nature Publishing Group.
76. Menon V, et al. (2020) Microstructural organization of human insula is linked to its macrofunctional circuitry and predicts cognitive control. *eLife* 9:e53470.
77. Trotti RL, et al. (2020) Electrophysiological correlates of emotional scene processing in bipolar disorder. *Journal of Psychiatric Research* 120:83–90.
78. Buitinck L, et al. (2013) API design for machine learning software: experiences from the scikit-learn project. *arXiv:1309.0238 [cs]*. arXiv: 1309.0238.
79. Hoyer S, Hamman JJ (2017) xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software* 5:10.
80. McKinney W (2010) Data structures for statistical computing in python in *Proceedings of the 9th Python in Science Conference*. (Austin, TX), Vol. 445, pp. 51–56.
81. Virtanen P, et al. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* 17(3):261–272. Number: 3 Publisher: Nature Publishing Group.
82. van der Walt S, Colbert SC, Varoquaux G (2011) The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* 13(2):22–30.
83. Rahim M, Thirion B, Bzdok D, Buvat I, Varoquaux G (2017) Joint prediction of multiple scores captures better individual traits from brain images. *NeuroImage* 158:145–154.
84. Song Y, Schreier PJ, Ramírez D, Hasija T (2016) Canonical correlation analysis of high-dimensional data with very small sample support. *Signal Processing* 128:449–458.