NeuroImage 216 (2020) 116745

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists

Hao-Ting Wang [a,b,*], Jonathan Smallwood [a], Janaina Mourao-Miranda [c,d], Cedric Huchuan Xia [e], Theodore D. Satterthwaite [e], Danielle S. Bassett [f,g,h,i], Danilo Bzdok [j,k,l,m,n,**]

[a] Department of Psychology, University of York, Heslington, York, United Kingdom
[b] Sackler Center for Consciousness Science, University of Sussex, Brighton, United Kingdom
[c] Centre for Medical Image Computing, Department of Computer Science, University College London, London, United Kingdom
[d] Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, United Kingdom
[e] Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA
[f] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, 19104, USA
[g] Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, 19104, USA
[h] Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA
[i] Department of Physics & Astronomy, School of Arts & Sciences, University of Pennsylvania, Philadelphia, PA, 19104, USA
[j] Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, Germany
[k] JARA-BRAIN, Jülich-Aachen Research Alliance, Germany
[l] Parietal Team, INRIA, Neurospin, Bat 145, CEA Saclay, 91191, Gif-sur-Yvette, France
[m] Department of Biomedical Engineering, Montreal Neurological Institute, Faculty of Medicine, McGill University, Montreal, Canada
[n] Mila - Quebec Artificial Intelligence Institute, Canada

ARTICLE INFO

Keywords:
Machine learning
Big data
Data science
Deep phenotyping
Modality fusion

ABSTRACT

The 21st century marks the emergence of "big data" with a rapid increase in the availability of datasets with multiple measurements. In neuroscience, brain-imaging datasets are more commonly accompanied by dozens or hundreds of phenotypic subject descriptors on the behavioral, neural, and genomic level. The complexity of such "big data" repositories offer new opportunities and pose new challenges for systems neuroscience. Canonical correlation analysis (CCA) is a prototypical family of methods that is useful in identifying the links between variable sets from different modalities. Importantly, CCA is well suited to describing relationships across multiple sets of data, such as in recently available big biomedical datasets. Our primer discusses the rationale, promises, and pitfalls of CCA.

1. Introduction

The parallel developments of large biomedical datasets and increasing computational power have opened new avenues with which to understand relationships among brain, cognition, and disease. Similar to the advent of microarrays in genetics, brain-imaging and extensive behavioral phenotyping yield datasets with tens of thousands of variables (Efron, 2010). Since the beginning of the 21st century, the improvements and availability of technologies, such as functional magnetic resonance imaging (fMRI), have made it feasible to collect large neuroscience datasets (Poldrack and Gorgolewski, 2014). At the same time, problems in reproducing the results of key studies in neuroscience and psychology have highlighted the importance of drawing robust conclusions based on rich datasets (Open Science Collaboration, 2015).

The UK Biobank, for example, is a prospective population study with 500,000 participants and comprehensive imaging data, genetic information, and environmental measures on mental disorders and other diseases (Allen et al., 2012; Miller et al., 2016). Similarly, the Human Connectome Project (van Essen et al., 2013) has recently completed brain-imaging of >1000 young adults, with high spatial and temporal resolution, featuring approximately 4 h of brain scanning per participant. Further, the Enhanced Nathan Kline Institute Rockland Sample (Nooner

* Corresponding author. Sackler Center for Consciousness Science, University of Sussex, Brighton, United Kingdom.
** Corresponding author. Department of Biomedical Engineering, Montreal Neurological Institute, McGill University, Mila - Quebec Artificial Intelligence Institute, Montreal, Canada.
E-mail addresses: H.Wang@bsms.ac.uk (H.-T. Wang), danilo.bzdok@mcgill.ca (D. Bzdok).

https://doi.org/10.1016/j.neuroimage.2020.116745
Received 2 December 2019; Received in revised form 12 February 2020; Accepted 12 March 2020
Available online 8 April 2020

et al., 2012) and the Cambridge Centre for Aging and Neuroscience (Shafto et al., 2014; Taylor et al., 2017) offer cross-sectional studies (n > 700) across the lifespan (18–87 years of age) in large population samples. By providing rich datasets that include measures of brain imaging, measures of cognition, demographics, and neuropsychological assessments, such studies can help quantify developmental trajectories in cognition as well as the relationships between brain structure and function. While "deep" phenotyping and such unprecedented sample sizes provide opportunities for more robust descriptions of subtle population variability, the abundance of measurement for each subject does not come without challenges.

Modern datasets often provide more variables than observations of these variable sets (Bzdok et al., 2019; Bzdok and Yeo, 2017; Smith and Nichols, 2018). In this situation, classical statistical approaches can often fail to fully capitalize on the potential of these datasets. For example, even with large samples the number of participants is often smaller than the number of brain locations that have been sampled in high-resolution brain scans. On the other hand, in datasets with a particularly high number of participants, traditional statistical approaches will identify associations that are highly statistically significant but may only account for a small fraction of the variation in the data (Miller et al., 2016; Smith and Nichols, 2018). In such scenarios, investigators who aim to exploit the full potential of big datasets to reveal important relationships between brain, cognition, genes and disease require techniques that are better suited to the nature of their data than are many of the traditional statistical tools.

Canonical correlation analysis (CCA) is one tool that is useful in unlocking the complex relationships among many variables in large datasets. A key strength of CCA is that it can simultaneously evaluate two different sets of variables, without assuming any particular form of precedence or directionality (such as in many types of partial least squares analyses, cf. section 4.2). For example, CCA allows a data matrix of brain measurements (e.g., connectivity links between a set of brain regions) to be simultaneously analyzed with respect to a second data matrix of behavioral measurements (e.g., participant responses to items on various questionnaires). In other words, CCA identifies the sources of common variation in two high-dimensional sets of variables.

CCA is a multivariate statistical method that was introduced in the 1930s (Hotelling, 1936). However, CCA is more computationally expensive than many other common analysis tools and so is only recently becoming more popular in biomedical research. Moreover, the ability to accommodate two variable sets allows the identification of patterns that describe *many-to-many* relations. CCA, therefore, opens interpretational opportunities that go beyond techniques that map one-to-one relations (e.g., Pearson's correlation) or many-to-one relationships (e.g., ordinary multiple regression).

Early applications of CCA to neuroimaging data focused initially on its ability of filtering spatial signals (Cordes et al., 2012; Friman et al., 2004, 2003; 2001; Zhuang et al., 2017), classification (Hardoon et al., 2007), and more recently focused on the ability to combine different imaging modalities together (see Calhoun and Sui, 2016; Correa et al., 2010a for review). These include functional MRI and EEG (Sui et al., 2014) and grey and white matter (Lottman et al., 2018). Previous work used CCA to bring multiple imaging modalities to the same table, a process often referred to as *multi-modal fusion* (see Yang et al., 2019 for a review). CCA has also been applied to group-level analysis in task-based fMRI data instead of the traditional mass-univariate approach to improve the sensitivity to detect neural activation (Zhuang et al., 2019). However, with the recent trend towards rich phenotyping and massive cohort data collections, the imaging community has now increasingly recognized the capacity for CCA to provide compact multivariate solutions to big datasets. In this regime, CCA can efficiently chart links between brain, cognition, genes and disease (Calhoun and Sui, 2016; Hu et al., 2019, 2018; Liu and Calhoun, 2014; Marquand et al., 2017; Smith et al., 2015; Tsvetanov et al., 2016; Vatansever et al., 2017; Wang et al., 2018a; Xia et al., 2018).

Our conceptual primer describes how CCA can deepen understanding of large multi-variable datasets in fields such as cognitive neuroscience. We consider the modeling principles behind CCA and the circumstances in which this method can be useful by considering several recent applications of CCA in brain to behavior. Next, we consider the types of scientific conclusions that can be drawn from applications of the CCA algorithm, with a focus on the scope and limitations of applying this technique. Finally, we provide a set of practical guidelines for the implementation of CCA in scientific investigations.

2. Modeling intuitions

One way to appreciate the utility of CCA is by viewing it as an extension of principal component analysis (PCA). This widespread matrix decomposition technique identifies a set of latent dimensions as a linear approximation of the main components of variation that underlie the information contained in the original set of observations. In other words, PCA can re-express a set of correlated variables in a smaller number of hidden factors of variation. These latent sources of variation are not always directly observable in the original measurements, but together explain dominant motifs of how the actual observations are intrinsically organized.

PCA and other matrix-decomposition approaches have been used frequently in the area of personality research. For example, the 'Big Five' construct describes a set of personality traits that are identified by latent patterns that are revealed when PCA is applied to how people describe other people's time-enduring behavioral tendencies (Barrick and Mount, 1991). This approach tends to produce five reliable components that explain a substantial amount of meaningful variation in data gathered by personality assessments. A strength of a decomposition method like PCA is that it can produce a parsimonious description of the original dataset by re-expressing the set of variables as a few dimensional representations. These can often be amenable to human interpretation (such as the concept of introversion). The ability to re-express the original data in a more compact form, therefore, has both computational and statistical appeal (because it reduces the number of variables), as well as because it can also aid our interpretations of the problem (as it did by highlighting that the 'Big Five' are important dimensions of personality traits).

Although akin to PCA, CCA maximizes the linear correspondence between *two* sets of variables. The CCA algorithm, therefore, seeks dominant dimensions that describe shared variation across different sets of measures. In this way, CCA is particularly applicable when describing observations that bridge several levels of observation. Examples include charting correspondences between i) genetics and behavior, ii) brain and behavior, or iii) brain and genetics. In order to fully appreciate these qualities of CCA, it is helpful to consider how the assessment of the association between high-dimensional variable sets is achieved.

2.1. Mathematical notions

Canonical correlation analysis (Hotelling, 1936) determines the relationship between variable sets from two domains of measurement. Given X and Y of dimensions p and q from the same set of n observations, the first CCA mode is reflected in a linear combination of the variables in X and another linear combination of the variables in Y

$$U = a^T X, a \in R^p$$

$$V = b^T Y, b \in R^q$$

that maximize the first mode's linear association, that is, correlation

$$\rho = corr(U, V) = corr\left(a^T X, b^T Y\right).$$

In addition to optimizing the correspondence between U and V as the first canonical mode, it is possible to continue to seek additional pairs of

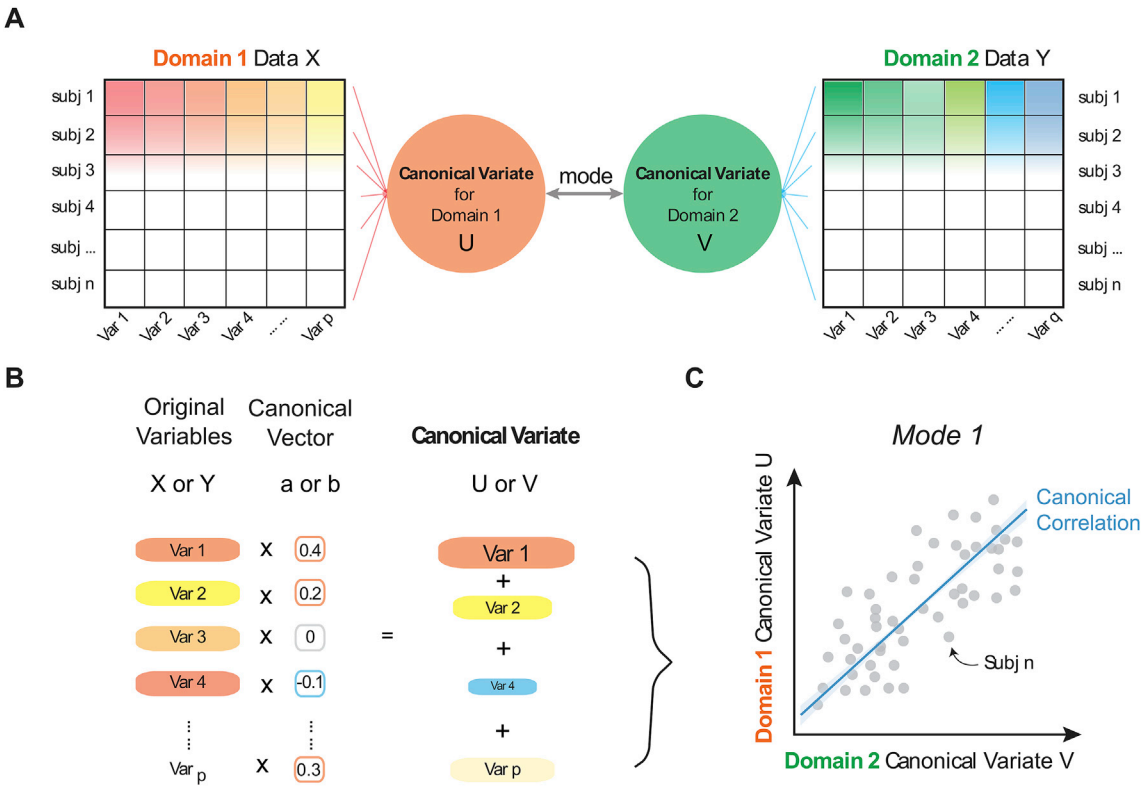

Fig. 1. A general schematic for Canonical Correlation Analysis (CCA).
(A) Multiple domains of data, with p and q variables respectively, measured in the same sample of participants can be submitted to co-decomposition by CCA. The algorithm seeks to re-express the datasets as multiple pairs of canonical variates that are highly correlated with each other across subjects. Each pair of the latent embedding of the left and right variable set is often referred to as 'mode' (Smith et al., 2015; Kernbach et al., 2018). (B) In each domain of data, the resulting canonical variate is composed of the weighted sum of variables by the canonical vector. (C) In a two-way CCA setting, each subject can thus be parsimoniously described by two canonical variates per mode, which are maximally correlated as represented here on the scatter plot. The linear correspondence between these two canonical variates is the canonical correlation - the chief performance metric used to estimate the parameters of a CCA model.

linear combinations that are uncorrelated with the first canonical mode(s). This process may be continued up to $min(p, q)$ times. In this primer, we will refer to a and b as the *canonical vectors*, and we will refer to U and V as the *canonical variates*. The *canonical correlation* denotes the correlation coefficient ρ of the canonical variates (see Fig. 1).

Let $\Sigma_{XX} = Cov(X,X) = X^T X$ *and* $\Sigma_{YY} = Cov(Y,Y) = Y^T Y$

and $\Sigma_{XY} = Cov(X,Y) = X^T Y$, so

$$\rho = \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a}\sqrt{b^T \Sigma_{YY} b}}$$

We can then reduce (1) to $\rho = a^T \Sigma_{XY} b$ (2) subject to the constraints above. Put differently, we define a change of basis (i.e., the coordinate system in which the data points live):

$$c = \Sigma_{xx}^{1/2} a$$

$$d = \Sigma_{yy}^{1/2} b$$

$$\rho = \frac{c^T \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} d}{\sqrt{c^T c}\sqrt{d^T d}}$$

The relationship between canonical vectors (a and b) and canonical variates (U and V) can also be expressed as:

$$U = c^T \Sigma_{xx}^{-1/2} X = a^T X$$

$$V = d^T \Sigma_{yy}^{-1/2} Y = b^T Y.$$

The relationship between the two original variable sets (X and Y) and the derived canonical variates U and V can be understood as the best way to rotate the left variable set and the right variable set from their original measurement spaces to new spaces in a way that maximizes their linear correlation. The fitted parameters of CCA thus describe the rotation of the coordinate systems: the canonical vectors encapsulating how to get from the original measurement coordinate system to the new latent space, the canonical variates encoding the embedding of each data point in that new space. This coordinate system rotation is formally related to singular value decomposition (SVD). SVD is perhaps the most common means to compute CCA (Healy, 1957). Assuming X and Y are centered, the CCA solution can be obtained by applying SVD to the correlation matrix $X^T Y / N$ (for detailed mathematical proof, see Uurtio et al., 2017).

From a practical application perspective, there are three properties of CCA that are perhaps particularly relevant for gaining insight into the variable-rich datasets that become more and more available in neuroscience: (1) joint-information compression, (2) multiplicity and (3) symmetry. We consider each of these properties in the following.

2.2. Joint information compression

A key feature of CCA is that it identifies the correspondence between two sets of variables, typically capturing two different levels of observation (e.g., brain and behavior). The salient relations among each set of variables is represented as a linear combination within each domain that together reflect common variation across both domains. Similar to PCA, CCA re-expresses data in form of high-dimensional linear representations (i.e., the canonical variates). Each resulting canonical variate is computed from the weighted sum of the original variable as indicated by

the canonical vector. Similar to PCA, CCA aims to compress the information within the relevant datasets by maximizing the linear correspondence between the low-rank projections from each set of observations, under the constraint of uncorrelated hidden dimensions (cf. multiplicity below). This means that the *canonical correlation quantifies the linear correspondence between the left and right variable sets based on Pearson's correlation between their canonical variates*; in other words how much the right and left variable set can be considered to approach each other in a common embedding space (Fig. 1). Canonical correlation, therefore, can be seen as a metric of successful joint information reduction between two variable arrays and, therefore, routinely serves as a performance measure for CCA that can be interpreted as the amount of achieved parsimony. Analogous to other multivariate modeling approaches, adding or removing even a single variable in one of the variable sets can lead to larger changes in the CCA solution (Hastie et al., 2001).

2.3. Symmetry

Another important feature of CCA is that the two co-analyzed variable sets can be exchanged without altering the nature of the solution. Many classical statistical approaches involve 'independent variables' or 'explanatory variables' which usually denote the model input (e.g., several questionnaire response items) as well as 'dependent variable' or 'response variable' which describes the model output (e.g., total working memory performance). However, such concepts lose their meaning in the context of CCA (Friston et al., 2008). Instead, the solutions provided by CCA reflect a description of how a unit change in one series of measurements is related to another series of measurements in another set of observations. These relationships are invariant to changes to which is the left vs. right flanking matrix to be jointly analyzed. In this way an important feature of CCA is that the two co-analyzed variable sets can be exchanged without altering the nature of the solution. We call this property of CCA 'symmetry'.

The symmetry in analysis and neuroscientific interpretation produced via CCA departs from many other multivariate methods, in which the dependent and independent variables play distinct roles in model estimation and domain interpretation. For instance, linear-regression-type methods account for the impact of a unit change in the (dependent) response variable as a function of the (independent) input variable. In this case, changing the dependent and independent variables can alter the nature of any specific result. A second important characteristic of CCA, therefore, is that the co-relationship between two sets of variables is determined in a symmetrical manner and describes mappings between each domain of data analyzed.

2.4. Multiplicity

As third important property of CCA is that this method can produce several modes (i.e., multiple pairs of canonical variates), each describing patterns of unique variation in the sets of variables. Each CCA mode carries a linear low-rank projection of the left variable set (one canonical variate associated with that mode) and a second linear low-rank projection of the right variable set (the other canonical variate associated with that mode). After extracting the first mode, which describes the largest variation in the observed data (cf. above), CCA determines the next pair of latent dimensions whose variation between both variable sets is not accounted for by the first mode. Since every new mode is found in the remaining variation in the observed data, the classical formulation of CCA optimizes the modes to be mutually uncorrelated with each other, a property known as *orthogonality*. The use of orthogonality to constraint CCA modes is analogous to what happens using PCA. Consequently, the different modes produced by CCA are ordered by the total variation explained in the domain-domain associations. To the extent that the unfolding modes are scientifically meaningful, interpretations can afford complex datasets to be considered as being made up of multiple overlapping descriptions underlying the observed data. For instance, much genetic variability in Europe can be jointly explained by orthogonal directions of variation along a north-south axis (i.e., one mode of variation) and a west-east axis (i.e., another mode of variation) (Moreno-Estrada et al., 2013). The ability for CCA to produce many pairs of canonical variates we refer to as 'multiplicity'.

Fig. 1 illustrates how the three core properties underlying CCA modeling and guidance of neuroscientific interpretation make it a particularly useful technique for the analysis of modern biomedical datasets – joint information compression, symmetry and multiplicity. First, CCA can provide a description that succinctly captures variation present across multiple variable sets. Second, CCA models are symmetrical in the sense that exchanging the two variable sets makes no difference to the results gained. Finally, we can estimate a collection of modes that describe the correspondence between two variable sets. As such, CCA modeling does not attempt to describe the "true" effects of any single variable (cf. below), instead targets the prominent correlation structure shared across dozens or potentially thousands of variables (Breiman and Friedman, 1997). Together these allow CCA to efficiently uncover symmetric linear relations that compactly summarize complex multivariate variable sets.

2.5. Examples of CCA in contemporary cognitive neuroscience

The suitability of CCA to big datasets available in modern neuroscience can be illustrated by considering examples of how it has been used to address specific questions that bear on the relationships between brain, cognition and disease. In the following section we consider 4 examples of how CCA can help describe the relationships between phenotypic measurements and neurobiological measurement such as functional brain activity.

Example 1. Smith et al. (2015) employed CCA to uncover brain-behavior modes of population co-variation in approximately 500 healthy participants from the Human Connectome Project (van Essen et al., 2013). These investigators aimed to discover whether specific patterns of whole-brain functional connectivity, on the one hand, are associated with specific sets of various demographics and behaviors on the other hand (see Fig. 2 for the analysis pipeline). Functional brain connectivity was estimated from resting state functional MRI scans measuring brain activity in the absence of a task or stimulus (Biswal et al., 1995). Independent component analysis (ICA; Beckmann et al., 2009) was used to extract 200 network nodes from fluctuations in neural activity. Next, functional connectivity matrices were calculated based on the pairwise correlation of the 200 nodes to yield a first variable set that quantified inter-individual variation in brain connectivity "fingerprints" (Finn et al., 2015). A rich set of phenotypic measures including descriptions of cognitive performance and demographic information provided a second variable set that captured inter-individual variation in behavior. The two variable arrays were submitted to CCA to gain insight into how latent dimensions of network coupling patterns present linear correspondences to latent dimensions underlying phenotypes of cognitive processing and life experience. The statistical robustness of the ensuing brain-behavior modes was determined via a non-parametric permutation approach in which the canonical correlation was the test statistic.

Smith and colleagues identified a single statistically significant CCA mode which included behavioral measures that varied along a positive-negative axis; measures of intelligence, memory, and cognition were located on the positive end of the mode, and measures of lifestyle (such as marijuana consumption) were located on the negative end of the mode. The brain regions exhibiting strongest contributions to coherent connectivity changes were reminiscent of the default mode network (Buckner et al., 2008). It is notable that prior work has provided evidence that regions composing the default mode network are associated with episodic and semantic memory, scene construction, and complex social

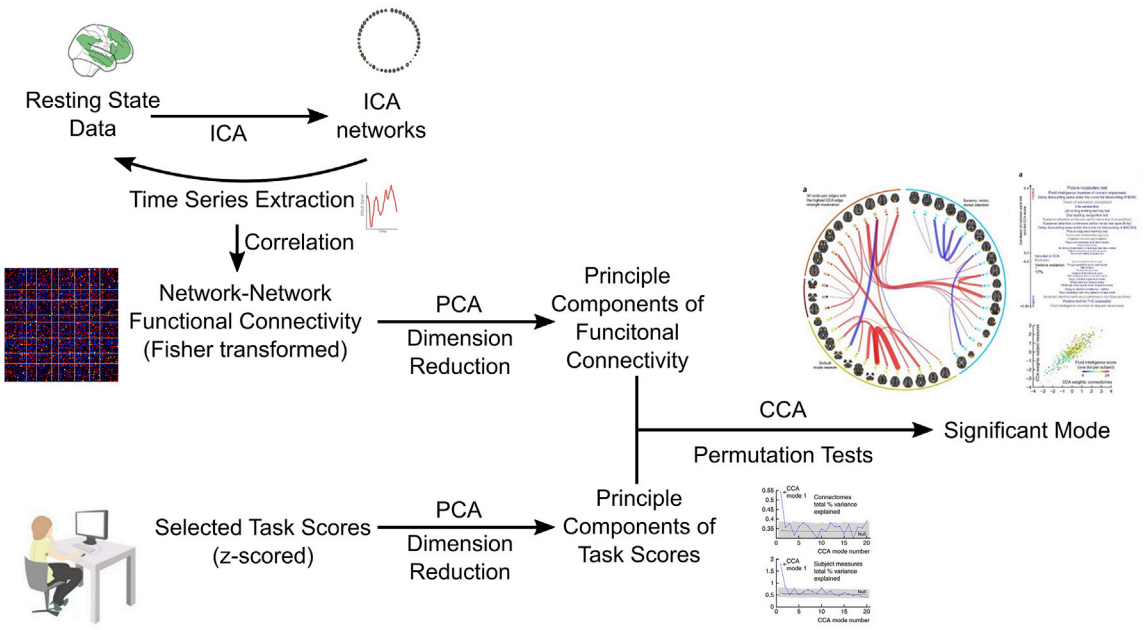

Fig. 2. The analysis pipeline of Smith et al. (2015).
These investigators aimed to discover whether specific patterns of whole-brain functional connectivity, on the one hand, are associated with specific sets of correlated demographics and behaviors on the other hand. The two domains of the input variables were transformed into principal components before the CCA model evaluation. The significant mode was determined by permutation tests. The finding of Smith et al. (2015)
provide evidence that functional connectivity in the default mode network is important for higher-level cognition and intelligent behaviors and are closely linked to positive life satisfaction.

reasoning such as theory of mind (Andrews-Hanna et al., 2010; Bzdok et al., 2012; Spreng et al., 2009). The finding of Smith and colleagues (Smith et al., 2015) provide evidence that functional connectivity in the default mode network is important for higher-level cognition and intelligent behaviors and that have important links to life satisfaction. This study illustrates the capacity of CCA for joint compression because it was able to successfully extract multivariate descriptions of datasets containing both brain measurements and a broad array of demographic and lifestyle indicators.

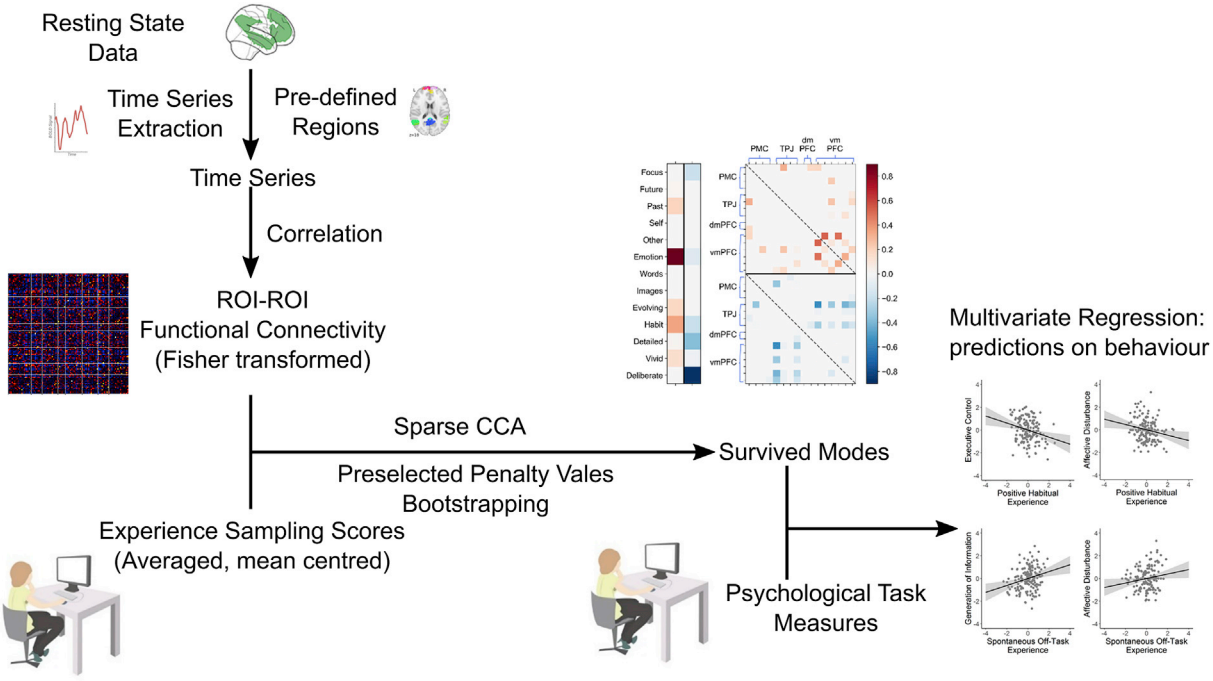

Fig. 3. The analysis pipeline of Wang et al. (2018b).Wang et al. (2018b)
used CCA to interrogate the hypothesis that various distinct aspects of ongoing thought can track distinct components of functional connectivity patterns within the default mode network. Sparse CCA was used to perform feature selection simultaneously with the model fitting on the brain-experience data. The identified CCA modes showed robust trait combinations of positive-habitual thoughts and spontaneous task-unrelated thoughts with linked patterns of connectivity fluctuations within the default mode network. The two modes were also related to distinct high-level cognitive profiles respectively.

Example 2. Another use of CCA has been to help understand the complex relationship between neural function and patterns of ongoing thought. In both the laboratory and in daily life, ongoing thought can often shift from the task at hand to other personally relevant characteristics - a phenomenon that is often referred to by the term 'mind-wandering' (Seli et al., 2018). Studies suggest there is a complex pattern of positive and negative associations between states of mind-wandering (Mooneyham and Schooler, 2013). This apparent complexity raises the possibility that mind-wandering is a heterogeneous rather than homogeneous state.

Wang and colleagues (2018b) used CCA to empirically explore this question by examining the links between connectivity within the default mode network and patterns of ongoing self-generated cognitive processes recorded in the laboratory (Fig. 3). Their analysis used patterns of functional connectivity within the default mode network as one set of observations, and patterns of self-reported descriptions recorded in the laboratory across multiple days as the second set of observations (Witten et al., 2009). The connectivity among 16 regions in the default mode network and 13 self-reported aspects on mind-wandering experience were fed into a sparse version of CCA (see Section 4.2 for further information on this variant of CCA). This analysis found two modes, one describing a pattern of positive-habitual thoughts, and a second that reflected spontaneous task-unrelated thoughts and both were associated with unique patterns of connectivity fluctuations within the default mode network. As a means to further validate the extracted brain-behavior modes in new data, follow-up analyses confirmed that the modes were uniquely related to aspects of cognition, such as executive control and the ability to generate information in a creative fashion, and the modes also independently distinguished well-being measures. These data suggest that the default mode network can contribute to ongoing thought in multiple ways, each with unique behavioral associations and underlying neural activity combinations. By demonstrating evidence for multiple brain-experience relationships within the default mode network, the authors (2018b) underline that greater specificity is needed when considering the links between brain activity and neural experience (see also Seli et al., 2018). This study illustrates the property of CCA for multiplicity because it was able to identify multiple different patterns of thought each of which could be validated based on their associations with other sets of observations.

Example 3. In the third example, Xia and colleagues (2018, see Fig. 4) mapped item-level psychiatric symptoms to brain connectivity patterns in brain networks using resting-state fMRI scans in a sample of roughly 1000 subjects from the Philadelphia Neurodevelopmental Cohort. Recognizing the high level of heterogeneity and comorbidity in existing diagnostic psychiatric diagnoses, these investigators were interested in how functional connectivity and individual symptoms can form linked dimensions of psychopathology and brain networks (Insel and Cuthbert, 2015). Notably, the study used a feature-selection step based on median absolute deviation to first reduce the dimensionality of the connectivity feature space prior to running CCA. As a result, about 3000 functional edges and 111 symptom items were analyzed in conjunction. As the number of features was still greater than the number of subjects, sparse CCA was used (Witten et al., 2009). This variant of the CCA family penalizes the number of features selected by the final CCA model. Based on covariation-explained and subsequent permutation testing (Mišić et al., 2016), the analysis identified four linked dimensions of psychopathology and functional brain connectivity – mood, psychosis, fear, and externalizing behavior. Through a resampling procedure that applied sparse CCA to different subsets of the data, the study identified stable clinical and connectional signatures that consistently contributed to each of the four modes. The resultant dimensions were relatively consistent with existing clinical diagnoses, but additionally cut across diagnostic boundaries to a significant degree. Furthermore, each of these dimensions were associated with a unique pattern of abnormal connectivity. However, a loss of network segregation was common to all dimensions, particularly

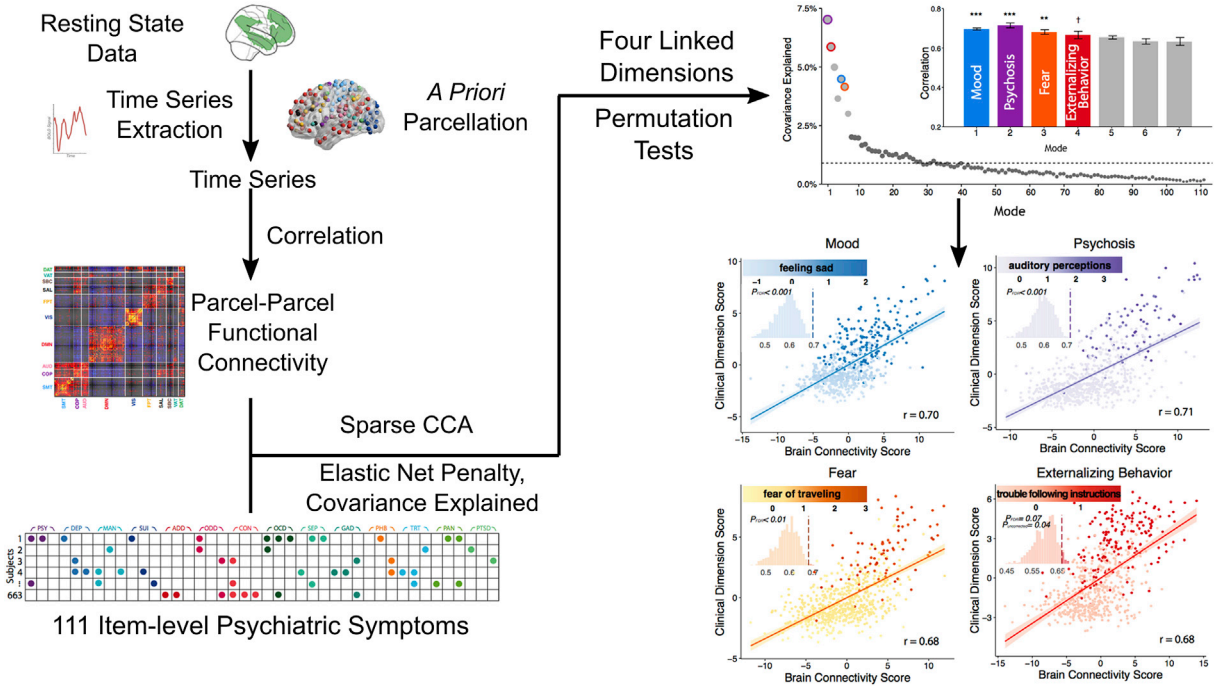

Fig. 4. The analysis pipeline of Xia et al. (2018). Xia and colleagues (2018) were interested in how functional connectivity and individual symptoms can form linked dimensions of psychopathology and brain networks. The study took a feature selection step based on median absolute deviation in preprocessing to first reduce the dimensionality of the functional connectivity measures. A sparse variation of CCA was applied to extract modes of linked dimensions of psychopathology and functional brain connectivity. Based on covariation-explained and subsequent non-parametric permutation statistical testing, the analysis identified four linked dimensions – mood, psychosis, fear, and externalizing behavior – each were associated with a unique pattern of abnormal brain connectivity. The results suggested that specific circuit-level abnormalities in the brain's functional network architecture may give rise to diverse psychiatric symptoms.

between executive networks and the default mode network. As network segregation is a normative feature of network development, loss of network segregation across all dimensions suggests that common neurodevelopmental abnormalities may be important for a wide range of psychiatric symptoms. Taking advantage of CCA's ability to capture common sources of variation in more than one datasets, these findings support the idea behind NIMH Research Domain Criteria that specific circuit-level abnormalities in the brain's functional network architecture may give rise to a diverse psychiatric symptoms (Cuthbert and Insel, 2013; Bzdok and Meyer-Lindenberg, 2018). This study illustrates the flexible use of CCA to reveal trans-diagnostic, continuous symptom dimensions based on whole-brain intrinsic connectivity fingerprints that can cut across existing disease boundaries in clinical neuroscience.

Example 4. The final example is a seminal study by Hu and colleagues (2018) who demonstrated the application of sparse multiple CCA (Witten and Tibshirani, 2009) to understand links between epigenetics and brain function. The multivariate nature of CCA makes it suitable as a common tool for examining data matrices with quite different features. For example, epigenetics can be characterized across mutiple biological levels including single nucleotide polymorphisms (SNPs), mRNA sequencing and DNA methylation of the primary tissue, or, organ level changes in the brain. In this context, sparse multiple CCA can help describe canonical correlations across three or more domains of variables, making it a pertinent tool for exploring genetics and imaging data. To address this issue, Hu and colleagues proposed an adaptively reweighted version of sparse multiple CCA (SMCCA, Witten et al., 2009). Conventional SMCCA can overlook smaller pairwise covariations and/or be over-influenced by data variables with large variation. To address this issue the authors adapted the algorithm variant to minimize the unfair combination through the adaptive introduction of weight coefficients. When applied to schizophrenia data, this multi-level analysis combined two genetic measures, genomic profiles from 9273 DNA methylation sites and genetic profiles from 777365 SNPs loci, for joint consideration with brain activity from resting state fMRI data. The functional neural data was parcellated across 116 anatomical regions using the AAL brain atlas (Tzourio-Mazoyer et al., 2002).

The authors selected model hyper-parameters using 5-fold cross validation, and during each step, the authors picked up one subgroup as testing sample and use the remaining 4 subgroups as training sample. A score quantifying fitting success was determined by the difference between the correlation of training sample and that of the test sample, which was used in this particular study to evaluate how successfully the sparsity parameters were selected. Next, a bootstrapping method was used to assess the most stable subset of variables (Meinshausen and Bühlmann, 2010) using a frequency cutoff based on Meinshausen and Bühlmann's work (2010). Using this novel methodology the authors found relationships between schizophrenia and both brain regions and genetic variants highlighted by past studies such as (i) hippocampus and fusiform in the fMRI data (Kircher and Thienel, 2005), (ii) SNPs related to brain development including BSX that has influence on methylation level (Park et al., 2007), PFTK1, which is relevant to brain degenerative diseases gene THR (Shibusawa et al., 2008), and AMIGO2 which is associated with hippocampus (Laeremans et al., 2013), and (iii) neuro tube development pathway in DNA methylation that is relevant to brain development (Kamburov et al., 2013). This study by Hu and colleagues highlights the utility of CCA in fusing data from many different domains in one coherent model to allow analysis of complex and heterogeneous features such as the relationship between epigenetics and brain architecture.

3. Interpretation and limitations of CCA

The modeling goal of CCA is to achieve a common decomposition of multiple matrices. This purpose makes the tool particularly useful for getting a handle on richly sampled descriptions of a population with observations that cross multiple levels of investigation. However, it remains a matter of ongoing debate whether this analysis technique corresponds more closely to a *descriptive* re-expression of the data (i.e., unsupervised modeling) or should be more readily understood as a form of *predictive* reduced-rank regression (i.e., supervised modeling, cf. Bach and Jordan, 2005; Bzdok et al., 2018; Breiman and Friedman, 1997; Witten et al., 2009). There are legitimate arguments in support of both views. A supervised algorithm depends on a designated modeling target to be predicted from an array of input variables, whereas an unsupervised algorithm aims to extract coherent patterns in observations without associated ground-truth labels that can be used during model estimation (Hastie et al., 2001). It is possible that as the dimensionality of one of the variable sets decreases to approach the single output of most linear-regression-type methods, in which case CCA may be more similar to a more supervised modeling approach. Conversely, with increasingly large variable sets on both sides, applying CCA is perhaps closer in spirit to an unsupervised modeling approach.

Whether the investigator considers CCA as either a supervised or unsupervised method has a consequence for both the interpretation of the results and their choice of eligible strategies to validate the model solutions (Bzdok and Ioannidis, 2019). For example, cross-validation is a technique that is commonly used for supervised model evaluation by comparing model-derived predictions in unseen data. In an unsupervised setting, however, there is typically no clear criterion for optimization (such as low residual sum of squares in supervised linear regression) that could be used for model selection or model evaluation, such as common in cross-validation schemes (Hastie et al., 2001). However, cross-validation is currently seldom used to buttress unsupervised model solutions, such as clustering methods like k-means or matrix decomposition techniques like PCA, because in these cases there is often no label upon which to evaluate performance (Bzdok, 2017; Hastie et al., 2001; Pereira et al., 2009). In situations when a CCA model describes the data without a known quantity to be predicted, cross-validation procedures can evaluate a CCA model by projecting data from new, previously unseen individuals using the canonical vectors observed from the initial sample. If this is not possible, an alternative validation strategy is to demonstrate whether the canonical variates of the obtained CCA solution are useful in capturing variation in other unseen measurements in the same set of individuals (e.g., Wang et al., 2018a). Yet another validation strategy for CCA is to show that the solutions it produces are robust when repeating the analysis on random subsets of the (already seen) individuals in so-called split-half analyses (Miller et al., 2016; Smith et al., 2015).

From a more formal perspective, the optimization objective governing parameter estimation during CCA fitting is unusual for a supervised model because this goal is based on Pearson's correlation metric. The majority of linear-regression-type predictive models have an optimization function that describes the degree of deviation from the ground-truth labels, including different residual-sum-of-squares loss functions (Casella and Berger, 2002; Hastie et al., 2001). Moreover, the symmetry of the variable sets in CCA is another reason why CCA may be considered an example of an unsupervised analysis tool. We are not aware of any existing supervised predictive model that would yield identical sets of model parameter fits after the independent and dependent variables have been swapped (if possible). To conclude, the CCA model is a relatively unique approach - "sitting between chairs" - that shares features of what are classical properties of supervised and unsupervised methods.

Another way to categorize statistical methods is based on their modeling goal: *estimation*, *prediction*, or *inference* (Bzdok and Ioannidis, 2019; Efron and Hastie, 2016; Hastie and Tibshirani, 1990). Model estimation refers to the process of adjusting randomly initialized parameters by fitting them to the data at hand; an intuitive example of these are beta parameters in classical linear regression. As model estimation can often be performed without applying the model to unseen observations or without assessing the fundamental trueness of the effects, some authors recently called this modeling regime "retrodiction" (McElreath, 2015; Pearl and Mackenzie, 2018). Prediction is concerned with

maximizing model fit in terms of optimizing its usefulness for predicting unseen data, such as data points acquired from future participants. Finally, drawing inferences on model fits has frequently been based on statistical null hypothesis testing and accompanying methodology (Wasserstein and Lazar, 2016). This form of drawing rigorous conclusions from data is especially useful in the classical analysis paradigm where the primary goal is to make precise statements about the contribution of single input variables.

In the context of this tripartite view of general modeling goals, CCA perhaps most naturally qualifies for the estimation category, rather than either primarily a predictive or inferential tool. Because of its exploratory nature, CCA can often be useful for applications focused on uncovering parsimonious principles in complex high-dimensional spaces as alternative descriptions of the observations at hand. Identifying the predictive value of individual variables in new data is not an integral part of the optimization objective underlying CCA. In fact, CCA applications often do not primarily seek to establish statistically significant links between *subsets* of the variables in each set. This is because the analytic goal is targeted at relevant patterns found across the entirety of both variable arrays. Even if p-values are obtained based on non-parametric null hypothesis testing in the context of CCA, the particular null hypothesis at play (commonly: the left and right variable matrix carry no corresponding information) is really centered on the *overall* robustness of the latent space correlations, as measured by the canonical correlations between the (projected) variable sets, and does not put a premium on specific single measurements; let alone on any particular link between one measurements from the left and one measurements from the right variable set. Thus, using CCA to pinpoint specific relations should only be done in a cautious manner. Stated in another way, CCA is not an optimal choice when the investigator wishes to make strong statements about the relevance and relationships of individual variables of the interrogated within a variable sets - a property shared with many other pattern-learning tools.

3.1. Limitations of CCA

Having considered the relationship between CCA and existing classifications of statistical techniques, we next consider some of the challenges that researchers may encounter when considering whether CCA is a good choice for a given data-analysis problem. We summarize the choices that a researcher is faced with in the form of a flowchart (see Fig. 5). As with many statistical approaches, the number of observations n in relation to the number of variables p is a key aspect when considering whether CCA is likely to be useful (Giraud, 2014; Hastie et al., 2015). Ordinary CCA can only be expected to yield useful model fits in data with more observations than the number of variables of the larger variable set (i.e., $n > max(p,q)$). Concretely, if the number of individuals included in the analysis is too close to the number of brain or behavior or genomic variables, then CCA will struggle to approximate any latent dimensions that do exist in the population (but see regularized CCA variants below). In these circumstances, even if CCA reaches a solution, without raising an error, the derived canonical vectors can be meaningless (Hastie et al., 2015). More formally, in such degenerate cases, CCA loses its usual ability to find unique identifiable solutions (despite being a non-convex optimization problem) that another laboratory with the same data and CCA implementation could also obtain (Jordan, 2018). Additionally, as an important note on reproducibility, with increasing number of variables in one or both sets, the ensuing canonical correlation often tends to increase due to the model's higher degrees of freedom (the number of model parameters is higher). An important consequence is that the canonical correlations obtained from CCA applications with differently sized variables sets cannot be directly used to decide which of the obtained CCA models are "better". The CCA solution is constrained by the participant sample as well as the number of variables in each set. As a cautionary note, the canonical correlation effect sizes obtained from the training data limit statements about how the obtained CCA solution at hand would perform on future data or other participants.

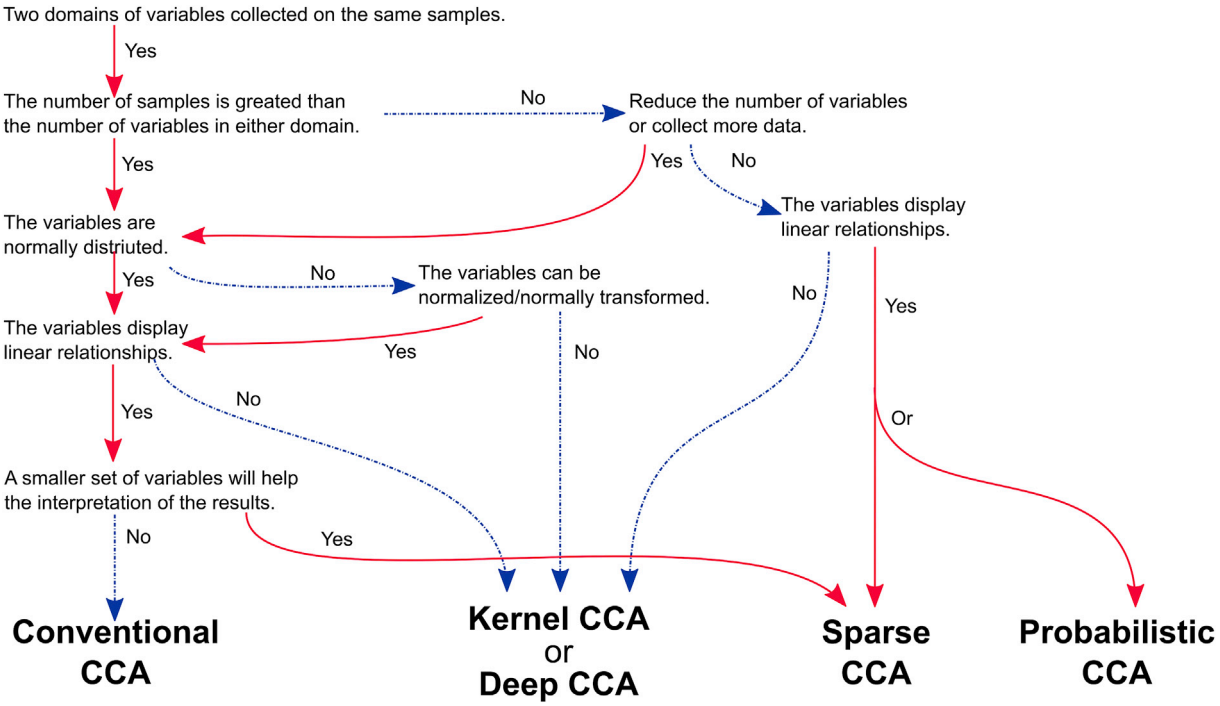

Fig. 5. A flowchart illustrating the choices when considering the application of CCA a dataset.

In a similar vein, smaller datasets offering measurements from only a few dozen individuals or observations may have difficulty in fully profiting from the strengths of multivariate procedure such as CCA. Moreover, the ground-truth effects in areas like psychology, neuroscience, and genetics are often small, which are hard to detect with insufficient sampling of the variability components. One practical remedy that can alleviate modeling challenges in small datasets is using data-reduction methods such as PCA or other methods for preprocessing each variable before applying CCA (e.g., Smith et al., 2015) or to adopt a sparse variant of CCA (see below). Reducing the variable sets according to their most important directions of linear variation can facilitate the CCA approach and the ensuing solution, including canonical variates, can be translated back to and interpreted within the original variable space (Bzdok et al., 2016). These considerations illustrate why CCA applications have long been less attractive in the context of many neuroscience studies, while its appeal and feasibility are now steadily growing as ever more rich, multi-modal, and open datasets become available (Davis et al., 2014).

A second limitation concerns the scope of the statistical relationships that CCA can discover and quantify in the underlying data. As a linear model, classical CCA imposes the assumption of additivity on the underlying relationships to unearth relevant linked co-variation patterns, thus ignoring more complicated variable-variable interactions that may exist in the data. CCA can accommodate any metric variable without strict dependence on normality. However, Gaussian normality in the data is desirable because CCA exactly operates on differences in averages and spreads that parameterize this data distribution. Before CCA is applied to the data, it is common practice that one evaluates the normality of the variable sets and possibly apply data an appropriate transformation, such as z-scoring (variable normalization by mean centering to zero and unit-spread scaling to one) or Box-Cox transformations (variable normalization involving logarithm and square-root operations). Finally, the relationships discovered by CCA solutions have been optimized to highlight those variables whose low-dimensional projection is most (linearly) coupled with the low-dimensional projection of the other variable set. As such, the derived canonical modes provide only one window into which multivariate relationships are most important *given the presence of the other variable set,* rather than identifying variable subsets that are important in the dataset per se.

This chart summarizes some of the practical choices faced by a researcher when considering whether to use CCA to analyze her data. Note that some of the choices in CCA workflows vary considerably depending on the interpretative goals and scientific context (i.e., conventional vs sparse CCA and sparse CCA vs probabilistic CCA).

3.2. Comparison to related methods and CCA extensions

CCA is probably the most general statistical approach to distill the relationships between two high-dimensional sources of quantitative measurements. In fact, CCA can be viewed as a broad class of methods that generalizes many more specialized approaches from the general linear model (GLM; Gelman and Hill, 2007). In fact, most of the linear models commonly used by behavioral scientists for parametric testing (including ANOVA, MANOVA, multiple regression, Pearson's correlation, and *t*-test) can be interpreted as special cases of CCA (Knapp, 1978; Thompson, 2015). Because these techniques are closely related, when evaluating CCA it will often be beneficial to more deeply understand the opportunities and challenges of similar approaches.

3.2.1. Related methods

i) **PCA** has certain similarities to CCA, although PCA performs unsupervised matrix decomposition of one variable set (Shlens, 2014a). A shared property of PCA and CCA is the orthogonality constraint imposed during structure discovery. As such, the set of uncovered sources of variation (i.e., modes) are assumed to be uncorrelated with each other in both methods. As an important difference, there are PCA formulations that minimize the reconstruction error between the original variable set and the back-projection of each observation from the latent dimensions of variation (Hastie et al., 2015). CCA instead directly optimizes the correspondence between the latent dimensions directly in the embedding space, rather than the reconstruction loss in the original variables incurred by the low-rank bottleneck. Moreover, PCA can be used for dimensionality reduction as a pre-processing step before CCA (e.g., Smith et al., 2015).

ii) Analogous to PCA and CCA, **independent component analysis (ICA)** also extracts hidden dimensions of variation in a potentially high-dimensional variable sets. While CCA is concerned with revealing multivariate sources of variation based on linear covariation structure, ICA can identify more complicated non-linear relationships in data that can capture statistical relationships that go beyond differences in averages and spreads (Shlens, 2014b). A second aspect that departs from CCA is the fact that latent dimensions obtained from ICA are not naturally ordered from highest to lowest contribution in reducing the reconstruction error, which needs to be computed in a later step. Another difference between CCA and ICA is how both approaches attempt to identify solutions featuring a form of uncorrelatedness. As described earlier, CCA's uses the constraint of orthogonality to obtain *uncorrelated* latent dimensions; in contrast, ICA optimizes the *independence* between the emerging hidden sources of variation. In this context independence between two variables implies their uncorrelatedness, but the lack of a *linear* correlation between the two variables does not ensure the lack of a *nonlinear* statistical relation between the two variables. Finally, it is worth mentioning that ICA can also be used as a post-processing step to further inspect effects in CCA solutions (Miller et al., 2016; Sui et al., 2010) (cf. below).

iii) **Partial least squares (PLS) regression** is more similar to CCA than to PCA or ICA. This is because PLS and CCA can identify latent dimensions of variation across *two* variable sets (McIntosh et al., 1996). A key distinctive feature of PLS is that the optimization goal is to minimize the *covariance* rather than the linear *correlation.*

While PLS is consistently viewed and used as a supervised method, it is controversial whether CCA should counted as part of the supervised or unsupervised family (see above) (Hastie et al., 2001). Further, many PLS and CCA implementations are similar in the sense that they impose an orthogonality constraint on the hidden sources of variation to be discovered. However, the two methods are also different in the optimization objective in the following sense: PLS maximizes the variance of the projected dimensions with the original variables of the designed response variables. Instead, CCA operates only in the embedding spaces of the left and right variable sets to maximize the correlation between the emerging low-rank projections, without correlation any of the original measurements directly. CCA thus indirectly identifies those canonical vectors whose ensuing canonical variates correlate most. In contrast to CCA, PLS is scale-variant (by reliance on the covariance), which leads to different results after transforming the variables.

As well as considering the alternative methods, there are also a number of important extensions to the CCA model, each of which are optimized with respect to specific analytic situations. These different model extensions are presented at the foot of Fig. 5.

3.2.2. Model extensions

i. **Probabilistic CCA** is a modification that motivates classical CCA as a generative model (Bach and Jordan, 2005; Klami et al., 2013). One advantage of this CCA variant is that it has a more principled

definition of the variation to be expected in the data and so has more opportunity to produce synthetic but plausible observations once the model has been fit. Additionally, because probabilistic CCA allows for the introduction of prior knowledge into the model specification, an advantageous aspect of many Bayesian models, this approach has been shown to yield more convincing results in small biomedical datasets which would otherwise be challenging to handle using ordinary CCA (e.g., Fujiwara et al., 2009; Huopaniemi et al., 2009).

ii. **Sparse CCA (SCCA** Witten et al., 2009) is a variant for identifying parsimonious sources of variation by encouraging exactly-zero contributions from many variables in each variable set. Besides facilitating interpretation of CCA solutions, the imposed l_1-norm penalty term is also effective in scaling CCA applications to higher-dimensional variable sets, where the number of variables can exceed the number of available observations (Hastie et al., 2015). One consequence of the introduction of the sparsity constraint is that this additionally introduced assumption can interfere with the orthogonality constraint of CCA. In neuroscience applications, the sparser the CCA modes that are generated, the more the canonical variates of the different modes can be correlated with one another. Additionally, it is important to note that the variation that each mode explains will not decrease in order from the first mode onwards as occurs in ordinary CCA. As a side note, other regularization schemes can also be an interesting extension to classical CCA. In particular, imposing an l_2-norm penalty term stabilizes CCA estimation in the wide-data setting using variable shrinkage, without the variable-selection property of the sparsity-inducing constraint (Witten and Tibshirani, 2009).

iii. **Multiset CCA** (Parra, 2018) or **multi-omics data fusion** (Hu et al., 2018) extent the analysis for more than two domains of data. In the field of neuroimaging, the application of multiset CCA is common blind source separation among subjects or among multiple imaging phenotypes (e.g., fMRI, structural MRI, and EEG). The advantage of multiset CCA is the flexibility in addressing variability in each domain of data without projecting data into a common space (c.f. ICA). The sparse variation of multiset CCA is also a popular choice to overcome the limitations when handling high number of variables. *Discriminative CCA*, or *Collaborative Regression* (Gross and Tibshirani, 2015; Luo et al., 2016), is a form of *multiset sparse CCA* (Hu et al., 2018; Witten and Tibshirani, 2009). In discriminative CCA, one data domain is a vector of labels. The labels help identify label/phenotype related cross-data associations in the other two domains, hence created a supervised version of CCA.

iv. **Kernel CCA** (KCCA; Hardoon et al., 2004) is an extension of CCA designed to capture more complicated nonlinear relationships. Kernels are mapping functions that implicitly express the variable sets in richer feature spaces, without ever having to explicitly compute the mapping, a method known as the 'kernel trick' (Hastie et al., 2001). KCCA first projects the data into this enriched virtual variable space before performing CCA in that enlarged input space. It is advantageous that KCCA allows for the detection of complicated non-linear relationships in the data. The drawback is that the interpretation of variable contributions in the original variable space is typically more challenging and in certain cases impossible. Further, KCCA is a nonparametric method; hence, the quality of the model fit scales poorly with the size of the training set.

v. **Deep CCA** (DCCA Andrew et al., 2013) is a variant of CCA that capitalizes on recent advances in "deep" neural-network algorithms (Jordan and Mitchell, 2015; LeCun et al., 2015). A core property of many modern neural network architectures is the capacity to learn representations in the data that emerge through multiple nested non-linear transformations. By analogy, DCCA simultaneously learns two deep neural network mappings of the two variable sets to maximize the correlation of their (potentially highly abstract) latent dimensions, which may remain opaque to human intuition.

4. Practical considerations

After these conceptual considerations, we next consider the actual implementation of CCA workflows. The computation of CCA solutions is possible by built-in libraries in MATLAB (canocorr), R (cancor or the PMA package), and the Python machine-learning library scikit-learn (sklearn.cross_decomposition.CCA). The sparse CCA mentioned in the examples is implemented in R package PMA. These code implementations provide comprehensive documentation for how to deploy CCA. For readers interested in reading more on detailed technical comparisons and discussions of CCA variants, please refer to the texts in Table 1.

4.1. Preprocessing

Some minimal data preprocessing is usually required as for most machine-learning methods. CCA is *scale-invariant* in that standardizing the data should not change the resulting canonical correlations. This property is inherited from Pearson's correlation defined by the degree of simultaneous unit change between two variables, with implicit standardization of the data variables. Nevertheless, z-scoring of each variable of the measurement sets is still recommended before performing CCA to facilitate the model numerical process of model estimation and to enhance domain interpretability. To avoid outliers skewing CCA estimation, it is recommended that one applies outlier detection and other common data-cleaning techniques (Gelman and Hill, 2007). Several readily applicable heuristics exist to identify unlikely variable values, such as replacing extreme values with 5th and 95th percentiles of the respective input dimension, a statistical transformation known as 'winsorizing'. Missing data is a common occurrence in large dataset. It is recommended to exclude observations with too many missing variables (e.g., those missing a whole domain of a questionnaire). Alternatively, missing variables can be "filled in" with mean or median when the proportion of missing data is small, or more sophisticated data-imputation techniques.

Besides unwarranted extreme and missing values, it is often necessary to account for potential nuisance influences on the variable sets. Deconfounding procedures are a preprocessing step in many neuroimaging data analysis settings to reduce the risk of finding non-meaningful modes of variation (such as motion) (cf. Bzdok et al., 2020). The same procedures that are commonly applied prior to the use of linear-regression analyses can also be useful in the context of CCA. Note that deconfounding is typically performed as an independent preceding step because the CCA model itself has no explicit noise component. Deconfounding is often carried by creating a regression model that captures the variation in the original data that can be explained by the confounder. The residuals of such regression modeling will be the new "cleaned" data with potential confound information removed. In neuroimaging, for example, head motion, age, sex, and total brain volume have frequently been considered unwanted sources of influence in many analysis contexts (Baum et al., 2018; Ciric et al., 2017; Kernbach et al., 2018; Miller et al., 2016; Smith et al., 2015). While some previous studies have submit one variable set to a nuisance-removal procedure, in the majority of the analysis scenarios the identical deconfounding step should probably be applied on each of the variable sets.

4.2. Data reduction

When the number of variables exceeds the number of participants, dimensionality-reduction techniques can provide useful data compression before performing CCA. The main techniques include features selection based on statistical dispersion, such as mean or median absolute deviation, and matrix factorizing methods, such as PCA and ICA. The

Table 1
Further reading on variations of CCA.

Fusion CCA	Calhoun and Sui (2016), Correa et al. (2010), Sui et al. (2014)
CCA application to signal processing	Cordes et al. (2012), Friman et al. (2001, 2002, 2003), Lottman et al. (2018), Yang et al. (2018), Zhuang et al. (2017)
Multiple CCA/Multi-omics data fusion	Correa et al. (2010), Hu et al. (2018)
CCA vs multivariate methods	Hair et al. (2010), Le Floch et al. (2012), Liu and Calhoun (2014), Parra, (2018), Pituch and Stevens (2015), Sui et al. (2012)
CCA vs PLS	Grellmann et al. (2015), Sun et al. (2009), Uurtio et al. (2017)

application of PCA to *pre*-process the various variables in each matrix to a smaller set of most explanatory dimension prior to performing CCA can allow this technique to be applied to smaller, computationally more feasible set of variables (besides a potentially beneficial denoising effect). To interpret the CCA solutions in the original data, some authors have related the canonical variates with the original data to recover the relevant variate relationships with the original variables as captured by each CCA mode. A potential limitation of performing the PCA first before CCA is that the assumptions implicit in the PCA application carry over into constraining the CCA solutions.

Another attractive analysis strategy is to *post*-process the CCA solution via ICA (Miller et al., 2016; Sui et al., 2010). Such an analysis tactic can overcome some issues of projecting the PCA-compressed data back into the original variable space. After CCA has been fitted, the ensuing canonical variates of both the left and right side can be concatenated across participants into one array (number of observations x 2 x number of modes). ICA is then applied to the aggregated canonical mode expressions to recover the *independent* sources of the variation *between observations expressed in the embedding space*. While incurring additional computational load, this approach can be advantageous because CCA can only disentangle latent directions of variation in the data up to a random rotation (Miller et al., 2016, p. 18). That is, orthogonal rotations between the obtained modes could have given an equivalently valid CCA solution (a weakness shared with PCA). The latent dimensions described by the obtained canonical vectors and canonical variate embeddings can be further disambiguated by the post-hoc ICA step (Sui et al., 2010, p. 20). Going beyond discovery of *uncorrelated* sources of variation, the ICA post-processing is especially useful in the detection of *independent* components that contribute to the common solution extracted from the two variable sets. The CCA + ICA hybrid approach could zoom in more on relationships between the two original variable sets in some cases. Yet, additional application of PCA preprocessing could influence the outcome of the CCA + ICA approach (Sui et al., 2010).

4.3. Model selection

CCA allows multiple modes to be calculated from the observed data leading to the obvious question of how to choose the optimal number of latent sources of variation to be extracted. While various strategies have been proposed, currently there is little consensus so far. The ambiguity regarding how to choose the number of CCA modes is closely related to issues of choosing the number of clusters in k-means and other clustering procedures as well as choosing the number of components in PCA, ICA, and other matrix decomposition techniques (Eickhoff et al., 2015).

To select a specific number of useful modes, several quality metrics can be used for quantifying the variation that can be explained with respect to a notion of the optimal sources of variation, without a clear default. Since the canonical variates represent the compressed (i.e., projected) information of the original data, the canonical modes should bear relation to the original data. Other alternatives include assessing the decrease in that reconstruction error metric with the canonical variates of one domain to predict or associate the original variables with increasing number of modes (Wang et al., 2018b). A drop in the overall data variation captured after adding yet another mode for modeling k+1 sources of variation indicates a candidate cut-off at k. An important overarching property in this context is that because the modes are constrained by their orthogonality, computing classical CCA with 5 or 50 modes produces the same first 5 canonical modes.

Another tactic relies on determining how many of the extracted modes from CCA are statistically robust as indicated by non-parametric permutation tests (e.g., Kernbach et al., 2018; Smith et al., 2015). An empirical distribution of canonical correlation of each mode can be computed under the null hypothesis that there is no coherent relation between the left and right variable set – in which case the canonical correlation should fluctuate around chance level. The permutation procedure proceeds by random shuffling of the rows or columns of the two variable sets to break any existing relationships between the ensuing low-rank projections of the two variable sets across observations (Efron, 2012; Nichols and Holmes, 2002). If the relation between the two variable sets is random, all derived modes should be meaningless. The first mode can be viewed as the strictest measure of null hypothesis, because it extracts the highest direction of variation explained in a null sample (e.g., Smith et al., 2015). Following many iterations of this process, the extracted perturbed mode from the permutation datasets serve to compute the chance level of associations between the two variable sets. Each canonical mode whose original canonical correlation exceeds the 95% level (significance at $p < 0.05$) or 99.9% level (significance at $p < 0.001$) can be certified as robust under the null hypothesis of absent linkage between the left and right variable set. If the investigator wishes to add an explicit correction for multiple comparisons, the p-value threshold can for instance be divided by the number of modes (i.e., Bonferroni's method) or false-discovery rate (FDR) can be used to reduce possible type I errors. This approach hence yields one p-value for each of the originally obtained CCA modes. Again, please note that no statistical null hypothesis testing is performed on any individual variable in this way, illustrating CCA's native inability to make targeted statements about specific isolated input variables.

Moreover, a hold-out framework has been proposed to determine the generalizability and statistical significance of discovered CCA modes in sufficiently large samples (Ferreira et al., 2018; Monteiro et al., 2016). This analysis scheme starts by randomly separating the data into a training set and a holdout set. A CCA model is then fitted based on the training set. The data from held-out individuals is then projected to the previously obtained CCA embedding (i.e., using the precomputed canonical vectors to obtain new canonical variate embeddings in the hold-out data) to generate independent hold-out correlations. Then, a permutation test is performed on the test data against the left-out correlations. This validation framework can be used to explicitly measure the pattern-generalization performance and obtain a p-value for the mode. A possible limitation lies in the need to have a reasonably sized hold-out set.

Finally, to explicitly evaluate the contribution of each individual input variable to the overall modeling solution, a sensitivity analysis has been performed for CCA (Kernbach et al., 2018). The impact of each variable was isolated by selectively removing all information from a given input variable, including for instance the functional connectivity strengths derived from that same brain region, and reiterating the CCA procedure based on the reduced data of one variable set and the original data from the other variable set. This analysis strategy issued a perturbed set of canonical variates under the assumption that, one-by-one, a particular input dimension may not have been important to obtain the original canonical modes. The degree of alteration in the canonical correlations was quantified by computing Pearson's correlation coefficient between the original and perturbed canonical variates. In addition to these point estimates after variable deletion, the induced statistical uncertainty was quantified by carrying out bootstrapping analysis. "Shaken

up" bootstrap datasets were generated from the original participant sample by randomly drawing individuals with replacement. In each of these alternative datasets, the perturbed CCA was fitted and evaluated in identical fashion. This robustness assessment provided population-level uncertainty intervals and hence enabled extrapolation of statements on variable importance to data we would observe in the future. High correlation between the original canonical variates and the canonical variates obtained without the contribution of a specific variable indicated that the variable in question was not vital to estimating the original CCA correspondence between the two data modalities. This is because removing the given variable (and any related information) incurred no dramatic change of the original CCA performance metrics. Instead, low correlations pointed towards variables that were of special relevance for deriving the co-variation between the two levels of observations. This generally applicable variable-deletion scheme can determine interpretable contributions of single input variables that play disproportionately important roles in highly multivariate analysis tools such as CCA.

5. Code availability

MATLAB, Python, and R code of several published CCA project can be found on several different repositories. Analysis by Smith and colleagues (2015) is available providing details of the CCA implementation in MATLAB and the permutation test (https://www.fmrib.ox.ac.uk/datasets/HCP-CCA/). The Python code of nested cross-validation scheme of Wang and colleague (2018a) is available on GitHub (https://github.com/htwangtw/patterns-of-thought) along with pre-modeling data scaling. Xia et al. (2018) made their R codes on grid search parameter selection and permutation test procedures on GitHub (https://github.com/cedricx/sCCA/tree/master/sCCA/code/final). We provide a reusable Conda environment file containing the key Python and R libraries for CCA and the model selection for users to explore the methods (https://gist.github.com/htwangtw/492ef08a07b0995049bc76a797dd18bf).

6. Conclusions

In contemporary biomedical research, complex multivariate relationships are expected among body, brain, cognition, and genes (Bzdok et al., 2019; Bzdok and Ioannidis, 2019). These together are likely to provide insight into the cause of diseases and other societal problems. CCA provides a simple, effective method for describing the correspondences between two variable sets that can be instrumental in describing complex relationships in neuroimaging. The appeal of CCA is likely to increase as the detail and quality of multi-modal datasets in neuroscience and other biomedical sciences increases. CCA has already started to be useful in two of the currently largest brain-imaging collections - the Human Connectome Project and UK Biobank. In many of these applications, CCA serves as the centerpiece of the analysis workflow. Given its versatility, CCA has the capacity to become a core building block of more elaborated data analysis pipelines (Calhoun and Sui, 2016; Correa et al., 2010b; Liu and Calhoun, 2014), instead of being the goal of the analysis itself (Smith et al., 2015). In this way, we hope the present primer will help encourage scientists to employ CCA, when appropriate, to quantify the multi-form and multi-faceted relationships that underscore many important phenomena of the human condition.

CRediT authorship contribution statement

Hao-Ting Wang: Conceptualization, Writing - original draft. **Jonathan Smallwood:** Writing - review & editing. **Janaina Mourao-Miranda:** Writing - review & editing. **Cedric Huchuan Xia:** Writing - review & editing. **Theodore D. Satterthwaite:** Writing - review & editing. **Danielle S. Bassett:** Writing - review & editing. **Danilo Bzdok:** Conceptualization, Writing - original draft.

Acknowledgment

JS and **H-TW** are funded by European Research Council Consolidator (WANDERINGMINDS – 646927). **DB** is funded by the Deutsche Forschungsgemeinschaft (DFG, BZ2/2-1, BZ2/3-1, and BZ2/4-1; International Research Training Group IRTG2150), Amazon AWS Research Grant, the German National Academic Foundation, and the START-Program of the Faculty of Medicine, RWTH Aachen. **JM-M** is supported by the Wellcome Trust (WT102845/Z/13/Z). **CHX** is supported by the Blavatnik Family Foundation and Medical Scientist Training Program (MSTP). **TDS** is supported by grants from National Institute of Mental Health (R01MH107703, R01MH112847, R21MH106799, R01MH113550). **DSB** acknowledges support from the John D. and Catherine T. MacArthur Foundation, the Alfred P. Sloan Foundation, the ISI Foundation, the Paul Allen Foundation, the Army Research Laboratory (W911NF-10-2-0022) , the Army Research Office (Bassett-W911NF-14-1-0679, Grafton-W911NF-16-1-0474, DCIST-W911NF-17-2-0181) , the Office of Naval Research (2-R01-DC-009209-11), the National Institute of Mental Health (R01 – MH112847, R01-MH107235, R21-M MH-106799), the National Institute of Child Health and Human Development (1R01HD086888-01), National Institute of Neurological Disorders and Stroke (R01 NS099348), and the National Science Foundation (BCS-1441502, BCS-1430087, NSF PHY-1554488 and BCS-1631550). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.116745.

References

Friman, O., Cedefamn, J., Lundberg, P., Borga, M., Knutsson, H., 2001. Detection of neural activity in functional MRI using canonical correlation analysis. Magn. Reson. Med. 45, 323–330. https://doi.org/10.1002/1522-2594(200102)45:2<323::AID-MRM1041>3.0.CO;2-#.

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., Gallacher, J., Green, J., Matthews, P.M., Pell, J., Sprosen, T., Collins, R., 2012. UK Biobank: current status and what it means for epidemiology. Health Pol. Technol. 1, 123–126. https://doi.org/10.1016/j.hlpt.2012.07.003.

Andrew, G., Arora, R., Bilmes, J., Livescu, K., 2013. Deep canonical correlation analysis. Proc. 30th Int. Conf. Mach. Learn. 28, 1247–1255.

Andrews-Hanna, J.R., Reidler, J.S., Sepulcre, J., Poulin, R., Buckner, R.L., 2010. Functional-anatomic fractionation of the brain's default network. Neuron 65, 550–562. https://doi.org/10.1016/j.neuron.2010.02.005.

Bach, F.R., Jordan, M.I., 2005. A probabilistic interpretation of canonical correlation analysis.

Barrick, M.R., Mount, M.K., 1991. The big five personality dimensions and job performance: a meta-analysis. Person. Psychol. 44.

Baum, G.L., Roalf, D.R., Cook, P.A., Ciric, R., Rosen, A.F.G., Xia, C., Elliott, M.A., Ruparel, K., Verma, R., Tunç, B., Gur, R.C., Gur, R.E., Bassett, D.S., Satterthwaite, T.D., 2018. The impact of in-scanner head motion on structural connectivity derived from diffusion MRI. Neuroimage 173, 275–286. https://doi.org/10.1016/j.neuroimage.2018.02.041.

Beckmann, C.F., Mackay, C.E., Filippini, N., Smith, S.M., 2009. Group comparison of resting-state FMRI data using multi-subject ICA and dual regression. Neuroimage 47, S148.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo???planar mri. Magn. Reson. Med. 34, 537–541. https://doi.org/10.1002/mrm.1910340409.

Breiman, L., Friedman, J.H., 1997. Predicting multivariate responses in multiple linear regression. J. R. Stat. Soc. Ser. B (Stat. Methodol. 59, 3–54.

Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L., 2008. The brain's default network: anatomy, function, and relevance to disease. Ann. N. Y. Acad. Sci. 1124, 1–38. https://doi.org/10.1196/annals.1440.011.

Bzdok, D., 2017. Classical statistics and statistical learning in imaging neuroscience. Front. Neurosci. 11, 543. https://doi.org/10.3389/fnins.2017.00543.

Bzdok, D., Varoquaux, G., Grisel, O., Eickenberg, M., Poupon, C., Thirion, B., 2016. Formal models of the network co-occurrence underlying mental operations. PLoS Comput. Biol. https://doi.org/10.1371/journal.pcbi.1004994.

Bzdok, D., Yeo, B.T.T., 2017. Inference in the age of big data: future perspectives on neuroscience. Neuroimage 155, 549–564. https://doi.org/10.1016/j.neuroimage.2017.04.061.

Bzdok, D., Altman, N., Krzywinski, M., 2018. Statistics versus machine learning. Nat. Methods 15, 233–234.
Bzdok, D., Floris, D.L., Marquand, A.F., 2020. Analyzing brain networks in population neuroscience: A case for the Bayesian philosophy. Philos. Trans. Roy. Soc. B: Biol. Sci. https://doi.org/10.1098/rstb.2019.0661.
Bzdok, D., Ioannidis, J.P.A., 2019. Exploration, inference and prediction in neuroscience and biomedicine. Trends Neurosci. 42 (4), 251–262.
Bzdok, D., Meyer-Lindenberg, A., 2018. Machine learning for precision psychiatry: Opportunites and challenges. Biol. Psychiatr.: Cognit. Neurosci. Neuroimaging 3 (3), 223–230.
Bzdok, D., Nichols, T.E., Smith, S.M., 2019. Towards algorithmic analytics for large-scale datasets. Nat. Mach. Intell. 1, 296–306.
Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A.R., Langner, R., Eickhoff, S.B., 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. Brain Struct. Funct. 217, 783–796. https://doi.org/10.1007/s00429-012-0380-y.
Calhoun, V.D., Sui, J., 2016. Multimodal fusion of brain imaging data: a key to finding the missing link(s) in complex mental illness. Biol. Psychiatr. Cogn. Neurosci. Neuroimag. https://doi.org/10.1016/j.bpsc.2015.12.005.
Casella, G., Berger, R.L., 2002. Statistical Inference. Duxbury Pacific Grove, CA.
Ciric, R., Wolf, D.H., Power, J.D., Roalf, D.R., Baum, G.L., Ruparel, K., Shinohara, R.T., Elliott, M.A., Eickhoff, S.B., Davatzikos, C., Gur, R.C., Gur, R.E., Bassett, D.S., Satterthwaite, T.D., 2017. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. Neuroimage 154, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020.
Cordes, D., Jin, M., Curran, T., Nandy, R., 2012. Optimizing the performance of local canonical correlation analysis in fMRI using spatial constraints. Hum. Brain Mapp. 33, 2611–2626. https://doi.org/10.1002/hbm.21388.
Correa, N.M., Adali, T., Li, Y., Calhoun, V.D., 2010a. Canonical correlation analysis for data fusion and group inferences. IEEE Signal Process. Mag. 27, 39–50. https://doi.org/10.1109/MSP.2010.936725.
Correa, N.M., Eichele, T., Adali, T., Li, Y.O., Calhoun, V.D., 2010b. Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI. Neuroimage. https://doi.org/10.1016/j.neuroimage.2010.01.062.
Cuthbert, B.N., Insel, T.R., 2013. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med. 11 https://doi.org/10.1186/1741-7015-11-126.
Davis, T., Xue, G., Love, B.C., Preston, A.R., Poldrack, R.A., 2014. Global neural pattern similarity as a common basis for categorization and recognition memory. J. Neurosci. 34, 7472–7484. https://doi.org/10.1523/JNEUROSCI.3376-13.2014.
Efron, B., 2010. The future of indirect evidence. Stat. Sci. 25, 145–157. https://doi.org/10.1214/09-STS308.
Efron, B., 2012. Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press.
Efron, B., Hastie, T., 2016. Computer-Age Statistical Inference. Cambridge University Press.
Eickhoff, S.B., Thirion, B., Varoquaux, G., Bzdok, D., 2015. Connectivity-based parcellation: critique and implications. Hum. Brain Mapp. 36, 4771–4792. https://doi.org/10.1002/hbm.22933.
Ferreira, F.S., Rosa, M.J., Moutoussis, M., Dolan, R., Shawe-Taylor, J., Ashburner, J., Mourao-Miranda, J., 2018. Sparse PLS hyper-parameters optimisation for investigating brain-behaviour relationships. In: 2018 International Workshop on Pattern Recognition in Neuroimaging, PRNI 2018. IEEE, pp. 1–4. https://doi.org/10.1109/PRNI.2018.8423947.
Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., Todd Constable, R., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18, 1–11. https://doi.org/10.1038/nn.4135.
Friman, O., Borga, M., Lundberg, P., Knutsson, H., 2002. Detection of neural activity in fMRI using maximum correlation modeling. Neuroimage 15, 386–395. https://doi.org/10.1006/nimg.2001.0972.
Friman, O., Borga, M., Lundberg, P., Knutsson, H., 2003. Adaptive analysis of fMRI data. Neuroimage 19, 837–845. https://doi.org/10.1016/S1053-8119(03)00077-6.
Friman, O., Borga, M., Lundberg, P., Knutsson, H., 2004. Detection and detrending in fMRI data analysis. Neuroimage 22, 645–655. https://doi.org/10.1016/j.neuroimage.2004.01.033.
Friston, K.J., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2008. Bayesian decoding of brain images. Neuroimage 39, 181–205. https://doi.org/10.1016/j.neuroimage.2007.08.013.
Fujiwara, Y., Miyawaki, Y., Kamitani, Y., 2009. Estimating image bases for visual image reconstruction from human brain activity. In: Advances in Neural Information Processing Systems, pp. 576–584.
Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multi-Level Hierarchical Models. Cambridge University Press, New York, NY, USA.
Giraud, C., 2014. Introduction to High-Dimensional Statistics. CRC Press.
Grellmann, C., Bitzer, S., Neumann, J., Westlye, L.T., Andreassen, O.A., Villringer, A., Horstmann, A., 2015. Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of MRI and genetic data. Neuroimage 107, 289–310. https://doi.org/10.1016/j.neuroimage.2014.12.025.
Gross, S.M., Tibshirani, R., 2015. Collaborative regression. Biostatistics 16, 326–338. https://doi.org/10.1093/biostatistics/kxu047.
Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., 2010. Multivariate Data Analysis. Vectors. https://doi.org/10.1016/j.ijpharm.2011.02.019.
Hardoon, D.R., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: an overview with application to learning methods. Neural Comput. 16, 2639–2664. https://doi.org/10.1162/0899766042321814.
Hardoon, D.R., Mourão-Miranda, J., Brammer, M., Shawe-Taylor, J., 2007. Unsupervised analysis of fMRI data using kernel canonical correlation. Neuroimage. https://doi.org/10.1016/j.neuroimage.2007.06.017.
Hastie, T., Tibshirani, R.J., 1990. Generalized additive models. In: Monographs on Statistics and Applied Probability. Chapman & Hall, London.
Hastie, T., Tibshirani, R.J., Friedman, J.H., 2001. The Elements of Statistical Learning. Springer Series in Statistics, Heidelberg, Germany.
Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: the Lasso and Generalizations. CRC Press. https://doi.org/10.1201/b18401-1.
Healy, M.J.R., 1957. A rotation method for computing canonical correlations. Math. Comput. 11 https://doi.org/10.1090/s0025-5718-1957-0085600-6, 83–83.
Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321. https://doi.org/10.2307/2333955.
Hu, W., Lin, D., Cao, S., Liu, J., Chen, J., Calhoun, V.D., Wang, Y.P., 2018. Adaptive sparse multiple canonical correlation analysis with application to imaging (Epi)Genomics study of schizophrenia. IEEE Trans. Biomed. Eng. 65, 390–399. https://doi.org/10.1109/TBME.2017.2771483.
Hu, W., Zhang, A., Cai, B., Calhoun, V., 2019. Distance canonical correlation analysis with application to an imaging-genetic study. J. Med. Imaging 6, 1. https://doi.org/10.1117/1.jmi.6.2.026501.
Huopaniemi, I., Suvitaival, T., Nikkilä, J., Orešič, M., Kaski, S., 2009. Two-way analysis of high-dimensional collinear data. Data Min. Knowl. Discov. 19, 261–276.
Insel, T.R., Cuthbert, B.N., 2015. Brain disorders? Precisely. Science (80-.) 348, 499–500. https://doi.org/10.1126/science.aab2358.
Jordan, M.I., 2018. On gradien-based optimization: accelerated, nonconvex and stochastic. Talk.
Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science (80-.) 349, 255–260.
Kamburov, A., Stelzl, U., Lehrach, H., Herwig, R., 2013. The ConsensusPathDB interaction database: 2013 Update. Nucleic Acids Res. https://doi.org/10.1093/nar/gks1055.
Kernbach, J.M., Yeo, B.T.T., Smallwood, J., Margulies, D.S., Thiebaut de Schotten, M., Walter, H., Sabuncu, M.R., Holmes, A.J., Gramfort, A., Varoquaux, G., Thirion, B., Bzdok, D., 2018. Subspecialization within default mode nodes characterized in 10,000 UK Biobank participants. Proc. Natl. Acad. Sci. U. S. A.
Kircher, T.T.J., Thienel, R., 2005. Functional brain imaging of symptoms and cognition in schizophrenia. Prog. Brain Res. https://doi.org/10.1016/S0079-6123(05)50022-0.
Klami, A., Virtanen, S., Kaski, S., 2013. Bayesian canonical correlation analysis. J. Mach. Learn. Res. 14, 965–1003.
Knapp, T.R., 1978. Canonical correlation analysis: a general parametric significance-testing system. Psychol. Bull. 85, 410–416. https://doi.org/10.1037/0033-2909.85.2.410.
Laeremans, A., Nys, J., Luyten, W., D'Hooge, R., Paulussen, M., Arckens, L., 2013. AMIGO2 mRNA expression in hippocampal CA2 and CA3a. Brain Struct. Funct. https://doi.org/10.1007/s00429-012-0387-4.
Le Floch, É., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., Dehaene, S., Thirion, B., Poline, J.B., Duchesnay, É., 2012. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. Neuroimage. https://doi.org/10.1016/j.neuroimage.2012.06.061.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.
Liu, J., Calhoun, V.D., 2014. A review of multivariate analyses in imaging genetics. Front. Neuroinf. https://doi.org/10.3389/fninf.2014.00029.
Lottman, K.K., White, D.M., Kraguljac, N.V., Reid, M.A., Calhoun, V.D., Catao, F., Lahti, A.C., 2018. Four-way multimodal fusion of 7 T imaging data using an mCCA+jICA model in first-episode schizophrenia. Hum. Brain Mapp. 39, 1475–1488. https://doi.org/10.1002/hbm.23906.
Luo, C., Liu, J., Dey, D.K., Chen, K., 2016. Canonical variate regression. Biostatistics 17, 468–483. https://doi.org/10.1093/biostatistics/kxw001.
Marquand, A.F., Haak, K.V., Beckmann, C.F., 2017. Functional corticostriatal connection topographies predict goal-directed behaviour in humans. Nat. Hum. Behav. 1, 0146 https://doi.org/10.1038/s41562-017-0146.
McElreath, R., 2015. Statistical Rethinking.
McIntosh, A.R., Bookstein, F.L., Haxby, J.V., Grady, C.L., 1996. Spatial pattern analysis of functional brain images using partial least squares. Neuroimage 3, 143–157.
Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc. Ser. B Stat. Methodol. https://doi.org/10.1111/j.1467-9868.2010.00740.x.
Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19, 1523–1536. https://doi.org/10.1038/nn.4393.
Mišić, B., Betzel, R.F., de Reus, M.A., van den Heuvel, M.P., Berman, M.G., McIntosh, A.R., Sporns, O., 2016. Network-level structure-function relationships in human neocortex. Cerebr. Cortex 26, 3285–3296. https://doi.org/10.1093/cercor/bhw089.
Monteiro, J.M., Rao, A., Shawe-Taylor, J., Mourão-Miranda, J., 2016. A multiple hold-out framework for sparse partial least squares. J. Neurosci. Methods 271, 182–194. https://doi.org/10.1016/j.jneumeth.2016.06.011.
Mooneyham, B.W., Schooler, J.W., 2013. The costs and benefits of mind-wandering: a review. Can. J. Exp. Psychol. 67, 11–18. https://doi.org/10.1037/a0031569.
Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., Eng, C., Sandoval, K., Acevedo-Acevedo, S., Norman, P.J., Layrisse, Z., Parham, P., Martínez-Cruzado, J.C., Burchard, E.G., Cuccaro, M.L., Martin, E.R., Bustamante, C.D., 2013. Reconstructing

the population genetic history of the caribbean. PLoS Genet. 9 https://doi.org/10.1371/journal.pgen.1003925.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.
Nooner, K.B., Colcombe, S.J., Tobe, R.H., Mennes, M., Benedek, M., Moreno, A.L., Panek, L.J., Brown, S., Zavitz Stephen, T.T., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R.T., Anwar, S.M.K., Hinz, C.M., Kaplan, M.S., Rachlin, A.B., Adelsberg, S., Cheung, B., Khanuja, R., Yan, C.-G., Craddock, R.C., Courtney, W., King, M., Wood, D., Cox, C.L., Kelly, A.M.C., Petkova, E., Reiss, P.T., Duan, N., Thomsen, D., Biswal, B.B., Coffey, B., Hoptman, M.J., Javitt, D.C., Pomara, N., Sidtis, J.J., Koplewicz, H.S., Castellanos, F.X., Leventhal, B.L., Milham, M.P., 2012. The NKI-Rockland sample: a model for accelerating the pace of discovery science in psychiatry. Front. Neurosci. 6, 152. https://doi.org/10.3389/fnins.2012.00152.
Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science (80-.) 349, aac4716.
Park, S.Y., Kim, J.B., Han, Y.-M., 2007. REST is a key regulator in brain-specific homeobox gene expression during neuronal differentiation. J. Neurochem. https://doi.org/10.1111/j.1471-4159.2007.04947.x.
Parra, L.C., 2018. Multi-set Canonical Correlation Analysis Simply Explained arXiv: 1802.03759 [stat].
Pearl, J., Mackenzie, D., 2018. The Book of Why: the New Science of Cause and Effect. Basic Books.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, 199–209. https://doi.org/10.1016/j.neuroimage.2008.11.007.
Pituch, K.A., Stevens, J.P., 2015. Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM's SPSS. Routledge. https://doi.org/10.1017/CBO9781107415324.004.
Poldrack, R.A., Gorgolewski, K.J., 2014. Making big data open: data sharing in neuroimaging. Nat. Neurosci. 17, 1510–1517. https://doi.org/10.1038/nn.3818.
Seli, P., Kane, M.J., Smallwood, J., Schacter, D.L., Maillet, D., Schooler, J.W., Smilek, D., 2018. Mind-wandering as a natural kind: a family-resemblances view. Trends Cognit. Sci. 22, 479–490. https://doi.org/10.1016/j.tics.2018.03.010.
Shafto, M.A., Tyler, L.K., Dixon, M., Taylor, J.R., Rowe, J.B., Cusack, R., Calder, A.J., Marslen-Wilson, W.D., Duncan, J., Dalgleish, T., Henson, R.N.A., Brayne, C., Bullmore, E.T., Campbell, K., Cheung, T., Davis, S.W., Geerligs, L., Kievit, R.A., McCarrey, A., Price, D., Samu, D., Treder, M., Tsvetanov, K.A., Williams, N., Bates, L., Emery, T., Erzinçlioglu, S., Gadie, A., Gerbase, S., Georgieva, S., Hanley, C., Parkin, B., Troy, D., Allen, J., Amery, G., Amunts, L., Barcroft, A., Castle, A., Dias, C., Dowrick, J., Fair, M., Fisher, H., Goulding, A., Grewal, A., Hale, G., Hilton, A., Johnson, F., Johnston, P., Kavanagh-Williamson, T., Kwasniewska, M., McMinn, A., Norman, K., Penrose, J., Roby, F., Rowland, D., Sargeant, J., Squire, M., Stevens, B., Stoddart, A., Stone, C., Thompson, T., Yazlik, O., Barnes, D., Hillman, J., Mitchell, J., Villis, L., Matthews, F.E., 2014. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. BMC Neurol. 14, 204. https://doi.org/10.1186/s12883-014-0204-1.
Shibusawa, N., Hashimoto, K., Yamada, M., 2008. Thyrotropin-releasing hormone (TRH) in the cerebellum. Cerebellum. https://doi.org/10.1007/s12311-008-0033-0.
Shlens, J., 2014a. A Tutorial on Principal Component Analysis arXiv Prepr. arXiv1404.1100.
Shlens, J., 2014b. A Tutorial on Independent Component Analysis arXiv Prepr. arXiv1404.2986.
Smith, S.M., Nichols, T.E., 2018. Statistical challenges in "big data" human neuroimaging. Neuron 97, 263–268. https://doi.org/10.1016/j.neuron.2017.12.018.
Smith, S.M., Nichols, T.E., Vidaurre, D., Winkler, A.M., Behrens, T.E.J., Glasser, M.F., Ugurbil, K., Barch, D.M., Van Essen, D.C., Miller, K.L., 2015. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nat. Neurosci. 18, 1565–1567. https://doi.org/10.1038/nn.4125.
Spreng, R.N., Mar, R.A., Kim, A.S., 2009. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. J. Cognit. Neurosci. 21, 489–510. https://doi.org/10.1162/jocn.2008.21029.
Sui, J., Adali, T., Pearlson, G., Yang, H., Sponheim, S.R., White, T., Calhoun, V.D., 2010. A CCA+ICA based model for multi-task brain imaging data fusion and its application to schizophrenia. Neuroimage 51, 123–134. https://doi.org/10.1016/j.neuroimage.2010.01.069.
Sui, J., Adali, T., Yu, Q., Chen, J., Calhoun, V.D., 2012. A review of multivariate methods for multimodal fusion of brain imaging data. J. Neurosci. Methods. https://doi.org/10.1016/j.jneumeth.2011.10.031.
Sui, J., Castro, E., He, H., Bridwell, D., Du, Y., Pearlson, G.D., Jiang, T., Calhoun, V.D., 2014. Combination of FMRI-SMRI-EEG data improves discrimination of schizophrenia patients by ensemble feature selection. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC 2014, pp. 3889–3892. https://doi.org/10.1109/EMBC.2014.6944473.
Sun, L., Ji, S., Yu, S., Ye, J., 2009. On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In: Proceedings of the 21st International Jont Conference on Artifical Intelligence, IJCAI'09. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1230–1235.
Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam-Can, Henson, R.N.A., 2017. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. Neuroimage 144, 262–269. https://doi.org/10.1016/j.neuroimage.2015.09.018.
Thompson, B., 2015. The case for using the general linear model as a unifying conceptual framework for teaching statistics and psychometric theory. J. Methods Meas. Soc. Sci. 6, 30–41. https://doi.org/10.2458/azu_jmmss_v6i1_blair.
Tsvetanov, K.A., Henson, R.N.A., Tyler, L.K., Razi, A., Geerligs, L., Ham, T.E., Rowe, J.B., Cam-Can, 2016. Extrinsic and intrinsic brain network connectivity maintains cognition across the lifespan despite accelerated decay of regional brain activation. J. Neurosci. 36, 3115–3126. https://doi.org/10.1523/JNEUROSCI.2733-15.2016.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. https://doi.org/10.1006/nimg.2001.0978.
Uurtio, V., Monteiro, J.M., Kandola, J., Shawe-Taylor, J., Fernandez-Reyes, D., Rousu, J., 2017. A tutorial on canonical correlation methods. ACM Comput. Surv. 50, 1–33. https://doi.org/10.1145/3136624.
van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The Wu-minn human connectome project: an overview. Neuroimage 80, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041.
Vatansever, D., Bzdok, D., Wang, H.-T., Mollo, G., Sormaz, M., Murphy, C.E., Karapanagiotidis, T., Smallwood, J., Jefferies, E., 2017. Varieties of semantic cognition revealed through simultaneous decomposition of intrinsic brain connectivity and behaviour. Neuroimage 158, 1–11. https://doi.org/10.1016/j.neuroimage.2017.06.067.
Wang, H.-T., Bzdok, D., Margulies, D.S., Craddock, R.C., Milham, M.P., Jefferies, E., Smallwood, J., 2018a. Patterns of thought: population variation in the associations between large-scale network organisation and self-reported experiences at rest. Neuroimage 176, 518–527. https://doi.org/10.1016/j.neuroimage.2018.04.064.
Wang, H.-T., Poerio, G.L., Murphy, C.E., Bzdok, D., Jefferies, E., Smallwood, J., 2018b. Dimensions of experience: exploring the ontology of the wandering mind. Psychol. Sci. 29, 56–71. https://doi.org/10.1177/0956797617728727.
Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on p-values: context, process, and purpose. Am. Statistician 70, 129–133.
Witten, D.M., Tibshirani, R.J., 2009. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat. Appl. Genet. Mol. Biol. 8, 29. https://doi.org/10.2202/1544-6115.1470.
Witten, D.M., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534. https://doi.org/10.1093/biostatistics/kxp008.
Xia, C.H., Ma, Z., Ciric, R., Gu, S., Betzel, R.F., Kaczkurkin, A.N., Calkins, M.E., Cook, P.A., García de la Garza, A., Vandekar, S.N., Cui, Z., Moore, T.M., Roalf, D.R., Ruparel, K., Wolf, D.H., Davatzikos, C., Gur, R.C., Gur, R.E., Shinohara, R.T., Bassett, D.S., Satterthwaite, T.D., 2018. Linked dimensions of psychopathology and connectivity in functional brain networks. Nat. Commun. 9, 3003. https://doi.org/10.1038/s41467-018-05317-y.
Yang, Z., Zhuang, X., Sreenivasan, K., Mishra, V., Curran, T., Byrd, R., Nandy, R., Cordes, D., 2018. 3D spatially-adaptive canonical correlation analysis: local and global methods. Neuroimage 169, 240–255. https://doi.org/10.1016/j.neuroimage.2017.12.025.
Yang, Z., Zhuang, X., Bird, C., Sreenivasan, K., Mishra, V., Banks, S., Cordes, D., 2019. Performing sparse regularization and dimension reduction simultaneously in multimodal data fusion. Front. Neurosci. https://doi.org/10.3389/fnins.2019.00642.
Zhuang, X., Yang, Z., Curran, T., Byrd, R., Nandy, R., Cordes, D., 2017. A family of locally constrained CCA models for detecting activation patterns in fMRI. Neuroimage 149, 63–84. https://doi.org/10.1016/j.neuroimage.2016.12.081.
Zhuang, X., Yang, Z., Sreenivasan, K.R., Mishra, V.R., Curran, T., Nandy, R., Cordes, D., 2019. Multivariate group-level analysis for task fMRI data with canonical correlation analysis. Neuroimage. https://doi.org/10.1016/j.neuroimage.2019.03.030.