NeuroImage 249 (2022) 118854

Contents lists available at ScienceDirect

NeuroImage

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

A hierarchical Bayesian model to find brain-behaviour associations in incomplete data sets

Fabio S. Ferreira[a,b,*], Agoston Mihalik[a,b], Rick A. Adams[a,b,c], John Ashburner[c], Janaina Mourao-Miranda[a,b]

[a] Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK
[b] Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, UK
[c] Wellcome Centre for Human Neuroimaging, University College London, London, UK

ARTICLE INFO

Keywords:
Multivariate methods
Group factor analysis
Bayesian inference
Missing data
Brain connectivity
Behaviour

ABSTRACT

Canonical Correlation Analysis (CCA) and its regularised versions have been widely used in the neuroimaging community to uncover multivariate associations between two data modalities (e.g., brain imaging and behaviour). However, these methods have inherent limitations: (1) statistical inferences about the associations are often not robust; (2) the associations within each data modality are not modelled; (3) missing values need to be imputed or removed. Group Factor Analysis (GFA) is a hierarchical model that addresses the first two limitations by providing Bayesian inference and modelling modality-specific associations. Here, we propose an extension of GFA that handles missing data, and highlight that GFA can be used as a predictive model. We applied GFA to synthetic and real data consisting of brain connectivity and non-imaging measures from the Human Connectome Project (HCP). In synthetic data, GFA uncovered the underlying shared and specific factors and predicted correctly the non-observed data modalities in complete and incomplete data sets. In the HCP data, we identified four relevant shared factors, capturing associations between mood, alcohol and drug use, cognition, demographics and psychopathological measures and the default mode, frontoparietal control, dorsal and ventral networks and insula, as well as two factors describing associations within brain connectivity. In addition, GFA predicted a set of non-imaging measures from brain connectivity. These findings were consistent in complete and incomplete data sets, and replicated previous findings in the literature. GFA is a promising tool that can be used to uncover associations between and within multiple data modalities in benchmark datasets (such as, HCP), and easily extended to more complex models to solve more challenging tasks.

1. Introduction

In the past few years, there has been a substantial increase in the application of multivariate methods, such as Canonical Correlation Analysis (CCA) (Hotelling, 1936), to identify associations between two data modalities (e.g., brain imaging and behaviour). CCA uncovers underlying associations between two sets of variables by finding linear combinations of variables from each modality that maximise the correlation between them. This is particularly relevant in brain imaging research, where different types of data (e.g., brain structural/functional data, behavioural and cognitive assessments) are collected from the same individuals to investigate the population variability. Moreover, the unsupervised nature of CCA has made it increasingly popular in fields such as psychiatric neuroscience, where there is a lack of objective biomarkers of illness and the diagnostic categories are not reliable (Bzdok and Meyer-Lindenberg, 2017; Insel et al., 2010).

CCA and regularised variants of CCA, such as sparse CCA (Lê Cao et al., 2009; Waaijenborg et al., 2008; Witten et al., 2009), have been used to identify associations, for instance, between brain connectivity and cognitive/psychopathology measures (Drysdale et al., 2017; Mihalik et al., 2019; Xia et al., 2018), brain connectivity and general lifestyle, demographic and behavioural measures (Alnæs et al., 2020; Bijsterbosch et al., 2018; Lee et al., 2019; Li et al., 2019; Smith et al., 2015), brain structure, demographic and behavioural measures (Mihalik et al., 2020; Monteiro et al., 2016) and between different brain imaging modalities (Sui et al., 2012).

Nonetheless, these methods have some limitations. First, they do not provide an inherent robust inference approach to infer the relevant associations. This is usually done by assessing the statistical significance of the associations using permutation inference (Winkler et al., 2020) or hold-out sets (Mihalik et al., 2020; Monteiro et al., 2016). Second, the associations within data modalities, which might explain important vari-

* Corresponding author at Centre for Medical Imaging Computing (CMIC), 90 High Holborn, Holborn, London WC1V 6LJ.
E-mail address: fabio.ferreira.16@ucl.ac.uk (F.S. Ferreira).

https://doi.org/10.1016/j.neuroimage.2021.118854.
Received 8 March 2021; Received in revised form 30 November 2021; Accepted 22 December 2021
Available online 29 December 2021.

ance in the data, are not modelled. Finally, CCA assumes data pairing between data modalities, which is problematic when values are missing in one or both data modalities. This is a common issue in clinical and neuroimaging datasets, in which the missing values usually need to be imputed or removed before applying the models.

One potential way to address the limitations mentioned above is to solve the CCA problem within a probabilistic framework. Bach and Jordan (2006) proposed a probabilistic interpretation of CCA, but showed that the maximum likelihood estimates are equivalent to the solution that standard CCA finds. Nevertheless, probabilistic CCA provided an initial step towards robust inference by allowing estimation of the uncertainty of the parameters and it could be used as building block for more complex models, such as Bayesian CCA proposed by Klami and Kaski (2007) and Chong Wang (2007). In both papers, the authors introduced a hierarchical Bayesian extension of CCA by adding suitable prior distributions over the model parameters to automatically infer the number of relevant latent components (i.e., relevant associations) using Bayesian inference.

Bayesian CCA has some limitations, however: it is not able to uncover associations within data modalities and, in high dimensional data sets, it can be computationally infeasible (Klami et al., 2013). Virtanen and colleagues (Klami et al., 2013; Virtanen et al., 2011) proposed an extension of Bayesian CCA to solve these two limitations, whilst removing irrelevant latent components (i.e., components that explain little variance). This model was further extended to include more than two data modalities (termed "groups") and was named Group Factor Analysis (GFA) (Klami et al., 2015; Virtanen et al., 2012). Examples of GFA applications are still scarce: it has mostly been used on genomics data (Klami et al., 2013; Suvitaival et al., 2014; Zhao et al., 2016), drug response data (Khan et al., 2014; Klami et al., 2015) and task-based fMRI (Klami et al., 2015; Virtanen et al., 2011; 2012). To the best of our knowledge, GFA has not been applied to uncover associations between brain connectivity and behaviour, especially using high dimensional data.

The original GFA implementation still does not address the third limitation mentioned above, i.e., it cannot be applied to data modalities with missing data. Therefore, we propose an extension of GFA that can handle missing data and allows more flexible assumptions about noise. We first applied our GFA extension to synthetic data to assess whether it can find known associations among data modalities. We then applied it to data from the Human Connectome Project (HCP) to uncover associations between brain connectivity and non-imaging measures (e.g., demographics, psychometrics and other behavioural measures). We evaluated the consistency of the findings across different experiments with complete and incomplete data sets. Finally, even though the GFA model was proposed for unsupervised tasks, it can also be used as a predictive model: we applied our GFA implementation to synthetic and HCP data to assess whether it was able to predict missing data and non-observed data modalities from those observed, in incomplete data sets.

To illustrate the differences between GFA and CCA, we also applied CCA to both datasets. First, we hypothesised that GFA would replicate previous CCA findings using broadly the same HCP dataset, where previous investigators identified a single mode of population covariation representing a "positive-negative" factor linking lifestyle, demographic and psychometric measures to specific patterns of brain connectivity (Smith et al., 2015). Second, we expected CCA to show poorer performance when data was missing, whereas GFA results would be consistent across experiments with complete and incomplete data sets. Due to its flexibility and robustness, the proposed extension of GFA provides an integrative framework that can be used to uncover associations among multiple data modalities in benchmark neuroimaging datasets.

2. Materials and methods

We first describe the link between CCA and GFA (Section 2.1), then we explain how we modified the GFA model to accommodate missing data (Section 2.2) and used it to make predictions (Section 2.3). These subsections are followed by descriptions of experiments where we assess the effectiveness of the model on synthetic data (Section 2.4.1), as well as on HCP data (Section 2.4.2).

2.1. From CCA to GFA

In this section, we show that the probabilistic extension of CCA serves as a building block for GFA. We begin by describing CCA (Section 2.1.1), which is followed by the descriptions of probabilistic (Section 2.1.2) and Bayesian CCA (Section 2.1.3). We finish this section by describing the GFA model and its inference (Section 2.1.4).

2.1.1. CCA

Canonical Correlation Analysis was introduced by Hotelling (1936) and is a classical method for seeking maximal correlations between linear combinations of two multivariate data sets, which can be seen as two different data modalities from the same observations or individuals. This can be illustrated using the following notation: $\mathbf{X}^{(1)} \in \mathbb{R}^{D_1 \times N}$ and $\mathbf{X}^{(2)} \in \mathbb{R}^{D_2 \times N}$ are two data matrices containing multivariate data from the same N observations, where D_1 and D_2 denote the number of variables of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively. CCA finds pairs of weight vectors $\mathbf{u}_k \in \mathbb{R}^{D_1 \times 1}$ and $\mathbf{v}_k \in \mathbb{R}^{D_2 \times 1}$ that maximise the correlation between the corresponding projections $\mathbf{u}_k^T \mathbf{X}^{(1)}$ and $\mathbf{v}_k^T \mathbf{X}^{(2)}$ (also known as canonical scores), $k = 1, \dots, K$ (where K is the number of canonical directions, also called CCA modes). This is achieved by solving:

$$\begin{aligned} &\max_{\mathbf{u}_k, \mathbf{v}_k} \mathbf{u}_k^T \mathbf{X}^{(1)} \mathbf{X}^{(2)T} \mathbf{v}_k, \\ &\text{subject to } \mathbf{u}_k^T \mathbf{X}^{(1)} \mathbf{X}^{(1)T} \mathbf{u}_k = 1 \text{ and } \mathbf{v}_k^T \mathbf{X}^{(2)} \mathbf{X}^{(2)T} \mathbf{v}_k = 1, \end{aligned} \tag{1}$$

where the variables (i.e., rows of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$) are considered to be standardised to zero mean and unit variance. The optimisation problem in Eq. (1) can be solved using a standard eigenvalue solution (Hotelling, 1936), singular value decomposition (SVD) (Uurtio et al., 2017), alternating least squares (ALS) (Golub and Zha, 1994) or nonlinear iterative partial least squares (NIPALS) (Wegelin, 2000).

As mentioned above, CCA lacks robust inference methods and does not model the associations within data modalities. Probabilistic approaches, such as probabilistic CCA, might be used to overcome these limitations, in which the generative nature of the models offers straightforward extensions to novel models through simple changes of the generative description, and more robust inference methods (e.g., Bayesian inference).

2.1.2. Probabilistic CCA

The probabilistic interpretation of CCA (Bach and Jordan, 2006) assumes that N observations of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ (similarly defined as above) are generated by the same latent variables $\mathbf{Z} \in \mathbb{R}^{K \times N}$ capturing the associations between data modalities (Fig. 1a), where K corresponds to the number of components (which are equivalent to the CCA modes described in Section 2.1.1):

$$\begin{aligned} \mathbf{z}_n &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \\ \mathbf{x}_n^{(1)} &\sim \mathcal{N}(\mathbf{A}^{(1)} \mathbf{z}_n + \boldsymbol{\mu}^{(1)}, \boldsymbol{\Psi}^{(1)}), \\ \mathbf{x}_n^{(2)} &\sim \mathcal{N}(\mathbf{A}^{(2)} \mathbf{z}_n + \boldsymbol{\mu}^{(2)}, \boldsymbol{\Psi}^{(2)}), \end{aligned} \tag{2}$$

where $\mathcal{N}(\cdot)$ represents the multivariate normal distribution, $\mathbf{A}^{(1)} \in \mathbb{R}^{D_1 \times K}$ and $\mathbf{A}^{(2)} \in \mathbb{R}^{D_2 \times K}$ are the projection matrices (also known as loading matrices) that represent the transformations of the latent variables $\mathbf{z}_n \in \mathbb{R}^{K \times 1}$ (which corresponds to a column vector of $\mathbf{Z}$) into the input space. The projection matrices are equivalent to the (horizontal) concatenation of all pairs of weight vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ that CCA finds (see Section 2.1.1). $\boldsymbol{\Psi}^{(1)} \in \mathbb{R}^{D_1 \times D_1}$, $\boldsymbol{\Psi}^{(2)} \in \mathbb{R}^{D_2 \times D_2}$ denote the noise covariance matrices.

Bach and Jordan proved that the maximum likelihood estimates of the parameters in Eq. (2) lead to the same canonical directions as standard CCA up to a rotation (Bach and Jordan, 2006), i.e., the posterior

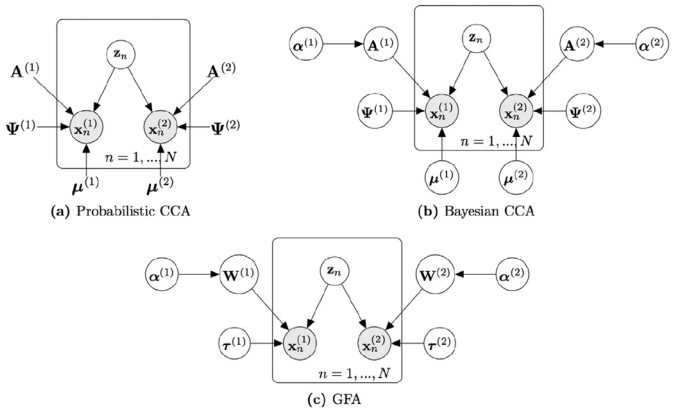

Fig. 1. Graphical representation of (a) Probabilistic CCA, (b) Bayesian CCA and (c) GFA. A separate mean parameter is not included for GFA, but it assumes zero-mean data without loss of generality (Section 2.1.4).

expectations $E(\mathbf{Z}|\mathbf{X}^{(1)})$ and $E(\mathbf{Z}|\mathbf{X}^{(2)})$ lie in the same subspace that standard CCA finds, where the subspace is represented by the canonical scores $\mathbf{U}^T\mathbf{X}^{(1)}$ and $\mathbf{V}^T\mathbf{X}^{(2)}$, where $\mathbf{U} \in \mathbb{R}^{D_1 \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_2 \times K}$. Moreover, an equivalent representation of the latent variables $\mathbf{Z}$ can be obtained - for CCA - by averaging the canonical scores obtained for each data modality (Klami et al., 2013).

Although probabilistic CCA does not provide an explicit inference approach to infer the number of relevant components, it was used as a building block for Bayesian CCA that - as described in the next section - provides a solution for this limitation.

2.1.3. Bayesian CCA

Klami and Kaski (2007) and Chong Wang (2007) proposed a hierarchical Bayesian extension of CCA by giving full Bayesian treatment to the probabilistic CCA model, introducing suitable prior distributions over the model parameters, which can be inferred using Bayesian inference.

The goal of Bayesian inference is to provide a procedure for incorporating our prior beliefs with any evidence (i.e., data) that we can collect to obtain an updated posterior belief. This is done using the Bayes' theorem: $p(\boldsymbol{\Theta}|\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})/p(\mathbf{X})$, where $p(\boldsymbol{\Theta})$ represents the prior distributions over the model parameters $\boldsymbol{\Theta}$ (here, $\boldsymbol{\Theta}$ denotes the model parameters $\{\mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{\mu}\}$ and latent variables $\mathbf{Z}$), $p(\mathbf{X}|\boldsymbol{\Theta})$ represents the likelihood and $p(\boldsymbol{\Theta}|\mathbf{X})$ represents the joint posterior distribution that expresses the uncertainty about the model parameters after accounting for the prior knowledge and data. $p(\mathbf{X})$ represents the model evidence, or marginal likelihood, which is usually considered a normalising constant. In this way, Bayes' theorem is formulated as: $p(\boldsymbol{\Theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})$, which means that the posterior distribution is proportional to the likelihood times the prior.

In the Bayesian CCA model (represented in Fig. 1b), the observations $\mathbf{X}^{(m)}$ are assumed to be generated by Eq. (2). The joint probabilistic distribution of the model is given by Chong Wang (2007):

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{\mu}) = \prod_{m=1}^{M} \Big[p(\mathbf{X}^{(m)}|\mathbf{Z}, \mathbf{A}^{(m)}, \boldsymbol{\Psi}^{(m)}, \boldsymbol{\mu}^{(m)}) \times p(\mathbf{A}^{(m)}|\boldsymbol{\alpha}^{(m)}) p(\boldsymbol{\alpha}^{(m)}) p(\boldsymbol{\Psi}^{(m)}) p(\boldsymbol{\mu}^{(m)}) \Big] p(\mathbf{Z}), \tag{3}$$

where M is the number of data modalities, $\mathbf{A}^{(m)}$ and $\mathbf{Z}$ are defined as in Eq. (2) and $\boldsymbol{\alpha}^{(m)} \in \mathbb{R}^{1 \times K}$. The prior distributions are chosen to be conjugate (i.e., the posterior distribution has the same functional form as the prior distribution) which simplifies the inference:

$$p(\mathbf{A}^{(m)}|\boldsymbol{\alpha}^{(m)}) = \prod_{j=1}^{D_m} \prod_{k=1}^{K} \mathcal{N}(a_{jk}^{(m)}|0, (\alpha_k^{(m)})^{-1}), \quad p(\boldsymbol{\alpha}^{(m)}) = \prod_{k=1}^{K} \Gamma(\alpha_k^{(m)}|a_{\alpha}^{(m)}, b_{\alpha}^{(m)}),$$
$$p(\boldsymbol{\mu}^{(m)}) = \mathcal{N}(\boldsymbol{\mu}^{(m)}|0, (\beta^{(m)})^{-1}\mathbf{I}), \quad p(\boldsymbol{\Psi}^{(m)}) = \mathcal{W}^{-1}(\boldsymbol{\Psi}^{(m)}|\mathbf{S}_0^{(m)}, \nu_0^{(m)}), \tag{4}$$

where $\mathbf{S}_0^{(m)}$ is a symmetric positive definite matrix, $\nu_0^{(m)}$ denotes the degrees of freedom for the inverse Wishart distribution ($\mathcal{W}^{-1}(\cdot)$) and $\Gamma(\cdot)$ represents the Gamma distribution. The prior over the projection matrices $\mathbf{A}^{(m)}$ is the Automatic Relevance Determination (ARD) prior (Mackay, 1995), which is used to find the relevant latent components (i.e., rows of $\mathbf{Z}$). This is done by allowing some $\alpha_k^{(m)}$ to be pushed towards infinity, which consequently drives the loadings (i.e., elements of the projection/loading matrices) of the k columns of $\mathbf{A}^{(m)}$ close to zero and the corresponding irrelevant latent components k to be pruned out during inference.

For learning the Bayesian CCA model, we need to infer the model parameters and latent variables from data, which can be done by estimating the posterior distribution $p(\mathbf{Z}, \mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{\mu}|\mathbf{X})$ and marginalising out uninteresting variables. However, these marginalisations are often analytically intractable, and therefore the posterior distribution needs to be approximated. This can be done using mean-field variational Bayes (Chong Wang, 2007) or Gibbs sampling (Klami and Kaski, 2007), since all conditional distributions are conjugate. However, the inference of the Bayesian CCA model is difficult for high dimensional data as the posterior distribution needs to be estimated over large covariance matrices $\boldsymbol{\Psi}^{(m)}$ (Klami et al., 2013). The inference algorithms usually need to invert those matrices in every step, which results in $O(D_m^3)$ complexity, leading to long computational times. Moreover, Bayesian CCA does not account for the modality-specific associations.

Virtanen et al. (2011) proposed an extension of Bayesian CCA to impose modality-wise sparsity to separate associations between data modalities from those within data modalities. Moreover, this model assumes spherical noise covariance matrices ($\boldsymbol{\Psi}^{(m)} = \sigma^{(m)^2}\mathbf{I}$, where $\sigma^{(m)^2}$ corresponds to the noise variance of data modality m) for more efficient inference. The same authors proposed a further extension of the model to uncover associations between more than two groups (e.g., data modalities), called Group Factor Analysis (GFA) (Klami et al., 2015; Virtanen et al., 2012).

2.1.4. Group factor analysis

In the GFA problem, we assume that a collection of N observations, stored in $\mathbf{X} \in \mathbb{R}^{D \times N}$, have disjoint M partitions of variables D_m called groups. In this and the following two sections (Sections 2.2 and 2.3), we refer to a given data modality as a group of variables of $\mathbf{X}$ ($\mathbf{X}^{(m)} \in \mathbb{R}^{D_m \times N}$ for the m-th group), in accordance with the GFA nomenclature. Moreover, we introduce the concept "factor" that corresponds to the loadings in a given column k of the loading matrices (represented as $\mathbf{W}$ in Fig. 1c). The latent factors correspond to the rows of the latent variables $\mathbf{Z} \in \mathbb{R}^{K \times N}$ (equivalent to a latent component in probabilistic and Bayesian CCA).

GFA finds the set of K latent factors that can separate the associations between groups (i.e., shared factors) from those within groups (i.e., group-specific factors) by considering a joint factor model (Fig. 1c), where each m-th group is generated as follows Klami et al. (2015); Virtanen et al. (2012):

$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K),$$
$$\mathbf{x}_n^{(m)} \sim \mathcal{N}(\mathbf{W}^{(m)}\mathbf{z}_n, \mathbf{T}^{(m)^{-1}}), \tag{5}$$

where $\mathbf{T}^{(m)^{-1}}$ is a diagonal covariance matrix ($\mathbf{T}^{(m)} = \text{diag}(\boldsymbol{\tau}^{(m)})$, where $\boldsymbol{\tau}^{(m)}$ represents the noise precision, i.e., inverse noise variance of the m-th group), $\mathbf{W}^{(m)} \in \mathbb{R}^{D_m \times K}$ is the loading matrix of the m-th group and $\mathbf{z}_n \in \mathbb{R}^{K \times 1}$ is the latent variable for a given observation $\mathbf{x}_n^{(m)}$ (i.e., column of $\mathbf{X}^{(m)}$). The model assumes zero-mean data without loss of generality. Alternatively, a separate mean parameter could have been included; however, its estimate would converge close to the empirical mean, which can be subtracted from the data before estimating the model parameters (Klami et al., 2013).

If we consider $M = 2$ (also known as Bayesian CCA via group sparsity (Virtanen et al., 2011) or Bayesian inter-battery factor analysis (Klami et al., 2013)), the noise covariance matrix is given by $\mathbf{T} =$

$\begin{pmatrix} \mathbf{T}^{(1)} & 0 \\ 0 & \mathbf{T}^{(2)} \end{pmatrix}$ and $\mathbf{W} = \begin{bmatrix} \mathbf{A}^{(1)} & \mathbf{B}^{(1)} & \mathbf{0} \\ \mathbf{A}^{(2)} & \mathbf{0} & \mathbf{B}^{(2)} \end{bmatrix}$, where $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ represent the loading matrices containing the shared factors and $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ correspond to the loading matrices containing the group-specific factors. The structure of $\mathbf{W}$ and the corresponding latent structure (represented by $\mathbf{Z}$) is learned automatically by imposing group-wise sparsity on the factors, i.e., the matrices $\mathbf{A}$ and $\mathbf{B}$ are not explicitly specified (Klami et al., 2013). This is achieved by assuming independent ARD priors to encourage sparsity over the groups (Klami et al., 2013; Virtanen et al., 2011):

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{m=1}^{M}\prod_{j=1}^{D_m}\prod_{k=1}^{K}\mathcal{N}(w_{jk}^{(m)}|0,(\alpha_k^{(m)})^{-1}), \quad p(\boldsymbol{\alpha}) = \prod_{m=1}^{M}\prod_{k=1}^{K}\Gamma(\alpha_k^{(m)}|a_{\boldsymbol{\alpha}^{(m)}},b_{\boldsymbol{\alpha}^{(m)}}), \tag{6}$$

which is a simple extension of the single ARD prior used by Chong Wang (2007). Here, a separate ARD prior is used for each $\mathbf{W}^{(m)}$, which are chosen to be uninformative to enable the automatic pruning of irrelevant latent factors. $\Gamma(\cdot)$ represents a gamma distribution with shape parameter $a_{\boldsymbol{\alpha}^{(m)}}$ and rate parameter $b_{\boldsymbol{\alpha}^{(m)}}$. These separate priors cause groups of variables to be pushed close to zero for some factors k ($\mathbf{w}_k^{(m)} \to 0$) by driving the corresponding $\alpha_k^{(m)}$ towards infinity. If the loadings of certain factors are pushed towards zero for all groups, the underlying latent factor is deemed inactive and pruned out. Klami et al. (2013). Finally, the prior distributions over the noise and latent variables $\mathbf{Z}$ are:

$$p(\boldsymbol{\tau}) = \prod_{m=1}^{M}\prod_{j=1}^{D_m}\Gamma(\tau_j^{(m)}|a_{\boldsymbol{\tau}^{(m)}},b_{\boldsymbol{\tau}^{(m)}}), \quad p(\mathbf{Z}) = \prod_{k=1}^{K}\prod_{n=1}^{N}\mathcal{N}(z_{kn}|0,1), \tag{7}$$

where $\Gamma(\cdot)$ represents a gamma distribution with shape parameter $a_{\boldsymbol{\tau}^{(m)}}$ and rate parameter $b_{\boldsymbol{\tau}^{(m)}}$. The hyperparameters $a_{\boldsymbol{\alpha}^{(m)}}, b_{\boldsymbol{\alpha}^{(m)}}, a_{\boldsymbol{\tau}^{(m)}}, b_{\boldsymbol{\tau}^{(m)}}$ can be set to a very small number (e.g., 10^{-14}), resulting in uninformative priors. The joint distribution $p(\mathbf{X},\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\tau})$ is hence given by:

$$p(\mathbf{X},\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\tau}) = p(\mathbf{X}|\mathbf{Z},\mathbf{W},\boldsymbol{\tau})p(\mathbf{Z})p(\mathbf{W}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\boldsymbol{\tau}). \tag{8}$$

As mentioned in Section 2.1.3, the calculations needed to infer the model parameters and latent variables from data are often analytically intractable. Therefore, the posterior distribution needs to be approximated by applying, for instance, mean field variational approximation (similarly to Bayesian CCA (Chong Wang, 2007)). This involves approximating the true posterior $p(\boldsymbol{\theta}|\mathbf{X})$ by a suitable factorized distribution $q(\boldsymbol{\theta})$ (Bishop, 1999). The marginal log-likelihood ($\ln p(\mathbf{X})$) can be decomposed as follows Bishop (2006):

$$\begin{aligned} &\ln p(\mathbf{X}) = \mathcal{L}(q) + D_{KL}(q||p), \\ &\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \ln \tfrac{p(\mathbf{X},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}, \\ &D_{KL}(q||p) = \int q(\boldsymbol{\theta}) \ln \tfrac{p(\boldsymbol{\theta}|\mathbf{X})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}, \end{aligned} \tag{9}$$

where $D_{KL}(q||p)$ is the Kullback-Leibler divergence between $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{X})$ and $\mathcal{L}(q)$ is the lower bound of the marginal log-likelihood. Since $\ln p(\mathbf{X})$ is constant, maximising the lower bound $\mathcal{L}(q)$ is equivalent to minimising the KL divergence $D_{KL}(q||p)$, which means $q(\boldsymbol{\theta})$ can be used to approximate the true posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ (Bishop, 1999). Assuming that $q(\boldsymbol{\theta})$ can be factorised such that $q(\boldsymbol{\theta}) = \prod_i q_i(\boldsymbol{\theta}_i)$, the $\mathcal{L}(q)$ can be maximised with respect to all possible distributions $q_i(\boldsymbol{\theta}_i)$ as follows Bishop (1999, 2006):

$$\ln q_i(\boldsymbol{\theta}_i) = \langle \ln p(\mathbf{X},\boldsymbol{\theta}) \rangle_{j\neq i} + \text{const}, \tag{10}$$

where $\langle\cdot\rangle_{j\neq i}$ denotes the expectation taken with respect to $\prod_{j\neq i} q_j(\boldsymbol{\theta}_j)$ for all $j \neq i$. In GFA, the full posterior is approximated by:

$$q(\boldsymbol{\theta}) = q(\mathbf{Z})\prod_{m=1}^{M}\left[q(\mathbf{W}^{(m)})q(\boldsymbol{\alpha}^{(m)})q(\boldsymbol{\tau}^{(m)})\right], \tag{11}$$

where $\boldsymbol{\theta}$ denotes the model parameters and latent variables ($\boldsymbol{\theta} = \{\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\tau}\}$). As conjugate priors are used, the free-form optimisation of $q(\boldsymbol{\theta})$ (using Eq. (10)) results in the following analytically tractable distributions:

$$\begin{aligned} &q(\mathbf{Z}) = \textstyle\prod_{n=1}^{N}\mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_{\mathbf{z}_n},\boldsymbol{\Sigma}_{\mathbf{z}_n}), \quad q(\mathbf{W}^{(m)}) = \textstyle\prod_{j=1}^{D_m}\mathcal{N}(\mathbf{W}_{j,*}^{(m)}|\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}},\boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}}), \\ &q(\boldsymbol{\alpha}^{(m)}) = \textstyle\prod_{k=1}^{K}\Gamma(\alpha_k^{(m)}|\tilde{a}_{\boldsymbol{\alpha}^{(m)}},\tilde{b}_{\boldsymbol{\alpha}^{(m)}}^{(k)}), \quad q(\boldsymbol{\tau}^{(m)}) = \textstyle\prod_{j=1}^{D_m}\Gamma(\tau_j^{(m)}|\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)},\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)}), \end{aligned} \tag{12}$$

where $\mathbf{z}_n$ is the n-th column of $\mathbf{Z}$ and $\mathbf{W}_{j,*}^{(m)}$ denotes the j-th row of $\mathbf{W}^{(m)}$. The optimisation is done using variational Expectation-Maximization (EM), where the parameters in Eq. (12) are updated sequentially until convergence, which is achieved when a relative change of the evidence lower bound (ELBO) $\mathcal{L}(q)$ falls below an arbitrary low number (e.g., 10^{-6}). The recommended choice for the maximal number of latent factors is $K = \min(D_1, D_2)$, but in some settings this leads to large K and consequently long computational times (Klami et al., 2013). In practice, a K value that leads to the removal of some irrelevant latent factors should be a reasonable choice (Klami et al., 2013). In our experiments with synthetic data, we initialised the model with different values of K and the results were consistent across the different experiments (Supplementary Fig. 1).

2.2. Our proposed GFA extension

Here, we propose an extension of the GFA model to handle missing data by modifying the inference algorithm of variational factor analysis proposed by Luttinen and Ilin (2010). The extended GFA model assumes independent noise for each variable (i.e., diagonal noise) within a group ($p(\boldsymbol{\tau}) = \prod_{m=1}^{M}\prod_{j=1}^{D_m}\Gamma(\tau_j^{(m)}|a_{\boldsymbol{\tau}^{(m)}},b_{\boldsymbol{\tau}^{(m)}})$). This assumption enables a more flexible model because a noise variance parameter can be computed for each variable (which is useful to inform us about the uncertainty of each variable). Furthermore, we use only the noise parameters of non-missing variables when updating the parameters of the posterior distribution.

In summary, the proposed inference algorithm (Algorithm 1) starts by updating the parameters of the distribution over each latent variable ($q(\mathbf{z}_n)$) using the loadings and noise parameters of the non-missing variables of the n-th sample/subject ($j \in O_n^{(m)}$, where $O_n^{(m)}$ is the set of indices in the n-th column of $\mathbf{X}^{(m)}$ that are not missing). After that, the parameters of the distribution over each row of the loading matrices are computed using the updated latent variables of the non-missing samples of the j-th variable ($n \in O_j^{(m)}$, where $O_j^{(m)}$ is the set of indices in the j-th row of $\mathbf{X}^{(m)}$ that are not missing). The parameters of the distribution over $\boldsymbol{\alpha}^{(m)}$ and $\boldsymbol{\tau}^{(m)}$ are then updated using the updated latent variables and loading matrices. Finally, the ELBO is calculated with the updated parameters. These update steps are repeated until convergence, i.e., when a relative change of the ELBO falls below an arbitrarily low number (10^{-6} in our implementation). The derivations of the variational update rules and ELBO calculations can be found in Appendix A and Appendix B, respectively.

Although we just show here examples of our GFA extension being applied to two data modalities, our Python implementation (Section 2.5) can be used for more than two data modalities.

2.3. Multi-output and missing data prediction

As mentioned above, GFA can be used as a predictive model. As the groups are generated by the same latent variables, the unobserved group of new (test) observations ($\mathbf{X}^{(m)\star}$) can be predicted from the observed ones on the test set ($\mathbf{X}^{-(m)\star}$) using the predictive distribution $p(\mathbf{X}^{(m)\star}|\mathbf{X}^{-(m)\star})$ (Klami et al., 2015). This distribution is analytically intractable, but its expectation can be approximated using the parameters learned during the variational approximation (Appendix B) as follows Klami et al. (2015):

$$\begin{aligned} \mathbb{E}[\mathbf{X}^{(m)\star}|\mathbf{X}^{-(m)\star}] = &\quad \langle \mathbf{W}^{(m)}\mathbf{Z}\rangle_{q(\mathbf{W}^{(m)}),\, q(\mathbf{Z}|\mathbf{X}^{-(m)\star})}, \\ = &\quad \langle \mathbf{W}^{(m)}\rangle \boldsymbol{\Sigma}_Z^{\star} \langle \mathbf{W}^{-(m)^T}\rangle \mathbf{T}^{\star}\mathbf{X}^{-(m)\star}, \end{aligned} \tag{13}$$

Algorithm 1 Pseudocode of the variational updates of GFA to handle missing data.

repeat

▷ Update $q(\mathbf{Z})$

for $n = 1 \ldots N$ **do**

$$\boldsymbol{\Sigma}_{\mathbf{z}_n} \leftarrow \left[\mathbf{I}_K + \sum_{m=1}^{M} \sum_{j \in O_n^{(m)}} \frac{\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)}}{\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)}} \left(\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}^T \boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}} + \boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}} \right)\right]^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{z}_n} \leftarrow \boldsymbol{\Sigma}_{\mathbf{z}_n} \left(\sum_{m=1}^{M} \sum_{j \in O_n^{(m)}} \frac{\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)}}{\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)}} \boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}^T x_{j,n}^{(m)} \right)$$

end for

for $m = 1 \ldots M$ **do**

▷ Update $q(\mathbf{W}^{(m)})$

$$\mathbf{H}^{(m)} \leftarrow \mathrm{diag}\left(\frac{\tilde{a}_{\boldsymbol{\alpha}^{(m)}}}{\tilde{\mathbf{b}}_{\boldsymbol{\alpha}^{(m)}}} \right)$$

for $j = 1 \ldots J$ **do**

$$\boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}} \leftarrow \left[\mathbf{H}^{(m)} + \frac{\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)}}{\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)}} \sum_{n \in O_j^{(m)}} \left(\boldsymbol{\mu}_{\mathbf{z}_n} \boldsymbol{\mu}_{\mathbf{z}_n}^T + \boldsymbol{\Sigma}_{\mathbf{z}_n} \right) \right]^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}} \leftarrow \frac{\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)}}{\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)}} \left(\sum_{n \in O_j^{(m)}} x_{j,n}^{(m)} \boldsymbol{\mu}_{\mathbf{z}_n}^T \right) \boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}}$$

end for

▷ Update $q(\boldsymbol{\alpha}^{(m)})$

$$\mathbf{C}^{(m)} \leftarrow \sum_{j=1}^{D_m} \left(\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}^T \boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}} + \boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}} \right)$$

$$\tilde{a}_{\boldsymbol{\alpha}^{(m)}} \leftarrow a_{\boldsymbol{\alpha}^{(m)}} + \frac{1}{2} D_m$$

$$\tilde{b}_{\boldsymbol{\alpha}^{(m)}}^{(k)} \leftarrow b_{\boldsymbol{\alpha}^{(m)}} + \frac{1}{2} c_{k,k}^{(m)}$$

▷ Update $q(\boldsymbol{\tau}^{(m)})$

for $j = 1 \ldots J$ **do**

$$\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)} \leftarrow a_{\boldsymbol{\tau}^{(m)}} + \frac{1}{2} N_j^{(m)}$$

$$\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)} \leftarrow b_{\boldsymbol{\tau}^{(m)}} + \frac{1}{2} \sum_{n \in O_j^{(m)}} \left(x_{j,n}^{(m)2} - 2 x_{j,n}^{(m)} \boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}} \boldsymbol{\mu}_{\mathbf{z}_n} \right) + \frac{1}{2} \sum_{n \in O_j^{(m)}} \mathrm{Tr}\left[\left(\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}^T \boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}} + \boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}} \right) \left(\boldsymbol{\mu}_{\mathbf{z}_n} \boldsymbol{\mu}_{\mathbf{z}_n}^T + \boldsymbol{\Sigma}_{\mathbf{z}_n} \right) \right]$$

end for

end for

until convergence

where $\langle \cdot \rangle$ denotes expectations, $\boldsymbol{\Sigma}_Z^\star = \mathbf{I}_K + \sum_{l \neq m} \sum_j^{D_l} \langle \tau_j^{(l)} \rangle \langle \mathbf{W}_{j,*}^{(l)T} \mathbf{W}_{j,*}^{(l)} \rangle$, $\langle \mathbf{W}_{j,*}^{(l)T} \mathbf{W}_{j,*}^{(l)} \rangle = \boldsymbol{\Sigma}_{\mathbf{w}_j^{(l)}} + \boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(l)}}^T \boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}$ ($\boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}}$ and $\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}$ are the variational parameters obtained for $q(\mathbf{W}^{(m)})$ in Eq. A.11) and $\mathbf{T}^\star = \{\mathrm{diag}(\langle \boldsymbol{\tau}^{(l)} \rangle)\}_{l \neq m}$. In all experiments, $\mathbb{E}[\mathbf{X}^{(m)\star} | \mathbf{X}^{-(m)\star}]$ was used for prediction.

Additionally, the missing data can be predicted using Eq. (13) where, in this case, the observed groups $\mathbf{X}^{-(m)\star}$ correspond to the training observations in group m and the missing data is represented as $\mathbf{X}^{(m)\star} = \mathbf{X}_{nj \in O_{nj}^{(m)}}^{(m)\star}$.

2.4. Experiments

We begin this section by detailing the experiments that we ran on synthetic data (Section 2.4.1), which is followed by the description of the experiments on the HCP dataset (Section 2.4.2).

2.4.1. Synthetic data

We validated the extended GFA model on synthetic data drawn from Eq. (5). We generated $N = 500$ observations for two data modalities with $D_1 = 50$ ($\mathbf{X}^{(1)} \in \mathbb{R}^{50 \times 500}$) and $D_2 = 30$ ($\mathbf{X}^{(2)} \in \mathbb{R}^{30 \times 500}$), respectively. The data modalities were generated from two shared and two modality-specific latent factors, which were manually specified, similarly to the examples generated in Klami et al. (2013) (Fig. 2). The shared factors correspond to latent factor 1 and 2, the latent factor specific to $\mathbf{X}^{(1)}$ is represented in latent factor 4 and the latent factor specific to $\mathbf{X}^{(2)}$ is represented in latent factor 3. The $\boldsymbol{\alpha}^{(m)}$ parameters were set to 1 for the active factors and 10^6 for the inactive ones. The loading matrices $\mathbf{W}^{(m)}$ were drawn from the prior (Eq. (6)) and diagonal noise with fixed precisions ($\tau_1 = 5\mathbf{I}_{D_1}$ and $\tau_2 = 10\mathbf{I}_{D_2}$) was added to the observations.

We ran experiments with the proposed extension of GFA on the following selections of synthetic data:

1. *Complete data*. In this experiment, we compared the extended GFA model to the vanilla GFA implementation of Klami et al. (2015).
2. *Incomplete data*:
   (a) 20% of the elements of $\mathbf{X}^{(2)}$ were randomly removed.
   (b) 20% of the observations (i.e., rows) in $\mathbf{X}^{(1)}$ were randomly removed.

In all experiments, the model was initialised with $K = 15$ (number of latent factors) to assess whether it can learn the true latent factors while automatically removing the irrelevant ones. We ran additional experiments with complete data where the model was initialised with $K = 30$ to assess whether it could still converge to a good solution when the number of latent factors were overestimated in low and high dimensional data (Supplementary Fig. 1).

As the variational approximations for GFA are deterministic, and the model converges to a local optimum that depends on the initialisation, all experiments were randomly initialised 10 times. The initialisation with the largest variational lower bound was considered to be the best one. For visualization purposes, we matched the true and inferred latent factors by calculating the maximum similarity (using Pearson's correlation) between them, in all experiments. If a correlation value was negative, the corresponding inferred factor was multiplied by -1. The inferred factors with correlations greater than 0.70 were visually compared with the true ones.

For each random initialisation, in all experiments, the data was split into training (80%) and test (20%) sets. The model performance was assessed by predicting one data modality from the other on the test set (e.g., predict $\mathbf{X}^{(2)}$ from $\mathbf{X}^{(1)}$) using Eq. (13). The mean and standard deviation of the mean squared error (MSE) (calculated between the true and predicted values of the non-observed data modality on the test set) was calculated across the different initialisations. The chance level of each experiment was obtained by calculating the MSE between the observations on the test set and the means of the corresponding variables on the training set.

In the incomplete data experiments, the missing data was predicted using Eq. (13). We calculated the mean and standard deviation (across initialisations) of the Pearson's correlations between the true and predicted missing values to assess the ability of the model to predict missing data. To compare our results with a common strategy for data imputation in the incomplete data experiments, we ran GFA with complete data, after imputing the missing values using the median of the respective variable. We ran additional experiments with missing data (see Supplementary Materials and Methods), including when values from the tails of the distribution of $\mathbf{X}^{(2)}$ were randomly removed (Supplementary Fig. 2a) and when values in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ were missing for low (Supplementary Fig. 2b) and high dimensional data (Supplementary Fig. 2c). Furthermore, the performance of the proposed extension of GFA was assessed with increasing percentages of missing data when values of $\mathbf{X}^{(2)}$ were missing from the tails of the distribution (Supplementary Fig. 5a) or randomly (Supplementary Fig. 5b). For each of these settings, we ran experiments with no missing data in $\mathbf{X}^{(1)}$ and 20% missing rows in $\mathbf{X}^{(1)}$ (blue and orange lines in Supplementary Fig.5, respectively). Finally, we also ran experiments applying CCA to complete and incomplete data (Supplementary Fig. 4).

2.4.2. HCP Data

We applied our GFA extension to the publicly available resting-state functional MRI (rs-fMRI) and non-imaging measures (e.g., demographics, psychometrics and other behavioural measures) obtained from 1003

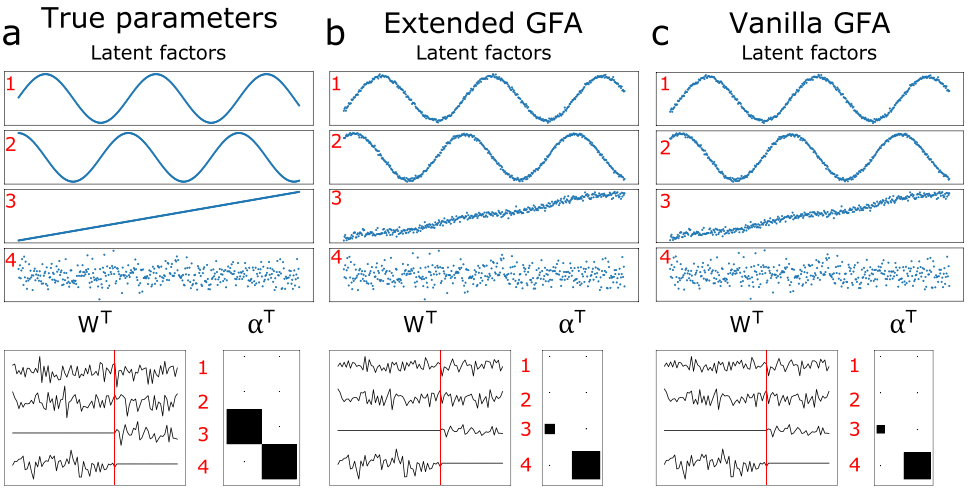

Fig. 2. Complete data experiment: (a) true latent factors and model parameters; (b) inferred latent factors and model parameters obtained with our GFA extension; (c) inferred latent factors and model parameters obtained with the vanilla GFA implementation of Klami et al. (2015). The latent factors and parameters used to generate the data are plotted on the left-hand side, and the ones inferred by the model are plotted on the right-hand side. The four rows on the top represent the four latent factors. The loading matrices of the first and second data modality are represented on the left and right-hand side of the red line in $\mathbf{W}^T$, respectively. The alphas of the first and second data modality are shown in the form of a Hinton diagram in the first and second columns of $\boldsymbol{\alpha}^T$, respectively, where the alphas are proportional to the area of the squares. The small black dots and big black squares represent active and inactive factors, respectively.

subjects (only these had rs-fMRI data available) of the 1200-subject data release of the HCP (https://www.humanconnectome.org/study/hcp-young-adult/data-releases). Two subjects were missing the family structure information that we needed to perform the restricted permutations in the CCA analysis, so were excluded.

In particular, we used the brain connectivity features of the extensively processed rs-fMRI data using pairwise partial correlations between 200 brain regions from a parcellation estimated by independent component analysis. The data processing was identical to Smith et al. (2015), yielding 19,900 brain variables for each subject (i.e., the lower triangular part of the brain connectivity matrix containing pair-wise connectivity among all 200 regions). The vectors were concatenated across subjects to form $\mathbf{X}^{(1)} \in \mathbb{R}^{19900\times1001}$. We used 145 items of the non-imaging measures used in Smith et al. (2015) as the remaining measures (SR_Aggr_Pct, ASR_Attn_Pct, ASR_Intr_Pct, ASR_Rule_Pct, ASR_Soma_Pct, ASR_Thot_Pct, ASR_Witd_Pct, DSM_Adh_Pct, DSM_Antis_Pct, DSM_Anxi_Pct, DSM_Avoid_Pct, DSM_Depr_Pct, DSM_Somp_Pct) were not available in the 1200-subject data release. The non-imaging matrix contained 145 variables from 1001 subjects ($\mathbf{X}^{(2)} \in \mathbb{R}^{145\times1001}$).

Similarly to Smith et al. (2015), nine confounding variables (acquisition reconstruction software version, summary statistic quantifying average subject head motion during acquisition, weight, height, blood pressure systolic, blood pressure diastolic, hemoglobin A1C measured in blood, the cube-root of total brain and intracranial volumes estimated by FreeSurfer) were regressed out from both data modalities. Finally, each variable was standardised to have zero mean and unit variance. For additional details of the data acquisition and processing, see Smith et al. (2015).

We ran GFA experiments on the following selections of HCP data:

1. *Complete data*.
2. *Incomplete data*:
   (a) 20% of the elements of $\mathbf{X}^{(2)}$ were randomly removed.
   (b) 20% of the subjects were randomly removed from $\mathbf{X}^{(1)}$.

In all experiments, the model was initialised with $K = 80$ latent factors. As in the experiments with synthetic data, all experiments were randomly initialised 10 times and the data was randomly split into training (80%) and test (20%) sets. The initialisation with the largest variational lower bound was considered to be the best one.

As a considerable number of relevant factors might remain after automatically pruning out the noisy ones, showing all factors is not possible due to space constraints. Furthermore, as the number of brain connectivity variables is much greater than non-imaging measures (~100 times more brain connectivity variables than non-imaging measures), using the percentage of variance explained by each factor is not a good criterion, because the factors explaining most variance in the data are most likely brain-specific (Supplementary Fig. 6a). Therefore, we propose a criterion to identify the most relevant factors by calculating the relative variance explained (rvar) by each factor k within each data modality m (i.e., k-th column of $\mathbf{W}^{(m)}$):

$$\text{rvar}_k^{(m)} = \frac{\mathbf{w}_k^{(m)T}\mathbf{w}_k^{(m)}}{\text{Tr}(\mathbf{W}^{(m)}\mathbf{W}^{(m)T})}, \tag{14}$$

where $\text{Tr}(\cdot)$ represents the trace of the matrix. The factors explaining more than 7.5% variance within any data modality were considered most relevant. Then, in order to decide whether a given most relevant factor was modality-specific or shared, the ratio between the variance explained (var) by the non-imaging and brain loadings of the k-th factor was computed:

$$r_k = \frac{\text{var}_k^{(2)}}{\text{var}_k^{(1)}}, \tag{15}$$

where $\text{var}_k^{(m)} = \frac{\mathbf{w}_k^{(m)T}\mathbf{w}_k^{(m)}}{\text{Tr}(\mathbf{W}^{(m)}\mathbf{W}^{(m)T}+\mathbf{T}^{(m)^{-1}})}$, and $\mathbf{T}^{(m)^{-1}}$ is the diagonal covariance matrix in Eq. (5). A factor was considered shared if $0.001 \leq r_k \leq 300$, non-imaging specific if $r_k > 300$ or brain-specific if $r_k < 0.001$ (Supplementary Fig. 6b illustrates how many factors would be considered shared or specific in the complete HCP data using these thresholds). These values were selected taking into account that there was an imbalance in the total number of variables across the data modalities. These thresholds were validated in high dimensional synthetic data (Supplementary Table 1).

To assess whether the missing data affected the estimation of the most relevant factors, we calculated the Pearson's correlations between the factors obtained in the complete data experiment and the incomplete data experiments. In the multi-output prediction task, all non-imaging measures were predicted from brain connectivity on the test set. The model performance was assessed by calculating the mean and standard deviation of the relative MSE (rMSE) between the true and predicted values of each non-imaging measure on the test set, across the different initialisations:

$$\text{rMSE}_j = \frac{\frac{1}{N}\sum_{n=1}^{N}(x_{nj}^{(2)} - x_{nj}^{(2)^*})^2}{\frac{1}{N}\sum_{n=1}^{N}(x_{nj}^{(2)})^2}, \tag{16}$$

where N is the number of subjects, $x_{nj}^{(2)}$ and $x_{nj}^{(2)^*}$ are the true and predicted non-imaging measure j on the test set. The chance level was obtained by calculating the relative MSE between each non-imaging mea-

Table 1
Most relevant shared and modality-specific factors obtained with complete data according to the proposed criteria. Factors explaining more than 7.5% variance within any data modality were considered most relevant. A factor was considered shared if $0.001 \leq r_k \leq 300$, non-imaging (NI) specific if $r_k > 300$ or brain-specific if $r_k < 0.001$. rvar - relative variance explained; var - variance explained; r_k - ratio between the variance explained by the non-imaging and brain loadings in factor k.

	rvar (%)		**var** (%)		$\mathbf{r_k}$
Factors	Brain	NI	Brain	NI	$\text{var}_{\text{NI}}/\text{var}_{\text{brain}}$
Shared	0.096	8.103	0.007	0.028	4.03
b	0.032	17.627	0.002	0.061	26.22
c	0.011	9.869	7.65×10^{-4}	0.034	44.32
d	0.008	33.336	5.46×10^{-4}	0.114	209.65
Brain	14.267	2.311×10^{-9}	1.028	7.93×10^{-12}	7.72×10^{-12}
b	11.407	0.036	0.822	1.23×10^{-4}	1.50×10^{-4}

sure in the test set and the mean of the corresponding non-imaging measure in the training data.

Similarly to the incomplete data experiments on synthetic data, the missing data was predicted using Eq. (13) and the mean and standard deviation (across initialisations) of the Pearson's correlations between the true and predicted missing values were calculated.

2.5. Data and code availability

The data used in this study was downloaded from the Human Connectome Project website (https://www.humanconnectome.org/study/hcp-young-adult/document/extensively-processed-fmri-data-documentation). The GFA models and experiments were implemented in Python 3.9.1 and are available here: https://github.com/ferreirafabio80/gfa. The CCA experiments (Supplementary Materials and Methods) were run in a MATLAB toolkit that will be made publicly available in an open-access platform soon.

2.6. Ethics statement

All authors involved in data curation and analysis agreed to the HCP open and restricted access data use terms and were granted access. The study was approved by the UCL Research Ethics Committee (Project No. 4356/003).

3. Results

In this section, we present the results of the experiments on synthetic data (Section 3.1) and real data from the Human Connectome Project (Section 3.2).

3.1. Synthetic data

In this section, we applied the proposed extension of GFA to the synthetic data described in Section 2.4.1. We ran separate experiments using three different selections of synthetic data: no missing data (complete data experiment), when data was missing randomly (20% of the elements of $\mathbf{X}^{(2)}$ missing) and one group/modality was missing for some observations (20% of the rows of $\mathbf{X}^{(1)}$ missing). Fig. 2 shows the results of the extended GFA model applied to complete data. The model correctly inferred the factors, identifying two of them as shared and the other two as modality-specific. These factors were all considered most relevant based on the rvar metric (Eq. (14)) and were all correctly assigned as shared or modality-specific based on the ratio r_k (Eq. (15)). The structure of the inferred latent factors was similar to those used for generating the data (Fig. 2). The results were robust to initialisation, i.e., the model converged to similar solutions across the different initialisations. Furthermore, the irrelevant latent factors were correctly pruned out during inference. The noise parameters were also inferred correctly (i.e., the average values of τs were close to the real ones ($\tau_1 = 5\mathbf{I}_{D_1}$ and $\tau_2 = 10\mathbf{I}_{D_2}$): $\hat{\tau}^{(1)} \approx 5.08$ and $\hat{\tau}^{(2)} \approx 10.07$). Furthermore, our GFA extension showed very similar results to the vanilla GFA implementation of Klami et al. (2015) (Fig. 2c).

Fig. 3 and 4 display the results of the incomplete data experiments when data was missing randomly (20% of the elements of $\mathbf{X}^{(2)}$ missing), and one group was missing for some observations (20% of the rows of $\mathbf{X}^{(1)}$ missing), respectively. The parameters inferred using our GFA extension (middle column) were compared to those obtained using the median imputation approach (right column). The results were comparable when the amount of missing data was small (Fig. 3), i.e., both approaches were able to infer the model parameters fairly well. Even so, the model misses completely the true value of the noise parameter of $\mathbf{X}^{(2)}$ ($\hat{\tau}^{(1)} \approx 5.14$ and $\hat{\tau}^{(2)} \approx 5.22$) when the median imputation approach is used. Whereas, the noise parameters were correctly recovered ($\hat{\tau}^{(1)} \approx 5.15$ and $\hat{\tau}^{(2)} \approx 10.17$) when the proposed extension of GFA was applied. The parameters were not inferred correctly by the median imputation approach (although the noise parameters were recovered fairly well, $\hat{\tau}^{(1)} \approx 6.24$ and $\hat{\tau}^{(2)} \approx 10.20$), when the number of missing observations was considerable (Fig. 4). This was not observed when our GFA extension was applied ($\hat{\tau}^{(1)} \approx 5.04$ and $\hat{\tau}^{(2)} \approx 10.24$).

The extended GFA model predicted missing data consistently well in both incomplete data experiments. The averaged Pearson's correlation obtained between the missing and predicted values across initialisations was $\rho = 0.868 \pm 0.016$ when data was missing randomly, and $\rho = 0.680 \pm 0.039$ when one group was missing for some observations.

In the multi-output prediction task, we showed that the model could make reasonable predictions when the data was missing randomly or one modality was missing for some observations, i.e., the MSEs were similar across experiments and below chance level (Fig. 5). Moreover, there seems to be no improvement in prediction between using the proposed extension of GFA or imputing the median before training the model.

In additional experiments (presented in the Supplementary Materials and Methods), we showed that the extended GFA model outperforms the median imputation approach (in inferring the model parameters and predicting one unobserved data modality from the other), when values from the tails of the data distribution are missing (Supplementary Fig. 2a and 3). The proposed extension of GFA also outperformed the median imputation approach, when both data modalities were generated with missing values in low (Supplementary Fig. 2b) and high dimensional (Supplementary Fig. 2c) data.

3.2. HCP Data

In this section, we applied the proposed extension of GFA to the HCP data described in Section 2.4.2. We ran separate experiments using three different selections of HCP data: no missing data (complete data experiment), when data was missing randomly (20% of the elements of the non-imaging matrix missing) and when one data modality was missing for some subjects (20% of the subjects missing from the brain connectivity matrix). In the complete data experiment, the model converged to a solution comprising 75 latent factors, i.e., five factors were inactive for both data modalities (the loadings were close to zero) and were consequently pruned out. The model converged to similar solutions across different initialisations, i.e., the number of inferred latent factors was consistent across initialisations. The total percentage of variance explained by the latent factors ($\sum_{m=1}^{2}\sum_{k=1}^{75} \text{var}_k^{(m)}$) corresponded to $\sim 7.55\%$, leaving 92.45% of the variance captured by residual error. Within the variance explained, six factors were considered most relevant ($\text{rvar}_k^{(m)} > 7.5\%$), which captured $\sim 27.8\%$ of the variance explained by the total number of factors (Table 1). Based on the ratio between the

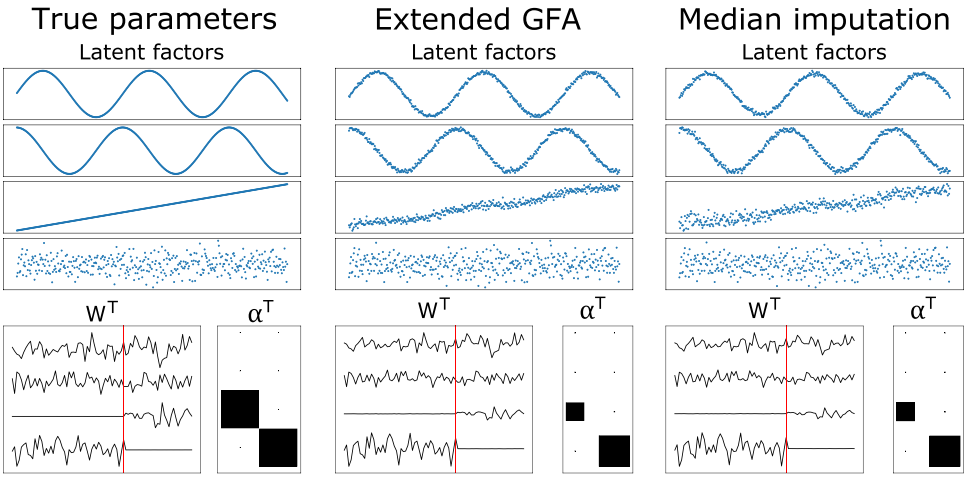

Fig. 3. True and inferred latent factors and model parameters obtained when data is missing randomly (20% of the elements of $\mathbf{X}^{(2)}$ missing). (Left column) latent factors and parameters used to generate the data. (Middle column) Latent factors and parameters inferred using the proposed extension of GFA. (Right column) Latent factors and parameters inferred using the median imputation approach. The loading matrices ($\mathbf{W}^T$) and alphas ($\boldsymbol{\alpha}^T$) can be interpreted as in Fig. 2.

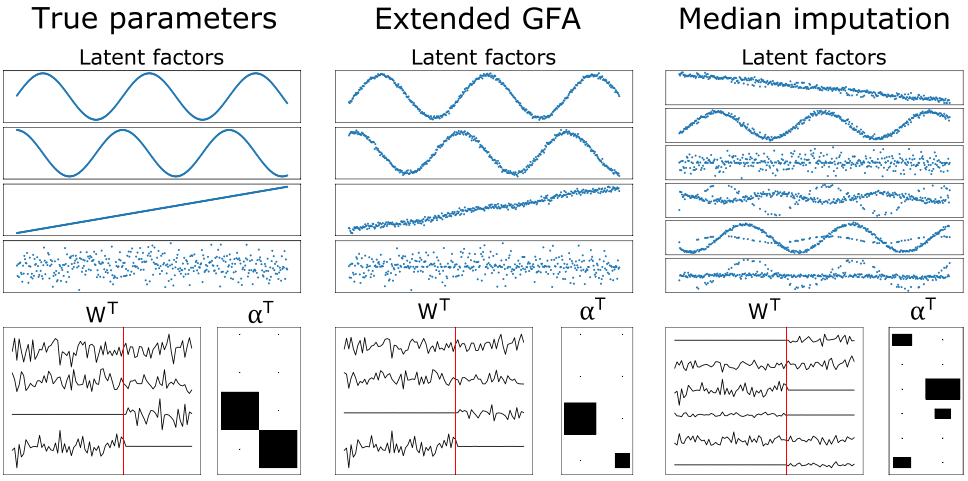

Fig. 4. True and inferred latent factors and model parameters obtained when one group was missing for some observations (20% of the rows of $\mathbf{X}^{(1)}$ were randomly removed). (Left column) latent factors and parameters used to generate the data. (Middle column) latent factors and parameters inferred using the proposed extension of GFA. (Right column) latent factors and parameters inferred using the median imputation approach (the latent factors were not ordered because the model did not converge to the right solution). The loading matrices ($\mathbf{W}^T$) and alphas ($\boldsymbol{\alpha}^T$) can be interpreted as in Fig. 2.

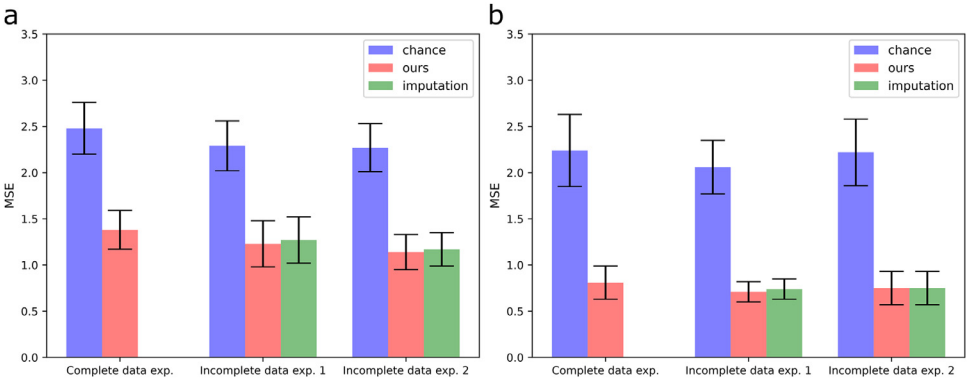

Fig. 5. Prediction errors of the multi-output prediction tasks. The bars and error bars correspond to the mean and standard deviation of the MSEs across 10 initialisations, respectively. (a) MSEs between the test observations $\mathbf{X}^{(1)^\star}$ and the mean predictions $\mathbb{E}[\mathbf{X}^{(1)^\star}|\mathbf{X}^{(2)^\star}]$ are shown for all experiments; (b) MSEs between $\mathbf{X}^{(2)^\star}$ and $\mathbb{E}[\mathbf{X}^{(2)^\star}|\mathbf{X}^{(1)^\star}]$ are shown for all experiments. ours - the proposed extension of GFA; imputation - median imputation approach; chance - chance level. Incomplete data exp. 1 - 20% of the elements of $\mathbf{X}^{(2)}$ missing; incomplete data exp. 2 -20% of the rows of $\mathbf{X}^{(1)}$ missing.

variance explained by the non-imaging and brain factors r_k (Eq. (15)), we identified four shared factors (displayed in Fig. 6) and two brain-specific factors (displayed in Fig. 7), ordered from the highest to the lowest ratio r_k (Table 1). Using the variance explained as a criterion to select the most relevant factors leads to the selection of mostly brain-specific factors due to the imbalance in the number of brain connectivity features and non-imaging measures (see Supplementary Fig. 6a-b).

In Fig. 6, we display the loadings of the shared GFA factors obtained with complete data. To aid interpretation, the loadings of the brain factors were multiplied by the sign of the population mean correlation to obtain a measure of edge strength increase or decrease (as in Smith et al. (2015)). The first factor (Fig. 6a) relates cognitive performance (loading positively), smoking and drug use (loading negatively) to the default mode and frontoparietal control networks (loading positively) and insula (loading negatively). The second shared factor (Fig. 6b) relates negative mood, long-term frequency of alcohol use (loading negatively) and short-term alcohol consumption (loading positively) to the default mode and dorsal and ventral attentional networks (loading negatively), and frontoparietal networks loading in the opposite direction. The third shared factor (Fig. 6c) is dominated by smoking behaviour (loading negatively) and, with much lower loadings, externalising in the opposite direction, which are related to the somatomotor and frontotemporal networks (loading positively). The fourth shared factor (Fig. 6d) seems to relate emotional functioning, with strong negative

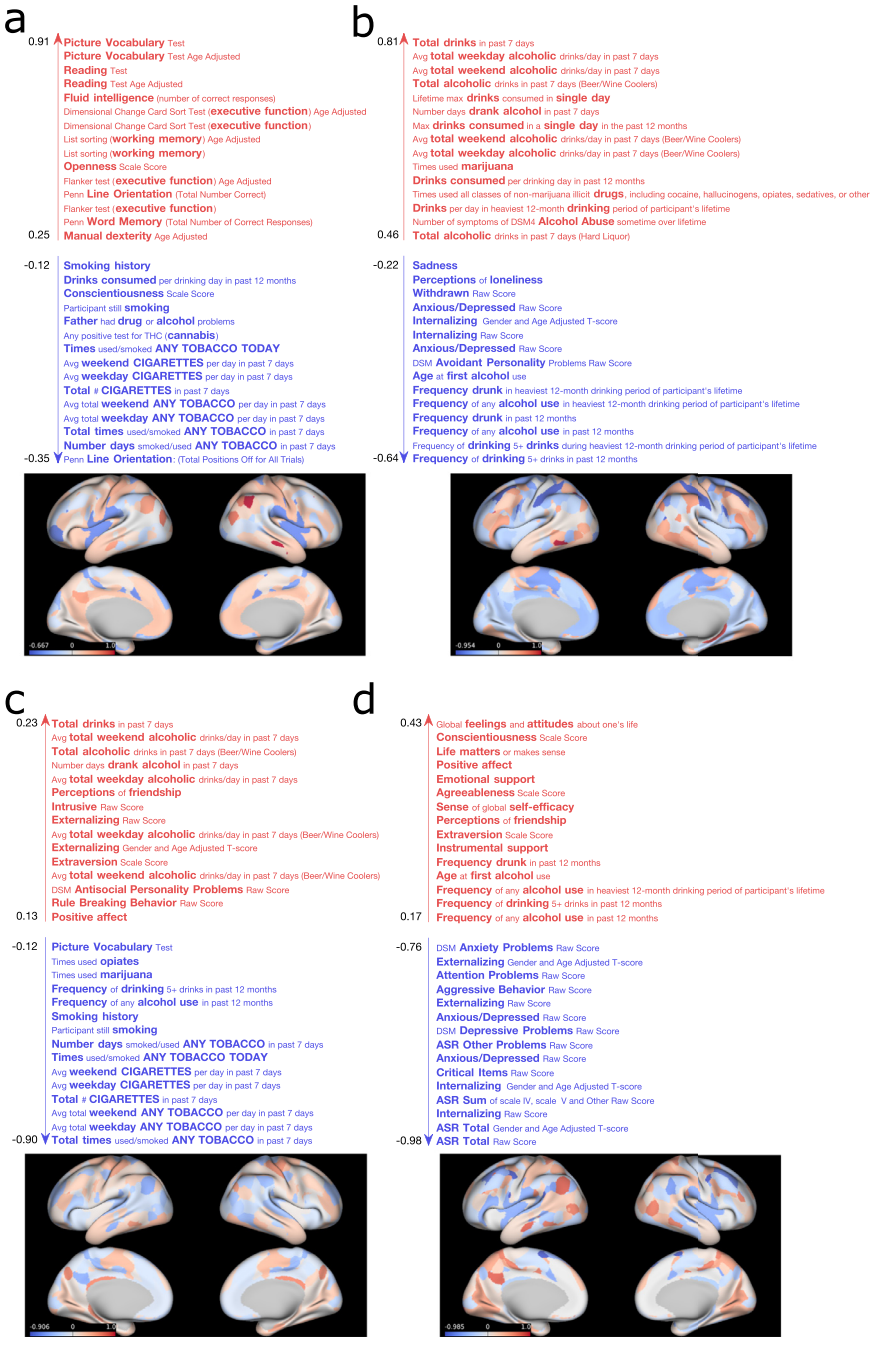

Fig. 6. Non-imaging measures and brain networks described by the first (a), second (b), third (c) and fourth (d) shared GFA factors obtained in the complete data experiment. For illustrative purposes, the top and bottom 15 nonimaging measures of each factor are shown. The brain surface plots represent maps of brain connection strength increases/decreases, which were obtained by weighting each node's parcel map with the GFA edge-strengths summed across the edges connected to the node (for details, see the Supplementary Materials and Methods). Separate thresholded maps of brain connection strength increases and decreases can be found in Supplementary Fig. 10.

loadings on a variety of psychopathological aspects (including both internalising and externalising symptoms), and positive loadings on traits such as conscientiousness and agreeableness and other aspects of well-being to cingulo-opercular network (loading negatively), and the left sided default mode network (loading positively).

Fig. 7 shows the loadings of the brain-specific factors obtained with complete data. The first factor (Fig. 7a) contains positive loadings on many areas within the frontoparietal control network, including dorsolateral prefrontal areas and inferior frontal gyrus, supramarginal gyrus, posterior inferior temporal lobe and parts of the cingulate and superior frontal gyrus. The second factor (Fig. 7b) includes positive loadings on many default mode network areas, such as medial prefrontal, posterior cingulate and lateral temporal cortices, and parts of angular and inferior frontal gyri. These factors show that there is great variability in these networks across the sample, however this variability was not linked to the non-imaging measures included in the model.

The model converged to a similar solution when data was missing randomly (20% of the elements of the non-imaging matrix were randomly removed), which included 73 factors and the total percentage of variance explained by these was ~ 7.60%. The number of most relevant factors, based on the rvar metric (Eq. (14)), was six, and they were similar to those obtained in the complete data experiment (Table 2), capturing ~ 28.2% of the variance explained by all factors (Supplementary Table 2). Four of these were considered shared factors (Supplementary Fig. 7) and two were considered brain-specific (Supplementary Fig. 9a,c). When one modality was missing for some subjects (20% of the subjects were randomly removed from the brain connectivity matrix), the model converged to a solution containing 63 factors and that explained

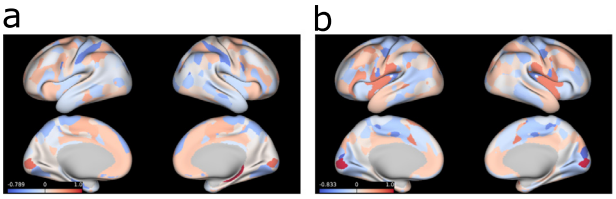

Fig. 7. Brain networks associated with the brain-specific GFA factors obtained in the complete data experiment. The brain surface plots represent maps of brain connection strength increases/decreases, which were obtained by weighting each node's parcel map with the GFA edge-strengths summed across the edges connected to the node (for details, see the Supplementary Materials and Methods).

Table 2
Similarity (measured by Pearsons correlation) between the most relevant factors obtained in the complete and the most relevant factors obtained when data was missing randomly (incomplete data exp. 1) and one modality was missing for some subjects (incomplete data exp. 2) (first and second row, respectively). The shared factors obtained with complete data are displayed in Fig. 6. and those obtained with incomplete data are shown in Supplementary Fig. 7–8. The brain-specific factors obtained with complete data are presented in Fig. 7 and those identified with incomplete data are shown in Supplementary Fig. 9.

	Shared factors				Brain factors	
	a	b	c	d	a	b
Incomplete data exp. 1	0.896	0.964	0.954	0.989	0.974	0.974
Incomplete data exp. 2	0.907	0.973	0.954	0.995	0.941	0.942

$\sim 5.21\%$ of the total variance. Although more factors were removed and a loss of variance explained was noticeable, the most relevant factors were similar to those obtained in the other experiments (Table 2, Supplementary Fig. 8 and Supplementary Fig. 9b,d), capturing $\sim 33.2\%$ of the variance explained by all factors (Supplementary Table 3).

In the multi-output prediction task, the extended GFA model predicted several non-imaging measures better than chance (Fig. 8) using complete data. The top 10 predicted variables corresponded to those with the highest loadings obtained mainly in the first shared factor (Fig. 6a) and were consistent across the incomplete data experiments (Supplementary Fig. 11). Finally, our GFA extension failed to predict the missing values in both incomplete data experiments: $\rho = 0.112 \pm 0.011$ (experiment 1, 20% of the elements of the non-imaging matrix missing); $\rho = 0.003 \pm 0.007$ (experiment 2, 20% of the subjects missing in the brain connectivity matrix).

4. Discussion

In this study, we proposed an extension of the Group Factor Analysis (GFA) model that can uncover multivariate associations among multiple data modalities, even when these modalities have missing data. We showed that our proposed GFA extension can: (1) find associations between high dimensional brain connectivity data and non-imaging measures (e.g., demographics, psychometrics, and other behavioural measures) and (2) predict non-imaging measures from brain connectivity when either data is missing at random or one modality is missing for some subjects. Moreover, we replicated previous findings obtained in a subset of the HCP dataset using CCA (Smith et al., 2015).

We showed, using synthetic data, that our GFA extension can correctly learn the underlying latent structure, i.e., it separates the shared factors from the modality-specific ones, when data is missing. In addition, it obtained very similar results to those obtained by the vanilla GFA (Klami et al., 2015) (Fig. 2). Moreover, the proposed extension of GFA inferred the model parameters better than the median imputation approach in different incomplete data scenarios. Whereas, CCA was only able to recover the shared latent factors and identified spurious latent factors when the values of the tails of the data distribution were missing (Supplementary Fig. 4). These findings underline the importance of using approaches that can handle missing data and model the modality-specific associations. Interestingly, in the multi-output prediction task, our GFA extension only outperformed the median imputation approach when the most informative values of the data (i.e., the values on the tails of the data distribution) were missing (Supplementary Fig. 3). This indicates that these values might be driving the predictions, and the model fails to predict one data modality from the other when these values are not carefully imputed. The proposed GFA extension performed worse when the percentage of missing values in the tails of the distribution increased (especially when it was greater than 32%, Supplementary Fig. 5a), whereas the performance remained constant when the percentage of random missing values increased (Supplementary Fig. 5b). Finally, our GFA extension was able to predict the missing values in different incomplete data scenarios.

In applying the proposed GFA extension to the HCP dataset, we identified 75 relevant factors. Although all factors are relevant (i.e., the highest ELBO is obtained when all factors are included in the model, see Supplementary Fig. 6c), it is challenging to interpret all of them, especially when most of them are brain-specific (Supplementary Fig. 6b). In addition, the variance explained by each factor alone is not an informative criterion to select the most relevant factors, because there is a considerable imbalance between the number of brain connectivity features and non-imaging features, and it is expected that variability within the functional brain connectivity is not necessarily related to the non-imaging measures included in this study. Therefore, if the most relevant factors were based on the variance explained by each factor, most of them would probably be considered brain-specific. As can be seen in Supplementary Figs. 6a-b, the top 14 factors that explained most variance were brain-specific. Based on the criteria proposed to overcome this issue, we obtained six most relevant factors: four describing associations between brain connectivity and non-imaging measures and two describing associations within brain connectivity. Importantly, these were consistent across the experiments with complete and incomplete data sets. Of note, only a small proportion of the variance was captured by the GFA latent structure, which may be explained by two main reasons: the brain connectivity data is noisy and/or the shared variance between the included non-imaging measures and the brain connectivity measures is relatively small with respect to the overall variance in brain connectivity.

Interestingly, most of the featured domains of non-imaging measures were not unique to particular factors, but appeared in different arrangements across the four factors. For instance, alcohol use appeared in three out of four factors: in the first, it loads in the opposite direction to cognitive performance, in the second, its frequency loads in the same direction as low mood and internalising, and in the third, its total amount loads in the same direction as externalising. For a more detailed discussion about the alcohol use loadings, see Supplementary Results. The first GFA factor was almost identical to the first CCA mode (Supplementary Fig. 12 and Supplementary Table 4), which resembled the CCA mode obtained using a subset of this data set (Smith et al., 2015). The second and third CCA modes presented similar most positive and negative non-imaging measures to the first GFA factor (for a more detailed description of the CCA modes, see the Supplementary Results). A possible explanation of the differences observed between the CCA and GFA results is that we had to apply principal component analysis to reduce the dimensionality of the data before applying CCA. This extra preprocessing step makes the CCA approach less flexible because the model cannot explore all variance in the data, whereas in GFA this does not happen because no dimensionality reduction technique is needed. For more details about the HCP experiments using CCA, see Supplementary Materials and Methods.

The brain-specific factors were difficult to interpret - as would be expected due to the inherent complexity of this data modality. Their partial similarity to known functional connectivity networks (frontopari-

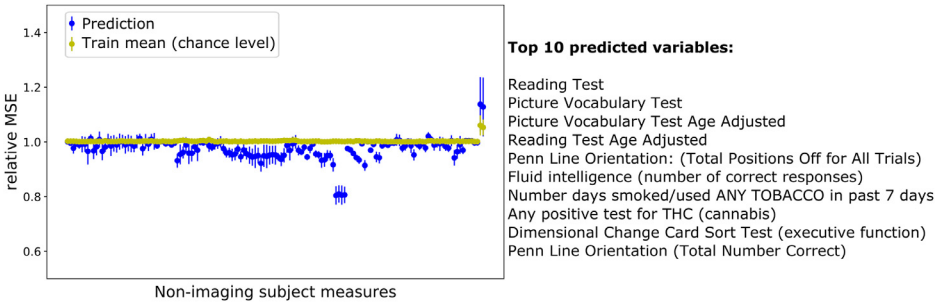

Fig. 8. Multi-output predictions of the non-imaging measures using complete data. The top 10 predicted variables are described on the right-hand side. For each non-imaging measure, the mean and standard deviation of the relative MSE (Eq. (16)) between the true and predicted values on the test set was calculated across different random initialisations of the experiments.

etal and default mode) indicates, unsurprisingly, that there are aspects of these networks that are not related to the non-imaging measures included here. Interestingly, the second brain factor (Fig. 7b) showed a few similarities ($\rho \approx 0.39$, Supplementary Table 4) with the fifth CCA mode (Supplementary Fig. 12e), which indicates that this mode could be either a spurious association or a brain-specific factor that CCA is not able to explicitly identify. This finding indicates the importance of separating the shared factors from the modality-specific ones and the use of more robust inference methods. Furthermore, the relevance of the modality-specific associations would have been more evident if we had included more than two data modalities, where associations within subsets of data modalities could be identified.

Finally, our GFA extension was able to predict a few non-imaging measures from brain connectivity in incomplete data sets. Even though the relative MSE values were modest, the model could predict several measures better than chance. Importantly, the best predicted measures corresponded to the loadings most informative in the shared factors (i.e., the highest absolute loadings), which demonstrates the potential of GFA as a predictive model.

Although the findings from both synthetic and real datasets were robust, there are still a few inherent limitations in our GFA extension. Firstly, the number of initial latent factors K needs to be chosen; however, we have shown in synthetic data that the model can still converge to a good solution even if the number of latent factors is overestimated (Supplementary Fig. 1). Secondly, although the criteria used to select the most relevant factors were validated on synthetic data, these can be further improved, e.g., by including the stability of the factors across multiple initialisations. Thirdly, our GFA extension is computationally demanding to run experiments with incomplete data sets (e.g., the CPU time was approximately 50 hours per initialisation in the HCP experiments).

Future work should investigate GFA with more data modalities, which could potentially uncover other interesting multivariate associations and improve the predictions of the non-observed data modalities and missing data. Moreover, strategies to improve the interpretability of the factor loadings (e.g., adding additional priors to impose sparsity simultaneously on the group and variable-level) could be implemented. Additionally, automatic inference methods such as Hamiltonian Monte Carlo or Automatic Differentiation Variational Inference could be implemented, as these would provide a more flexible framework, permitting new extensions of the model without the need to derive new inference equations. Finally, further extensions of the generative description of GFA could be investigated to improve its predictive accuracy.

5. Conclusions

In this study, we have shown that GFA provides an integrative and robust framework that can be used to explore associations among multiple data modalities (in benchmark datasets, such as HCP) and/or predict non-observed data modalities from the observed ones, even if data is missing in one or more data modalities. Due to its Bayesian nature, GFA provides great flexibility to be extended to more complex models to solve more complex tasks, for instance, in neuroscience.

Conflicts of interest

The authors do not have any conflicts of interest to disclose.

Credit authorship contribution statement

Fabio S. Ferreira: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. **Agoston Mihalik:** Software, Data curation, Writing – review & editing, Visualization. **Rick A. Adams:** Conceptualization, Supervision, Writing – review & editing. **John Ashburner:** Conceptualization, Writing – review & editing, Supervision. **Janaina Mourao-Miranda:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

FSF was supported by Fundação para a Ciência e a Tecnologia (Ph.D. fellowship No. SFRH/BD/120640/2016). AM, and JM-M were supported by the Wellcome Trust under Grant No. WT102845/Z/13/Z. RAA was supported by a Medical Research Council (MRC) Skills Development Fellowship (Grant No. MR/ S007806/1). The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Appendix A. Variational updates of GFA

The variational updates of the model parameters are derived by writing the log of the joint distribution $p(\mathbf{X}, \boldsymbol{\theta})$ with respect to all other variational posteriors (Eq. (10)). Considering Eq. (8), the log of the joint distribution is defined as follows:

$$\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = \ln[p(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau})p(\mathbf{Z})p(\mathbf{W}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\boldsymbol{\tau})] + \text{const}, \tag{A.1}$$

where the individual log-densities (considering the priors in Eq. (6) and Eq. (7)) are given by:

$$\begin{aligned}
\ln p(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau}) &= \sum_{m=1}^{M}\left[\frac{N}{2}\sum_{j=1}^{D_m}(\ln \tau_j^{(m)} - \ln(2\pi)) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n^{(m)} - \mathbf{W}^{(m)}\mathbf{z}_n)^T\mathbf{T}^{(m)}(\mathbf{x}_n^{(m)} - \mathbf{W}^{(m)}\mathbf{z}_n)\right],\\
\ln p(\mathbf{Z}) &= -\frac{1}{2}\sum_{n=1}^{N}\mathbf{z}_n^T\mathbf{z}_n - \frac{NK}{2}\ln(2\pi),\\
\ln p(\mathbf{W}|\boldsymbol{\alpha}) &= \sum_{m=1}^{M}\left[\frac{D_m}{2}\sum_{k=1}^{K}\ln \alpha_k^{(m)} - \frac{1}{2}\sum_{k=1}^{K}\alpha_k^{(m)}\mathbf{w}_k^{(m)T}\mathbf{w}_k^{(m)} + \frac{D_m K}{2}\ln(2\pi)\right],\\
\ln p(\boldsymbol{\alpha}) &= \sum_{m=1}^{M}\sum_{k=1}^{K}\left[a_{\boldsymbol{\alpha}^{(m)}}\ln b_{\boldsymbol{\alpha}^{(m)}} - \ln\Gamma(a_{\boldsymbol{\alpha}^{(m)}}) + (a_{\boldsymbol{\alpha}^{(m)}} - 1)\ln \alpha_k^{(m)} - b_{\boldsymbol{\alpha}^{(m)}}\alpha_k^{(m)}\right],\\
\ln p(\boldsymbol{\tau}) &= \sum_{m=1}^{M}\sum_{j=1}^{D_m}\left[a_{\boldsymbol{\tau}^{(m)}}\ln b_{\boldsymbol{\tau}^{(m)}} - \ln\Gamma(a_{\boldsymbol{\tau}^{(m)}}) + (a_{\boldsymbol{\tau}^{(m)}} - 1)\ln \tau_j^{(m)} - b_{\boldsymbol{\tau}^{(m)}}\tau_j^{(m)}\right],
\end{aligned} \tag{A.2}$$

where $\mathbf{T}^{(m)} = \text{diag}(\boldsymbol{\tau}^{(m)})$, $\mathbf{z}_n$ is the n-th column of $\mathbf{Z}$, $\mathbf{x}_n^{(m)}$ is the n-th column of $\mathbf{X}^{(m)}$, $\mathbf{w}_k^{(m)}$ is a column vector representing the k-th column of $\mathbf{W}^{(m)}$ and $a_{\boldsymbol{\alpha}^{(m)}}, b_{\boldsymbol{\alpha}^{(m)}}, a_{\boldsymbol{\tau}^{(m)}}, b_{\boldsymbol{\tau}^{(m)}}$ are the hyperparameters of the Gamma distributions in Eqs. 6–7.

A1. $q(\mathbf{Z})$ Distribution

The optimal log-density for $q(\mathbf{Z})$, given the other variational distributions is calculated using Eq. (10):

$$\begin{aligned}
\ln q(\mathbf{Z}) &= \mathbb{E}_{q(\mathbf{W}),q(\boldsymbol{\tau})}[\ln p(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau}) + \ln p(\mathbf{Z})],\\
&= \sum_{n=1}^{N}\left[-\frac{1}{2}\sum_{m=1}^{M}\langle(\mathbf{x}_n^{(m)} - \mathbf{W}^{(m)}\mathbf{z}_n)^T\mathbf{T}^{(m)}(\mathbf{x}_n^{(m)} - \mathbf{W}^{(m)}\mathbf{z}_n)\rangle - \frac{1}{2}\mathbf{z}_n^T\mathbf{z}_n\right],\\
&= \sum_{n=1}^{N}\left[\mathbf{z}_n^T\sum_{m=1}^{M}\sum_{j\in O_n^{(m)}}\langle\tau_j^{(m)}\rangle\langle\mathbf{W}_{j,*}^{(m)T}\rangle x_{j,n}^{(m)} - \frac{1}{2}\mathbf{z}_n^T\left(\sum_{m=1}^{M}\sum_{j\in O_n^{(m)}}\langle\tau_j^{(m)}\rangle\langle\mathbf{W}_{j,*}^{(m)T}\mathbf{W}_{j,*}^{(m)}\rangle\right)\mathbf{z}_n - \frac{1}{2}\mathbf{z}_n^T\mathbf{z}_n\right],\\
&= \sum_{n=1}^{N}\left[\mathbf{z}_n^T\sum_{m=1}^{M}\sum_{j\in O_n^{(m)}}\langle\tau_j^{(m)}\rangle\langle\mathbf{W}_{j,*}^{(m)T}\rangle x_{j,n}^{(m)} - \frac{1}{2}\mathbf{z}_n^T\left(\mathbf{I}_K + \sum_{m=1}^{M}\sum_{j\in O_n^{(m)}}\langle\tau_j^{(m)}\rangle\langle\mathbf{W}_{j,*}^{(m)T}\mathbf{W}_{j,*}^{(m)}\rangle\right)\mathbf{z}_n\right],
\end{aligned} \tag{A.3}$$

where $\langle\cdot\rangle = \mathbb{E}_{q(\mathbf{W}),q(\boldsymbol{\tau})}[\cdot]$ represents expectations, $\mathbf{W}_{j,*}^{(m)}$ denotes the j-th row of $\mathbf{W}^{(m)}$, $\langle\tau_j^{(m)}\rangle = \frac{\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)}}{\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)}}$ ($\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)}$ and $\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)}$ are the variational parameters obtained for $q(\boldsymbol{\tau}^{(m)})$ in Eq. (A.17)) and $\langle\mathbf{W}_{j,*}^{(m)T}\mathbf{W}_{j,*}^{(m)}\rangle = \boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}} + \boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}^T\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}$ ($\boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}}$ and $\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}$ are the variational parameters obtained for $q(\mathbf{W}^{(m)})$ in Eq. (A.11)). $O_n^{(m)}$ is the set of indices in the n-th column of $\mathbf{X}^{(m)}$ ($x_{(:,n)}^{(m)}$) that are not missing. In Eq. (A.3) omits constant terms that do not depend on $\mathbf{Z}$. Taking the exponential of the log density, the optimal $q(\mathbf{Z})$ is a multivariate normal distribution:

$$q(\mathbf{Z}) = \prod_{n=1}^{N}q(\mathbf{z}_n) = \prod_{n=1}^{N}\mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_{\mathbf{z}_n}, \boldsymbol{\Sigma}_{\mathbf{z}_n}). \tag{A.4}$$

The updates equations for $q(\mathbf{Z})$ are:

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{z}_n} &= \left[\mathbf{I}_K + \sum_{m=1}^{M}\sum_{j\in O_n^{(m)}}\langle\tau_j^{(m)}\rangle\langle\mathbf{W}_{j,*}^{(m)T}\mathbf{W}_{j,*}^{(m)}\rangle\right]^{-1},\\
\boldsymbol{\mu}_{\mathbf{z}_n} &= \boldsymbol{\Sigma}_{\mathbf{z}_n}\left(\sum_{m=1}^{M}\sum_{j\in O_n^{(m)}}\langle\tau_j^{(m)}\rangle\langle\mathbf{W}_{j,*}^{(m)T}\rangle x_{j,n}^{(m)}\right).
\end{aligned} \tag{A.5}$$

A2. $q(\mathbf{W}^{(m)})$ Distribution

The optimal log-density for $q(\mathbf{W}^{(m)})$, given the other variational distributions, is obtained by calculating:

$$\begin{aligned}
\ln q(\mathbf{W}^{(m)}) &= \mathbb{E}_{q(\mathbf{Z}),q(\boldsymbol{\alpha}^{(m)}),q(\boldsymbol{\tau}^{(m)})}[\ln p(\mathbf{X}^{(m)}|\mathbf{Z}, \mathbf{W}^{(m)}, \boldsymbol{\tau}^{(m)}) + \ln p(\mathbf{W}^{(m)}|\boldsymbol{\alpha}^{(m)})],\\
&= -\frac{1}{2}\sum_{n=1}^{N}\langle(\mathbf{x}_n - \mathbf{W}^{(m)}\mathbf{z}_n)^T\mathbf{T}^{(m)}(\mathbf{x}_n - \mathbf{W}^{(m)}\mathbf{z}_n)\rangle - \frac{1}{2}\sum_{k=1}^{K}\langle\alpha_k^{(m)}\mathbf{w}_k^{(m)T}\mathbf{w}_k^{(m)}\rangle,
\end{aligned} \tag{A.6}$$

where $\langle\cdot\rangle = \mathbb{E}_{q(\mathbf{Z}),q(\boldsymbol{\alpha}^{(m)}),q(\boldsymbol{\tau}^{(m)})}[\cdot]$. The constant term was omitted. The first term of Eq. (A.6) can be expanded as follows:

$$\begin{aligned}
&-\frac{1}{2}\sum_{n=1}^{N}\langle\left(\mathbf{x}_n-\mathbf{W}^{(m)}\mathbf{z}_n\right)^T\mathbf{T}^{(m)}\left(\mathbf{x}_n-\mathbf{W}^{(m)}\mathbf{z}_n\right)\rangle = \\
&=\sum_{j=1}^{D_m}\langle\tau_j^{(m)}\rangle\left(\sum_{n\in O_j^{(m)}}x_{j,n}^{(m)}\langle\mathbf{z}_n^T\rangle\right)\mathbf{W}_{j,*}^{(m)^T}+\sum_{j=1}^{D_m}-\frac{1}{2}\mathbf{W}_{j,*}^{(m)}\left(\langle\tau_j^{(m)}\rangle\sum_{n\in O_j^{(m)}}\langle\mathbf{z}_n\mathbf{z}_n^T\rangle\right)\mathbf{W}_{j,*}^{(m)^T},
\end{aligned} \tag{A.7}$$

where $\langle\mathbf{z}_n\mathbf{z}_n^T\rangle = \boldsymbol{\Sigma}_{\mathbf{z}_n}+\boldsymbol{\mu}_{\mathbf{z}_n}\boldsymbol{\mu}_{\mathbf{z}_n}^T$ ($\boldsymbol{\Sigma}_{\mathbf{z}_n}$ and $\boldsymbol{\mu}_{\mathbf{z}_n}$ are the variational parameters of $q(\mathbf{Z})$ in Eq. A.5) and $O_j^{(m)}$ is the set of indices in the j-th row of $\mathbf{X}^{(m)}$ ($x_{(j,:)}^{(m)}$) that are not missing. The second term of Eq. (A.6) is given by:

$$-\frac{1}{2}\sum_{k=1}^{K}\langle\alpha_k^{(m)}\mathbf{w}_k^{(m)T}\mathbf{w}_k^{(m)}\rangle=-\frac{1}{2}\sum_{j=1}^{D_m}\mathbf{W}_{j,*}^{(m)}\langle\mathbf{H}^{(m)}\rangle\mathbf{W}_{j,*}^{(m)^T}, \tag{A.8}$$

where $\langle\mathbf{H}^{(m)}\rangle = \mathrm{diag}(\langle\boldsymbol{\alpha}^{(m)}\rangle)$ and $\langle\boldsymbol{\alpha}^{(m)}\rangle = \frac{\tilde{a}_{\boldsymbol{\alpha}^{(m)}}}{\tilde{\mathbf{b}}_{\boldsymbol{\alpha}^{(m)}}}$ ($\tilde{a}_{\boldsymbol{\alpha}^{(m)}}$ and $\tilde{\mathbf{b}}_{\boldsymbol{\alpha}^{(m)}}$ are the variational parameters of $q(\boldsymbol{\alpha}^{(m)})$ in Eq. (A.14)). Putting both terms together we get:

$$\begin{aligned}
\ln q\left(\mathbf{W}^{(m)}\right) = &\sum_{j=1}^{D_m}\left[\langle\tau_j^{(m)}\rangle\left(\sum_{n\in O_j^{(m)}}x_{j,n}^{(m)}\langle\mathbf{z}_n^T\rangle\right)\mathbf{W}_{j,*}^{(m)^T}\right. \\
&\left.-\frac{1}{2}\mathbf{W}_{j,*}^{(m)}\left(\langle\mathbf{H}^{(m)}\rangle+\langle\tau_j^{(m)}\rangle\sum_{n\in O_j^{(m)}}\langle\mathbf{z}_n\mathbf{z}_n^T\rangle\right)\mathbf{W}_{j,*}^{(m)^T}\right].
\end{aligned} \tag{A.9}$$

Taking the exponential of the log density, the optimal $q(\mathbf{W}^{(m)})$ is a multivariate normal distribution:

$$q(\mathbf{W}^{(m)})=\prod_{j=1}^{D_m}q(\mathbf{W}_{j,*}^{(m)})=\prod_{j=1}^{D_m}\mathcal{N}(\mathbf{W}_{j,*}^{(m)}|\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}},\boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}}). \tag{A.10}$$

Then the updates equations for $q(\mathbf{W}^{(m)})$ are:

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}} &= \left[\langle\mathbf{H}^{(m)}\rangle+\langle\tau_j^{(m)}\rangle\sum_{n\in O_j^{(m)}}\langle\mathbf{z}_n\mathbf{z}_n^T\rangle\right]^{-1}, \\
\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}} &= \langle\tau_j^{(m)}\rangle\sum_{n\in O_j^{(m)}}\left(x_{j,n}^{(m)}\langle\mathbf{z}_n^T\rangle\right)\boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}}.
\end{aligned} \tag{A.11}$$

A3. $q\alpha^m$ distribution

The optimal log-density for $q(\boldsymbol{\alpha}^{(m)})$, given the other variational distributions is obtained by calculating:

$$\begin{aligned}
\ln q(\boldsymbol{\alpha}^{(m)}) &= \mathbb{E}_{q(\mathbf{W}^{(m)})}[\ln p(\mathbf{W}^{(m)}|\boldsymbol{\alpha}^{(m)})+\ln p(\boldsymbol{\alpha}^{(m)})], \\
&=\sum_{k=1}^{K}\left[\frac{D_m}{2}\ln\alpha_k^{(m)}-\frac{1}{2}\alpha_k^{(m)}\langle\mathbf{w}_k^{(m)^T}\mathbf{w}_k^{(m)}\rangle+(a_{\boldsymbol{\alpha}^{(m)}}-1)\ln\alpha_k^{(m)}-b_{\boldsymbol{\alpha}^{(m)}}\alpha_k^{(m)}\right], \\
&=\sum_{k=1}^{K}\left(\frac{D_m}{2}+a_{\boldsymbol{\alpha}^{(m)}}-1\right)\ln\alpha_k^{(m)}-\sum_{k=1}^{K}\left(b_{\boldsymbol{\alpha}^{(m)}}+\frac{1}{2}\langle\mathbf{w}_k^{(m)^T}\mathbf{w}_k^{(m)}\rangle\right)\alpha_k^{(m)},
\end{aligned} \tag{A.12}$$

where $\langle\cdot\rangle = \mathbb{E}_{q(\mathbf{W}^{(m)})}[\cdot]$. Constant terms that do not depend on $\boldsymbol{\alpha}$ are omitted. Taking the exponential of the log density, the optimal $q(\boldsymbol{\alpha}^{(m)})$ is a Gamma distribution:

$$q(\boldsymbol{\alpha}^{(m)})=\prod_{k=1}^{K}q(\boldsymbol{\alpha}_k^{(m)})=\prod_{k=1}^{K}\Gamma(\alpha_k^{(m)}|\tilde{a}_{\boldsymbol{\alpha}^{(m)}},\tilde{b}_{\boldsymbol{\alpha}^{(m)}}^{(k)}). \tag{A.13}$$

And the update equations for $q(\boldsymbol{\alpha}^{(m)})$ are:

$$\begin{aligned}
\tilde{a}_{\boldsymbol{\alpha}^{(m)}} &= a_{\boldsymbol{\alpha}^{(m)}}+\frac{1}{2}D_m, \\
\tilde{b}_{\boldsymbol{\alpha}^{(m)}}^{(k)} &= b_{\boldsymbol{\alpha}^{(m)}}+\frac{1}{2}\langle\mathbf{w}_k^{(m)^T}\mathbf{w}_k^{(m)}\rangle,
\end{aligned} \tag{A.14}$$

where $\langle\mathbf{w}_k^{(m)T}\mathbf{w}_k^{(m)}\rangle = \sum_{j=1}^{D_m}\left(\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}^T\boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}+\boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}}\right)_{(k,k)}$.

A4. $q\tau^m$ distribution

The optimal log-density for $q(\boldsymbol{\tau}^{(m)})$, given the other variational distributions is obtained in the following way:

$$\begin{aligned}
\ln q(\boldsymbol{\tau}^{(m)}) &= \mathbb{E}_{q(\mathbf{Z}),q(\mathbf{W}^{(m)})}[\ln p(\mathbf{X}^{(m)}|\mathbf{Z},\mathbf{W}^{(m)},\boldsymbol{\tau}^{(m)}) + \ln p(\boldsymbol{\tau}^{(m)})] \\
&= -\frac{1}{2}\sum_{j=1}^{D_m}\left[\tau_j^{(m)}\sum_{n\in O_j^{(m)}}\left(x_{j,n}^{(m)2} - 2x_{j,n}^{(m)}\langle \mathbf{W}_{j,*}^{(m)}\rangle\langle \mathbf{z}_n\rangle + \mathrm{Tr}[\langle \mathbf{W}_{j,*}^{(m)T}\mathbf{W}_{j,*}^{(m)}\rangle\langle \mathbf{z}_n\mathbf{z}_n^T\rangle]\right)\right. \\
&\quad \left. + \frac{N_j^{(m)}}{2}\ln\tau_j^{(m)} + (a_{\boldsymbol{\tau}^{(m)}}-1)\ln\tau_j^{(m)} - b_{\boldsymbol{\tau}^{(m)}}\tau_j^{(m)}\right] \\
&= \sum_{j=1}^{D_m}\left(a_{\boldsymbol{\tau}^{(m)}} + \frac{N_j^{(m)}}{2} - 1\right)\ln\tau_j^{(m)} - \sum_{j=1}^{D_m}\left(b_{\boldsymbol{\tau}^{(m)}} + \frac{1}{2}\sum_{n\in O_j^{(m)}} x_{j,n}^{(m)2} - 2x_{j,n}^{(m)}\langle \mathbf{W}_{j,*}^{(m)}\rangle\langle \mathbf{z}_n\rangle\right. \\
&\quad \left. + \mathrm{Tr}[\langle \mathbf{W}_{j,*}^{(m)T}\mathbf{W}_{j,*}^{(m)}\rangle\langle \mathbf{z}_n\mathbf{z}_n^T\rangle]\right)\tau_j^{(m)},
\end{aligned} \tag{A.15}$$

where $N_j^{(m)}$ is the number of non-missing observations in the j-th row of $\mathbf{X}^{(m)}$ and $\langle\cdot\rangle = \mathbb{E}_{q(\mathbf{Z}),q(\mathbf{W}^{(m)})}[\cdot]$. Constant terms that do not depend on τ are omitted. Taking the exponential of the log density, the optimal $q(\boldsymbol{\tau}^{(m)})$ is a Gamma distribution:

$$q(\boldsymbol{\tau}^{(m)}) = \prod_{j=1}^{D_m} q(\boldsymbol{\tau}_j^{(m)}) = \prod_{j=1}^{D_m}\Gamma(\tau_j^{(m)}|\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)}, \tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)}), \tag{A.16}$$

where the variational parameters are calculated by:

$$\begin{aligned}
\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)} &= a_{\boldsymbol{\tau}^{(m)}} + \frac{1}{2}N_j^{(m)}, \\
\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)} &= b_{\boldsymbol{\tau}^{(m)}} + \frac{1}{2}\sum_{n\in O_j^{(m)}} x_{j,n}^{(m)2} - 2x_{j,n}^{(m)}\langle \mathbf{W}_{j,*}^{(m)}\rangle\langle \mathbf{z}_n\rangle + \mathrm{Tr}[\langle \mathbf{W}_{j,*}^{(m)T}\mathbf{W}_{j,*}^{(m)}\rangle\langle \mathbf{z}_n\mathbf{z}_n^T\rangle].
\end{aligned} \tag{A.17}$$

Finally, to solve the rotation and scaling ambiguity known to be present in factor analysis models, we used a similar approach previously proposed by Virtanen and colleagues (Klami et al., 2013; Virtanen et al., 2011; 2012), which consists of maximising the variational lower bound with respect to a linear transformation $\mathbf{R}$ of the latent space, after each round of variational EM updates. This also improves convergence and speeds up the learning.

Appendix B. Evidence lower bound of GFA

Considering Eq. (9), the lower bound of $\ln p(\mathbf{X})$ is given by:

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}[\ln p(\mathbf{X},\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\tau})] - \mathbb{E}[\ln q(\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\tau})] \\
&= \mathbb{E}[\ln p(\mathbf{X}|\mathbf{Z},\mathbf{W},\boldsymbol{\tau})] + \mathbb{E}[\ln p(\mathbf{Z})] + \mathbb{E}[\ln p(\mathbf{W}|\boldsymbol{\alpha})] + \mathbb{E}[\ln p(\boldsymbol{\alpha})] + \mathbb{E}[\ln p(\boldsymbol{\tau})] \\
&\quad - \mathbb{E}[\ln q(\mathbf{Z})] + \mathbb{E}[\ln q(\mathbf{W})] + \mathbb{E}[\ln q(\boldsymbol{\alpha})] + \mathbb{E}[\ln q(\boldsymbol{\tau})],
\end{aligned} \tag{B.1}$$

where the expectations of the $\ln p(\cdot)$ terms are given by (see Eq. (A.2)):

$$\mathbb{E}_{q(\boldsymbol{\theta})}[\ln p(\mathbf{X}|\mathbf{Z},\mathbf{W},\boldsymbol{\tau})] = \sum_{m=1}^{M}\left[\sum_{j=1}^{D_m}\left(\frac{N_j^{(m)}}{2}(\langle\ln\tau_j^{(m)}\rangle - \ln(2\pi)) - \langle\tau_j^{(m)}\rangle(\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)} - b_{\boldsymbol{\tau}^{(m)}})\right)\right], \tag{B.2}$$

$$\mathbb{E}[\ln p(\mathbf{Z})] = -\frac{1}{2}\sum_{n=1}^{N}\mathrm{Tr}[\langle\mathbf{z}_n\mathbf{z}_n^T\rangle] - \frac{NK}{2}\ln(2\pi), \tag{B.3}$$

$$\mathbb{E}_{q(\boldsymbol{\alpha})}[\ln p(\mathbf{W}|\boldsymbol{\alpha})] = \sum_{m=1}^{M}\left[\frac{D_m}{2}\sum_{k=1}^{K}\langle\ln\alpha_k^{(m)}\rangle - \sum_{k=1}^{K}\mathrm{Tr}[\langle\alpha_k^{(m)}\rangle\langle\mathbf{w}_k^{(m)T}\mathbf{w}_k^{(m)}\rangle] + \frac{D_m K}{2}\ln(2\pi)\right], \tag{B.4}$$

$$\mathbb{E}[\ln p(\boldsymbol{\alpha})] = \sum_{m=1}^{M}\sum_{k=1}^{K}\left[a_{\boldsymbol{\alpha}^{(m)}}\ln b_{\boldsymbol{\alpha}^{(m)}} - \ln\Gamma(a_{\boldsymbol{\alpha}^{(m)}}) + (a_{\boldsymbol{\alpha}^{(m)}}-1)\langle\ln\alpha_k^{(m)}\rangle - b_{\boldsymbol{\alpha}^{(m)}}\langle\alpha_k^{(m)}\rangle\right], \tag{B.5}$$

$$\mathbb{E}[\ln p(\boldsymbol{\tau})] = \sum_{m=1}^{M}\sum_{j=1}^{D_m}\left[a_{\boldsymbol{\tau}^{(m)}}\ln b_{\boldsymbol{\tau}^{(m)}} - \ln\Gamma(a_{\boldsymbol{\tau}^{(m)}}) + (a_{\boldsymbol{\tau}^{(m)}}-1)\langle\ln\tau_j^{(m)}\rangle - b_{\boldsymbol{\tau}^{(m)}}\langle\tau_j^{(m)}\rangle\right], \tag{B.6}$$

where $q(\boldsymbol{\theta}) = q(\mathbf{Z})q(\mathbf{W})q(\boldsymbol{\tau})$, $\langle\ln\tau_j^{(m)}\rangle = \psi(\tilde{a}_{\boldsymbol{\tau}^{(m)}}^{(j)}) - \ln\tilde{b}_{\boldsymbol{\tau}^{(m)}}^{(j)}$, $\langle\ln\alpha_k^{(m)}\rangle = \psi(\tilde{a}_{\boldsymbol{\alpha}^{(m)}}) - \ln\tilde{b}_{\boldsymbol{\alpha}^{(m)}}^{(k)}$ and $\psi(\cdot)$ is a digamma function. $\langle\tau_j^{(m)}\rangle$, $\langle\mathbf{z}_n\mathbf{z}_n^T\rangle$, $\langle\alpha_k^{(m)}\rangle$ and $\langle\mathbf{w}_k^{(m)T}\mathbf{w}_k^{(m)}\rangle$ are calculated as in Eq. (A.3), Eq. (A.7), Eq. (A.8) and Eq. (A.14), respectively.

The terms involving expectations of the logs of the $q(\cdot)$ distributions simply represent the negative entropies of those distributions (Bishop, 2006):

$$\mathbb{E}[\ln q(\mathbf{Z})] = -\frac{1}{2}\left[\sum_{n=1}^{N} \ln|\Sigma_{\mathbf{z}_n}| + K(1+\ln(2\pi))\right], \tag{B.7}$$

$$\mathbb{E}[\ln q(\mathbf{W})] = \sum_{m=1}^{M} -\frac{1}{2}\left[\sum_{j=1}^{D_m} \ln|\Sigma_{\mathbf{W}^{(m)}_{j,*}}| + K(1+\ln(2\pi))\right], \tag{B.8}$$

$$\mathbb{E}[\ln q(\boldsymbol{\alpha})] = \sum_{m=1}^{M}\sum_{k=1}^{K}\left[\tilde{a}_{\boldsymbol{\alpha}^{(m)}} \ln \tilde{b}^{(k)}_{\boldsymbol{\alpha}^{(m)}} - \ln\Gamma(\tilde{a}_{\boldsymbol{\alpha}^{(m)}}) + (\tilde{a}_{\boldsymbol{\alpha}^{(m)}} - 1)\langle \ln \alpha_k^{(m)}\rangle - \tilde{b}^{(k)}_{\boldsymbol{\alpha}^{(m)}}\langle \alpha_k^{(m)}\rangle\right], \tag{B.9}$$

$$\mathbb{E}[\ln q(\boldsymbol{\tau})] = \sum_{m=1}^{M}\sum_{j=1}^{D_m}\left[\tilde{a}^{(j)}_{\boldsymbol{\tau}^{(m)}} \ln \tilde{b}^{(j)}_{\boldsymbol{\tau}^{(m)}} - \ln\Gamma(\tilde{a}^{(j)}_{\boldsymbol{\tau}^{(m)}}) + (\tilde{a}^{(j)}_{\boldsymbol{\tau}^{(m)}} - 1)\langle \ln \tau_j^{(m)}\rangle - \tilde{b}^{(j)}_{\boldsymbol{\tau}^{(m)}}\langle \tau_j^{(m)}\rangle\right]. \tag{B.10}$$

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neuroimage.2021.118854

References

Alnæs, D., Kaufmann, T., Marquand, A.F., Smith, S.M., Westlye, L.T., 2020. Patterns of sociocognitive stratification and perinatal risk in the child brain. Proc. Natl. Acad. Sci. U.S.A. 117 (22), 12419–12427. doi:10.1073/pnas.2001517117.

Bach, F.R., Jordan, M.I., 2006. A Probabilistic Interpretation of Canonical Correlation Analysis. Technical Report. University of California, Berkeley. https://statistics.berkeley.edu/tech-reports/688

Bijsterbosch, J.D., Woolrich, M.W., Glasser, M.F., Robinson, E.C., Beckmann, C.F., Van Essen, D.C., Harrison, S.J., Smith, S.M., 2018. The relationship between spatial configuration and functional connectivity of brain regions. Elife 7. doi:10.7554/eLife.32992.

Bishop, C., 1999. Variational principal components. In: 9th International Conference on Artificial Neural Networks: ICANN '99. IEE, pp. 509–514. doi:10.1049/cp:19991160.

Bishop, C.M., 2006. Pattern recognition and machine learning (information science and statistics). Springer-Verlag, Berlin, Heidelberg. https://dl.acm.org/doi/book/10.5555/1162264.

Bzdok, D., Meyer-Lindenberg, A., 2017. Machine learning for precision psychiatry: opportunities and challenges. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging doi:10.1016/j.bpsc.2017.11.007.

Chong Wang, 2007. Variational bayesian approach to canonical correlation analysis. IEEE Trans. Neural Networks 18 (3), 905–910. doi:10.1109/TNN.2007.891186.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., Schatzberg, A.F., Sudheimer, K., Keller, J., Mayberg, H.S., Gunning, F.M., Alexopoulos, G.S., Fox, M.D., Pascual-Leone, A., Voss, H.U., Casey, B., Dubin, M.J., Liston, C., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med. 23 (1), 28–38. doi:10.1038/nm.4246.

Golub, G.H., Zha, H., 1994. Perturbation analysis of the canonical correlations of matrix pairs. Linear Algebra Appl 210 (C), 3–28. doi:10.1016/0024-3795(94)90463-4.

Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28 (3–4), 321–377. doi:10.1093/biomet/28.3-4.321.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., Wang, P., 2010. Research domain criteria (RDoc): toward a new classification framework for research on mental disorders. American Journal of Psychiatry 167 (7), 748–751. doi:10.1176/appi.ajp.2010.09091379.

Khan, S.A., Virtanen, S., Kallioniemi, O.P., Wennerberg, K., Poso, A., Kaski, S., 2014. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. Bioinformatics 30 (17), i497–i504. doi:10.1093/bioinformatics/btu456.

Klami, A., Kaski, S., 2007. Local dependent components. In: Proceedings of the 24th international conference on Machine learning - ICML '07. ACM Press, New York, New York, USA, pp. 425–432. doi:10.1145/1273496.1273550.

Klami, A., Virtanen, S., Kaski, S., 2013. Bayesian canonical correlation analysis. Journal of Machine Learning Research 14, 965–1003. doi:10.5555/2567709.2502612.

Klami, A., Virtanen, S., Leppäaho, E., Kaski, S., 2015. Group factor analysis. IEEE Trans Neural Netw Learn Syst 26 (9), 2136–2147. doi:10.1109/TNNLS.2014.2376974.

Lê Cao, K.A., Martin, P.G., Robert-Granié, C., Besse, P., 2009. Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinformatics 10 (10). doi:10.1186/1471-2105-10-34.

Lee, W.H., Moser, D.A., Ing, A., Doucet, G.E., Frangou, S., 2019. Behavioral and health correlates of resting-State metastability in the human connectome project. Brain Topogr 32 (1), 80–86. doi:10.1007/s10548-018-0672-5.

Li, J., Bolt, T., Bzdok, D., Nomi, J.S., Yeo, B.T., Spreng, R.N., Uddin, L.Q., 2019. Topography and behavioral relevance of the global signal in the human brain. Sci Rep 9 (1), 1–10. doi:10.1038/s41598-019-50750-8.

Luttinen, J., Ilin, A., 2010. Transformations in variational bayesian factor analysis to speed up learning. Neurocomputing 73 (7–9), 1093–1102. doi:10.1016/j.neucom.2009.11.018.

Mackay, D.J.C., 1995. Probable networks and plausible predictions a review of practical bayesian methods for supervised neural networks. Network: Computation in Neural Systems 6 (3), 469–505. doi:10.1088/0954-898X_6_3_011.

Mihalik, A., Ferreira, F.S., Moutoussis, M., Ziegler, G., Adams, R.A., Rosa, M.J., Prabhu, G., de Oliveira, L., Pereira, M., Bullmore, E.T., Fonagy, P., Goodyer, I.M., Jones, P.B., Hauser, T., Neufeld, S., Romero-Garcia, R., St Clair, M., Vértes, P.E., Whitaker, K., Inkster, B., Ooi, C., Toseeb, U., Widmer, B., Bhatti, J., Villis, L., Alrumaithi, A., Birt, S., Bowler, A., Cleridou, K., Dadabhoy, H., Davies, E., Firkins, A., Granville, S., Harding, E., Hopkins, A., Isaacs, D., King, J., Kokorikou, D., Maurice, C., McIntosh, C., Memarzia, J., Mills, H., O'Donnell, C., Pantaleone, S., Scott, J., Fearon, P., Suckling, J., van Harmelen, A.L., Kievit, R., Shawe-Taylor, J., Dolan, R., Mourão-Miranda, J., 2020. Multiple holdouts with stability: improving the generalizability of machine learning analyses of brainbehavior relationships. Biol. Psychiatry 87 (4), 368–376. doi:10.1016/j.biopsych.2019.12.001.

Mihalik, A., Ferreira, F.S., Rosa, M.J., Moutoussis, M., Ziegler, G., Monteiro, J.M., Portugal, L., Adams, R.A., Romero-Garcia, R., Vértes, P.E., Kitzbichler, M.G., Váša, F., Vaghi, M.M., Bullmore, E.T., Fonagy, P., Goodyer, I.M., Jones, P.B., Dolan, R., Mourão-Miranda, J., 2019. Brain-behaviour modes of covariation in healthy and clinically depressed young people. Sci Rep 9 (1), 11536. doi:10.1038/s41598-019-47277-3.

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 (271), 182–194. doi:10.1016/j.jneumeth.2016.06.011.

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 (11), 1565–1567. doi:10.1038/nn.4125.

Sui, J., Adali, T., Yu, Q., Chen, J., Calhoun, V. D., 2012. A review of multivariate methods for multimodal fusion of brain imaging data. 10.1016/j.jneumeth.2011.10.031

Suvitaival, T., Parkkinen, J.A., Virtanen, S., Kaski, S., 2014. Cross-organism toxicogenomics with group factor analysis. Systems Biomedicine 2 (4), 71–80. doi:10.4161/sysb.29291.

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 (6), 1–33. doi:10.1145/3136624.

Virtanen, S., Klami, A., Kaski, S., 2011. Bayesian CCA via Group Sparsity. In: Proceedings of the 28th International Conference on Machine Learning, pp. 457–464. https://dl.acm.org/doi/10.5555/3104482.3104540

Virtanen, S., Klami, A., Khan, S., Kaski, S., 2012. Bayesian Group Factor Analysis. In: Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1269–1277. http://proceedings.mlr.press/v22/virtanen12.html

Waaijenborg, S., Verselewel de Witt Hamer, P.C., Zwinderman, A.H., 2008. Quantifying the association between gene expressions and DNA-Markers by penalized canonical correlation analysis. Stat Appl Genet Mol Biol 7 (1). doi:10.2202/1544-6115.1329.

Wegelin, J.A., 2000. A survey of partial least squares (PLS) methods, with emphasis on the two-block case. Technical Report. University of Washington.

Winkler, A.M., Renaud, O., Smith, S.M., Nichols, T.E., 2020. Permutation inference for canonical correlation analysis. Neuroimage 220 (April), 117065. doi:10.1016/j.neuroimage.2020.117065.

Witten, D.M., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10 (3), 515–534. doi: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 (1), 3003. doi:10.1038/s41467-018-05317-y.

Zhao, S., Gao, C., Mukherjee, S., Engelhardt Zhao, B.E., 2016. Bayesian group factor analysis with structured sparsity. Technical Report. https://dl.acm.org/doi/10.5555/2946645.3053478