Multi-view machine learning methods to uncover brain-behaviour associations

Fábio S. Ferreira

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science
University College London

4th December 2021

I, Fábio S. Ferreira, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

The heterogeneity of neurological and mental disorders has been a key confound in disease understanding and treatment outcome prediction, as the study of patient populations typically includes multiple subgroups that do not align with the diagnostic categories. The aim of this thesis is to investigate and extend classical multivariate methods, such as Canonical Correlation Analysis (CCA), and latent variable models, e.g., Group Factor Analysis (GFA), to uncover associations between brain and behaviour that may characterize patient populations and subgroups of patients.

In the first contribution of this thesis, we applied CCA to investigate brain-behaviour associations in a sample of healthy and depressed adolescents and young adults. We found two positive-negative brain-behaviour modes of covariation, capturing externalisation/ internalisation symptoms and well-being/distress. In the second contribution of the thesis, I applied sparse CCA to the same dataset to present a regularised approach to investigate brain-behaviour associations in high dimensional datasets. Here, I compared two approaches to optimise the regularisation parameters of sparse CCA and showed that the choice of the optimisation strategy might have an impact on the results. In the third contribution, I extended the GFA model to mitigate some limitations of CCA, such as handling missing data. I applied the extended GFA model to investigate links between high dimensional brain imaging and non-imaging data from the Human Connectome Project, and predict non-imaging measures from brain functional connectivity. The results were consistent between complete and incomplete data, and replicated previously reported findings. In the final contribution of this thesis, I proposed two extensions of GFA to uncover brain behaviour associations that characterize subgroups of subjects in an unsupervised and supervised way, as well as explore within-group variability at the individual level. These extensions were demonstrated using a dataset of patients with genetic frontotemporal dementia.

In summary, this thesis presents multi-view methods that can be used to deepen our understanding about the latent dimensions of disease in mental/neurological disorders and potentially enable patient stratification.

Impact Statement

This thesis explores different implementations and applications of multi-view machine learning approaches to uncover associations among multiple data modalities, namely brain imaging data and behavioural data. The work presented in this thesis shows how these methods can shine light on the latent dimensions of abnormal states of the healthy or diseased brain and be used to improve patient stratification. The methods explored and proposed in this thesis can be easily applied to other neuroimaging applications, as well as other fields of research.

The work presented in Chapters 3 and 4 shows the potential of classical multi-view methods to deepen our understanding of the underlying dimensions of depression in adolescence and young adulthood. The method proposed in Chapter 5 is a more robust alternative to the classical methods used to uncover associations among multiple data modalities in high dimensional incomplete data sets. The work presented in Chapter 6 provides an extension of these methods to improve model interpretability, characterize subgroups of a population and explore the heterogeneity of the subgroups.

The studies presented in this thesis have been disseminated through journal (Chapters 3, 5) and conference (Chapter 4) publications, as well as international conference presentations (Chapter 4 and 5). Lastly, the work presented in Chapter 6 is an important proof of concept that might have an impact beyond academia by providing a set of models that are interpretable, able to explore variability within subgroups of patients and contribute to personalised medicine by computing individual patient outcomes.

Acknowledgements

There are many great people who have contributed for this PhD project to become a reality. I would like to start by thanking my supervisors, Janaina Mourao-Miranda and John Ashburner, for giving me the opportunity to do a PhD at UCL and for their support, advice and guidance throughout the last four years.

I am grateful for visiting the probabilistic machine learning group at Aalto University, co-led by Samuel Kaski and Aki Vehtari, and thankful to Sami for his advice and guidance in the last part of my PhD. My thanks to Aki as well for the insightful discussions regarding the regularised horseshoe prior. I would also like to thank John Shawe-Taylor for the feedback and advice in the beginning of my PhD, as well as my collaborators Michael Moutoussis, Gabriel Ziegler, Jonathan Rohrer and Arabella Bouzigues, for their valuable input and discussions.

I would also like to thank my lab mates for the nice chats and discussions: Rick, Maria, Anil, James, Najiba, Richard and Konstantinos. Special thanks to Agoston, Cemre, João and Jessica not only for helping me throughout the various stages of my PhD, but also for making this long and arduous journey more pleasant and for making me believe that it is possible to find great people (who I can call friends) far away from home. Thank you for all the chats, board game nights, pub drinks, picnics and other social events.

I am also very thankful to my CMIC mates Maura, Giuseppe, Pere, Marco, Rica, Kiko, Chris, Burcu and Michele for all the chats, pub events and table football/tennis matches. Special thanks to Ruaridh, Nooshin and Maria for your support, friendship and for making my experience in London much more enjoyable! Thanks a lot for the dancing/board games nights and football matches (at the pub or the pitch).

I would like to acknowledge the NeuroScience in Psychiatry Network consortium, Human Connectome Project and the Genetic Frontotemporal demen-

tia initiatives for collecting, organising and providing the data used in this thesis. I would also like to thank FCT (*Fundação para a Ciência e a Tecnologia*) for funding my PhD project (SFRH/BD/120640/2016).

Agora na língua de Camões e Fernando Pessoa que tinham muito mais jeito para escrever em Português do que eu. Quero começar por agradecer ao gangue de Holloway, João, Liane e Marciano por me fazerem sentir mais "em casa" numa cidade como Londres, por todos os momentos de verdadeira tradição académica portuguesa e também por me aturarem quando eu precisava de desabafar e descarregar todas as minhas frustações. Queria não deixar de agradecer à minha malta do Seixo (Diogo, André, Flávio, Hugo, Marcelo, Júlio, Nelson e João) por não me deixarem esquecer as minhas origens, por todas as mensagens diárias que me ajudam a esquecer a distância e, como é óbvio, por todas as conversas da bola. Um agradecimento especial à Daniela por tudo o que já passamos durante estes quase 25 anos (sim, estamos velhos). Pouco há a dizer, apenas um grande obrigado por estares sempre disponível quando um ombro amigo é preciso. Um especial e sentido FRA aos meus mates de Coimbra, Gonçalo, Adriana, Mafalda, Fernando e Levita (eu não me esqueço de ti mesmo que nunca atendas as chamadas) por continuarem bem "perto" e fazerem da distância uma coisa insignificante. Tem sido um prazer fazer esta viagem convosco, que assim continue por muito tempo. Obrigado meus putos!

Queria agradecer ao meu ponto de abrigo nesta longa e árdua aventura: à Ana por todo o apoio, confidência e paciência que, mesmo durante algum tempo à distância, foram essenciais para enfrentar as dificuldades do dia a dia. Obrigado minha pequena por todos os momentos passados nestes quase quatro anos!

Por fim, queria deixar um agradecimento especial à minha família, em particular àqueles sem os quais esta experiência não teria sido possível: Mãe, Pai e Mana. Eu sei que não há necessidade de dizer muito, mas queria agradecer por tudo o que me dão todos os dias, por todos os esforços que tiveram de fazer para que eu aqui chegasse e espero que se sintam tão orgulhosos como eu me sinto abençoado por ter nascido e crescido convosco, adoro-vos!

List of Publications

Journal papers

- **F. S. Ferreira**, A. Bouzigues, J. Ashburner, J. D. Rohrer[1], S. Kaski[1], J. Mourao-Miranda[1]. Uncovering brain-behaviour associations in subgroups of patients in genetic FTD using GFA (in preparation).

- A. Mihalik, J. Chapman, R. A. Adams, **F. S. Ferreira**, J. Shawe-Taylor, J. Mourao-Miranda. Tutorial on Canonical Correlation Analysis and Partial Least Squares for identifying brain-behaviour associations. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (submitted).

- **F. S. Ferreira**, A. Mihalik, R. A. Adams, J. Ashburner[1], J. Mourao-Miranda[1]. A hierarchical Bayesian model to find brain-behaviour associations in incomplete data sets. *NeuroImage* (under revision). Available on: https://arxiv.org/abs/2103.06845.

- A. Mihalik, **F. S. Ferreira**, M. Moutoussis, G. Ziegler, R. A. Adams, M. J. Rosa, G. Prabhu, L. de Oliveira, M. Pereira, E. T. Bullmore, P. Fonagy, I. M. Goodyer, P. B. Jones, NSPN Consortium, J. Shawe-Taylor, R. Dolan, J. Mourao-Miranda. Multiple holdouts with stability: improving the generalizability of machine learning analyses of brain-behavior relationships. *Biological Psychiatry* (2020). Available on: https://doi.org/10.1016/j.biopsych.2019.12.001.

- A. Mihalik[2], **F. S. Ferreira**[2], M. J. Rosa, M. Moutoussis, G. Ziegler, J. M. Monteiro, L. Portugal, R. A. Adams, R. Romero-Garcia, P. E. Vertes, M. G. Kitzbichler, F. Vasa, M. M. Vaghi, E. T. Bullmore, P. Fonagy, I. M. Goodyer, P. B. Jones, NSPN Consortium, R. Dolan, J.

[1]Joint senior authors
[2]Joint first authors

Mourao-Miranda. Brain-behaviour modes of covariation in healthy and clinically depressed young people. *Scientific Reports* (2019). Available on: https://www.nature.com/articles/s41598-019-47277-3.

Conference papers

- N.P. Oxtoby[2], **F. S. Ferreira**[2], A. Mihalik, T. Wu, M. Brudfors, H. Lin, A. Rau, S. B. Blumberg, M. Robu, C. Zor, M. Tariq, M.D.M.E Garcia, B. Kanber, D. I. Nikitichev, J. Mourao-Miranda. ABCD Neurocognitive Prediction Challenge 2019: Predicting Individual Residual Fluid Intelligence Scores from Cortical Grey Matter Morphology. In: Pohl K., Thompson W., Adeli E., Linguraru M. (eds) Adolescent Brain Cognitive Development Neurocognitive Prediction. *Lecture Notes in Computer Science* (2019). Available on: https://doi.org/10.1007/978-3-030-31901-4_14.

- A. Mihalik[2], M. Brudfors[2], M. Robu, **F. S. Ferreira**, H. Lin, A. Rau, T. Wu, S. B. Blumberg, B. Kanber, M. Tariq, M.D.M.E. Garcia, C. Zor, D. I. Nikitichev, J. Mourao-Miranda[1], N. P. Oxtoby[1]. ABCD Neurocognitive Prediction Challenge 2019: Predicting Individual Fluid Intelligence Scores from Structural MRI Using Probabilistic Segmentation and Kernel Ridge Regression. In: Pohl K., Thompson W., Adeli E., Linguraru M. (eds) Adolescent Brain Cognitive Development Neurocognitive Prediction. *Lecture Notes in Computer Science* (2019). Available on: https://doi.org/10.1007/978-3-030-31901-4_16.

- **F. S. Ferreira**, M. J. Rosa, M. Moutoussis, R. Dolan, J. Shawe-Taylor, J. Ashburner, J. Mourao-Miranda. Sparse PLS hyperparameter optimisation for investigating brain-behaviour relationships. *International Workshop on Pattern Recognition in Neuroimaging* (2018). Available on: https://discovery.ucl.ac.uk/id/eprint/10058961/.

Software

- Group Factor Analysis - Python implementation available on: https://github.com/ferreirafabio80/gfa (Main contributor).

- Supervised GFA - Python implementation available on: https://github.com/ferreirafabio80/svgfa (Main contributor).

- PLS/CCA Toolkit - MATLAB implementation available on: https://github.com/anaston/PLS_CCA_framework (One of the contributors).

Acronyms

ABCD Adolescent Brain Cognitive Development
ARD Automatic Relevance Determination
CCA Canonical Correlation Analysis
FTD Frontotemporal Dementia
GFA Group Factor Analysis
GENFI Genetic Frontotemporal dementia Initiative
HMC Hamiltonian Monte Carlo
HCP Human Connectome Project
ICA Independent Component Analysis
MCMC Markov Chain Monte Carlo
MRI Magnetic Resonance Imaging
MSE Mean Squared Error
NSPN NeuroScience in Psychiatry Network
NUTS No-U-Turn Sampler
PCA Principal Component Analysis
PLS Partial Least Squares

Mathematical Notation

Greek Symbols

$\boldsymbol{\mu}^{(m)}$	Mean parameters of view m (Chapter 2)
$\boldsymbol{\alpha}^{(m)}$	ARD parameters of view m (Chapters 2 and 5)
τ	Noise parameter (Chapters 2 and 5), global shrinkage parameter of the horseshoe priors (Chapter 6)
$\boldsymbol{\Lambda}^{(m)}$	Local parameters of the horseshoe priors for the m-th view (Chapter 6)
$\sigma^{(m)}$	Noise precision of view m (Chapter 6)
$\Psi_{s,n}$	Probability of sample n belonging to class s (Chapter 6)

Matrices and vectors

$\mathbf{X}^{(m)}$	Input data matrix of view m
$\mathbf{x}_n$	Column vector of $\mathbf{X}$
$\mathbf{u}_k$, $\mathbf{v}_k$	Weight vectors of the k-th CCA/PLS mode
$\mathbf{Z}$	Latent variables' matrix
$\mathbf{z}_n$	Column vector of $\mathbf{Z}$
$\mathbf{I}$	Identity matrix
$\boldsymbol{\Phi}^{(m)}$, $\mathbf{T}^{(m)^{-1}}$	Noise covariance matrix of view m
$\mathbf{A}^{(m)}$, $\mathbf{B}^{(m)}$, $\mathbf{W}^{(m)}$	Loading matrices
$\mathbf{W}^{(m)}_{j,*}$	Row vector of $\mathbf{W}^{(m)}$
$\mathbf{w}^{(m)}_{k}$	Column vector of $\mathbf{W}^{(m)}$
$\boldsymbol{\lambda}^{(z)}_{n}$	Column vector of $\boldsymbol{\Lambda}^{(z)}$
$\mathbf{Y}$	Observed labels
$\mathbf{A}$	Data matrix with class probabilities (Chapter 6)

Distributions

$\mathcal{N}(\cdot)$	Normal distribution
$\mathcal{W}^{-1}(\cdot)$	Inverse Wishart distribution
$\Gamma(\cdot)$	Gamma distribution
$C^{+}(\cdot)$	Half-Cauchy distribution

Other variables

N	Number of samples
M	Number of views/data modalities
S	Number of classes/subtypes
D_m	Number of features of view m
K	Number of latent components or CCA/PLS modes
c_u, c_v	Regularisation parameters of sparse CCA

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning methods have been applied in several fields for automatic detection of patterns in data, which can be defined as any relations, regularities or structure. These patterns can then be used, for instance, to classify unseen observations into different categories (Bishop, 2006). Since the beginning of this century, these methods have been extensively applied to brain imaging data, e.g., structural and functional Magnetic Resonance Imaging (MRI), to distinguish two groups of subjects (e.g., healthy controls from patients) or cognitive states (Mourão-Miranda et al., 2005; Ecker et al., 2010; Nouretdinov et al., 2011; Orrù et al., 2012; Mateos-Pérez et al., 2018). Although these classification approaches have demonstrated the biomarker potential of neuroimaging in psychiatry and neurology, they rely on the quality of the diagnostic categories. This might be a hindrance because these populations are usually heterogeneous, and, in the case of mental disorders, their diagnosis is based on signs and symptoms rather than objective biomarkers of illness. Indeed, it has been shown that diagnostic categories in psychiatry do not align with findings emerging from clinical neuroscience and genetics, and fail to predict treatment response (Insel et al., 2010; Bzdok and Meyer-Lindenberg, 2018).

The lack of understanding of the underlying dimensions of disease in such studies opened a window of opportunity for exploratory multivariate methods, such as Canonical Correlation Analysis (CCA) (Hotelling, 1936) and Partial Least Squares (PLS) (Wegelin, 2000). CCA and PLS have been used to uncover multivariate associations between multiple sets of data (also termed as different *views*), e.g., different data modalities, without relying on diagnostic categories. These approaches are particularly relevant for brain imaging research, where different types of data (e.g., structural and functional MRI, behavioural and/or cognitive assessments) are collected from the same individuals to have a better

understanding of the brain diseases and cognitive processes.

In high dimensional data (e.g., neuroimaging datasets), i.e., when the number of samples is much smaller than the number of features, solving the CCA optimisation problem is not possible (Uurtio et al., 2017). To address this issue, dimensionality reduction methods, such as Principal Component Analysis (PCA), can be used to reduce the number of features before applying CCA (Smith et al., 2015) or alternatively regularisation methods, such as sparse CCA (Witten et al., 2009; Parkhomenko et al., 2009; Waaijenborg et al., 2008; Lê Cao et al., 2008) can be applied. However, both methods contain hyperparameters (i.e., number of principal components and regularisation parameters) that must be carefully tuned to optimise the trade-off between the variance explained and model overfitting. In this thesis, I compare two different strategies to optimise the number of principal components and the regularisation parameters of sparse CCA.

Although these methods have been successfully applied to different scenarios in neuroimaging to uncover associations between brain connectivity/structure, demographic and behavioural features (Smith et al., 2015; Monteiro et al., 2016; Drysdale et al., 2017; Xia et al., 2018; Mihalik et al., 2019), they 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 on the whole data set (Smith et al., 2015; Winkler et al., 2020) or on hold-out sets of the data (Monteiro et al., 2016; Mihalik et al., 2020). Second, the variability within data modalities, which might explain important variance in the data, are not modelled. Finally, they assume data pairing between views, which is problematic when values are missing in one or both views. This is a common issue in clinical and neuroimaging datasets and the missing values usually need to be imputed, or the samples removed before applying the models.

One potential way to address the limitations mentioned above is to consider CCA as a latent variable model (Bach and Jordan, 2006), in which the data is assumed to be generated by the same latent variables. Bach and Jordan (2006) used a maximum likelihood approach to estimate the model's parameters and showed that the latent space found by probabilistic CCA is equivalent to the subspace that standard CCA finds, and therefore the limitations mentioned above are not solved using probabilistic CCA alone. Nevertheless, probabilistic CCA could be used as a building block for more complex mod-

els, such as Bayesian CCA (Klami and Kaski, 2007; Chong Wang, 2007), to assess the uncertainty of the model parameters and impose regularisation, by adding appropriate priors over the model parameters, in order to remove latent components that explain little variance. Bayesian CCA has some limitations, however: it is not able to uncover associations within data modalities and, in high dimensional data, it can be computationally infeasible (Klami et al., 2013). Virtanen et al. (2011) and Klami et al. (2013) proposed an extension of Bayesian CCA to overcome these two limitations. This model was further extended to include more than two data modalities (termed *groups*) and was named Group Factor Analysis (GFA) (Virtanen et al., 2012; Klami et al., 2015). GFA does not address the third limitation mentioned above, i.e., it cannot be applied to data modalities with missing data. In this thesis, I extend GFA to handle missing data. This extension required re-writing the variational update rules and allow more flexible assumptions about noise.

Bunte et al. (2016) proposed a sparse extension of GFA to impose feature and sample-wise sparsity by adding shrinkage priors (i.e., spike-and-slab priors) over the loading matrices and latent variables. This model finds subsets of samples sharing associations across multiple data modalities, which is particularly useful in clinical applications where the populations are usually heterogeneous. In this thesis, I propose a new sparse GFA model by replacing the spike-and-slab priors with regularised horseshoe priors (Piironen and Vehtari, 2017), which allow a more efficient inference using automatic methods, such as Hamiltonian Monte Carlo (Hoffman and Gelman, 2014). In addition, I propose supervised GFA by including a discriminative module to the sparse GFA model to identify brain-behaviour associations that characterise pre-defined/underlying subtypes. Both models explore heterogeneity within these subtypes and provide information about how these associations are expressed at the individual level.

1.1 Thesis outline and contributions

The aim of this thesis is to apply and propose new extensions of machine learning methods that can extract and combine information from different data modalities (e.g., brain structural/functional MRI and cognitive/behavioural assessments) to uncover multivariate brain-behaviour associations that might provide a better understanding of the underlying dimensions of disease and characterisation of subgroups of patients to improve patient stratification. The thesis is structured as follows:

- Chapter 2 gives an overview of the machine learning concepts relevant for this thesis, which is followed by the descriptions of the background theory on the classical multi-view methods and latent variable models used in this thesis. The chapter also includes an overview of the data modalities used in this thesis.

- Chapter 3 presents an application of CCA coupled with PCA to a clinical dataset of healthy and depressed young subjects and shows how these methods can shed some light on the underlying dimensions of abnormal behaviour by investigating multivariate associations between individual patterns of functional brain connectivity and individual sets of psychometric/IQ/demographic data. In the same study, two different strategies to optimise the number of principal components were compared. Publication associated with this chapter: Mihalik et al. (2019).

- Chapter 4 describes a comparison between two frameworks to optimise the sparse CCA hyperparameters using the dataset mentioned above. It also describes how these different strategies might affect the brain-behaviour associations identified. Publication associated with this chapter: Ferreira et al. (2018).

- Chapter 5 presents a novel extension of GFA to uncover multivariate associations among multiple data modalities with missing data. The GFA extension was applied to synthetic data and data from the Human Connectome Project (HCP) (Van Essen et al., 2013) to uncover associations between high dimensional brain functional connectivity and non-imaging features (e.g., demographics, psychometrics and other behavioural features). Publication associated with this chapter: Ferreira et al. (2021).

- Chapter 6 describes novel sparse and supervised GFA models, which were applied to the Genetic Frontotemporal dementia Initiative (GENFI) dataset. Sparse and supervised GFA can find sparse associations that may describe pre-defined subtypes/subgroups of patients, explore within-subtype variability and provide information about how these associations are expressed at the individual level.

- Chapter 7 summarises the work presented in this thesis, and proposes directions for future work.

Chapter 2

Background

This chapter provides the background theory relevant for the work presented in this thesis. Section 2.1 introduces the machine learning framework, distinct types of learning, as well as techniques for model selection. Sections 2.2 and 2.3 present the classical multi-view methods and latent variable models, respectively, used for the approaches presented in this thesis. Finally, an overview of the brain imaging and non-imaging data used in this thesis is provided in Section 2.4.

2.1 Machine Learning overview

Machine learning is a subfield of Artificial Intelligence that provides statistical methods to automatically infer hidden patterns or associations in data that can be used to predict unseen data (Bishop, 2006). Machine learning approaches are divided into two main phases: a *learning* or *training* phase, where a large set of data called a *training set* is used to learn the parameters of an adaptive model; a *testing* phase, where the generalisability/predictive performance of the trained model is assessed on a new data set, the *test set* (Bishop, 2006). The observations/samples in both sets are assumed to be generated from the same distribution. Finally, a *decision* step can also be considered, i.e., in practice people often need to make decisions based on the model's predictions. For instance, a doctor needs to decide whether to give a treatment to a patient or not, given the class that the patient was assigned (or more likely to belong) to. This area of research is often known as decision analysis and the goal is to choose the best action from a set of candidate actions, e.g., choosing the action that maximises the expected utility (Bishop, 2006; Gelman et al., 2013).

Machine learning is usually divided into three broad, distinct types of learning. In the *supervised learning* approach, the goal is to learn a func-

tion/the model's parameters, during the training phase, that maps the inputs $\mathbf{X} \in \mathbb{R}^{D \times N}$ (D represents the number of features and N is the number of training observations) to the outputs or response variables $\mathbf{y} \in \mathbb{R}^{N \times 1}$ of the training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ (where $\mathbf{x}_i$, y_i represent a labelled input-output pair). When the outputs are categorical, the problem is known as *classification* (e.g., object classification) and when the outputs are continuous variables, the problem is known as *regression* (e.g., age prediction). The performance of the supervised models are often assessed by comparing the predictions of the outputs y_i with observed values in the test set (e.g., using accuracy or mean squared error). Examples of supervised learning models are logistic regression, support vector machines, decision trees or neural networks.

The second main type of learning is the *unsupervised learning* approach, where only the inputs ($\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$) are available, and the goal is to find "relevant" patterns or associations in data. The problem is less well-defined, since the models are less constrained with respect to what kind of patterns to look for and often there is no obvious performance metric to use. Example of unsupervised learning models are clustering (e.g., k-means or Gaussian mixture models), dimensionality reduction (e.g., Principal Component Analysis) or latent variable models (such as, factor analysis or variational autoencoders).

The third type is known as *reinforcement learning* (RL). The goal in RL is to take actions in an environment in order to maximise the cumulative reward. The RL agent (i.e., machine learning model) learns by receiving positive or negative rewards occasionally to avoid bad behaviours and prioritize good behaviours. It is less commonly used and more challenging, but it has become increasingly popular in recent years, for instance in gaming (Silver et al., 2017) or healthcare (Coronato et al., 2020) applications.

More recently, a fourth type of learning has emerged: *semi-supervised learning*, which falls between the supervised and unsupervised learning. Semi-supervised algorithms usually use a large amount of unlabelled data in conjunction with a small amount of labelled data. This approach is very useful for applications where the labelling process is challenging or expensive (e.g., in neuroimaging (Honnorat et al., 2019; Wen et al., 2021)).

In this thesis, I use methods that can be considered supervised and unsupervised learning approaches, such as Canonical Correlation Analysis (CCA) and Group Factor Analysis (GFA). We use CCA in an unsupervised manner only (Chapters 3 and 4) to uncover associations among multiple views, but it can also be used as a supervised approach, e.g., to predict one view from

the other. GFA was used as a supervised and unsupervised model to uncover associations among multiple views (Chapters 5 and 6), predict behaviour from brain functional connectivity (Chapter 5) and predict the probabilities of subjects to belong to the underlying subtypes (Chapter 6).

2.1.1 Model selection and prediction

As mentioned above, the generalisability of a machine learning model relates to its predictive performance on a test set, since the model performance on the training set is not a good indicator due to the problem of overfitting (Bishop, 2006; Hastie et al., 2009). In practice, the data available is sometimes limited, and therefore it is difficult to choose a good proportion of training and test observations to simultaneously learn the model's parameters properly and obtain a useful estimate of the model's predictive performance. The procedure commonly applied to address this practical issue is called *cross-validation*, where the data is first split into k distinct partitions, one partition is used as test set and the remaining ones are used as training set. A model is fit to the training set (i.e., the model's parameters are estimated based on the training set) and evaluated k times on the test set (each time a different partition is used as test set). The model's predictive performances obtained for all test sets (e.g., accuracy) are averaged at the end of the procedure. When only a few different partitions are considered ($k = 5, 10$), the procedure is called k-fold cross validation (Bishop, 2006; Hastie et al., 2009). Finally, the statistical significance of the results can be assessed using *permutation tests* (Fisher, 1935; Pitman, 1938). The goal of permutation tests is to generate "randomised" data by permuting either the inputs or outputs to break the associations between them, and test whether the predictive performance metric on the test set could have happened by chance. This is achieved by computing a p-value to assess whether the predictive performance metric obtained with the non-permuted/original data is larger than the null distribution of those obtained with the permuted data sets. A small p-value means that it is unlikely that the predictive performance of the model was obtained by chance.

In most machine learning applications, the researchers need to train a range of different models or the same model with different hyperparameters (e.g., regularisation parameters). If enough data is available, a separate independent set of the data, called a *validation set*, should be considered to select the best model or set of hyperparameters. As the model selection step is usually performed multiple times, the risk of the model overfitting the val-

idation set is high, therefore the generalisability of the best model should be assessed on the test set, which is held-out during model selection (Hastie et al., 2009). This procedure is often incorporated into a cross-validation scheme for hyperparameter search and is known as *nested cross-validation*, which consists of two loops of cross-validation. In the outer loop, the data is divided into training and test sets. In the inner loop, the training set is further divided into training and validation sets, where the model is fit to the training set and the best hyperparameters are chosen to maximise the predictive performance metric on the validation sets. In the outer loop, the model is fit to the training set using the best hyperparameters, chosen in the inner loop, and the model's predictive performance is assessed on the held-out test sets. In summary, the inner loop is used to optimise the hyperparameters and the outer loop is used to access the generalisability of the model.

The use of cross-validation might be computationally expensive in cases where the training of different models is itself demanding, or multiple hyperparameters need to be tuned. In deterministic approaches, researchers often use a more statistical framework, in which the model is trained on the whole data set and the statistical significance of the results is assessed using permutation tests, as explained above. Although a small p-value indicates that the results may have not happen by chance, the predictive performance of the model cannot be assessed and there is the risk of overfitting because no independent test set is used.

In probabilistic approaches, it is possible to choose the best model on a single training run (i.e., no validation set is needed) using approaches, such as the Akaike Information Criterion (Akaike, 1974) or Bayesian Information Criterion (Schwarz, 1978). In these approaches, penalty terms are included to control for model complexity and to avoid overfitting. However, in practice, these metrics usually favour very simple models and do not take into account the uncertainty of the model's parameters (Bishop, 2006; Hastie et al., 2009). These limitations can be addressed by applying a fully Bayesian approach to model selection. In Bayesian model selection, we assume a set of candidate models $\mathcal{M}_m$, $m = 1, ..., M$ and corresponding model parameters $\boldsymbol{\theta}_m$, and we are uncertain which model is the best. Bayes' theorem allow us to incorporate our prior knowledge with any evidence (i.e., data) to obtain an updated posterior belief. Given a training set $\mathcal{D}$ and assuming the prior distribution $p(\boldsymbol{\theta}_m|\mathcal{M}_m)$ over the parameters of each model $\mathcal{M}_m$, we can compute the posterior probability of a given model using Bayes' theorem (Bishop, 2006; Hastie

et al., 2009):

$$\begin{aligned} p(\mathcal{M}_m|\mathcal{D}) &\propto p(\mathcal{M}_m)p(\mathcal{D}|\mathcal{M}_m), \\ &\propto p(\mathcal{M}_m)\int p(\mathcal{D}|\boldsymbol{\theta}_m,\mathcal{M}_m)p(\boldsymbol{\theta}_m|\mathcal{M}_m)d\boldsymbol{\theta}_m, \end{aligned} \tag{2.1}$$

where $p(\mathcal{M}_m)$ represents the prior probability of each m-th model and $p(\mathcal{D}|\mathcal{M}_m)$ is the *model evidence*, which can be interpreted as the probability of the data being generated by a given model $\mathcal{M}_m$, under the prior belief about the model parameters $\boldsymbol{\theta}_m$. The model evidence is also called the *marginal likelihood* because it can be seen as a likelihood function over the space of models, in which the parameters have been marginalised out, using the sum and product rules of probability, as shown in Equation 2.1 (Bishop, 2006). Due to the marginalisation over the model parameters, we can control the model complexity and avoid overfitting or underfitting by favouring models with intermediate complexity (Bishop, 2006). Finally, to compare two models $\mathcal{M}_m$ and $\mathcal{M}_j$ we can compute the ratio of their posterior probabilities:

$$\frac{p(\mathcal{M}_m|\mathcal{D})}{p(\mathcal{M}_j|\mathcal{D})} = \frac{p(\mathcal{M}_m)}{p(\mathcal{M}_j)}\frac{p(\mathcal{D}|\mathcal{M}_m)}{p(\mathcal{D}|\mathcal{M}_j)}, \tag{2.2}$$

where the ratio of the model evidences (i.e., the rightmost quantity) is known as the *Bayes factor* (Kass and Raftery, 1995). If we assume equal prior probability for all different models, the model $\mathcal{M}_m$ is considered the best if the Bayes factor is greater than one, i.e. $p(\mathcal{D}|\mathcal{M}_m)$ is larger than $p(\mathcal{D}|\mathcal{M}_j)$. In practice, if multiple models are compared, the model with the highest model evidence is usually considered the best. As the model evidence might be sensitive to the choice of the prior distribution, the model's predictive performance should still be evaluated in an independent test set (Bishop, 2006).

2.2 Classical multi-view methods

In this section, I begin by briefly introducing Canonical Correlation Analysis (CCA) (Section 2.2.1), Partial Least Squares (PLS) (Section 2.2.2), and describing sparse CCA (Section 2.2.3). In Section 2.2.4, I present some applications of these models to neuroimaging. These methods were used in the studies presented in Chapters 3 and 4.

2.2.1 Canonical Correlation Analysis

CCA was introduced by Hotelling (1936) and it is a multivariate statistical method for exploring linear associations between two views. Let the two views be denoted by $\mathbf{X}^{(1)} \in \mathbb{R}^{D_1 \times N}$ and $\mathbf{X}^{(2)} \in \mathbb{R}^{D_2 \times N}$ (D_1 and D_2 denote the number of features of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively), where the N samples are assumed to be generated from a common latent process. CCA finds pairs of weight vectors $\mathbf{u}_k \in \mathbb{R}^{D_1 \times 1}$ and $\mathbf{v}_k \in \mathbb{R}^{D_2 \times 1}$, $k = 1, \ldots, K$ (where K is the number of canonical directions, also called CCA modes) that maximise the (canonical) correlation between the corresponding projections $\mathbf{u}_k^T \mathbf{X}^{(1)}$ and $\mathbf{v}_k^T \mathbf{X}^{(2)}$ (also known as canonical variates). This is achieved by solving the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{u}_k, \mathbf{v}_k} \mathbf{u}_k^T \mathbf{X}^{(1)} \mathbf{X}^{(2)T} \mathbf{v}_k, \\ &\text{s.t. } \mathbf{u}_k^T \mathbf{X}^{(1)} \mathbf{X}^{(1)T} \mathbf{u}_k = 1 \text{ and } \mathbf{v}_k^T \mathbf{X}^{(2)} \mathbf{X}^{(2)T} \mathbf{v}_k = 1, \end{aligned} \tag{2.3}$$

where the features of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are considered to be standardised to zero mean and unit variance. The optimisation problem in Equation (2.3) can be solved using a standard eigenvalue solution (Hotelling, 1936), singular value decomposition (SVD) (Uurtio et al., 2017), alternating least squares (Golub and Zha, 1994) or non-linear iterative partial least squares (Wegelin, 2000). The description of these approaches is beyond the scope of this thesis. For more details, please refer to Uurtio et al. (2017).

2.2.2 Partial Least Squares

PLS was introduced by Wold (1985) and, similarly to CCA, it is a multivariate statistical method to uncover associations between two views ($\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, defined as in Section 2.2.1). PLS computes a pair of weight vectors $\mathbf{u}_k$ and $\mathbf{v}_k$, $k = 1, \ldots, K$ (K is the number of PLS modes), such that the projections $\mathbf{u}_k^T \mathbf{X}^{(1)}$ and $\mathbf{v}_k^T \mathbf{X}^{(2)}$ have maximum covariance (Wegelin, 2000). This is achieved by solving the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{u}_k, \mathbf{v}_k} \mathbf{u}_k^T \mathbf{X}^{(1)} \mathbf{X}^{(2)T} \mathbf{v}_k, \\ &\text{s.t. } ||\mathbf{u}_k||_2 = 1 \text{ and } ||\mathbf{v}_k||_2 = 1, \end{aligned} \tag{2.4}$$

where $||\cdot||_2$ represents the L_2-norm (which is defined as the square root of the sum of the squares of the components of a vector in a space). The optimisation problem in Equation 2.4 can be solved by performing the rank-1 approximation of ${\mathbf{X}^{(1)}}^T \mathbf{X}^{(2)}$ using SVD (Wegelin, 2000). After the first weight vectors ($\mathbf{u}_1$ and $\mathbf{v}_1$) are computed, the association explained by these vectors is removed from

the data by performing matrix deflation of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ (Monteiro et al., 2016). The process is repeated to find the subsequent weight vector pairs. For more details about PLS variants, see Wegelin (2000); Rosipal and Krämer (2006).

2.2.3 Sparse CCA

Sparse CCA is a regularised variant of CCA (Lê Cao et al., 2008; Parkhomenko et al., 2009; Waaijenborg et al., 2008; Witten et al., 2009) that enables CCA to be applied to high dimensional data and improves the interpretability of the results. In this section, the method proposed by Witten et al. (2009) is explained in more detail, as this is the method used in Chapter 4. The comparison with other versions of sparse CCA is beyond the scope of this thesis. For more details regarding other sparse CCA methods, see Lê Cao et al. (2008); Waaijenborg et al. (2008); Parkhomenko et al. (2009)

Sparse CCA computes sparse weight vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ that maximise the correlation between $\mathbf{u}_k^T\mathbf{X}^{(1)}$ and $\mathbf{v}_k^T\mathbf{X}^{(2)}$ (similarly defined as in Section 2.2.1). This is achieved by adding sparsity penalties to the norm of weight vectors (e.g., a L_1-norm, represented as $||\cdot||_1$, which corresponds to the sum of the magnitudes of a vector in a space) to the CCA optimisation problem (Equation 2.3). Due to the geometry of the L_1-norm, these penalties will allow feature selection by shrinking some "irrelevant" weights (e.g., collinear weights) to zero, which is a useful property to regularise the models in high dimensional spaces. In the case of sparse CCA, this corresponds to the following optimisation problem (Witten et al., 2009):

$$\begin{aligned} &\max_{\mathbf{u}_k,\mathbf{v}_k} \mathbf{u}_k^T\mathbf{X}^{(1)}\mathbf{X}^{(2)T}\mathbf{v}_k, \\ \text{s.t. } &\mathbf{u}_k^T\mathbf{X}^{(1)}\mathbf{X}^{(1)T}\mathbf{u}_k \leq 1,\ \mathbf{v}_k^T\mathbf{X}^{(2)}\mathbf{X}^{(2)T}\mathbf{v}_k \leq 1, ||\mathbf{u}_k||_1 \leq c_u, ||\mathbf{v}_k||_1 \leq c_v. \end{aligned} \tag{2.5}$$

In high dimensional data, the calculations of $\mathbf{X}^{(1)}\mathbf{X}^{(1)^T}$ and $\mathbf{X}^{(2)}\mathbf{X}^{(2)^T}$ are computationally demanding; therefore (Witten et al., 2009) proposed to replace $\mathbf{X}^T\mathbf{X}$ and $\mathbf{Y}^T\mathbf{Y}$ by identity matrices:

$$\begin{aligned} &\max_{\mathbf{u}_k,\mathbf{v}_k} \mathbf{u}_k^T\mathbf{X}^{(1)}\mathbf{X}^{(2)T}\mathbf{v}_k, \\ \text{s.t. } &\|\mathbf{u}_k\|_2^2 \leq 1, \|\mathbf{v}_k\|_2^2 \leq 1, \|\mathbf{u}_k\|_1 \leq c_u, \|\mathbf{v}_k\|_1 \leq c_v, \end{aligned} \tag{2.6}$$

where c_u and c_v are the regularisation parameters (or sparse CCA hyperparameters) that control the L_1 penalties of $\mathbf{u}_k$ and $\mathbf{v}_k$, respectively. If c_u and c_v are sufficiently small, the L_1 penalties impose sparsity on the weights and, con-

sequently, fewer features are included in the model. The pair of regularisation parameters can be optimised using permutation tests to assess the significance of the canonical correlation on the whole data set, (Witten and Tibshirani, 2009) or using multiple training and validation sets to obtain an out-of-sample performance metric (Monteiro et al., 2016). A comparison between these approaches is presented in Chapter 4. Although this method is referred to as "diagonal penalised CCA" in Witten et al. (2009), what is being maximised is no longer the correlation between $\mathbf{u}_k^T\mathbf{X}^{(1)}$ and $\mathbf{v}_k^T\mathbf{X}^{(2)}$, but the covariance between them, therefore the optimisation problem becomes equivalent to sparse PLS.

The sparse weight vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ are obtained in an iterative manner by generating 1-rank approximations of the covariance matrix, where a soft-thresholding operator is applied in each iteration (Witten et al., 2009). As in PLS, the associations explained by the weight vectors are removed from the data by performing matrix deflation of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, iteratively. There are different deflation methods, such as Hotelling's deflation (Witten et al., 2009), projection deflation (Mackey, 2008; Monteiro et al., 2016) or mode-A deflation (Wegelin, 2000; Mihalik et al., 2020). The projection deflation, which is used in the study presented in Chapter 4, removes the association explained by $\mathbf{u}_k$ and $\mathbf{v}_k$ from $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ by subtracting from each view the projection of each view onto the space spanned by the corresponding weight vector:

$$\begin{aligned} \mathbf{X}_{k+1}^{(1)} &= \mathbf{X}_k^{(1)} - \mathbf{u}_k(\mathbf{u}_k^T\mathbf{X}^{(1)}), \\ \mathbf{X}_{k+1}^{(2)} &= \mathbf{X}_k^{(2)} - \mathbf{v}_k(\mathbf{v}_k^T\mathbf{X}^{(2)}). \end{aligned} \tag{2.7}$$

The description of the rest of the deflation techniques is beyond the scope of this thesis. For a more detailed description of the sparse CCA algorithm, see Witten et al. (2009).

2.2.4 Applications to neuroimaging

Due to the high dimensionality of neuroimaging datasets, CCA has to be applied jointly with a dimensionality reduction or regularisation technique. Recently, CCA has been applied jointly with Principal Component Analysis to investigate associations between brain connectivity, demographics and behaviour in healthy population (Smith et al., 2015; Bijsterbosch et al., 2018; Li et al., 2019), healthy and clinically depressed young adolescents (Mihalik et al., 2019) and children (Alnæs et al., 2020).

PLS and its variants have also been widely applied to neuroimaging data

(Krishnan et al., 2011). For instance, PLS has been used in studies of emotional processing (Keightley et al., 2003), memory (Della-Maggiore et al., 2000) and behavioural (Vallesi et al., 2009) tasks to explore associations between brain imaging features (e.g. functional MRI) and task measurements (e.g. pictures or colour patterns). It has also been used to find brain-behaviour associations in schizophrenia (Nestor et al., 2002). For a more extended review of PLS applications in neuroimaging, see Krishnan et al. (2011).

Different sparse CCA algorithms were proposed to uncover associations between different types of genomics data (Waaijenborg et al., 2008; Witten et al., 2009; Witten and Tibshirani, 2009; Parkhomenko et al., 2009) and omics data (Lê Cao et al., 2008). Sparse CCA has also been widely applied to brain imaging to identify associations between genetic polymorphisms and brain activity during a cognitive functional MRI task (Le Floch et al., 2012), and between single nucleotide polymorphisms and brain activity in schizophrenic patients (Lin et al., 2014). Avants et al. (2010) applied sparse CCA to explore associations between brain structure and diffusion tensor imaging of Alzheimer's disease and frontotemporal dementia patients. Monteiro et al. (2016) used sparse CCA to uncover associations between brain structure and demographic and clinical/cognitive data in a sample of healthy controls and patients with Alzheimer disease and mild cognitive impairment. More recently, sparse CCA was also applied to find associations between behavioural, clinical, and multimodal imaging phenotypes in psychosis (Moser et al., 2018), and between functional connectivity and psychiatric symptoms in a large sample of young people (Xia et al., 2018). Finally, Mihalik et al. (2020) used sparse CCA to identify associations between brain structure, demographic and behavioural measures.

2.3 Latent variable models

Here, I begin by briefly introducing the probabilistic and Bayesian CCA models (Section 2.3.1 and Section 2.3.2, respectively) and explaining their connection to Group Factor Analysis (GFA, Section 2.3.3). Then, I describe a sparse extension of GFA (Section 2.3.4) and finalise by presenting some applications of these models to neuroimaging datasets. These models were used in the studies presented in Chapters 5 and 6.

2.3.1 Probabilistic CCA

The probabilistic interpretation of CCA (Figure 2.1) assumes that N samples of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ (similarly defined as in Section 2.2.1) are generated by the same latent variables $\mathbf{Z} \in \mathbb{R}^{K \times N}$ that explain the associations between data modalities (Bach and Jordan, 2006), where K corresponds to the number of components (which are equivalent to the CCA modes described in Section 2.2.1):

$$\begin{aligned} \mathbf{z}_n &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \\ \mathbf{x}_n^{(1)} &\sim \mathcal{N}(\mathbf{A}^{(1)}\mathbf{z}_n + \boldsymbol{\mu}^{(1)}, \boldsymbol{\Phi}^{(1)}), \\ \mathbf{x}_n^{(2)} &\sim \mathcal{N}(\mathbf{A}^{(2)}\mathbf{z}_n + \boldsymbol{\mu}^{(2)}, \boldsymbol{\Phi}^{(2)}), \end{aligned} \tag{2.8}$$

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 loading matrices (also known as projection matrices) that represent the transformations of the latent variables $\mathbf{z}_n \in \mathbb{R}^{K \times 1}$ into the input space. $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ are equivalent to the (horizontal) concatenation of all $\mathbf{u}_k$ and $\mathbf{v}_k$, respectively, that CCA finds (see Section 2.2.1). $\boldsymbol{\Phi}^{(1)} \in \mathbb{R}^{D_1 \times D_1}$ and $\boldsymbol{\Phi}^{(2)} \in \mathbb{R}^{D_2 \times D_2}$ denote the noise covariance matrices, and $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ are the mean parameters.

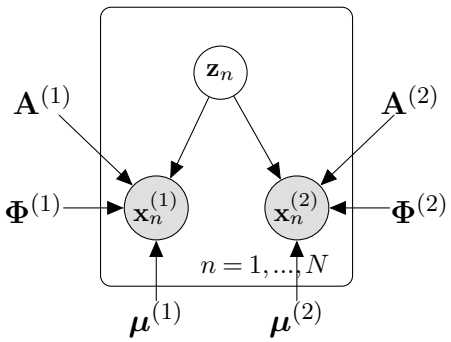

Figure 2.1: Graphical representation of the probabilistic CCA model.

Bach and Jordan proved that the maximum likelihood estimates of the parameters in Equation 2.8 lead to the same canonical directions as classical CCA up to a rotation (Bach and Jordan, 2006), i.e., the posterior expectations $\mathbb{E}[\mathbf{Z}|\mathbf{X}^{(1)}]$ and $\mathbb{E}[\mathbf{Z}|\mathbf{X}^{(2)}]$ lie in the same subspace that classical CCA finds (which is represented by the canonical variates $\mathbf{U}^T\mathbf{X}^{(1)}$ and $\mathbf{V}^T\mathbf{X}^{(2)}$, where $\mathbf{U} \in \mathbb{R}^{D_1 \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_2 \times K}$). An equivalent representation of the latent variables $\mathbf{Z}$ can be obtained - for CCA - by averaging the canonical variates obtained for each view (Klami et al., 2013).

Although probabilistic CCA does not provide an explicit inference ap-

proach to estimate 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.

2.3.2 Bayesian CCA

Klami and Kaski (2007) and Chong Wang (2007) proposed a hierarchical Bayesian extension of CCA by introducing suitable prior distributions over the model parameters, which can then be inferred using Bayesian inference. The goal of Bayesian inference is to provide a procedure for incorporating our prior beliefs about unknown random variables $\boldsymbol{\theta}$ (e.g., latent variables and model parameters) with any evidence (e.g., a data set $\mathcal{D}$) to obtain an updated posterior belief. This is done using Bayes' theorem: $p(\boldsymbol{\theta}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathcal{D})$, where $p(\boldsymbol{\theta})$ represents the prior distributions over $\boldsymbol{\theta}$, $p(\mathcal{D}|\boldsymbol{\theta})$ represents the likelihood and $p(\boldsymbol{\theta}|\mathcal{D})$ represents the joint posterior distribution that expresses the uncertainty about $\boldsymbol{\theta}$ after accounting for the prior knowledge and data. $p(\mathcal{D})$ represents the model evidence (Equation 2.1) and is usually considered a normalising constant. In this way, Bayes' theorem is formulated as: $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

In the Bayesian CCA model (represented graphically in Figure 2.2), the samples of $\mathbf{X}^{(m)} \in \mathbb{R}^{D_m \times N}$ are assumed to be generated by Equation 2.8. The joint probabilistic distribution of the model is given by (Chong Wang, 2007):

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}, \mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\Phi}, \boldsymbol{\mu}) = \prod_{m=1}^{M} \Big[& p(\mathbf{X}^{(m)}|\mathbf{Z}, \mathbf{A}^{(m)}, \boldsymbol{\Phi}^{(m)}, \boldsymbol{\mu}^{(m)}) \times \\ & p(\mathbf{A}^{(m)}|\boldsymbol{\alpha}^{(m)}) p(\boldsymbol{\alpha}^{(m)}) p(\boldsymbol{\Phi}^{(m)}) p(\boldsymbol{\mu}^{(m)}) \Big] p(\mathbf{Z}), \end{aligned} \tag{2.9}$$

where M is the number of views, $\mathbf{A}^{(m)}$ and $\mathbf{Z}$ are defined as in Equation (2.8) 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:

$$\begin{aligned} p(\mathbf{A}^{(m)}|\boldsymbol{\alpha}^{(m)}) = \prod_{j=1}^{D_m} \prod_{k=1}^{K} \mathcal{N}(a_{j,k}^{(m)}|0, (\alpha_k^{(m)})^{-1}), \; p(\boldsymbol{\alpha}^{(m)}) = \prod_{k=1}^{K} \Gamma(\alpha_k^{(m)}|a_{\boldsymbol{\alpha}}^{(m)}, b_{\boldsymbol{\alpha}}^{(m)}), \\ p(\boldsymbol{\mu}^{(m)}) = \mathcal{N}(\boldsymbol{\mu}^{(m)}|0, (\beta^{(m)})^{-1}\mathbf{I}), \; p(\boldsymbol{\Phi}^{(m)}) = \mathcal{W}^{-1}(\boldsymbol{\Phi}^{(m)}|\mathbf{S}_0^{(m)}, \nu_0^{(m)}), \end{aligned} \tag{2.10}$$

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 loading 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 achieved by allowing some $\alpha_k^{(m)}$ to be pushed towards infinity, which consequently drives the loadings (i.e., elements of the loading matrices) of the corresponding kth columns of $\mathbf{A}^{(m)}$ close to zero. The corresponding irrelevant latent components k are then pruned out during inference.

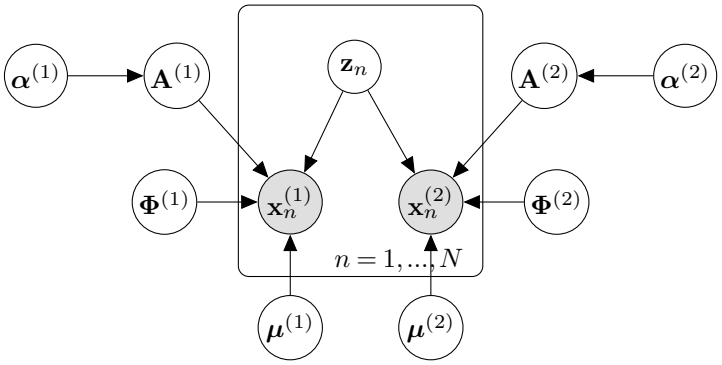

Figure 2.2: Graphical representation of the Bayesian CCA model.

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 $p(\mathbf{Z}, \mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\Phi}, \boldsymbol{\mu}|\mathbf{X})$ and marginalise out uninteresting variables. However, these marginalisations are often analytically intractable, so the posterior distribution needs to be approximated. This can be achieved using mean-field variational approximation (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 $\boldsymbol{\Phi}^{(m)}$ (Klami et al., 2013). The inference algorithms usually need to invert those matrices in every step, leading to long computational times. Moreover, Bayesian CCA does not account for view-specific associations.

Virtanen et al. (2011) proposed an extension of Bayesian CCA to impose view-wise sparsity to separate associations between views from those within each view of the data. Moreover, this model assumes spherical noise covariance matrices ($\boldsymbol{\Phi}^{(m)} = {\sigma^{(m)}}^2\mathbf{I}$, where ${\sigma^{(m)}}^2$ corresponds to the noise variance of the m-th view) for more efficient inference. The same authors proposed a further extension of the model to uncover associations between more than two views,

called Group Factor Analysis (GFA) (Virtanen et al., 2012; Klami et al., 2015).

2.3.3 Group Factor Analysis

In the GFA problem, it is assumed that a collection of N samples, stored in $\mathbf{X} \in \mathbb{R}^{D \times N}$, have disjoint M partitions of features D_m called groups (in this thesis, I refer to distinct groups of features as views, e.g., different data modalities), $\mathbf{X}$ ($\mathbf{X}^{(m)} \in \mathbb{R}^{D_m \times N}$ for the mth view). Moreover, the latent components correspond to the rows of $\mathbf{Z} \in \mathbb{R}^{K \times N}$ (as in probabilistic and Bayesian CCA).

GFA finds a set of K components that can separate the associations between views (i.e., shared components) from those within views (i.e., view-specific components) by considering a joint component model (Figure 2.3), where each mth view is generated as follows (Virtanen et al., 2012; Klami et al., 2015):

$$\begin{aligned} \mathbf{z}_n &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \\ \mathbf{x}_n^{(m)} &\sim \mathcal{N}(\mathbf{W}^{(m)}\mathbf{z}_n, {\mathbf{T}^{(m)}}^{-1}), \end{aligned} \tag{2.11}$$

where ${\mathbf{T}^{(m)}}^{-1}$ is a diagonal covariance matrix ($\mathbf{T}^{(m)} = \tau^{(m)}\mathbf{I}$), where $\tau^{(m)}$ represents the noise precisions, i.e., inverse noise variances, of the mth view), $\mathbf{W}^{(m)} \in \mathbb{R}^{D_m \times K}$ is the loading matrix of view m and $\mathbf{z}_n \in \mathbb{R}^{K \times 1}$ is the latent variable for a given observation $\mathbf{x}_n^{(m)}$ (i.e., row 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 training the model (Klami et al., 2013).

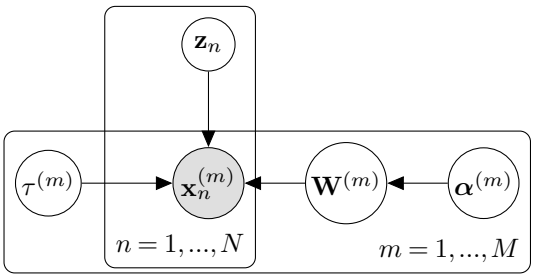

Figure 2.3: Graphical representation of the GFA model.

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 con-

taining the shared components (equivalent to those obtained by probabilistic and Bayesian CCA) and $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ correspond to the loading matrices containing the view-specific components. The structure of $\mathbf{W}$ and the corresponding latent structure (represented by $\mathbf{Z}$) is learned automatically by imposing a view-wise sparsity on the components (Virtanen et al., 2011), which is achieved by assuming independent ARD priors to encourage sparsity over the views (Virtanen et al., 2012; Klami et al., 2015):

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{m=1}^{M}\prod_{j=1}^{D_m}\prod_{k=1}^{K}\mathcal{N}(w_{j,k}^{(m)}|0,(\alpha_k^{(m)})^{-1}),\ p(\boldsymbol{\alpha}) = \prod_{m=1}^{M}\prod_{k=1}^{K}\Gamma(\alpha_k^{(m)}|a_{\boldsymbol{\alpha}^{(m)}},b_{\boldsymbol{\alpha}^{(m)}}), \tag{2.12}$$

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 is chosen to be uninformative to enable the automatic pruning of irrelevant latent components. $\Gamma(\cdot)$ represents a gamma distribution with shape parameter $a_{\boldsymbol{\tau}^{(m)}}$ and rate parameter $b_{\boldsymbol{\tau}^{(m)}}$. These separate priors cause features of some views to be pushed close to zero for some components k ($\mathbf{w}_k^{(m)} \to 0$) by driving the corresponding $\alpha_k^{(m)}$ towards infinity. If the loadings of certain components are pushed towards zero for all views, the underlying latent component is deemed inactive and pruned out. (Klami et al., 2013). Finally, the prior distributions over the noise and latent variables $\mathbf{Z}$ are:

$$p(\boldsymbol{\tau}) = \prod_{m=1}^{M}\Gamma(\tau^{(m)}|a_{\tau^{(m)}},b_{\tau^{(m)}}),\quad p(\mathbf{Z}) = \prod_{k=1}^{K}\prod_{n=1}^{N}\mathcal{N}(z_{k,n}|0,1), \tag{2.13}$$

with shape parameter $a_{\tau^{(m)}}$ and rate parameter $b_{\tau^{(m)}}$ of the gamma distribution. The hyperparameters $a_{\boldsymbol{\alpha}^{(m)}}, b_{\boldsymbol{\alpha}^{(m)}}, a_{\tau^{(m)}}, b_{\tau^{(m)}}$ 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:

$$p(\mathbf{X},\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\tau}) = p(\mathbf{X}|\mathbf{Z},\mathbf{W},\boldsymbol{\tau})p(\mathbf{Z})p(\mathbf{W}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\boldsymbol{\tau}). \tag{2.14}$$

As mentioned in Section 2.3.2, 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 using, for instance, similarly to Bayesian CCA, mean-field variational approximation. This involves approximating the true posterior $p(\boldsymbol{\theta}|\mathcal{D})$ by a suitable factorized distribution $q(\boldsymbol{\theta})$ (Bishop, 1999). The marginal log-likelihood ($\ln p(\mathcal{D})$) can be

decomposed as follows (Bishop, 2006):

$$\begin{aligned} \ln p(\mathcal{D}) &= \mathcal{L}(q) + D_{KL}(q||p), \\ \mathcal{L}(q) &= \int q(\boldsymbol{\theta}) \ln \frac{p(\mathcal{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}, \\ D_{KL}(q||p) &= \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}|\mathcal{D})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}, \end{aligned} \tag{2.15}$$

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 $\ln p(\mathcal{D})$ is constant, maximising $\mathcal{L}(q)$ is equivalent to minimising $D_{KL}(q||p)$, 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):

$$\ln q_i(\boldsymbol{\theta}_i) = \langle \ln p(\mathbf{X}, \boldsymbol{\theta}) \rangle_{j \neq i} + \text{const}, \tag{2.16}$$

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:

$$q(\boldsymbol{\theta}) = q(\mathbf{Z}) \prod_{m=1}^{M} \left[q(\mathbf{W}^{(m)}) q(\boldsymbol{\alpha}^{(m)}) q(\boldsymbol{\tau}^{(m)}) \right], \tag{2.17}$$

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 Equation 2.16) results in the following analytically tractable distributions:

$$\begin{aligned} q(\mathbf{Z}) &= \prod_{n=1}^{N} \mathcal{N}(\mathbf{z}_n | \boldsymbol{\mu}_{\mathbf{z}_n}, \boldsymbol{\Sigma}_{\mathbf{z}_n}), \quad q(\mathbf{W}^{(m)}) = \prod_{j=1}^{D_m} \mathcal{N}(\mathbf{W}^{(m)}_{j,*} | \boldsymbol{\mu}_{\mathbf{W}^{(m)}_{j,*}}, \boldsymbol{\Sigma}_{\mathbf{W}^{(m)}_{j,*}}), \\ q(\boldsymbol{\alpha}^{(m)}) &= \prod_{k=1}^{K} \Gamma(\alpha_k^{(m)} | \tilde{a}_{\boldsymbol{\alpha}^{(m)}}, \tilde{b}^{(k)}_{\boldsymbol{\alpha}^{(m)}}), \quad q(\boldsymbol{\tau}^{(m)}) = \Gamma(\tau^{(m)} | \tilde{a}_{\boldsymbol{\tau}^{(m)}}, \tilde{b}_{\boldsymbol{\tau}^{(m)}}), \end{aligned} \tag{2.18}$$

where $\mathbf{z}_n$ is the n-th column of $\mathbf{Z}$ and $\mathbf{W}^{(m)}_{j,*}$ denotes the j-th row of $\mathbf{W}^{(m)}$. The optimisation is done using variational Expectation-Maximization, where the parameters in Equation 2.18 are updated sequentially until convergence, which is achieved when a relative change of the lower bound $\mathcal{L}(q)$ falls below an arbitrary low number (e.g., 10^{-6}) (Klami et al., 2013).

2.3.4 Sparse GFA

As described in the previous section, GFA provides view-wise sparsity; however, in some applications only a few subsets of features within each view might be associated with features in other views. Therefore, feature-wise sparsity is important to improve the interpretability of the model. This can be achieved by adding a popular shrinkage prior in sparse Bayesian estimation, i.e, the spike-and-slab prior (Mitchell and Beauchamp, 1988), over the loading matrices (Khan et al., 2014; Bunte et al., 2016):

$$\begin{aligned} w_{j,k}^{(m)}|h_{j,k}^{(m)},\alpha_k^{(m)} &\sim h_{j,k}^{(m)}\mathcal{N}\left(0,(\alpha_k^{(m)})^{-1}\right)+(1-h_{j,k}^{(m)})\delta_0, \\ h_{j,k}^{(m)}|\pi_k^{(m)} &\sim \text{Bernoulli}(\pi_k^{(m)}), \quad \pi_k^{(m)} \sim \text{Beta}(a_\pi,b_\pi), \end{aligned} \tag{2.19}$$

where $h_{j,k}^{(m)}$ is binary and determines whether the component k is active in the j-th feature of $\mathbf{X}^{(m)}$, $\pi_k^{(m)}$ represents the probability of $h_{j,k}^{(m)}=1$ and $\alpha_k^{(m)}$ is sampled from the prior in Equation 2.12 and determines the scale of the component k in view m. If the Gamma prior is uninformative, as explained in Section 2.3.3, it implements view-wise sparsity. In this way, assuming these priors and those in Equation 2.13 for the noise precisions and latent variables, the model imposes view and feature-wise sparsity, which enables finding associations between subsets of features among multiple data modalities.

The spike-and-slab prior can also be applied to the latent variables to impose sample-sparsity. This is relevant if one assumes that there are associations present only in subsamples of the data (e.g. subgroups of patients). The prior over the latent variables is defined as follows (Bunte et al., 2016):

$$\begin{aligned} z_{n,k}|h_{n,k}^{(z)},\alpha_k^{(z)} &\sim h_{n,k}^{(z)}\,\mathcal{N}\left(0,(\alpha_k^{(z)})^{-1}\right)+(1-h_{n,k}^{(z)})\delta_0, \\ h_{n,k}^{(z)}|\pi_k^{(z)} \sim \text{Bernoulli}(\pi_k^{(z)}), \quad \pi_k^{(z)} &\sim \text{Beta}(a_\pi,b_\pi), \quad \alpha_k^{(z)} \sim \Gamma(a_\alpha,b_\alpha). \end{aligned} \tag{2.20}$$

Assuming the priors described above, the sparse GFA model (displayed in Figure 2.4) can identify subpopulations in data that share common characteristics (e.g., brain-behaviour associations), which may be described as associations between subsets of features of two or more views. The model parameters and latent variables are inferred using Gibbs sampling (Bunte et al., 2016). The description of Gibbs sampling is beyond the scope of this thesis, for more details see e.g., Bishop (2006); Gelman et al. (2013).

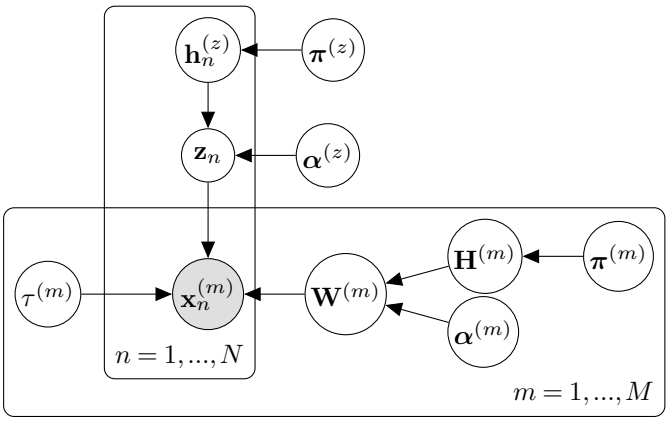

Figure 2.4: Graphical representation of sparse GFA using spike-and-slab priors.

2.3.5 Applications to neuroimaging

Sparse versions of Bayesian CCA have been used to reconstruct visual images from brain activity patterns (measured with functional MRI) (Fujiwara et al., 2009, 2013), uncover links between brain structure and single-nucleotide polymorphisms (Grellmann et al., 2015), and analyse relationships between brain activity and natural music stimuli (Virtanen et al., 2011). Mixture of Bayesian CCA models (where an additional multinomial latent variable was included to let each mixture cluster to model different kind of associations between the views) has also been applied to find associations between brain activity (measured with magnetoencephalography) and autonomic nervous system response under emotional sound stimuli (Viinikanoja et al., 2010) or speech segments (Koskinen et al., 2013).

Examples of GFA applications to neuroimaging are scarce. Although it has mostly been applied to genomics data (Klami et al., 2013; Suvitaival et al., 2014; Zhao et al., 2016; Bunte et al., 2016) and drug response data (Khan et al., 2014; Klami et al., 2015), GFA has also been used to uncover associations between brain activity and audio (Klami et al., 2015) or audiovisual stimuli (Virtanen et al., 2012; Remes et al., 2013). In our recent study (see Chapter 5), I applied GFA to explore associations between high-dimensional brain functional connectivity data, demographics, psychometrics and other behavioural measures in a sample of healthy people (Ferreira et al., 2021).

2.4 Data modalities

The models described in this thesis were applied to brain imaging data and non-imaging data (e.g., self-report questionnaires and cognitive tests). Here, I

briefly describe the two different brain imaging modalities used in this thesis: resting-state functional MRI (Section 2.4.1) and structural brain MRI (Section 2.4.2); as well as the non-imaging measures (Section 2.4.3). The description of the image acquisition and preprocessing techniques are beyond the scope of this thesis, but a brief description of these steps is provided in the Chapters where the datasets are introduced.

2.4.1 Resting-state functional MRI

Resting-state functional MRI (rs-fMRI) allows us to measure the brain activity during rest (i.e, no explicit task is being performed) by detecting local changes in cerebral blood flow, which can be quantified by measuring a blood-oxygen-level dependent (BOLD) signal using MRI. When neurons in a particular region of the brain are more active, there is an increased regional blood flow and oxygen supply that can be detected by comparing the relative levels of oxyhaemoglobin and deoxyhaemoglobin using their different magnetic properties (Lv et al., 2018). One way of analysing rs-fMRI data is to estimate the functional connectivity between different brain regions, which has shown the potential to provide biomarkers for better characterising brain disorders (Du et al., 2018). To achieve this, the rs-fMRI scans are parcellated into different regions or components, which can be done using pre-defined brain functional atlases (e.g., see Glasser et al. (2016)) or using data-driven models (e.g, Independent Component Analysis (ICA)). In Chapter 3 and 4, we use the former approach to parcellate the brain scans into approximately 350 regions. The regional time-series signal is estimated as the average of the time-series signal of all voxels within each region. Finally, brain functional connectivity can be estimated, for each subject, as the pairwise Pearson's correlation between the averaged signal of each possible pair of regions. In Chapter 5, I use the latter approach, where 200 brain parcellations are extracted using ICA and a brain functional connectivity matrix, for each subject, is calculated using pairwise partial correlations between all parcellations.

2.4.2 Structural brain MRI

Structural brain MRI is an imaging technique used to examine the anatomy and pathology of the brain. As the MRI signal varies across the different tissues, the brain MRI scans can be separated into four main components: grey matter that consists mostly of cell bodies (e.g., neurons and glial cells); white matter which is composed of long-range nerve fibres (myelinated axons) connecting the neurons, along with supporting glial cells; the cerebrospinal fluid,

a clear and colourless fluid providing mechanical and immunological protection to the brain; hard tissue (e.g., skull) (Symms et al., 2004). MR images with different types of contrast between tissues can be obtained using different type of MRI sequences, e.g., T1-weighted (which provides a good contrast between grey and white matter) and T2-weighted images (which shows good contrast between brain tissue and the cerebrospinal fluid) (Symms et al., 2004). The former is often more used to detect anatomical changes in grey matter, while the latter is used to detect white matter changes. In Chapter 6, I use grey matter volume extracted from T1-weighted MRI images, which is then parcellated into different cortical and subcortical regions using anatomical brain atlases.

2.4.3 Non-imaging data

As non-imaging data, we use item-level measures or total scores from self-report questionnaires (e.g., questionnaires completed by the patients), informant questionnaires (completed by primary caregivers), cognitive tests/neuropsychological tasks (completed by patients) and medical assessments. We also include demographic measures (e.g., age, gender and education) in the set of non-imaging measures. In Chapter 3 and 4, we use measures that assess psychopathological symptoms, personality characteristics, mental well-being and fluid intelligence (see Appendix A.1.1). In Chapter 5, we use demographics (age, sex, income and substance use), psychometrics (IQ, language performance) and other behavioural measures to assess, for instance, rule-breaking behaviour, mental well-being and personality (see Table B.5). In Chapter 6, we use measures of cognitive/neuropsychological tests to assess memory and language performance, medical assessments of disease severity and other behavioural measures to assess mood, self-care and abnormal behaviour (see Table C.1).

Chapter 3

Brain-behaviour modes of covariation in healthy and clinically depressed young people

The content of this chapter is based on a journal article published in *Scientific Reports* where I was joint first author with Agoston Mihalik (Mihalik et al., 2019). I have rewritten to avoid repetition with the background content described in Chapter 2. Agoston Mihalik, Maria J. Rosa and I implemented the learning frameworks. Agoston Mihalik and I ran the CCA analyses, prepared the results and wrote the paper. Michael Moutoussis, Edward T. Bullmore, Peter Fonagy, Ian M. Goodyer, Peter B. Jones, Raymond Dolan and the NeuroScience in Psychiatry Network (NSPN) consortium collected the data. Agoston Mihalik jointly with other co-authors preprocessed the MRI and questionnaire data. Agoston Mihalik, I, Michael Moutoussis and Rick A. Adams interpreted the results. Janaina Mourao-Miranda designed and supervised the study, and revised the manuscript.

3.1 Introduction

Adolescence and early adulthood are periods of high risk for the onset of many psychiatric disorders (Kessler et al., 2007; Paus et al., 2008), with up to a fifth of 18 to 25-year-olds seeking professional help for psychological distress (Lipari et al., 2014). Despite this, there are still no biological measures that inform early diagnosis and treatment. Neuroimaging techniques, especially resting-state functional MRI (Smith et al., 2013), enable researchers to relate

biological measures, such as patterns of functional brain connectivity, to the continuum of healthy to pathological states (Bassett and Bullmore, 2009).

As discussed in Chapter 2, CCA can be used to explore associations between multiple data modalities collected from the same individuals. Moreover, the fact that CCA can be used in an unsupervised manner has made it increasingly popular in several fields of neuroscience, such as psychiatry, where the diagnostic categories are not reliable (Insel et al., 2010; Bzdok and Meyer-Lindenberg, 2018). In this study, we applied CCA (coupled with PCA to reduce the dimensionality of the data) to resting-state fMRI and non-imaging measures (i.e., items of self-report questionnaires and demographics) to uncover associations between individual patterns of functional brain connectivity and individual sets of psychometrics/demographics during a key developmental period. We used permutation tests on the whole data set to assess the statistical significance of the CCA modes, as in Smith et al. (2015), and optimise the number of principal components. In addition, we propose a new framework to assess the generalisability of the statistically significant CCA modes and optimise the number of principal components using a similar multiple hold-out framework to that proposed by Monteiro et al. (2016).

Given the age range of our sample, we expected a strong age (or developmental) effect on the brain-behaviour modes of covariation. We predicted that variation in these modes would also be related to the presence of depression (Buckholtz and Meyer-Lindenberg, 2012), given that our sample also included approximately 9% of depressed subjects. Finally, we hypothesised that psychopathological symptoms might be associated with a core set of abnormal functional brain networks, incorporating default mode, frontoparietal and limbic networks as suggested by recent literature (Buckholtz and Meyer-Lindenberg, 2012; Menon, 2011).

3.2 Methods

3.2.1 Data

In total, 2406 healthy subjects and 50 subjects clinically diagnosed with depression (diagnosis and referral made by the subject's NHS GP) aged 14 to 24 years were recruited from schools, colleges, National Health Service (NHS) primary care and mental health services, and via direct advertisement in London and Cambridgeshire (Kiddle et al., 2018). This was carried out by the University College London and University of Cambridge NeuroScience in Psy-

chiatry Network research initiative. An MRI cohort was subsampled from the primary cohort, comprising a healthy cohort of 318 subjects and a depression cohort of 37 subjects. Furthermore, a demographically balanced cohort of 297 subjects was subsampled from the healthy cohort, with approximately 60 subjects in each of five age-defined strata: 14-15 years inclusive, 16-17, 18-19, 20-21, and 22-24 years.

Of the healthy cohort, two subjects were excluded due to low quality images, one was excluded due to gross radiological abnormalities, four were excluded due to missing convergence in multi-echo ICA preprocessing, and nine were excluded due to excessive motion during their resting-state functional MRI acquisitions (five subjects with maximum framewise displacements larger than 1.3 mm and four subjects with mean framewise displacements of 0.3 mm using a calculation by Power et al. (2012)). Of the depression cohort, three subjects were excluded due to low quality anatomical scans, one was excluded due to radiological artefacts, four were excluded due to motion-induced Freesurfer reconstruction errors, one was excluded due to lack of convergence, one was excluded due to extremely low explained variance ($< 20\%$) and two were excluded due to excessive motion during their resting-state functional MRI (the same criteria as for the healthy cohort was applied). These exclusion criteria produced a final healthy cohort consisting of 281 subjects (mean age=19.13, SD=2.88, 144 females) and a final depression cohort comprising 25 subjects (mean age=16.80, SD=1.15, 21 females).

Written informed consent was obtained for all subjects over the age of 16 years. For subjects under 16 years old, a written informed assent was obtained from their parent/legal guardian. The study was ethically approved by the Cambridge Central Research Ethics Committee and conducted in accordance with the NHS research governance standards.

3.2.1.1 Structural MRI data

All MRI data was acquired on three identical 3T whole-body MRI systems (Magnetom TIM Trio; VB17 software version; Siemens Healthcare): two located in Cambridge and one located in London. Between-site reliability and tolerability of all MRI procedures were satisfactorily assessed by a pilot study of five healthy volunteers at each site (Weiskopf et al., 2013). Only scans at the baseline visit were included in the current study. Structural MRI scans were acquired using a multi-echo acquisition protocol with six equidistant echo times (TE) between 2.2 and 14.7 ms and averaged to form a single image of

increased signal-to-noise ratio (Weiskopf et al., 2013). Apparent longitudinal relaxation rate R1 (R1=1/T1w) was calculated using previously developed models to create quantitative R1 maps (Helms et al., 2008a,b; Weiskopf et al., 2011). Other acquisition parameters included: temporal resolution of 18.70 ms, spatial resolution 1.0 mm isotropic, field of view (FOV) = 256 × 256, 176 sagittal slices and parallel imaging using GRAPPA factor 2 in the anterior-posterior phase-encoding direction.

The R1 images were used to perform a surface reconstruction of each subject using Freesurfer *recon-all* (Dale et al., 1999) (https://surfer.nmr.mgh.harvard.edu/). Freesurfer average subject (fsaverage) was parcellated using a multimodal scheme that subdivides the cortex into 360 bilaterally symmetric regions based on Human Connectome Project (HCP) data (Glasser et al., 2016). HCP parcellation was transformed from fsaverage space to the cortical surface of each individual subject using Freesurfer *mri_surf2surf*. In addition, 16 regions were used from the subcortical segmentation of Freesurfer (thalamus-proper, caudate, putamen, pallidum, hippocampus, amygdala, accumbens-area and ventral diencephalon for each hemisphere).

3.2.1.2 Resting-state functional MRI data

The resting-state functional MRI data was acquired using a multi-echo acquisition protocol with three echo times TE = 13, 31, 48 ms, temporal resolution (TR) of 2.420 s, spatial resolution 3.8 mm isotropic with 10% gap, sequential slice acquisition, FOV = 240 × 240 mm, 34 oblique slices; bandwidth 1/4 2,368 Hz/pixel and matrix size = 64 × 64 × 34.

The data was preprocessed using multi-echo ICA (Kundu et al., 2013, 2015). Multi-echo ICA identifies BOLD components that scale linearly with TE and discards remaining components to reduce motion-related artefacts. Only BOLD components were optimally combined to generate the denoised time-series of each voxel. A wavelet filtering was used to focus on the physiologically relevant frequency range of 0.025-0.111 Hz (scales 2 and 3). Wavelet-based methods have shown significant advantages in terms of signal preservation and denoising over other filtering approaches (e.g., bandpass filtering), but the choice of method, filter type and length should be considered carefully as they might have an impact on the reliability of the estimates (Zhang et al., 2016). Functional scans were coregistered with each individual's structural R1 image for time-series extraction. Regional time-series were estimated as

the average time-series of all the voxels included in each of the 360 cortical and 16 subcortical regions. 28 regions (mostly near the frontal and temporal pole) were excluded due to low regional mean signal in at least one subject (z-score across regions within subject, $z < -1.96$), resulting in a total of 348 retained regions. Functional connectivity was calculated as the pairwise Pearson-correlation (also known as full correlation) between time-series of each possible pair of regions, resulting in a total of 60,378 brain connectivity features per subject ($\mathbf{x}_n^{(1)} \in \mathbb{R}^{60378\times 1}$ for each subject n). The vectors of the 306 subjects were concatenated to form the brain connectivity matrix ($\mathbf{X}^{(1)} \in \mathbb{R}^{60378\times 306}$).

3.2.1.3 Non-imaging features

Subjects completed self-report questionnaires and cognitive tests as part of the NSPN data acquisition (Kiddle et al., 2018). We used a subset of these features that assess psychopathological symptoms, personality characteristics, mental well-being and IQ: Antisocial Behaviours Checklist; Antisocial Process Screening Device; Barratt Impulsivity Scale; Child and Adolescent Dispositions Scale; Child Trauma Questionnaire; Drugs Alcohol and Self-Injury; Inventory of Callous-Unemotional Traits; Kessler Psychological Distress Scale; Leyton Obsessional Inventory; Moods and Feelings Questionnaire; Revised Children's Manifest Anxiety Scale; Rosenberg Self-Esteem Scale; Schizotypal Personality Questionnaire; Wechsler Abbreviated Scale of Intelligence; Warwick Edinburgh Mental Wellbeing Scale. A description of each questionnaire can be found in Appendix A.1.1. We removed eight items for which more than 95% of the subjects had the same value. Finally, we added four demographic features (age, sex, and socioeconomic deprivation index), resulting in a total of 364 non-imaging features per subject ($\mathbf{x}_n^{(2)} \in \mathbb{R}^{364\times 1}$). The vectors of the 306 subjects were concatenated to form the non-imaging matrix ($\mathbf{X}^{(2)} \in \mathbb{R}^{364\times 306}$). The missing data was imputed by the median of the respective feature across subjects.

3.2.2 Additional data preprocessing

We identified two main confounding variables, which were regressed out from both brain and non-imaging data: mean frame-wise displacement, which is a summary statistic quantifying average subject head motion during the resting-state functional MRI acquisitions (Power et al., 2012), and site. Finally, we standardised each non-imaging and brain connectivity feature to have zero mean and unit variance before applying CCA.

3.2.3 CCA experiments

We apply CCA to the NSPN dataset to explore the associations between patterns of brain connectivity and the non-imaging features. As mentioned in Chapter 2, CCA cannot be applied to high dimensional data sets without using regularisation or dimensionality reduction techniques. Here, we first reduced the dimensionality of the data using PCA. In summary, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ were first decomposed into $\mathbf{X}_d^{(1)} \in \mathbb{R}^{d \times N}$ and $\mathbf{X}_d^{(2)} \in \mathbb{R}^{d \times N}$, where N is the number of subjects and d represents the number of principal components. The optimal d was chosen using permutation tests (Section 3.2.4.1) from a set of 9 different values of $d = \{5, 10, 25, 50, 75, 100, 125, 150, 200\}$.

After this step, $\mathbf{X}_d^{(1)}$ and $\mathbf{X}_d^{(2)}$ are fed into CCA, which outputs d CCA modes. Each CCA mode is represented by a pair of weight vectors $\mathbf{u} \in \mathbb{R}^{D_1 \times 1}$ (where D_1 is the number of brain connectivity features) and $\mathbf{v} \in \mathbb{R}^{D_2 \times 1}$ (where D_2 is the number of non-imaging features), which indicate the direction of maximum brain-behaviour correlation; as well as a pair of canonical variates $\mathbf{P}_{\mathbf{X}^{(1)}} \in \mathbb{R}^{N \times 1}$ and $\mathbf{P}_{\mathbf{X}^{(2)}} \in \mathbb{R}^{N \times 1}$ obtained by projecting $\mathbf{X}_d^{(1)}$ and $\mathbf{X}_d^{(2)}$ onto $\mathbf{u}$ and $\mathbf{v}$, respectively. The correlation between $\mathbf{P}_{\mathbf{X}^{(1)}}$ and $\mathbf{P}_{\mathbf{X}^{(1)}}$ corresponds to the canonical correlation.

The statistical significance of the CCA modes was assessed using permutation tests on the whole data set, described in Section 3.2.4.1. To find the brain connectivity features and non-imaging features most strongly associated with the CCA modes, we correlated $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ with $\mathbf{P}_{\mathbf{X}^{(1)}}$ and $\mathbf{P}_{\mathbf{X}^{(2)}}$, respectively (similarly to Smith et al. (2015)). Alternatively, $\mathbf{P}_{\mathbf{X}^{(1)}}$ and $\mathbf{P}_{\mathbf{X}^{(2)}}$ could be first averaged (which corresponds to a shared latent component of probabilistic CCA, as explained in Section 2.3.1) to reduce potential within-view overfitting. For illustration purposes, we selected the 20 most (positively and negatively) correlated brain connectivity features (Figures 3.3-3.4) and non-imaging features (Figure 3.2) with the statistically significant CCA modes.

3.2.4 Learning frameworks

Here, we describe the statistical (Section 3.2.4.1) and machine learning frameworks (Section 3.2.4.2) used in this study to optimise the number of principal components and assess the statistical significance of the CCA modes.

3.2.4.1 Statistical framework

We used permutation tests on the whole data set to optimise the number of principal components and calculate a corrected p-value to assess the significance of each CCA mode. The algorithm proceeds as follows (Figure A.1):

1. For a given number of principal components (e.g., $d = 5$), the reduction step is performed on $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ resulting in the reduced data matrices $\mathbf{X}_5^{(1)} \in \mathbb{R}^{5\times N}$ and $\mathbf{X}_5^{(2)} \in \mathbb{R}^{5\times N}$. Then, these are fed into CCA to compute a vector of "true" canonical correlations $\mathbf{q} \in \mathbb{R}^{5\times 1}$.
2. The columns of $\mathbf{X}_5^{(2)}$ are permuted ($\mathbf{X}_5^{(2)^*}$). CCA is run again with $\mathbf{X}_5^{(1)}$ and $\mathbf{X}_5^{(2)^*}$ and a vector with "permuted" canonical correlations $\mathbf{q}^* \in \mathbb{R}^{5\times 1}$ is obtained. This procedure is repeated 10,000 times, resulting in a matrix of "permuted" canonical correlations $\mathbf{Q}^* \in \mathbb{R}^{5\times 10000}$.
3. For each row $i = 1, \ldots, 5$ of $\mathbf{Q}^*$, a p-value is computed as the fraction of permuted canonical correlations (in row i) exceeding the first "true" canonical correlation (the canonical correlations are ordered), which is equivalent to a maximum statistics approach. At the end of this procedure, a vector of p-values $\mathbf{p} \in \mathbb{R}^{5\times 1}$ is obtained, one per CCA mode. This allows one to estimate the number of significant CCA modes (i.e. any CCA mode with $p < 0.05$ is considered statistically significant). The p-value of the first CCA component (i.e. the first element of $\mathbf{p}$) is used to choose the optimal number of PCA components.
4. Steps 1-3 are repeated for other number of principal components ($d = \{10, 25, 50, 75, 100, 125, 150, 200\}$)

The obtained p-value of each d is corrected for multiple comparisons using a Bonferroni correction (i.e. $\alpha = \frac{0.05}{9} = 0.0056$), which means that only those p-values smaller than or equal to α ($p_{\text{corr}} \leq 0.0056$) are considered statistically significant. The optimal number of PCA components is chosen based on the lowest p_{corr}.

3.2.4.2 Machine learning framework

We used a multiple hold-out framework to choose the optimal number of PCA components and assess the statistical significance of the CCA modes based on permutation tests on the held-out sets. This was implemented based on the framework proposed by Monteiro et al. (2016) and it is similar to the framework displayed in Figures 4.1 and 4.2:

1. The data matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are randomly split into an optimisation set (80% of the data), $\mathbf{X}_{\text{op}}^{(1)}$ and $\mathbf{X}_{\text{op}}^{(2)}$, and a hold-out/test set (20% of the data), $\mathbf{X}_{\text{ho}}^{(1)}$ and $\mathbf{X}_{\text{ho}}^{(2)}$.
2. For each d principal components, $\mathbf{X}_{\text{op}}^{(1)}$ and $\mathbf{X}_{\text{op}}^{(2)}$ are randomly split 50 times into a training set (80% of the optimisation set), $\mathbf{X}_{\text{op}_{\text{tr}}}^{(1)}$ and $\mathbf{X}_{\text{op}_{\text{tr}}}^{(2)}$,

and a validation set (20% of the optimisation set), $\mathbf{X}^{(1)}_{\text{op}_{\text{val}}}$ and $\mathbf{X}^{(2)}_{\text{op}_{\text{val}}}$. For each random split, PCA and CCA are applied to obtain the combined PCA+CCA vectors $\mathbf{U}_{\text{tr}} \in \mathbb{R}^{D_1 \times d}$ and $\mathbf{V}_{\text{tr}} \in \mathbb{R}^{D_1 \times d}$. The canonical correlations on the validation sets $\mathbf{Q}_{\text{val}} \in \mathbb{R}^{d \times 50}$ (each row corresponds to a CCA mode) are computed between the projections $\mathbf{X}^{(1)}_{\text{op}_{\text{val}}}\mathbf{U}_{\text{tr}}$ and $\mathbf{X}^{(2)}_{\text{op}_{\text{val}}}\mathbf{V}_{\text{tr}}$. The optimal number of principal components d^* is chosen based on the maximal averaged canonical correlation of the first CCA mode across the validation sets.

3. PCA (using d^* components) and CCA are applied to $\mathbf{X}^{(1)}_{\text{op}}$ and $\mathbf{X}^{(2)}_{\text{op}}$ to compute the combined PCA+CCA vectors, $\mathbf{U}_{\text{op}}$ and $\mathbf{V}_{\text{op}}$. The "true" hold-out canonical correlations, $\mathbf{q}_{\text{ho}} \in \mathbb{R}^{d^* \times 1}$, are computed between the projections $\mathbf{X}^{(1)}_{\text{ho}}\mathbf{U}_{\text{op}}$ and $\mathbf{X}^{(2)}_{\text{ho}}\mathbf{V}_{\text{op}}$.

4. For assessing the statistical significance of the CCA modes, the columns of $\mathbf{X}^{(2)}_{\text{op}}$ (after applying PCA with d^* components) are permuted and CCA is applied to compute the "permuted" weight vectors $\mathbf{U}^*_{\text{op}}$ and $\mathbf{V}^*_{\text{op}}$. The "permuted" hold-out canonical correlations $\mathbf{q}^*_{\text{ho}} \in \mathbb{R}^{d^* \times 1}$ are calculated between $\mathbf{X}^{(1)}_{\text{ho}}\mathbf{U}^*_{\text{op}}$ and $\mathbf{X}^{(2)^*}_{\text{ho}}\mathbf{V}^*_{\text{op}}$. The permutation approach is repeated 10,000 times resulting in a matrix of "permuted" hold-out canonical correlations $\mathbf{Q}^*_{\text{ho}} \in \mathbb{R}^{d^* \times 10000}$. For CCA mode k, a p-value is computed as the fraction of permuted canonical correlations (kth-row of $\mathbf{Q}^*_{\text{ho}}$) exceeding the maximal "true" hold-out canonical correlation (the correlations are not ordered in the hold-out set). At the end of this procedure, a vector of p-values is obtained ($p \in \mathbb{R}^{d^* \times 1}$, where any CCA mode with $p < 0.05$ is considered statistically significant).

5. Steps 1-4 are repeated 9 more times (i.e., 10 different hold-out sets in total). The obtained p-values for each hold-out set are corrected for multiple comparisons using Bonferroni correction (i.e. $\alpha = \frac{0.05}{10} = 0.005$), which means that only the hold-out sets with a $p_{\text{corr}} \leq 0.005$ are considered statistically significant. Finally, the best hold-out set is chosen based on the lowest p_{corr}. As analyses from different dimensionalities can be correlated, this approach might be over-conservative. Alternatively, we could calculate a p-value through pooled permutation testing to potentially obtain greater sensitivity.

3.3 Results

We found two significant modes of covariation (Figure 3.1) between patterns of functional brain connectivity and sets of non-imaging features (or behavioural features, for simplicity) using the statistical framework. The optimal number of principal components obtained was 25 ($d = 25$), which explained 53% and 56% of the non-imaging features and brain connectivity features variance, respectively. The first and second CCA modes yielded canonical correlations of $q = 0.62$, $p < 0.0001$ (the mean null canonical correlation $q_{\text{null}} = 0.52$) and $q = 0.58$, $p < 0.0134$ (the mean null canonical correlation $q_{\text{null}} = 0.48$), respectively.

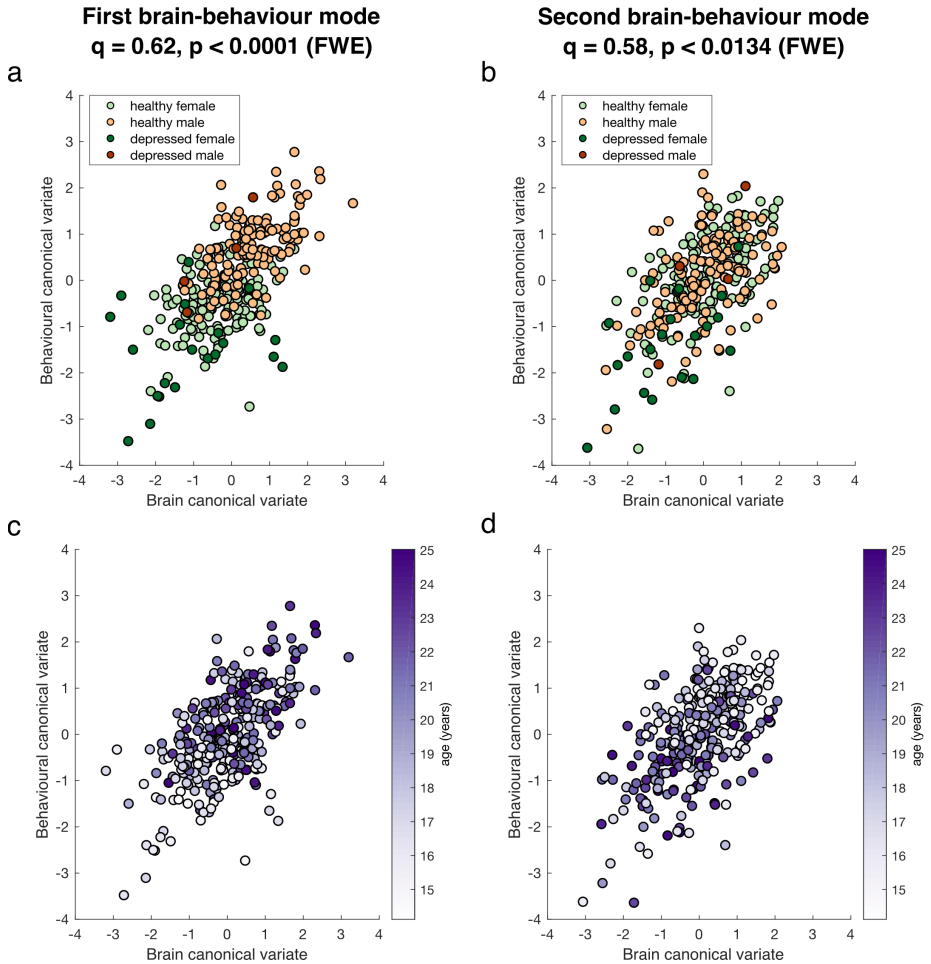

Figure 3.1: Significant brain-behaviour modes of covariation. Scatter plots showing the brain and behaviour scores for the first (**a** and **c**) and second (**b** and **d**) modes, where each dot represents an individual subject. Subjects are colour coded by: sex and clinical diagnosis (**a** and **b**); age (**c** and **d**). The canonical correlation, q, and the corresponding p-value are shown on the top of each plot.

Figure 3.1 shows the two significant brain-behaviour modes of covariation, representing the correlation between brain and behaviour scores of individual subjects. The first mode was associated with sex and has an interaction with depression, with healthy males clustering towards higher scores and depressed females clustering towards lower scores (Figure 3.1a). Additionally, younger adolescents can be seen to have lower scores, whereas older ones were distributed more towards higher scores (Figure 3.1c). The characteristics of the second mode were qualitatively different. Although depressed females seem to cluster towards lower scores (Figure 3.1b) again, both males and females were evenly distributed along this mode. Moreover, younger adolescents presented higher scores, whereas older ones were more distributed towards lower scores (Figure 3.1d).

To interpret the association captured by each mode, we correlated the brain connectivity and behavioural features with the brain and behaviour scores, respectively. Figure 3.2a shows that the first CCA mode was positively associated with age, being male, measures of impulsivity, sensation seeking, drinking habits, and negatively associated with being female, depression-related symptoms and suicidal thoughts. Thus, the first mode has characteristic of an externalization/internalization axis, where extreme positive and negative scores represent vulnerability for males and females, respectively. Importantly, sex was weakly associated with the other top identified non-imaging features suggesting that these are present due to an association with brain connectivity and not because of their association with sex (Figure A.2). The brain connections most positively correlated with the first CCA mode (denoted by red edges in Figure 3.3) included regions within the dorsal and ventral attention networks and somatomotor network; brain connections most negatively correlated (denoted by blue edges in Figure 3.3) included nodes of the default mode, limbic and frontoparietal networks. Similar overall patterns were observed using different thresholds on the top connections (Figure A.3). In addition, when looking at the 0.5% most negatively correlated connections (top 302 connections), the subcortical network (mostly thalamus and caudate nucleus) also appeared negatively correlated with the first mode (including subcortical-subcortical connections and cortical connections with the default mode network, Figure A.3). The list of the 20 brain connections most positively/negatively associated with the first mode and their assignment to anatomical regions are displayed on Figure A.4.

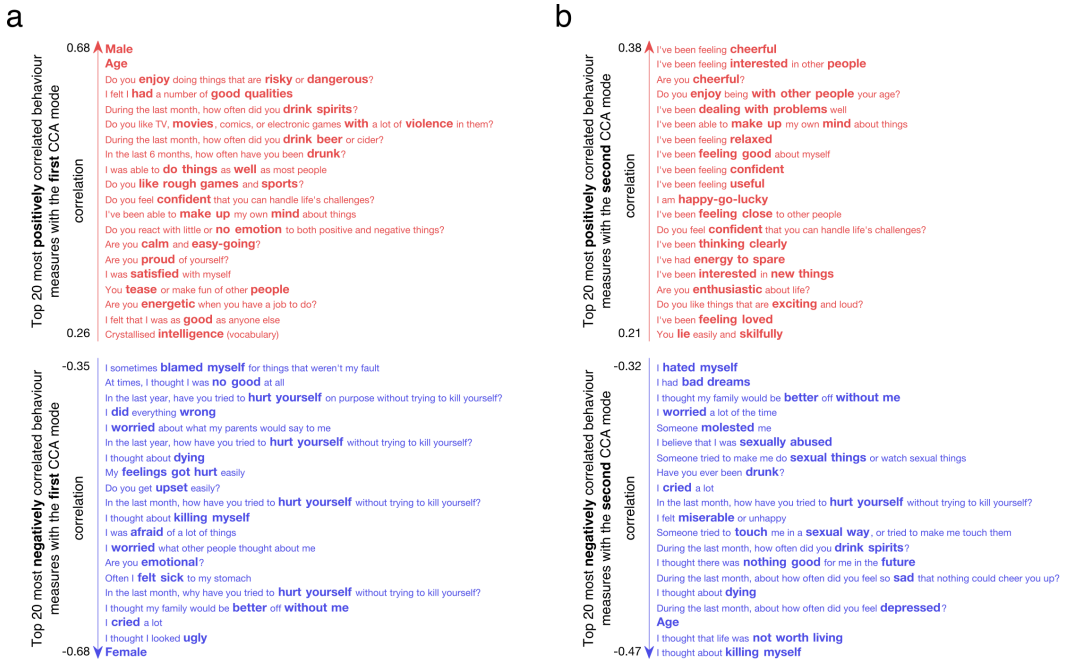

Figure 3.2: Correlations between the behavioural features and the behavioural canonical variate of the first (**a**) and second (**b**) CCA modes. The top 20 most positively and top 20 most negatively correlated features are shown.

For the second mode, the most positively correlated behavioural features (Figure 3.2b) related to measures of mental well-being, self-esteem and confidence, while the most negatively associated related to age, depression-related symptoms, drinking habits, suicidal thoughts and sexual abuse. Thus, this second mode captures a well-being/distress axis, along which individuals vary from high mental well-being to distress. The brain connections most positively correlated (depicted in red edges in Figure 3.4) with this CCA mode included nodes involving mainly the default mode and subcortical networks (thalamus); brain connections most negatively correlated (depicted in blue edges in Figure 3.4) included nodes within the dorsal and ventral attention networks and the visual and somatomotor networks. A largely similar overall pattern of networks was observed using different thresholds on the top connections (Figure A.3). In addition, when looking at the 0.5% most positively correlated connections (top 302 connections), the limbic and frontoparietal networks also appeared positively correlated with the second mode (including cortico-cortical connections and subcortical connections mostly with the thalamus, putamen and accumbens nucleus, Figure A.3). The top 20 most positively/negatively brain connections associated with the second mode and their assignment to

anatomical regions are displayed on Figure A.5.

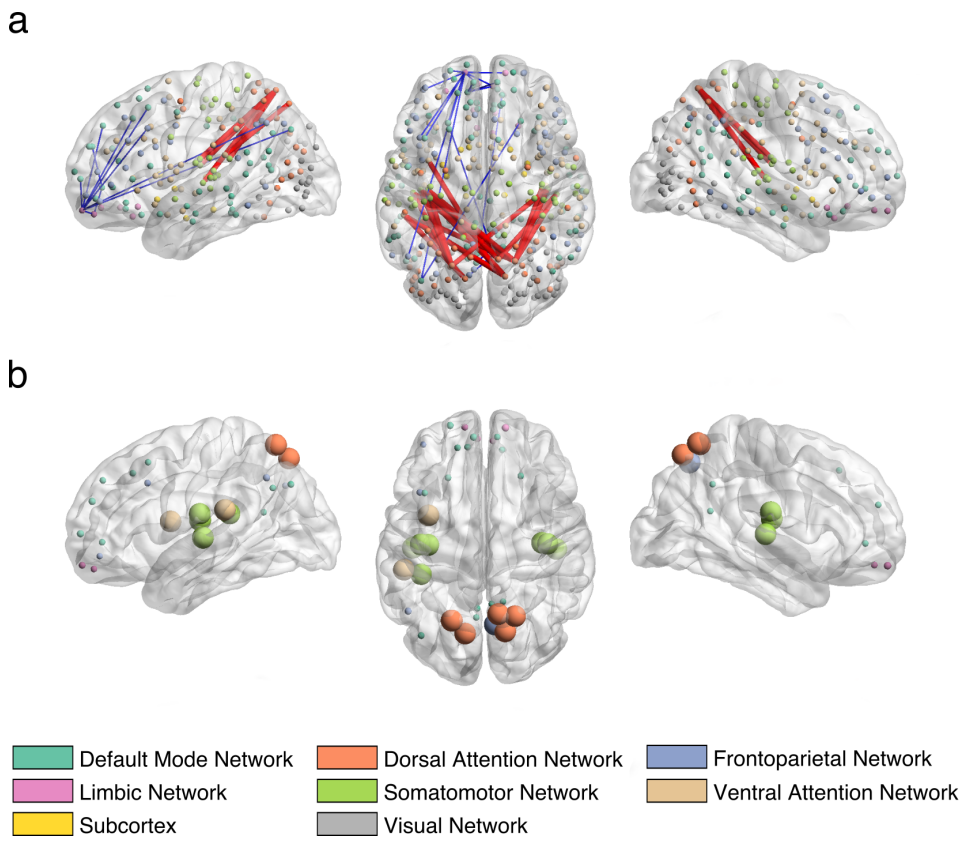

Figure 3.3: Correlations between the brain connectivity features and the brain canonical variate of the first CCA mode in sagittal (left and right) and axial views (middle). (**a**) Top 20 most positively and top 20 most negatively correlated brain connections. The thickness of the edges is proportional to the absolute correlation (red for positive correlations and blue for negative correlations). (**b**) Top 20 most positively and top 20 most negatively correlated brain connections, summarised by nodes. The node size is proportional to the mean absolute correlation. Nodes are colour coded by resting state networks, assigning each node to one of the seven cortical networks (based on the maximal surface based intersection) described in Thomas Yeo et al. (2011) or the subcortex.

Additionally, we applied a multiple hold-out framework (Section 3.2.4.2) and obtained one brain-behaviour mode of covariation ($q_{\text{ho}} = 0.46$ ($p < 0.0008$) and the mean null canonical correlation $q_{\text{null}} = 0.00016$) (Figure A.6). The optimal number of principal components was ten ($d = 10$), which explained 40% and 47% of the behaviour and brain connectivity variance, respectively. Importantly, the distribution of subjects along the CCA main axis showed the same trend in the training and test sets (Figure A.7). The overall ranking of the brain connectivity and behavioural features was similar to those

obtained using the statistical framework described in Section 3.2.4.1 (Figure A.8). Although the overlap was not large when only the top 20 most positively/negatively correlated behavioural and brain features were considered, it was more pronounced when the top 5% most positively/negatively correlated features were selected (Figure A.8). This might be explained by the fact that the correlations of the very top features only differ from each other on the fourth decimal place.

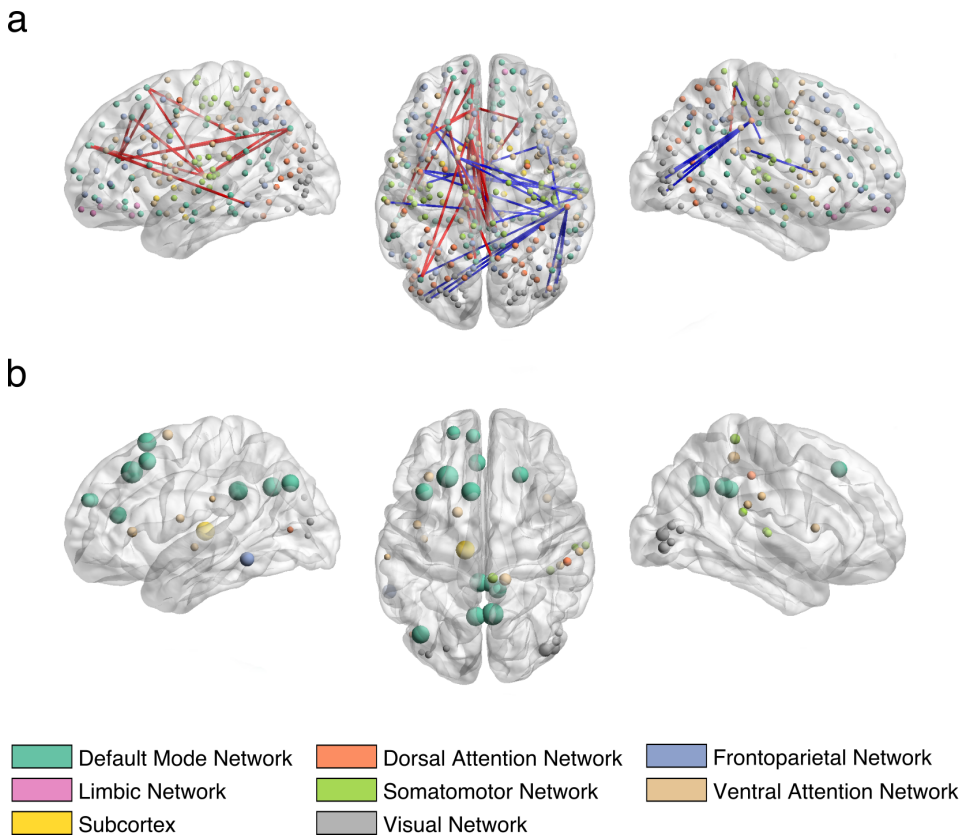

Figure 3.4: Correlations between the brain connectivity features and the brain canonical variate of the second CCA mode in sagittal (left and right) and axial views (middle). (**a**) Top 20 most positively and top 20 most negatively correlated brain connections. The thickness of the edges is proportional to the absolute correlation (red for positive correlations and blue for negative correlations). (**b**) Top 20 most positively and top 20 most negatively correlated brain connections, summarised by nodes. The node size is proportional to the mean absolute correlation. The nodes were coloured as explained in Figure 3.3.

3.4 Discussion and Conclusion

In summary, leveraging from both resting-state fMRI and non-imaging features (e.g., demographics, psychometrics and other behavioural features, which we

also refer to as behavioural features for simplicity) within a multivariate analysis framework, we identified two brain-behaviour modes of covariation in a sample of 306 adolescents and young adults. The first CCA mode related to an externalization/internalization axis, which was associated with sex. Specifically, it suggests that males might be more susceptible to disruptive behaviour and alcohol use, whilst females might be more susceptible to depression and self-harm. The second CCA mode related to a well-being/distress axis which covered positive symptoms of well-being on one side and negative symptoms related to depression, suicidal thoughts, history of sexual abuse and alcohol use on the other side. Both modes were also associated with age, which could be expected considering that the sample age range covers an important developmental period. Importantly, the brain networks related to both CCA modes align well with models of brain development that highlight the sequential maturation of subcortical and cortical regions in adolescence (Casey et al., 2008; Casey, 2015) and models of psychopathology (Buckholtz and Meyer-Lindenberg, 2012; Menon, 2011).

Both CCA modes were conceptually associated with broadly described depressive psychopathology, and can hence be seen as helping refine this clinical concept. It is therefore important to understand whether they capture distinctions in brain connectivity profiles alone or in descriptive psychopathology. At first glance, the behavioural features common to both modes of depression, such as "*...life was not worth living*", "*I thought about dying*", "*I cried a lot*" seem to support the former hypothesis. Nevertheless, there are three clear differences:

First, the first mode was associated with a more anxious, agitated and behaviourally-activated expression of depression (four self-harm measures, "*I felt sick...*", "*I worried...*", "*I was afraid...*", "*Are you emotional?*"). Conversely, the second mode was associated with a more anhedonic and a motivational state (negatively correlated with "*...life was not worth living*", "*...nothing good for me in the future*", "*...feel so sad...*", and positively correlated with "*...feeling interested in other people*"). Interestingly, similar "anxious" and "anhedonic" axes have been found in other large data-driven depression studies (Drysdale et al., 2017; Chekroud et al., 2017).

Second, the first CCA mode was strongly correlated with sex, but the second mode was not. Thus, the latter was a more sex independent dimension of psychopathology. Additionally, the depression-related features of the first mode were associated with younger age, whilst depression-related features of

the second mode were associated with older age (Figure 3.1c-d and Figure 3.2). Accordingly, depression in the first CCA mode was related to behavioural features, such as "*...I looked ugly*", "*...my family would be better off without me*", "*I worried about what my parents would say...*", which are more likely to be hallmarks of depression at a younger age. On the contrary, distress in the second CCA mode was related to measures more likely to characterise depression at an older age (e.g., "*I thought about killing myself*", being drunk and drinking spirits).

Third, depression in the second mode was associated with sexual abuse and was negatively associated with feeling loved, confident and close to other people, perhaps indicating that sexual abuse affects these traits (although causal attributions are not possible in this dataset).

The strong association between sex and the first CCA mode is striking in light of recent findings that there is $< 10\%$ overlap in gene expression changes in the brains of male and female humans with depression – at least in the prefrontal cortex and insula (other cortical areas were not sampled) (Labonté et al., 2017). Moreover, the authors demonstrated that a similar lack of overlap between the sexes also exists in a chronic variable stress mouse model (Labonté et al., 2017). It is interesting that both insula and the prefrontal cortex dominate the connections of the first CCA mode, being either positively (insula) or negatively (prefrontal cortex) correlated with depression. This suggests that sex interacts with depression risk in these (and likely other) areas in a way that might be fundamental to the disorder.

Adolescence and early adulthood is the peak age of onset for many psychiatric disorders (Kessler et al., 2007; Paus et al., 2008), therefore understanding the vulnerability of individuals at this age is of particular relevance. Importantly, most measures correlated with the CCA modes were related to psychopathology, and so the identified CCA modes might represent a two-dimensional space not only related to current depressive symptoms (or their absence), but to a latent vulnerability to psychopathology. Deeper understanding of this vulnerability may powerfully inform biologically informed interventions in young people (Lee et al., 2014).

Substance use is highly correlated with psychiatric disorders (Alterman, 1985; Brent, 1989), and it is especially detrimental in adolescence. Personality traits have an etiological role in the development of alcohol and substance use, and a vast body of research implicates two broad personality domains with opposing action tendencies, namely inhibition and disinhibition (Castellanos-

Ryan and Conrod, 2012; Carver et al., 2000). Our results concur with such a model. Alcohol usage was associated with both of our CCA modes in opposing directions. Behavioural features resembling to a disinhibited personality (first CCA mode) were positive correlations with, for instance, "*... enjoy doing things that are risky and dangerous?*", "*... like TV, movies, comics, or electronic games with a lot of violence in them?*" or "*Do you like rough games and sports?*"; whereas measures suggestive of an inhibited personality (second CCA mode) are negative correlations with, for instance, being interested in or enjoying the company of other people, or being interested in new things.

As discussed above, age was associated with both CCA modes. The first CCA mode correlated positively with age (depicted in red in Figure 3.2), attentional and frontoparietal networks (depicted in red in Figure A.3) and negatively with subcortical-subcortical connections as well as connections within the limbic system (depicted in blue in Figure A.3). These results are consistent with models of adolescent brain development, demonstrating that subcortical and limbic regions mature in early adolescence followed by the maturation of cortico-cortical connections (Casey et al., 2008; Casey, 2015). The second CCA mode was negatively correlated with age, connections within and between attentional networks (depicted in blue in Figure A.4) and was positively correlated with various subcortical-cortical connections (depicted in red in Figure A.4). Again, these results corroborate the aforementioned models of adolescent brain development. In particular, the results of the two CCA modes substantiate the sequential maturation of brain circuits, namely, the fine-tuning of circuits from subcortical-subcortical (early adolescence) to cortico-subcortical (late adolescence) and cortico-cortical (young adulthood) (Casey et al., 2016). Furthermore, the sequential maturation of brain circuits might be a risk factor for alcohol use (Spear, 2018), which aligns well with the strong positive correlation between alcohol use and age found in both CCA modes (Figure 3.2).

Our brain connectivity results were also consistent with recent literature suggesting that most psychiatric disorders emerge as a result of impairments within a few core brain circuits and networks (Xia et al., 2018; Buckholtz and Meyer-Lindenberg, 2012; Menon, 2011). In particular, the first mode was negatively correlated with depression and connections of the default mode, frontoparietal and limbic networks (Figure 3.3); whilst the second mode was negatively correlated with depression and positively correlated with many default mode areas (Figure 3.4). These networks underlie core social, executive and affective cognition, respectively, and dysfunctions in these networks

might result in specific domains of symptoms (e.g. alterations in default mode network connectivity resulting in impaired self-representation and social functioning) (Buckholtz and Meyer-Lindenberg, 2012). Interestingly, due to the strong interplay between these networks, the aberrant functioning in any of these could cause impairments of the others. For example, excessive coupling between the limbic and default mode networks could mean that initial dysfunction in the former may propagate to the latter, causing depressive symptoms (Xia et al., 2018; Berman et al., 2011; Cooney et al., 2010). Conversely, a default mode network that can only dominate but cannot reciprocally communicate with the limbic network could prevent positive mood being established by the latter (Admon and Pizzagalli, 2015).

In this study, we also applied CCA embedded in a multiple hold-out framework (Section 3.2.4.2) which was proposed by Monteiro et al. (2016). We found one mode of covariation, which was comparable to the first mode obtained using the statistical framework (Figure A.6). The second mode was not found with the hold-out framework, potentially due to the small sample size and the strictness of the framework. The most striking finding obtained with this approach was that the distribution of the subjects along the CCA main mode on the test set was very similar to the training set (Figure A.7), which means that the CCA mode generalised well on the test set.

Finally, we acknowledge limitations to the current study. Methodological limitations relate to the pipeline choice, which includes use of an atlas and full correlation as a connectivity metric. Full correlation is a robust and fast method, but when the number of nodes is large (here, >300 brain regions) the biological interpretability of a connection between two nodes might be poor because this method does not take into account the indirect connections between the nodes (Smith et al., 2013). Further work exploring other approaches to parcellate the data (e.g., independent component analysis (Smith et al., 2015)) and measures to estimate the resting state connectivity between nodes (such as, partial correlation (Smith et al., 2011), although it has not shown improved performances when the HCP atlas is used (Sala-Llonch et al., 2019)). Moreover, although the PCA dimensionality reduction step is needed to apply CCA to high dimensional data and avoid overfitting, it might also remove a significant amount of signal variability of potential interest if the number of principal components are not chosen carefully. Regularised variants of CCA could be investigated in future work to overcome potential limitations of the current pipeline. In addition, although we have used a multiple hold-out

framework, we should ideally use an independent replication sample to validate our model.

In conclusion, our results showed that identifying brain-behaviour modes of covariation in healthy and depressed young people provide a better understanding of the latent dimensions of abnormal mental states and behaviour (Insel et al., 2010), and brings new insights into the mediation of vulnerability to mental disorders.

Chapter 4

Hyperparameter optimisation in sparse CCA

The content of this chapter is based on a comparison study published in the *2018 International Workshop on Pattern Recognition in Neuroimaging* (Ferreira et al., 2018). I have rewritten some sections and changed the figures to be consistent with the nomenclature and notation used in this thesis. In this study, I compared two different frameworks to optimise the regularisation parameters (hyperparameters) of sparse CCA.

4.1 Introduction

As mentioned in the previous chapters, CCA is a multi-view method which has been widely used in neuroimaging to investigate associations between different types of data. However, neuroimaging datasets are typically high dimensional and include only a few hundred subjects, which prevents CCA from being applied to these datasets. As described in Chapter 2, regularised versions of CCA, such as sparse CCA, have been proposed to address this issue by adding regularisation terms to penalise the norm of the weight vectors (Lê Cao et al., 2008; Parkhomenko et al., 2009; Waaijenborg et al., 2008; Witten et al., 2009). L_1 and L_2-norm penalties are included in the sparse CCA optimisation problem to simultaneously regularise and impose sparsity on the weight vectors. These changes make the learning feasible for high dimensional datasets and improve the interpretability of the results by allowing feature selection. Moreover, the L_1 penalties help to solve the issue of arbitrary rotations that CCA suffers by constraining the model to converge to a unique solution. However, each L_1-norm penalty has a parameter that controls the degree of sparsity, which affects the number of features selected in each view, and therefore should be

carefully optimised.

In this study, I compared two frameworks to optimise the sparse CCA hyperparameters: a statistical framework proposed by Witten and Tibshirani (2009) and a machine learning framework proposed by Monteiro et al. (2016). In the former, permutation tests are applied on the whole training set and the hyperparameters are chosen based on the lowest p-value. In the latter, the training set is divided multiple times into training and validation sets and the hyperparameters are chosen to maximise the averaged correlation across the validation sets. The two approaches were compared in terms of the features selected in each view, and the generalisability of the sparse CCA modes using a hold-out framework.

4.2 Methods

4.2.1 Data

In this study, I used the dataset described in Section 3.2.1. In summary, the dataset comprises resting-state fMRI and extensive item-level questionnaire data and demographics (which I will refer to as non-imaging features) of 306 (281 healthy and 25 depressed) participants (adolescents and young adults: 14-24 years old) from NSPN study (Kiddle et al., 2018). For a more detailed description of the dataset and data preprocessing steps, see Section 3.2.1. All experiments were run using two data modalities: brain functional connectivity (represented by $\mathbf{X}^{(1)} \in \mathbb{R}^{D_1 \times N}$) and non-imaging features (represented by $\mathbf{X}^{(2)} \in \mathbb{R}^{D_2 \times N}$), where N is the number of subjects, D_1 is the number of brain connectivity features and D_2 is the number of non-imaging features.

4.2.2 Sparse CCA

Sparse CCA finds sparse weight vectors such that the covariance between the projections of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ onto these vectors is maximised:

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T \mathbf{X}^{(1)} \mathbf{X}^{(2)T} \mathbf{v}, \\ &\text{s.t. } \|\mathbf{u}\|_2^2 \leq 1, \|\mathbf{v}\|_2^2 \leq 1, \|\mathbf{u}\|_1 \leq c_u, \|\mathbf{v}\|_1 \leq c_v, \end{aligned} \tag{4.1}$$

The regularisation parameters c_u and c_v (Equation 4.1) control the L_1-norm penalties of $\mathbf{u} \in \mathbb{R}^{D_1 \times 1}$ and $\mathbf{v} \in \mathbb{R}^{D_2 \times 1}$, respectively. If c_u and c_v are sufficiently small, the L_1-norm penalties impose sparsity on the corresponding view and consequently fewer features are included in the model. The values of c_u and c_v must be chosen in $1 \leq c_u \leq \sqrt{D_1}$ and $1 \leq c_v \leq \sqrt{D_2}$ for both L_1

and L_2-norms to be active (Witten et al., 2009). For more details, see Section 2.2.3.

4.2.3 Learning framework

The learning framework consists of three parts: hyperparameter optimisation, statistical inference of the associations and matrix deflation. To evaluate the generalisability of each hyperparameter optimisation, the data matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are randomly split into an optimisation set, $\mathbf{X}^{(1)}_{\text{op}}$ and $\mathbf{X}^{(2)}_{\text{op}}$ (80% of the data) and a hold-out set, $\mathbf{X}^{(1)}_{\text{ho}}$ and $\mathbf{X}^{(2)}_{\text{ho}}$ (20% of the data). The former is used for optimising the hyperparameters, and the latter is used for assessing the generalisability of the associations. All ρs mentioned in this section represent Pearson's correlations.

4.2.3.1 Hyperparameter optimisation

The hyperparameters were optimised using a grid-search, in which 20 equidistant points in $1 \leq c_u \leq \sqrt{D_1}$ and $1 \leq c_v \leq \sqrt{D_2}$ were defined.

Statistical framework

For each c_u and c_v pair, the weight vectors $\mathbf{u}$ and $\mathbf{v}$ are computed using the optimisation set. Then, the correlation ρ between the projections $\mathbf{X}^{(1)T}_{\text{op}}\mathbf{u}$ and $\mathbf{X}^{(2)T}_{\text{op}}\mathbf{v}$ ($\rho = \text{corr}(\mathbf{X}^{(1)T}_{\text{op}}\mathbf{u}, \mathbf{X}^{(2)T}_{\text{op}}\mathbf{v})$) is computed (Figure 4.1a).

The rows of $\mathbf{X}^{(2)}_{\text{op}}$ are then randomly permuted to obtain $\mathbf{X}^{(2)^b}_{\text{op}}$ (where $b = 1, \ldots, B$). For each data permutation b, the weight vectors $\mathbf{u}^b$ and $\mathbf{v}^b$ are computed and the correlations ρ^b between $\mathbf{X}^{(1)T}_{\text{op}}\mathbf{u}^b$ and $\mathbf{X}^{(2)^bT}_{\text{op}}\mathbf{v}^b$ are computed (Figure 4.1a). The procedure is repeated $B = 1000$ times in total, and the p-value for ρ is calculated as follows (Witten and Tibshirani, 2009):

$$p = \frac{1 + \sum_{b=1}^{B} 1_{\ \rho^b \geq \rho}}{B + 1} \tag{4.2}$$

The hyperparameter pair with the lowest p-value ($p < 0.001$) is chosen. However, it is likely that several combinations have the same p-value and then a second criterion needs to be used: the hyperparameter pair with the largest distance between the true correlation and the null distribution of the "permuted" correlations ($d = \frac{\rho - \frac{1}{B}\sum_{b=1}^{B}\rho^b}{sd(\rho^B)}$, where $sd(\rho^B)$ indicates the standard deviation of $\rho^1, \ldots, \rho^B$) is chosen (Witten and Tibshirani, 2009). The best hyperparameter pair is finally passed for use in the statistical inference step

(Figure 4.2).

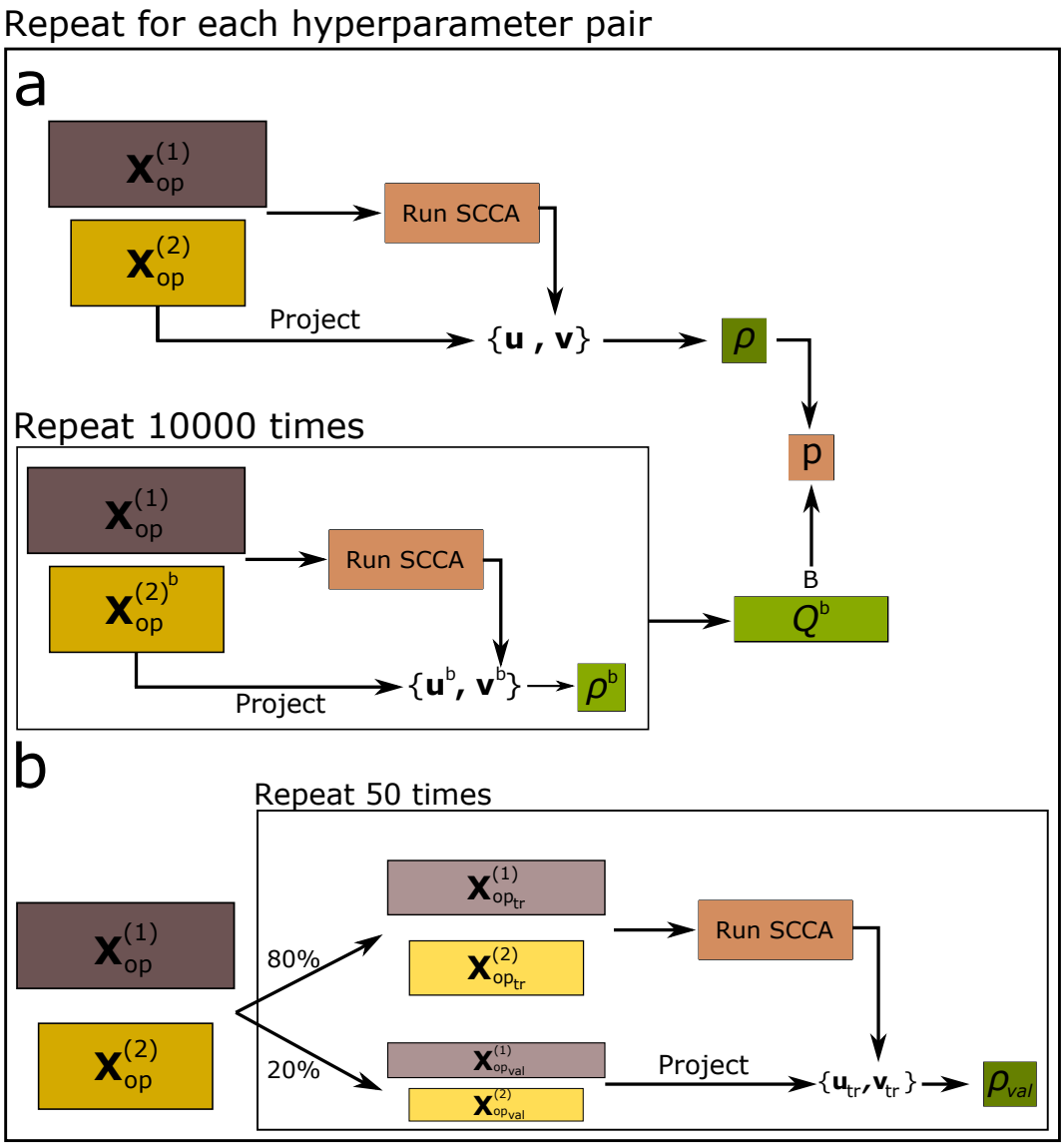

Figure 4.1: Hyperparameter optimisation step for (a) statistical and (b) machine learning framework.

Machine learning framework

For each c_u and c_v pair, the optimisation set is randomly split 50 times into a training (80%) and test sets (20%). For each split, the weight vectors $\mathbf{u}_{\text{tr}}$ and $\mathbf{v}_{\text{tr}}$ are computed using the training set ($\mathbf{X}^{(1)}_{\text{op}_{\text{tr}}}$ and $\mathbf{X}^{(2)}_{\text{op}_{\text{tr}}}$) and the canonical correlation is computed by projecting the validation set ($\mathbf{X}^{(1)}_{\text{op}_{\text{val}}}$ and $\mathbf{X}^{(2)}_{\text{op}_{\text{val}}}$) onto these weight vectors ($\rho_{\text{val}} = \text{corr}(\mathbf{X}^{(1)T}_{\text{op}_{\text{val}}}\mathbf{u}_{\text{tr}}, \mathbf{X}^{(2)T}_{\text{op}_{\text{val}}}\mathbf{v}_{\text{tr}})$ (Figure 4.1b). The

50 test correlations are averaged across splits to obtain $\bar{\rho}_{\text{val}}$, for each hyperparameter pair. The hyperparameter pair with the highest $\bar{\rho}_{\text{val}}$ is selected. If multiple hyperparameter combinations have the same $\bar{\rho}_{\text{val}}$, the sparsest pair among those is chosen (Monteiro et al., 2016). The best hyperparameter pair is then passed for use in the statistical significance step (Figure 4.2).

4.2.3.2 Statistical inference

The statistical significance of the sparse CCA modes is assessed using multiple hold-out sets (Figure 4.2). First, the model is trained with the best hyperparameter pair using the optimisation set to compute $\mathbf{u}_{\text{op}}$ and $\mathbf{v}_{\text{op}}$. Second, the hold-out set ($\mathbf{X}_{\text{ho}}^{(1)}$ and $\mathbf{X}_{\text{ho}}^{(2)}$) is projected onto $\mathbf{u}_{\text{op}}$ and $\mathbf{v}_{\text{op}}$ and the hold-out correlation is calculated ($\rho_{\text{ho}} = \text{corr}(\mathbf{X}_{\text{ho}}^{(1)T}\mathbf{u}_{\text{op}}, \mathbf{X}_{\text{ho}}^{(2)T}\mathbf{v}_{\text{op}})$). Third, the rows of $\mathbf{X}_{\text{ho}}^{(2)}$ are permuted to obtain $\mathbf{X}_{\text{ho}}^{(2)^m}$ (where $m = 1, \ldots, M$). For each permutation m, the model is trained with the best hyperparameter pair and the hold-out set is projected onto $\mathbf{u}_{\text{op}}^m$ and $\mathbf{v}_{\text{op}}^m$ (Figure). The hold-out correlation between the projections is calculated as follows: $\rho_{\text{ho}}^m = \text{corr}(\mathbf{X}_{\text{ho}}^{(1)T}\mathbf{u}_{\text{op}}^m, \mathbf{X}_{\text{ho}}^{(2)^m T}\mathbf{v}_{\text{op}}^m)$. The process is run $M = 10000$ times and a p-value for ρ_{ho} is computed using Equation 4.2.

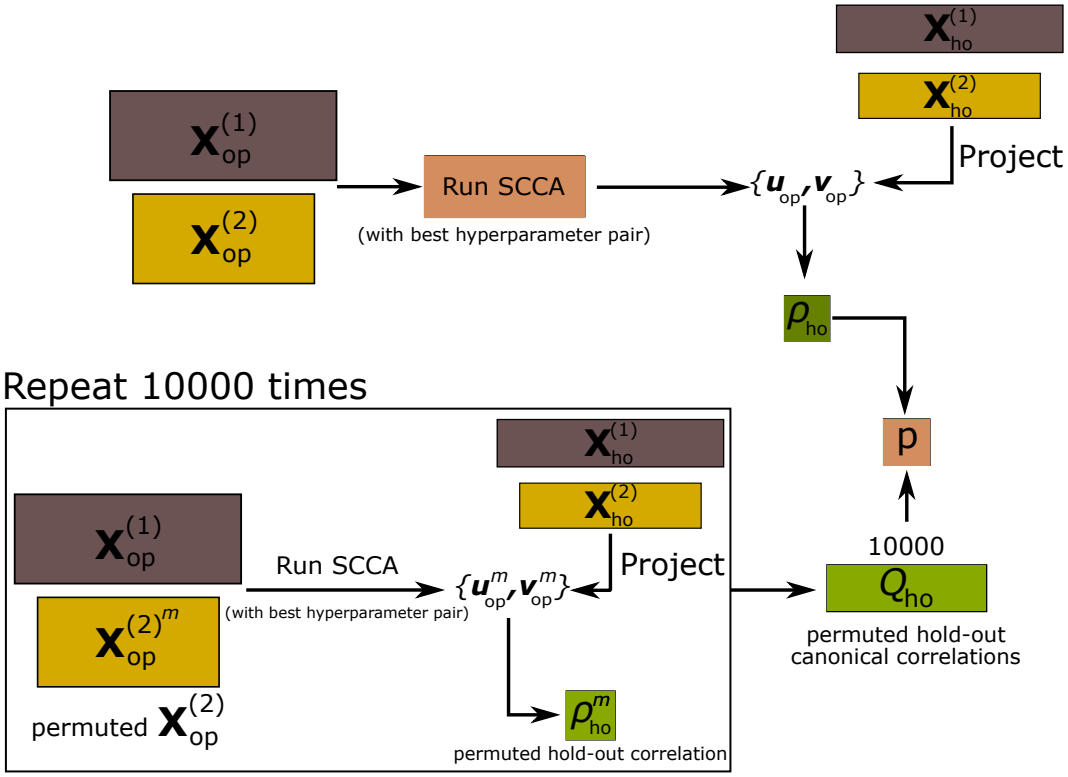

Figure 4.2: Statistical significance evaluation step.

In neuroimaging datasets, the sample sizes are usually small and therefore

few samples may be included in the hold-out set, which can lead to unstable findings since the validation is dependent on how the data is split. To make the model validation more robust, multiple hold-out sets (here ten random splits of the data into optimisation and hold-out sets) are used. As multiple hold-out sets were used, a criterion is needed to determine if any of the sparse CCA modes were statistically significant. Here, we assume the *ominbus* hypothesis, which is a statistical test to assess multiple hypotheses n_h at the same time by assuming that all of them are true. The *ominbus* hypothesis is rejected if any of the hypothesis $H_i,\ i = 1, ..., n_h$ is rejected, which means, in this case, that if any of the hold-out sets is considered statistically significant, the *ominbus* hypothesis is rejected. The p-values are corrected for multiple comparisons using a Bonferroni correction, which is a technique that correct for multiple comparisons when several statistical tests are performed simultaneously by setting the significance level α (usually 0.05) to $\alpha = \alpha/N_t$ (where N_t is the number of tests). A sparse CCA mode is considered statistically significant, if any hold-out set is considered statistically significant: $p_{\text{corr}} \leq 0.05/10$. The sparse CCA mode (among those statistically significant) with the lowest p-value is chosen to deflate the data matrices.

4.2.3.3 Matrix deflation

If any of the weight vector pairs is considered statistically significant, the brain-behaviour association explained by those weights is removed from the data for allowing new associations to be found. This process is known as *matrix deflation*. Here, I used the projection deflation method proposed by Mackey (2008) and tested in Monteiro et al. (2016) for sparse CCA (see Equation 2.7).

4.3 Results

The frameworks were compared in terms of the brain-behaviour associations identified and the generalisability of these, measured by hold-out correlation.

4.3.1 Brain-behaviour associations

Three statistically significant brain-behaviour associations were obtained using the statistical framework ($q_1 = 0.60$ ($p < 0.0001$), $q_2 = 0.51$ ($p < 0.0001$) and $q_3 = 0.33$ ($p < 0.0047$), and the corresponding mean null canonical correlations $q_1^* = -0.003$, $q_2^* = 0.009$ and $q_3^* = 0.003$, respectively), and only one was obtained using the machine learning framework ($q_1 = 0.60$ ($p < 0.0001$) and the mean null canonical correlation $q_1 = -0.001$). Figure 4.3 shows the non-imaging and brain weights of the first sparse CCA mode obtained using

both frameworks. Figure 4.4 shows the weights of the second and third sparse CCA modes identified by the statistical framework. For visualisation purposes, only the top 20 brain connectivity features associated with the modes (the features with the highest absolute weights) are shown. The first brain and non-imaging (NI) weight vectors were similar across frameworks ($\rho_{\text{brain}} = 0.70$ and $\rho_{\text{NI}} = 0.98$) as can be seen in Figure 4.3. Moreover, the distribution of the weights was similar across frameworks (Figure 4.5).

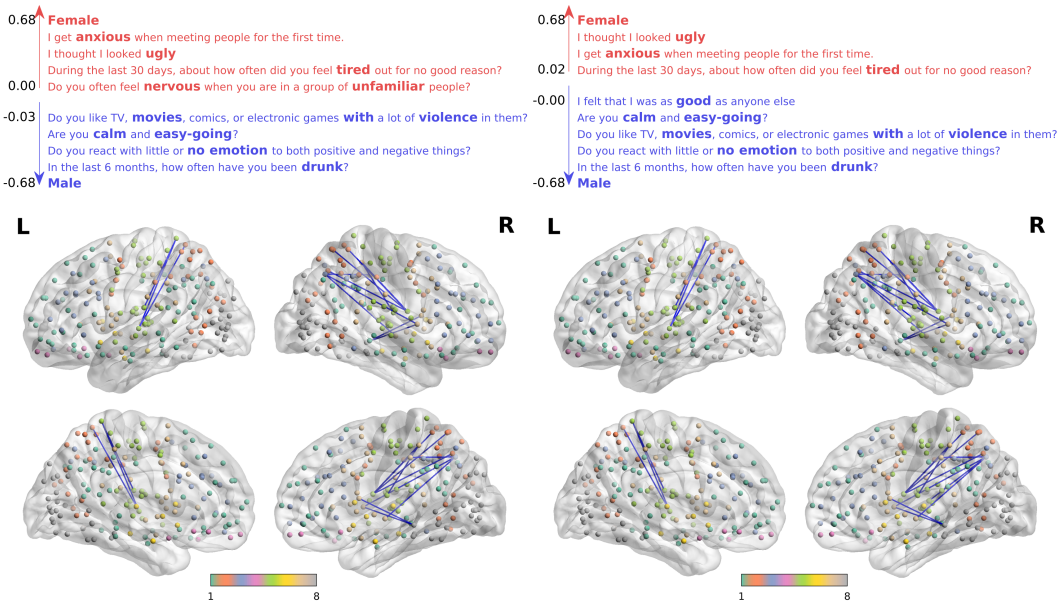

Figure 4.3: All non-zero non-imaging features **(top)** and top 20 brain connectivity features **(bottom)** associated with the first sparse CCA mode, obtained using the statistical **(left)** and machine learning **(right)** frameworks. L - left hemisphere; R - right hemisphere.

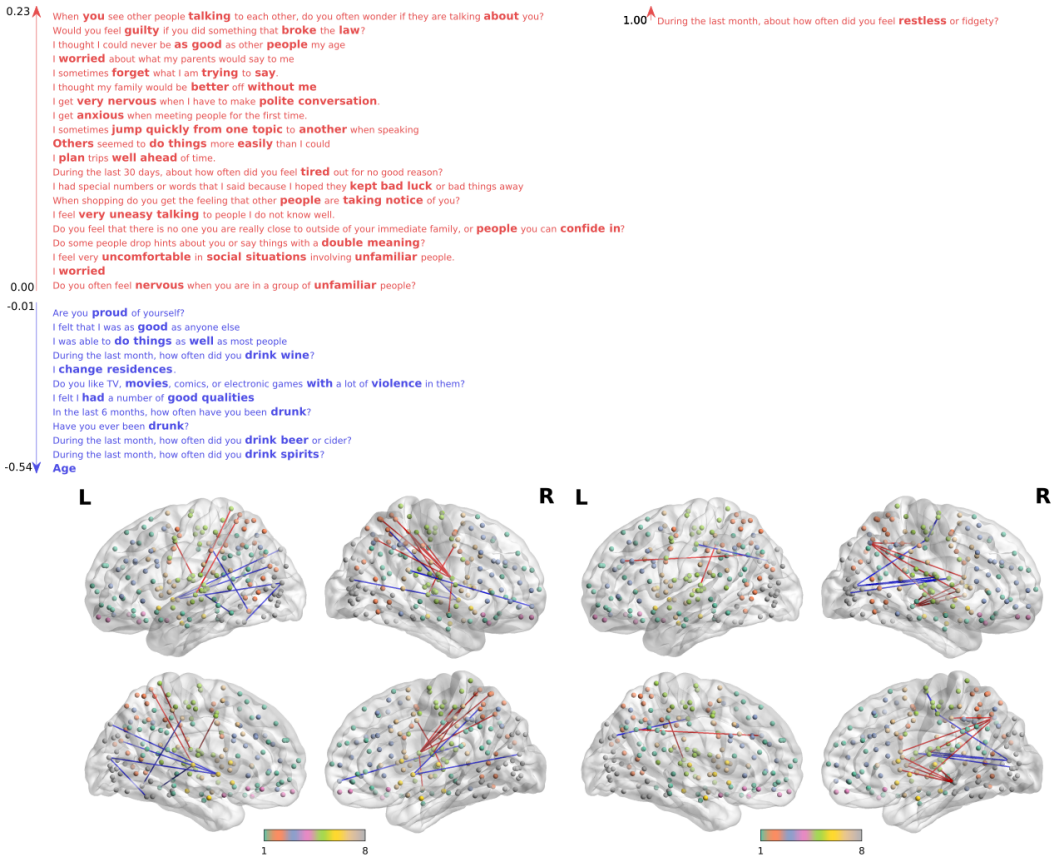

Figure 4.4: Non-imaging features **(top)** and top 20 brain connectivity features **(bottom)** associated with the second **(left)** and third **(right)** sparse CCA modes, obtained using the statistical framework. L - left hemisphere; R - right hemisphere.

4.3.2 Generalisability of the frameworks

Table 4.1 shows the hold-out correlations of the ten different splits of the data for the first and second sparse CCA mode, for each framework. The first statistically significant mode was similar across both frameworks. A second and third modes statistically significant were obtained using the statistical framework (Figure 4.4), but not with the machine learning framework.

Table 4.1: Hold-out correlations and p-values of the ten different splits of the data for the first and second sparse CCA modes obtained by each framework, and the third one for the statistical framework. The statistically significant splits are shown in bold.

	Machine learning framework		**Statistical framework**		
	First mode	Second mode	First mode	Second mode	Third mode
Split	ρ_{ho} (p_{ho})	ρ_{ho} (p_{ho})	ρ_{ho} (p_{ho})	ρ_{ho} (p_{ho})	ρ_{ho} (p_{ho})
1	**0.51 (0.0001)**	-0.03 (0.5906)	**0.51 (0.0001)**	0.11 (0.1940)	0.08 (0.2625)
2	0.33 (0.0052)	0.03 (0.4179)	**0.35 (0.0031)**	0.20 (0.0567)	0.24 (0.0289)
3	**0.45 (0.0005)**	0.11 (0.1900)	**0.45 (0.0002)**	0.20 (0.0576)	0.09 (0.2367)
4	0.27 (0.0161)	0.20 (0.0601)	**0.35 (0.0020)**	0.24 (0.0296)	0.22 (0.0406)
5	**0.36 (0.0030)**	0.02 (0.4349)	**0.51 (0.0001)**	**0.41 (0.0005)**	0.32 (0.0063)
6	**0.39 (0.0007)**	0.14 (0.1376)	**0.41 (0.0003)**	0.25 (0.0249)	0.21 (0.0487)
7	**0.43 (0.0003)**	0.22 (0.0436)	**0.45 (0.0002)**	0.29 (0.0135)	0.27 (0.0166)
8	**0.35 (0.0024)**	0.07 (0.3040)	**0.37 (0.0019)**	0.12 (0.1719)	0.11 (0.2070)
9	**0.60 (0.0001)**	0.21 (0.0499)	**0.60 (0.0001)**	**0.33 (0.0046)**	**0.33 (0.0047)**
10	**0.38 (0.0011)**	0.23 (0.0327)	**0.47 (0.0001)**	**0.35 (0.0034)**	0.30 (0.0091)

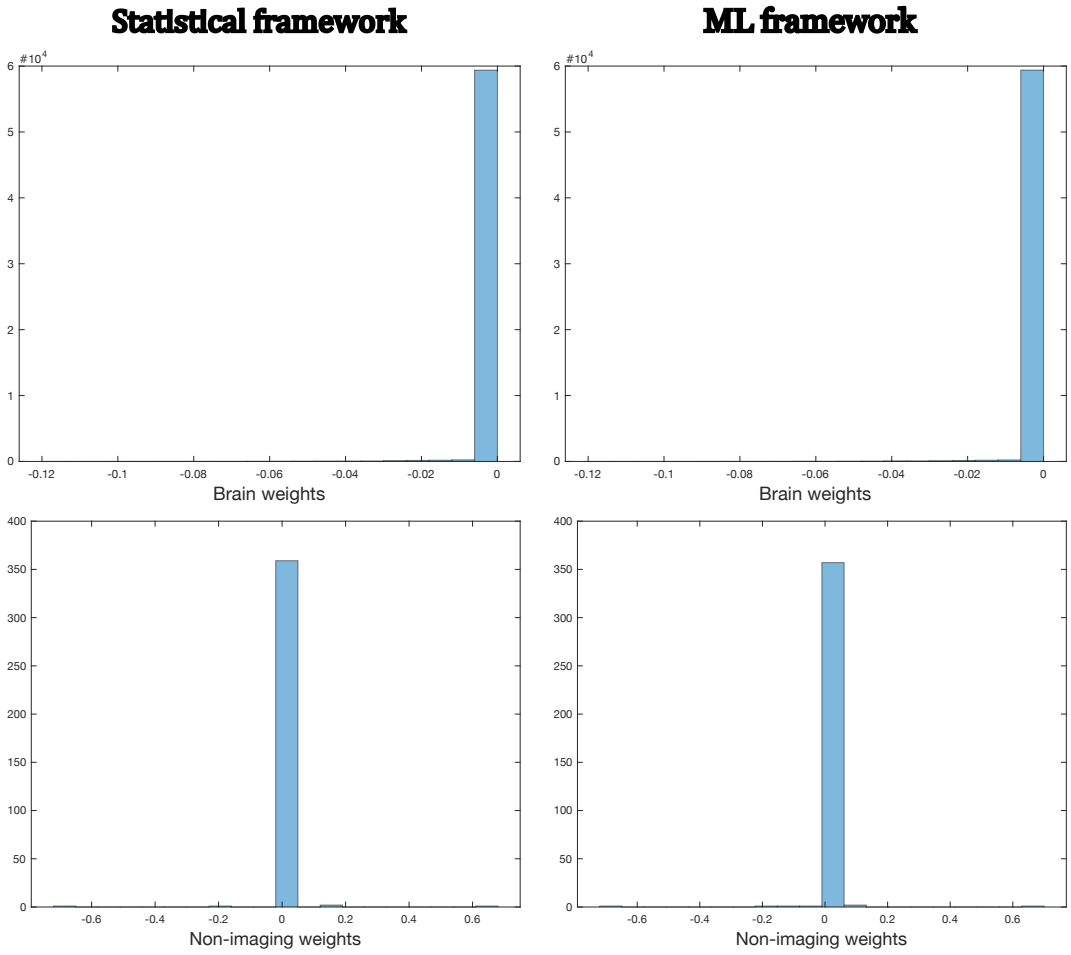

Figure 4.5: Distribution of the weights of the first sparse CCA mode obtained with the **(left)** statistical framework and **(right)** machine learning framework.

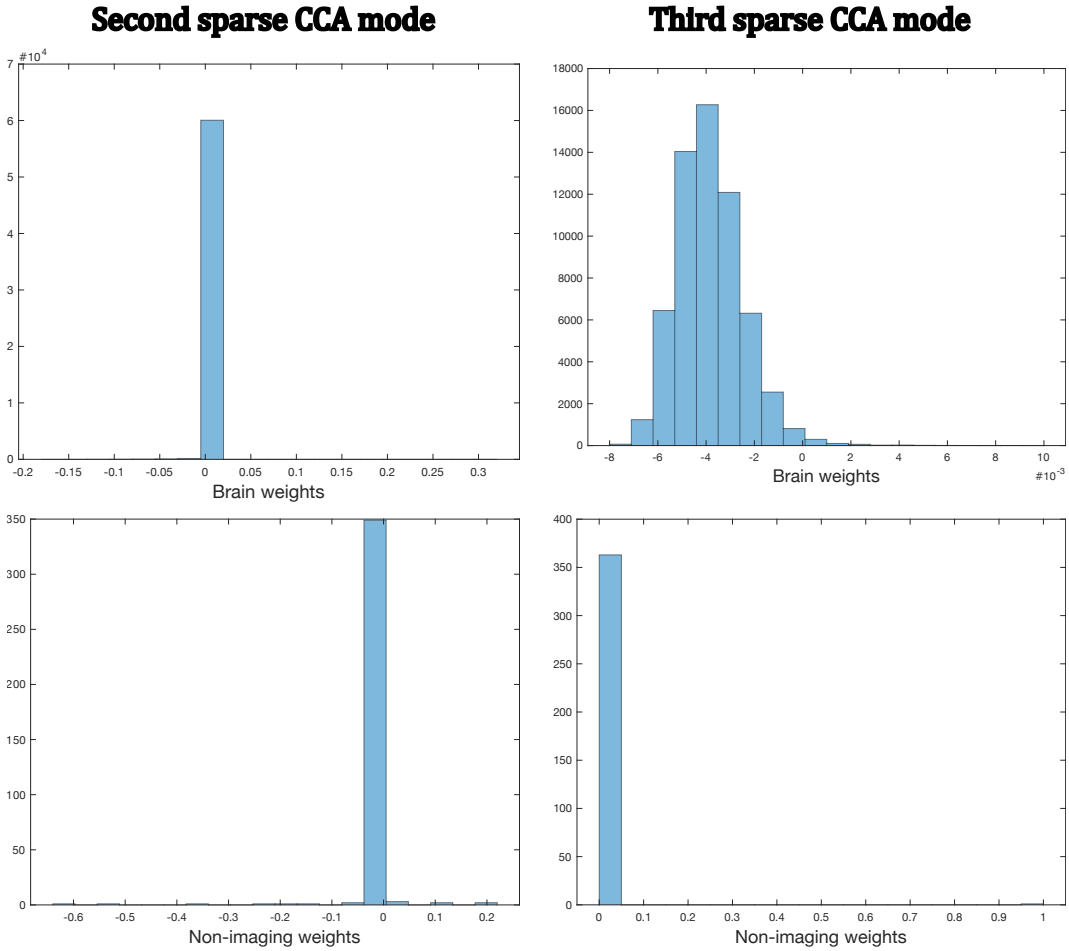

Figure 4.6: Distribution of the weights of the **(left)** second and **(right)** third sparse CCA modes obtained with the statistical framework.

4.4 Discussion

The statistical framework, proposed by Witten and Tibshirani (2009), uses p-values computed using permutation tests on the whole training set to optimise the sparse CCA hyperparameters, which can be seen as a limitation because no out-of-sample metric is used. Monteiro et al. (2016) proposed a machine learning framework to calculate out-of-sample correlations based on multiple validation sets. In this study, we compared both frameworks for optimising the sparse CCA hyperparameters using a hold-out framework to access the generalisability of the frameworks (based on hold-out correlation).

Both frameworks were able to identify at least one brain-behaviour association. The weight vectors obtained showed great similarity across frameworks (Figures 4.3 and 4.5), which indicates that both frameworks are able to generalise to different hold-out sets (Table 4.1). However, only the stat-

istical framework was able to find more brain-behaviour associations. On one hand, these results suggest that the statistical framework might be prone to overfitting and false positive findings. In fact, the third mode (Figure 4.4, right) is likely to be a false positive finding, because only one split of the data was rendered as statistically significant with a p-value close to the defined threshold ($p = 0.0047$, Table 4.1) and most of the brain features were included in the model (i.e., obtained non-zero weights, Figure 4.6). On the other hand, these results suggest that the machine learning framework might be a strict approach that may lead to false negative findings. Indeed, as we only considered the single most statistically significant held-out solution, the Bonferroni correction might be overconservative. Moreover, we might be losing power by not pooling the results before the statistical test. However, there is no trivial solution for this issue because the different splits of the data might obtain different sparse weight vectors. In terms of computational cost, the machine learning framework is much more efficient than the statistical framework (i.e. 200 times faster).

In summary, the sparse CCA hyperparameters should be carefully optimised, because different criteria and frameworks might have a strong influence on the results. As expected, optimising the hyperparameters using a metric based on test data (i.e., test correlation) leads to stricter approaches than those based on the whole data.

Chapter 5

Identifying brain-behaviour associations in incomplete data sets using GFA

The content of this chapter is based on the study recently submitted to *NeuroImage* (Ferreira et al., 2021). I have rewritten some sections to avoid repeating content presented in previous Chapters, and to be consistent with the nomenclature and notation used in this thesis. In this study, I extended Group Factor Analysis (GFA) (Section 2.3.3) to uncover associations among multiple data modalities and predict the features in one view (i.e., multi-output prediction) from other views observed in the test set (e.g., predict behavioural measures from brain functional connectivity) in incomplete data sets.

5.1 Introduction

As mentioned throughout the thesis, CCA and equivalent methods have been successfully applied to identify associations between two views (e.g., data modalities). Nonetheless, these methods have some limitations. First, they do not provide an inherently robust inference approach to infer the relevant associations. This is usually done by assessing the statistical significance of the associations using permutation tests on the whole data set (Smith et al., 2015; Winkler et al., 2020) or on hold-out sets (Monteiro et al., 2016; Mihalik et al., 2020). Second, the associations within data modalities, which might explain important variance 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 modalities. This is a common issue in clinical and neuroimaging datasets, in which the missing values usually need to be imputed or the

samples with missing values need to be removed before fitting the models.

In this chapter, I address these limitations by proposing an extension of GFA (described in Section 2.3.3) that can handle missing data. I first applied this extended GFA to synthetic data to assess whether it can find known associations among different views. I then applied it to data from the Human Connectome Project (HCP) (Van Essen et al., 2013) to uncover associations between brain connectivity and non-imaging features (e.g., demographics, psychometrics and other behavioural features). I evaluated the consistency of the findings across different experiments with complete and incomplete data sets. Finally, even though the GFA model was proposed as an unsupervised approach, it can also be used as a predictive model (Klami et al., 2015): I applied the proposed GFA implementation to synthetic and HCP data to assess whether it was able to predict missing data and non-observed data modalities from the observed ones in incomplete data sets.

To illustrate the differences between GFA and CCA, CCA was also applied to both datasets. First, I 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" component linking lifestyle, demographic and psychometric measures to specific patterns of brain connectivity (Smith et al., 2015). Second, I expected CCA to show poorer performance when data is missing, whereas GFA results would be more consistent across experiments with complete and incomplete data sets.

5.2 Materials and Methods

I first explain how the GFA model was modified to accommodate missing data (Section 5.2.1) and used to make predictions (Section 5.2.3). These subsections are followed by descriptions of experiments on synthetic data (Section 5.2.4), as well as on HCP data (Section 5.2.5).

5.2.1 Proposed GFA extension

To handle missing data, I modified the original GFA model (see Section 2.3.3) by assuming independent noise for each feature (i.e., diagonal noise) within a view:

$$p(\boldsymbol{\tau}) = \prod_{m=1}^{M} \prod_{j=1}^{D_m} \Gamma(\tau_j^{(m)} | a_{\boldsymbol{\tau}^{(m)}}, b_{\boldsymbol{\tau}^{(m)}}) \tag{5.1}$$

where $a_{\boldsymbol{\tau}^{(m)}}$ and $b_{\boldsymbol{\tau}^{(m)}}$ shape parameter and rate parameter of the gamma distribution, respectively. In addition, I modified the variational update rules similarly to what has been proposed by Luttinen and Ilin (2010) for variational Bayesian factor analysis. The derivations of the variational EM update rules are described in Section 5.2.2. The derivations of the lower bound can be found in Appendix B.1. Although I only present applications of GFA to two modalities/views in this thesis, the GFA extension can be applied to more than two data modalities (see our Python implementation: https://github.com/ferreirafabio80/gfa).

5.2.2 Variational updates of GFA

The variational updates of the model parameters are derived by writing the expectation of the log joint distribution $p(\mathbf{X}, \boldsymbol{\theta})$ with respect to all other variational parameters (Equation 2.16). Considering Equation 2.14, the log joint distribution is defined as follows:

$$\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = \ln \Big[p(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau}) p(\mathbf{Z}) p(\mathbf{W}|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\boldsymbol{\tau}) \Big] + \text{const} \tag{5.2}$$

where the individual log-densities (considering the priors in Equations 2.12 and 2.13) are given by:

$$\begin{aligned} \ln p(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau}) = \sum_{m=1}^{M} \Bigg[& \frac{N}{2} \sum_{j=1}^{D_m} (\ln \tau_j^{(m)} - \ln(2\pi)) \\ & - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n^{(m)} - \mathbf{W}^{(m)} \mathbf{z}_n)^T \mathbf{T}^{(m)} (\mathbf{x}_n^{(m)} - \mathbf{W}^{(m)} \mathbf{z}_n) \Bigg] \end{aligned} \tag{5.3}$$

$$\ln p(\mathbf{Z}) = -\frac{1}{2} \sum_{n=1}^{N} \mathbf{z}_n^T \mathbf{z}_n - \frac{NK}{2} \ln(2\pi) \tag{5.4}$$

$$\ln p(\mathbf{W}|\boldsymbol{\alpha}) = \sum_{m=1}^{M} \left[\frac{D_m}{2} \sum_{k=1}^{K} \ln \alpha_k^{(m)} - \frac{1}{2} \sum_{k=1}^{K} \alpha_k^{(m)} \mathbf{w}_k^{(m)T} \mathbf{w}_k^{(m)} + \frac{D_m K}{2} \ln(2\pi) \right] \tag{5.5}$$

$$\ln p(\boldsymbol{\alpha}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \left[a_{\boldsymbol{\alpha}^{(m)}} \ln b_{\boldsymbol{\alpha}^{(m)}} - \ln \Gamma(a_{\boldsymbol{\alpha}^{(m)}}) + (a_{\boldsymbol{\alpha}^{(m)}} - 1) \ln \alpha_k^{(m)} - b_{\boldsymbol{\alpha}^{(m)}} \alpha_k^{(m)} \right] \tag{5.6}$$

$$\ln p(\boldsymbol{\tau}) = \sum_{m=1}^{M} \sum_{j=1}^{D_m} \left[a_{\boldsymbol{\tau}^{(m)}} \ln b_{\boldsymbol{\tau}^{(m)}} - \ln \Gamma(a_{\boldsymbol{\tau}^{(m)}}) + (a_{\boldsymbol{\tau}^{(m)}} - 1) \ln \tau_j^{(m)} - b_{\boldsymbol{\tau}^{(m)}} \tau_j^{(m)} \right] \tag{5.7}$$

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 $a_{\boldsymbol{\alpha}^{(m)}}, b_{\boldsymbol{\alpha}^{(m)}}, a_{\tau^{(m)}}, b_{\tau^{(m)}}$ are the hyperparameters of the Gamma distributions in Equations 2.12-2.13.

5.2.2.1 $q(\mathbf{Z})$ distribution

The optimal log-density for $q(\mathbf{Z})$, given the other variational distributions is calculated using Equation 2.16:

$$\begin{aligned}
\ln q(\mathbf{Z}) &= \mathbb{E}_{q(\mathbf{W}),q(\boldsymbol{\tau})}[\ln p(\mathbf{X}|\mathbf{Z},\mathbf{W},\tau) + \ln p(\mathbf{Z})] \\
&= \sum_{n=1}^{N}\left[-\frac{1}{2}\sum_{m=1}^{M}\langle(\mathbf{x}_n^{(m)} - \mathbf{W}^{(m)}\mathbf{z}_n)^T\mathbf{T}^{(m)}(\mathbf{x}_n^{(m)} - \mathbf{W}^{(m)}\mathbf{z}_n)\rangle - \frac{1}{2}\mathbf{z}_n^T\mathbf{z}_n\right] \\
&= \sum_{n=1}^{N}\left[\mathbf{z}_n^T\sum_{m=1}^{M}\sum_{j\in O_n^{(m)}}\langle\tau_j^{(m)}\rangle\langle\mathbf{W}_{j,*}^{(m)^T}\rangle x_{j,n}^{(m)}\right. \\
&\quad \left. -\frac{1}{2}\mathbf{z}_n^T\left(\sum_{m=1}^{M}\sum_{j\in O_n^{(m)}}\langle\tau_j^{(m)}\rangle\langle\mathbf{W}_{j,*}^{(m)^T}\mathbf{W}_{j,*}^{(m)}\rangle\right)\mathbf{z}_n - \frac{1}{2}\mathbf{z}_n^T\mathbf{z}_n\right] \\
&= \sum_{n=1}^{N}\left[\mathbf{z}_n^T\sum_{m=1}^{M}\sum_{j\in O_n^{(m)}}\langle\tau_j^{(m)}\rangle\langle\mathbf{W}_{j,*}^{(m)^T}\rangle x_{j,n}^{(m)}\right. \\
&\quad \left. -\frac{1}{2}\mathbf{z}_n^T\left(\mathbf{I}_K + \sum_{m=1}^{M}\sum_{j\in O_n^{(m)}}\langle\tau_j^{(m)}\rangle\langle\mathbf{W}_{j,*}^{(m)^T}\mathbf{W}_{j,*}^{(m)}\rangle\right)\mathbf{z}_n\right]
\end{aligned} \tag{5.8}$$

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 Equation 5.22) 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 Equation 5.16). $O_n^{(m)}$ is the set of indices in the n-th column of $\mathbf{X}^{(m)}$ ($x_{(:,n)}^{(m)}$) that are not missing. Equation 5.8 omits any 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:

$$q(\mathbf{Z}) = \prod_{n=1}^{N} q(\mathbf{z}_n) = \prod_{n=1}^{N}\mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_{\mathbf{z}_n}, \boldsymbol{\Sigma}_{\mathbf{z}_n}) \tag{5.9}$$

The variational parameters for $q(\mathbf{Z})$ are:

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{z}_n} &= \left[\mathbf{I}_K + \sum_{m=1}^{M} \sum_{j \in O_n^{(m)}} \langle \tau_j^{(m)} \rangle \langle \mathbf{W}_{j,*}^{(m)^T} \mathbf{W}_{j,*}^{(m)} \rangle \right]^{-1} \\ \boldsymbol{\mu}_{\mathbf{z}_n} &= \boldsymbol{\Sigma}_{\mathbf{z}_n} \sum_{m=1}^{M} \sum_{j \in O_n^{(m)}} \langle \tau_j^{(m)} \rangle \langle \mathbf{W}_{j,*}^{(m)^T} \rangle x_{j,n}^{(m)}\end{aligned} \tag{5.10}$$

5.2.2.2 $q(\mathbf{W}^{(m)})$ distribution

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

$$\begin{aligned}\ln q(\mathbf{W}^{(m)}) &= \mathbb{E}_{q(\mathbf{Z}),q(\boldsymbol{\alpha}^{(m)}),q(\boldsymbol{\tau}^{(m)})}[\ln p(\mathbf{X}^{(m)}|\mathbf{Z}, \mathbf{W}^{(m)}, \boldsymbol{\tau}^{(m)}) + \ln p(\mathbf{W}^{(m)}|\boldsymbol{\alpha}^{(m)})] \\ &= -\frac{1}{2} \sum_{n=1}^{N} \langle (\mathbf{x}_n - \mathbf{W}^{(m)} \mathbf{z}_n)^T \mathbf{T}^{(m)} (\mathbf{x}_n - \mathbf{W}^{(m)} \mathbf{z}_n) \rangle \\ &\quad - \frac{1}{2} \sum_{k=1}^{K} \langle \alpha_k^{(m)} \mathbf{w}_k^{(m)T} \mathbf{w}_k^{(m)} \rangle\end{aligned} \tag{5.11}$$

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 Equation 5.11 can be expanded as follows:

$$\begin{aligned}-\frac{1}{2} \sum_{n=1}^{N} \langle (\mathbf{x}_n^{(m)} - \mathbf{W}^{(m)} \mathbf{z}_n)^T \mathbf{T}^{(m)} (\mathbf{x}_n^{(m)} - \mathbf{W}^{(m)} \mathbf{z}_n) \rangle &= \sum_{j=1}^{D_m} \langle \tau_j^{(m)} \rangle \left(\sum_{n \in O_j^{(m)}} x_{j,n}^{(m)} \langle \mathbf{z}_n^T \rangle \right) \\ \times \mathbf{W}_{j,*}^{(m)^T} + \sum_{j=1}^{D_m} -\frac{1}{2} \mathbf{W}_{j,*}^{(m)} \left(\langle \tau_j^{(m)} \rangle \sum_{n \in O_j^{(m)}} \langle \mathbf{z}_n \mathbf{z}_n^T \rangle \right) \mathbf{W}_{j,*}^{(m)^T}\end{aligned} \tag{5.12}$$

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 Equation 5.10) 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 Equation 5.11 is given by:

$$-\frac{1}{2} \sum_{k=1}^{K} \langle \alpha_k^{(m)} \mathbf{w}_k^{(m)T} \mathbf{w}_k^{(m)} \rangle = -\frac{1}{2} \sum_{j=1}^{D_m} \mathbf{W}_{j,*}^{(m)} \langle \mathbf{A}_\alpha^{(m)} \rangle \mathbf{W}_{j,*}^{(m)^T} \tag{5.13}$$

where $\langle \mathbf{A}_\alpha^{(m)} \rangle = \text{diag}(\langle \boldsymbol{\alpha}^{(m)} \rangle)$ 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 Equation 5.19). Putting both terms together,

we get:

$$\begin{aligned}\ln q(\mathbf{W}^{(m)}) = \sum_{j=1}^{D_m} \Bigg[\langle \tau_j^{(m)} \rangle \Bigg(\sum_{n \in O_j^{(m)}} x_{j,n}^{(m)} \langle \mathbf{z}_n^T \rangle \Bigg) \mathbf{W}_{j,*}^{(m)^T} \\ - \frac{1}{2} \mathbf{W}_{j,*}^{(m)} \Bigg(\langle \mathbf{A}_\alpha^{(m)} \rangle + \langle \tau_j^{(m)} \rangle \sum_{n \in O_j^{(m)}} \langle \mathbf{z}_n \mathbf{z}_n^T \rangle \Bigg) \mathbf{W}_{j,*}^{(m)^T} \Bigg]\end{aligned} \tag{5.14}$$

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

$$q(\mathbf{W}^{(m)}) = \prod_{j=1}^{D_m} q(\mathbf{W}_{j,*}^{(m)}) = \prod_{j=1}^{D_m} \mathcal{N}(\mathbf{W}_{j,*}^{(m)} | \boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}}, \boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}}) \tag{5.15}$$

Then the variational update rules for $q(\mathbf{W}^{(m)})$ are:

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}} &= \Bigg[\langle \mathbf{A}_\alpha^{(m)} \rangle + \langle \tau_j^{(m)} \rangle \sum_{n \in O_j^{(m)}} \langle \mathbf{z}_n \mathbf{z}_n^T \rangle \Bigg]^{-1} \\ \boldsymbol{\mu}_{\mathbf{W}_{j,*}^{(m)}} &= \langle \tau_j^{(m)} \rangle \sum_{n \in O_j^{(m)}} \Bigg(x_{j,n}^{(m)} \langle \mathbf{z}_n^T \rangle \Bigg) \boldsymbol{\Sigma}_{\mathbf{W}_{j,*}^{(m)}}\end{aligned} \tag{5.16}$$

5.2.2.3 $q(\boldsymbol{\alpha}^{(m)})$ distribution

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

$$\begin{aligned}\ln q(\boldsymbol{\alpha}^{(m)}) &= \mathbb{E}_{q(\mathbf{W}^{(m)})}[\ln p(\mathbf{W}^{(\mathbf{m})} | \alpha^{(\mathbf{m})}) + \ln p(\boldsymbol{\alpha}^{(m)})] \\ &= \sum_{k=1}^{K} \Bigg[\frac{D_m}{2} \ln \alpha_k^{(m)} - \frac{1}{2} \alpha_k^{(m)} \langle \mathbf{w}_k^{(m)^T} \mathbf{w}_k^{(m)} \rangle + (a_{\boldsymbol{\alpha}^{(m)}} - 1) \ln \alpha_k^{(m)} \\ &\quad - b_{\boldsymbol{\alpha}^{(m)}} \alpha_k^{(m)} \Bigg] \\ &= \sum_{k=1}^{K} \Bigg(\frac{D_m}{2} + a_{\boldsymbol{\alpha}^{(m)}} - 1 \Bigg) \ln \alpha_k^{(m)} - \sum_{k=1}^{K} \Bigg(b_{\boldsymbol{\alpha}^{(m)}} + \frac{1}{2} \langle \mathbf{w}_k^{(m)^T} \mathbf{w}_k^{(m)} \rangle \Bigg) \alpha_k^{(m)}\end{aligned} \tag{5.17}$$

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

distribution:

$$q(\boldsymbol{\alpha}^{(m)}) = \prod_{k=1}^{K} q(\boldsymbol{\alpha}_k^{(m)}) = \prod_{k=1}^{K} \Gamma(\alpha_k^{(m)} | \tilde{a}_{\boldsymbol{\alpha}^{(m)}}, \tilde{b}^{(k)}_{\boldsymbol{\alpha}^{(m)}}) \tag{5.18}$$

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

$$\begin{aligned} \tilde{a}_{\boldsymbol{\alpha}^{(m)}} &= a_{\boldsymbol{\alpha}^{(m)}} + \frac{1}{2} D_m \\ \tilde{b}^{(k)}_{\boldsymbol{\alpha}^{(m)}} &= b_{\boldsymbol{\alpha}^{(m)}} + \frac{1}{2} \langle \mathbf{w}_k^{(m)^T} \mathbf{w}_k^{(m)} \rangle \end{aligned} \tag{5.19}$$

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

5.2.2.4 $q(\boldsymbol{\tau}^{(m)})$ distribution

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

$$\begin{aligned} \ln q(\boldsymbol{\tau}^{(m)}) &= \mathbb{E}_{q(\mathbf{Z}), q(\mathbf{W}^{(m)})} [\ln p(\mathbf{X}^{(m)} | \mathbf{Z}, \mathbf{W}^{(m)}, \boldsymbol{\tau}^{(m)}) + \ln p(\boldsymbol{\tau}^{(m)})] \\ &= -\frac{1}{2} \sum_{j=1}^{D_m} \Bigg[\tau_j^{(m)} \sum_{n \in O_j^{(m)}} \Bigg(x_{j,n}^{(m)2} - 2 x_{j,n}^{(m)} \langle \mathbf{W}_{j,*}^{(m)} \rangle \langle \mathbf{z}_n \rangle \\ &\quad + \mathrm{Tr}[\langle \mathbf{W}_{j,*}^{(m)^T} \mathbf{W}_{j,*}^{(m)} \rangle \langle \mathbf{z}_n \mathbf{z}_n^T \rangle] \Bigg) + \frac{N_j^{(m)}}{2} \ln \tau_j^{(m)} \\ &\quad + (a_{\boldsymbol{\tau}^{(m)}} - 1) \ln \tau_j^{(m)} - b_{\boldsymbol{\tau}^{(m)}} \tau_j^{(m)} \Bigg] \\ &= \sum_{j=1}^{D_m} \left(a_{\boldsymbol{\tau}^{(m)}} + \frac{N_j^{(m)}}{2} - 1 \right) \ln \tau_j^{(m)} - \sum_{j=1}^{D_m} \Bigg(b_{\boldsymbol{\tau}^{(m)}} + \frac{1}{2} \sum_{n \in O_j^{(m)}} \Big[x_{j,n}^{(m)2} \\ &\quad - 2 x_{j,n}^{(m)} \langle \mathbf{W}_{j,*}^{(m)} \rangle \langle \mathbf{z}_n \rangle + \mathrm{Tr}[\langle \mathbf{W}_{j,*}^{(m)^T} \mathbf{W}_{j,*}^{(m)} \rangle \langle \mathbf{z}_n \mathbf{z}_n^T \rangle] \Big] \Bigg) \tau_j^{(m)} \end{aligned} \tag{5.20}$$

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

$$q(\boldsymbol{\tau}^{(m)}) = \prod_{j=1}^{D_m} q(\boldsymbol{\tau}_j^{(m)}) = \prod_{j=1}^{D_m} \Gamma(\tau_j^{(m)} | \tilde{a}^{(j)}_{\boldsymbol{\tau}^{(m)}}, \tilde{b}^{(j)}_{\boldsymbol{\tau}^{(m)}}) \tag{5.21}$$

where the variational parameters are:

$$\tilde{a}^{(j)}_{\boldsymbol{\tau}^{(m)}} = a_{\boldsymbol{\tau}^{(m)}} + \frac{1}{2} N^{(m)}_j$$
$$\tilde{b}^{(j)}_{\boldsymbol{\tau}^{(m)}} = b_{\boldsymbol{\tau}^{(m)}} + \frac{1}{2} \sum_{n \in O^{(m)}_j} x^{(m)2}_{j,n} - 2x^{(m)}_{j,n} \langle \mathbf{W}^{(m)}_{j,*} \rangle \langle \mathbf{z}_n \rangle + \mathrm{Tr}[\langle \mathbf{W}^{(m)T}_{j,*} \mathbf{W}^{(m)}_{j,*} \rangle \langle \mathbf{z}_n \mathbf{z}^T_n \rangle] \tag{5.22}$$

Finally, to solve the rotation and scaling ambiguity known to be present in factor analysis models, I used a similar approach previously proposed by Virtanen et al. (2011), which consists of maximising the variational lower bound with respect to a separate parameter matrix $\mathbf{R}$ (that is a linear transformation applied to $\mathbf{W}$), after each round of variational EM updates. As it is shown by Virtanen et al. (2011), maximising the lower bound with respect to $\mathbf{R}$ forces the model to find components that are maximally independent of each other, and provides a deterministic choice for the rotation. This not only solves the rotation and scaling ambiguity, but also improves convergence and speeds up the learning.

5.2.3 Multi-output and missing data prediction

As mentioned above, GFA can be used as a predictive model. As the views are generated by the same latent variables, the unobserved view 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 EM as follows (Klami et al., 2015):

$$\begin{aligned} \langle \mathbf{X}^{(m)\star} | \mathbf{X}^{-(m)\star} \rangle &= \langle \mathbf{W}^{(m)} \mathbf{Z} \rangle_{q(\mathbf{W}^{(m)}),\, q(\mathbf{Z}|\mathbf{X}^{-(m)\star})} \\ &= \langle \mathbf{W}^{(m)} \rangle \boldsymbol{\Sigma}^{\star}_Z \langle \mathbf{W}^{-(m)T} \rangle \mathbf{T}^{\star} \mathbf{X}^{-(m)\star} \end{aligned} \tag{5.23}$$

where $\langle \cdot \rangle$ denotes expectations, $\boldsymbol{\Sigma}^{\star}_Z = \mathbf{I}_K + \sum_{l \neq m} \sum_j^{D_l} \langle \tau^{(l)}_j \rangle \langle \mathbf{W}^{(l)T}_{j,*} \mathbf{W}^{(l)}_{j,*} \rangle$, $\langle \mathbf{W}^{(l)T}_{j,*} \mathbf{W}^{(l)}_{j,*} \rangle = \boldsymbol{\Sigma}_{\mathbf{w}^{(l)}_j} + \boldsymbol{\mu}^T_{\mathbf{W}^{(l)}_{j,*}} \boldsymbol{\mu}_{\mathbf{W}^{(m)}_{j,*}}$ and $\mathbf{T}^{\star} = \{\mathrm{diag}(\langle \boldsymbol{\tau}^{(l)} \rangle)\}_{l \neq m}$. In all experiments, $\langle \mathbf{X}^{(m)\star} | \mathbf{X}^{-(m)\star} \rangle$ was used for prediction.

Additionally, the missing data can be predicted using Equation 5.23 where, in this case, the observed views $\mathbf{X}^{-(m)\star}$ correspond to the training observations in view m and the missing data is represented as: $\mathbf{X}^{(m)\star} = \mathbf{X}^{(m)\star}_{n,j \in O^{(m)}_{n,j}}$, where $O^{(m)}_{n,j}$ is the set of indices (n, j) for which the corresponding $x^{(m)\star}_{n,j}$ are missing.

5.2.4 Synthetic data

I validated the extended GFA model on synthetic data generated from $\mathbf{x}_n^{(m)} \sim \mathcal{N}(\mathbf{W}^{(m)}\mathbf{z}_n, \mathbf{T}^{(m)^{-1}})$ (Equation 2.11). I generated $N = 500$ observations for two views with $D_1 = 50$ ($\mathbf{X}^{(1)} \in \mathbb{R}^{50\times500}$) and $D_2 = 30$ ($\mathbf{X}^{(2)} \in \mathbb{R}^{30\times500}$), respectively. The data was generated from two shared and two view-specific latent components, which were manually specified, similarly to the toy example generated in Klami et al. (2013) (Figure 5.1). The $\boldsymbol{\alpha}^{(m)}$ parameters were set to 1 for the active components and 10^6 for the inactive ones. The loading matrices $\mathbf{W}^{(m)}$ were drawn from $\mathbf{W}^{(m)} \sim \mathcal{N}(0, (\boldsymbol{\alpha}^{(m)})^{-1})$ (Equation 2.12) 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.

I ran GFA experiments on the following selections of synthetic data:

1. *Complete data.*
2. *Incomplete data*:
   (a) 20% of the elements (i.e., entries) 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 components) to assess whether it can learn the true latent components while automatically pruning out the irrelevant ones. The recommended choice for the maximal number of latent components is $K = \min(D_1, D_2)$, but in high dimensional data sets this leads to large K and consequently to long computational times. In practice, a K value that leads to the removal of some irrelevant latent components should be a reasonable choice (Klami et al., 2013). In Figure B.1, I show that the model still converges to the right solution when the number of latent components is overestimated ($K = 30$) in low and high dimensional data.

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, I matched the true and inferred latent components by calculating the maximum similarity (using Pearson's correlation) between them, in all experiments. If a correlation value was negative, the corresponding inferred component was multiplied by -1. The inferred components 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 view from the other on the test set (e.g., predict $\mathbf{X}^{(2)}$ from $\mathbf{X}^{(1)}$) using Equation 5.23. The mean and standard deviation of the mean squared error (MSE) (calculated between the true and predicted values of the non-observed view on the test set) was calculated across the different initialisations (Table 5.1). 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 features on the training set.

In the incomplete data experiments, the missing data was predicted using Equation 5.23. I 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, I ran GFA with complete data, after imputing the missing values using the median of the respective feature.

Finally, in order to assess the CCA performance in complete and incomplete data sets, I ran additional experiments with CCA (Section B.2.2).

5.2.5 HCP dataset

I applied the proposed GFA extension to the publicly available resting-state fMRI and non-imaging features (e.g., demographics, psychometrics and other behavioural features) obtained from 1003 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, I 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 (ICA). This processing was identical to Smith et al. (2015), yielding 19,900 brain features 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 $\mathbf{X}^{(1)} \in \mathbb{R}^{19900\times 1001}$. I used 145 items of the non-imaging features used in Smith et al. (2015) as the remaining features (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 description of the non-imaging measures can be found in Table B.5. The non-imaging matrix included 145 features 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, haemoglobin 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 feature was standardised to have zero mean and unit variance. For additional details on the data acquisition and processing, see Smith et al. (2015).

I 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 components. As in the experiments with synthetic data, all experiments were randomly initialised ten 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.

The number of components obtained in all experiments was greater than 60. Therefore, to facilitate interpretability, I selected the most relevant components by calculating the relative variance explained (rvar) by each component k within each data modality m (i.e., k-th column of $\mathbf{W}^{(m)}$) with respect to the total variance in the data modality:

$$\text{rvar}_k^{(m)} = \frac{\mathbf{w}_k^{(m)T}\mathbf{w}_k^{(m)}}{\text{Tr}(\mathbf{W}^{(m)}\mathbf{W}^{(m)T})} \times 100 \tag{5.24}$$

where $\text{Tr}(\cdot)$ represents the trace of the matrix. The components explaining more than 7.5% variance within any data modality were considered most relevant. In order to decide whether a given most relevant component was

modality-specific or shared, the ratio between the variance explained (var) by each component on the non-imaging and brain data was computed:

$$r_k = \frac{\text{var}_k^{(2)}}{\text{var}_k^{(1)}} \tag{5.25}$$

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 Equation 2.11. A component was considered shared if $0.001 \leq r_k \leq 300$, specific to the non-imaging modality if $r_k > 300$ or brain-specific if $r_k < 0.001$. These values were selected taking into account the imbalance of the total number of features across data modalities (∼100 times more brain connectivity features than non-imaging features). These thresholds were validated in high dimensional synthetic data (Table B.2).

To assess whether the missing data affected the estimation of the most relevant components, I calculated the Pearson's correlations between the components obtained in the complete data experiment and the components obtained in the incomplete data ones (Table 5.3).

All non-imaging features were predicted from brain connectivity on the test set to show the potential of GFA to be used for multi-output prediction. 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 feature on the test set, across the different initialisations:

$$\text{rMSE}_j = \frac{\frac{1}{N}\sum_{n=1}^{N}(x_{nj}^{(2)} - x_{nj}^{(2)^*})^2}{\frac{1}{N}\sum_{n=1}^{N}(x_{nj}^{(2)})^2} \tag{5.26}$$

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

Similarly to the incomplete data experiments on synthetic data, the missing data was predicted using Equation 5.23 and the mean and standard deviation (across initialisations) of the Pearson's correlations between the true and predicted missing values were calculated.

Finally, to compare our GFA results with CCA, we applied a similar CCA analysis proposed by Smith et al. (2015) to the HCP dataset used in the

GFA experiments. In summary, we reduced the dimensionality of both data modalities using PCA (using 100 principal components for each data modality) and applied CCA to these reduced data matrices. The statistical significance of the CCA modes was estimated by permutation inference, in which the subjects of the non-imaging matrix were permuted 10,000 times respecting the family structure of the data (Winkler et al., 2015). For each CCA mode, we compute a p-value to assess whether the "true" canonical correlation (i.e., the canonical correlation of the respective CCA mode obtained without permuting the data) was larger than the null distribution of permuted canonical correlations. If a CCA mode is statistically significant ($p < 0.05$), we remove its effect from the data using matrix deflation (Shawe-Taylor and Cristianini, 2004). These steps are repeated to compute subsequent CCA modes until no more statistically significant modes are obtained.

5.3 Results

In this section, I present the results of the experiments on synthetic data (Section 5.3.1) and real data from the Human Connectome Project (Section 5.3.2).

5.3.1 Synthetic data

Figure 5.1 shows the results of the extended GFA model applied to complete data (experiment 1). The model correctly inferred the components, identifying two of them as shared and the other two as view-specific. These components were all considered most relevant based on the *rvar* threshold (Equation 5.24) and were all correctly assigned as shared or view-specific based on the ratio r_k (Equation 5.25). The structure of the inferred latent components was similar to those used for generating the data (Figure 5.1). The results were robust to initialisation, i.e., the model converged to similar solutions across the different initialisations. Furthermore, the irrelevant latent components 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, the model performed well in the multi-output prediction task, i.e., the averaged MSE was lower than chance level when predicting $\mathbf{X}^{(1)}$ from $\mathbf{X}^{(2)}$, and vice-versa (Table 5.1).

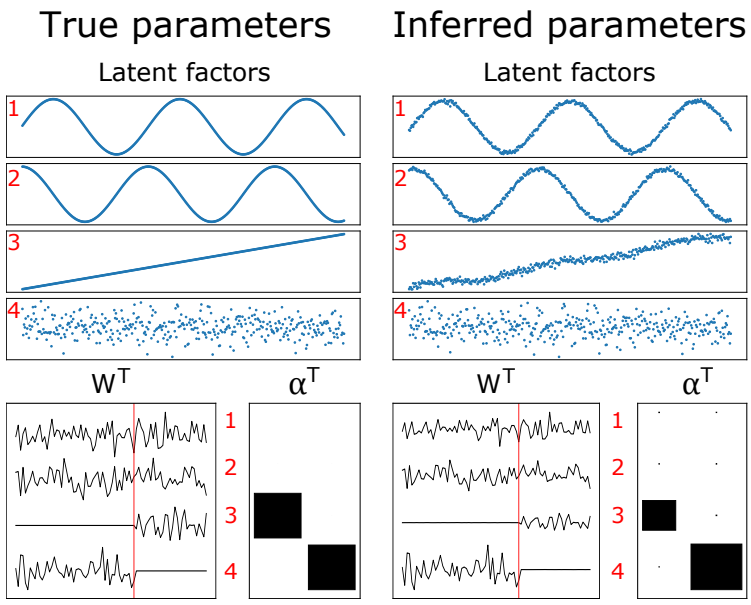

Figure 5.1: True and inferred latent components and model parameters obtained in the complete data experiment. The latent components and parameters used to generate the data are plotted on the left-hand side, and the ones inferred by the model are plotted on the right-hand side. The four rows on the top represent the two shared (1 and 2) and two view-specific (3 and 4) latent components. The loading matrices of the first and second view 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 view are shown on the first and second column of $\boldsymbol{\alpha}^T$, respectively. The small black dots and big black squares represent active and inactive components, respectively.

Figures 5.2a and 5.2b show the results of the incomplete data experiments 2a (20% of the elements of $\mathbf{X}^{(2)}$ missing) and 2b (20% of the rows of $\mathbf{X}^{(1)}$ missing), respectively. The parameters inferred using our GFA extension (Figure 5.2, 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 (Figure 5.2a), i.e., both approaches were able to infer the model parameters fairly well. Even so, the model misses 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 GFA approach 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 (Figure 5.2b). This was not observed when the extended GFA

was applied ($\hat{\tau}^{(1)} \approx 5.04$ and $\hat{\tau}^{(2)} \approx 10.24$).

Finally, the proposed GFA model predicted missing data consistently well in both experiments: $\rho = 0.868 \pm 0.016$ and $\rho = 0.680 \pm 0.039$ (where ρ represents the averaged Pearson's correlation between the missing and predicted values across initialisations) for the incomplete data experiments 2a and 2b, respectively.

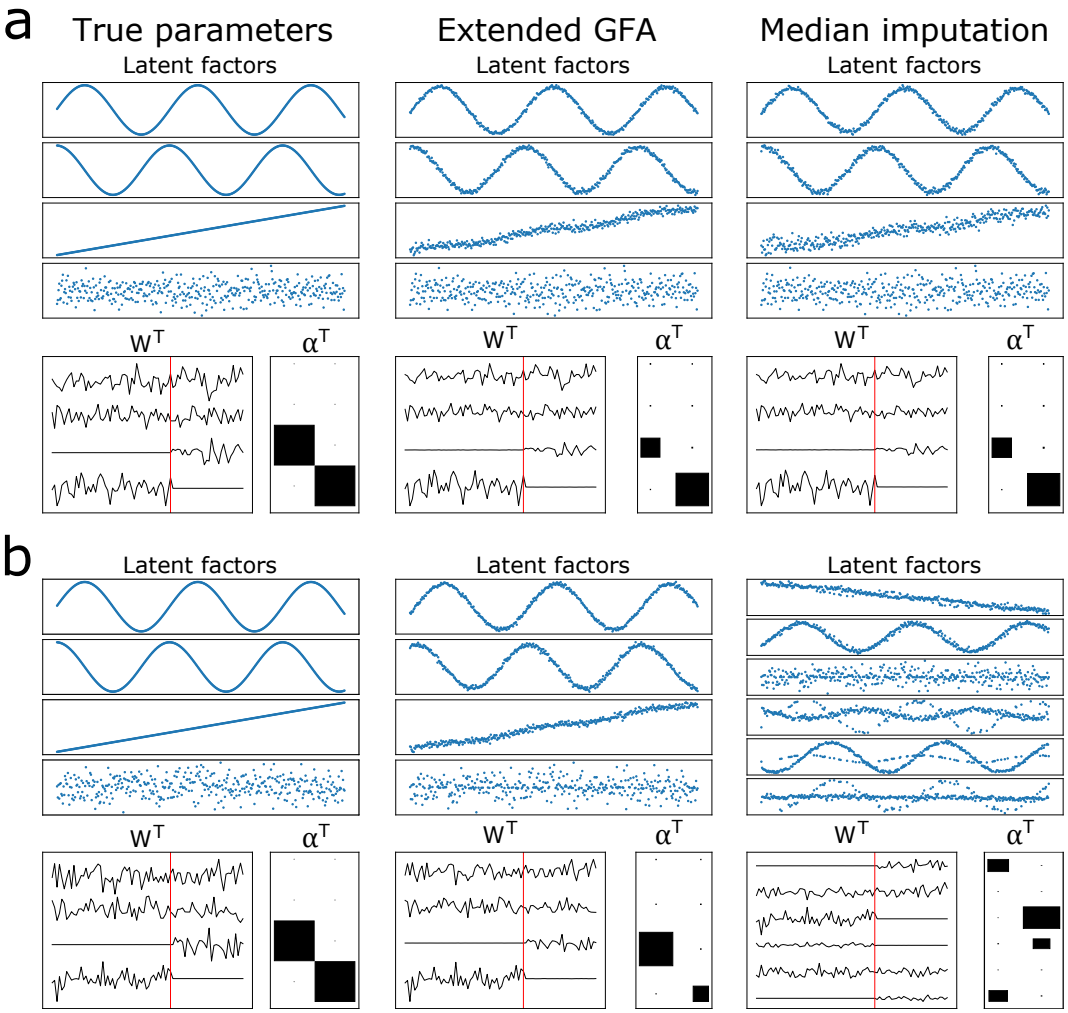

Figure 5.2: True and inferred latent components and model parameters obtained in the incomplete data experiments 2a **(a)** and 2b **(b)**. **(Left column)** latent components and parameters used to generate the data. **(Middle column)** latent components and parameters inferred using the proposed GFA approach. **(Right column)** latent components and parameters inferred using the median imputation approach (the latent components 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 Figure 5.1.

I showed that the model can make reasonable predictions when the data

was missing randomly or one data modality was missing for some observations, i.e., the MSEs were similar across experiments and below chance level (Table 5.1). Moreover, there seems to be no improvement between using the proposed GFA approach or imputing the median before training the model.

Table 5.1: Prediction errors of the multi-output prediction tasks. The values correspond to the mean and standard deviation of the MSEs across 10 initialisations. The first (second) column shows the MSE between the test observations $\mathbf{X}^{(1)\star}$ ($\mathbf{X}^{(2)\star}$) and the mean predictions $\langle \mathbf{X}^{(1)\star} | \mathbf{X}^{(2)\star} \rangle$ ($\langle \mathbf{X}^{(2)\star} | \mathbf{X}^{(1)\star} \rangle$). ours - proposed GFA approach; imputation - median imputation approach; chance - chance level. Exp. 1 - complete data; Exp. 2a - 20% of the elements of $\mathbf{X}^{(2)}$ missing; Exp. 2b - 20% of the rows of $\mathbf{X}^{(1)}$ missing.

		Predict $\mathbf{X}^{(1)}$ from $\mathbf{X}^{(2)}$	Predict $\mathbf{X}^{(2)}$ from $\mathbf{X}^{(1)}$
Exp. 1	ours	1.38 ± 0.21	0.81 ± 0.18
	chance	2.48 ± 0.28	2.24 ± 0.39
Exp. 2a	ours	1.23 ± 0.25	0.71 ± 0.11
	imputation	1.27 ± 0.25	0.74 ± 0.11
	chance	2.29 ± 0.27	2.06 ± 0.29
Exp. 2b	ours	1.14 ± 0.19	0.75 ± 0.18
	imputation	1.17 ± 0.18	0.75 ± 0.18
	chance	2.27 ± 0.26	2.22 ± 0.36

In additional experiments, I showed that the proposed GFA approach outperforms the median imputation approach (in inferring the model parameters and predicting one unobserved view from the observed one), when values from the tails of the data distribution are missing (Figure B.2a and Table B.2). The proposed GFA extension also outperformed the median imputation approach, when both data modalities were generated with missing values in low (Figure B.2b) and high dimensional (Figure B.2c) data.

5.3.2 HCP data

In the complete data experiment, the model converged to a solution comprising 75 latent components, i.e., five components 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 components was consistent across initialisations. The

total percentage of variance explained by the components ($\sum_{m=1}^{2}\sum_{k=1}^{75}\text{var}_k^{(m)}$) corresponded to ∼7.55%, leaving 92.45% of the variance captured by residual error. Within the variance explained, six components were considered most relevant ($\text{rvar}_k^{(m)} > 7.5\%$), which captured ∼27.8% of the variance explained by the total number of components (Table 5.2). Based on the ratio between the variance explained by the non-imaging and brain components r_k (Equation 5.25), we identified four shared components (displayed in Figure 5.3) and two brain-specific components (displayed in Figure 5.4), ordered from the highest to the lowest ratio r_k (Table 5.2).

Table 5.2: Most relevant shared and modality-specific components obtained with complete data according to the proposed criteria. Components explaining more than 7.5% variance within any data modality were considered most relevant. A component was considered shared if $0.001 \leq r_k \leq 300$, specific to non-imaging (NI) modality 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 k-th component in the non-imaging and brain data.

	Components	**rvar** (%)		**var** (%)		$\mathbf{r_k}$
		Brain	NI	Brain	NI	$\text{var}_{\text{NI}}/\text{var}_{\text{brain}}$
Shared	a	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	a	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}

Figure 5.3 shows the loadings of the shared GFA components obtained with complete data. To aid interpretation, the loadings of the brain components 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 shared component (Figure 5.3a) 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 component (Figure 5.3b) relates negative mood, the 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 component

(Figure 5.3c) 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 component (Figure 5.3d) seems to relate emotional functioning, with strong negative 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 wellbeing to cingulo-opercular network (loading negatively), and the left sided default mode network (loading positively).

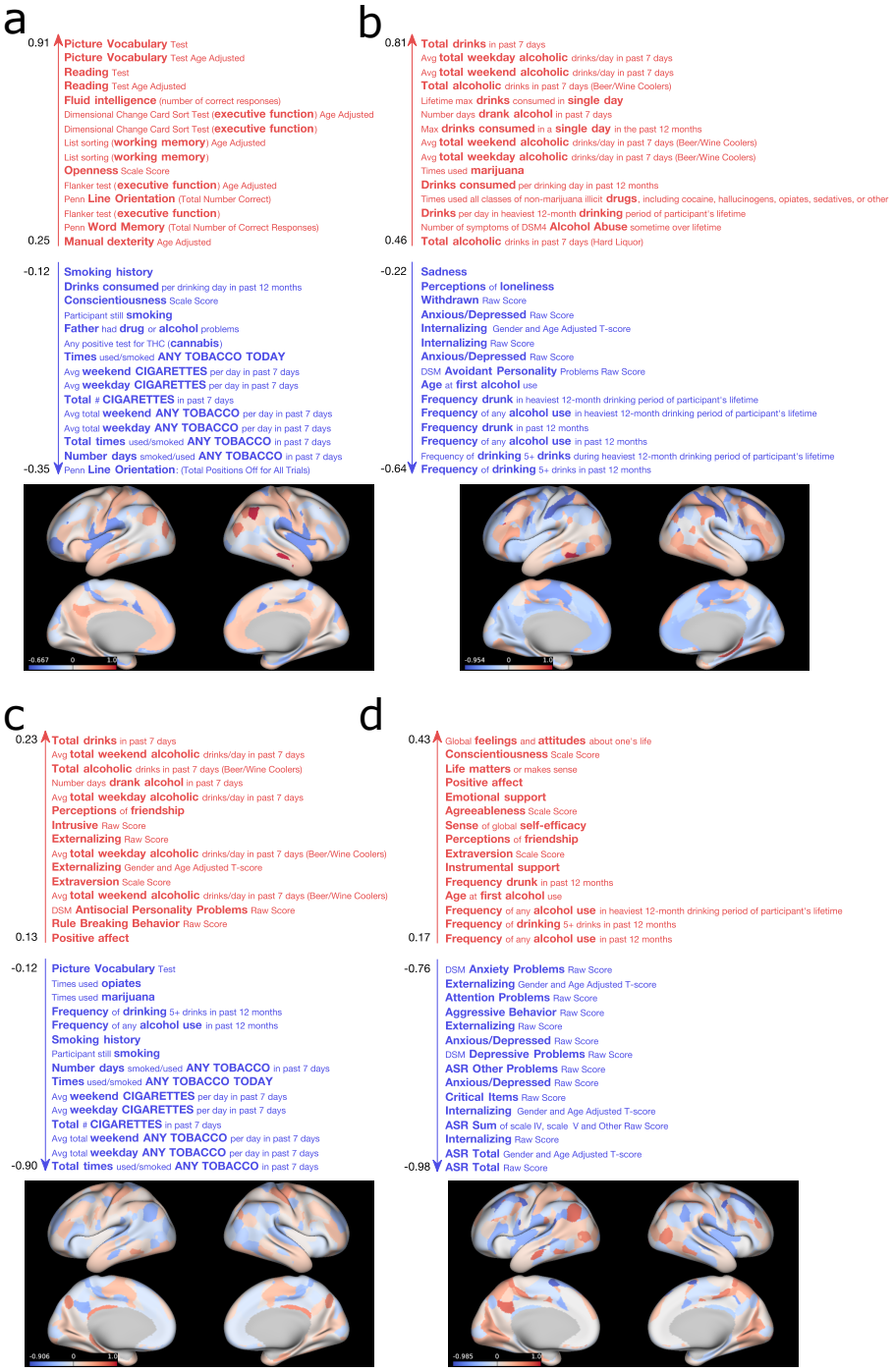

Figure 5.3: Non-imaging features and brain networks described by the first **(a)**, second **(b)**, third **(c)** and fourth **(d)** shared GFA components obtained in the complete data experiment. For illustrative purposes, the top and bottom 15 non-imaging features of each component are shown. The brain 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 Appendix B.2.3). Separate thresholded maps of brain connection strength increases and decreases can be found in Figure B.7.

Figure 5.4 shows the loadings of the brain-specific components obtained with complete data. The first component (Figure 5.4a) 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 component (Figure 5.4b) 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 components show that there is great variability in these networks across the sample, but this variability was not linked to the non-imaging features included in the model.

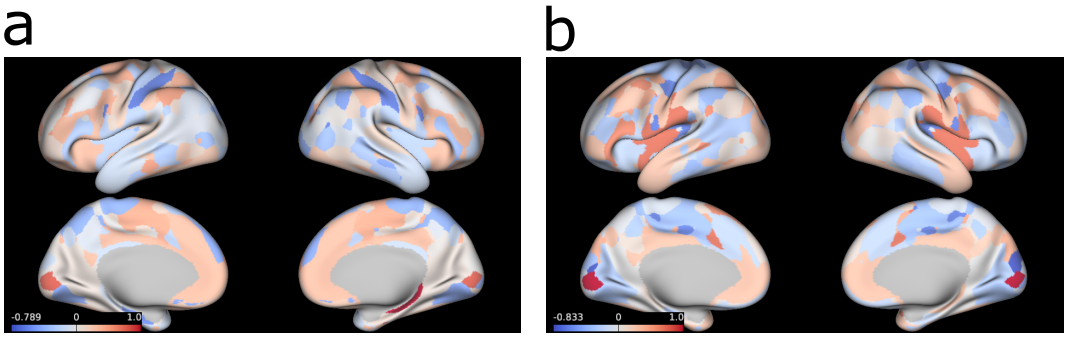

Figure 5.4: Brain networks associated with the brain-specific GFA components obtained in the complete data experiment. The brain surface plots represent maps of brain connection strength increases/decreases, which were obtained by weighting each node's parcel map with the GFA edge-strengths summed across the edges connected to the node (see Appendix B.2.3).

The model converged to a similar solution in the incomplete data experiment 2a (20% of the elements of the non-imaging matrix missing), which included 73 components and the total percentage of variance explained by these was ∼7.60%. The number of most relevant components, based on the rvar metric (Equation 5.24), was six, and they were similar to those obtained in the complete data experiment (Table 5.3), capturing ∼28.2% of the variance explained by all components (Table B.3). Four of these were considered shared components (Figure B.4) and two were considered brain-specific (Figure B.6a, c). In the incomplete data experiment 2b (20% of the subjects missing in the brain connectivity matrix), the model converged to a solution containing 63 components and the total percentage of variance explained corresponded to ∼5.21%. Although more components were removed and a loss of variance

explained was noticeable, the most relevant components were similar to those obtained in the other experiments (Table 5.3, Figure B.5 and Figure B.6b, d), capturing ∼33.2% of the variance explained by all components (Table B.4).

Table 5.3: Similarity (measured by Pearson's correlation) between the most relevant components obtained in the complete and those obtained in the incomplete data experiment 2a and 2b (first and second row, respectively).

	Shared components				**Brain components**	
	a	b	c	d	a	b
Experiment 2a	0.896	0.964	0.954	0.989	0.974	0.974
Experiment 2b	0.907	0.973	0.954	0.995	0.941	0.942

In the multi-output prediction task, the model predicted several non-imaging features better than chance (Figure 5.5) using complete data. The top 10 predicted features corresponded to those with the highest loadings obtained mainly in the first shared component (Figure 5.3a) and were consistent across the incomplete data experiments (Figure B.8). Finally, the model failed to predict the missing values in both incomplete data experiments: $\rho = 0.112 \pm 0.011$ (where ρ represents the averaged Pearson's correlation across initialisations) for experiment 2a; $\rho = 0.003 \pm 0.007$ for experiment 2b.

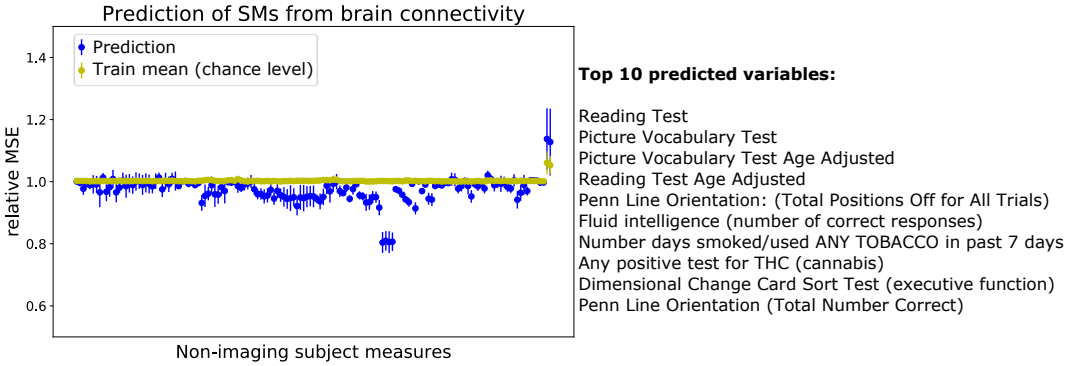

Figure 5.5: Multi-output predictions of the non-imaging features using complete data. The top 10 predicted features are described on the right-hand side. For each non-imaging feature, the mean and standard deviation of the relative MSE (Equation 5.26) between the true and predicted values on the test set was calculated across different random initialisations of the experiments.

To highlight the differences between GFA and CCA, we compared the CCA modes to the GFA components obtained using complete data (Figure 5.6 and Table 5.4). To interpret the association captured by each CCA mode,

we correlated the non-imaging measures and brain connectivity variables with the canonical scores obtained for each data modality (as in Smith et al. (2015)), respectively. The first, second and third CCA modes (Figure 5.6a,b,c) were similar to the top and bottom non-imaging measures obtained in the first GFA component (Figure 5.3a). However, the positive brain loadings on the postero-lateral and medial default mode networks in the first GFA component are split between the first and third CCA modes, respectively. The fifth CCA mode related most strongly to inattention, aggression and antisocial behaviour to positive loadings on posterior insula, and inferior, superior and medial frontal regions. The fourth GFA component contained these non-imaging measures and low mood/internalising as well. Finally, the brain loadings in lateral pre-frontal and insular cortex were similar across the fifth CCA mode and the fourth GFA component.

Finally, the first GFA component (related to CCA modes 1-3) replicates the findings found by Smith et al. (2015) using CCA applied to approximately 500 subjects (first release of the HCP dataset). Both of these contained loadings related to cognitive performance and tobacco or cannabis use, and brain loadings on default mode areas. Some remaining non-imaging measures in Smith et al.'s mode appeared in our fourth GFA component (related to life satisfaction and aggression), which strongly related to different forms of psychopathology.

Table 5.4: Pearson's correlations between the most relevant GFA components and the CCA modes obtained in the HCP experiment with complete data. The values in bold represent the highest absolute correlations between a given CCA mode and the GFA components.

		GFA					
		Shared				**Brain-specific**	
		a	b	c	d	a	b
CCA	a	**0.605**	0.011	0.105	0.064	0.093	0.347
	b	**0.380**	0.112	0.050	0.190	0.093	0.081
	c	0.231	0.112	0.206	0.065	**0.299**	0.048
	d	0.009	**0.191**	0.039	0.061	0.083	0.036
	e	0.052	0.092	0.115	0.173	0.031	**0.386**

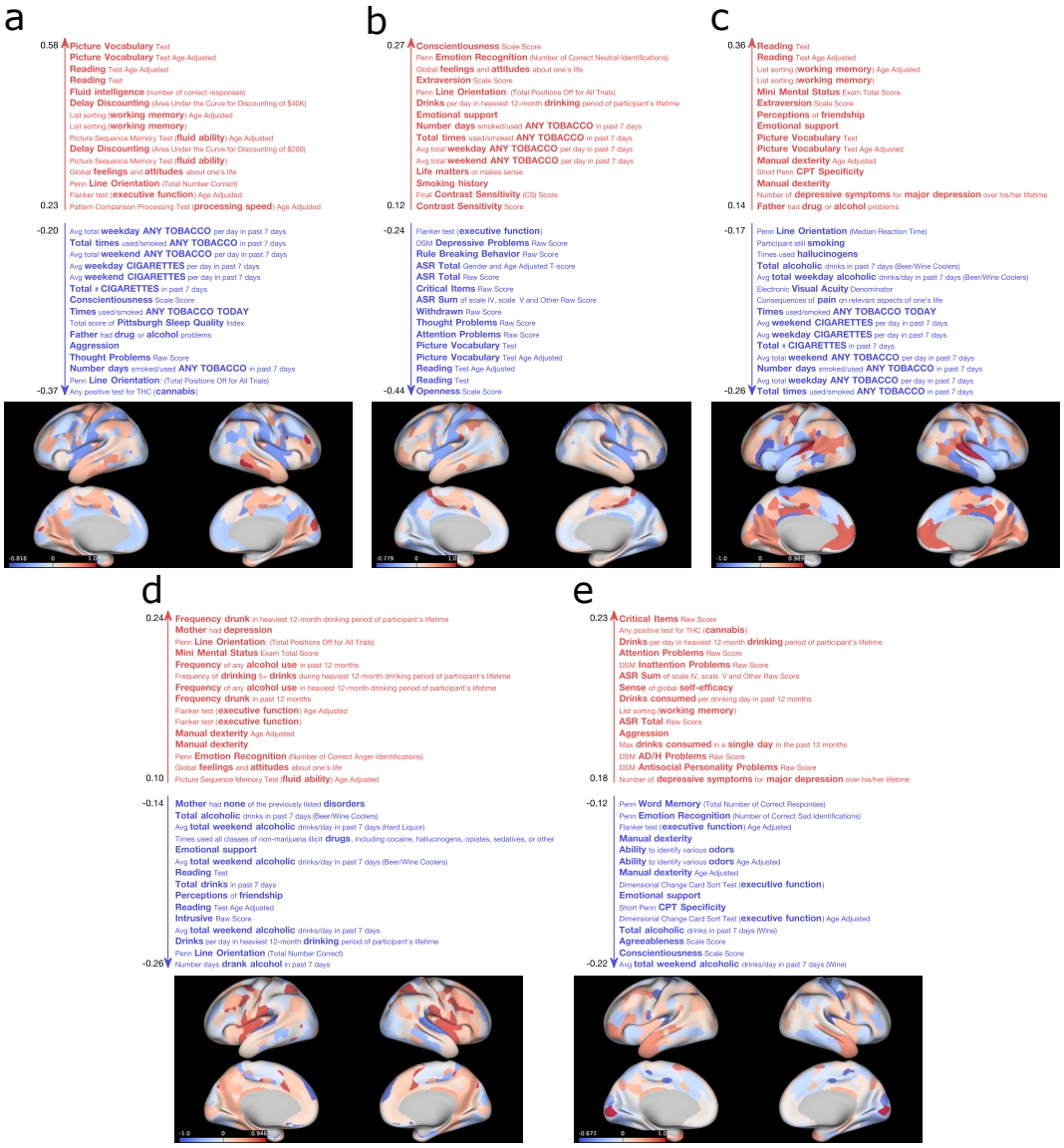

Figure 5.6: Non-imaging measures and brain networks correlated with the CCA modes obtained in the HCP experiment with complete data. The top and bottom 15 non-imaging measures for each component are shown. The brain surface plots represent maps of brain connection strength increases/decreases, which were obtained as described in Section B.2.3.

5.4 Discussion

We showed, using synthetic data, that the proposed GFA extension can correctly learn the underlying latent structure, i.e., it separates the shared components from the view-specific ones, when data is missing. Moreover, it inferred the model parameters better than the median imputation approach in different incomplete data scenarios. CCA, on the other hand, was only able to recover the shared latent components and identified spurious latent compon-

ents when the values of the tails of the data distribution were missing (Figure B.3). These findings highlight the importance of using approaches that can handle missing data and model the view-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 (Table B.1). This indicates that these values might be driving the predictions, and the model fails to predict one view from the other when these values are not carefully imputed. Finally, our GFA extension was able to predict, on synthetic data, the missing values in different incomplete data scenarios.

The application of the extended GFA model to the HCP dataset led to the identification of six most relevant components: four describing associations between brain connectivity and non-imaging features 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 and the brain connectivity features is relatively small with respect to the overall variance in brain connectivity. Interestingly, most of the featured domains of non-imaging features were not unique to particular components, but appeared in different arrangements across the four components. For instance, alcohol use appeared in three out of four components: 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. The second GFA component (Figure 5.3b) has puzzlingly opposing loadings of frequency of alcohol use versus total alcohol drunk in the last seven days. This is probably because the distributions of "total amount" answers are very skewed, with most subjects reporting zero, hence a lot of variance can be explained by this rather paradoxical set of loadings (Figure B.9). Alternatively, it might be that these alcohol use items represent two different behaviours, where "total amount" answers are related to a more short-term alcohol use and the "frequency" answers might represent more long-term and consistent alcohol use.

The first GFA component was almost identical to the first CCA mode (Figure 5.6 and Table 5.4), 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 features to the first GFA component. A possible explanation of the differences observed between the CCA and GFA results is that we had to apply principal component analysis (using 100 principal components for each data modality) 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. Moreover, the lack of correspondence between CCA modes and GFA factors might be due to the different assumptions on noise across the two models, i.e., CCA assumes full covariance matrices, while GFA assumes diagonal covariance matrices. Curiously, sparse CCA (which also assumes diagonal covariance matrices, and it does not need a PCA step beforehand) was applied to this dataset (results not show here), and it uncovered modes of covariation similar to GFA's factors. Finally, this difference might also be explained due to the random rotations of CCA, however this is unlikely reason here because we ran CCA iteratively (see Section 5.2.5), where each mode is force to be orthogonal to the subsequent ones and in this way we potentially fix the issue of rotation ambiguity in CCA.

The brain-specific components 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 (frontoparietal and default mode) indicates, unsurprisingly, that there are aspects of these networks that are not related to the non-imaging features included here. Interestingly, the second brain component (Figure 5.4b) showed a few similarities ($\rho \approx 0.39$, Table 5.4) with the fifth CCA mode (Figure 5.6e), which indicates that this mode could be either a spurious association or a brain-specific component that CCA is not able to explicitly identify. This finding indicates the importance of separating the shared components from the modality-specific ones and the use of more robust inference methods. Finally, 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 features from brain connectivity in incomplete data sets. Even though the relative MSE values were modest, the model could predict several features better than chance. Importantly, the best predicted features corresponded to the loadings most informative in the shared components (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 components 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 components is overestimated. Secondly, although the criteria used to select the most relevant components were validated on synthetic data, these can be further improved, e.g., by assessing the robustness of the components. 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). Finally, although we were not proposing an imputation method, our GFA extension could have been compared to better methods than median imputation, e.g., using PCA-low-rank-soft-shrinkage (Cai et al., 2010) or an iterated GFA-imputation extension.

In summary, I proposed an extension of Group Factor Analysis (GFA) that can uncover associations among multiple data modalities, even when these modalities have missing data. We showed that our proposed GFA approach can: (1) find associations between high dimensional brain connectivity data and non-imaging features (e.g., demographics, psychometrics and other behavioural features) and (2) predict non-imaging features 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). Due to its Bayesian nature, GFA provides great flexibility to be extended to more complex models that can potentially solve more complex tasks in neuroimaging studies.

Chapter 6

Uncovering multivariate associations in subgroups of patients with genetic FTD using GFA

In this chapter, I present two extensions of the GFA model: a new sparse GFA model and supervised GFA. These models include sparsity over the features and samples to allow feature selection to improve model interpretability and sample selection to identify latent components that characterise subgroups of subjects at the individual subject level. The study presented in this chapter was a collaborative work with Samuel Kaski (Department of Computer Science, Aalto University), Jonathan Roher and Arabella Bouzigues (UCL Dementia Research Centre). Samuel Kaski and Janaina Mourao-Miranda supervised the project. Jonathan Rohrer and Arabella Bouzigues provided the preprocessed GENFI dataset and contributed to the interpretation and discussion of the results.

6.1 Introduction

As described in the previous chapter, GFA can be used to model associations among multiple views, separating the associations between views from those within views. However, the model does not allow feature-wise sparsity, which is particularly useful for feature selection and model interpretation in high dimensional data sets. Moreover, the model is unable to identify associations only present in population subgroups, which might be relevant in clinical applications where populations are usually heterogeneous (e.g., neurological and

psychiatric disorders).

Bunte et al. (2016) proposed a sparse extension of GFA to find biclusters, which are defined as sets of rows that are similar for sets of columns in a data matrix, and vice versa. These biclusters are inferred by adding shrinkage priors (e.g., spike-and-slab priors) over the loading matrices and latent variables to impose sparsity over samples and features, respectively (see Section 2.3.4). For instance, the biclusters can be interpreted as subsets of individuals sharing associations among subsets of features in multiple data modalities.

The spike-and-slab priors have been widely used in sparse Bayesian models and have shown good performance in practice (Piironen and Vehtari, 2017; van Erp et al., 2019). However, due to its discrete nature, the model inference may be quite slow. Alternative continuous shrinkage priors, such as the horseshoe prior (Carvalho et al., 2009), have been proposed to provide more efficient inference using automatic methods (e.g., Hamiltonian Monte Carlo (HMC) (Neal, 2011; Betancourt and Girolami, 2013)), while obtaining similar performance to the spike-and-slab prior in practice (Piironen and Vehtari, 2017; van Erp et al., 2019). In this study, I implemented a new sparse GFA method by replacing the spike-and-slab priors of Bunte et al.'s model with regularised horseshoe priors (Piironen and Vehtari, 2017) to uncover sparse associations among multiple data modalities and identify components that characterise subgroups of subjects at the individual subject level. In addition, I extended the sparse GFA model by including a discriminative module to find latent components that describe pre-defined subtypes and explore the heterogeneity of the subtypes. This new model was termed *supervised GFA*.

The sparse and supervised GFA were applied to synthetic data and the Genetic Frontotemporal dementia Initiative (GENFI) dataset, which includes patients with genetic frontotemporal dementia (FTD). In FTD, a large proportion of cases are caused by mutations in progranulin (*GRN*), microtubule-associated protein tau (*MAPT*) and chromosome 9 open reading frame 72 (*C9orf72*) (Snowden et al., 2012; Koskinen et al., 2013). *GRN* and *MAPT* mutations are associated with distinct phenotypes representing more homogeneous groups, whereas *C9orf72* is known to be a heterogeneous group (Mahoney et al., 2012). Therefore, GENFI serves as a real data set with a partially known ground truth for validating the proposed models. Here, I used sparse and supervised GFA to: (1) uncover associations between brain structure and non-imaging data (i.e., behaviour, disease severity and cognitive measures) in genetic FTD; (2) identify latent components that may describe the distinct

subtypes and explore within-subtype variability; (3) assess how the components are expressed at the individual level.

6.2 Methods

In this section, I first describe our new implementation of sparse GFA using regularised horseshoe priors (Section 6.2.1), which is followed by descriptions of the supervised GFA model (Section 6.2.2), model inference and implementation (Section 6.2.3) and how supervised GFA can be used for prediction (Section 6.2.4). I then present a brief description of the generated synthetic data (Section 6.2.5) and the GENFI data (Section 6.2.6). I end this section by describing the approach used to assess the robustness of the inferred components (Section 6.2.7).

6.2.1 Sparse GFA using regularised horseshoe priors

The horseshoe prior (Carvalho et al., 2009) is a popular shrinkage prior for Bayesian regression that ensures that small coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_D)^T$ (where D is the number of features) are heavily shrunk towards zero, while large coefficients remain large. However, this property might be harmful in practice when the coefficients are weakly identified. Piironen and Vehtari (2017) proposed a regularised extension of the horseshoe prior to ensure that large βs are shrunk at least by a small amount. These priors are often termed as global-local shrinkage priors, because there is a global parameter τ that shrinks all coefficients towards zero, while the local parameters $\boldsymbol{\lambda}$ allow some of these to escape complete shrinkage. The regularised horseshoe prior is defined as follows (Piironen and Vehtari, 2017):

$$\begin{aligned} \beta_j|\tilde{\lambda}_j, \tau \sim \mathcal{N}(0, \tau^2\tilde{\lambda}_j^2), \quad \tilde{\lambda}_j^2 = \frac{c^2\lambda_j^2}{c^2 + \tau^2\lambda_j^2}, \quad \lambda_j \sim C^+(0,1), \quad j = 1, ..., D, \\ c^2 \sim \text{Inv-Gamma}(\nu/2, \nu s^2/2), \quad \tau \sim C^+(0, \tau_0^2), \quad \tau_0 = \frac{p_0}{D - p_0}\frac{\sigma}{\sqrt{N}}, \end{aligned} \tag{6.1}$$

where p_0 is the prior guess of the number of relevant features, σ is the noise standard deviation, N is the number of samples and C^+ represents a half-Cauchy distribution (which represents a Cauchy distribution truncated to only have non-zero probability density for values greater than or equal to the location of the peak). The inverse-Gamma distribution over c^2 corresponds to a Student-t$_\nu(0, s^2)$ (with ν degrees of freedom and scale s^2) slab for coefficients far from zero (Piironen and Vehtari, 2017). If the degrees of freedom ν are

small enough, the prior will have heavy tails that ensure robust shrinkage of the large coefficients. In this way, the local parameters $\tilde{\boldsymbol{\lambda}}$ cause small coefficients to shrink close to zero, while also regularising the largest ones. For more details on shrinkage priors for Bayesian regression, see Piironen and Vehtari (2017) and van Erp et al. (2019).

We replaced the spike-and-slab priors of the sparse GFA model (Equations 2.19-2.20) with regularised horseshoe priors (Figure 6.1). The priors over the loading matrices ($\mathbf{W}^{(m)} \in \mathbb{R}^{D_m \times K}$, where D_m is the number of features in the m-th view and K is the number of components) are defined as follows:

$$
\begin{aligned}
w_{j,k}^{(m)} | \tilde{\lambda}_{j,k}^{(m)}, \tau_w^{(m)} &\sim \mathcal{N}(0, (\tau_w^{(m)})^2 (\tilde{\lambda}_{j,k}^{(m)})^2), \quad j = 1, ..., D_m, \quad k = 1, ..., K, \\
(\tilde{\lambda}_{j,k}^{(m)})^2 &= \frac{(c_k^{(m)})^2 (\lambda_{j,k}^{(m)})^2}{(c_k^{(m)})^2 + (\tau_w^{(m)})^2 (\lambda_{j,k}^{(m)})^2}, \quad \lambda_{j,k}^{(m)} \sim C^+(0,1), \\
\tau_w^{(m)} \sim C^+(0, \tau_0^2), \quad \tau_0^{(m)} &= \frac{p_0^{(m)}}{D_m - p_0^{(m)}} \frac{(\sqrt{\sigma^{(m)}})^{-1}}{\sqrt{N}}, \quad \sigma^{(m)} \sim \Gamma(a, b),
\end{aligned}
\tag{6.2}
$$

where $p_0^{(m)}$ and $\sigma^{(m)}$ are the prior guess of the number of relevant features and noise precision of the m-th view, respectively. The prior over $(c_k^{(m)})^2$ is an inverse-Gamma distribution as defined in Equation 6.1. The prior allows a different number of relevant features for each view because a global shrinkage parameter $\tau_w^{(m)}$ is specified for each m-th view. Moreover, as a c value is specified for each component k within a view, the prior implements sparsity over the views (i.e. $c_k^{(m)} \to 0$ leads to $\mathbf{w}_k^{(m)} \to 0$). A component is deemed "irrelevant" if the loadings of that component are pushed close to zero for all views. Finally, the prior also implements feature-wise sparsity because some loadings escape shrinkage due to large local parameters in $\boldsymbol{\Lambda}^{(m)}$. An analogous prior is used over the latent variables ($\mathbf{Z} \in \mathbb{R}^{K \times N}$) to include sparsity over the samples:

$$
\begin{aligned}
z_{k,n} | \tilde{\lambda}_{k,n}^{(z)}, \tau_k^{(z)} &\sim \mathcal{N}(0, (\tau_k^{(z)})^2 (\tilde{\lambda}_{k,n}^{(z)})^2), \quad n = 1, ..., N, \quad k = 1, ..., K, \\
(\tilde{\lambda}_{k,n}^{(z)})^2 = \frac{(c_k^{(z)})^2 (\lambda_{k,n}^{(z)})^2}{(c_k^{(z)})^2 + (\tau_k^{(z)})^2 (\lambda_{k,n}^{(z)})^2}, &\quad \tau_k^{(z)} \sim C^+(0,1), \quad \lambda_{k,n}^{(z)} \sim C^+(0,1),
\end{aligned}
\tag{6.3}
$$

where the prior over $c_k^{(z)}$ is also defined as in Equation 6.1. The regularised horseshoe prior implements different levels of sparsity across the latent components by assuming different $\tau_k^{(z)}$. The interpretation of the effects of $\mathbf{c}^{(z)}$

and $\mathbf{\Lambda}^{(z)}$ is equivalent to that described above for the loading matrices. Each view is then generated from the following model:

$$\mathbf{x}_n^{(m)} \sim \mathcal{N}(\mathbf{W}^{(m)}\mathbf{z}_n, {\sigma^{(m)}}^{-1}). \tag{6.4}$$

The joint probability distribution of sparse GFA is then given by:

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}, \mathbf{W}, \mathbf{\Lambda}_w, \boldsymbol{\tau}_w, \mathbf{c}_w, \mathbf{\Lambda}^{(z)}, \boldsymbol{\tau}^{(z)}, \mathbf{c}^{(z)}, \boldsymbol{\sigma}) = \prod_{m=1}^{M} \Bigg[p(\mathbf{W}^{(m)} | \mathbf{\Lambda}_w^{(m)}, \boldsymbol{\tau}_w^{(m)}, \mathbf{c}_w^{(m)}) \\ p(\mathbf{\Lambda}_w^{(m)}) p(\boldsymbol{\tau}_w^{(m)}) p(\mathbf{c}_w^{(m)}) p(\sigma^{(m)}) \prod_{n=1}^{N} \Bigg(p(\mathbf{z}_n | \boldsymbol{\lambda}_n^{(z)}, \boldsymbol{\tau}^{(z)}, \mathbf{c}^{(z)}) p(\boldsymbol{\lambda}_n^{(z)}) \\ p(\mathbf{x}_n | \mathbf{z}_n, \mathbf{W}^{(m)}, \sigma^{(m)}) \Bigg) \Bigg] p(\boldsymbol{\tau}^{(z)}) p(\mathbf{c}^{(z)}), \end{aligned} \tag{6.5}$$

where $\boldsymbol{\lambda}_n^{(z)}$ is the n-th column of $\mathbf{\Lambda}^{(z)}$.

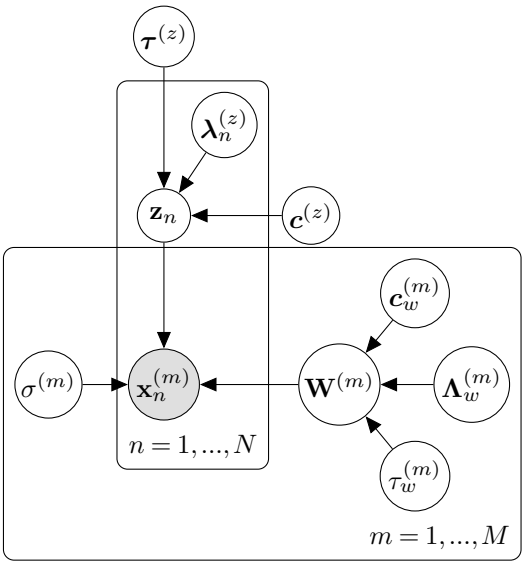

Figure 6.1: Graphical representation of sparse GFA using regularised horseshoe priors.

6.2.2 Supervised GFA

To assess whether the latent components are related to a given class (e.g., a subtype), we assume that observed labels $\mathbf{Y} \in \mathbb{R}^{N \times S}$ (where S is the number of classes) are generated by the same latent variables that generate $\mathbf{X}$ (Figure 6.2). In addition, the model includes a matrix $\mathbf{A}$ that stores the probabilities of a given subtype s ($\mathbf{a}_{(s,:)}$) being associated with any of the latent components. We added a regularised horseshoe prior over $\mathbf{A}$ to include sparsity over the

subtypes and, in this way, find latent components that are associated with a given subtype:

$$\begin{aligned} a_{s,k}|\tilde{\lambda}_{s,k}^{(a)}, \tau_k^{(a)} &\sim \mathcal{N}(0, (\tau_k^{(a)})^2(\tilde{\lambda}_{s,k}^{(a)})^2), \quad s = 1, ..., S, \quad k = 1, ..., K, \\ (\tilde{\lambda}_{s,k}^{(a)})^2 &= \frac{(c_k^{(a)})^2(\lambda_{s,k}^{(a)})^2}{(c_k^{(a)})^2 + (\tau_k^{(a)})^2(\lambda_{s,k}^{(a)})^2}, \quad \tau_k^{(a)} \sim C^+(0,1), \quad \lambda_{s,k}^{(a)} \sim C^+(0,1), \end{aligned} \tag{6.6}$$

where $\mathbf{c}^{(a)}$, $\mathbf{\Lambda}^{(a)}$ and $\boldsymbol{\tau}^{(a)}$ can be similarly interpreted as the parameters of the regularised horseshoe prior used over $\mathbf{Z}$. $\mathbf{A}$ and $\mathbf{Z}$ are then normalized, where a large absolute value $|z_{k,n}|$ represents a high probability of a sample n being associated with a latent component k, and a large positive value $a_{s,k}$ represent a high probability of a subtype s being related to a latent component k. These are calculated as follows:

$$\bar{z}_{k,n} = \frac{|z_{k,n}|}{\sum_{j=1}^{K} |z_{j,n}|}, \quad \bar{a}_{s,k} = \frac{\exp(a_{s,k})}{\sum_{j=1}^{K} \exp(a_{s,j})}, \tag{6.7}$$

Using these two matrices, one can calculate the probability of a given sample n to belong to a subtype s, $\Psi_{n,s}$:

$$\Psi_{n,s} = \sum_{k=1}^{K} \bar{a}_{s,k}\bar{z}_{k,n}, \quad y_{n,s} \sim \text{Bernoulli}(\Psi_{n,s}), \tag{6.8}$$

where the label $y_{n,s}$ is assumed to be binary and generated from a Bernoulli distribution. The joint probability distribution of the supervised GFA model is then defined as follows:

$$\begin{aligned} p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{A}, \mathbf{\Theta}) = \prod_{n=1}^{N} \Big[p(\mathbf{y}_n|\mathbf{A})\Big] p(\mathbf{A}|\mathbf{\Lambda}^{(a)}, \boldsymbol{\tau}^{(a)}, \mathbf{c}^{(a)}) p(\mathbf{\Lambda}^{(a)}) p(\boldsymbol{\tau}^{(a)}) p(\mathbf{c}^{(a)}) \\ \times p(\mathbf{X}, \mathbf{Z}, \mathbf{\Theta}), \end{aligned} \tag{6.9}$$

where $\mathbf{\Theta} = \{\mathbf{W}, \mathbf{\Lambda}_w, \boldsymbol{\tau}_w, \mathbf{c}_w, \mathbf{\Lambda}^{(z)}, \boldsymbol{\tau}^{(z)}, \mathbf{c}^{(z)}, \boldsymbol{\sigma}\}$, $p(\mathbf{X}, \mathbf{Z}, \mathbf{\Theta})$ was defined in Equation 6.5 and $\mathbf{x}_n^{(m)}$ is generated from Equation 6.4.

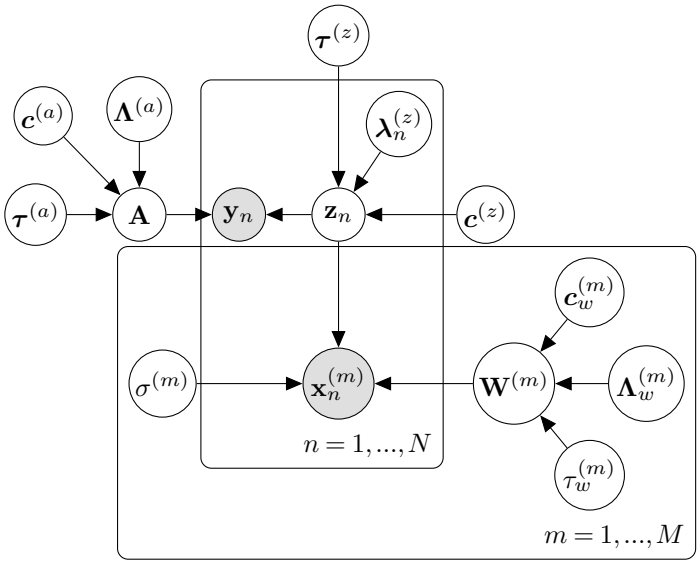

Figure 6.2: Graphical representation of supervised GFA.

6.2.2.1 Extended supervised GFA in GENFI dataset

As the GENFI dataset (Section 6.2.6) included presymptomatic and symptomatic mutation carriers, we decided to extend the model to account for both the patients' subtype category and to the fact the patients could be presymptomatic or symptomatic. This was achieved by inferring two subtype probabilities matrices: $\mathbf{A}$ (as in Equations 6.6-6.7) to infer the subtype probabilities within the affected/symptomatic group; $\mathbf{P}$ to infer the subtype probabilities within the presymptomatic group. A regularised horseshoe prior was added over $\mathbf{P}$ (as it was done for $\mathbf{A}$ in Equation 6.6), which was then normalised, as for $\mathbf{A}$ in Equation 6.7. The probabilities of the subjects to belong to the symptomatic and presymptomatic mutation carriers were estimated in the following way:

$$\begin{aligned} \Psi_{s,n}^{(A)} &= \sum_{k=1}^{K} \bar{a}_{s,k}\bar{z}_{k,n}, \quad y_{s,n}^{(A)} \sim \text{Bernoulli}(\Psi_{s,n}^{(A)}), \\ \Psi_{s,n}^{(P)} &= \sum_{k=1}^{K} \bar{p}_{s,k}\bar{z}_{k,n}, \quad y_{s,n}^{(P)} \sim \text{Bernoulli}(\Psi_{s,n}^{(P)}), \end{aligned} \tag{6.10}$$

where $y_{s,n}^{(A)} = 1$ if a symptomatic individual belongs to subtype s, and $y_{s,n}^{(A)} = 0$ otherwise. $y_{s,n}^{(P)} = 1$ if a presymptomatic individual belongs to subtype s, and $y_{s,n}^{(P)} = 0$ otherwise.

6.2.3 Model inference and implementation

As in GFA and Bayesian CCA (Section 2.3), exact inference of sparse (unsupervised) and supervised GFA is analytically intractable. Here, I used Hamilto-

nian Monte Carlo (HMC) (Neal, 2011; Betancourt and Girolami, 2013), which is a sampling technique, to approximate the posterior distribution of both models. Briefly, sampling algorithms, such as Markov Chain Monte Carlo (MCMC) methods, approximate the posterior distribution by drawing samples from it to compute posterior estimates (e.g., posterior expectation and variance).

MCMC methods use the properties of a Markov Chain to stochastically explore (using Markov transitions) the space around the mode of the posterior distribution (i.e., the region with high probability mass) (Betancourt, 2018). However, constructing appropriate transitions is very important to use these methods efficiently, which is often a challenging process. The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) was proposed to mitigate this issue by considering a proposal and a correction step. In the former, a candidate sample is randomly generated from a proposal probability distribution (i.e., a probability distribution that is compared to the posterior distribution), which is accepted if it falls into regions close to the high posterior probability mass, or rejected otherwise in the correction step (Bishop, 2006; Betancourt, 2018). However, this random "walk" is problematic in high dimensional spaces, where, due to its geometry, most of the proposal samples fall far away from the high probability regions. The rejection rate is therefore high, making the procedure very slow and with the possibility of several regions of the target distribution not being explored. It is possible to induce a larger acceptance rate by shrinking the covariance of the proposal distribution, but the transitions would be very small, which leads to extremely slow exploration. Moreover, even if the posterior distribution is well explored, the slow exploration yields large autocorrelations and imprecise estimates (Betancourt, 2018). Hamiltonian Monte Carlo (HMC) is a MCMC method that makes use of Hamiltonian dynamics to improve the exploration step and increase the acceptance rate in high dimensional spaces (Betancourt, 2018). In this way, approximate inference can be run more efficiently to compute good estimates of the model's parameters. The description of these methods and its variants are beyond the scope of this thesis. For more details, see e.g. Gelman et al. (2013); Betancourt (2018).

In recent years, several probabilistic programming libraries, such as Stan (Stan Development Team, 2019), Edward (Tran et al., 2016), PyMC3 (Salvatier et al., 2016) and NumPyro (Phan et al., 2019), have been developed to provide high-performance probabilistic modelling and inference. These libraries use, for instance, HMC and its extensions (e.g., the No-U-Turn Sampler

(NUTS) (Hoffman and Gelman, 2014)) to run full Bayesian statistical inference efficiently even in high dimensional datasets. Here, I used NumPyro to implement both models. NumPyro is a Python library that uses automatic differentiation and end-to-end compilation to run HMC (with a NUTS implementation) on CPU/GPU, which allows efficient automatic inference, i.e, the user does not need to derive the inference equations.

The models were fitted using HMC with four sampling chains and 5,000 samples (the first 1,000 were discarded as warm-up) and randomly initialised ten times. The best initialisation was selected to maximise the expected log joint posterior density.

6.2.4 Predictive inference

Supervised GFA can be used to predict the probabilities of the subjects on the test set to belong to a given class/subtype of each sample, using its posterior predictive distribution:

$$\begin{aligned} p(\tilde{\mathbf{Y}}|\tilde{\mathbf{X}},\mathbf{Y},\mathbf{X}) &= \int_{\boldsymbol{\Theta}}\int_{\tilde{\mathbf{Z}}} p(\tilde{\mathbf{Y}}|\tilde{\mathbf{X}},\tilde{\mathbf{Z}},\boldsymbol{\Theta})\cdot p(\tilde{\mathbf{Z}}|\tilde{\mathbf{X}},\boldsymbol{\Theta})\cdot p(\boldsymbol{\Theta}|\mathbf{X},\mathbf{Y})\mathrm{d}\boldsymbol{\Theta}\mathrm{d}\tilde{\mathbf{Z}} \\ &\approx \frac{1}{L}\frac{1}{B}\sum_{l=1}^{L}\sum_{b=1}^{B} p(\tilde{\mathbf{Y}}|\tilde{\mathbf{X}},\tilde{\mathbf{Z}}^{(b)},\boldsymbol{\Theta}^{(l)}), \end{aligned} \tag{6.11}$$

where $\boldsymbol{\Theta}^{(l)} \sim p(\boldsymbol{\Theta}|\mathbf{X},\mathbf{Y})$, $\tilde{\mathbf{Z}}^{(b)} \sim p(\tilde{\mathbf{Z}}|\tilde{\mathbf{X}},\boldsymbol{\Theta}^{(l)})$, L is the total number of samples drawn from $p(\boldsymbol{\Theta}|\mathbf{X},\mathbf{Y})$ and B is the total number of samples drawn from $p(\tilde{\mathbf{Z}}|\tilde{\mathbf{X}},\boldsymbol{\Theta}^{(l)})$ using HMC. $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Z}}$ correspond to the input data and latent variables on the test set, respectively, and $\boldsymbol{\Theta} = \{\mathbf{Z},\mathbf{W},\boldsymbol{\Lambda}_w,\boldsymbol{\tau}_w,\mathbf{c}_w,\boldsymbol{\Lambda}^{(z)},\boldsymbol{\tau}^{(z)},\mathbf{c}^{(z)},\boldsymbol{\Lambda}^{(A)},\boldsymbol{\tau}^{(A)},\mathbf{c}^{(A)}\boldsymbol{\sigma}\}$.

6.2.5 Synthetic data

I created a toy example consisting of $N = 150$ samples with three different views ($D_1 = 60$, $D_2 = 40$ and $D_3 = 15$) generated from the following model $\mathbf{x}_n^{(m)} \sim \mathcal{N}(\mathbf{W}^{(m)}\mathbf{z}_n,\sigma^{(m)})$, where the priors over $\mathbf{W}^{(m)}$ and $\mathbf{z}_n$ are defined in Equations 6.2 and 6.3, respectively, and the noise precision of each view was fixed ($\sigma^{(1)} = 3$, $\sigma^{(2)} = 6$ and $\sigma^{(3)} = 4$). The parameters to construct the regularised horseshoe priors over $\mathbf{W}$ and $\mathbf{Z}$ were defined as follows: $(c_k^{(m)})^2$ and $(c_k^{(z)})^2$ were sampled from the inverse-Gamma prior in Equation 6.1 with $\nu = 2$ and $s = 2$ (as proposed by Piironen and Vehtari (2017)); $\boldsymbol{\Lambda}^{(m)}$ and $\boldsymbol{\Lambda}^{(z)}$ were set manually, i.e., the values of the indices of the relevant features/samples were set to 50 and the remainder to 0.01; $\tau^{(m)} = \tau_0^{(m)}$, where $p_0^{(m)} = D_m/3$ (i.e., one third of the features in a given view m were considered

relevant); $\tau_k^{(z)} = \frac{p_k^{(z)}}{N-p_k^{(z)}}\sigma_z$, where $\sigma_z = 1$ and $p_k^{(z)} = N/3$ (i.e., one third of the samples in a given latent component k had non-zero values). The data was generated with $K_{\text{true}} = 3$ latent components that were defined to represent three distinct subtypes ($S = 3$): $\bar{\mathbf{A}} = \mathbf{I}_{S \times K_{\text{true}}}$, where $\mathbf{I}$ represents an identity matrix. The subtypes were defined as independent subsets of samples, i.e.: $S_1 = \{\mathbf{X}_{1,:}, \ldots, \mathbf{X}_{50,:}\}$, $S_2 = \{\mathbf{X}_{51,:}, \ldots, \mathbf{X}_{100,:}\}$ and $S_3 = \{\mathbf{X}_{101,:}, \ldots, \mathbf{X}_{150,:}\}$, where $\mathbf{X}$ corresponds to the concatenation of the three views. The first subtype (Figure 6.3a) is defined by associations between subsets of features of $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$. The second subtype (Figure 6.3b) represent the subset of samples with associations within subsets of features of $\mathbf{X}^{(1)}$. The third subtype (Figure 6.3c) is described by associations between subsets of features of all views. Forty samples of each subtype were chosen for training, and the remaining ten of each subtype were used for testing. Finally, all sampling chains were initialised with $K = 5$.

6.2.6 GENFI dataset

We used cross-sectional brain structural MRI data and non-imaging data (e.g., psychometrics and other behavioural features) of 473 subjects from GENFI (https://www.genfi.org/) recruited across 13 centres in the United Kingdom, Canada, Italy, The Netherlands, Sweden and Portugal. 60 subjects were removed because they had more than one third of the non-imaging features missing. The 413 participants included 296 presymptomatic (114 *C9orf72*, 134 *GRN* and 48 *MAPT*) and 117 symptomatic (58 *C9orf72*, 41 *GRN* and 18 *MAPT*) mutation carriers.

The structural MR images were parcellated into different cortical and subcortical regions using a multi-atlas segmentation propagation approach (Cardoso et al., 2015) to calculate grey matter volumes of the left and right frontal, temporal, parietal, occipital, cingulate and insula cortices. An estimate of the volume of the cerebellum was also included. The subcortical volumes included left and right amygdala, caudate, hippocampus, pallidum, putamen and thalamus. In addition to regional volumetric measures, a measure of volume asymmetry was calculated as proposed in Young et al. (2018), i.e., the absolute value of the difference between the volumes of the right and left hemispheres, normalised by the total volume of both hemispheres. This asymmetry measure was log transformed to improve normality. The total number of brain imaging features was 28 ($\mathbf{X}^{(1)} \in \mathbb{R}^{28 \times 413}$). For more details on the acquisition and preprocessing procedures of the structural MRI data, see Rohrer et al. (2015).

As non-imaging data, we included measures from informant questionnaires (which were completed by primary caregivers) assessing behaviour and disease severity, neuropsychological tasks (completed by patients) and medical assessments of disease severity. A brief description of each non-imaging feature is provided in Table C.1. Measures with more than 10% of missing data were excluded, and the missing values that still remained were imputed by the median of the respective feature across subjects (as the percentage of missing data for most of the remaining features was below 1%). Four confounding variables were regressed out from both data modalities: age, sex, education and total intracranial volume. All features were standardised to have zero mean and unit variance. After the preprocessing step, 34 non-imaging features were included in the second view ($\mathbf{X}^{(2)} \in \mathbb{R}^{34\times413}$).

All presymptomatic individuals and 90% of the symptomatic individuals were randomly selected for training the model. The remaining 10% symptomatic individuals were used for testing. The mean and confidence intervals of the individual subject probabilities on the test set were computed using Equation 6.11. Finally, all sampling chains were initialised with $K = 15$.

6.2.7 Robust data components

The inferred data components correspond to the latent components mapped back to the input/data space (i.e., $\mathbf{X}_k = \mathbf{w}_{(:,k)}\mathbf{z}_{(k,:)}$, $k = 1, ..., K$). To minimise the risk of obtaining components that might have occurred by chance, I used a similar approach as in Bunte et al. (2016) to search and select components that were consistent across the different sampling chains. Briefly, as the components indices can be arbitrarily permuted across different sampling chains, they need to be matched with similar components across the sampling chains. The components were first averaged over the posterior samples within a chain (which can be done because the component indices are stable within a chain), and then compared with the components in other sampling chains using cosine similarity. Two components were considered to be similar if the highest cosine similarity measure was greater than 0.80. Finally, a component was considered robust if it had appeared in more than half of all sampling chains.

6.3 Results

In this section, I present the results of the experiments on synthetic data using supervised GFA (Section 6.3.1) and on the GENFI dataset (Section 6.3.2) using sparse (unsupervised) and supervised GFA.

6.3.1 Synthetic data

Figure 6.3 shows the generated and inferred subtypes (which are represented by three distinct components). Supervised GFA was able to infer the structure of each subtype correctly, i.e., the sparsity over the samples and features was correctly inferred. The number of latent components was correctly estimated, where most of the elements of the "irrelevant" components were very close to zero (these components were not considered robust components using the approach described in Section 6.2.7). Moreover, the individual sample probabilities $\mathbf{\Psi}$ were well inferred by supervised GFA (Figure 6.5a-b) by correctly assigning most samples to the generated subtype, i.e., a high probability was calculated if a sample belonged to a given subtype. The probabilities of the subtypes being associated with the latent components were also properly inferred:

$$\bar{\mathbf{A}} = \begin{bmatrix} 0.984 \pm 0.020 & 0.003 \pm 0.006 & 0.003 \pm 0.005 \\ 0.003 \pm 0.006 & 0.984 \pm 0.020 & 0.003 \pm 0.006 \\ 0.003 \pm 0.005 & 0.003 \pm 0.005 & 0.984 \pm 0.020 \end{bmatrix}$$

where each subtype was associated with (i.e., had a high probability of being related to) a distinct latent component. The posterior distribution over the remainder of the model's parameters included the values used to generate the data (see, e.g., Figure 6.4). Finally, the model predicted the probabilities of the test samples belonging to each class/subtype reasonably well (Figures 6.5c-d), i.e., a high probability of belonging to the correct subtype was correctly estimated for most of the samples.

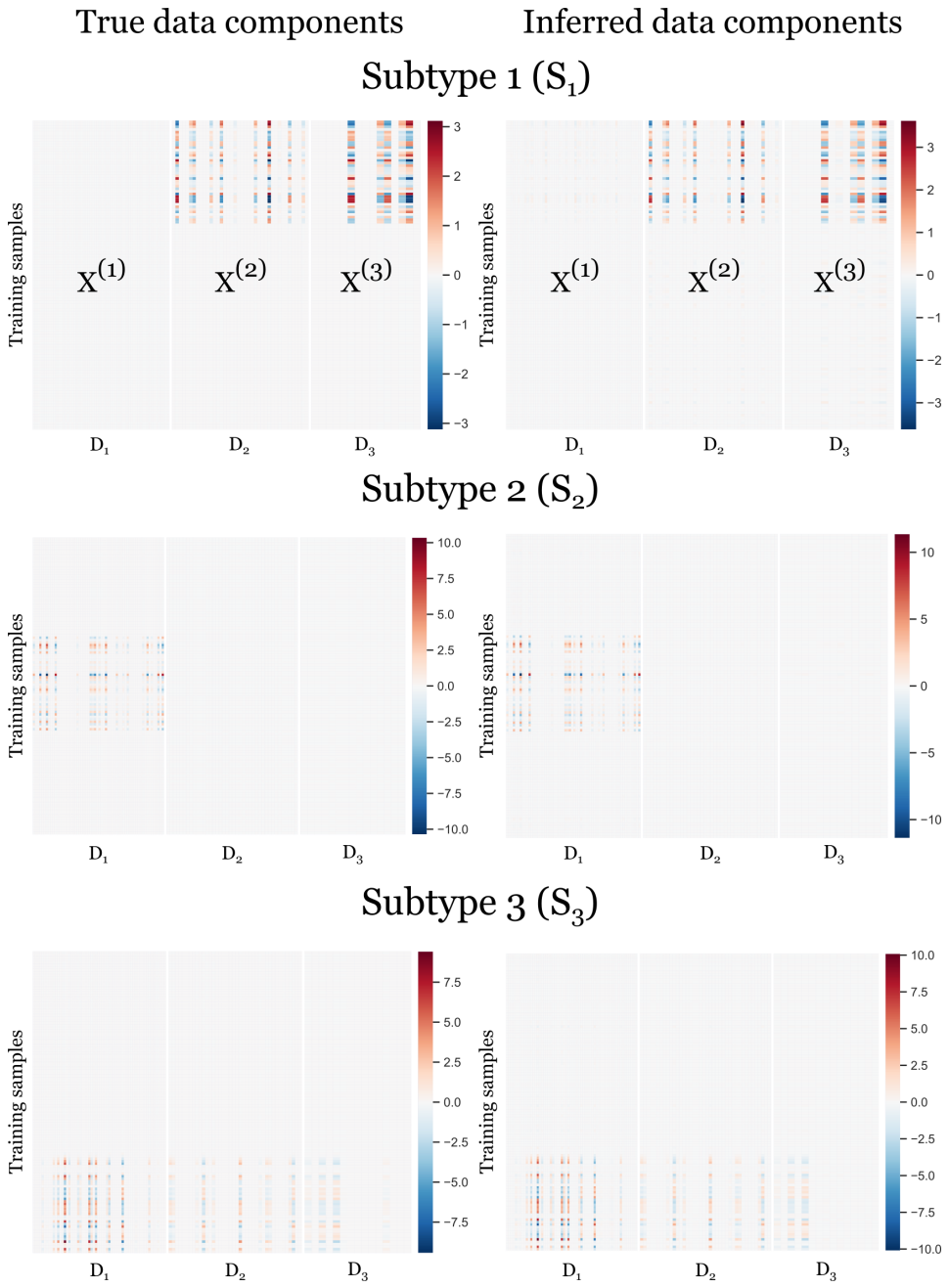

Figure 6.3: Generated and inferred data components representing the underlying subtypes. Subtype 1 is described by associations between subsets of features of $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$. Subtype 2 is described by association within subsets of features of $\mathbf{X}^{(1)}$. Subtype 3 is characterised by associations between subsets of features of all views. The components were transposed for visualisation purposes, where the training samples are represented on the rows and the features on the columns.

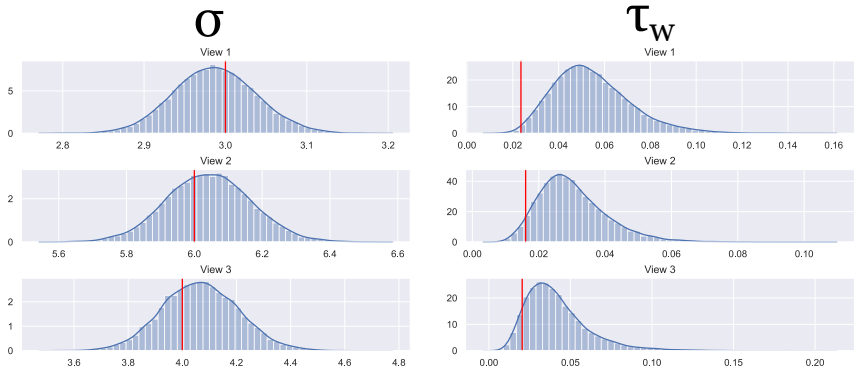

Figure 6.4: Histogram of the posterior samples of σ (left) and τ_w (right) obtained when running supervised GFA on synthetic data. The vertical red line indicates the true value of each parameter.

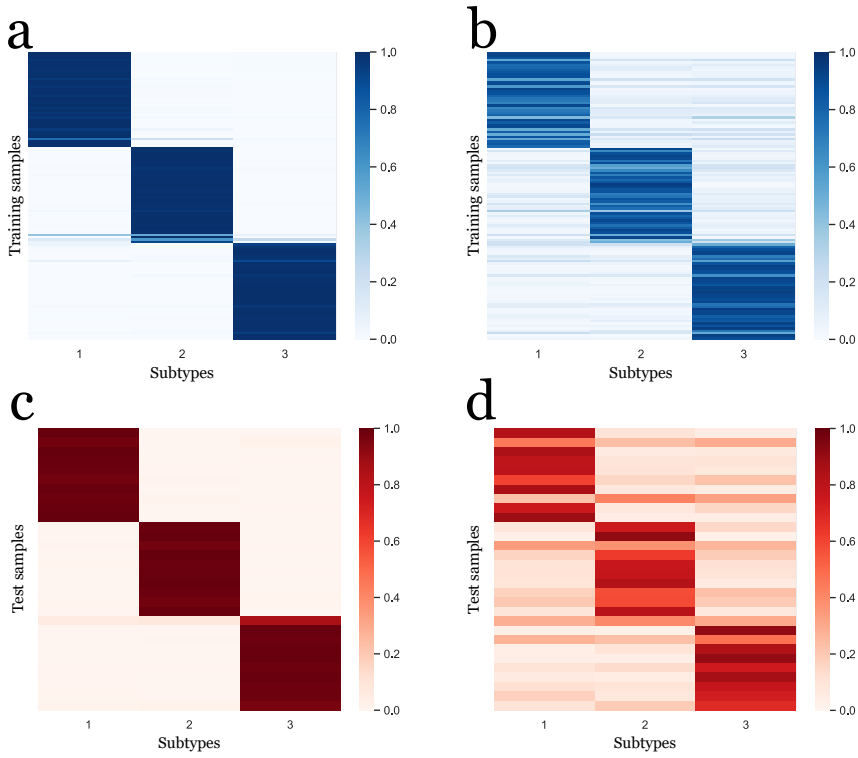

Figure 6.5: Probabilities of samples to belong to the subtypes ($\mathbf{\Psi}$). True **(a)** and inferred **(b)** $\mathbf{\Psi}$ on the training set; True **(c)** and predicted **(d)** $\mathbf{\Psi}$ on the test set. The samples are represented on the rows and the subtypes on the columns.

6.3.2 GENFI dataset

In this section, I present the results obtained using sparse GFA (Section 6.3.2.1) and supervised GFA (Section 6.3.2.2) applied to the GENFI dataset.

6.3.2.1 Sparse GFA

Figures 6.6a-d show the four robust components obtained by sparse GFA that explained more variance ($\approx 40.04\%$) in the data (i.e., Components 15, 7, 9 and 3 in the scree plot shown in Figure 6.7b). In these heatmaps, the rows represent the brain and non-imaging features and the columns represent patients, coloured by genetic group (*C9orf72*, *GRN* and *MAPT*) and status (symptomatic and presymptomatic). Component 15 (Figure 6.6a) explained more than 20% of the variance, and it seems to separate the presymptomatic individuals from the symptomatic ones. Component 7 (Figure 6.6b) did not seem to be associated with any specific subtype, but obtained greater values on only non-imaging measures, mostly for symptomatic *C9orf72* and *MAPT* carriers. Component 9 (Figure 6.6c) shows mostly symmetric subcortical changes, and it was also not specifically associated with any subtype. Component 3 (Figure 6.6d) obtained greater absolute values on a few brain regions for symptomatic *MAPT* carriers. Although, the values of these brain regions were not very close to zero for individuals of the other subtypes, it is likely that this component might be associated with the symptomatic *MAPT* carriers.

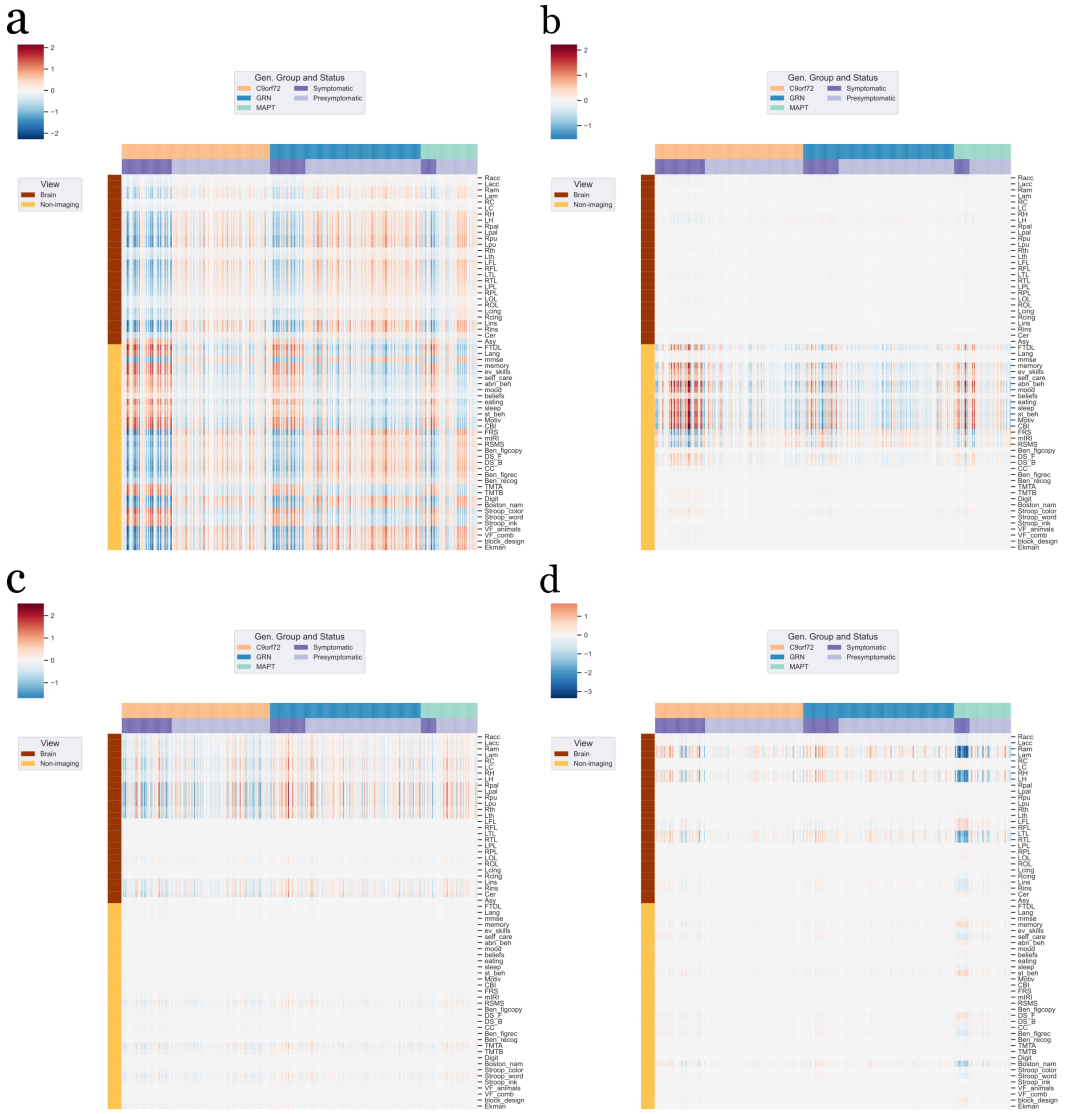

Figure 6.6: Robust components obtained using sparse GFA. **(a)** Component 15; **(b)** Component 7; **(c)** Component 9; **(d)** Component 3. The rows represent the brain (brown) and non-imaging (yellow) features. The columns represent the patients, coloured by genetic group and status. The task name of the abbreviated non-imaging labels can be found in Table C.1 and the full label of the brain imaging features can be found in Table C.2.

To investigate whether any of the remaining robust components were associated with a specific subtype, we considered a component to be related to a specific subtype if it had non-zero values (i.e., $|x_{(d,n)}| > 0.01$) for more than 10% individuals in a single subtype only. Only Component 5 (Figure 6.7a), which explained 2.14% variance, was selected based on this threshold, and it seems to be associated with the symptomatic *GRN* carriers (i.e., most of the non-zero values were specific to these individuals).

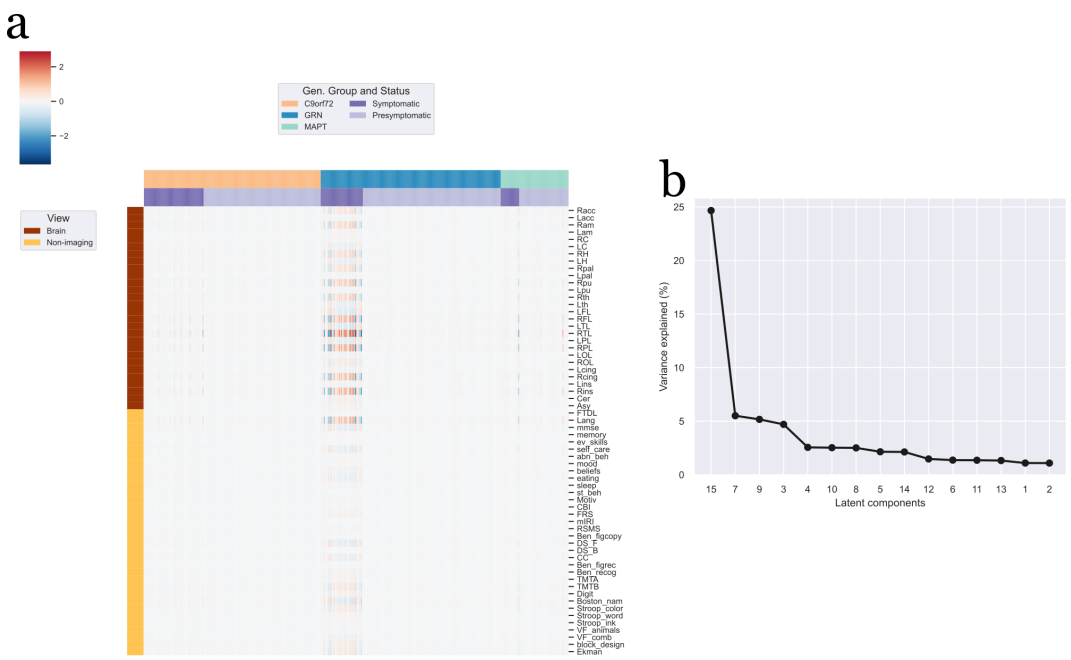

Figure 6.7: Component related to symptomatic *GRN* carriers obtained using sparse GFA and total variance explained by each component. **(a)** Component 5; (b) scree plot of the variance explained by each inferred component. The rows represent the brain (brown) and non-imaging (yellow) features. The columns represent the patients, coloured by genetic group and status.

6.3.2.2 Supervised GFA

Figures 6.8a and 6.8b show the inferred probabilities of the subtypes being associated with the latent components of the symptomatic ($\bar{\mathbf{A}}$) and presymptomatic ($\bar{\mathbf{P}}$) individuals, respectively. 12 (out of 15) components were considered robust using the approach described in Section 6.2.7. These components explained 51.52% of the total variance on the training data. Figure 6.8c shows the percentage of the total variance explained by each component. Here, I will focus on the interpretation of the robust components that showed a probability of being associated with a subtype above 15% (i.e., Components 1, 4, 5, 7 and 9), which together explained 39.54% of variance. Components 4, 5, 7 and 9 explained more variance in the data (Figure 6.8c). Component 1 (which explained 1.97% of variance) obtained a high probability (0.98) of being associated with the symptomatic *GRN* mutation carriers. Component 9 (which explained 4.39% of variance) showed a 0.95 probability of being related to the symptomatic *MAPT* mutation carriers. Component 9 was also associated with presymptomatic *MAPT* (probability of 0.28). Components 4 and 7 (23.42% and 5.46% of variance explained, respectively) obtained probabil-

ities of 0.58 and 0.19, respectively, of being associated with the symptomatic *C9orf72* mutation carriers. In addition, Component 4 showed a 0.77 probability of being related to the presymptomatic *GRN* mutation carriers, but was also slightly associated with the presymptomatic *C9orf72* and *MAPT* mutation carriers (probabilities of 0.19 and 0.17, respectively). Finally, Component 5 explained 4.30% of variance, and was more associated with the presymptomatic *C9orf72* mutation carriers (probability of 0.74), and slightly related to *GRN* and *MAPT* individuals (probability of 0.12 for both groups).

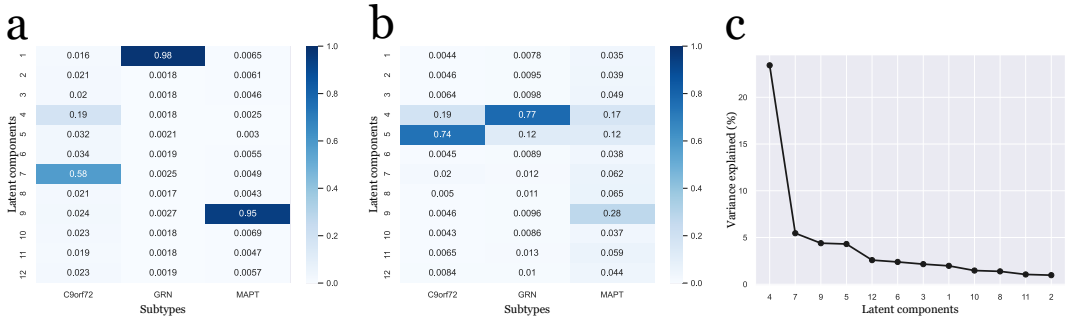

Figure 6.8: The probabilities of the subtypes being associated with the latent components for the **(a)** symptomatic ($\bar{\mathbf{A}}$) and **(b)** presymptomatic ($\bar{\mathbf{P}}$) individuals, and the **(c)** scree plot of the variance explained by each inferred component.

Figure 6.9 shows the component related to the symptomatic *GRN* mutation carriers (which is very similar to Component 5 obtained by sparse GFA, Figure 6.7a). Most of these individuals showed greater grey matter volume (or values above average) in the right frontal lobe, parietal lobe, temporal lobe, insula, cingulate and right subcortical regions (hippocampus, putamen, pallidum, thalamus, amygdala and accumbens) while the same regions on the left hemisphere showed less grey matter volume (or values below average). In terms of non-imaging measures, these *GRN* individuals had higher scores in language, speech and neuropsychological tasks such as trail making and Stroop tasks, and lower scores in self-care, cognitive performance (mini mental state examination (MMSE)), digit span forwards and Boston naming task. Finally, a few other symptomatic *GRN* carriers showed an opposite pattern on the same brain and non-imaging features, i.e., the positive and negative values were flipped.

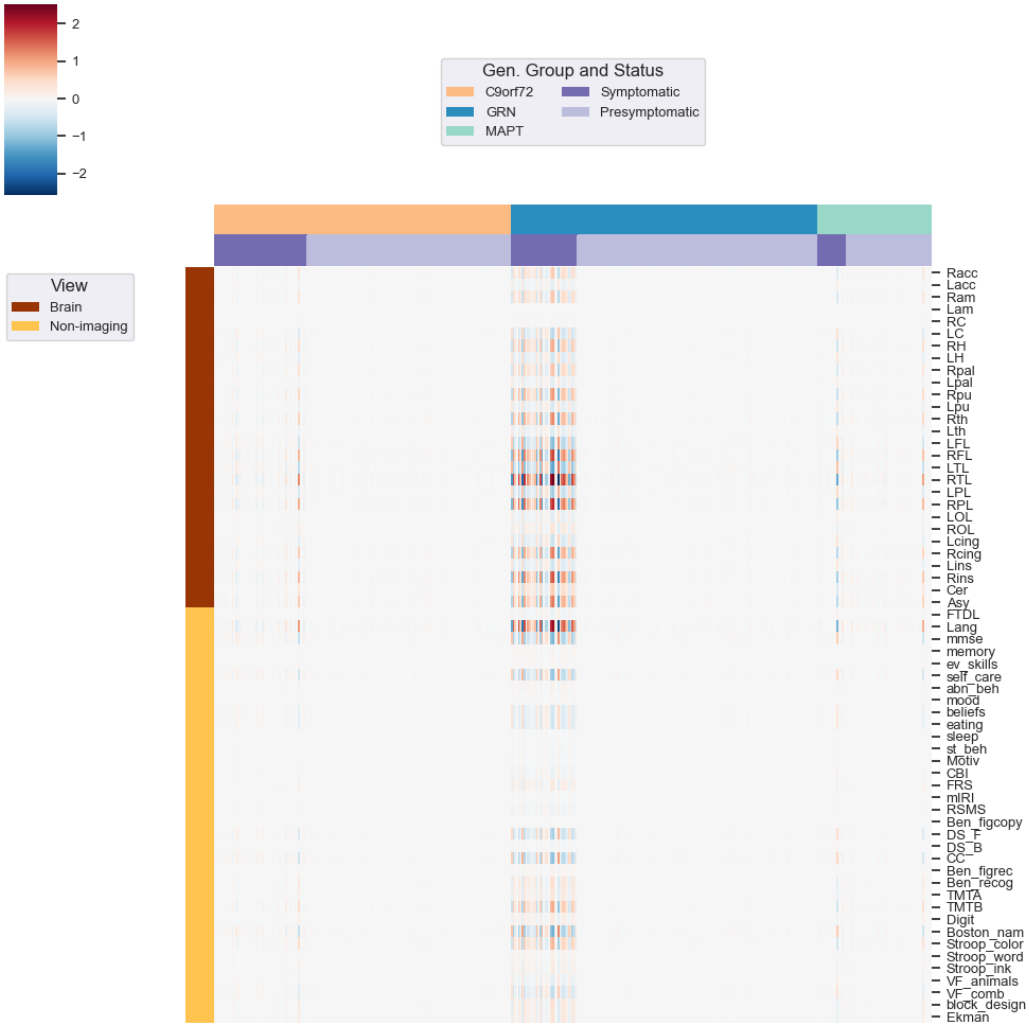

Figure 6.9: Component related to the symptomatic *GRN* mutation carriers. The rows represent the brain (brown) and non-imaging (yellow) features. The columns represent the patients, coloured by genetic group and status. The task name of the abbreviated non-imaging labels can be found in Table C.1 and the full label of the brain imaging features can be found in Table C.2.

Figure 6.10 shows Component 9, which was associated with the symptomatic (probability of 0.95) and presymptomatic (probability of 0.28) *MAPT* mutation carriers, and it is identical to Component 3 obtained by sparse GFA, Figure 6.6d). The symptomatic individuals showed predominantly smaller grey matter volume (values below average) in the right and left temporal lobes, amygdala and hippocampus, and lower scores in the Boston naming task. For some presymptomatic *MAPT* carriers and symptomatic carriers of other genetic groups, the sign of the features described above were flipped.

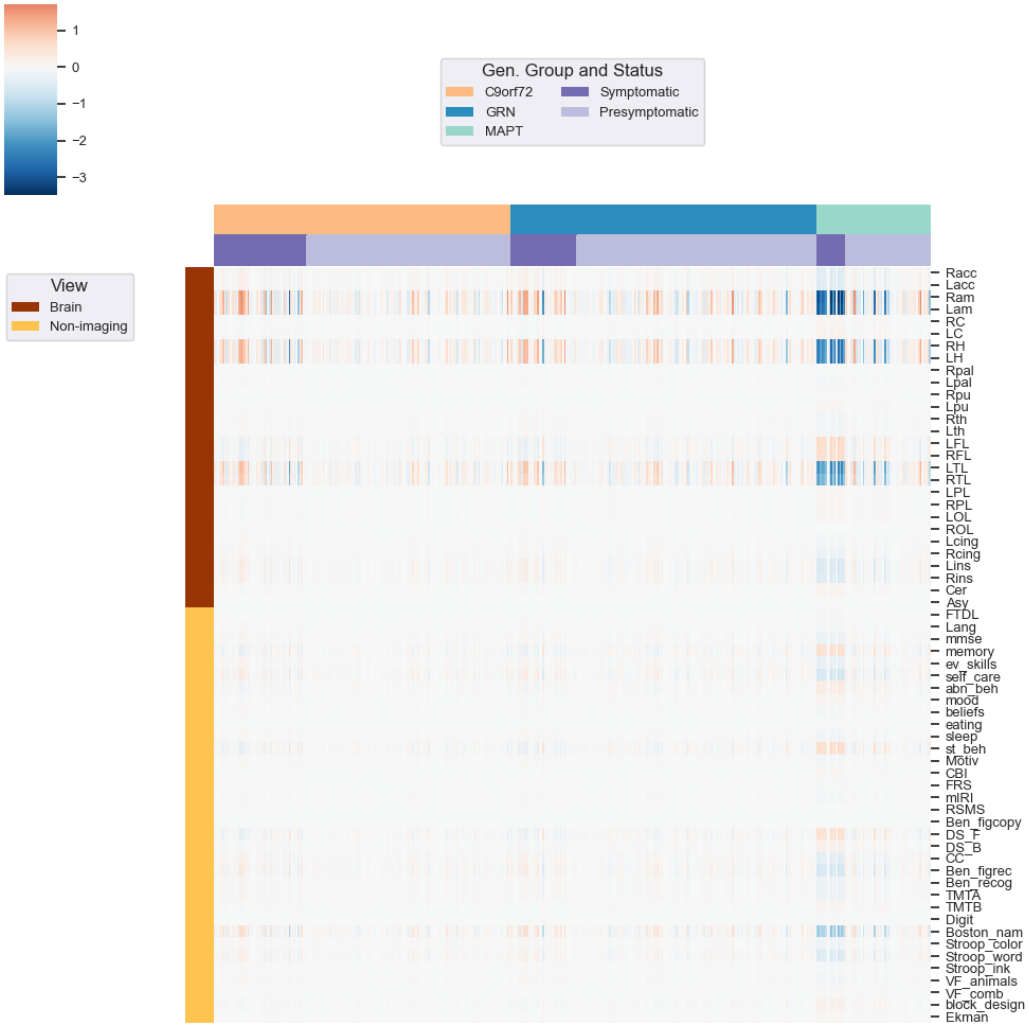

Figure 6.10: Component related to the symptomatic *MAPT* mutation carriers. The rows represent the brain (brown) and non-imaging (yellow) features. The columns represent the patients, coloured by genetic group and status. The task name of the abbreviated non-imaging labels can be found in Table C.1 and the full label of the brain imaging features can be found in Table C.2.

Figure 6.11 shows the component associated with the symptomatic *C9orf72* individuals (similar to Component 7, Figure 6.6b), which is represented by non-imaging features only, i.e., the values of the brain imaging features are very close to zero. The symptomatic *C9orf72* mutation carriers showed higher scores in several behavioural measures (e.g., abnormal, stereotypic behaviour, CBI (general behaviour), mood, motivation, memory, eating, self-care, beliefs and everyday skills), clinical dementia rating (FTDL) and neuropsychological tasks (digit span forwards and backwards). The lower scores for these

individuals were mostly related to disease severity (FRS), social behaviour (RSMS), and cognitive and emotional empathy (mIRI).

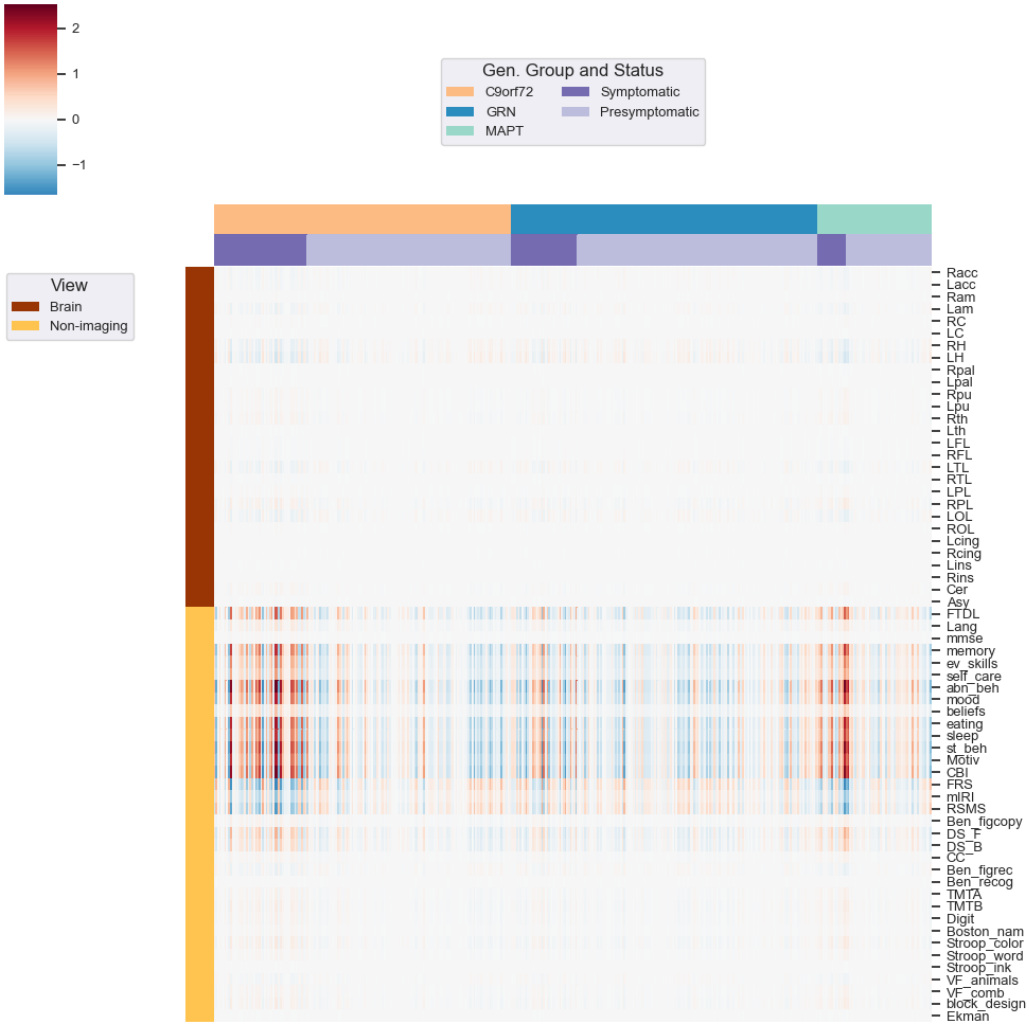

Figure 6.11: Component associated with the symptomatic *C9orf72* mutation carriers. The rows represent the brain (brown) and non-imaging (yellow) features. The columns represent the patients, coloured by genetic group and status. The task name of the abbreviated non-imaging labels can be found in Table C.1 and the full label of the brain imaging features can be found in Table C.2.

Component 4 is shown in Figure 6.12a, and it seems to separate the symptomatic individuals (similar to Component 15 obtained by sparse GFA, Figure 6.6a) from the presymptomatic individuals, where the former showed higher scores mostly in behavioural measures (e.g., CBI, memory, motivation, eating, everyday skills, self-care, abnormal and stereotypic behaviour), executive function tasks (e.g., Stroop task, trail making task), clinical dementia rating

(FTDL), speech and language assessments, and lower scores in most neuropsychological tasks (except for the Stroop and trail masking tasks), the FTD rating scale (disease severity), modified interpersonal reactivity index (cognitive and emotional empathy), revised self monitoring scale (social behaviour) and MMSE. In addition, the symptomatic individuals showed greater asymmetry between left and right hemispheres, smaller grey matter volume in the frontal, temporal, parietal lobes and insula, as well as in several subcortical regions (e.g., hippocampus, putamen and thalamus). The presymptomatic individuals showed an inverted pattern on the same brain and non-imaging features. Figure 6.12b (similar to Component 9 obtained by sparse GFA, Figure 6.6c) displays Component 5, which was more associated with the presymptomatic carriers (particularly *C9orf72*) and reveals smaller grey matter volume in several subcortical regions for these individuals (e.g., hippocampus, caudate, thalamus, amygdala, putamen and pallidum), insula and cerebellum, and lower scores in the revised self monitoring scale, Stroop and trail making tasks for some presymptomatic individuals and an opposite pattern for others.

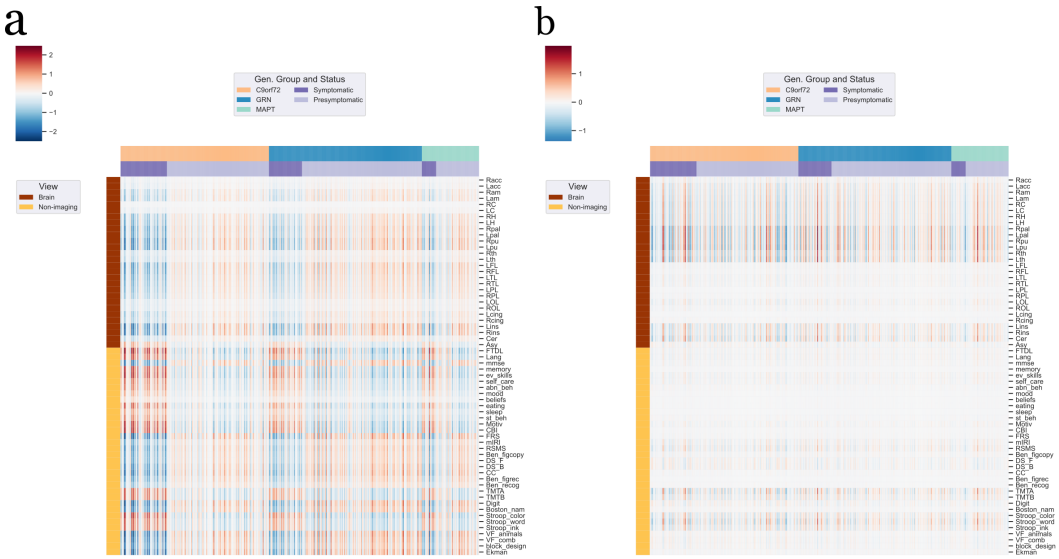

Figure 6.12: Components mostly associated with presymptomatic individuals. **(a)** Component mostly associated with the presymptomatic carriers and symptomatic *C9orf72* carriers (Component 4). **(b)** Component mostly associated with the presymptomatic *C9orf72* individuals (Component 5). The rows represent the brain (brown) and non-imaging (yellow) features. The columns represent the patients, coloured by genetic group and status. The task name of the abbreviated non-imaging labels can be found in Table C.1 and the full label of the brain imaging features can be found in Table C.2.

The predicted probabilities of the symptomatic individuals on the test set are shown in Figure 6.13. The *GRN* mutation carriers were fairly well predicted, i.e., a probability greater than 0.65 was obtained for three *GRN* carriers. Moreover, half of the *C9orf72* carriers obtained a probability greater than 0.50, whereas for the remaining *C9orf72* individuals and *MAPT* carriers the probabilities were below 0.40. Finally, the model was more uncertain about the predictions of the *C9orf72* and *MAPT* carriers.

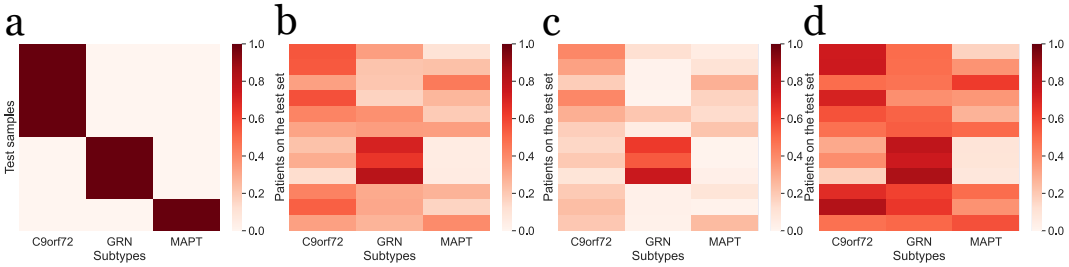

Figure 6.13: Probabilities of the symptomatic individuals on the test set to belong to the underlying subtypes. The test samples are represented on the rows and the subtypes on the columns. **(a)** Test labels; posterior **(b)** mean and quantiles at **(c)** 0.05 and **(d)** 0.95 of the posterior distribution of $\mathbf{\Psi}^{(A)}$ on the test set.

6.4 Discussion

In this study, we proposed a sparse extension of GFA using regularised horseshoe priors to impose feature and view-sparsity, as well as sparsity over the samples, to improve model interpretability and identify components that may characterise different subgroups of patients in the data. We then proposed supervised GFA by including a discriminative module to find components that may describe the pre-defined/underlying subtypes. Although the supervised GFA model is still at the relatively preliminary stage, it seemed to successfully uncover associations between brain structure and non-imaging data (i.e., behaviour, disease severity and cognitive measures) in genetic FTD that characterised the different subtypes, which aligned with the components inferred by sparse GFA. Moreover, both models further identified and characterised within-subtype heterogeneity for the heterogeneous group, i.e., *C9orf72* carriers.

We first showed in synthetic data that supervised GFA is able to correctly infer the data structure of each individual subtype (Figure 6.3) and the model's parameters (Figure 6.4). Supervised GFA correctly inferred the probabilities of a given subtype being associated with latent components (represented in

$\bar{\mathbf{A}}$), which means that supervised GFA can correctly identify the latent components that are associated with subtypes, as well as the probabilities of the samples to belong to a given subtype ($\boldsymbol{\Psi}$, Figure 6.5a-b). In addition, supervised GFA predicted well the probabilities of test samples to belong to the subtypes (Figure 6.5c-d), although the probabilities of some samples obtained higher uncertainty (Figure C.1). In addition, the model was able to infer the true relevant features in each view even when the prior guess of $p_0^{(m)}$ did not match the true value (Figure C.3). Finally, although our sparse (unsupervised) GFA model cannot estimate whether a latent component is associated with an underlying subtype, it can still infer the correct data structure of each subtype and the model's parameters (Figure C.2).

In applying sparse and supervised GFA to the GENFI dataset, we identified five robust components that were associated with the subtypes, which together explained approximately 40% of variance in the data. Although we cannot calculate probabilities of the individuals to belong to the underlying subtypes with sparse GFA, it can identify similar components to those obtained with the supervised model. These findings show that components that are related to subgroups in the data can be identified even if the classes/subtypes are not known or are unreliable.

Two distinct components were associated with the symptomatic *GRN* (Figures 6.7a and 6.9) and *MAPT* (Figures 6.6d and 6.10) mutation carriers, which are known to be homogeneous groups (Mahoney et al., 2012). The symptomatic *GRN* carriers showed a prominent asymmetrical pattern of atrophy in either the left or the right frontal, temporal and parietal lobes, which replicates previous findings (Rohrer et al., 2010; Mahoney et al., 2012; Gordon et al., 2016). These individuals also showed worse executive function (longer TMTB times, longer Stroop colour times), attention (lower digit span forwards), language (impaired performance on Boston Naming task, verbal fluency and C&C tasks as well as a higher score on the PASS language scale), overall cognitive abilities (MMSE) and - not as impaired - on certain everyday behaviours (self-care, beliefs, eating). These results are in line with previous literature which suggests that symptomatic GRN carriers show a diverse range of behavioural, cognitive and language deficits (Rohrer et al., 2010). Language problems are common (in left sided cases) in this mutation group compared to the others. The symptomatic *MAPT* carriers showed a very symmetrical pattern of atrophy involving the temporal lobe volume, as well as in the amygdala and hippocampus, as reported in previous findings (Rohrer et al., 2010; Rohrer

and Rosen, 2013; Whitwell et al., 2009). The *MAPT* individuals showed worse object naming than other genetic groups (as measured with the Boston Naming Task) which has been also found in previous work (Rohrer et al., 2015) and can even dissociate this mutation carrier group from the others at a presymptomatic stage (Bouzigues et al., in prep.). The symptomatic *C9orf72* mutation carriers (Figures 6.6b and 6.11), showed worse overall behaviour as measured by the clinical dementia rating, FTD rating scale and Cambridge Behavioural Index (CBI, where the overall score was worse, as well as the scores of nine out of ten individual items). In addition, the symptomatic *C9orf72* carriers showed worse social cognition (as measured by the revised self monitoring scale (RSMS) and modified interpersonal reactivity index (mIRI)) and working memory (digit span backwards) and attention (digit span forwards). These results are in keeping with previous studies, where *C9orf72* carriers have been shown to have empathy deficits (measured by the mIRI), social cognition impairments (measured by the RSMS) and emotion recognition impairments in the prodromal phase (Russell et al., 2020; Franklin et al., 2021) which the other genetic groups do not show. *C9orf72* carriers are a very heterogeneous group, which might explain why the probability of this component is slightly lower, and no brain regions were found to particularly dissociate them from the other genetic groups.

We also identified two more components that were mostly associated to presymptomatic individuals. The component shown in Figures 6.6a and 6.12a explained almost half of the total variance explained by all robust components, and it seems to dissociate the symptomatic individuals from the presymptomatic individuals. The symptomatic individuals showed worse overall cognition and behaviour (CBI scores, FTLD, Stroop colour time and TMT times), which is expected, for instance, in the symptomatic *C9orf72* carriers as they usually show overall heterogeneous cognitive and behavioural changes, as mentioned above. In addition, these individuals showed asymmetrical widespread brain atrophy involving left and right cortical (e.g., frontal and temporal lobes), subcortical regions (e.g., thalamus) and cerebellum, which has been reported in previous studies (Mahoney et al., 2012; Rohrer and Rosen, 2013). Finally, the component shown in Figures 6.6c and 6.12b characterises presymptomatic carriers according to their executive function, attention processing and social cognition, and grey matter volume changes in the left and right subcortical structures as well as insula and cerebellum. Early subcortical involvement, particularly in *C9orf72* carriers, is in line with previous work (Bocchetta et al.,

2021). Bocchetta and colegues found that in all three groups, subcortical involvement can be identified early in the disease course, particularly in *C9orf72* and *MAPT* mutation carriers, involving volume changes in thalamic subnuclei, cerebellum, hippocampus, amygdala and hypothalamus. *C9orf72* carriers were found to show the earliest and most widespread changes, including the thalamus, basal ganglia and medial temporal lobe. These components do not dissociate presymptomatic carriers, most likely because presymptomatic changes are subtle, and the early changes differ according to the genetic group, as shown in previous works (Rohrer et al., 2015; Young et al., 2018). These findings altogether show that sparse and supervised GFA can successfully identify latent components that are characteristic of specific subtypes, as well as reveal heterogeneous characteristics within a specific subtype.

The supervised GFA model's predictions of the individual probabilities on the test set for the *GRN* group were reasonably good, but they were not as good for the other two subtypes (Figure 6.13). The model was also more uncertain about the predictions of the non-*GRN* mutation carriers. This may be explained by the fact that the *C9orf72* subtype is more heterogeneous, and the sample size of this study was small. Moreover, there is recent evidence that distinct groups might characterise the *MAPT* subtype (Young et al., 2021).

In summary, sparse and supervised GFA can be used to uncover latent dimensions of brain-behaviour associations that provide insights about diseases mechanisms and improve the characterisation of diseases subtypes. In addition, these models enable within-subtype heterogeneity to be characterised, potentially leading to the identification of new subtypes. Supervised GFA may be used to improve patient stratification because it can output individualised predictions of the patients, and estimate uncertainty about them. Finally, both models are interpretable as they select subsets of features within each view, can be easily extended to more complex models, and can be applied to other neuroimaging tasks or fields of research.

Chapter 7

Conclusions

This chapter presents a summary of the main contributions of this thesis (Section 7.1), along with future research directions (Section 7.2).

7.1 Summary of the Main Contributions

This thesis presents classical multi-view methods and latent variable models to uncover associations among multiple views, which can provide insights about underlying dimensions of disease. Moreover, I have shown in this thesis that these methods can be used to identify latent components that characterize disease subtypes and estimate uncertainty about individualized predictions to improve patient stratification. These models could potentially be used in the future to identify disease subtypes.

In Chapter 3, we applied CCA (coupled with PCA) to find associations between resting-state functional MRI and non-imaging data (including behavioural, cognitive and demographic measures) in a sample of healthy and clinically depressed adolescents and young adults. We identified two positive-negative brain-behaviour modes of covariation: the first mode related externalisation/ internalisation symptoms, age and sex to attentional and frontoparietal networks, as well as to subcortical and limbic regions; the second mode related well-being/distress and age to many default mode regions. This work shows the potential of classical multi-view methods, such as CCA, to provide a better understanding of the underlying dimensions of depression in adolescence and young adulthood.

In Chapter 4, I applied sparse CCA to the same dataset and compared two approaches to optimise its regularisation parameters. In this work, I showed that the choice of the optimisation strategy and criterion might influence the results of sparse methods, where approaches based on the optimisation of met-

rics on validation sets are stricter than those based on the whole data set. The experiments showed that the latter approach may be prone to overfitting and false positive findings, whereas the former may lead to false negative findings.

In Chapter 5, I proposed an extension of GFA to address some limitations of CCA, such as control for model complexity (i.e., infer the number of relevant/robust associations), explore variability within views and handle missing data. I showed that the proposed GFA model is able to handle missing data in different scenarios, and it can: (1) uncover associations between high dimensional brain functional connectivity data and non-imaging measures (e.g., demographics, psychometrics and other behavioural measures); (2) predict non-imaging measures from brain functional connectivity. Moreover, we were able to replicate previous findings obtained in a subset of the Human Connectome Project dataset using CCA (Smith et al., 2015).

In Chapter 6, we proposed a sparse extension of GFA to impose feature and view-sparsity, as well as sparsity over the samples, to uncover sparse associations among multiple views and identify components that characterised subsets of samples and could be expressed at the individual level. In addition, we proposed supervised GFA by including a discriminative module to find latent components that describe pre-defined subtypes and explore within-subtype variability. Sparse and supervised GFA uncovered associations between brain structure and non-imaging data (i.e., behaviour, disease severity and cognitive measures) in genetic FTD, identified latent components that described known genotypes and explored within-genotype variability. Moreover, supervised GFA predicted individual probabilities of the patients belonging to the genotypes and estimated uncertainty about those, which shows the potential of supervised GFA to improve patient stratification.

7.2 Future Research Directions

Here, I will describe potential future research directions in terms of further applications of the GFA methods proposed in this thesis (Section 7.2.1), and provide suggestions for further methodological improvements, along with ideas for new methods (Section 7.2.2).

7.2.1 Applications

In this thesis, I have presented GFA applications with two data modalities only (e.g, structural/functional brain MRI and behavioural/cognitive assessments), but these models can be applied to more than two data modalities. For in-

stance, more brain imaging modalities (e.g., diffusion MRI) or other type of data (e.g., genetic or other "omics" data) can be included to potentially uncover more interesting associations between, e.g., multiple imaging modalities (brain structure, function and connectivity) and non-imaging measures (e.g., genetics and behaviour). This would likely increase the predictive power of the models, to predict either missing views or underlying subtypes. Moreover, other type of outcome variables can be used to classify the patients into different categories (e.g. treatment outcome), e.g., one can use sparse GFA to find latent components that describe different subgroups of patients and assess which subgroups respond better to specific treatments.

Both GFA models could also be applied to other datasets, such as the ABCD study (Volkow et al., 2018) or UK Biobank (Miller et al., 2016), to explore more variability in the general population and potentially find different dimensions of psychopathology or population subgroups at risk of developing certain diseases. For example, the new sparse GFA model could be applied to mental health datasets to potentially find subtypes, e.g., in mood or psychotic disorders. In addition, supervised GFA could be applied to datasets, in which treatment response or other relevant outcome measures are collected, for instance, to compute the probability of a given patient responding to one or several treatments. Lastly, in future studies we should ideally use independent replication samples to assess the out-of-distribution generalisation of the models, i.e., assess whether the models are robust to population (or other) shifts on the test set.

Finally, the methods presented and proposed in this thesis can be easily applied to other subfields of neuroscience and other fields of research.

7.2.2 Methodological developments

Future research could also focus on further improvements of the methods. First, in all studies we have assumed that all features were continuous; however, different priors should be included in the models to handle different type of data, e.g., categorical variables. Second, the assumption over the observed labels of the supervised GFA could be extended to allow continuous variables to be predicted, i.e., to assess whether the latent components could be predictive of a clinical score. Third, the supervised GFA formulation is still at relatively preliminary stage. In future work, a fully generative approach could be implemented, for instance, by considering some form of mixture of horseshoe priors over the latent variables so that these could differ among different subgroups.

Moreover, a baseline comparison between supervised GFA and other models should be done in future work. Fourth, supervised GFA could also be extended to incorporate longitudinal data, for instance, to assess how the individual subject probabilities of belonging to each underlying subtype change over time. This could be achieved by considering a Hidden Markov Model, where the probability of each subject transitioning from one subtype to another between different time points could be modelled (Vogelsmeier et al., 2019; Chien et al., 2020). This would allow modelling subjects' disease trajectories and therefore potentially predict disease progression for each subject or identify people at risk of developing illnesses. Lastly, future studies could also explore the development and application of deep multi-view learning methods, such as a deep variational information bottleneck approach for incomplete multi-view observations (DeepIMV) (Lee and van der Schaar, 2021), to integrate more complex representations of the data in incomplete data sets.

Appendix A

Complements to Chapter 3

A.1 Methods

A.1.1 Self-report questionnaires

- **Antisocial Behaviours Checklist**: self-report questionnaire for symptoms of antisocial behaviour based on DSM-IV conduct disorder items. The questionnaire was designed solely for the purpose of the NSPN project (11 items).
- **Antisocial Process Screening Device**: self-report scale measuring psychopathic traits and antisocial behaviour (20 items).
- **Barratt Impulsive Scale**: self-report questionnaire assessing personality and behavioural constructs of impulsiveness (30 items).
- **Child and Adolescent Dispositions Scale**: self-report measure of the three underlying dimensions of cognitive control of behaviour (prosociability, negative emotionality and daring) (57 items).
- **Child Trauma Questionnaire**: self-report inventory screening for histories of abuse and neglect, which covers five types of maltreatment: emotional, physical, and sexual abuse, and emotional and physical neglect (28 items).
- **Drugs Alcohol and Self-Injury**: self-report measure assessing the frequency of drug and alcohol use as well as the frequency, methods and motives of non-suicidal self-harm acts. The questionnaire was designed solely for the purpose of the NSPN project (16 items).

- **Inventory of Callous-Unemotional Traits**: self-report inventory of assessing 3 domains of callous and unemotional traits: callousness, uncaring, and unemotional (24 items).
- **Kessler Psychological Distress Scale**: self-report measurement of psychological distress (10 items).
- **Leyton Obsessional Inventory**: self-report questionnaire measuring obsessional and anxiety symptoms (11 items).
- **Moods and Feelings Questionnaire**: self-report questionnaire measuring depressive symptoms in the last 2 weeks (33 items).
- **Revised Children's Manifest Anxiety Scale**: self-report questionnaire measuring anxiety symptoms (28 items).
- **Rosenberg Self-Esteem Scale**: self-report questionnaire measuring global self-esteem or feelings of self-worth and self-acceptance (10 items).
- **Schizotypal Personality Questionnaire**: self-report scale measuring schizotypal personality traits (74 items).
- **Wechsler Abbreviated Scale of Intelligence**: matrix reasoning and vocabulary subsets of the Wechsler Abbreviated Scale of Intelligence designed to assess fluid and crystallized intelligence, respectively (2 items).
- **Warwick Edinburgh Mental Wellbeing Scale**: self-report instruments spanning the theoretical distribution of common mental symptoms and wellbeing (14 items).

A.2 Figures

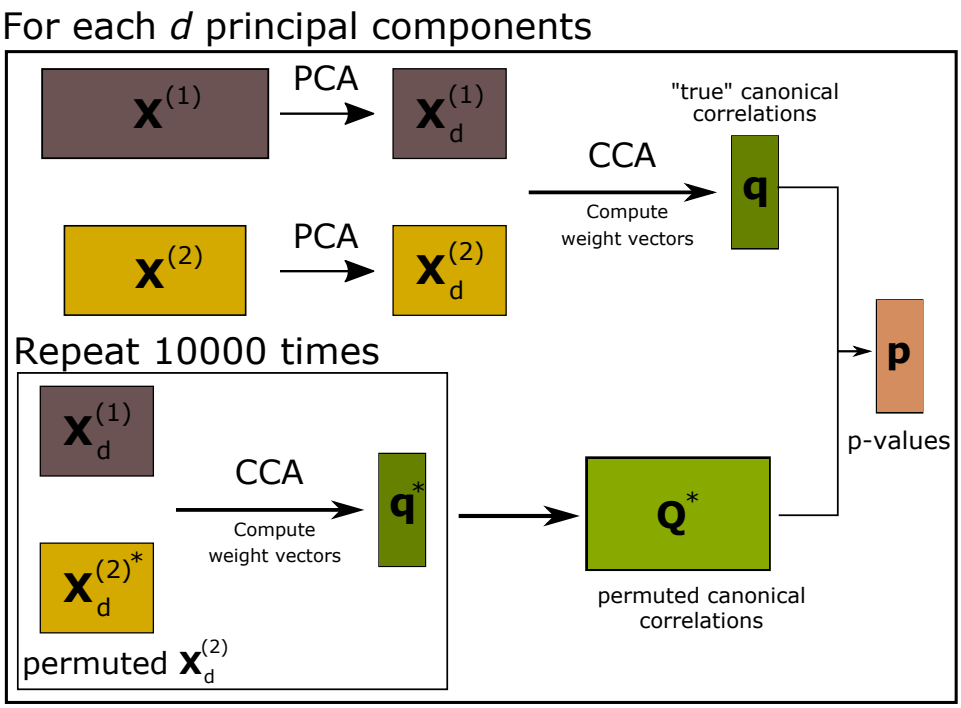

Figure A.1: Statistical framework to jointly optimise the number of principal components and estimate the statistical significance of the CCA modes.

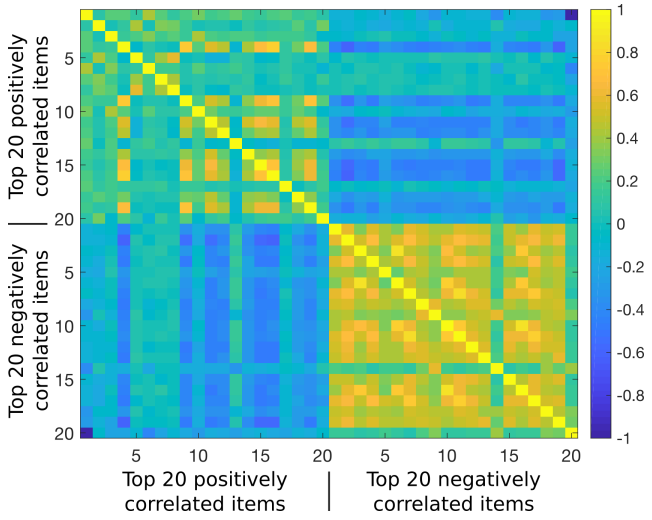

Figure A.2: Correlations between the top 20 positive and top 20 negative behavioural items of the first CCA mode. Male gender (first item) is weakly associated with the other positive behavioural items (items 2-19 with mean correlation=0.20). Female gender (last item) is weakly associated with the other negative behavioural items (items 1-19 with mean correlation=0.17).

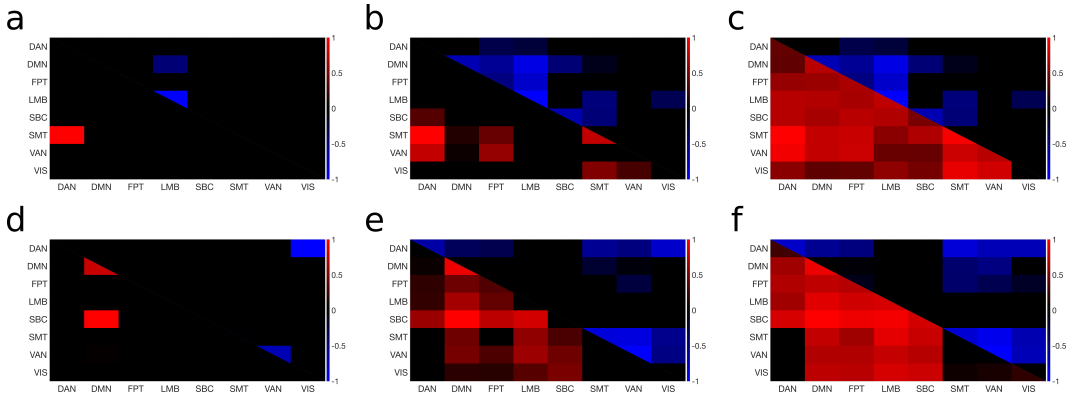

Figure A.3: Mean correlations between and within resting-state networks for the first **(a-c)** and the second **(d-f)** CCA mode at three different levels of top connections: top 20 **(a, d)**, top 0.5% **(b, e)** and top 5% **(c, f)** of most positively/negatively correlated connections. Positive correlations (red) and negative correlations (blue) are summarized separately in the lower and upper triangular matrices, respectively. The mean absolute correlations are log-transformed and normalized for easier comparison between the three levels. Dorsal Attention Network (DAN); Default Mode Network (DMN); Frontoparietal Network (FPT); Limbic Network (LMB); Subcortex (SBC); Somatomotor Network (SMT); Ventral Attention Network (VAN); Visual Network (VIS).

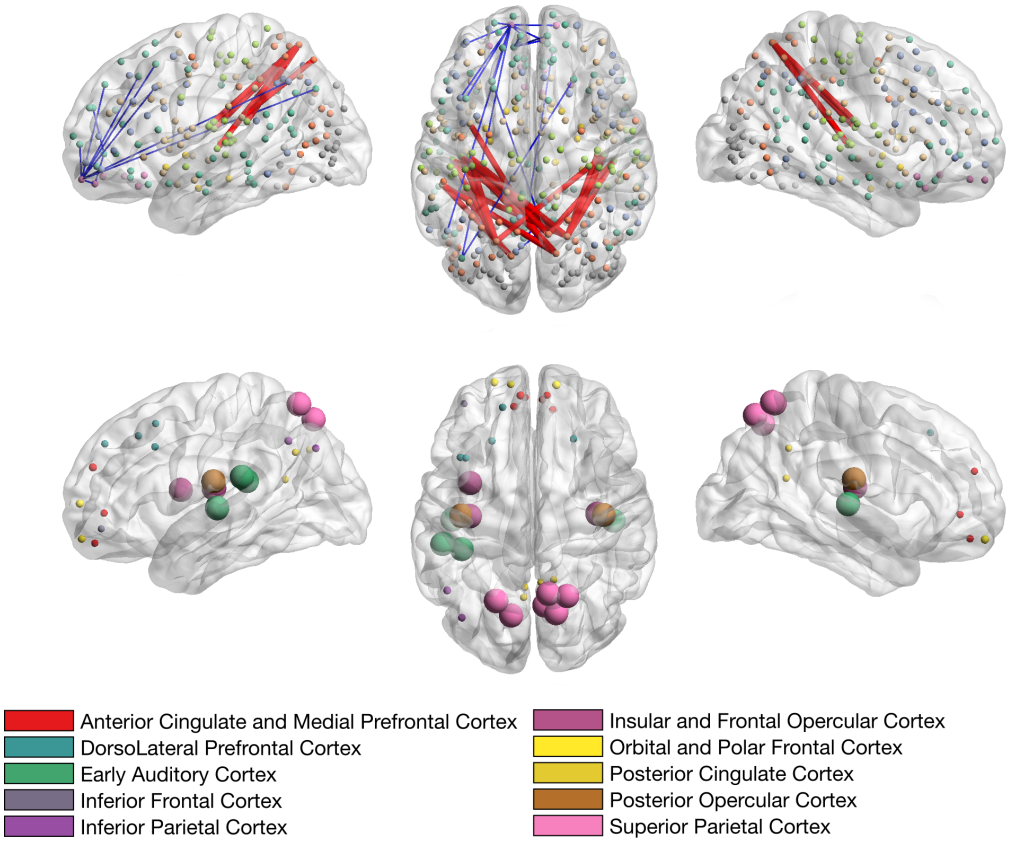

Figure A.4: Correlations between the brain connectivity variables and the brain canonical variate of the first CCA mode in sagittal (left and right) and axial views (middle). Notations are as in Figure 3.3 except that nodes are colour coded by gross anatomical regions used in Glasser et al. (2016).

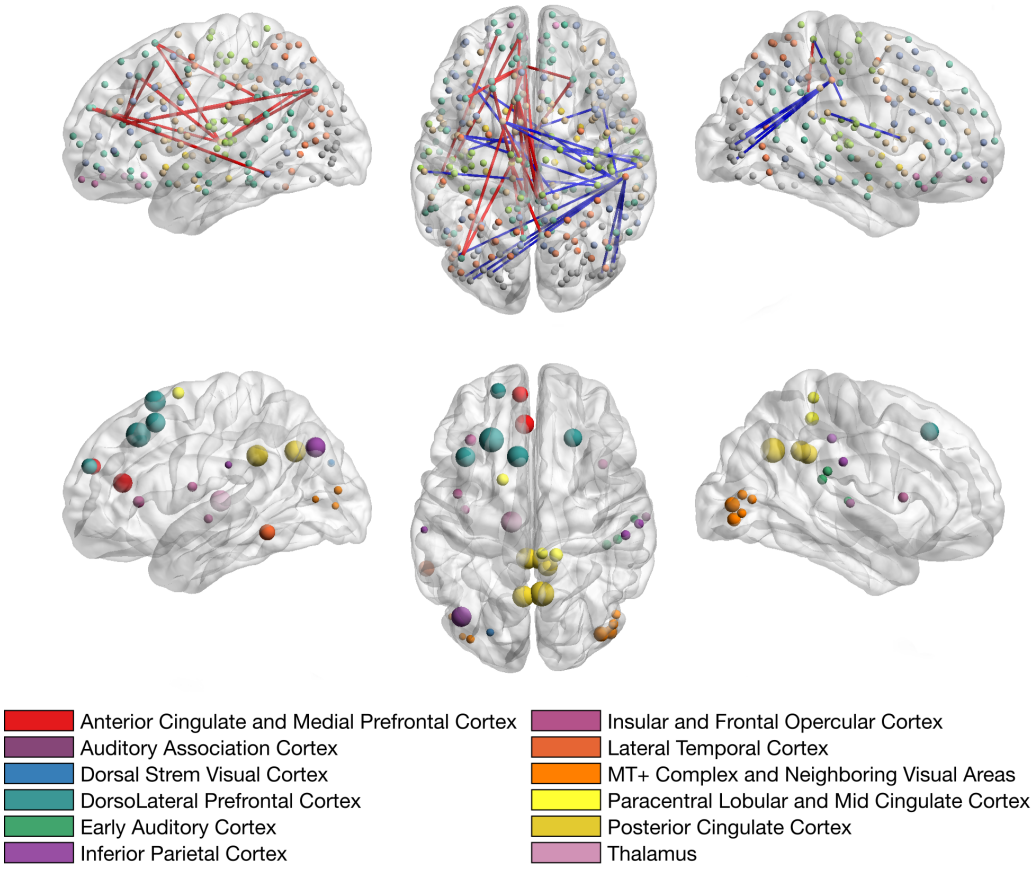

Figure A.5: Correlations between the brain connectivity variables and the brain canonical variate of the second CCA mode in sagittal (left and right) and axial views (middle).

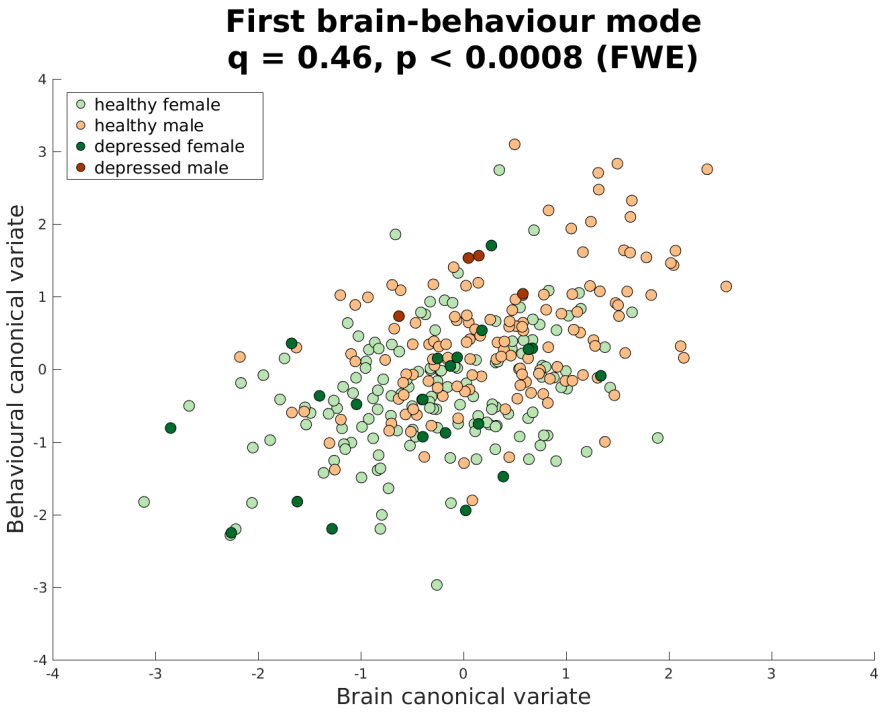

Figure A.6: Brain-behaviour mode of covariation obtained using the machine learning framework. The scatter plot shows the brain and behaviour scores of the first CCA mode (each dot represents a subject). Subjects are colour coded by gender and clinical diagnosis. The canonical hold-out correlation, q, and corresponding p-value are shown on the top of the plot.

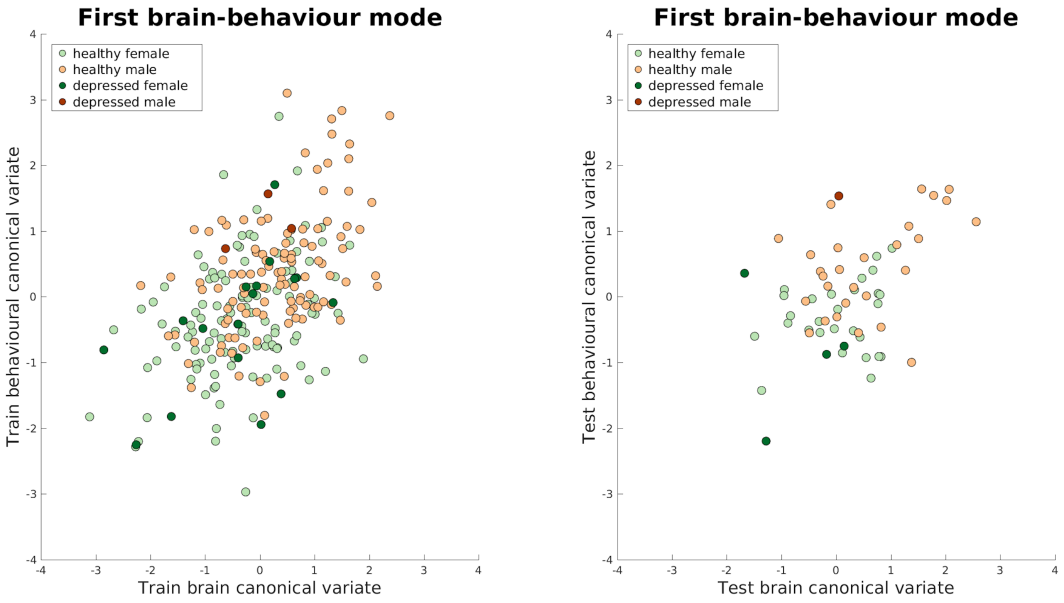

Figure A.7: Brain-behaviour mode of population covariation obtained in the training (left) and test set (right) using the machine learning framework. All the conventions are as in Figure A.6.

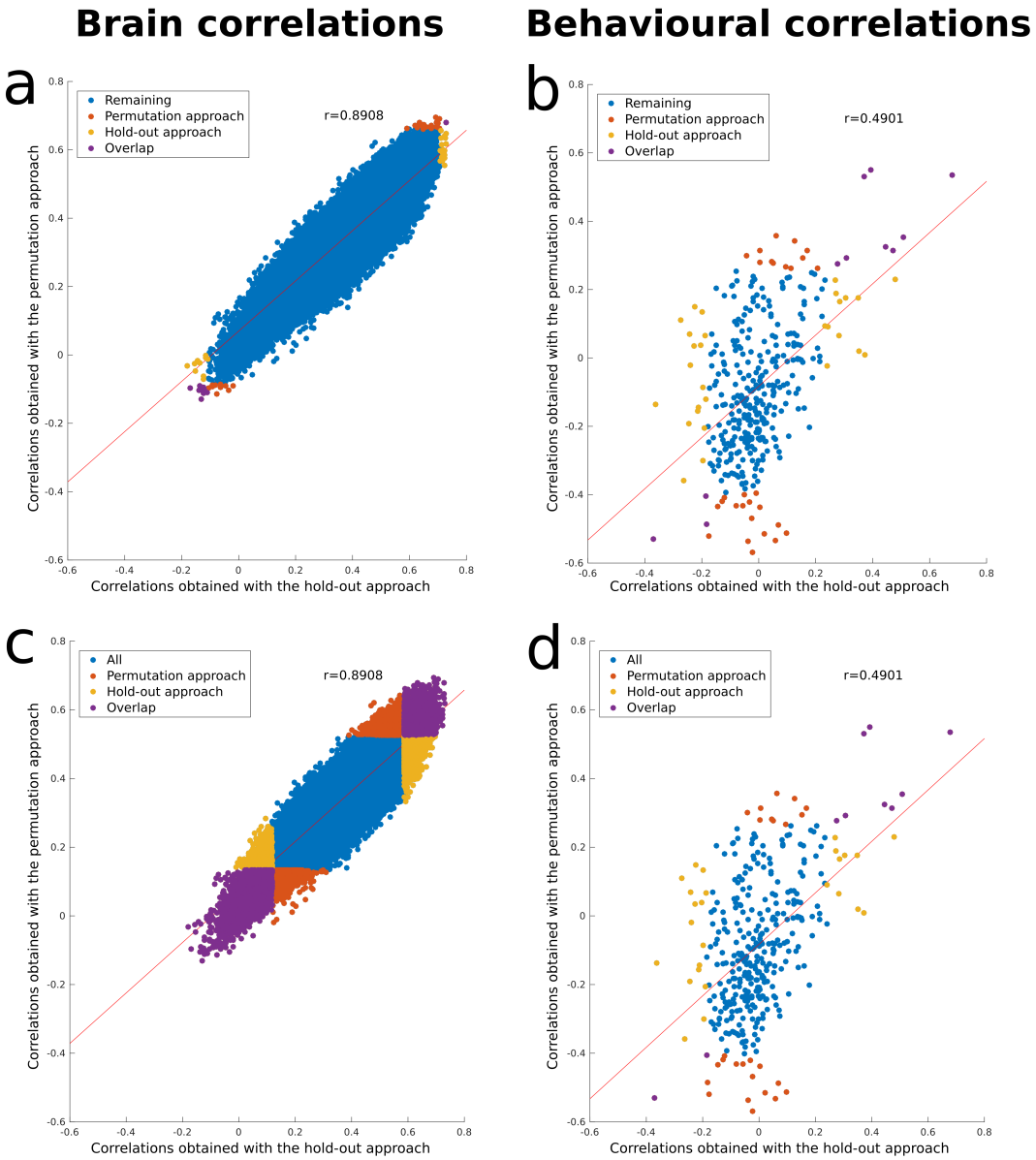

Figure A.8: Brain **(a,c)** and behaviour **(b,d)** correlations of the first CCA mode obtained using the statistical ("permutation approach") and machine learning ("hold-out approach") frameworks. **(a,b)** The overlap (purple) between the top 20 most positively/negatively correlated variables obtained with the statistical (orange) and machine learning (yellow) frameworks is shown; **(c,d)** the same colour scheme is used to show the overlap between the top 5% most positively/negatively correlated variables obtained with the statistical and machine learning frameworks. Blue denotes the remaining variables.

Appendix B

Complements to Chapter 5

B.1 Lower bound for GFA

Considering Equations 2.14-2.15, the lower bound of $\ln p(\mathbf{X})$ is given by:

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

where the expectations of the $\ln p(\cdot)$ terms are given by (see Equations 5.3-5.7):

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

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

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

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

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

where $q(\boldsymbol{\theta}) = q(\mathbf{Z})q(\mathbf{W})q(\boldsymbol{\tau})$, $\langle\ln\tau_j^{(m)}\rangle = \psi(\tilde{a}^{(j)}_{\boldsymbol{\tau}^{(m)}}) - \ln\tilde{b}^{(j)}_{\boldsymbol{\tau}^{(m)}}$, $\langle\ln\alpha_k^{(m)}\rangle = \psi(\tilde{a}_{\boldsymbol{\alpha}^{(m)}}) - \ln\tilde{b}^{(k)}_{\boldsymbol{\alpha}^{(m)}}$, $\Gamma(\cdot)$ is a Gamma function 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 Equations 5.8, 5.12, 5.13 and 5.18, respectively.

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

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

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

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

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

B.2 Methods

B.2.1 Additional GFA experiments on synthetic data

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

1. *Complete data* (all models were initialised with $K = 30$):
   (a) low dimensional data ($D_1 = 50$ and $D_2 = 30$) was generated using the same parameters described in Section 5.2.4.
   (b) high dimensional data was generated ($D_1 = 20000$ and $D_2 = 200$) using the same parameters described in Section 5.2.4.
2. *Incomplete data* (all models were initialised with $K = 15$):

(a) the elements of $\mathbf{X}^{(2)}$ deviating more than 1σ (i.e., standard deviation) from the mean (i.e, $x_{dn} > \mu + \sigma$ and $x_{dn} < \mu - \sigma$) were removed from the synthetic data generated in experiment 1a, which led to approximately 30% of missing values in $\mathbf{X}^{(2)}$.

(b) 10% of the rows of $\mathbf{X}^{(1)}$ and 20% of the elements of $\mathbf{X}^{(2)}$ were randomly removed from the synthetic data generated in experiment 1a.

(c) 10% of the rows of $\mathbf{X}^{(1)}$ and 20% of the elements of $\mathbf{X}^{(2)}$ were randomly removed from the high dimensional data generated in experiment 1b.

B.2.2 CCA experiments on synthetic data

In order to assess the CCA performance in complete and incomplete data sets, we generated data using the parameters described in Section 5.2.4 and ran experiments on the following selections of the data:

- *Complete data*
- *Incomplete data*
  - 20% of the elements of $\mathbf{X}^{(1)}$ and 40% of the elements of $\mathbf{X}^{(2)}$ were randomly removed.
  - the elements of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ deviating more than 1σ from the mean were removed, which led to approximately 30% of missing values in each data modality.

The missing values were imputed using the median. The statistical significance of the CCA modes was estimated by permutation inference, in which the rows of $\mathbf{X}^{(2)}$ were permuted 1000 times and CCA was run after each permutation. For each CCA mode, we compute a p-value to assess whether the "true" canonical correlation (i.e., the canonical correlation of the respective CCA mode obtained without permuting the data) was larger than the null distribution of permuted canonical correlations of the first CCA mode. To obtain an equivalent representation of a single latent variable for CCA (comparable to a latent component in GFA), 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 $\mathbf{V} \in \mathbb{R}^{D_2 \times K}$, were averaged. These experiments were also run using our GFA extension without imputing the values. The incomplete data experiments were different from those described in Section 5.2.4 because we

wanted to show the potential of GFA to handle missing data when more than one modality has missing values. Moreover, it would be of little interest to run CCA with missing rows because, in practice, one would not impute values but rather remove the rows in both data modalities.

B.2.3 Surface plots

The surface plots illustrate maps of brain connection strength increases/decreases, which were obtained by weighting each node's parcel map by the GFA/CCA edge-strengths (the loadings were multiplied by the sign of the population mean correlation) summed across the edges connected to the node. We used the node's parcel maps provided as a cifti file (named *melodic_IC_ftb.dlabel.nii*) in the group ICA folder (named *groupICA_3T_HCP1200_MSMAll_d200.ica*). In this file, one can find the number of the ICA component that each vertex is most likely to belong to.

B.3 Results

B.3.1 Additional GFA experiments on synthetic data

The model parameters were correctly inferred using low (Figure B.1a) ($\hat{\tau}^{(1)} \approx 5.10$ and $\hat{\tau}^{(2)} \approx 9.98$) and high (Figure B.1b) ($\hat{\tau}^{(1)} \approx 5.01$ and $\hat{\tau}^{(2)} \approx 9.97$) dimensional synthetic data, when the model was initialised with $K = 30$. The most relevant shared and view-specific components were correctly estimated in both experiments.

In the experiment 2a, taking into account the difficulty of the task our GFA approach recovered the model parameters fairly well ($\hat{\tau}^{(1)} \approx 5.04$ and $\hat{\tau}^{(2)} \approx 11.72$), whereas the median imputation approach failed to estimate the noise parameter of the second view ($\hat{\tau}^{(1)} \approx 5.03$ and $\hat{\tau}^{(2)} \approx 6.95$) and the third component (i.e. the component specific to $\mathbf{X}^{(2)}$) was erroneously identified (Figure B.2a). Furthermore, our GFA extension performed better in the multi-output prediction task (Table B.1). The model predicted missing data accurately ($\rho = 0.929 \pm 0.021$).

In the experiments 2b (Figure B.2b) and 2c (Figure B.2c), the proposed GFA approach inferred the model parameters correctly in low ($\hat{\tau}^{(1)} \approx 5.01$ and $\hat{\tau}^{(2)} \approx 10.15$) and high dimensional ($\hat{\tau}^{(1)} \approx 5.03$ and $\hat{\tau}^{(2)} \approx 9.97$) data sets, respectively. The median imputation approach failed to infer the model parameters in both experiments ($\hat{\tau}^{(1)} \approx 6.23$ and $\hat{\tau}^{(2)} \approx 6.39$ in experiment 2b (Figure B.2b); $\hat{\tau}^{(1)} \approx 6.33$ and $\hat{\tau}^{(2)} \approx 3.95$ in experiment 2c (Figure B.2c)). The performance of both approaches in the multi-output prediction task was sim-

ilar and below chance level (Table B.1). The model predicted reasonably well the missing observations in both views (experiment 2b: $\rho = 0.675 \pm 0.031$ and $\rho = 0.779 \pm 0.022$ for the missing values in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively; experiment 2c: $\rho = 0.627 \pm 0.012$ and $\rho = 0.859 \pm 0.003$ for the missing values in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively).

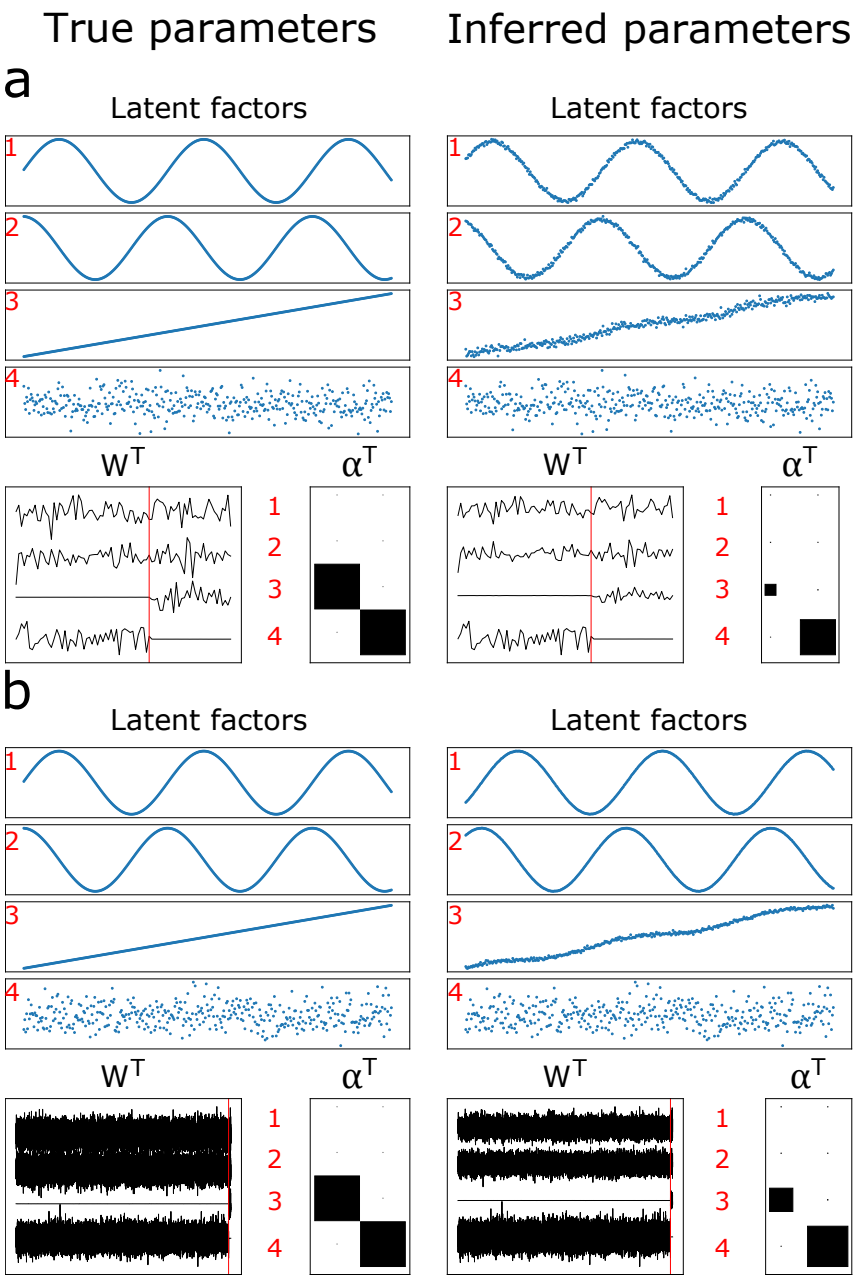

Figure B.1: True and inferred latent components and model parameters obtained in the experiments 1a **(a)** and 1b **(b)** described in Section B.2.1. The latent components and parameters used to generate the data are plotted on the left-hand side and those inferred by the model are plotted on the right-hand side. The four rows on the top represent the four latent components. 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 on the first and second column of $\boldsymbol{\alpha}^T$, respectively. The small black dots and big black squares represent active and inactive latent components, respectively.

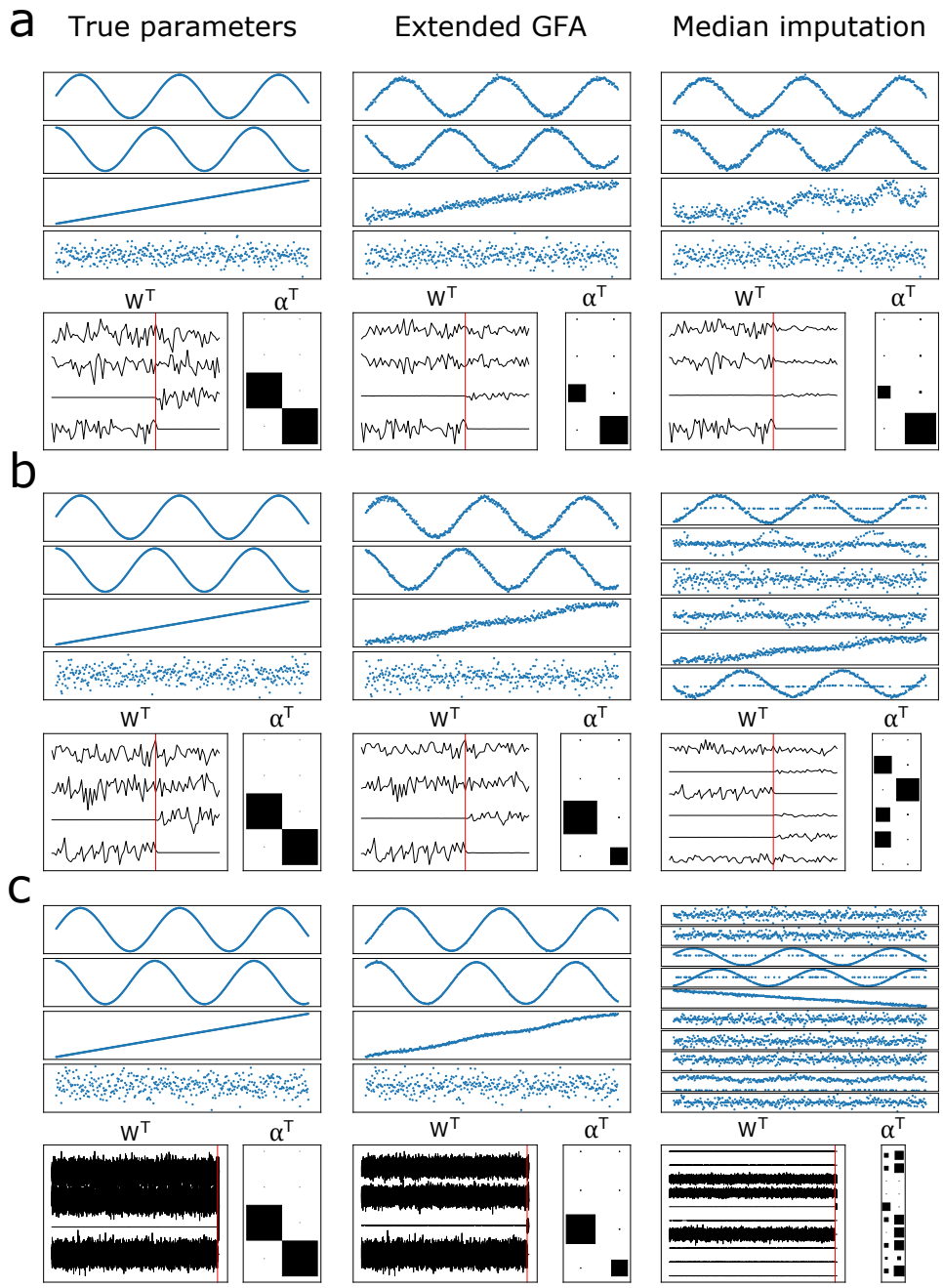

Figure B.2: True and inferred latent components and model parameters obtained in the experiments 2a **(a)**, 2b **(b)** and 2c **(c)** described in Section B.2.1. **(Left column)** latent components and model parameters used to generate the data. **(Middle column)** latent components and parameters inferred using the proposed GFA approach. **(Right column)** latent components and parameters inferred using the median imputation approach. The loading matrices ($\mathbf{W}^T$) and alphas ($\boldsymbol{\alpha}^T$) can be interpreted as in Figure B.1.

Table B.1: Prediction errors of the multi-output prediction tasks obtained in the experiments 2a-c described in Section B.2.1. The values correspond to the mean and standard deviation of the MSEs across 10 initialisations. The first (second) column shows the MSE between the test observations $\mathbf{X}^{(1)^\star}$ ($\mathbf{X}^{(2)^\star}$) and the mean predictions $\langle \mathbf{X}^{(1)^\star} | \mathbf{X}^{(2)^\star} \rangle$ ($\langle \mathbf{X}^{(2)^\star} | \mathbf{X}^{(1)^\star} \rangle$). ours - proposed GFA approach; imputation - median imputation approach; chance - chance level.

		Predict $\mathbf{X}^{(1)}$ from $\mathbf{X}^{(2)}$	Predict $\mathbf{X}^{(2)}$ from $\mathbf{X}^{(1)}$
Exp. 2a	ours	**1.35 ± 0.34**	**0.92 ± 0.16**
	imputation	2.87 ± 1.01	1.56 ± 0.28
	chance	2.33 ± 0.45	2.04 ± 0.36
Exp. 2b	ours	1.21 ± 0.10	0.78 ± 0.15
	imputation	1.16 ± 0.07	0.79 ± 0.15
	chance	2.23 ± 0.12	2.34 ± 0.36
Exp. 2c	ours	1.26 ± 0.06	0.84 ± 0.04
	imputation	1.17 ± 0.03	0.85 ± 0.04
	chance	2.27 ± 0.02	2.35 ± 0.10

Table B.2: Most relevant shared and view-specific components obtained with the complete high dimensional synthetic data (experiment 1b in Section B.2.1) according to the proposed criteria. Components explaining more than 7.5% variance within any view were considered most relevant. A component was considered shared if $0.001 \leq r_k \leq 300$, specific to $\mathbf{X}^{(2)}$ if $r_k > 300$ or specific $\mathbf{X}^{(1)}$ if $r_k < 0.001$. rvar - relative variance explained; var - variance explained; r_k - ratio between the variance explained by $\mathbf{w}_k^{(2)}$ and $\mathbf{w}_k^{(1)}$.

	Components	**rvar** (%)		**var** (%)		$\mathbf{r_k}$
		$\mathbf{X}^{(1)}$	$\mathbf{X}^{(2)}$	$\mathbf{X}^{(1)}$	$\mathbf{X}^{(2)}$	$\text{var}_{\mathbf{w}_k^{(2)}}/\text{var}_{\mathbf{w}_k^{(1)}}$
Shared	1	25.09	46.03	15.39	0.19	0.01
	2	25.47	35.44	15.62	0.15	9.6×10^{-3}
Specific	3	2.88×10^{-4}	18.53	1.76×10^{-4}	0.08	442.85
	4	49.44	8.80×10^{-5}	30.32	3.71×10^{-7}	1.22×10^{-8}

B.3.2 CCA experiments on synthetic data

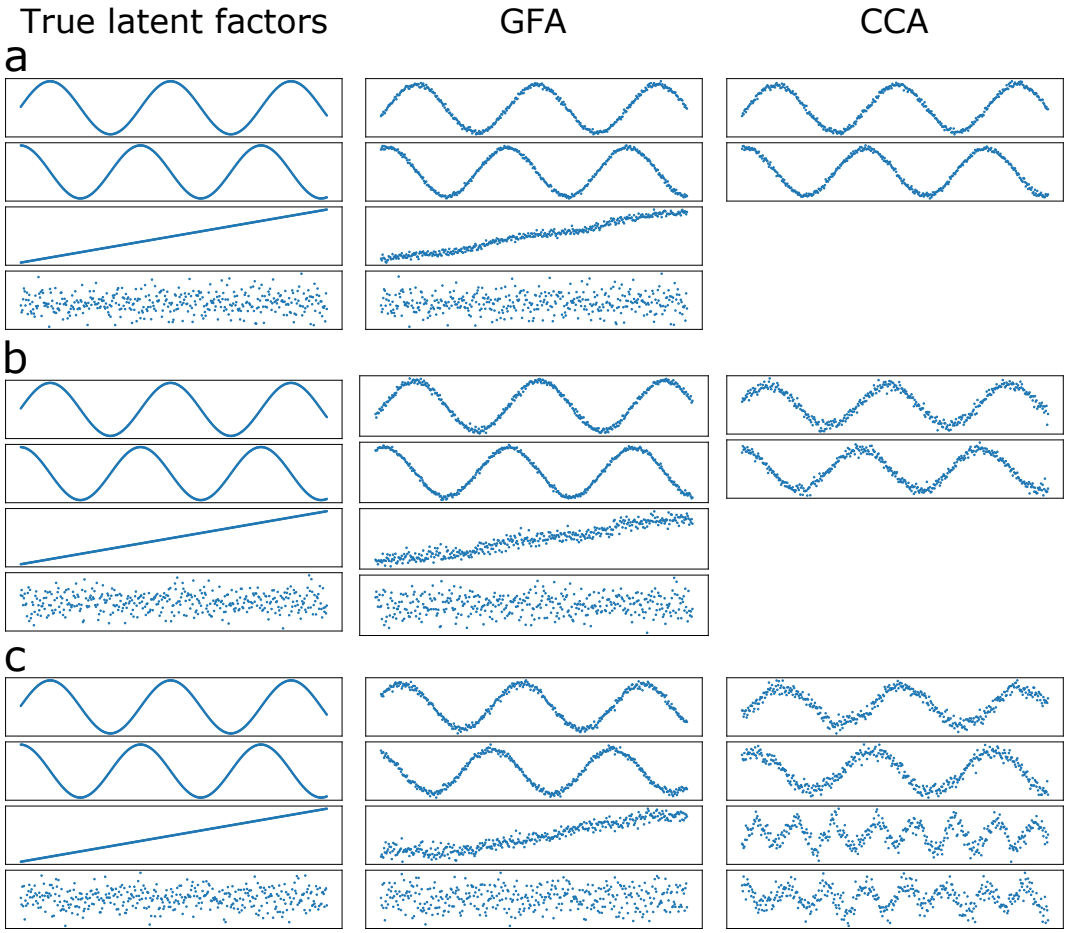

Figure B.3: True and inferred latent components obtained using CCA and GFA in synthetic data. **(Left column)** latent components used to generate the data. **(Middle column)** latent components inferred using the proposed GFA approach. **(Right column)** latent components inferred using CCA. **(a)** experiment using complete data; **(b)** experiments using incomplete data, where 20% of the elements of $\mathbf{X}^{(1)}$ and 40% of the elements of $\mathbf{X}^{(2)}$ were randomly removed; **(c)** experiment using incomplete data, where the elements of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ deviating more than 1σ from the mean were removed.

B.3.3 GFA experiments on the HCP data

Table B.3: Most relevant shared and modality-specific components obtained in the HCP experiment 2a (Section 5.2.5) according to the proposed criteria. Components explaining more than 7.5% variance within any data modality were considered most relevant. A component was considered shared if $0.001 \leq r_k \leq 300$, specific to non-imaging (NI) measures 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 k-th component in the non-imaging and brain data.

	components	**rvar** (%) Brain	**rvar** (%) NI	**var** (%) Brain	**var** (%) NI	$\mathbf{r_k}$ $\text{var}_{\text{NI}}/\text{var}_{\text{brain}}$
Shared	a	0.159	9.44	0.012	0.028	2.42
	b	0.065	18.152	0.005	0.005	11.32
	c	0.036	10.539	0.003	0.031	12.04
	d	0.015	39.330	0.001	0.117	105.10
Brain	a	13.531	6.60×10^{-5}	0.988	1.97×10^{-7}	1.99×10^{-7}
	b	12.269	0.001	0.896	4.19×10^{-6}	4.68×10^{-6}

Table B.4: Most relevant shared and modality-specific components obtained in the HCP experiment 2b (Section 5.2.5). Components explaining more than 7.5% variance within any data modality were considered most relevant. A component was considered shared if $0.001 \leq r_k \leq 300$, specific to non-imaging (NI) measures 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 k-th component in the non-imaging and brain data.

	Components	**rvar** (%) Brain	**rvar** (%) NI	**var** (%) Brain	**var** (%) NI	$\mathbf{r_k}$ $\text{var}_{\text{NI}}/\text{var}_{\text{brain}}$
Shared	a	0.149	7.643	0.007	0.028	3.83
	b	0.034	16.550	0.002	0.060	36.19
	c	0.016	8.670	7.82×10^{-4}	0.031	40.11
	d	0.019	31.255	8.99×10^{-4}	0.113	125.82
Brain	a	15.625	0.0350	0.758	1.27×10^{-4}	1.67×10^{-4}
	b	14.979	0.177	0.727	6.42×10^{-4}	8.83×10^{-4}

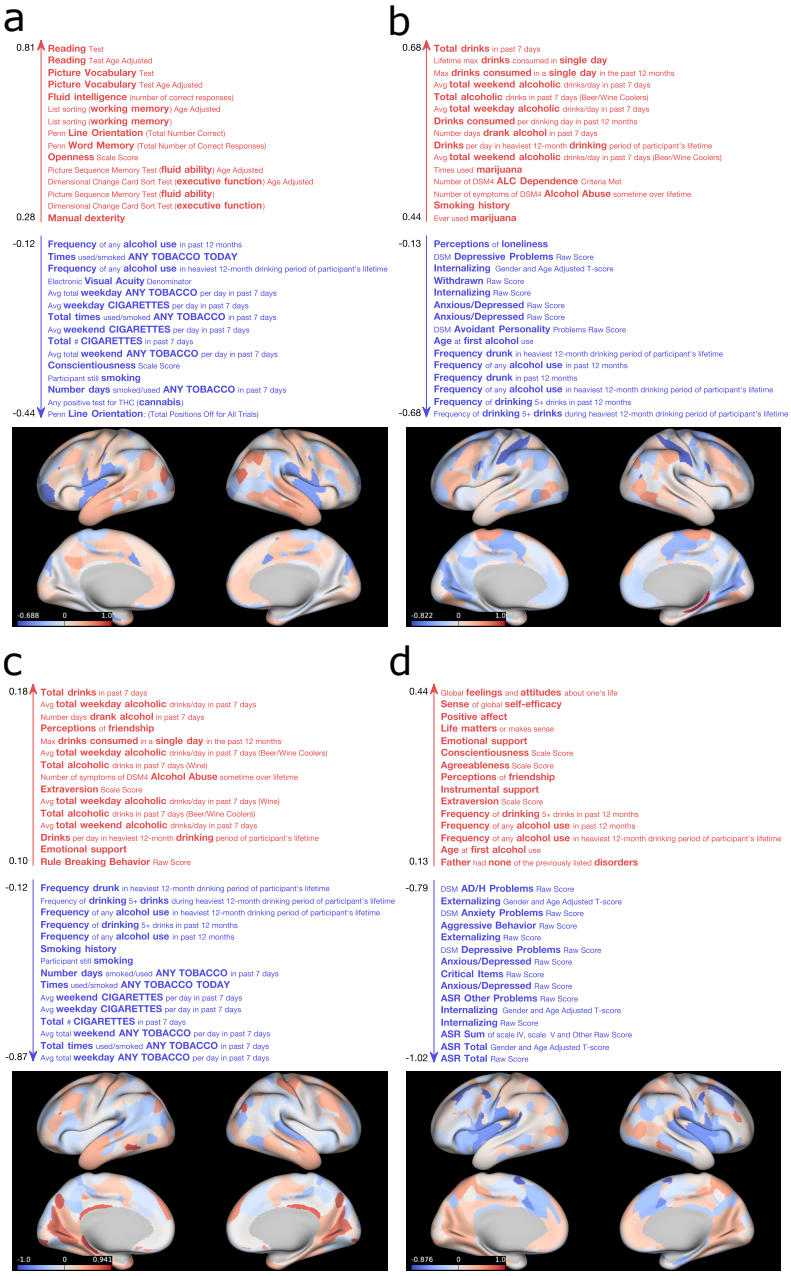

Figure B.4: Non-imaging measures and brain networks described by the first **(a)**, second **(b)**, third **(c)** and fourth **(d)** shared GFA components obtained in the incomplete data HCP experiment 2a described in Section 5.2.5. For illustrative purposes, the top and bottom 15 non-imaging measures for each component are shown. The brain surface plots represent maps of brain connection strength increases/decreases, which were obtained as described in Section B.2.3.

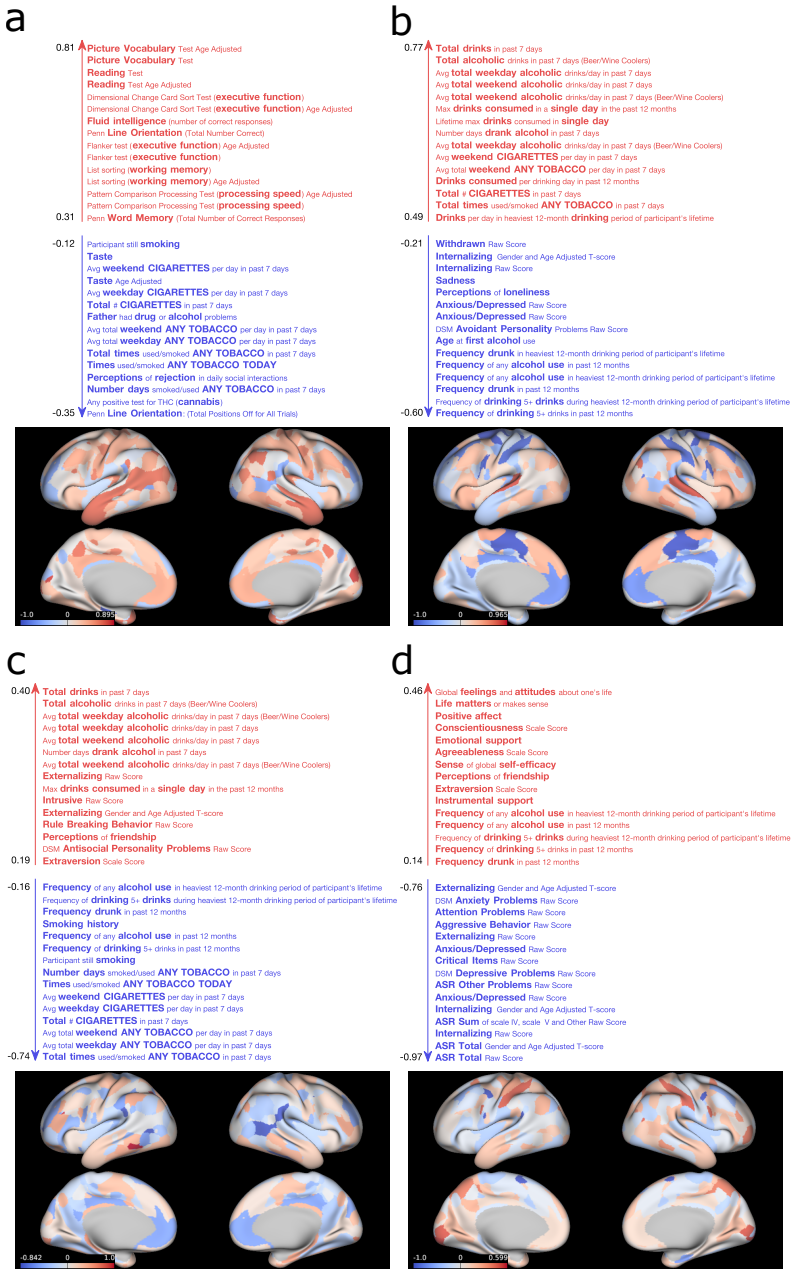

Figure B.5: Non-imaging measures and brain networks described by the first **(a)**, second **(b)**, third **(c)** and fourth **(d)** shared GFA components obtained in the incomplete data HCP experiment 2b described in Section 5.2.5. For illustrative purposes, the top and bottom 15 non-imaging measures for each component are shown. The brain surface plots represent maps of brain connection strength increases/decreases, which were obtained as described in Section B.2.3.

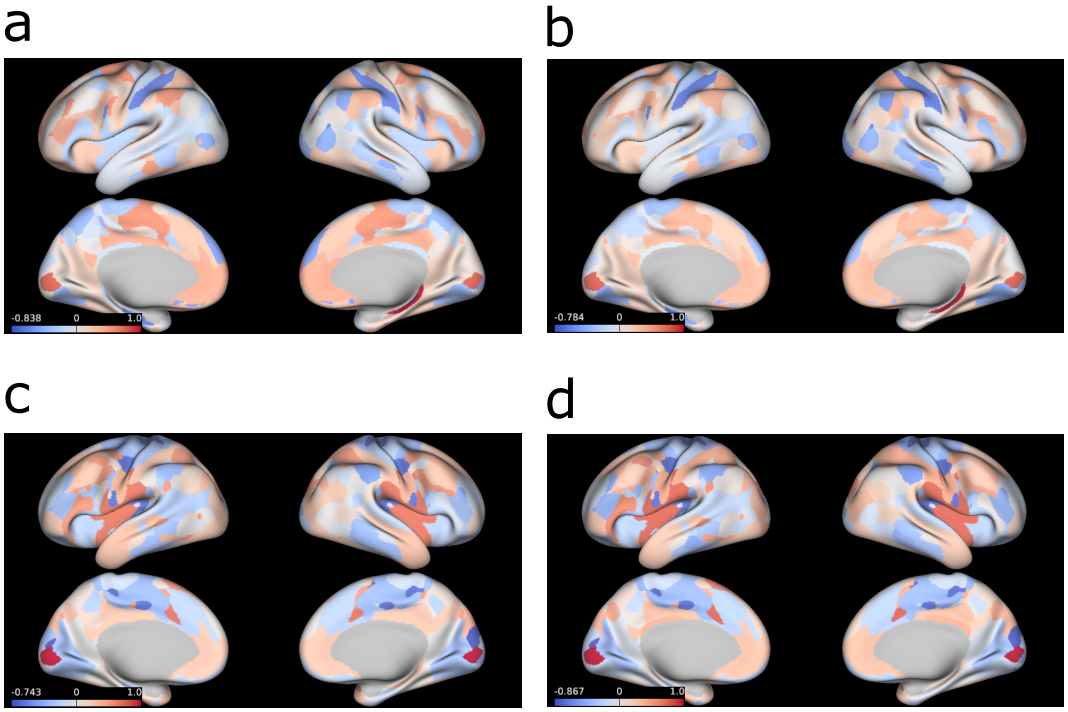

Figure B.6: Brain networks associated with the brain-specific GFA components obtained in the incomplete data HCP experiments 2a **(a,c)** and 2b **(b,d)** described in Section 5.2.5. The brain surface plots represent maps of brain connection strength increases/decreases, which were obtained as described in Section B.2.3.

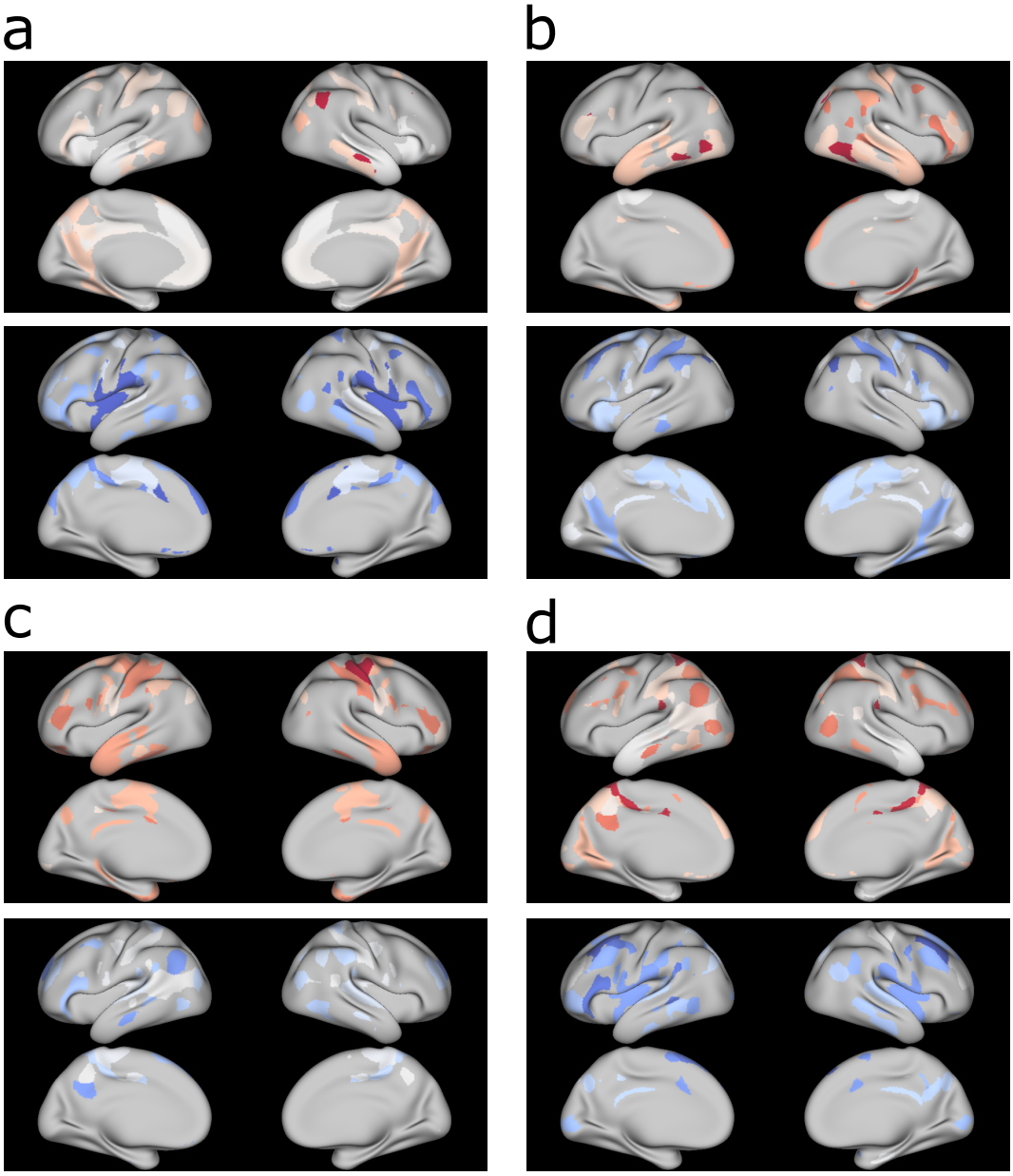

Figure B.7: Brain surface maps of the brain connection strength increases (red) and decreases (blue) of the first **(a)**, second **(b)**, third **(c)** and fourth **(d)** shared GFA components obtained in the HCP experiment with complete data (Figure 5.3). The distribution of the brain connection strengths was thresholded at the 80th (red) and 20th percentile (blue).

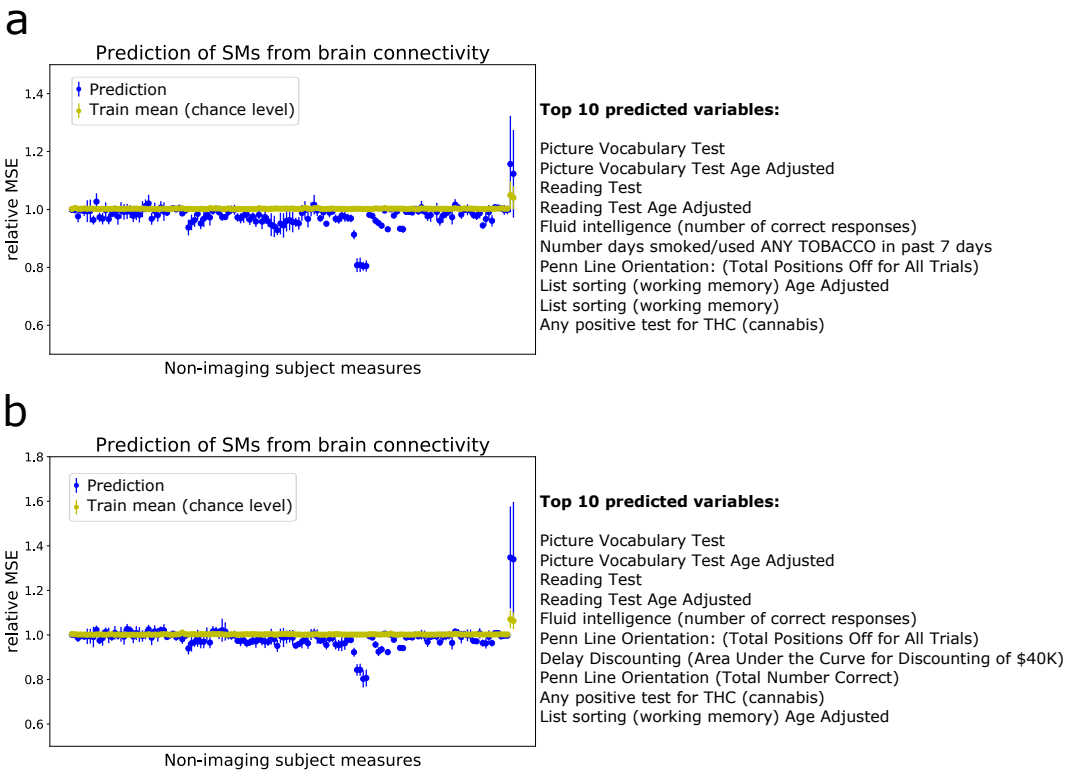

Figure B.8: Multi-output predictions of the non-imaging measures obtained in the incomplete data HCP experiments 2a **(a)** and 2b **(b)** described in Section 5.2.5. The top 10 predicted measures are displayed on the right. For each non-imaging measure, the mean and standard deviation of the relative MSE between the true and predicted values on the test set was calculated across different random initialisations of the experiments.

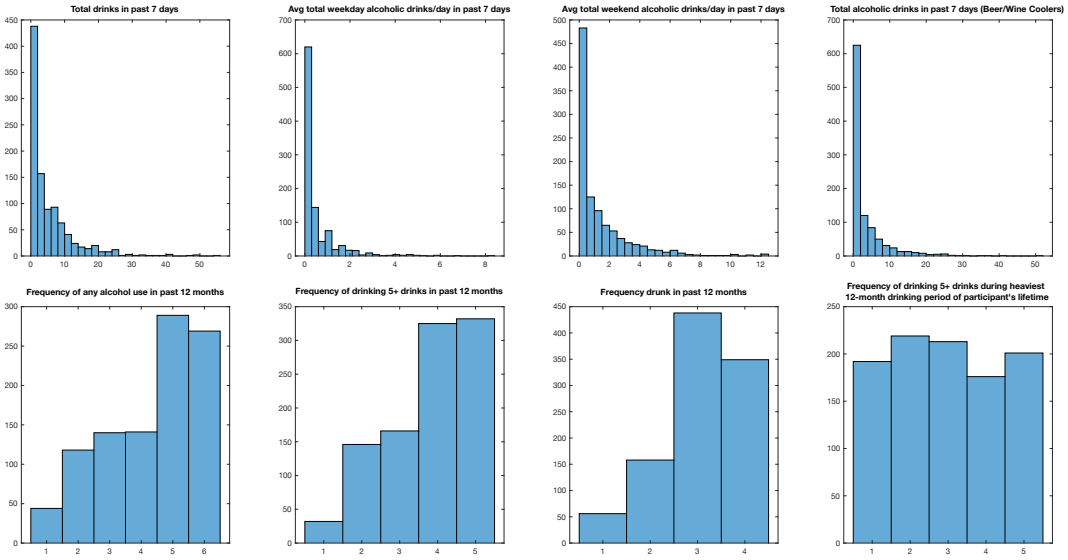

Figure B.9: Histograms of the top 4 variables **(top)** and bottom 4 variables **(bottom)** of the second shared GFA component, displayed in Figure 5.3.

B.3.4 Non-imaging measures from HCP

Table B.5: Description of the non-imaging measures from the HCP dataset used in Chapter 5

Item	Variable	Label
1	FamHist_Moth_Dep	Mother had depression
2	FamHist_Fath_Dep	Father had depression
3	FamHist_Fath_DrgAlc	Father had drug or alcohol problems
4	FamHist_Moth_None	Mother had none of the previously listed disorders
5	FamHist_Fath_None	Father had none of the previously listed disorders
6	ASR_Anxd_Raw	Anxious/Depressed Raw Score
7	ASR_Anxd_Pct	Anxious/Depressed Raw Score
8	ASR_Witd_Raw	Withdrawn Raw Score
9	ASR_Soma_Raw	Somatic Complaints Raw Score
10	ASR_Thot_Raw	Thought Problems Raw Score
11	ASR_Attn_Raw	Attention Problems Raw Score
12	ASR_Aggr_Raw	Aggressive Behaviour Raw Score
13	ASR_Rule_Raw	Rule Breaking Behaviour Raw Score
14	ASR_Intr_Raw	Intrusive Raw Score
15	ASR_Oth_Raw	ASR Other Problems Raw Score
16	ASR_Crit_Raw	Critical Items Raw Score
17	ASR_Intn_Raw	Internalizing Raw Score
18	ASR_Intn_T	Internalizing Gender and Age Adjusted T-score
19	ASR_Extn_Raw	Externalizing Raw Score
20	ASR_Extn_T	Externalizing Gender and Age Adjusted T-score
21	ASR_TAO_Sum	ASR Sum of scale IV, scale V and Other Raw Score
22	ASR_Totp_Raw	ASR Total Raw Score
23	ASR_Totp_T	ASR Total Gender and Age Adjusted T-score
24	DSM_Depr_Raw	DSM Depressive Problems Raw Score
25	DSM_Anxi_Raw	DSM Anxiety Problems Raw Score
26	DSM_Somp_Raw	DSM Somatic Problems Raw Score
27	DSM_Avoid_Raw	DSM Avoidant Personality Problems Raw Score

28	DSM_Adh_Raw	DSM AD/H Problems Raw Score
29	DSM_Inat_Raw	DSM Inattention Problems Raw Score
30	DSM_Hype_Raw	DSM Hyperactivity Problems Raw Score
31	DSM_Antis_Raw	DSM Antisocial Personality Problems Raw Score
32	SSAGA_Childhood Conduct	Number of of Childhood Conduct problems
33	SSAGA_PanicDisorder	Non-diagnostic screen of panic disorder
34	SSAGA_Agoraphobia	Non-diagnostic screen of agoraphobia
35	SSAGA_Depressive_Ep	DSMIV Major Depressive Episode over his/her lifetime
36	SSAGA_Depressive_Sx	Number of depressive symptoms for major depression over his/her lifetime
37	EVA_Denom	Electronic Visual Acuity Denominator
38	Correction	Eyeglass lens correction
39	THC	Any positive test for THC (cannabis)
40	Total_Drinks_7days	Total drinks in past 7 days
41	Num_Days_Drank_7days	Number days drank alcohol in past 7 days
42	Avg_Weekday_Drinks_ 7days	Avg total weekday alcoholic drinks/day in past 7 days
43	Avg_Weekend_Drinks_ 7days	Avg total weekend alcoholic drinks/day in past 7 days
44	Total_Beer_Wine_ Cooler_7days	Total alcoholic drinks in past 7 days (Beer/Wine Coolers)
45	Avg_Weekday_Beer_ Wine_Cooler_7days	Avg total weekday alcoholic drinks/day in past 7 days (Beer/Wine Coolers)
46	Avg_Weekend_Beer_ Wine_Cooler_7days	Avg total weekend alcoholic drinks/day in past 7 days (Beer/Wine Coolers)
47	Total_Wine_ 7days	Total alcoholic drinks in past 7 days (Wine)
48	Avg_Weekday_Wine_ 7days	Avg total weekday alcoholic drinks/day in past 7 days (Wine)
49	Avg_Weekend_Wine_ 7days	Avg total weekend alcoholic drinks/day in past 7 days (Wine)
50	Total_Hard_Liquor_ 7days	Total alcoholic drinks in past 7 days (Hard Liquor)
51	Avg_Weekend_Hard_ Liquor_7days	Avg total weekend alcoholic drinks/day in past 7 days (Hard Liquor)

52	SSAGA_Alc_D4_Dp_Sx	Number of DSM4 ALC Dependence Criteria Met
53	SSAGA_Alc_D4_Ab_Dx	DSM4 criteria for Alcohol Abuse sometime over his/her lifetime
54	SSAGA_Alc_D4_Ab_Sx	Number of symptoms of DSM4 Alcohol Abuse sometime over lifetime
55	SSAGA_Alc_12_Drinks_Per_Day	Drinks consumed per drinking day in past 12 months
56	SSAGA_Alc_12_Frq	Frequency of any alcohol use in past 12 months
57	SSAGA_Alc_12_Frq_5plus	Frequency of drinking 5+ drinks in past 12 months
58	SSAGA_Alc_12_Frq_Drk	Frequency drunk in past 12 months
59	SSAGA_Alc_12_Max_Drinks	Max drinks consumed in a single day in the past 12 months
60	SSAGA_Alc_Age_1st_Use	Age at first alcohol use
61	SSAGA_Alc_Hvy_Drinks_Per_Day	Drinks per day in the heaviest 12-month drinking period of participant's lifetime
62	SSAGA_Alc_Hvy_Frq	Frequency of any alcohol use in the heaviest 12-month drinking period of participant's lifetime
63	SSAGA_Alc_Hvy_Frq_5plus	Frequency of drinking 5+ drinks during the heaviest 12-month drinking period of participant's lifetime
64	SSAGA_Alc_Hvy_Frq_Drk	Frequency drunk in the heaviest 12-month drinking period of participant's lifetime
65	SSAGA_Alc_Hvy_Max_Drinks	Lifetime max drinks consumed in single day
66	Total_Any_Tobacco_7days	Total times used/smoked ANY TOBACCO in past 7 days
67	Times_Used_Any_Tobacco_Today	Times used/smoked ANY TOBACCO TODAY
68	Num_Days_Used_Any_Tobacco_7days	Number days smoked/used ANY TOBACCO in past 7 days
69	Avg_Weekday_Any_Tobacco_7days	Avg total weekday ANY TOBACCO per day in past 7 days
70	Avg_Weekend_Any_Tobacco_7days	Avg total weekend ANY TOBACCO per day in past 7 days

71	Total_Cigarettes_7days	Total number of CIGARETTES in past 7 days
72	Avg_Weekday_Cigarettes_7days	Avg weekday CIGARETTES per day in past 7 days
73	Avg_Weekend_Cigarettes_7days	Avg weekend CIGARETTES per day in past 7 days
74	SSAGA_TB_Smoking_History	Smoking history
75	SSAGA_TB_Still_Smoking	Participant still smoking
76	SSAGA_Times_Used_Illicits	Times used all classes of non-marijuana illicit drugs, including cocaine, hallucinogens, opiates, sedatives, or other
77	SSAGA_Times_Used_Cocaine	Times used cocaine
78	SSAGA_Times_Used_Hallucinogens	Times used hallucinogens
79	SSAGA_Times_Used_Opiates	Times used opiates
80	SSAGA_Times_Used_Sedatives	Times used sedatives
81	SSAGA_Times_Used_Stimulants	Times used stimulants
82	SSAGA_Mj_Use	Ever used marijuana
83	SSAGA_Mj_Ab_Dep	DSM criteria for Marijuana Dependence at some time over his/her lifetime
84	SSAGA_Mj_Times_Used	Times used marijuana
85	MMSE_Score	Mini Mental Status Exam Total Score
86	PSQI_Score	Total score of Pittsburgh Sleep Quality Index
87	PicSeq_Unadj	Picture Sequence Memory Test (fluid ability)
88	PicSeq_AgeAdj	Picture Sequence Memory Test (fluid ability) Age Adjusted
89	CardSort_Unadj	Dimensional Change Card Sort Test (executive function)
90	CardSort_AgeAdj	Dimensional Change Card Sort Test (executive function) Age Adjusted
91	Flanker_Unadj	Flanker test (executive function)

92	Flanker_AgeAdj	Flanker test (executive function) Age Adjusted
93	PMAT24_A_CR	Fluid intelligence (number of correct responses)
94	ReadEng_Unadj	Reading Test
95	ReadEng_AgeAdj	Reading Test Age Adjusted
96	PicVocab_Unadj	Picture Vocabulary Test
97	PicVocab_AgeAdj	Picture Vocabulary Test Age Adjusted
98	ProcSpeed_Unadj	Pattern Comparison Processing Test (processing speed)
99	ProcSpeed_AgeAdj	Pattern Comparison Processing Test (processing speed) Age Adjusted
100	DDisc_AUC_200	Delay Discounting (Area Under the Curve for Discounting of $200)
101	DDisc_AUC_40K	Delay Discounting (Area Under the Curve for Discounting of $40K)
102	VSPLOT_TC	Penn Line Orientation (Total Number Correct)
103	VSPLOT_CRTE	Penn Line Orientation (Median Reaction Time)
104	VSPLOT_OFF	Penn Line Orientation: (Total Positions Off for All Trials)
105	SCPT_SEN	Short Penn CPT Sensitivity
106	SCPT_SPEC	Short Penn CPT Specificity
107	IWRD_TOT	Penn Word Memory (Total Number of Correct Responses)
108	ListSort_Unadj	List sorting (working memory)
109	ListSort_AgeAdj	List sorting (working memory) Age Adjusted
110	ER40_CR	Penn Emotion Recognition (Number of Correct Responses)
111	ER40ANG	Penn Emotion Recognition (Number of Correct Anger Identifications)
112	ER40FEAR	Penn Emotion Recognition (Number of Correct Fear Identifications)
113	ER40NOE	Penn Emotion Recognition (Number of Correct Neutral Identifications)
114	ER40SAD	Penn Emotion Recognition (Number of Correct Sad Identifications)

115	AngAffect_Unadj	Anger
116	AngHostil_Unadj	Hostility and cynicism
117	AngAggr_Unadj	Aggression
118	FearAffect_Unadj	Fear and anxious misery
119	FearSomat_Unadj	Somatic symptoms of anxiety
120	Sadness_Unadj	Sadness
121	LifeSatisf_Unadj	Global feelings and attitudes about one's life
122	MeanPurp_Unadj	Life matters or makes sense
123	PosAffect_Unadj	Positive affect
124	Friendship_Unadj	Perceptions of friendship
125	Loneliness_Unadj	Perceptions of loneliness
126	PercHostil_Unadj	Perceptions of hostility in daily social interactions
127	PercReject_Unadj	Perceptions of rejection in daily social interactions
128	EmotSupp_Unadj	Emotional support
129	InstruSupp_Unadj	Instrumental support
130	PercStress_Unadj	Perception of stress
131	SelfEff_Unadj	Sense of global self-efficacy
132	Dexterity_Unadj	Manual dexterity
133	Dexterity_AgeAdj	Manual dexterity Age Adjusted
134	NEOFAC_A	Agreeableness Scale Score
135	NEOFAC_O	Openness Scale Score
136	NEOFAC_C	Conscientiousness Scale Score
137	NEOFAC_N	Neuroticism Scale Score
138	NEOFAC_E	Extraversion Scale Score
139	Odor_Unadj	Ability to identify various odours
140	Odor_AgeAdj	Ability to identify various odours Age Adjusted
141	PainInterf_Tscore	Consequences of pain on relevant aspects of one's life
142	Taste_Unadj	Taste
143	Taste_AgeAdj	Taste Age Adjusted
144	Mars_Log_Score	Contrast Sensitivity Score
145	Mars_Final	Final Contrast Sensitivity (CS) Score

Appendix C

Complements to Chapter 6

C.1 Results on synthetic data

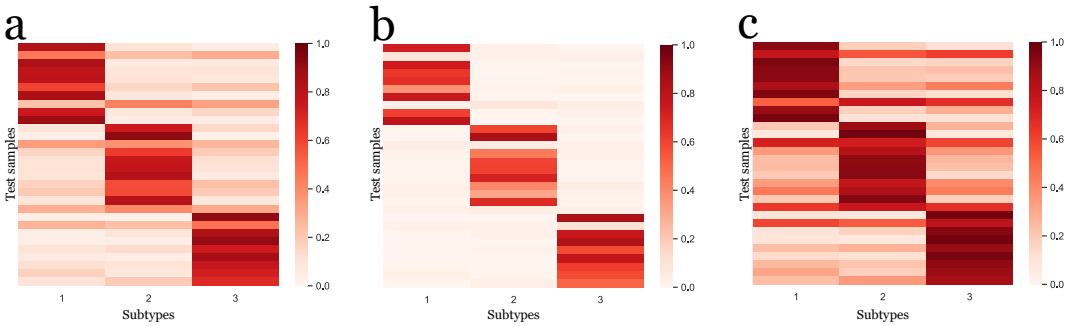

Figure C.1: Probabilities of the test samples to belong to the subtypes. **(a)** Posterior mean and quantiles at **(b)** 0.05 and **(c)** 0.95 of the posterior distribution of $\mathbf{\Psi}$ on the test set.

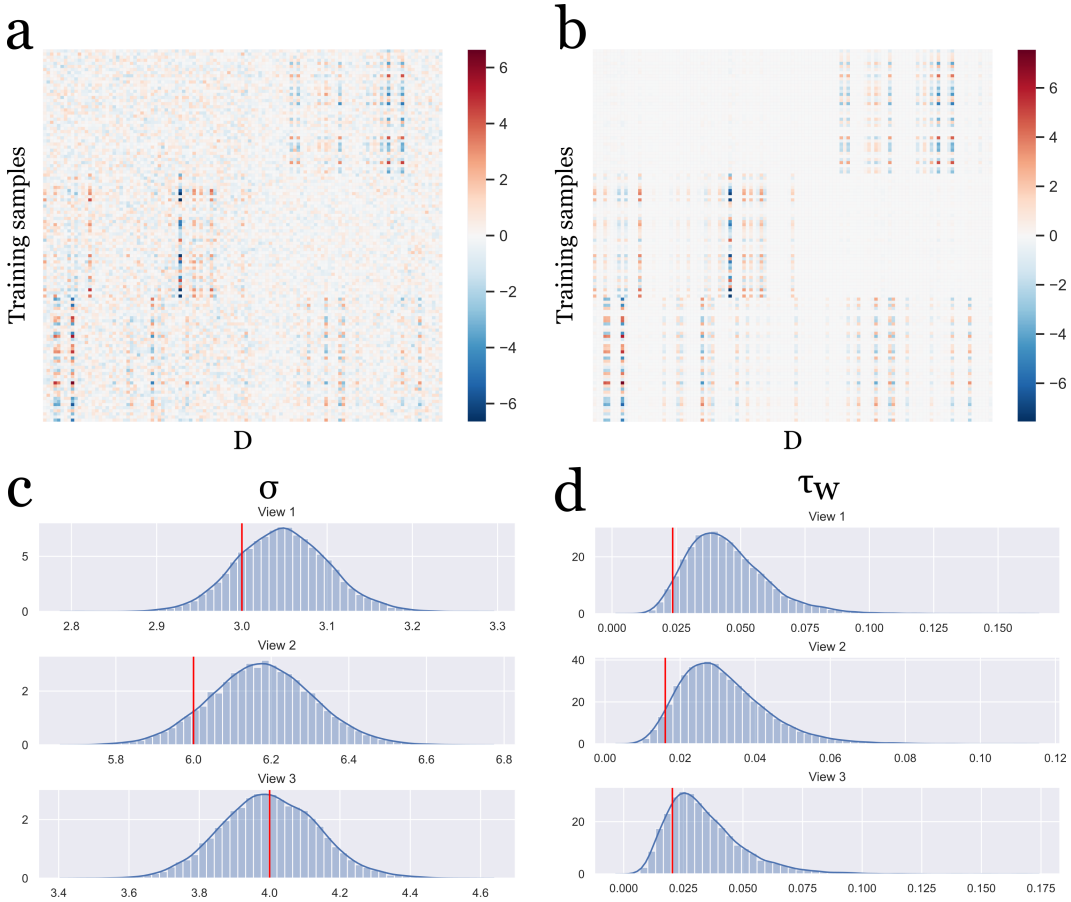

Figure C.2: Generated **(a)** and inferred **(b)** input data, and model's parameters when applying sparse GFA to the synthetic data described in Section 6.2.5. Histogram of the posterior samples of **(c)** σ and **(d)** τ_w. The vertical red line indicates the true value of each parameter.

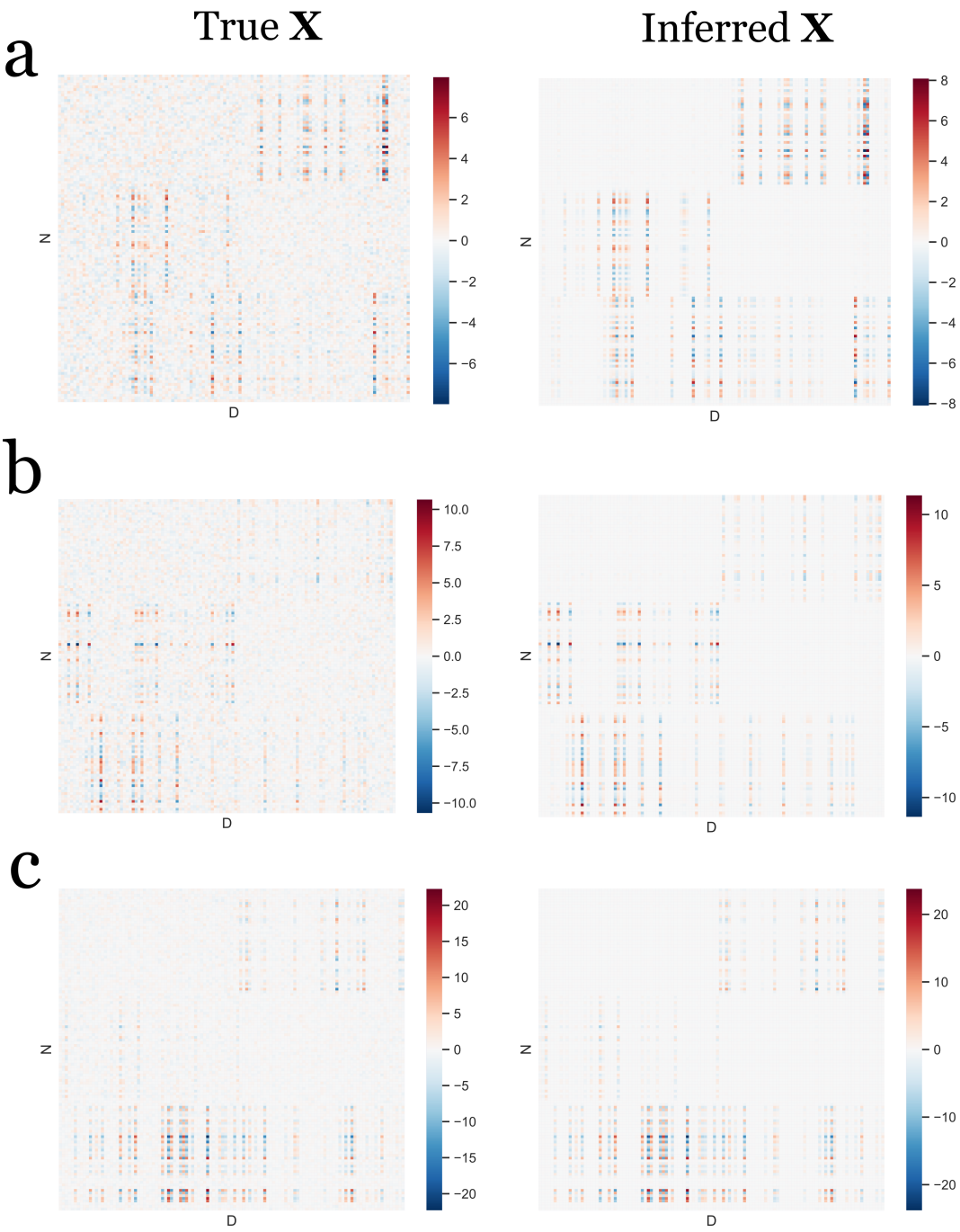

Figure C.3: Generated and inferred input data $\mathbf{X}$ with different values of $p_0^{(m)}$ using supervised GFA. **(a)** $p_0^{(m)} = D_m/5$; **(b)** $p_0^{(m)} = D_m/3$; **(c)** $p_0^{(m)} = D_m/2$. The samples are represented on the rows and the features on the columns.

C.2 GENFI data

Table C.1: Description of the non-imaging features from the GENFI dataset used in Chapter 6. Orange corresponds to medical assessments of the patients; blue corresponds to informant questionnaires completed by primary caregiver; red corresponds to neuropsychological tasks completed by patients. *a high score in the test means that the individual is more affected; **a high score in the test means that the individual is less affected.

Label	Task Name	Cognitive function
FTDL	Clinical Dementia Rating*	disease severity (FTD appropriate)
Lang	Progressive Aphasia Severity Scale*	speech and language
mmse	Mini Mental State Examination**	disease severity (general dementia)
memory	Cambridge Behavioural Index*	memory
ev_skills	Cambridge Behavioural Index*	everyday skills
self_care	Cambridge Behavioural Index*	self care
abn_beh	Cambridge Behavioural Index*	abnormal behaviour
mood	Cambridge Behavioural Index*	mood
beliefs	Cambridge Behavioural Index*	beliefs
eating	Cambridge Behavioural Index*	eating
sleep	Cambridge Behavioural Index*	sleep
st_beh	Cambridge Behavioural Index*	stereotypic behaviour
Motiv	Cambridge Behavioural Index*	motivation
CBI	Cambridge Behavioural Index (overall score)*	general behaviour
FRS	FTD Rating Scale**	disease severity
mIRI	Modified Interpersonal Reactivity Index**	cognitive and emotional empathy
RSMS	Revised Self Monitoring Scale**	social behaviour
Ben_figcopy	Benson figure copy**	visuospatial skills
Ben_figrec	Benson figure recall**	episodic memory
Ben_figrecog	Benson figure recognition**	visual memory
DS_F	Digit span forewards**	attention processing
DS_B	Digit span backwards**	working memory
CC	Camel and Cactus Test**	semantic memory
TMTA	Trail Making Task A*	attention processing
TMTB	Trail Making Task B*	executive function
Digit	Digit symbol**	processing speed
Boston_nam	Boston Naming Task**	object naming - word retrieval
Stroop_color	Stroop Task*	executive funtion
Stroop_word	Stroop Task*	executive function
Stroop_ink	Stroop Task*	executive function
VF_animals	Verbal fluency**	language fluency
VF_comb	Verbal fluency**	language fluency
Block_design	Block design**	visuospatial skills
Ekman	Facial emotion recognition**	emotion recognition

Table C.2: Description of the brain imaging features from the GENFI dataset.

Abbreviated Label	Variable Label
Racc	Right Accumbens Area
Lacc	Left Accumbens Area
Ram	Right Amygdala
Lam	Left Amygdala
RC	Right Caudate
LC	Left Caudate
RH	Right Hippocampus
LH	Left Hippocampus
Rpal	Right Pallidum
Lpal	Left Pallidum
Rpu	Right Putamen
Lpu	Left Putamen
Rth	Right Thalamus
Lth	Left Thalamus
LFL	Left Frontal lobe
RFL	Right Frontal lobe
LTL	Left Temporal lobe
RTL	Right Temporal lobe
LPL	Left Parietal lobe
RPL	Right Parietal lobe
LOL	Left Occipital lobe
ROL	Right Occipital lobe
Lcing	Left Cingulate
Rcing	Right Cingulate
Lins	Left Insula
Rins	Right Insula
Cer	Cerebellum
Asy	Asymmetry

Appendix D

Distributions

In this section, I present the probability density/mass function of the distributions used in this thesis.

D.1 Multivariate normal distribution

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \tag{D.1}$$

where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Sigma}$ is the covariance matrix, $|\cdot|$ is the determinant and D is the number of dimensions.

D.2 Gamma distribution

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^\alpha x^{\alpha-1}e^{-\beta x}, \tag{D.2}$$

where α is the shape parameter, $\Gamma(\cdot)$ is the gamma function and β is the rate parameter.

D.3 Inverse-Gamma distribution

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\frac{\beta^\alpha}{x^{\alpha+1}}e^{-\frac{\beta}{x}}, \tag{D.3}$$

where α is the shape parameter, $\Gamma(\cdot)$ is the gamma function and β is the rate parameter.

D.4 Half-Cauchy distribution

$$f(x;\mu,\sigma) = \begin{cases} \frac{2}{\pi\sigma}\frac{1}{1+(x-\mu)^2/\sigma}, & x \geq \mu \\ 0, & \text{otherwise,} \end{cases} \tag{D.4}$$

where μ is the location parameter and σ is the scale parameter.

D.5 Bernoulli distribution

$$f(x;\theta) = \begin{cases} 1-\theta, & x = 0 \\ \theta, & x = 1, \end{cases} \tag{D.5}$$

where θ is the probability that a trial is successful.

D.6 Beta distribution

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)x^{\alpha-1}(1-x)^{(\beta-1)}}{\Gamma(\alpha)\Gamma(\beta)}, \tag{D.6}$$

where α and β are shape parameters, and $\Gamma(\cdot)$ is the gamma function.

D.7 Inverse-Wishart distribution

$$f(\mathbf{x};\nu,\mathbf{\Psi}) = \frac{|\mathbf{\Psi}|^{\nu/2}}{2^{\nu D/2}\Gamma_D(\frac{\nu}{2})}|\mathbf{x}|^{-(\nu+D+1)/2}e^{-\frac{1}{2}\mathrm{Tr}(\mathbf{\Psi}\mathbf{x}^{-1})}, \tag{D.7}$$

where $\mathbf{\Psi}$ is the $D \times D$ scale matrix, ν is the degrees of freedom, $|\cdot|$ is the determinant and $\Gamma_D(\cdot)$ is the multivariate gamma function.

Bibliography

R. Admon and D. A. Pizzagalli. Corticostriatal pathways contribute to the natural time course of positive mood. *Nature Communications*, 6(1):10065, 2015. ISSN 2041-1723. doi: 10.1038/ncomms10065.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. ISSN 1558-2523. doi: 10.1109/TAC.1974.1100705.

D. Alnæs, T. Kaufmann, A. F. Marquand, S. M. Smith, and L. T. Westlye. Patterns of sociocognitive stratification and perinatal risk in the child brain. *Proceedings of the National Academy of Sciences of the United States of America*, 117(22):12419–12427, 2020. ISSN 10916490. doi: 10.1073/pnas.2001517117.

A. I. Alterman. *Substance abuse and psychopathology*. Applied Clinical Psychology. Springer US, Boston, MA, 1985. ISBN 978-1-4899-3643-1.

B. B. Avants, P. A. Cook, L. Ungar, J. C. Gee, and M. Grossman. Dementia induces correlated reductions in white matter integrity and cortical thickness: A multivariate neuroimaging study with sparse canonical correlation analysis. *NeuroImage*, 50(3):1004–1016, 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.01.041.

F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, University of California, Berkeley, 2006. URL https://statistics.berkeley.edu/tech-reports/688.

D. S. Bassett and E. T. Bullmore. Human brain networks in health and disease. *Current Opinion in Neurology*, 22(4):340–347, 2009. ISSN 1350-7540. doi: 10.1097/WCO.0b013e32832d93dd.

M. G. Berman, D. E. Nee, M. Casement, H. S. Kim, P. Deldin, E. Kross, R. Gonzalez, E. Demiralp, I. H. Gotlib, P. Hamilton, J. Joormann, C. Waugh, and J. Jonides. Neural and behavioral effects of interference resolution in depression and rumination. *Cognitive, Affective, & Behavioral Neuroscience*, 11(1):85–96, 2011. ISSN 1530-7026. doi: 10.3758/s13415-010-0014-x.

M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv*, 2018. URL https://arxiv.org/abs/1701.02434.

M. J. Betancourt and M. Girolami. Hamiltonian Monte Carlo for hierarchical models. 2013.

J. D. Bijsterbosch, M. W. Woolrich, M. F. Glasser, E. C. Robinson, C. F. Beckmann, D. C. Van Essen, S. J. Harrison, and S. M. Smith. The relationship between spatial configuration and functional connectivity of brain regions. *eLife*, 7, feb 2018. ISSN 2050084X. doi: 10.7554/eLife.32992.

C. M. Bishop. Variational principal components. In *IEE Conference Publication*, volume 1, pages 509–514. IEE, 1999. ISBN 0852967217. doi: 10.1049/cp:19991160.

C. M. Bishop. *Pattern recognition and machine learning*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

M. Bocchetta, E. G. Todd, G. Peakman, D. M. Cash, R. S. Convery, L. L. Russell, D. L. Thomas, J. Eugenio Iglesias, J. C. van Swieten, L. C. Jiskoot, H. Seelaar, B. Borroni, D. Galimberti, R. Sanchez-Valle, R. Laforce, F. Moreno, M. Synofzik, C. Graff, M. Masellis, M. Carmela Tartaglia, J. B. Rowe, R. Vandenberghe, E. Finger, F. Tagliavini, A. de Mendonça, I. Santana, C. R. Butler, S. Ducharme, A. Gerhard, A. Danek, J. Levin, M. Otto, S. Sorbi, I. Le Ber, F. Pasquier, J. D. Rohrer, S. Afonso, M. Rosario Almeida, S. Anderl-Straub, C. Andersson, A. Antonell, S. Archetti, A. Arighi, M. Balasa, M. Barandiaran, N. Bargalló, R. Bartha, B. Bender, A. Benussi, M. Bertoux, A. Bertrand, V. Bessi, S. Black, S. Borrego-Ecija, J. Bras, A. Brice, R. Bruffaerts, A. Camuzat, M. Cañada, V. Cantoni, P. Caroppo, M. Castelo-Branco, O. Colliot, T. Cope, V. Deramecourt, M. de Arriba, G. Di Fede, A. Díez, D. Duro, C. Fenoglio, C. Ferrari, C. B. Ferreira, N. Fox, M. Freedman, G. Fumagalli, A. Funkiewiez, A. Gabilondo, R. Gasparotti, S. Gauthier, S. Gazzina, G. Giaccone, A. Gorostidi,

C. Greaves, R. Guerreiro, C. Heller, T. Hoegen, B. Indakoetxea, V. Jelic, H.-O. Karnath, R. Keren, G. Kuchcinski, T. Langheinrich, T. Lebouvier, M. João Leitão, A. Lladó, G. Lombardi, S. Loosli, C. Maruta, S. Mead, L. Meeter, G. Miltenberger, R. van Minkelen, S. Mitchell, K. Moore, B. Nacmias, A. Nelson, J. Nicholas, L. Öijerstedt, J. Olives, S. Ourselin, A. Padovani, J. Panman, J. M. Papma, Y. Pijnenburg, C. Polito, E. Premi, S. Prioni, C. Prix, R. Rademakers, V. Redaelli, D. Rinaldi, T. Rittman, E. Rogaeva, A. Rollin, P. Rosa-Neto, G. Rossi, M. Rossor, B. Santiago, D. Saracino, S. Sayah, E. Scarpini, S. Schönecker, E. Semler, R. Shafei, C. Shoesmith, I. Swift, M. Tábuas-Pereira, M. Tainta, R. Taipa, D. Tang-Wai, P. Thompson, H. Thonberg, C. Timberlake, P. Tiraboschi, P. Van Damme, M. Vandenbulcke, M. Veldsman, A. Verdelho, J. Villanua, J. Warren, C. Wilke, I. Woollacott, E. Wlasich, H. Zetterberg, and M. Zulaica. Differential early subcortical involvement in genetic FTD within the GENFI cohort. *NeuroImage: Clinical*, 30:102646, jan 2021. ISSN 22131582. doi: 10.1016/j.nicl.2021.102646.

D. A. Brent. Comorbidity of substance abuse and other psychiatric disorders in adolescents. *American Journal of Psychiatry*, 146(9):1131–1141, 1989. ISSN 0002-953X. doi: 10.1176/ajp.146.9.1131.

J. W. Buckholtz and A. Meyer-Lindenberg. Psychopathology and the human connectome: toward a transdiagnostic model of risk for mental illness. *Neuron*, 74(6):990–1004, 2012. ISSN 08966273. doi: 10.1016/j.neuron.2012.06.002.

K. Bunte, E. Leppäaho, I. Saarinen, and S. Kaski. Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics*, 32(16):2457–2463, 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btw207.

D. Bzdok and A. Meyer-Lindenberg. Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3):223–230, 2018. ISSN 24519022. doi: 10.1016/j.bpsc.2017.11.007.

J. F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, mar 2010. ISSN 10526234. doi: 10.1137/080738970.

M. J. Cardoso, M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, and S. Ourselin. Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion. *IEEE Transactions on Medical Imaging*, 34(9):1976–1988, 2015. ISSN 1558254X. doi: 10.1109/TMI.2015.2418298.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In *Journal of Machine Learning Research*, volume 5, pages 73–80, 2009.

C. S. Carver, S. K. Sutton, and M. F. Scheier. Action, emotion, and personality: emerging conceptual integration. *Personality and Social Psychology Bulletin*, 26(6):741–751, 2000. ISSN 0146-1672. doi: 10.1177/0146167200268008.

B. Casey, R. M. Jones, and T. A. Hare. The adolescent brain. *Annals of the New York Academy of Sciences*, 1124(1):111–126, 2008. ISSN 00778923. doi: 10.1196/annals.1440.010.

B. Casey, A. Galván, and L. H. Somerville. Beyond simple models of adolescence to an integrated circuit-based account: A commentary. *Developmental Cognitive Neuroscience*, 17:128–130, 2016. ISSN 18789293. doi: 10.1016/j.dcn.2015.12.006.

B. J. Casey. Beyond simple models of self-control to circuit-based accounts of adolescent behavior. *Annual Review of Psychology*, 66(1):295–319, 2015. ISSN 0066-4308. doi: 10.1146/annurev-psych-010814-015156.

N. Castellanos-Ryan and P. Conrod. Personality and substance misuse: evidence for a four-factor model of vulnerability. In J. C. Verster, K. Brady, M. Galanter, and P. Conrod, editors, *Drug Abuse and Addiction in Medical Illness: Causes, Consequences and Treatment*, pages 1–573. Springer New York, 2012. ISBN 9781461433750.

A. M. Chekroud, R. Gueorguieva, H. M. Krumholz, M. H. Trivedi, J. H. Krystal, and G. McCarthy. Reevaluating the efficacy and predictability of antidepressant treatments. *JAMA Psychiatry*, 74(4):370, 2017. ISSN 2168-622X. doi: 10.1001/jamapsychiatry.2017.0025.

I. Chien, A. Enrique, J. Palacios, T. Regan, D. Keegan, D. Carter, S. Tschiatschek, A. Nori, A. Thieme, D. Richards, G. Doherty, and D. Belgrave. A

machine learning approach to understanding patterns of engagement with internet-delivered mental health Interventions. *JAMA Network Open*, 3(7): 2010791, 2020. ISSN 25743805. doi: 10.1001/jamanetworkopen.2020.10791.

Chong Wang. Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 18(3):905–910, 2007. ISSN 1045-9227. doi: 10.1109/TNN.2007.891186.

R. E. Cooney, J. Joormann, F. Eugène, E. L. Dennis, and I. H. Gotlib. Neural correlates of rumination in depression. *Cognitive, Affective, & Behavioral Neuroscience*, 10(4):470–478, 2010. ISSN 1530-7026. doi: 10.3758/CABN.10.4.470.

A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964, 2020. ISSN 18732860. doi: 10.1016/j.artmed.2020.101964.

A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis. *NeuroImage*, 9(2):179–194, 1999. ISSN 10538119. doi: 10.1006/nimg.1998.0395.

V. Della-Maggiore, A. B. Sekuler, C. L. Grady, P. J. Bennett, R. Sekuler, and A. R. McIntosh. Corticolimbic interactions associated with performance on a short-term memory task are modified by age. *J. Neurosci.*, 20(22):8410–8416, 2000. ISSN 1529-2401. doi: 20/22/8410[pii].

A. T. Drysdale, L. Grosenick, J. Downar, K. Dunlop, F. Mansouri, Y. Meng, R. N. Fetcho, B. Zebley, D. J. Oathes, A. Etkin, A. F. Schatzberg, K. Sudheimer, J. Keller, H. S. Mayberg, F. M. Gunning, G. S. Alexopoulos, M. D. Fox, A. Pascual-Leone, H. U. Voss, B. Casey, M. J. Dubin, and C. Liston. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, 23(1):28–38, 2017. ISSN 1078-8956. doi: 10.1038/nm.4246.

Y. Du, Z. Fu, and V. D. Calhoun. Classification and prediction of brain disorders using functional connectivity: promising but challenging. *Frontiers in Neuroscience*, 12:525, 2018. ISSN 1662-453X. doi: 10.3389/fnins.2018.00525.

C. Ecker, A. Marquand, J. Mourao-Miranda, P. Johnston, E. M. Daly, M. J. Brammer, S. Maltezos, C. M. Murphy, D. Robertson, S. C. Williams, and D. G. M. Murphy. Describing the brain in autism in five dimensions-Magnetic Resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach. *Journal of Neuroscience*, 30(32):10612–10623, 2010. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.5413-09.2010.

F. S. Ferreira, M. J. Rosa, M. Moutoussis, R. Dolan, J. Shawe-Taylor, J. Ashburner, and J. Mourao-Miranda. Sparse PLS hyperparameters optimisation for investigating brain-behaviour relationships. *2018 International Workshop on Pattern Recognition in Neuroimaging, PRNI 2018*, pages 1–4., 2018. ISSN 01413910.

F. S. Ferreira, A. Mihalik, R. A. Adams, J. Ashburner, and J. Mourao-Miranda. A hierarchical Bayesian model to find brain-behaviour associations in incomplete data sets. *arXiv*, mar 2021. URL http://arxiv.org/abs/2103.06845.

R. A. Fisher. *The design of experiments*. Oliver and Boyd, Edinburgh, 1935.

H. D. Franklin, L. Russell, G. Peakman, C. Greaves, M. Bocchetta, J. Nicholas, J. Poos, R. Convery, D. Cash, J. V. Swieten, L. Jiskoot, F. Moreno, R. Sanchez-Valle, B. Borroni, R. Laforce, M. Masellis, M. C. Tartaglia, C. Graff, D. Galimberti, J. Rowe, E. Finger, M. Synofzik, R. Vandenberghe, A. de Mendonça, F. Tagliavini, I. Santana, S. Ducharme, C. Butler, A. Gerhard, J. Levin, A. Danek, M. Otto, S. Sorbi, I. L. Ber, F. Pasquier, and J. Rohrer. The revised self-monitoring scale detects early impairment of social cognition in fenetic frontotemporal dementia within the GENFI cohort, 2021.

Y. Fujiwara, Y. Miyawaki, and Y. Kamitani. Estimating image bases for visual image reconstruction from human brain activity. In *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, pages 576–584, 2009. ISBN 9781615679119.

Y. Fujiwara, Y. Miyawaki, and Y. Kamitani. Modular encoding and decoding models derived from Bayesian canonical correlation analysis. *Neural Computation*, 25(4):979–1005, 2013. ISSN 08997667. doi: 10.1162/NECO_a_00423.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Number February. Chapman and Hall/CRC, 3rd edition, 2013. ISBN 9780429113079. doi: 10.1201/b16018.

M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. ISSN 0028-0836. doi: 10.1038/nature18933.

G. H. Golub and H. Zha. Perturbation analysis of the canonical correlations of matrix pairs. *Linear Algebra and Its Applications*, 210(C):3–28, 1994. ISSN 00243795. doi: 10.1016/0024-3795(94)90463-4.

E. Gordon, J. D. Rohrer, and N. C. Fox. Advances in neuroimaging in frontotemporal dementia, aug 2016. ISSN 14714159.

C. Grellmann, S. Bitzer, J. Neumann, L. T. Westlye, O. A. Andreassen, A. Villringer, and A. Horstmann. Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of MRI and genetic data. *NeuroImage*, 107:289–310, 2015. ISSN 10959572. doi: 10.1016/j.neuroimage.2014.12.025.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, 2009. ISBN 9780387848846.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444. doi: 10.2307/2334940.

G. Helms, H. Dathe, and P. Dechent. Quantitative FLASH MRI at 3T using a rational approximation of the Ernst equation. *Magnetic Resonance in Medicine*, 59(3):667–672, 2008a. ISSN 07403194. doi: 10.1002/mrm.21542.

G. Helms, H. Dathe, K. Kallenberg, and P. Dechent. High-resolution maps of magnetization transfer with inherent correction for RF inhomogeneity and T1 relaxation obtained from 3D FLASH MRI. *Magnetic Resonance in Medicine*, 60(6):1396–1407, 2008b. ISSN 07403194. doi: 10.1002/mrm.21732.

M. D. Hoffman and A. Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014. ISSN 15337928.

N. Honnorat, A. Dong, E. Meisenzahl-Lechner, N. Koutsouleris, and C. Davatzikos. Neuroanatomical heterogeneity of schizophrenia revealed by semi-supervised machine learning methods. *Schizophrenia Research*, 214:43–50, 2019. ISSN 15732509. doi: 10.1016/j.schres.2017.12.008.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4): 321–377, 1936. ISSN 0006-3444.

T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, D. S. Pine, K. Quinn, C. Sanislow, and P. Wang. Research Domain Criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7):748–751, 2010. ISSN 0002-953X. doi: 10.1176/appi.ajp.2010.09091379.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. ISSN 01621459.

M. Keightley, G. Winocur, S. J. Graham, H. S. Mayberg, S. Hevenor, and C. L. Grady. An fMRI study investigating cognitive modulation of brain regions associated with emotional processing of visual stimuli. *Neuropsychologia*, 41:585–596, 2003.

R. C. Kessler, G. P. Amminger, S. Aguilar-Gaxiola, J. Alonso, S. Lee, and T. B. Ustun. Age of onset of mental disorders: a review of recent literature. *Current Opinion in Psychiatry*, 20(4):359–364, 2007. ISSN 0951-7367. doi: 10.1097/YCO.0b013e32816ebc8c.

S. A. Khan, S. Virtanen, O. P. Kallioniemi, K. Wennerberg, A. Poso, and S. Kaski. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. *Bioinformatics*, 30 (17):i497–i504, 2014. ISSN 1460-2059. doi: 10.1093/bioinformatics/btu456.

B. Kiddle, B. Inkster, G. Prabhu, M. Moutoussis, K. J. Whitaker, E. T. Bullmore, R. J. Dolan, P. Fonagy, I. M. Goodyer, and P. B. Jones. Cohort Profile: The NSPN 2400 Cohort: a developmental sample supporting the Wellcome Trust NeuroScience in Psychiatry Network. *International Journal of Epidemiology*, 47(1):18–19g, 2018. ISSN 0300-5771. doi: 10.1093/ije/dyx117.

A. Klami and S. Kaski. Local dependent components. In *Proceedings of the 24th international conference on Machine learning - ICML '07*, pages 425–432, New York, New York, USA, 2007. ACM Press. ISBN 9781595937933. doi: 10.1145/1273496.1273550.

A. Klami, S. Virtanen, and S. Kaski. Bayesian Canonical Correlation Analysis. *Journal of Machine Learning Research*, 14:965–1003, 2013. ISSN 1532-4435.

A. Klami, S. Virtanen, E. Leppäaho, and S. Kaski. Group Factor Analysis. *IEEE transactions on neural networks and learning systems*, 26(9):2136–47, 2015. ISSN 2162-2388. doi: 10.1109/TNNLS.2014.2376974.

M. Koskinen, J. Viinikanoja, M. Kurimo, A. Klami, S. Kaski, and R. Hari. Identifying fragments of natural speech from the listener's MEG signals. *Human Brain Mapping*, 34(6):1477–1489, 2013. ISSN 10659471. doi: 10.1002/hbm.22004.

A. Krishnan, L. J. Williams, A. R. McIntosh, and H. Abdi. Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*, 56(2):455–475, 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.07.034.

P. Kundu, N. D. Brenowitz, V. Voon, Y. Worbe, P. E. Vertes, S. J. Inati, Z. S. Saad, P. A. Bandettini, and E. T. Bullmore. Integrated strategy for improving functional connectivity mapping using multiecho fMRI. *Proceedings of the National Academy of Sciences*, 110(40):16187–16192, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1301725110.

P. Kundu, B. E. Benson, K. L. Baldwin, D. Rosen, W.-M. Luh, P. A. Bandettini, D. S. Pine, and M. Ernst. Robust resting state fMRI processing for studies on typical brain development based on multi-echo EPI acquisition. *Brain Imaging and Behavior*, 9(1):56–73, 2015. ISSN 1931-7557. doi: 10.1007/s11682-014-9346-4.

B. Labonté, O. Engmann, I. Purushothaman, C. Menard, J. Wang, C. Tan, J. R. Scarpa, G. Moy, Y.-H. E. Loh, M. Cahill, Z. S. Lorsch, P. J. Hamilton, E. S. Calipari, G. E. Hodes, O. Issler, H. Kronman, M. Pfau, A. L. J. Obradovic, Y. Dong, R. L. Neve, S. Russo, A. Kasarskis, C. Tamminga, N. Mechawar, G. Turecki, B. Zhang, L. Shen, and E. J. Nestler. Sex-specific

transcriptional signatures in human depression. *Nature Medicine*, 23(9): 1102–1111, 2017. ISSN 1078-8956. doi: 10.1038/nm.4386.

K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse. A Sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008. ISSN 1544-6115. doi: 10.2202/1544-6115.1390.

É. Le Floch, V. Guillemot, V. Frouin, P. Pinel, C. Lalanne, L. Trinchera, A. Tenenhaus, A. Moreno, M. Zilbovicius, T. Bourgeron, S. Dehaene, B. Thirion, J.-B. Poline, and É. Duchesnay. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage*, 63(1):11–24, 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.06.061.

C. Lee and M. van der Schaar. A Variational Information Bottleneck Approach to Multi-Omics Data Integration. 2021.

F. S. Lee, H. Heimer, J. N. Giedd, E. S. Lein, N. Estan, D. R. Weinberger, and B. J. Casey. Adolescent mental health-opportunity and obligation. *Science*, 346(6209):547–549, 2014. ISSN 0036-8075. doi: 10.1126/science.1260497.

J. Li, T. Bolt, D. Bzdok, J. S. Nomi, B. T. Yeo, R. N. Spreng, and L. Q. Uddin. Topography and behavioral relevance of the global signal in the human brain. *Scientific Reports*, 9(1):14286, 2019. ISSN 20452322. doi: 10.1038/s41598-019-50750-8.

D. Lin, V. D. Calhoun, and Y.-P. Wang. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Medical Image Analysis*, 18(6):891–902, 2014. ISSN 13618415. doi: 10.1016/j.media.2013.10.010.

R. N. Lipari, S. L. Hedden, and A. Hughes. Substance use and mental health estimates from the 2013 national survey on drug use and health: overview of findings. In *The CBHSQ Report*. Rockville (MD): Substance Abuse and Mental Health Services Administration (US), 2014. ISBN HHS Publication No. (SMA) 11-4658.

J. Luttinen and A. Ilin. Transformations in variational Bayesian factor analysis to speed up learning. *Neurocomputing*, 73(7-9):1093–1102, 2010. ISSN 09252312. doi: 10.1016/j.neucom.2009.11.018.

H. Lv, Z. Wang, E. Tong, L. M. Williams, G. Zaharchuk, M. Zeineh, A. N. Goldstein-Piekarski, T. M. Ball, C. Liao, and M. Wintermark. Resting-state functional MRI: everything that nonexperts have always wanted to know. *American Journal of Neuroradiology*, 39(8):1390–1399, 2018. ISSN 0195-6108. doi: 10.3174/ajnr.A5527.

D. J. Mackay. Probable networks and plausible predictions - a review of practical bayesian methods for supervised neural networks, 1995. ISSN 0954898X. URL https://www.tandfonline.com/doi/abs/10.1088/0954-898X{_}6{_}3{_}011.

L. Mackey. Deflation methods for sparse PCA. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1017–1024, 2008. ISBN 9781605609492.

C. J. Mahoney, J. Beck, J. D. Rohrer, T. Lashley, K. Mok, T. Shakespeare, T. Yeatman, E. K. Warrington, J. M. Schott, N. C. Fox, M. N. Rossor, J. Hardy, J. Collinge, T. Revesz, S. Mead, and J. D. Warren. Frontotemporal dementia with the C9ORF72 hexanucleotide repeat expansion: Clinical, neuroanatomical and neuropathological features. *Brain*, 135(3): 736–750, 2012. ISSN 14602156. doi: 10.1093/brain/awr361.

J. M. Mateos-Pérez, M. Dadar, M. Lacalle-Aurioles, Y. Iturria-Medina, Y. Zeighami, and A. C. Evans. Structural neuroimaging as clinical predictor: A review of machine learning applications, 2018. ISSN 22131582.

V. Menon. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends in Cognitive Sciences*, 15(10):483–506, 2011. ISSN 13646613. doi: 10.1016/j.tics.2011.08.003.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114.

A. Mihalik, F. S. Ferreira, M. J. Rosa, M. Moutoussis, G. Ziegler, J. M. Monteiro, L. Portugal, R. A. Adams, R. Romero-Garcia, P. E. Vértes, M. G. Kitzbichler, F. Váša, M. M. Vaghi, E. T. Bullmore, P. Fonagy, I. M. Goodyer, P. B. Jones, T. Hauser, S. Neufeld, M. S. Clair, K. Whitaker, B. Inkster, G. Prabhu, C. Ooi, U. Toseeb, B. Widmer, J. Bhatti, L. Villis, A. Alrumaithi, S. Birt, A. Bowler, K. Cleridou, H. Dadabhoy, E. Davies, A. Firkins,

S. Granville, E. Harding, A. Hopkins, D. Isaacs, J. King, D. Kokorikou, C. Maurice, C. McIntosh, J. Memarzia, H. Mills, C. O'Donnell, S. Pantaleone, J. Scott, P. Fearon, J. Suckling, A. L. van Harmelen, R. Kievit, R. Dolan, and J. Mourão-Miranda. Brain-behaviour modes of covariation in healthy and clinically depressed young people. *Scientific Reports*, 9(1), 2019. ISSN 20452322. doi: 10.1038/s41598-019-47277-3.

A. Mihalik, F. S. Ferreira, M. Moutoussis, G. Ziegler, R. A. Adams, M. J. Rosa, G. Prabhu, L. de Oliveira, M. Pereira, E. T. Bullmore, P. Fonagy, I. M. Goodyer, P. B. Jones, T. Hauser, S. Neufeld, R. Romero-Garcia, M. St Clair, P. E. Vértes, K. Whitaker, B. Inkster, C. Ooi, U. Toseeb, B. Widmer, J. Bhatti, L. Villis, A. Alrumaithi, S. Birt, A. Bowler, K. Cleridou, H. Dadabhoy, E. Davies, A. Firkins, S. Granville, E. Harding, A. Hopkins, D. Isaacs, J. King, D. Kokorikou, C. Maurice, C. McIntosh, J. Memarzia, H. Mills, C. O'Donnell, S. Pantaleone, J. Scott, P. Fearon, J. Suckling, A. L. van Harmelen, R. Kievit, J. Shawe-Taylor, R. Dolan, and J. Mourão-Miranda. Multiple holdouts with stability: improving the generalizability of machine learning analyses of brain–behavior relationships. *Biological Psychiatry*, 87 (4):368–376, 2020. ISSN 18732402. doi: 10.1016/j.biopsych.2019.12.001.

K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. R. Andersson, L. Griffanti, G. Douaud, T. W. Okell, P. Weale, I. Dragonu, S. Garratt, S. Hudson, R. Collins, M. Jenkinson, P. M. Matthews, and S. M. Smith. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523–1536, 2016. ISSN 1097-6256. doi: 10.1038/nn.4393.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. ISSN 01621459. doi: 10.2307/2290129.

J. M. Monteiro, A. Rao, J. Shawe-Taylor, and J. Mourão-Miranda. A multiple hold-out framework for Sparse Partial Least Squares. *Journal of Neuroscience Methods*, 271(271):182–194, 2016. ISSN 01650270. doi: 10.1016/j.jneumeth.2016.06.011.

D. A. Moser, G. E. Doucet, W. H. Lee, A. Rasgon, H. Krinsky, E. Leibu, A. Ing, G. Schumann, N. Rasgon, and S. Frangou. Multivariate associations

among behavioral, clinical, and multimodal imaging phenotypes in patients with psychosis. *JAMA Psychiatry*, 75(4):386–395, 2018. ISSN 2168622X. doi: 10.1001/jamapsychiatry.2017.4741.

J. Mourão-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *NeuroImage*, 28(4):980–995, 2005. ISSN 10538119. doi: 10.1016/j.neuroimage.2005.06.070.

R. M. Neal. MCMC Using Hamiltonian Dynamics. In S. Brooks, A. Gelman, G. J. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*. CRC Press, New York, NY, USA, 2011.

P. G. Nestor, B. F. O'Donnell, R. W. McCarley, M. Niznikiewicz, J. Barnard, Z. Jen Shen, F. L. Bookstein, and M. E. Shenton. A new statistical method for testing hypotheses of neuropsychological/MRI relationships in schizophrenia: partial least squares analysis. *Schizophrenia Research*, 53 (1-2):57–66, 2002. ISSN 09209964. doi: 10.1016/S0920-9964(00)00171-7.

I. Nouretdinov, S. G. Costafreda, A. Gammerman, A. Chervonenkis, V. Vovk, V. Vapnik, and C. H. Fu. Machine learning classification with confidence: Application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. *NeuroImage*, 56(2):809–813, 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.05.023.

G. Orrù, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience & Biobehavioral Reviews*, 36(4):1140–1152, 2012. ISSN 01497634. doi: 10.1016/j.neubiorev.2012.01.004.

E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse Canonical Correlation Analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34, 2009. ISSN 1544-6115. doi: 10.2202/1544-6115.1406.

T. Paus, M. Keshavan, and J. N. Giedd. Why do many psychiatric disorders emerge during adolescence? *Nature Reviews Neuroscience*, 9(12):947–957, 2008. ISSN 1471-003X. doi: 10.1038/nrn2513.

D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in NumPyro, 2019. ISSN 23318422. URL http://arxiv.org/abs/1912.11554.

J. Piironen and A. Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11 (2):5018–5051, 2017. ISSN 19357524. doi: 10.1214/17-EJS1337SI.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, 29(3/4): 322–335, 1938. ISSN 00063444.

J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar, and S. E. Petersen. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, 59(3):2142–2154, 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2011.10.018.

S. Remes, A. Klami, and S. Kaski. Characterizing unknown events in MEG data with Group Factor Analysis. *Proceedings of the 3rd Workshop on Machine Learning and Interpretation in Neuroimaging (MLINI)*, 2013.

J. D. Rohrer and H. J. Rosen. Neuroimaging in frontotemporal dementia. *International Review of Psychiatry*, 25(2):221–229, apr 2013. ISSN 09540261. doi: 10.3109/09540261.2013.778822.

J. D. Rohrer, G. R. Ridgway, M. Modat, S. Ourselin, S. Mead, N. C. Fox, M. N. Rossor, and J. D. Warren. Distinct profiles of brain atrophy in frontotemporal lobar degeneration caused by progranulin and tau mutations. *NeuroImage*, 53(3):1070–1076, nov 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.12.088.

J. D. Rohrer, J. M. Nicholas, D. M. Cash, J. van Swieten, E. Dopper, L. Jiskoot, R. van Minkelen, S. A. Rombouts, M. J. Cardoso, S. Clegg, M. Espak, S. Mead, D. L. Thomas, E. De Vita, M. Masellis, S. E. Black, M. Freedman, R. Keren, B. J. MacIntosh, E. Rogaeva, D. Tang-Wai, M. C. Tartaglia, R. Laforce, F. Tagliavini, P. Tiraboschi, V. Redaelli, S. Prioni, M. Grisoli, B. Borroni, A. Padovani, D. Galimberti, E. Scarpini, A. Arighi, G. Fumagalli, J. B. Rowe, I. Coyle-Gilchrist, C. Graff, M. Fallström, V. Jelic, A. K. Ståhlbom, C. Andersson, H. Thonberg, L. Lilius, G. B. Frisoni, M. Pievani, M. Bocchetta, L. Benussi, R. Ghidoni, E. Finger, S. Sorbi,

B. Nacmias, G. Lombardi, C. Polito, J. D. Warren, S. Ourselin, N. C. Fox, and M. N. Rossor. Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the Genetic Frontotemporal dementia Initiative (GENFI) study: A cross-sectional analysis. *The Lancet Neurology*, 14(3):253–262, 2015. ISSN 14744465. doi: 10.1016/S1474-4422(14)70324-2.

R. Rosipal and N. Krämer. Overview and recent advances in Partial Least Squares. In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, editors, *Subspace, latent structure and feature selection*, pages 34–51, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-34138-3.

L. L. Russell, C. V. Greaves, M. Bocchetta, J. Nicholas, R. S. Convery, K. Moore, D. M. Cash, J. van Swieten, L. Jiskoot, F. Moreno, R. Sanchez-Valle, B. Borroni, R. Laforce, M. Masellis, M. C. Tartaglia, C. Graff, E. Rotondo, D. Galimberti, J. B. Rowe, E. Finger, M. Synofzik, R. Vandenberghe, A. de Mendonça, F. Tagliavini, I. Santana, S. Ducharme, C. Butler, A. Gerhard, J. Levin, A. Danek, M. Otto, J. D. Warren, J. D. Rohrer, M. N. Rossor, N. C. Fox, I. O. Woollacott, R. Shafei, C. Heller, R. Guerreiro, J. Bras, D. L. Thomas, S. Mead, L. Meeter, J. Panman, J. Papma, J. Poos, R. van Minkelen, Y. Pijnenburg, M. Barandiaran, B. Indakoetxea, A. Gabilondo, M. Tainta, M. de Arriba, A. Gorostidi, M. Zulaica, J. Villanua, Z. Diaz, S. Borrego-Ecija, J. Olives, A. Lladó, M. Balasa, A. Antonell, N. Bargallo, E. Premi, M. Cosseddu MPsych, S. Gazzina, A. Padovani, R. Gasparotti, S. Archetti, S. Black, S. Mitchell, E. Rogaeva, M. Freedman, R. Keren, D. Tang-Wai, L. Öijerstedt, C. Andersson, V. Jelic, H. Thonberg, A. Arighi, C. Fenoglio, E. Scarpini, G. Fumagalli, T. Cope, C. Timberlake, T. Rittman, C. Shoesmith, R. Bartha, R. Rademakers, C. Wilke, H. O. Karnarth, B. Bender, R. Bruffaerts, P. Vandamme, M. Vandenbulcke, C. B. Ferreira, G. Miltenberger, C. Maruta MPsych, A. Verdelho, S. Afonso, R. Taipa, P. Caroppo, G. Di Fede, G. Giaccone, C. Muscio, S. Prioni, V. Redaelli, G. Rossi, P. Tiraboschi, D. Duro NPsych, M. R. Almeida, M. Castelo-Branco, M. J. Leitão, M. Tabuas-Pereira, B. Santiago, S. Gauthier, P. Rosa-Neto, M. Veldsman, P. Thompson, T. Langheinrich, C. Prix, T. Hoegen, E. Wlasich, S. Loosli, S. Schonecker, E. Semler, and S. Anderl-Straub. Social cognition impairment in genetic frontotemporal dementia within the GENFI cohort. *Cortex*, 133:384–398, dec 2020. ISSN 19738102. doi: 10.1016/j.cortex.2020.08.023.

R. Sala-Llonch, S. M. Smith, M. Woolrich, and E. P. Duff. Spatial parcellations, spectral filtering, and connectivity measures in fMRI: Optimizing for discrimination. *Human Brain Mapping*, 40(2):407–419, feb 2019. ISSN 10970193. doi: 10.1002/hbm.24381. URL https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.24381https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.24381https://onlinelibrary.wiley.com/doi/10.1002/hbm.24381.

J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2016(4):e55, 2016. ISSN 23765992. doi: 10.7717/peerj-cs.55.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6 (2):461–464, 1978. doi: 10.1214/aos/1176344136.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, jun 2004. ISBN 9780521813976. doi: 10.1017/CBO9780511809682.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. Van Den Driessche, T. Graepel, and D. Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017. ISSN 14764687. doi: 10.1038/nature24270.

S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. Network modelling methods for fMRI. *NeuroImage*, 54(2):875–891, 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.08.063.

S. M. Smith, D. Vidaurre, C. F. Beckmann, M. F. Glasser, M. Jenkinson, K. L. Miller, T. E. Nichols, E. C. Robinson, G. Salimi-Khorshidi, M. W. Woolrich, D. M. Barch, K. Uğurbil, and D. C. Van Essen. Functional connectomics from resting-state fMRI. *Trends in Cognitive Sciences*, 17(12):666–682, 2013. ISSN 13646613. doi: 10.1016/j.tics.2013.09.016.

S. M. Smith, T. E. Nichols, D. Vidaurre, A. M. Winkler, T. E. Behrens, M. F. Glasser, K. Ugurbil, D. M. Barch, D. C. Van Essen, and K. L. Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior, 2015. ISSN 15461726.

J. S. Snowden, S. Rollinson, J. C. Thompson, J. M. Harris, C. L. Stopford, A. M. Richardson, M. Jones, A. Gerhard, Y. S. Davidson, A. Robinson, L. Gibbons, Q. Hu, D. DuPlessis, D. Neary, D. M. Mann, and S. M. Pickering-Brown. Distinct clinical and pathological characteristics of frontotemporal dementia associated with C9ORF72 mutations. *Brain*, 135(3): 693–708, 2012. ISSN 14602156. doi: 10.1093/brain/awr355.

L. P. Spear. Effects of adolescent alcohol consumption on the brain and behaviour. *Nature Reviews Neuroscience*, 19(4):197–214, 2018. ISSN 14710048. doi: 10.1038/nrn.2018.10.

Stan Development Team. Stan modeling language users guide and reference manual, version 2.26, 2019.

T. Suvitaival, J. A. Parkkinen, S. Virtanen, and S. Kaski. Cross-organism toxicogenomics with group factor analysis. *Systems Biomedicine*, 2(4):71–80, 2014. ISSN 2162-8130. doi: 10.4161/sysb.29291.

M. Symms, H. R. Jäger, K. Schmierer, and T. A. Yousry. A review of structural magnetic resonance neuroimaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(9):1235–1244, 2004. ISSN 0022-3050. doi: 10.1136/jnnp.2003.032714.

B. T. Thomas Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, B. Fischl, H. Liu, and R. L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011. ISSN 0022-3077. doi: 10.1152/jn.00338.2011.

D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. 2016. URL http://arxiv.org/abs/1610.09787.

V. Uurtio, J. M. Monteiro, J. Kandola, J. Shawe-Taylor, D. Fernandez-Reyes, and J. Rousu. A tutorial on Canonical Correlation Methods. *ACM Computing Surveys*, 50(6):1–33, 2017. ISSN 03600300. doi: 10.1145/3136624.

A. Vallesi, A. R. McIntosh, M. P. Alexander, and D. T. Stuss. fMRI evidence of a functional network setting the criteria for withholding a re-

sponse. *NeuroImage*, 45(2):537–548, 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2008.12.032.

S. van Erp, D. L. Oberski, and J. Mulder. Shrinkage priors for Bayesian penalized regression, 2019. ISSN 10960880.

D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, and K. Ugurbil. The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80:62–79, 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.05.041.

J. Viinikanoja, A. Klami, and S. Kaski. Variational Bayesian mixture of Robust CCA models. In J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 370–385, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15939-8.

S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. *Proceedings of the 28th International Conference on Machine Learning*, pages 457–464, 2011.

S. Virtanen, A. Klami, S. Khan, and S. Kaski. Bayesian Group Factor Analysis. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1269–1277, La Palma, Canary Islands, 2012. PMLR.

L. V. Vogelsmeier, J. K. Vermunt, E. van Roekel, and K. De Roover. Latent Markov Factor Analysis for fxploring measurement model changes in time-intensive longitudinal studies. *Structural Equation Modeling*, 26(4):557–575, 2019. ISSN 15328007. doi: 10.1080/10705511.2018.1554445.

N. D. Volkow, G. F. Koob, R. T. Croyle, D. W. Bianchi, J. A. Gordon, W. J. Koroshetz, E. J. Pérez-Stable, W. T. Riley, M. H. Bloch, K. Conway, B. G. Deeds, G. J. Dowling, S. Grant, K. D. Howlett, J. A. Matochik, G. D. Morgan, M. M. Murray, A. Noronha, C. Y. Spong, E. M. Wargo, K. R. Warren, and S. R. Weiss. The conception of the ABCD study: From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, 32:4–7, 2018. ISSN 18789293. doi: 10.1016/j.dcn.2017.10.002.

S. Waaijenborg, P. C. Verselewel de Witt Hamer, and A. H. Zwinderman. Quantifying the association between gene expressions and DNA-markers by penalized Canonical Correlation Analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008. ISSN 1544-6115. doi: 10.2202/1544-6115.1329.

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

N. Weiskopf, A. Lutti, G. Helms, M. Novak, J. Ashburner, and C. Hutton. Unified segmentation based correction of R1 brain maps for RF transmit field inhomogeneities (UNICORT). *NeuroImage*, 54(3):2116–2124, 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.10.023.

N. Weiskopf, J. Suckling, G. Williams, M. M. Correia, B. Inkster, R. Tait, C. Ooi, E. T. Bullmore, and A. Lutti. Quantitative multi-parameter mapping of R1, PD*, MT, and R2* at 3T: a multi-center validation. *Frontiers in Neuroscience*, 7:95, 2013. ISSN 1662-453X. doi: 10.3389/fnins.2013.00095.

J. Wen, E. Varol, A. Sotiras, Z. Yang, G. B. Chand, G. Erus, H. Shou, G. Hwang, F. the Alzheimer, D. Neuroimaging Initiative, and C. Davatzikos. Multi-scale semi-supervised clustering of brain images: deriving disease subtypes. *bioRxiv*, 2021. doi: 10.1101/2021.04.19.440501.

J. L. Whitwell, S. A. Przybelski, S. D. Weigand, R. J. Ivnik, P. Vemuri, J. L. Gunter, M. L. Senjem, M. M. Shiung, B. F. Boeve, D. S. Knopman, J. E. Parisi, D. W. Dickson, R. C. Petersen, C. R. Jack, and K. A. Josephs. Distinct anatomical subtypes of the behavioural variant of frontotemporal dementia: A cluster analysis study. *Brain*, 132(11):2932–2946, nov 2009. ISSN 14602156. doi: 10.1093/brain/awp232.

A. M. Winkler, M. A. Webster, D. Vidaurre, T. E. Nichols, and S. M. Smith. Multi-level block permutation. *NeuroImage*, 123:253–268, 2015. ISSN 10959572. doi: 10.1016/j.neuroimage.2015.05.092.

A. M. Winkler, O. Renaud, S. M. Smith, and T. E. Nichols. Permutation inference for canonical correlation analysis. *NeuroImage*, 220:117065, 2020. ISSN 10959572. doi: 10.1016/j.neuroimage.2020.117065.

D. M. Witten and R. J. Tibshirani. Extensions of sparse Canonical Correlation Analysis with applications to genomic data. *Statistical Applications*

in Genetics and Molecular Biology, 8(1):1–27, 2009. ISSN 1544-6115. doi: 10.2202/1544-6115.1470.

D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009. ISSN 1465-4644. doi: 10.1093/biostatistics/kxp008.

H. Wold. Partial Least Squares. In *Encyclopedia of the statistical sciences.*, pages 581–591. Wiley, 1985.

C. H. Xia, Z. Ma, R. Ciric, S. Gu, R. F. Betzel, A. N. Kaczkurkin, M. E. Calkins, P. A. Cook, A. García de la Garza, S. N. Vandekar, Z. Cui, T. M. Moore, D. R. Roalf, K. Ruparel, D. H. Wolf, C. Davatzikos, R. C. Gur, R. E. Gur, R. T. Shinohara, D. S. Bassett, and T. D. Satterthwaite. Linked dimensions of psychopathology and connectivity in functional brain networks. *Nature Communications*, 9(1):3003, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05317-y.

A. L. Young, R. V. Marinescu, N. P. Oxtoby, M. Bocchetta, K. Yong, N. C. Firth, D. M. Cash, D. L. Thomas, K. M. Dick, J. Cardoso, J. van Swieten, B. Borroni, D. Galimberti, M. Masellis, M. C. Tartaglia, J. B. Rowe, C. Graff, F. Tagliavini, G. B. Frisoni, R. Laforce, E. Finger, A. de Mendonça, S. Sorbi, J. D. Warren, S. Crutch, N. C. Fox, S. Ourselin, J. M. Schott, J. D. Rohrer, D. C. Alexander, C. Andersson, S. Archetti, A. Arighi, L. Benussi, G. Binetti, S. Black, M. Cosseddu, M. Fallström, C. Ferreira, C. Fenoglio, M. Freedman, G. G. Fumagalli, S. Gazzina, R. Ghidoni, M. Grisoli, V. Jelic, L. Jiskoot, R. Keren, G. Lombardi, C. Maruta, L. Meeter, S. Mead, R. van Minkelen, B. Nacmias, L. Öijerstedt, A. Padovani, J. Panman, M. Pievani, C. Polito, E. Premi, S. Prioni, R. Rademakers, V. Redaelli, E. Rogaeva, G. Rossi, M. Rossor, E. Scarpini, D. Tang-Wai, H. Thonberg, P. Tiraboschi, A. Verdelho, M. W. Weiner, P. Aisen, R. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowki, A. W. Toga, L. Beckett, R. C. Green, A. J. Saykin, J. Morris, L. M. Shaw, Z. Khachaturian, G. Sorensen, L. Kuller, M. Raichle, S. Paul, P. Davies, H. Fillit, F. Hefti, D. Holtzman, M. M. Mesulam, W. Potter, P. Snyder, A. Schwartz, T. Montine, R. G. Thomas, M. Donohue, S. Walter, D. Gessert, T. Sather, G. Jiminez, D. Harvey, M. Bernstein, P. Thompson, N. Schuff, B. Borowski, J. Gunter, M. Senjem, P. Vemuri, D. Jones, K. Kantarci, C. Ward, R. A. Koeppe, N. Foster, E. M. Reiman,

K. Chen, C. Mathis, S. Landau, N. J. Cairns, E. Householder, L. Taylor-Reinwald, V. Lee, M. Korecka, M. Figurski, K. Crawford, S. Neu, T. M. Foroud, S. Potkin, L. Shen, K. Faber, S. Kim, K. Nho, L. Thal, N. Buckholtz, M. Albert, R. Frank, J. Hsiao, J. Kaye, J. Quinn, B. Lind, R. Carter, S. Dolen, L. S. Schneider, S. Pawluczyk, M. Beccera, L. Teodoro, B. M. Spann, J. Brewer, H. Vanderswag, A. Fleisher, J. L. Heidebrink, J. L. Lord, S. S. Mason, C. S. Albers, D. Knopman, K. Johnson, R. S. Doody, J. Villanueva-Meyer, M. Chowdhury, S. Rountree, M. Dang, Y. Stern, L. S. Honig, K. L. Bell, B. Ances, M. Carroll, S. Leon, M. A. Mintun, S. Schneider, A. Oliver, D. Marson, R. Griffith, D. Clark, D. Geldmacher, J. Brockington, E. Roberson, H. Grossman, E. Mitsis, L. de Toledo-Morrell, R. C. Shah, R. Duara, D. Varon, M. T. Greig, P. Roberts, M. Albert, C. Onyike, D. D'Agostino, S. Kielb, J. E. Galvin, B. Cerbone, C. A. Michel, H. Rusinek, M. J. de Leon, L. Glodzik, S. De Santi, P. M. Doraiswamy, J. R. Petrella, T. Z. Wong, S. E. Arnold, J. H. Karlawish, D. Wolk, C. D. Smith, G. Jicha, P. Hardy, P. Sinha, E. Oates, G. Conrad, O. L. Lopez, M. A. Oakley, D. M. Simpson, A. P. Porsteinsson, B. S. Goldstein, K. Martin, K. M. Makino, M. S. Ismail, C. Brand, R. A. Mulnard, G. Thai, C. McAdams-Ortiz, K. Womack, D. Mathews, M. Quiceno, R. Diaz-Arrastia, R. King, M. Weiner, K. Martin-Cook, M. DeVous, A. I. Levey, J. J. Lah, J. S. Cellar, J. M. Burns, H. S. Anderson, R. H. Swerdlow, L. Apostolova, K. Tingus, E. Woo, D. H. Silverman, P. H. Lu, G. Bartzokis, N. R. Graff-Radford, F. Parfitt, T. Kendall, H. Johnson, M. R. Farlow, A. M. Hake, B. R. Matthews, S. Herring, C. Hunt, C. H. van Dyck, R. E. Carson, M. G. MacAvoy, H. Chertkow, H. Bergman, C. Hosein, B. Stefanovic, C. Caldwell, G. Y. R. Hsiung, H. Feldman, B. Mudge, M. Assaly, A. Kertesz, J. Rogers, C. Bernick, D. Munic, D. Kerwin, M. M. Mesulam, K. Lipowski, C. K. Wu, N. Johnson, C. Sadowsky, W. Martinez, T. Villena, R. S. Turner, K. Johnson, B. Reynolds, R. A. Sperling, K. A. Johnson, G. Marshall, M. Frey, B. Lane, A. Rosen, J. Tinklenberg, M. N. Sabbagh, C. M. Belden, S. A. Jacobson, S. A. Sirrel, N. Kowall, R. Killiany, A. E. Budson, A. Norbash, P. L. Johnson, J. Allard, A. Lerner, P. Ogrocki, L. Hudson, E. Fletcher, O. Carmichael, J. Olichney, C. DeCarli, S. Kittur, M. Borrie, T. Y. Lee, R. Bartha, S. Johnson, S. Asthana, C. M. Carlsson, S. G. Potkin, A. Preda, D. Nguyen, P. Tariot, S. Reeder, V. Bates, H. Capote, M. Rainka, D. W. Scharre, M. Kataki, A. Adeli, E. A. Zimmerman, D. Celmins, A. D. Brown, G. D. Pearlson, K. Blank, K. Anderson, R. B. Santulli, T. J. Kitzmiller, E. S.

Schwartz, K. M. Sink, J. D. Williamson, P. Garg, F. Watkins, B. R. Ott, H. Querfurth, G. Tremont, S. Salloway, P. Malloy, S. Correia, H. J. Rosen, B. L. Miller, J. Mintzer, K. Spicer, D. Bachman, S. Pasternak, I. Rachinsky, D. Drost, N. Pomara, R. Hernando, A. Sarrael, S. K. Schultz, L. L. Ponto, H. Shim, K. E. Smith, N. Relkin, G. Chaing, L. Raudin, A. Smith, K. Fargher, B. A. Raj, T. Neylan, J. Grafman, M. Davis, R. Morrison, J. Hayes, S. Finley, K. Friedl, D. Fleischman, K. Arfanakis, O. James, D. Massoglia, J. J. Fruehling, S. Harding, E. R. Peskind, E. C. Petrie, G. Li, J. A. Yesavage, J. L. Taylor, and A. J. Furst. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nature Communications*, 9(1):4273, 2018. ISSN 20411723. doi: 10.1038/s41467-018-05892-0.

A. L. Young, M. Bocchetta, L. L. Russell, R. S. Convery, G. Peakman, E. Todd, D. M. Cash, C. V. Greaves, J. van Swieten, L. Jiskoot, H. Seelaar, F. Moreno, R. Sanchez-Valle, B. Borroni, R. Laforce, M. Masellis, M. C. Tartaglia, C. Graff, D. Galimberti, J. B. Rowe, E. Finger, M. Synofzik, R. Vandenberghe, A. de Mendonça, F. Tagliavini, I. Santana, S. Ducharme, C. Butler, A. Gerhard, J. Levin, A. Danek, M. Otto, S. Sorbi, S. C. Williams, D. C. Alexander, and J. D. Rohrer. Characterizing the clinical features and atrophy patterns of MAPT - related frontotemporal dementia with disease progression modeling. *Neurology*, 2021. ISSN 0028-3878. doi: 10.1212/wnl.0000000000012410.

Z. Zhang, Q. K. Telesford, C. Giusti, K. O. Lim, and D. S. Bassett. Choosing wavelet methods, filters, and lengths for functional brain network construction. *PLoS ONE*, 11(6):e0157243, jun 2016. ISSN 19326203. doi: 10.1371/journal.pone.0157243. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157243.

S. Zhao, C. Gao, S. Mukherjee, and B. E. Engelhardt Zhao. Bayesian group factor analysis with structured sparsity. Technical report, 2016.