Multivariate Analysis of Neuroimaging data using Latent Variables

Valerie Bradley, Ana Ignatieva, Yuxi Jiang, James Thornton

September 24, 2019

Abstract

We review the latent factor analysis approach used by Miller et al. (2016) to analyse correlation structure within UK Biobank data. We examine several alternative methods of latent factor analysis and apply them to Human Connectome Project data.

1 Introduction

There are a large variety of methods used to identify latent structure in sets of data. The desire to identify an underlying structure in data is often driven by the need for dimensionality reduction, especially when the number of variables approaches or exceeds the number of observations. Latent factor analysis can uncover relationships between the observed variables, as well as allow for more meaningful downstream modelling and analysis.

First, we give the theory of several latent factor analysis methods. Then, we review the approach used by Miller et al. (2016) to analyse a subset of UK Biobank data. Further, we apply the methods discussed to Human Connectome Project (HCP) data and present the results.

2 Theory

2.1 PCA, ICA and CCA

Principal components analysis (PCA) and independent components analysis (ICA) are two common methods of latent factor analysis for data from a single mode (method of observation). Though these methods have existed for decades, Cunningham and Ghahramani (2015) developed a useful general framework for relating various latent factor analysis methods, which we will adopt here.

Generally, given a data matrix $\mathbf{X} \in R^{d\times n}$, ICA and PCA find linear transformations $\mathbf{P} \in R^{r\times d}$, where $r \leq d$, that optimizes some objective function, $f(\cdot)$. The transformed data is given as $\mathbf{Y} = \mathbf{PX} \in R^{r\times n}$. ICA and PCA differ from each other, and from other latent factor analysis methods, in the choice of objective function (Cunningham and Ghahramani (2015)).

PCA identifies orthogonal factors $\mathbf{M}$ that explain the maximum amount of variance in the data. This is equivalent to choosing the objective function to be the reconstruction error under the ℓ_2 norm:

$$f_X(M) = ||\mathbf{X} - \mathbf{M}^T\mathbf{MX}||_F^2$$

If $\mathbf{M}$ is chosen to minimize $f_{\mathbf{X}}(\mathbf{M})$, then $\mathbf{P} = \mathbf{M}^T$. The principal components $\mathbf{MY}$ explain the maximal amount of variance of $\mathbf{X}$ and are uncorrelated. PCA is generally used for dimensionality reduction, in which case $r << d$.

The critical difference between ICA and PCA is that the factors identified by ICA are by definition independent and additive, instead of simply uncorrelated (Cunningham and Ghahramani (2015)). ICA assumes that the data $\mathbf{X}$ is a mixture of independent sources. $\mathbf{P}$ is then considered to be the "unmixing" matrix. Generally, ICA aims to maximize the independence between factors of $\mathbf{P}$, which can be done either by minimizing factors' mutual information or by minimizing the non-Gaussianity of factors.

Unlike the usual application of PCA, ICA is usually implemented with $r = d$. If the number of independent components desired is less than the dimension of the original data, PCA can be applied to the data to first reduce the number of dimensions before ICA is run, as in Miller et al. (2016).

Canonical correlation analysis (CCA) is analogous to PCA for two co-occurring data sets. Say we have two data sets $\mathbf{X}_a \in R^{d_a\times n}$ and $\mathbf{X}_b \in R^{d_b\times n}$, CCA identifies transformations $\mathbf{P}_a$ and $\mathbf{P}_b$ of the data that maximize the correlation between (typically lower-dimensional) projections $\mathbf{Y}_a = \mathbf{P}_a\mathbf{X}_a$ and $\mathbf{Y}_b = \mathbf{P}_b\mathbf{X}_b$. CCA constrains the projections to have unit variance and be uncorrelated, or $\frac{1}{n}\mathbf{Y}_a\mathbf{Y}_a^T = I$, $\frac{1}{n}\mathbf{Y}_b\mathbf{Y}_b^T = I$ and $\frac{1}{n}\mathbf{Y}_a\mathbf{Y}_b^T = \mathbf{\Lambda}$ for a diagonal matrix $\mathbf{\Lambda}$. Most commonly, CCA identifies canonical variants by maximizing:

$$\text{tr}\left(M_a^T(X_xX_a^T)^{-1/2}X_aX_b^T(X_bX_b^T)^{-1/2}M_b\right)$$

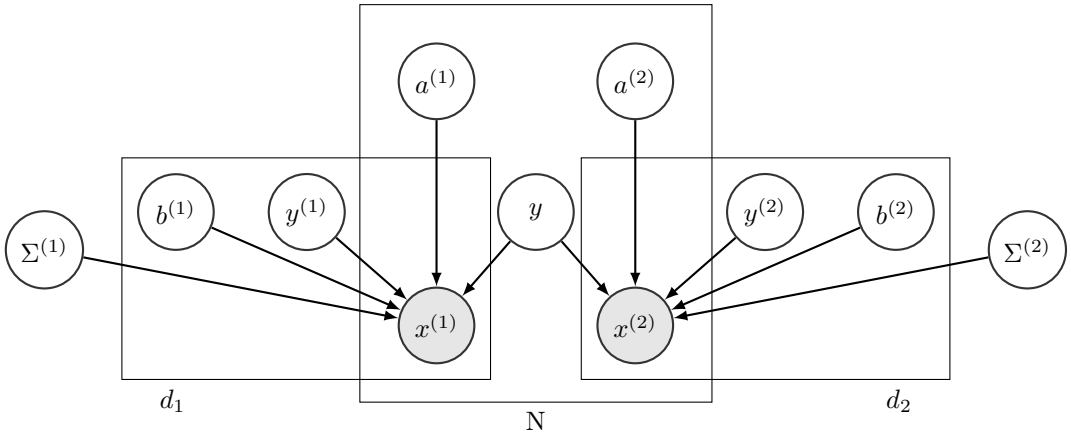

Figure 1: Data generating process for BCCA

subject to $M_a \in \mathcal{O}^{d_a \times r}$ and $M_b \in \mathcal{O}^{d_b \times r}$.

Traditionally, these methods are not generative, and are instead used for dimensionality reduction or exploratory analysis. The factors recovered do not distinguish between variance specific to features and common across features, or between variance from actual latent signal in the data and variance that is due solely to noise or measurement error. However, there are conceptualizations of these methods as generative processes, like Bayesian CCA and probabilistic PCA.

2.2 Bayesian CCA

Bayesian CCA, introduced by Klami et al. (2013), conceptualizes classical CCA as a fully Bayesian generative model. Bayesian CCA factorizes the residuals from traditional CCA into variation specific to each of the data sets being considered. One can think of Bayesian CCA of performing PCA on the variation not explained by the canonical variants.

We assume that there is some underlying shared random variable, $\mathbf{Y} \in R^{r \times n}$, that is mapped linearly into m different observation spaces. $\mathbf{X}^{(1)} \in R^{d_1 \times n}$ is the data in the first observation space (or data set), and more generally, $\mathbf{X}^{(m)} \in R^{d_m \times n}$ is the data in the m^{th} observation space, observed for the same n individuals, or samples. The variation in $\mathbf{X}^{(m)}$ that is not explained by $\mathbf{Y}$, and is specific to the m^{th} data set, is modeled as another latent variable, $\mathbf{Y}^{(m)}$.

The generative model for Bayesian CCA is:

$$\begin{aligned} \mathbf{Y} &\sim N(0, \mathbf{I}) \\ \mathbf{Y}^{(m)} &\sim N(0, \mathbf{I}) \\ \mathbf{X}^{(m)} &\sim N(\mathbf{A}^{(m)}\mathbf{Y} + \mathbf{B}^{(m)}\mathbf{Y}^{(m)}, \mathbf{\Sigma}^{(m)}) \end{aligned}$$

The $\mathbf{\Sigma}^{(m)} \in R^{d_m \times d_m}$ are diagonal matrices, indicating that noise across features is independent. $\mathbf{A}^{(m)} \in R^{d_m \times r}$ is the loading matrix that maps the shared latent variable $\mathbf{Y}$ onto the m^{th} observation space. Similarly, $\mathbf{B}^{(m)} \in R^{d_m \times r_m}$ is the loading matrix that maps the latent variable $\mathbf{Y}^{(m)}$ specific to the m^{th} observation space that space.

We could also define the Bayesian CCA in terms of only the shared latent variable $\mathbf{Y}$ by integrating out the observation-specific variables $\mathbf{Y}^{(m)}$ (Klami et al., 2013). This results in a simpler model that models the observation-specific latent factor only through correlated noise:

$$\begin{aligned} \mathbf{Y} &\sim N(0, \mathbf{I}) \\ \mathbf{X}^{(m)} &\sim N(\mathbf{A}^{(m)}\mathbf{Y}, \mathbf{\Psi}^{(m)}) \end{aligned}$$

Where $\mathbf{\Psi}^{(m)} = \mathbf{B}^{(m)}\mathbf{B}^{(m)^T} + \mathbf{\Sigma}^{(m)}$ and $\mathbf{\Psi}$ is a covariance matrix.

The maximum likelihood solution to this model results in the same canonical weights as classical CCA, up to a rotation. While the solutions are similar, Bayesian CCA identifies a single latent variable $\mathbf{Y}$ that explains the joint variation between data sets, while classical CCA obtains separate but correlated variables for each data set being examined. Though Bayesian CCA originated as the Bayesian formulation of classical 2-mode CCA, Bayesian group factor analysis (GFA) extends the theory to apply to multi-view, or $m > 2$ applications (Zhao et al., 2016).

2.3 Model 1 - ARD Prior

In order to perform inference on the Bayesian CCA, it is necessary to specify priors for the covariance matrix $\mathbf{\Psi}$ and the linear combinations $\mathbf{A}$. A natural choice of prior for $\mathbf{\Psi}^{(m)}$ is the inverse-Wishart distribution

$$\mathbf{\Psi}^{(m)} \sim IW(S_0, \nu_0)$$

with ν_0 degrees of freedom. The conjugate inverse-Wishart prior would draw positive definite matrices as long as the data dimensionality is less than the degrees of freedom.

For the linear mappings $\mathbf{A}^{(m)}$, both Klami and Kaski (2007) and Wang (2007) independently proposed the use of the automatic relevance determination (ARD; Neal (1996)) prior, given as

$$\begin{aligned}\mathbf{ARD}(\mathbf{A}^{(m)}, \alpha^{(m)}|\alpha_0, \beta_0) &= \prod_{k=1}^{K} p(\mathbf{a}_k^{(m)}|\alpha_k^{(m)}) p(\alpha_k^{(m)}|\alpha_0, \beta_0) \\ \alpha_k^{(m)} &\sim \text{Gamma}(\alpha_0, \beta_0) \\ \mathbf{a}_k^{(m)} &\sim \mathcal{N}(\mathbf{0}, (\alpha_k^{(m)})^{-1}\mathbf{I})\end{aligned}$$

the parameters α_0, β_0 are normally set to small values to ensure the prior has a wide support and relatively non-informative. By using the ARD prior, the posterior of the model allows automatic component selection by pushing the $\alpha_k^{(m)}$ of the unnecessary components towards infinity, so that the posterior mass of the corresponding $\mathbf{a}_k^{(m)}$ is highly concentrated around 0 (Klami et al., 2013).

Several inference techniques are available given the choices of priors above, such as the variational mean-field algorithm (Wang, 2007) and the Gibbs sampling algorithm (Klami and Kaski, 2007). However, the inference algorithms generally need to invert the covariance matrices $\mathbf{\Psi}^{(m)}$ in every step, which means the inference techniques are difficult to apply to data with large dimensionalities. Klami et al. (2013) introduced a solution to this by rewriting the model to

$$\begin{aligned}\mathbf{Y} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathbf{X} &\sim \mathcal{N}(\mathbf{WY}, \mathbf{\Sigma})\end{aligned}$$

where

$$\begin{aligned}\mathbf{W} &= \begin{bmatrix}\mathbf{W}^{(1)} \\ \mathbf{W}^{(2)}\end{bmatrix} = \begin{bmatrix}\mathbf{A}^{(1)} & \mathbf{B}^{(1)} & \mathbf{0} \\ \mathbf{A}^{(2)} & \mathbf{0} & \mathbf{B}^{(2)}\end{bmatrix} \\ \mathbf{\Sigma} &= \begin{bmatrix}\mathbf{\Sigma}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}^{(2)}\end{bmatrix}\end{aligned}$$

where $\mathbf{\Sigma}^{(m)} = \sigma_m^2 \mathbf{I}$ where a Gamma prior is given to the precision parameter $\tau_m = \sigma_m^{-2}$, and group-wise ARD is defined for the prior of $\mathbf{W}$ as

$$p(\mathbf{W}) = \prod_{m=1}^{2} \mathbf{ARD}(\mathbf{W}^{(m)}|\alpha_0, \beta_0)$$

which allows $\mathbf{W}$ to obtain the structure proposed for the model, and the inactive components of each view will all be pushed towards 0. Bayesian inference for this model is achieved through variational approximation. This inference method implicitly solves the rotational invariance, and helps the model to identify between shared and view-specific components (Klami et al., 2013). Multi-view analysis (where m can take values greater than 2) is also feasible as an extension of this setup by constructing $\mathbf{W}$ and $\mathbf{\Sigma}$ accordingly.

2.4 Model 2 - BASS

BASS uses three parameter beta (TPB) priors on global, factor-specific and local components, to extend the GFA model to enable element-wise and column-wise shrinkage, thus inducing a sparsity structure on the loading matrix (Zhao et al., 2016). The generative model for BASS is:

$$\begin{aligned}\mathbf{y_i} &= \mathbf{A}\mathbf{x_i} + \epsilon_\mathbf{i} \\ \mathbf{x_i} &\sim \mathcal{N}_k(0, \mathbf{I_k}) \\ \epsilon_\mathbf{i} &\sim \mathcal{N}_p(0, \mathbf{\Sigma})\end{aligned}$$

The TPB distribution for a random variable $Z \in (0,1)$ has the density:

$$f(z;a,b,\nu) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\nu^b z^{b-1}(1-z)^{a-1}\big(1+(\nu-1)z\big)^{-(a+b)}$$

We put TPB priors on the parameters as follows:

$$\begin{aligned}
\sigma_j^{-2} &\sim \text{Ga}(1, 0.3) \\
\rho &\sim \text{TPB}(e, f, \nu) \\
\psi_h &\sim \text{TPB}(c, d, \frac{1}{\rho} - 1) \\
\phi_{jh} &\sim \text{TPB}(a, b, \frac{1}{\psi_h} - 1) \\
\lambda_{jh} &\sim \mathcal{N}(0, \frac{1}{\phi_{jh}} - 1)
\end{aligned}$$

and set $a = b = c = d = e = f = 0.5$.

For interpreting factors when analysing the output of latent factor models, identifiability of the solution is a key issue. For instance, we have rotational invariance when right multiplying the joint loading matrix by an orthogonal matrix $\mathbf{Q^T}$ and left multiplying $\mathbf{x}$ by $\mathbf{Q}$ produces an identical covariance matrix and likelihood. The BASS model addresses rotational invariance as non-sparse rotations of the loading matrix would violate the prior structure (Zhao et al., 2016).

3 Miller et al. (2016) Methodology and Review

Miller et al. (2016) studied imaging data from the UK Biobank, with the goal of uncovering associations between brain imaging data and other subject measurements of interest. The approach can be broken down into data pre-processing, and the subsequent multivariate analysis.

3.1 Data pre-processing

As stated in the paper (Miller et al., 2016), it is difficult to identify useful insights directly from raw images and hence a number of pre-processing steps were carried out to remove artifacts, align images and combine results form multiple images, all in order to extract meaningful features for analysis. The outputs of this are referred to as image-derived phenotypes (IDPs), and number $p_{IDP} = 2,501$ in this analysis. Given that these processing steps are significant and complex, we will focus our review on the processing after useful the IDPs are produced.

At a high level, the data processing steps are as follows:

- Remove IDP variables with inadequate data
- Apply rank-based inverse Gaussian transformation on all variables
- Remove confounding variables via regression

Variable selection The number of non-IDP variables were reduced to $p_2 = 1,100$ based on a number of rules 'similar to' those detailed by Smith et al. (2015). The rules per Smith et al. (2015) were: select only variables with at least 250 non-null subject measurements, do not include variables that have extreme outliers (identified the largest L2 distance to median being more than 100 of the average L2 distance to median), do not include 'undesirable' measurements subjectively seen as 'minor', unrelated or already accounted for. Although code is provided, it is difficult to understand the reasoning for assessing whether some variables are 'minor', hence citing Smith et al. (2015) and expert judgment is insufficient for reproducibility, and does not appear to be a scalable variable selection method. Given the objective of the study was to identify associations in the variables, removing variables will not damage the associations that are discovered, however may limit discovery of unknown associations.

Normalisation of data All variables were passed through a rank-based inverse Gaussian transformation; this combined with the removal of variables with extreme values will coerce the data into being normal. The procedure detailed in Miller et al. (2016) can be viewed as linear transformations and truncations on the observations, hence preserving Gaussianity, and then ICA is performed. ICA on normal data has the undesired consequence of non-identifiablility as detailed by Hyvärinen and Oja (2000). The standard practice of coercing to Gaussianity may not be needed - especially for deterministic ICA and CCA. A discussion of the common but possibly unnecessary practice is detailed by Beasley et al. (2009).

Confounders Confounding variables including age, sex, head size and head motion were removed through regression (i.e. the residuals were taken forward). Removal of these variables appears justified given the strong yet uninteresting associations shown when the variables are included. The removal appears hypothesis lead as age and sex are described as common confounders in such studies, however the decision of regression and type of regression structure, versus other methods such as perhaps 'restriction' for gender, may also require further justification (see Pourhoseingholi et al. (2012)).

3.2 Multivariate CCA-ICA Study

After pre-processing, the approach taken by Miller et al. (2016) is of CCA+ICA, first described by Sui et al. (2010), with an extra step of reverse projection from PCA to the original feature space.

- Perform PCA on each of the IDP and non-IDP data-sets to reduce the dimensionality prior to CCA
- Deterministic CCA is performed on the blocks of IDP and non-IDP variables
- Permutation tests are used to identify significant CCA modes
- Project the CCA modes back to the observation space
- Perform deterministic joint ICA to promote sparsity in associations discovered

The motivation for this approach is clear. CCA is used to identify associations between blocks of data. Classical CCA requires a smaller number of variables than subjects to avoid a rank-deficient solution, which is achieved via PCA. Joint ICA is then used to group observed variables across blocks in aggregate, in a sparse manner for identifiability but based on the identified CCA associations in the mixing matrix.

PCA PCA was performed on both the IDP, $X^{\text{IDP}} \in R^{N \times P^{\text{IDP}}}$, and non-IDP variables, $X^{\text{non-IDP}} \in R^{N \times P^{\text{non-IDP}}}$, to return projections of the data in the principal component space. The projections shall be referred to as P^{IDP} and $P^{\text{non-IDP}}$, each of N rows and $r = 200$ columns. The cap on $r = 200$ components seems somewhat arbitrary, and it is not clear how much variance is lost because of this.

$$
\begin{aligned}
P^m &= X^m W^m \\
m &\in \{\text{IDP}, \text{non-IDP}\} \\
X^m \in R^{N \times p^m},\ W^m &\in R^{p^m \times r},\ P^m \in R^{N \times r}
\end{aligned}
$$

PCA variable-compression may have beneficial aspects as well as unusual complications. PCA is a typical pre-processing step for brain-imaging data, given the large noise to signal ratio and being high dimension (see Sui et al. (2010)). By compressing the variable dimension into uncorrelated components, it is possible that PCA reduces bilateral correlation among variables within each block (IDP and non-IDP), as highly correlated variables would be grouped into the same components. PCA may also encourage sparsity in the CCA map from latent signals to principal components in each block; if principal components are uncorrelated, one would hope latent signals would not be mapped to multiple components in each group. Ad-hoc removal of similar variables in pre-processing would also be somewhat redundant given PCA. In addition, removal of noise may also help the analysis, although it is unclear why this would not have been carried out in the early pre-processing steps.

PCA followed by CCA PCA is performed prior to CCA (right matrix multiplication of X^m by W^m) and then a reverse transform is applied to get a compressed data structure in the original observed variable space, prior to ICA. The hope is that joint ICA on the compressed data structure will promote sparsity and highlight some interpretable associations between variables.

However, the process as described by Miller et al. (2016) appears fundamentally flawed. Deterministic CCA is used to identify latent signals Z and mixing matrices A^m such that $X^m = A^m Z + \epsilon_m$, where ϵ_m are errors. Miller et al. (2016) then identify significant signals using permutation hypothesis tests. However, our concern arises when these signals (scores, Z) are then projected back to the observation space by multiplying by the original data blocks to produce two matrices R_m: "multiply the nine CCA subject-weight vectors into the original IDP and non-IDP data matrices" (Miller et al., 2016). This transformation seems very unusual and is difficult to justify. The R_m matrices are joined on on the variable dimension, and then Joint ICA is performed on this unusual compressed, re-weighted matrix of shape $K \times p$.

$$
\begin{aligned}
R_m &= Z^m X \\
m &\in \{\text{IDP}, \text{non-IDP}\} \\
R &= \{R_{\text{IDP}} | R_{\text{non-IDP}}\} \\
X^m \in R^{N \times p^m},\ Z^m &\in R^{K \times N}
\end{aligned}
$$

This may have been a typo or transcription error. A similar but more intuitive approach would be to project the mixing matrices discoveded from CCA, A^m, onto the observational space by multiplying out the principal components by loadings, W^m. This would also result in a $K \times p$ matrix, which can be fed into ICA.

$$
\begin{aligned}
R_m &= (W^m A^m)^T \\
m &\in \{\text{IDP}, \text{non-IDP}\} \\
R &= \{R_{\text{IDP}} | R_{\text{non-IDP}}\} \\
W^m &\in R^{p^m \times r},\ WA^m \in R^{r \times K}
\end{aligned}
$$

The $(i,j)^{th}$ entry of R_m is $\sum_k a^m_{k,i} w^m_{j,k}$ where $a^m_{k,i}$ and $w^m_{j,k}$ are the $(k,i)^{th}$ and $(j,k)^{th}$entries of A^m and W_m respectively. $a^m_{k,i}$ corresponds to the co-ordinate of the i^{th} canonical variate that projects onto the k^{th} principal component dimension and $w^m_{j,k}$ corresponds to the weight or loading of the j^{th} observed variable in the k^{th} principal component of block m. This means that each $(i,j)^{th}$ entry in R_m will be large in absolute terms if observed variable j has a large weight in a principal component which in turn has a large weight in one of the signals that link the two data sets. Concatenating in the variable direction gives a new feature matrix, R of shape $K \times p$, where $p = p_{IDP} + p_{\text{non-IDP}} = 2501 + 1100$. The rows of R correspond to the canonical variates and columns to the observed variables. Given the weighting interpretation of the R_m described above, joint ICA on R will group observed variables weighted by how significantly (in an informal meaning) each variable relates to a principal component of its block and how that principal component contributes to the linking of the two blocks, with the constraint of some proxy for independence. Intuitively, if a variable x^m_p in one block m is strongly associated to a principal component P^m_j which is related to canonical variate Z_k, and variable $x^{m'}_{p'}$ of group m' has a high weighting in component P^m_j which is also related to z_k, then it is hoped joint ICA would group x^m_p and $x^{m'}_{p'}$ together, in a sparse manner.

Alternative methods This PCA-CCA-ICA method is quite a complicated and it is not clear whether it is actually valid and meaningful compared to other methods such as Bayesian CCA. It appears PCA was used for just variable compression and then reversely projected for interpretability with added complication. Similarly it also appears ICA was used post CCA for interpretability and to produce an identifiable solution. It may be more meaningful to perform the reverse PCA transformation after the ICA step, rather than before or interpret the PCA solution. Indeed, Bayesian CCA would not necessarily require this PCA variable compression and the sparsity desired from ICA can be introduced via a prior on the mixing matrix with seemingly more control, understanding and in a much simpler way. The method proposed by Miller et al. (2016) requires hypothesis testing to determine the number of latent variables, which is an added complication. Bayesian methods, such as using the ARD prior described above, would handle this complexity and learn the number of hidden variables.

4 Application to Human Connectome Project (HCP)

We used the `CCAGFA` R package, and the C++ `BASS` package (Zhao et al., 2016) to analyse a dataset from the Human Connectome Project (https://www.humanconnectome.org/).

4.1 Data Processing

Pre-processed data was taken from the HCP. This totalled 13 relational tables:

- Structural data
- Behavioural data
- DTI data
- rFMRI connectivity (15, 25, 100 parcellations)
- tFMRI data (7 tables, one for each task)

In order to perform some basic analysis, some naive data processing was undertaken. Across the original 13 tables, the 'Subject' field was used as the observational index. For all tables, observations with missing variable measurements were removed. Each table was then filtered to the intersection of 'Subject' field across all tables so that the same subjects had a complete set of variables across the data sources.

Categorical variables such as gender and age group were naively removed for simplicity, and only 'Thickness' variables were taken from the 'Structural' data source (expert guidance). Columns of each table were then centred and scaled to unit variance before running through GFA and BASS. Only the 15 parcellation connectivity rFMRI table was used, in the interest of run-time.

We also attempted to run GFA and BASS on all of the behavioural and imaging variables (after removing missing data). Neither package ran as expected; `CCAGFA` terminated with a positive-definiteness related error, and `BASS` could not calculate the log-likelihood and terminated without producing any parameter estimates. This suggests limitations related to the relative or absolute number of variables in the two observation modes and possibly due to the approximation methods implemented in the model implementations. It can also be noted that both packages (particularly BASS) are not particularly well documented, making it difficult to troubleshoot errors.

4.2 Group Factor Analysis

We performed GFA on 4 modes of data (DTI, rfMRI, structural thickness, and, behavioral data), and a BCCA which combines the imaging data (DTI, rfMRI and structural thickness) as one view and behavioral data as the other view.

GFA identified 47 latent factors and the result is given in Figure 10. We can see that most of the latent factors identified only contains information for one or two modalities. Factor 3 in the figure is shown to be the only active latent factor for all of the four views, while the three imaging modes also shares signals from factors 17. In addition, factors 25 and 44 are active in the behaviour data and three other image views except 'structural thickness' and 'rfmri_15' respectively.

The BCCA on the other hand identified 46 latent factors with the mixing matrix shown in Figure 9. 9 latent factors are identified to be active in both views, and the figure indicates the presence of 29 latent factors within the imaging data. It can be seen in the figure that even though the brain-image views were grouped, the model identified distinct factors primarily corresponding to each of the views. These were factors 15, 21 and 36.

4.3 BASS

We applied the BASS model to the Human Connectome Project Data. We performed 2 main analyses - one on 3 modes of imaging data (DTI, rfMRI and structural thickness), and one on those 3 modes of imaging data plus behavioral data.

4.3.1 Imaging data

BASS requires prior specification of the number of latent factors (k) to try to detect - we ran BASS on the three modes of imaging data for $k \in (5, 15, 30, 50)$. In theory, if BASS is initialized to look for a large enough number of latent factors, the structured sparsity (specifically global sparsity) will eliminate extraneous factors, thus identifying the true number of latent factors. Figure 4 shows BASS performed for $k = 5$. In this case, BASS has been initialized to look for far too few latent factors, and the resulting loadings are quite dense and uninterpretable. Figures 5 and 6 show how the sparsity of the loading matrix increases as k increases to 15 and 30, respectively. We found that with $k \in (5, 15, 30)$, BASS identified the maximum number of factors allowed. However, with $k = 50$, BASS only identified 48 latent factors, eliminating 2 extraneous factors. For the rest of the image-data-only analysis, we will focus on the results for $k = 50$.

Figure 2 shows the result of the application of BASS to 3 modes of imaging data with $k = 50$. The sparsity of the resulting factors is immediately apparently - indicating that the structured sparsity imposed on latent variable estimation through the hierarchical three parameter Beta distribution priors has effectively induced sparsity within factors, across factors (clustering) and of factors (eliminating extraneous factors). 75% of the imaging features had non-zero loadings in fewer than 4 latent factors, and most of the latent factors identified had non-zero loadings for fewer than 10 imaging features. The few dense factors only contained non-zero loadings for observations from a single mode.

Most of the factors identified only contain features from a single modality, however there were 3 latent factors that had non-zero loadings for features from 2 different modes - F3, F6 and F13. Interestingly, there is a large degree of overlap between the three multi-modal latent factors. Figure 7 compares the loadings on the observed variables across the three multi-modal latent factors. F3 and F6 contain non-zero loadings for almost identical sets of observed variables, yet have inverse loading values for some specific subsets of the observed variables. For example, CST-R and CST-L have large negative loadings in F3 and large positive loadings in F6. Additionally, the rfMRI variables have positive loadings in F3 and negative loadings in F6. The processes identified in these two latent factors seem to be complementary, or even inverses, of one another.

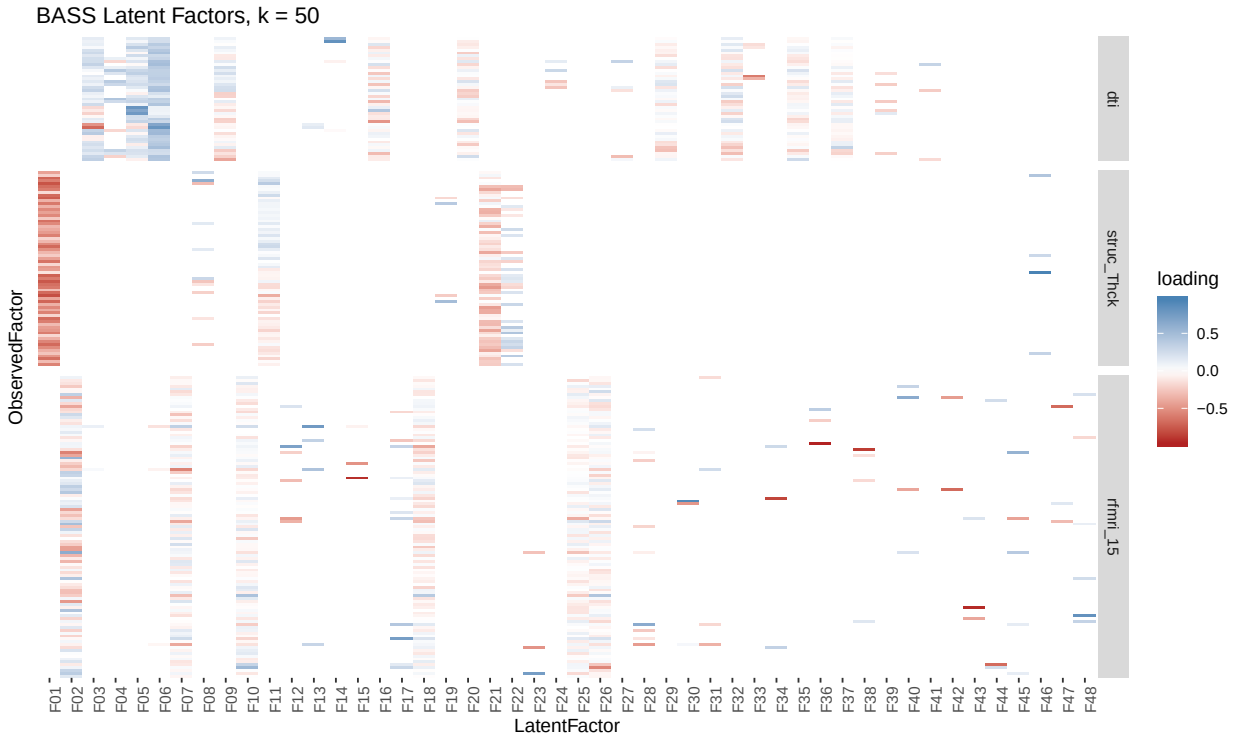

Figure 2: Heat map of loadings for BASS applied to 3 modes of imaging data (DTI, rfMRI and structural thickness). Initial number of latent variables set to 50. Rows represent observed variables grouped by data type, columns represent latent variables identified by BASS.

4.3.2 Imaging and behavioral Data

We ran BASS with an initial value of 50 latent factors on the imaging data, additionally including behavioural data. We view this as consisting of two data modalities (physical imaging variables vs. behavioural). The algorithm found 47 latent factors, again demonstrating the flexibility of the model in choosing the number of latent factors, due to the global sparsity induced by the prior structure.

The resulting loadings matrix has a sparsity structure as demonstrated in Figure 3. It can be noted that there is very little overlap between the behavioural and imaging modes; only four latent factors have non-zero loadings for both behavioural and imaging variables.

The latent factors which span both behavioural and imagine modalities are given in Figure 8. One latent factor relates behavioural data to DTI, while the other three factors link behavioural data with rfMRI measurements. It is interesting that one particular rfMRI variable (`rfmri_15_17`) occurs in all three of these factors.

5 Extensions

5.1 Prediction and Linear Modelling

As discussed above, latent factors are commonly used for exploratory analysis and dimensionality reduction. However, the recovered latent factors can also be used for classification of out-of-sample data. For example, Klami et al. (2013) gives an example of latent factors determined using BASS on a training set of news articles being used to classify similar news documents in an out-of-sample test set into meaningful subcategories. The analysis was able to accurately classify almost 75% of documents into the correct subcategory defined by latent factors.

Latent factors can also be useful in predictive modeling. Given a situation in which the number of predictors, p, is quite large relative to the number of observations, or where there is a high noise to signal ratio, latent factors are useful in identifying the underlying structures within the observed variables and capturing signal which is potentially predictive of an outcome. This could be considered automatic 'feature extraction' in the machine-learning literature. This helps avoid the need for extensive variable selection and lowers the risk of overfitting a model.

References

T Mark Beasley, Stephen Erickson, and David B Allison. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior genetics*, 39(5):580, 2009.

John P. Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16:2859–2900, 2015. URL `http://jmlr.org/papers/v16/cunningham15a.html`.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Arto Klami and Samuel Kaski. Local dependent components. pages 425–432, 2007.

Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(Apr):965–1003, 2013.

Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19 (11):1523, 2016.

Radford M Neal. *Bayesian learning for neural networks*. Springer-Verlag, 1996.

Mohamad Amin Pourhoseingholi, Ahmad Reza Baghestani, and Mohsen Vahedi. How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology from bed to bench*, 5(2):79, 2012.

Stephen M Smith, Thomas E Nichols, Diego Vidaurre, Anderson M Winkler, Timothy EJ Behrens, Matthew F Glasser, Kamil Ugurbil, Deanna M Barch, David C Van Essen, and Karla L Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature neuroscience*, 18(11):1565, 2015.

Jing Sui, Tülay Adali, Godfrey Pearlson, Honghui Yang, Scott R Sponheim, Tonya White, and Vince D Calhoun. A cca+ ica based model for multi-task brain imaging data fusion and its application to schizophrenia. *Neuroimage*, 51(1):123–134, 2010.

Chong Wang. Variational bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 18(3):905–910, 2007.

Shiwen Zhao, Chuan Gao, Sayan Mukherjee, and Barbara E Engelhardt. Bayesian group factor analysis with structured sparsity. *The Journal of Machine Learning Research*, 17(1):6868–6914, 2016.

A Plots and Figures

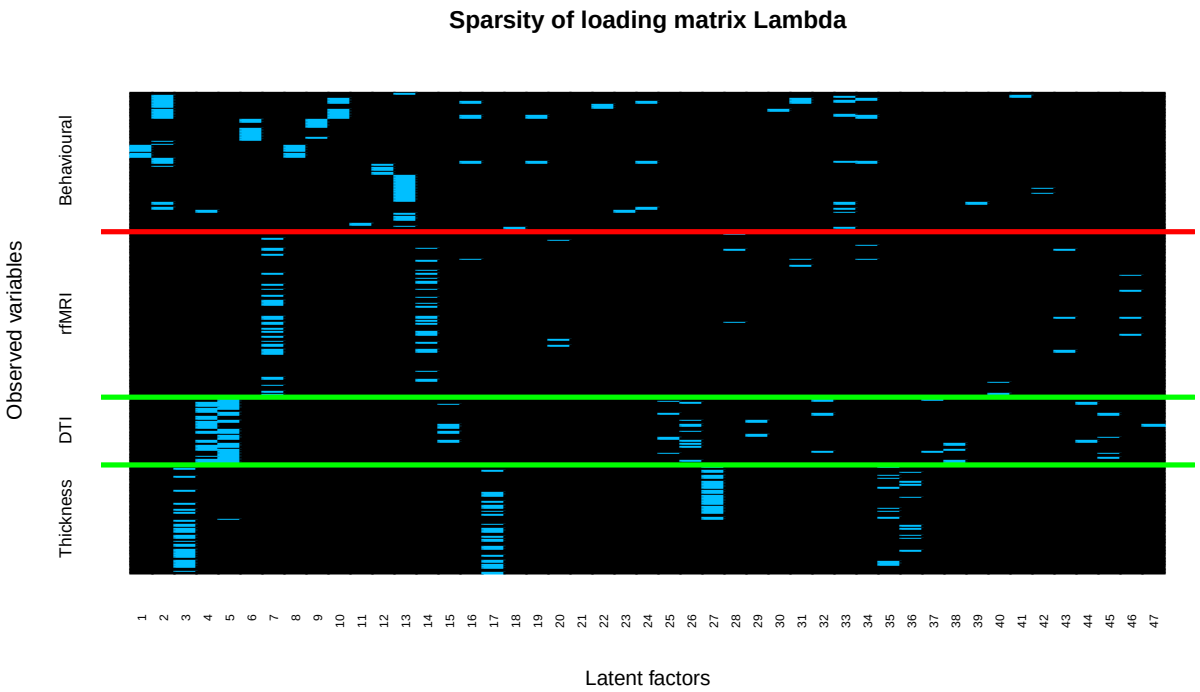

Figure 3: Sparsity map of the latent factor loadings matrix, with BASS applied to imaging data (DTI, rfMRI and structural thickness) and behavioural data. Blue lines show non-zero matrix entries. Columns represent the latent factors; rows represent the observed variables.

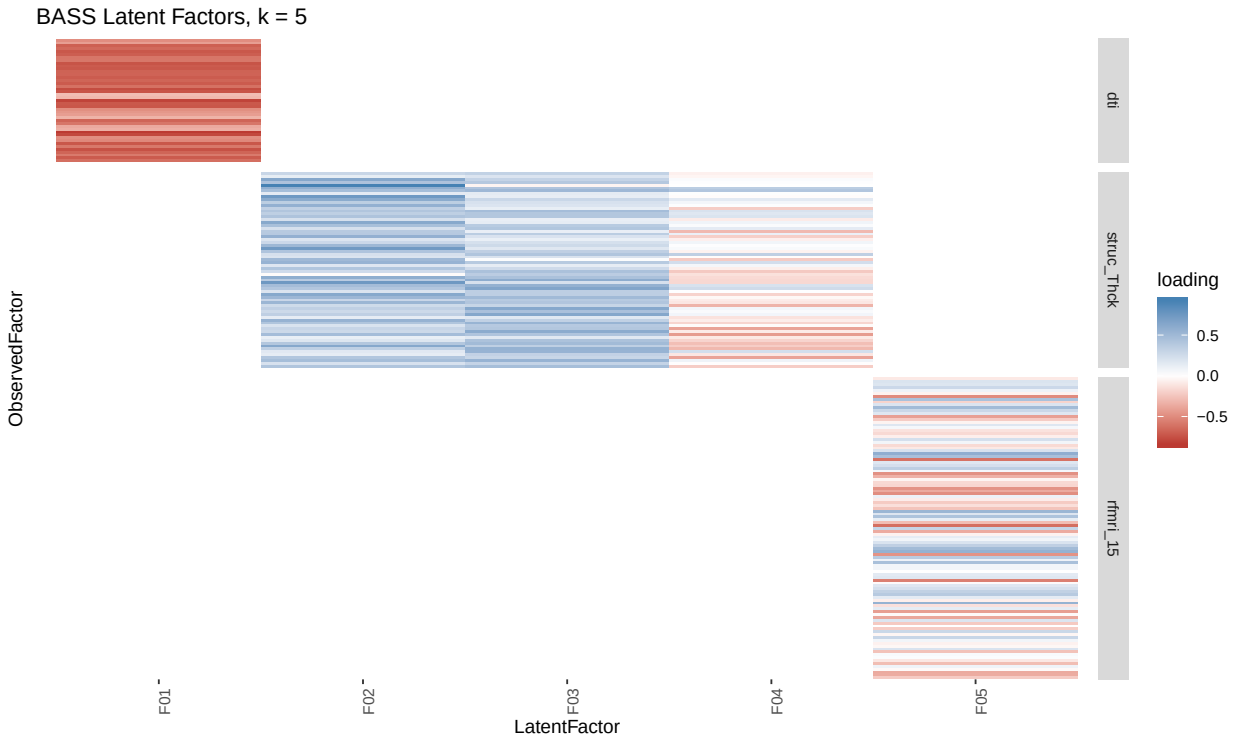

Figure 4: Heat map showing values of loadings for latent factors identified by BASS applied to 3 modes of imaging data (DTI, rfMRI and structural thickness). Initial number of latent factors was set to 5.

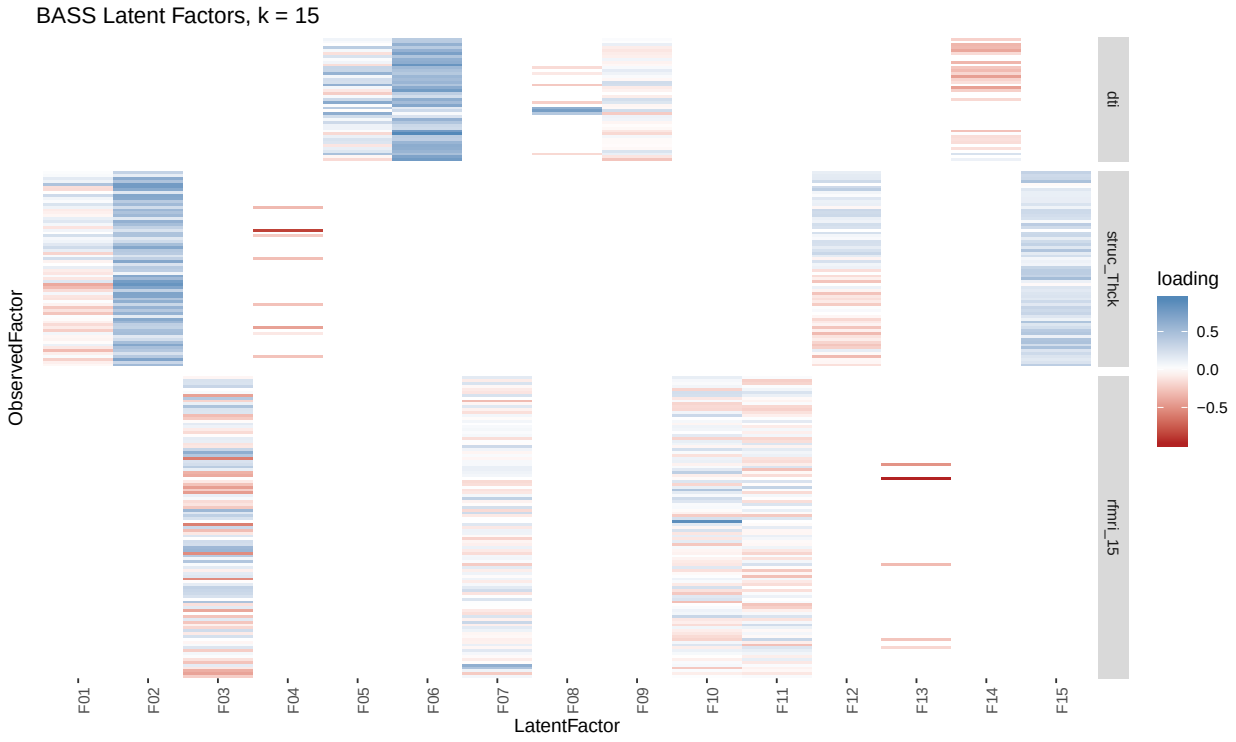

Figure 5: Heat map showing values of loadings for latent factors identified by BASS applied to 3 modes of imaging data (DTI, rfMRI and structural thickness). Initial number of latent factors was set to 15.

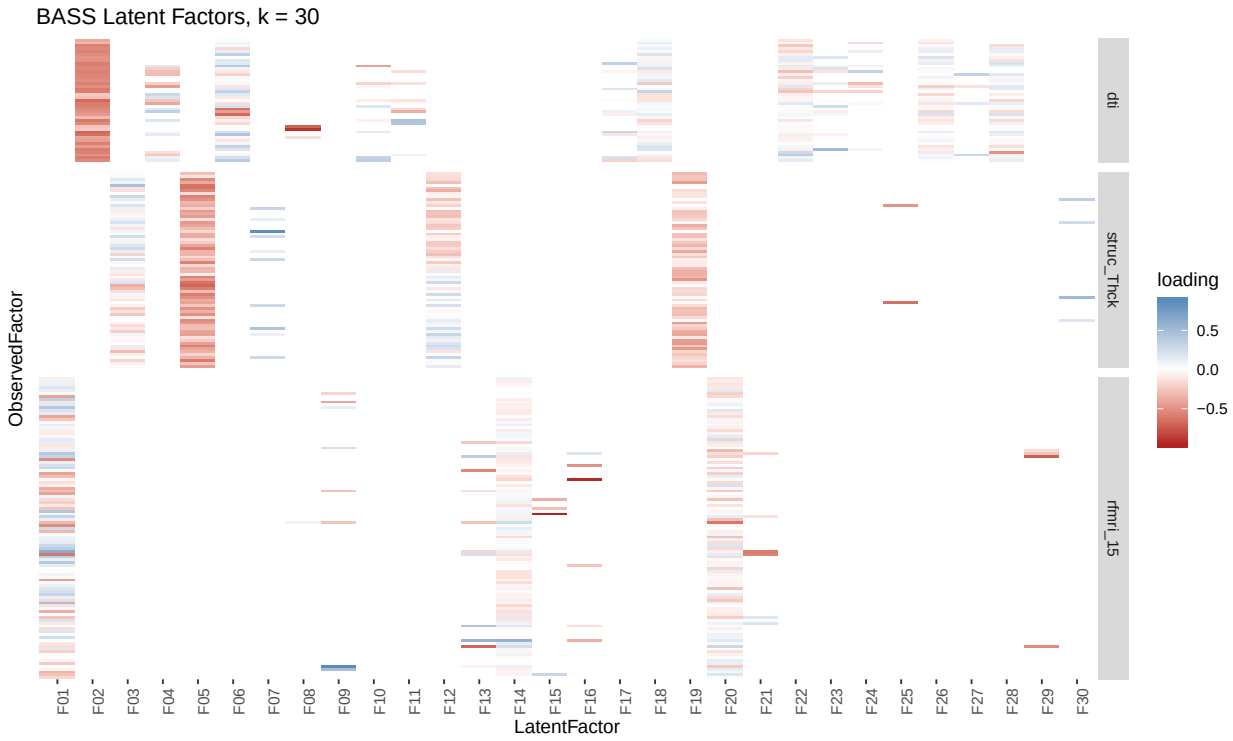

Figure 6: Heat map showing values of loadings for latent factors identified by BASS applied to 3 modes of imaging data (DTI, rfMRI and structural thickness).Initial number of latent factors was set to 30.

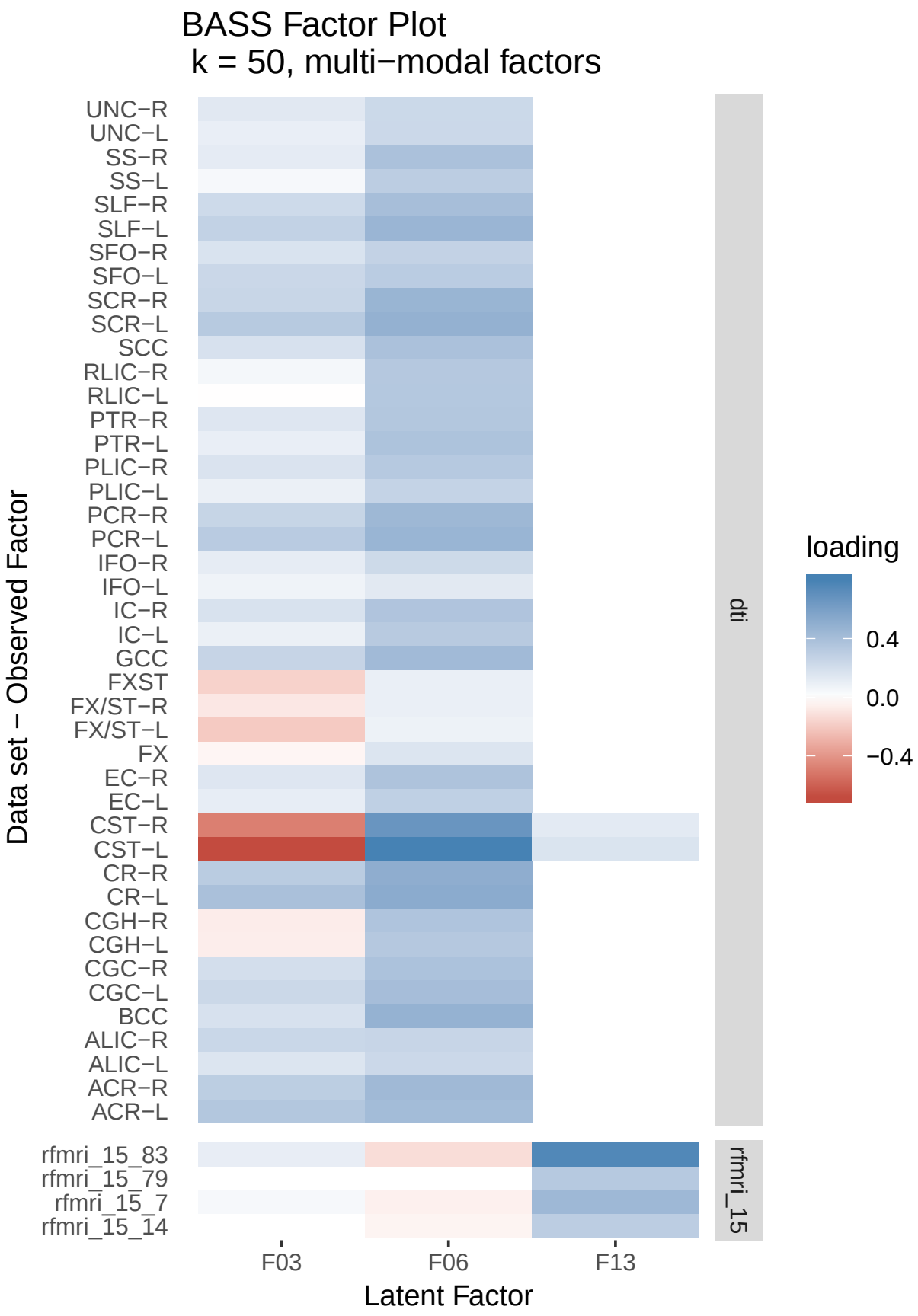

Figure 7: Heat map of latent factors from BASS with non-zero loadings from more than one data modality for $k = 50$ and imaging-only data.

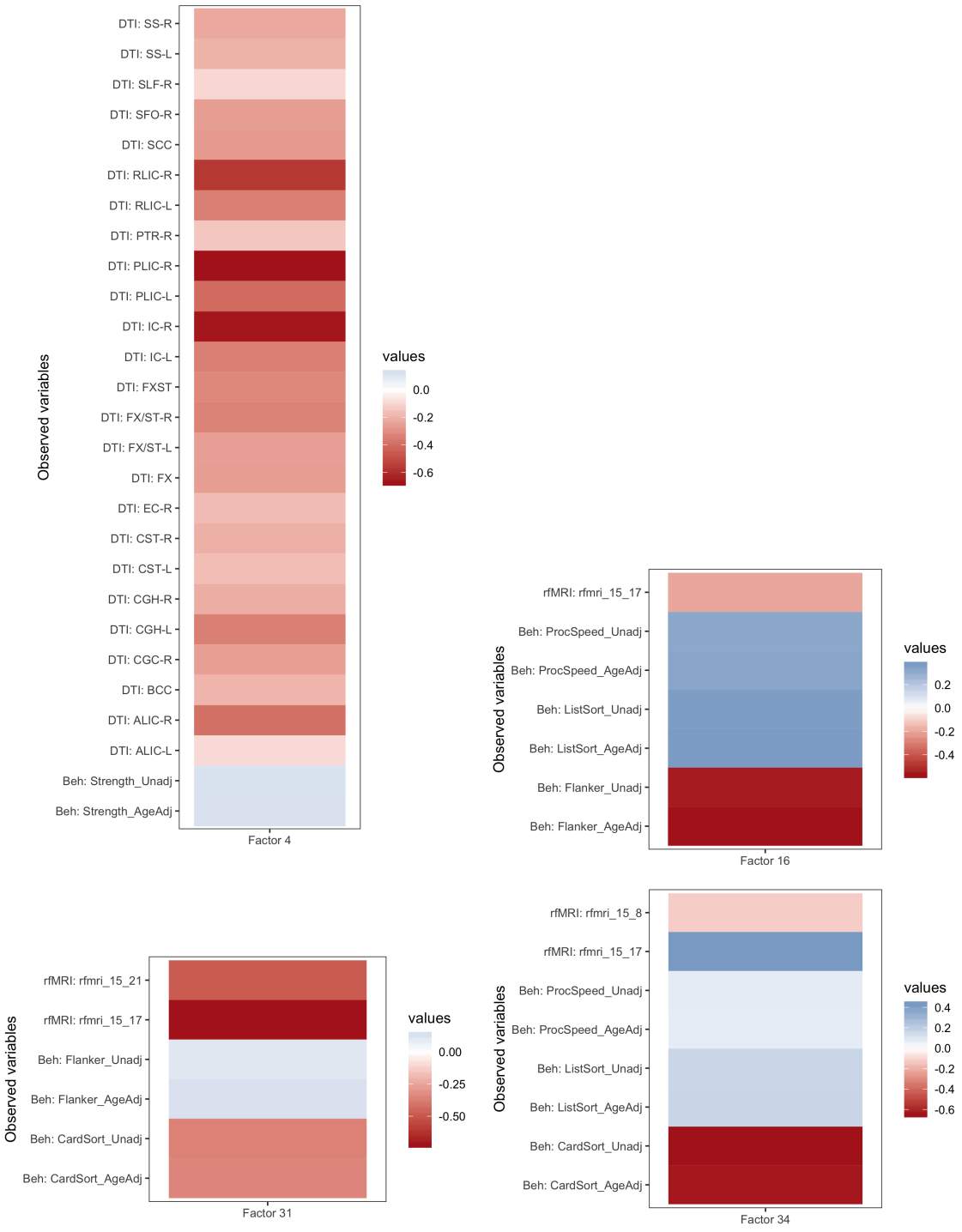

Figure 8: Heat map showing values of loadings for latent factors 4, 16, 31 and 34 which span both behavioural and imaging modalities

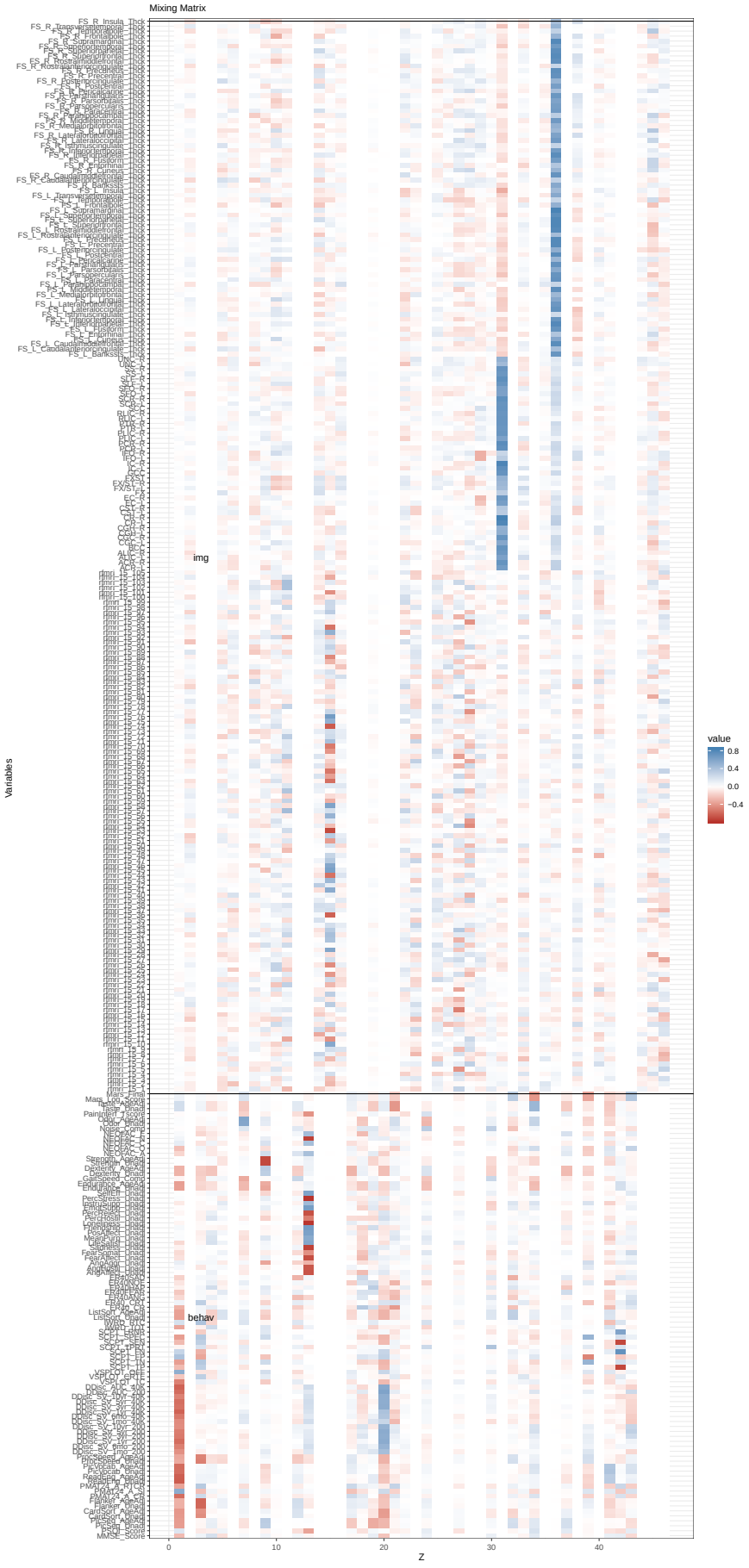

Figure 9: Heatmap of mixing matrix when CCA was performed using a combined view of brain image data, in addition to the behavioural data view

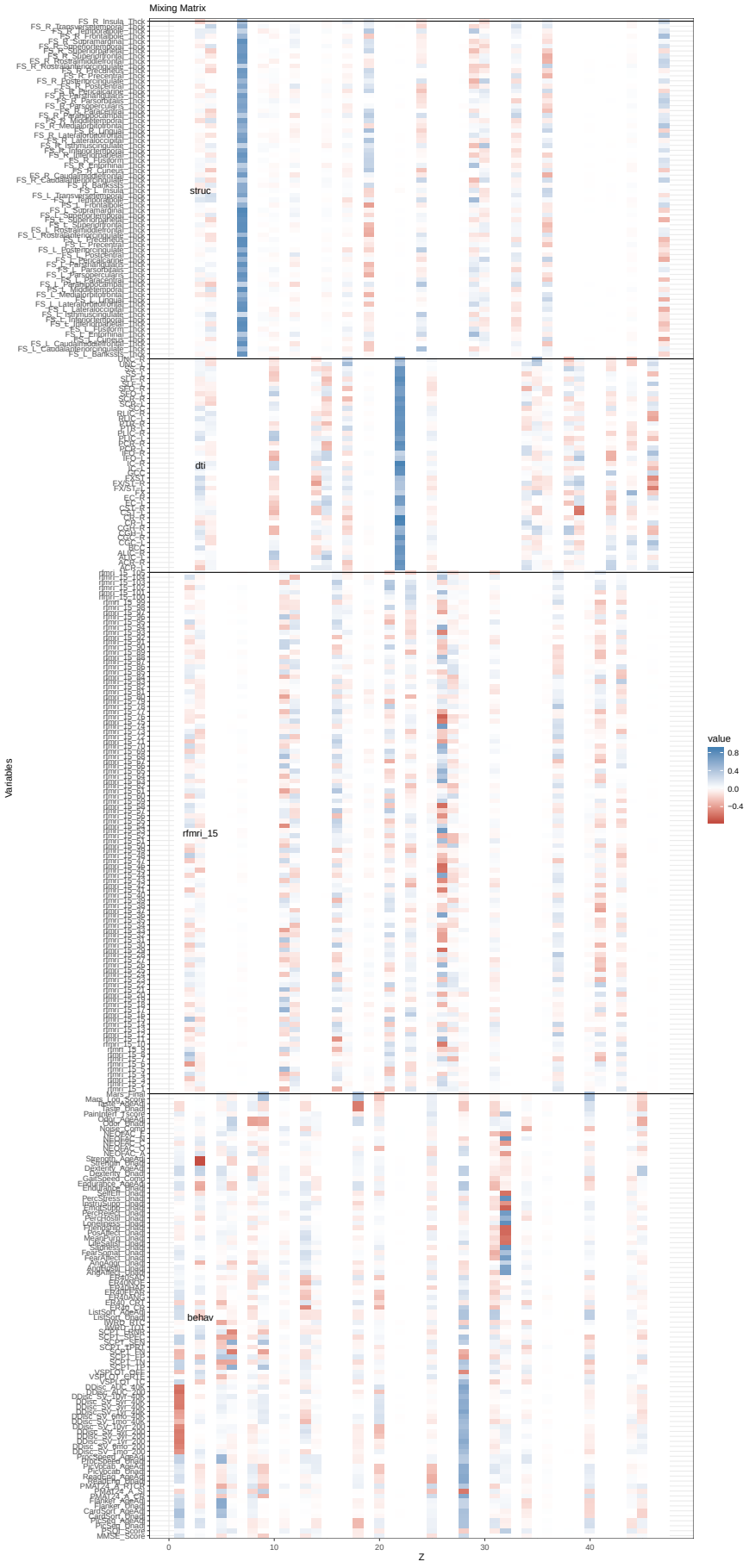

Figure 10: Heatmap of mixing matrix when GFA was performed using separate view of brain image data, in addition to the behavioural data view