The Extremes of Normative Modelling

Charlotte Fraza[1,2], Mariam Zabihi[1,2,5] Christian F. Beckmann[1,2,4], Andre F. Marquand[1,2,3]

[1] Donders Institute for Brain, Cognition, and Behavior, Radboud University, Nijmegen, the Netherlands
[2] Department of Cognitive Neuroscience, Radboud University Medical Center, Nijmegen, the Netherlands
[3] Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, London, UK
[4] Centre for Functional MRI of the Brain (FMRIB), Nuffield Department of Clinical Neurosciences, Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK
[5]MRC Unit for Lifelong Health & Ageing, University College London (UCL), London, United Kingdom

"Man can believe the impossible, but man can never believe the improbable" Oscar Wilde (Intentions, 1891)

Abstract

In this paper, we propose a new framework for understanding and modelling neurobiological extreme atypicalities for individual participants. We combine the strength of normative models, to make predictions for individual patients, with multivariate extreme value statistics, which allows us to model the outer centiles accurately, enabling accurate estimation of risk and detection of atypicality. In general, most statistical methods focus principally on estimating the mean of the distribution and are not appropriate to capture the tails of the distribution. We believe that the tails carry most of the information for predicting psychopathy in neurological and psychiatric disorders and should therefore be modelled carefully.
In the univariate case, we demonstrate how to fit a generalized Pareto distribution with the peaks over threshold method. For the multivariate case, we present a novel framework that allows us to capture the multivariate tails of the distribution fully. The model relies on the so-called tail pairwise dependence matrix (TPDM). The TPDM is similar to a covariance matrix, but specified for the tails. From the TPDM we can create a reduced basis of vectors that explain most of the structure underlying the extreme values. The method is similar to covariance principal component analysis, but uniquely made for the extreme values. To evaluate this framework, we use a range of neuroimaging-derived measures from the UK Biobank dataset, including structural, functional and diffusion image-derived phenotypes. We find that the first extreme principal component mostly consists of whole-brain volume and area measures and the second extreme principal component is mostly compromised of white matter tracts. We further demonstrate the link between extreme brain deviations in the individual from a reference cohort to behaviour.

Keywords: tail dependence; dimension reduction; Regular variation; UK biobank; Bayesian Linear Regression; Normative modelling; extreme value statistics;

Introduction

In this paper, we propose a new framework for modelling the extreme tail dependence in neuroimaging data. In many domains, understanding the extreme values of a given variable are of great importance since they have the most significant impact on outcomes of interest. For example, in climatology, extreme floods and storms cause the most substantial damage to human infrastructure (Marsh and Dale, 2002). Similarly, in finance, extreme deviations from an expected pattern can indicate market crashes (Zhao et al., 2010). In neuroimaging, we are similarly interested in the extreme values of a distribution, for example, detecting activation changes or conceptualizing psychopathology as an extreme deviation from an expected pattern. However, reliable estimates of the tails of a given distribution are difficult to obtain because that is where data are by definition the sparsest. Yet, accurate estimation of tail behaviour is crucial in order to obtain reliable estimates of risk. In addition, we are often interested in obtaining such estimates across multiple variables simultaneously, which further increases the difficulty in obtaining reliable estimates that properly account for correlations in the tails of the distribution.
Multivariate extreme value statistics tries to understand the structure underlying extreme probabilities. The basis for this framework is the multivariate extreme value model created by (Cooley and Thibaud, 2019). Modelling extreme values is almost unknown in clinical neuroimaging, but is extensively used in many fields, e.g. in meteorology to predict extreme weather events and in finance to predict stock market crashes (National Academies of Sciences, 2016; Poon et al., 2004). It has become clear that to understand these extreme phenomena we need to focus on the outer tails of the distribution. The underlying intuition is that central tendency (i.e. the average) is not useful to understand the limits of a complex system since the system is more likely to fail in extreme conditions than under average conditions. We will show that similar extreme value models can be estimated using brain data in order to obtain statistical estimates of extreme deviations from an expected pattern. To achieve this, we combine our extreme value statistics approach with normative modelling. Normative modelling provides statistical estimates of the distribution of a given brain measure across the population and can be used to estimate individual risk or deviation scores. Combining the strength of normative modelling with extreme value statistics will allow us to accurately estimate extreme deviation risk scores for the individual. The combination of these models enables us to establish brain-behaviour mappings that are more sensitive to the structure underlying the extreme deviations from an expected pattern (e.g. normative brain development). We often inaccurately estimate the effects of the tails of the distributions and therefore in neuroimaging we underestimate or overestimate the chance of an individual having these extreme brain deviations with regular statistics. We predict that the accurate understanding of small probabilities of extreme events is crucial for understanding several types of behaviour and psychiatric disorders.
In this paper, we propose a newly developed tool for understanding and exploring patterns in extreme brain deviations for the individual, based on extreme value statistics and normative modelling. First, we give a didactic introduction to the theory underlying normative modelling and extreme value statistics, in order to contextualise our approach. Then, we develop an exploratory statistical framework of multivariate extremes for individual participants. Our main contributions are:

(i) Developing the first framework for multivariate extreme value statistics in Neuroimaging data.

(ii) Applying and demonstrating it on a large imaging dataset UK biobank.
(iii) Establishing clear underlying tail pairwise dependence between several brain phenotypes.
(iv) Summarizing the tail pairwise dependence within a few extreme principal components.
(v) Relating these extreme principal components to several behavioural phenotypes.

Theoretical background

Normative models

We start with normative modelling in the multivariate extreme value framework. Normative modelling is a method created for understanding and modelling the heterogeneity in the population. Similar to growth charts, which are used to model the distribution of height or weight of children according to their age and sex, normative models can be used to model neuroimaging data according to a set of clinical or demographic covariates (e.g. age, gender and site) (Marquand et al., 2016a, 2016b). The output gives us the mean and centiles of variation for the data as a function of the covariates. Using normative models, we shift away from the case-control setting and are able to make inferences of deviation from the reference cohort for the individual. Although normative models have been used extensively to model the reference cohort and several psychiatric problems (Rutherford et al., 2022; Wolfers et al., 2020, 2018; Zabihi et al., 2019), it is not clear what the optimal approach is to understand how deviations in individual neurobiological measures relate to psychopathology at the subject level, especially if deviations are observed in multiple correlated variables. If we consider that atypical brain development underlies many brain disorders (Insel, 2014; Marquand et al., 2019), we can understand why mapping these extreme deviations becomes so important.

Conceptualising multivariate deviations using a watershed model

In most cases, it is important to capture the multivariate structure underlying the most extreme deviations from a reference cohort for an individual, rather than considering each measure separately. To illustrate this point, we give a simple analogy inspired by hydrology, where extreme value statistics are commonly used (Figure 1). In order to understand the effects of a flooding event, predicting the frequency and intensity of extreme values at different nodes on the river becomes much more important than predicting the change in the mean, because the extreme events are what cause the most damage. For example, when we consider the flooding of a river basin or watershed, it becomes crucial to map the dependency structure of the tails at several gauging stations. The extreme data from the gauging stations are correlated because they are part of a connected river network. We illustrate this in Figure 1, where we give a visual representation of the different types of extreme data that can exist at individual gauging stations that together cause a flood. It could be that one river in the network is having local water elevation, due to, for example, snow melting or it could be that the local water elevation is raised at multiple locations due to a storm across the land. When we try to understand the flooding of the river basin it is important that we model the extremal dependence structure between the different gauging stations and water flow correctly, as it is not only the local extremes, but also a co-occurrence of multivariate extremes that determine how severely the flooding will be

(Chiapino and Sabourin, 2017; Engelke and Ivanovs, 2020). Following Cannon's watershed model (Cannon and Keller, 2006), we can understand the connection with psychopathology. Pathophysiological brain states can manifest either due to local brain disruptions in certain brain regions or nodes in the network or via more widespread disruption in connected brain systems that cause psychopathology by a 'downstream' aggregation of different effects. Thus, there are many different ways that small causal effects (the tributaries) can be aggregated to yield the full-blown expression of a disorder (the river delta). Therefore, psychopathology can be understood as manifesting at the extreme end of continuous axes of variation across the population. Moreover, these atypicalities can be spread across multiple brain systems and need not overlap across individuals and also need not be reflected in only one data modality.

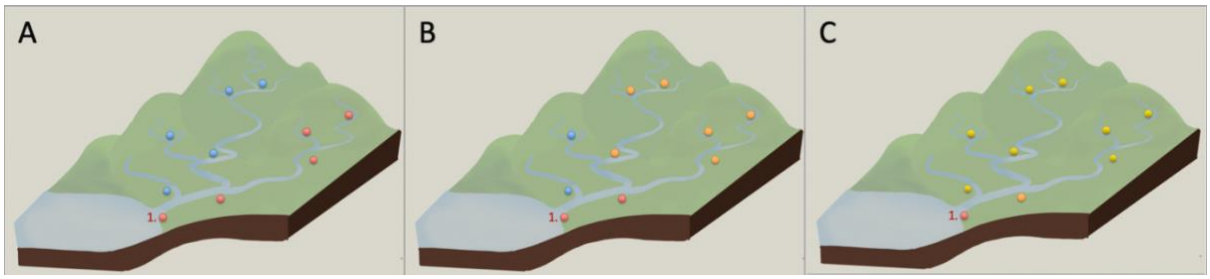


Figure 1 - Visual example of different types of extreme river elevations that can cause local flooding at location 1.

Similar generalisations of this model have also been proposed to understand cognitive function including genetic, environmental and other factors similar to the watershed model (Kievit et al., 2016). This model focuses on understanding how different complex behavioural traits can be understood as the downstream consequences of several upstream neural and genetic contributions. Our model has a similar conceptual ground only focusing our attention on extreme values, instead of the entire distribution.

Introduction to extreme value statistics

We will rely on extreme value statistics to accurately model behaviour in the tails of the distribution. We can apply univariate extreme value statistics when fitting the model for one brain imaging phenotype. There are classically two methods to achieve this, referred to in the extreme value statistics literature as (i) 'block maxima', which performs inference on the basis of the single maximum for each block or group of data (e.g. a period of time in a time-series or region of interest in a brain image) or (ii) 'exceedances' (a.k.a. 'peaks over threshold'), which performs inference over all variables exceeding a certain high threshold. The deviations scores from the reference cohort can be used to fit a generalized extreme value (GEV) distribution, using a block maxima approach or a generalized Pareto distribution (GPD), using the peaks over threshold approach (Coles, S., Bawa, J., Trenner, L., & Dorazio, P., 2001). The combination of univariate extreme value statistics with normative models has been explored in previous papers (Fraza et al., 2021; Marquand et al., 2016a; Zabihi et al., 2019). However, few studies have tried to fit a peak over threshold model to neuroimaging data and, currently, no studies have tried to fully capture the extreme dependence structure for multivariate variables. To capture and model the tail-wise dependence structure of multivariate variables will be a great leap forward.

For the multivariate extreme values, this work employs a framework that is specifically created for exploring extreme tail dependence when the dimensions become large. A new

tool is explored that analyses the extremes in a similar framework to standard correlation or covariance principal components analysis (PCA). This is motivated by the fact that covariance PCA is not well-suited to modelling multivariate tail dependency, because dependence in terms of the mean of a multivariate distribution may not be equivalent to dependence in the tails. To address this, we employ the tail-wise dependence matrix (TPDM) (Cooley and Thibaud, 2019) to understand dependence structure in the extremes. The TPDM is similar to a covariance matrix, but instead of describing the dependence structure for the mean of the distribution, the TPDM captures the dependence structure in the joint tails. Afterwards, the TPDM can be used with an eigendecomposition which gives us an ordered basis of extreme eigenvectors and values that capture, in order, the most important vectors for explaining the extreme dependence structure (Cooley and Thibaud, 2019; Jiang et al., 2020). We call the method 'extreme PCA' and the associated principal components (PCs) 'extreme PCs'. Similar to covariance PCA the extreme PCs can be interpreted by looking at the individual contributions of each variable to the PCs. This gives an indication of which variables are most important in understanding the extreme deviations.
We will apply the method to the UK biobank. We take the UK biobank as one of the most comprehensive datasets to test new frameworks on. It is one of the largest single population datasets to date and emphasizes strong quality control to reduce the noise of different artefacts. In general, the heterogeneity in the population can only be fully captured if our datasets are large enough (Wolfers et al., 2015). We use the data imaging derived phenotype (IDP) data from over 40,000 participants, which are widely used summary measures of brain structure and function (Alfaro-Almagro et al., 2018). Finally, we aim to demonstrate the biological plausibility of our extreme PCs by associating them with several behavioural measures. We always want to establish the correlation between brain and behaviour such that our found components could be useful for understanding regular as well as more complex behavioural phenotypes.

2. Materials and methods

2.1. Sample

The data used in the paper came from the UK Biobank imaging data set (Miller et al., 2016; Sudlow et al., 2015). We made use of the image-derived phenotypes (IDPs), which are pre-processed imaging data generated by an image-processing pipeline developed and run on behalf of the UK Biobank (Alfaro-Almagro et al., 2018). For the full details on the pre-processing steps and the design of the study, the following document can be read UK Biobank Brain Imaging Documentation (Smith et al., 2020). We used data from 43.893 participants between 40 and 80 years old, with around 53% females. For the exact distribution of the age range and the number of participants currently in our dataset, see the supplement. For this study, we used image-derived phenotypes (IDPs). We pre-processed them using FUNPACK (McCarthy, 2020). This is an automatic normalization, parsing and cleaning kit, which was developed at the Wellcome Centre for Integrative Neuroimaging. The IDPs contain several neuroimaging measures, including structural, functional and diffusion MR. We used 2006 IDPs in total, see the supplement for the spread of imaging modalities used.

2.2 Behavioral data

The UK Biobank dataset provides an extensive number of behavioural, clinical, lifestyle and cognitive variables (Miller et al., 2016). These non-imaging phenotypes (nIDP) are categorized into eight different groups presented in Table 1. We used the nIDPs that were extracted from the UK Biobank dataset to evaluate the behavioural associations with the extreme deviations from the normative model.

Non-imaging variables categories
Basic demographic
Cognitive phenotypes
Lifestyle environmental alcohol
Lifestyle environment tobacco
Lifestyle environment general
Lifestyle environment exercise work
Mental health
Age, sex, site

Table 1 - UK Biobank non-imaging variables categories

We aimed to determine whether our found extreme PCs were also associated with more severe neurological problems in a categorical sense. Therefore, we also added the binary data from several neurodegenerative disorders: 'motor neuron disease', 'myasthenia gravis', 'multiple sclerosis', 'Parkinson's disease', 'dementia/Alzheimer/cognitive-impairment' and 'other demyelinating disease'. The IDP codes can be found here: https://biobank.ndph.ox.ac.uk/showcase/coding.cgi?id=6. In total, we had 238 subjects with any of these neurodegenerative disorders and brain imaging data presented and binarized these subjects under one column called 'Neurodegenerative disorders'.

2.3 Multivariate extreme normative modelling

In Figure 2, we present the full overview of all the steps in the multivariate extreme framework. We will briefly explain the framework here. For all the details, refer to the different method sections down below. First, we estimated a normative model for all the IDPs that explain the brain data using age, gender and site as our covariates. We applied a warped Bayesian Linear Regression (BLR) model, which helps us to predict the neuroimaging data from the covariates and also provides us with a predictive confidence interval for each prediction. Afterwards, we obtained the deviation or z-scores, which we combined into one large matrix for all the IDPs. We transformed the marginals, which will be explained in more detail below. Then, we thresholded the data in polar coordinates. Afterwards, the TPDM was calculated, based on the extreme quantiles of the data. On this matrix, we performed an eigenvalue decomposition, which gave us an eigenbasis that explains most of the extremal structure in the first few eigenvectors. Finally, we correlated the first few extreme PCs back to behaviour, to establish a brain-behaviour relationship.

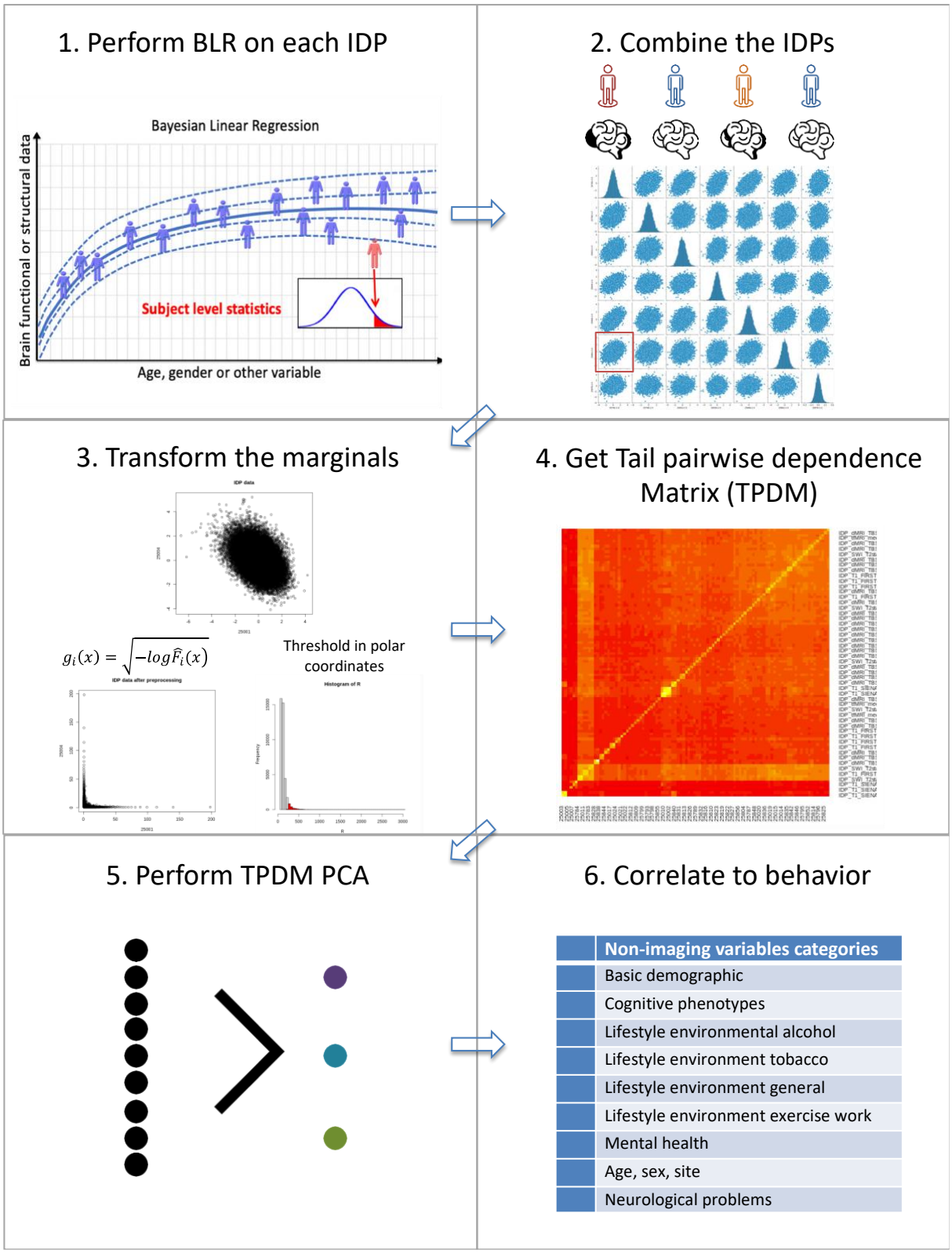


Figure 2 - Overview of the proposed extreme normative modelling approach showing all the steps in the pipeline. 1. Estimate a normative model with likelihood warped Bayesian Linear regression. This allows us to create deviation or z-scores from the mean trend for each participant. 2. Combine the z-scores of all the participants into the correct matrix. 3. Transform the marginals of the data such that the marginal distribution is varying with tail index = 2 and threshold the data in polar coordinates at the 95-quantile. *4. Calculate the tail-wise dependence matrix (TPDM) based on the outer 95-quantile of the z-scores. This matrix captures the extreme structure in the outer tails. 5. Perform an eigenvalue decomposition on the TPDM, creating an independent basis of which the first few eigenvectors contain most of the extremal structure information. 6. Correlate the first few extreme principal components to behaviour to indicate how multivariate extreme values influence behaviour. See the text for full details.*

2.3.1 Normative model formulation

We estimated all the normative models using python 3.8.3 and the PCNtoolkit version 0.20. We performed a normative model with the Bayesian linear regression model (BLR) and likelihood warping. For a full explanation and details of the mathematical framework behind this model see (Fraza et al., 2021). We used the covariates age, sex and site. Here we will give a short overview of the method. If we assume y a vector of participants, for each IDP:

$$y = w^T \phi(\boldsymbol{x}) + \epsilon_s$$

Where, w^{T} is the estimated vector of weights, $\phi(\mathbf{x})$ a basis expansion of the covariate vector **x**. In our case, a cubic B-spline basis expansion with 5 evenly spaced knots was chosen. Empirically, this was enough to capture the curvature in space caused by the age covariate. $\epsilon_s = \eta(0, \beta)$ is a Gaussian noise distribution for site s with mean zero and noise precision term β (the reciprocal of the variance). For the likelihood warping a Sinarcsinh function was employed:

$$\varphi_{Sinarcsinh(y,\gamma)} = \sinh\,(b * \operatorname{arcsinh}(y) - a)$$

With γ the found parameters for the warping function, in our case a and b. This method has shown to be able to model Gaussian as well as non-Gaussian distributions (Dinga et al., 2021; Fraza et al., 2021). We captured the site variation using a fixed-effects model, according to (Kia et al., 2021). We performed the optimization using Powell's conjugate direction method by minimizing the negative log-likelihood.
Afterwards, the z-scores were calculated in the warped space for each subject, *n*, and variable, *d*, in the test set as:

$$Z_{nd} = \frac{y_{nd} - \hat{y}_{nd}}{\sqrt{\sigma_d^2 + (\sigma_*^2)_d}} \qquad (2)$$

Where, $\mathrm{y_{nd}}$ is the true response, $\hat{\mathrm{y}}_{\mathrm{nd}}$ is the predicted mean, σ_{d}^2 is the estimated noise variance (reflecting uncertainty in the data), and $(\sigma_*^2)_{\mathrm{d}}$ is the variance attributed to modelling uncertainty (Marquand et al., 2016a). We evaluated the model fits according to several model criteria: explained variance (R^2), mean squared log-loss (MSLL), and skew and kurtosis. Together these criteria allowed us to assess the central tendency, variance and the performance of the warping function. Online several tutorials can be found for estimating a normative model and a large pre-trained model has also been curated, such that a new user can upload a smaller dataset and see how their subjects deviate from a reference cohort, see: https://github.com/predictive-clinical-neuroscience/braincharts (Rutherford et al., 2022). Afterwards, we used the z-scores to estimate the extreme value distributions.

2.3.2 Extreme value theory

Most of the univariate and multivariate extreme values statistics were performed in R version 4.1.0 as there are several dedicated packages for extreme value theory created in R, see https://cran.r-project.org/web/views/ExtremeValue.html. 'Ordinary' statistics focusses on the main part of the distribution. However, we are interested in the far upper tail or the extreme quantiles of a distribution and therefore have to apply extreme value statistics. Extreme value statistics allows us to model the tails accurately, see Figure 3.

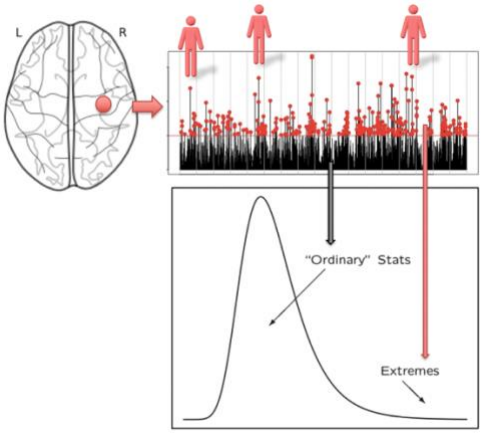


Figure 3 - Extreme value statistics vs. 'ordinary' statistics.

The extreme events in the tails have a small probability to occur, but usually exercise a large influence on their environment. In weather data, they are the main parts of the distribution that drive floods or devastating storms (National Academies of Sciences, 2016). In economics, the tails are mainly used to understand the stock market dependence structure and financial crashes (Poon et al., 2004). We suspect that the extremes in brain data will also exercise quite a large influence on our health and behaviour. However, the small occurrence of variables in the tails makes it hard to model them accurately. In the univariate case, two methods can be employed that we will explain briefly, i.e. 'block maxima' or 'exceedances' (a.k.a. 'peaks over threshold').
Under the block maxima approach, if we consider a set of independent variables X_n with common distribution function F. We want to focus on the maxima for each variable:

$$M_n = \max\,(X_1, \dots, X_n)$$

Usually X_n is measured with some kind of logical time frame, such as hours, months, years or a spatial frame such as street blocks and M_n then represents the maxima for each of those spatial/timeframes. We want to estimate the probability distribution function. It has been proven that the probability distribution of the rescaled sample maxima M_n converges to a generalized extreme value distribution (Fisher and Tippett, 1928; Gnedenko, 1943):

$$\Pr\{\frac{\mathrm{M_n} - b_n}{a_n}\} \to G(z), \quad n \to \infty$$

For constants $a_n > 0,\ b_n \in R$ and z the upper end of the tails.

$$G(z) = \exp\{-\left[1 + \xi\,\frac{(z-\mu)}{\sigma}\right]^{-\frac{1}{\xi}}\}$$

Which is defined by the shape ξ, location μ and scale σ parameters. The extreme value distributions can be further separated into Gumbel, Fréchet and Weibull families. For more details see (Coles, S., Bawa, J., Trenner, L., & Dorazio, P., 2001). This approach, although easy to employ, is not very efficient because it ignores large parts of the data that could be relevant (for example, the second- and third-highest values in each block which are potentially also informative about extreme behaviour).

An alternative method that takes this information into account is to estimate the probability distribution function is with the peaks over threshold approach:

$$M_n^k = kth\ largest\ of\ (X_1, \dots, X_n)$$

With k an appropriately large threshold. The probability distribution of this method converges to a generalized Pareto distribution:

$$H(y|X_n > k) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}}$$

which is defined by the shape ξ and scale σ parameters. For more details on the derivations see (Coles, S., Bawa, J., Trenner, L., & Dorazio, P., 2001).

In this method a threshold has to be decided, that is not too high as then there will be too few values to model the tails of the distribution accurately, but also not too low, as then the shape and scale parameters will be incorrect. One way to approach this problem is by using graphical tools with several thresholds and deciding on the best threshold based on their outcomes. In the result section, we take you through one IDP example for deciding the optimal threshold and modelling the shape and scale parameters using the peaks-over threshold method.

2.3.3 Multivariate extreme value theory

With univariate extreme value statistics, we can model the tails at one location. However, in many circumstances, we are interested in multiple variables simultaneously, which calls for a multivariate approach. For example, there is a strong tail dependence between certain locations in the brain or between different imaging phenotypes derived from the data. In multivariate extremes, our objective is to characterize and understand the dependence structure in the far upper joint tail, see Figure 4.

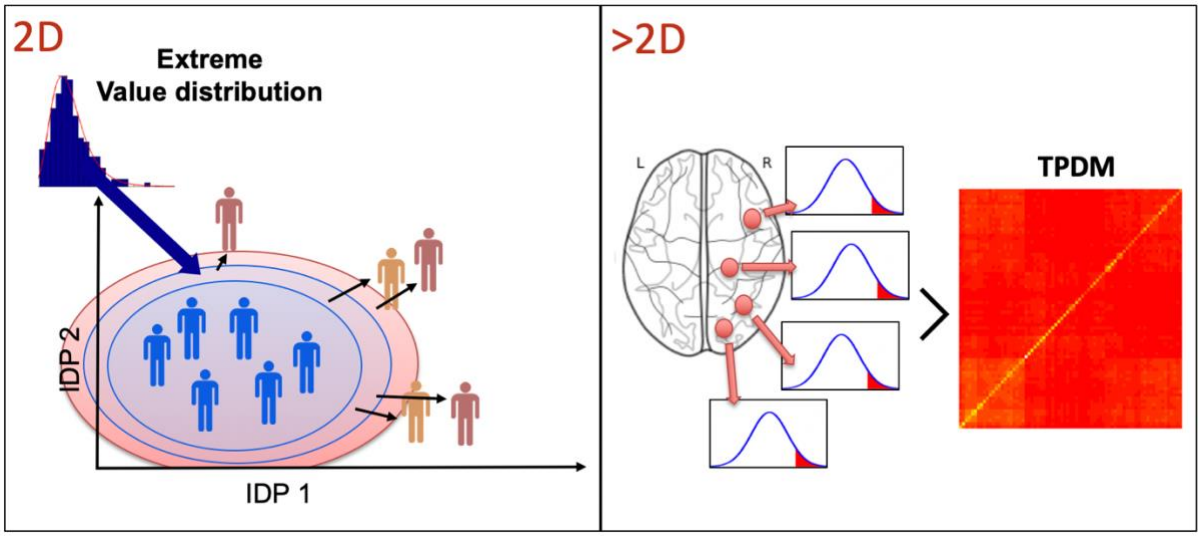


Figure 4 - Overview of the multivariate extreme value method. For two dimensions, bivariate extreme value theory can be used. For >2D variables, we need to summarize the tail dependence with a so-called tail pairwise dependence matrix (TPDM).

In the bivariate case, we can fit a bivariate extreme value distribution, which tries to estimate the probability of the variable being large whilst the other is large. We again have

to decide on an appropriate threshold similar to the univariate case (Coles, S., Bawa, J., Trenner, L., & Dorazio, P., 2001).
In this paper, we want to look at multivariate extremes beyond $D = 2$. There are several methods to model the joint tails, for a full overview see (Engelke and Ivanovs, 2020). Briefly, the first method tries to identify variables that are large together and find concomitant extremes. Identifying these extremes is quite hard and several methods have been proposed (Goix et al., 2017), (Chiapino et al., 2018). The second method relies on structure learning and uses tree-like graphs to find the underlying graph structure of the extreme values. This graphical extreme value method can also be used for causal inference (Engelke and Ivanovs, 2020), (Gnecco et al., 2020). However, we do not naturally expect the brain data to follow a tree-like structure and even if that is true, inferring the correct structure can be challenging. Lastly, methods based on unsupervised learning can be used, such as PCA and clustering analysis. In our case, the multivariate extreme value method is based on PCA (Cooley and Thibaud, 2019). The tail-wise dependence is summarized in a matrix on which an eigendecomposition can be performed, giving us a lower-dimensional linear subspace that accounts for most of the extremal dependence. Full details of the method will be explained below. For a deeper understanding of extreme value theory, the following textbooks can be considered:

- An Introduction to Statistical Modeling of Extreme Values (Coles, S., Bawa, J., Trenner, L., & Dorazio, P., 2001)
- Statistics of Extremes: Theory and Applications (Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. L., 2004)
- Extreme value theory: an introduction (De Haan, L., Ferreira, A., & Ferreira, A., 2006)
- Heavy-tail phenomena: probabilistic and statistical modelling (Resnick, S. I., 2007)

2.3.4 Framework for the tail pairwise dependence structure

We will describe the framework used in this paper for understanding the tail pairwise dependence structure briefly. For a more in-depth mathematical understanding see (Cooley and Thibaud, 2019; Jiang et al., 2020).
The method is based on the framework of regular variation. A random vector is multivariate regularly varying if its joint tail decays like a power function (Resnick, S. I., 2007). We take $\boldsymbol{X} = (X_1, \dots, X_d)$ as a regularly varying random vector for all d IDPs. In general, for multivariate regular variation it is convenient to consider the variables in polar coordinates:

$$\boldsymbol{X} \rightarrow \left(\|\boldsymbol{X}\|, \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} \right) = (\boldsymbol{r}, \boldsymbol{w})$$

With $w \epsilon N$ the direction of the vector and r stating the length of the vector (Resnick, S. I., 2007). Then for a given A, a set of large values away from the origin, we can calculate the probability of $\boldsymbol{X}$ as:

$$P(\boldsymbol{X} \in A) \approx \int_{(r,\boldsymbol{w}) \epsilon A} \alpha r^{-\alpha-1} dr dH(\boldsymbol{w})$$

With $\approx$ signifying "approximately proportional to", $\alpha > 0$, r the magnitude or radial component of the location, $\boldsymbol{w}$ the location on the unit sphere S, and H the angular measure on the unit sphere S. In the integrant, r follows a power-law distribution according to α the reciprocal of the shape parameter of the GEV (Coles, S., Bawa, J., Trenner, L., & Dorazio, P., 2001), (Jiang et al., 2020).

2.3.5 Tail pairwise dependence matrix TPDM

The angular measure H in lower dimensions can be modelled parametrically. However, as the dimension size increases there is not sufficient information in the extremes to capture H. Instead of fully modelling the dependence structure contained in H an approximation can be made with the tail pairwise dependency matrix or TPDM. If we assume $\boldsymbol{X}$ to be a d-dimensional regularly varying random vector with tail index $\alpha = 2$ and angular measure H, with d the number of IDPs. Then each element in the TPDM Σ_X is specified according to:

$$\sigma_{i,j} = \int_S w_i w_j dH(w), \quad i,j = 1, \dots, d$$

The TPDM has similar interpretation properties as a standard covariance matrix only specified for the extremes (Cooley and Thibaud, 2019). Thus, if two variables are independent in the outer tails $\sigma_{i,j} = 0$ and the diagonal element $\sigma_{i,i} = b^2$, with b the scale. The TPDM is positive definite and symmetric and therefore the eigenvalues are positive and the eigenvalues real (Jiang et al., 2020).

2.3.6 Estimation of the TPDM

We want to estimate the dependence in the tails via the TPDM. For this framework, the variables should be heavy-tailed with a common tail index $\alpha = 2$. To ensure this we first have to transform the marginals. Transforming the marginals is a common procedure in extreme value statistics (Resnick, S. I., 2007). We first use a rank transformation to obtain F, the estimated marginal cumulative distribution function (CDF) from the data, but other transformations can be considered, see (Resnick, S. I., 2007). Then we transform our F with $\mathrm{G}(x) = \sqrt{-logF(x)}$ the CDF of a Frechet random variable with $\alpha = 2$ (Cooley and Thibaud, 2019). Afterwards, a generalized polar coordinate transformation is applied and a threshold of the 95th quantile is chosen for the kth largest radius. Finally, we can estimate the TPDM according to:

$$\hat{\sigma}_{Xij} = 2n_{exc}^{-1} \sum_{n=1}^{n} w_{ni} w_{nj} \mathbb{I}\,(r_n > r_0)$$

Where r_0 the threshold for the radial component and n_{exc} the number of observations that pass the threshold.

2.3.7 Choosing a threshold

Choosing the optimal threshold for the TPDM estimation includes all the usual difficulties with selecting a threshold. Commonly, it is selected with a visual inspection of all the graphs presented in the univariate result section. However, when the dimensions become large this becomes infeasible. Therefore, as suggested in (Jiang et al., 2020) we selected a couple of IDPs and manually went through a few thresholds. We decided the 0.95 quantile was a good common threshold above which the examined parameters seemed relatively constant.

2.3.8 TPDM PCA

In (Cooley and Thibaud, 2019) it has been shown that an eigendecomposition can be applied to the TPDM yielding an orthonormal basis:

$$\widehat{U}\widehat{D}\widehat{U}^T = \widehat{\Sigma}_X$$

With D the diagonal matrix of eigenvalues $(\lambda_1, \dots, \lambda_d)$ and $U = (u_1, \dots, u_d)$ the matrix of eigenvectors. We can select the top eigenvectors by ordering the eigenvalues that explain the largest scale, which is similar to ordering normal eigenvalues which explain the most variance. Likewise, as with covariance or correlation PCA, the top extreme PCs can then be visualized and interpreted. Finally, we correlate the extreme PCs with behaviour to investigate how the extreme PCs map back to the behavioural phenotype.

3. Results

We split the available IDP data into a training set and a test set with a 10-90 split, giving us 4405 participants for training and 39488 participants for testing. We want a larger amount of testing subjects to implement the new multivariate extreme value framework on, as the UK Biobank dataset is very large the number of training subjects is still acceptable, but with a smaller dataset, we would recommend a 50-50 split. Afterwards, we performed a simple quality control, where we removed participants with absolute z-scores > 7. We fitted a normative BLR model with a Sinarcsinh likelihood warping and Powell's conjugate direction method. For the covariates, we used age, sex and fixed effects for the sites to predict the IDP values. We created a B-spline basis transformation for the train and test data with 5-knots. We can consider different types of B-spline transformations or basis transformations that could also generate equivalent performance (e.g. fractional polynomials). However, the B-spline has proven sufficient to model the types of data that we expect to see in neuroimaging, (ie. relatively slow varying trends) and performs better than vanilla polynomials, which induce a global curvature (Fraza et al., 2021). In Figure 5, we present an example BLR fit for one IDP showing the mean and centiles of variation.

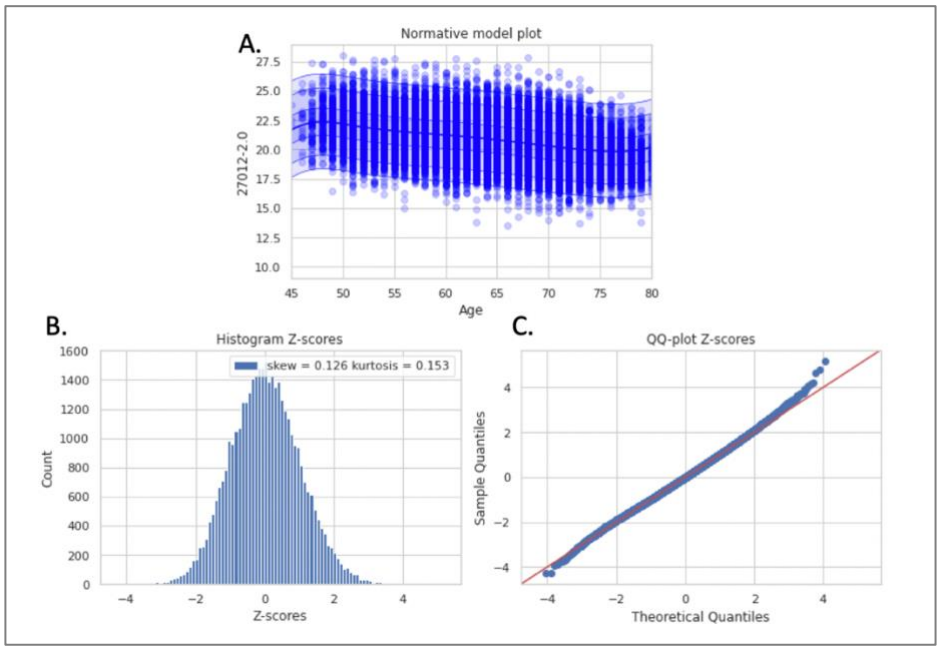


Figure 5 - IDP 27012 modelled as a function of age using a Bayesian Linear Regression (BLR) model. A. Shows the normative model fit with the centiles of variation. B. shows the histogram of z-scores. C. Shows the quantile-quantile (QQ) plot of the z-scores.

We evaluated the performance of the likelihood warping by looking at the explained variance, MSLL, kurtosis and skewness, see Figure 6. If fitted properly the skewness and kurtosis should be around zero. We removed 42 IDPs that had a kurtosis above 10 showing an extremely bad fit. We could optimize the model to fit the IDPs even better (e.g. by evaluating different warping functions or parameterisations), but since our objective for this paper was to develop a new method based on extreme value statistics this would increase the complexity of the analytical pipeline, effectively yielding different base models for different IDPs, therefore we do not pursue this here. Afterwards, we collected the deviation or z-scores for each participant and all the IDPs into one z-scores matrix.

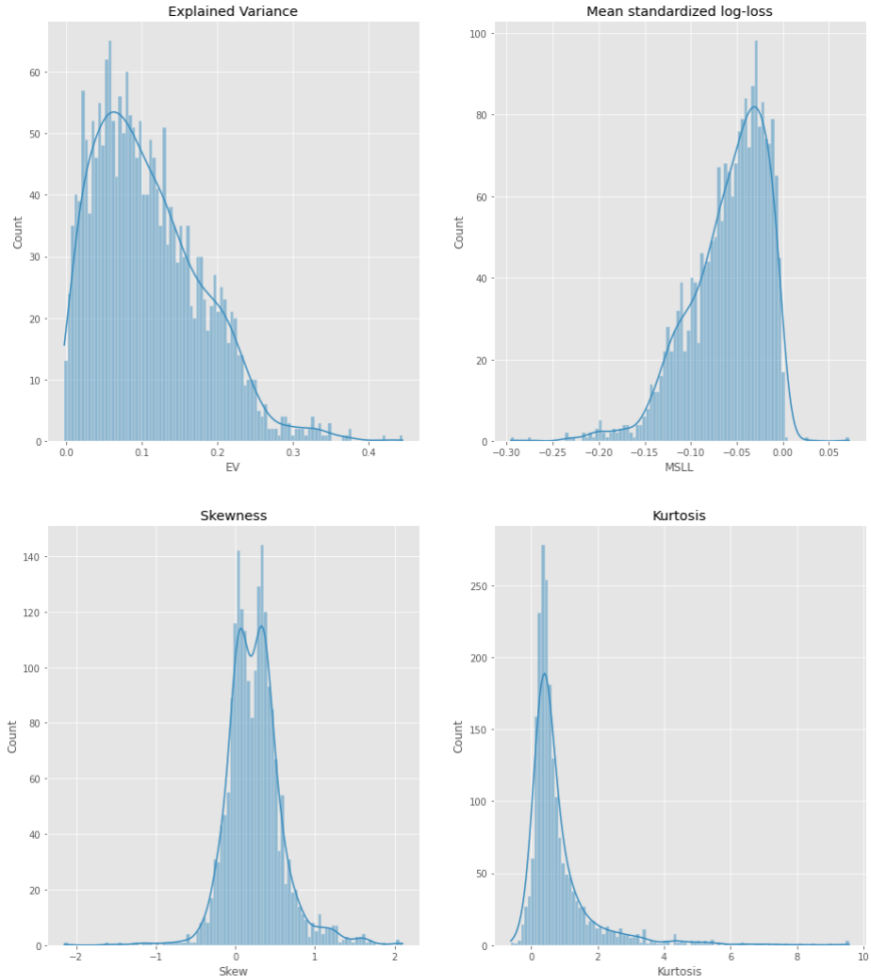


Figure 6 - showing the performance measures for the different IDPs using a warped BLR model.

3.1 Univariate extreme value theory results

In this section, we will take you through the peak over threshold method for one IDP and show how an optimal threshold can be decided and the tails accurately modelled. In Figure

7, we show the different possible thresholds for one IDP, white matter hyperintensity (WMH). To decide between the different thresholds several graphical tools can be used.

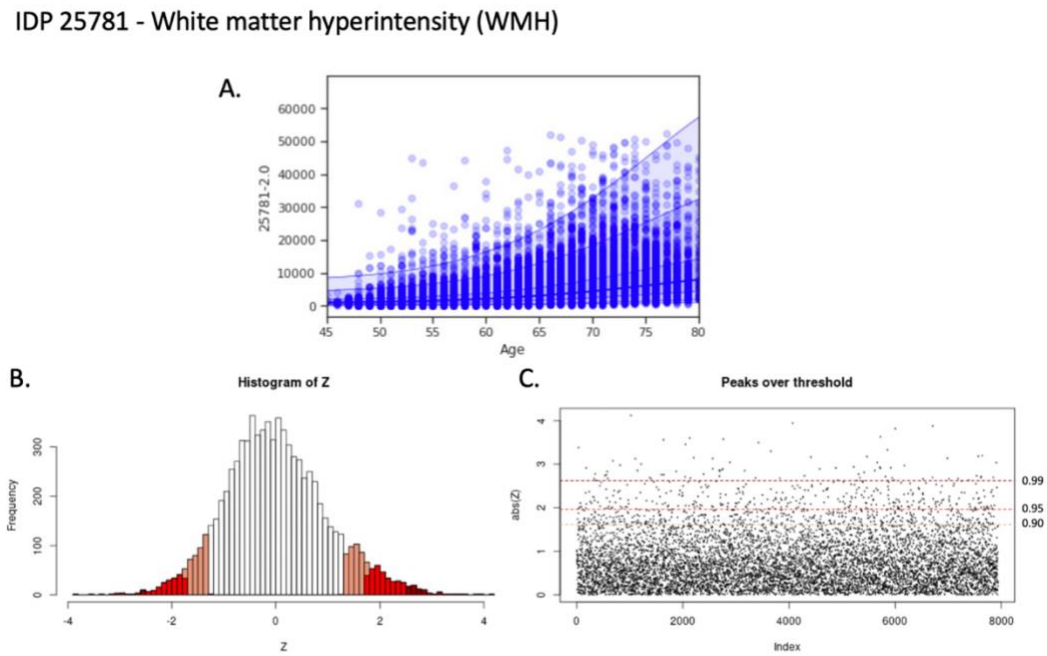


Figure 7 Showing the univariate peak over threshold method for one IDP. A. Warped BLR normative model result for the white matter hyperintensity IDP, demonstrating the mean and centiles of variation. B. Histogram of the z-scores showing different thresholds that could be employed in the peaks over threshold method. Scatterplot for the peaks over threshold method.

One graphical tool that can be employed is the mean residual life plot (MRL-plot) (Davison and Smith, 1990), which involves plotting the value of a threshold, u, against the 'mean excess', i.e. the exceedances at u, minus u. In Figure 8, we present the MRL plot for IDP WMH. To help interpret this plot, it is assumed that the threshold should be placed where the plot starts diverging from the linear trend. In this case, we can observe that the plot is relatively linear until the 0.95 mark

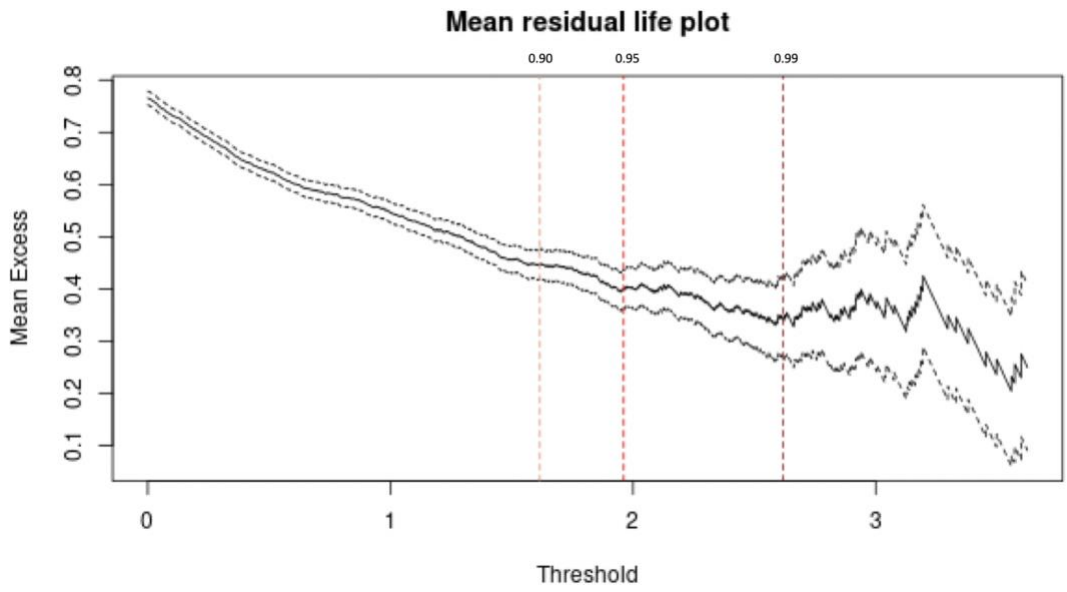


Figure 8 – shows the mean residual life plot of the WMH data. We can observe that the plot is relatively linear up to the 0.95 threshold. Indicating that this could be a good threshold.

Another graphical tool that can be employed is the parameter or threshold stability plot (Scarrott and MacDonald, 2012), which involves plotting model parameters as a function of the threshold. We look at different estimates of the parameters for the GDP and decide on an optimal threshold based on the lowest value where the plots are approximately stable. In Figure 9, we show the parameter threshold stability plot of the modified scale and shape parameters. The dots demonstrate the scale and shape parameters for the GPD at different thresholds. The lines show the 95% confidence interval. The plot was created using the tcplot function from the evd package in R. We can see that the threshold should be around the 0.9 to 0.95 mark.

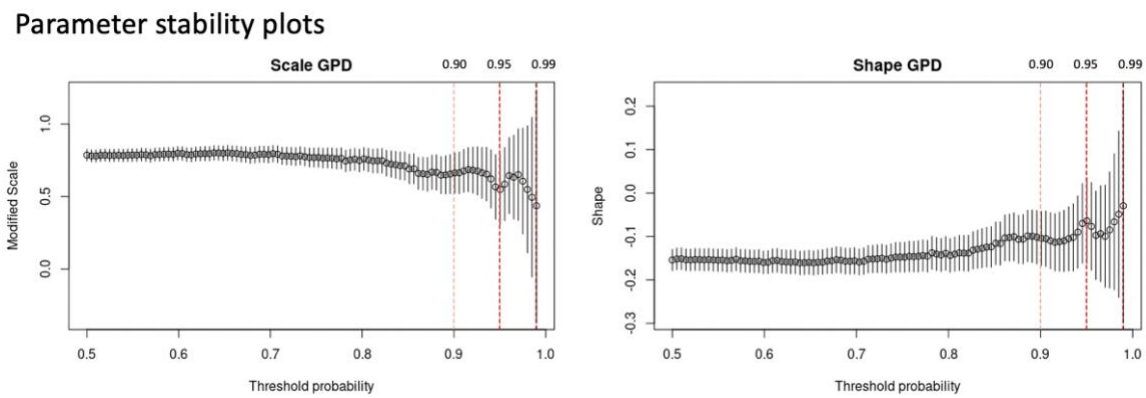


Figure 9 - Showing the parameter threshold stability plot of the modified scale and shape parameters. The dots demonstrate the scale and shape parameters for the generalized Pareto distribution (GPD) at different thresholds. The lines show the 95% confidence interval.

It is evident that it is difficult to get a clear cut-off threshold even using these graphical tools and there are several more methods to decide on an optimal threshold, see (Scarrott and MacDonald, 2012). Furthermore, sensitivity analyses can be employed to give confidence that the derived findings are not crucially dependent on the choice of threshold. For now,

we will set the threshold at 0.95 quantile as we estimate this is low enough to give us an adequate amount of data and high enough to estimate the GPD parameters accurately. Once the threshold is decided, we can estimate the parameters using a maximum likelihood optimization. We employed the fpot function in the evd package In R. In Table 2 the GPD parameter and standard error estimates are shown.

Threshold:	**1.9613**
Exceedances	397
Scale estimate	0.42516
Standard error scale estimate	0.03009
Shape estimate	-0.06503
Standard error shape estimate	0.04999

Table 2 - The scale and shape parameter estimates of the Generalized Pareto distribution (GPD), estimated using a maximum likelihood optimization with the fpot function in the evd package in R.

The shape estimate is <0 for the GPD which indicates a Weibull distribution.
In Figure 10, we plotted the profile likelihood from which the scale and shape parameters are decided. The more peaked the log-likelihood is the higher our confidence in our estimates. We also show the QQ-plot and return level plot, yielding straight plots. The straight line indicates that the fitted function is appropriate for the data (Resnick, S. I., 2007).

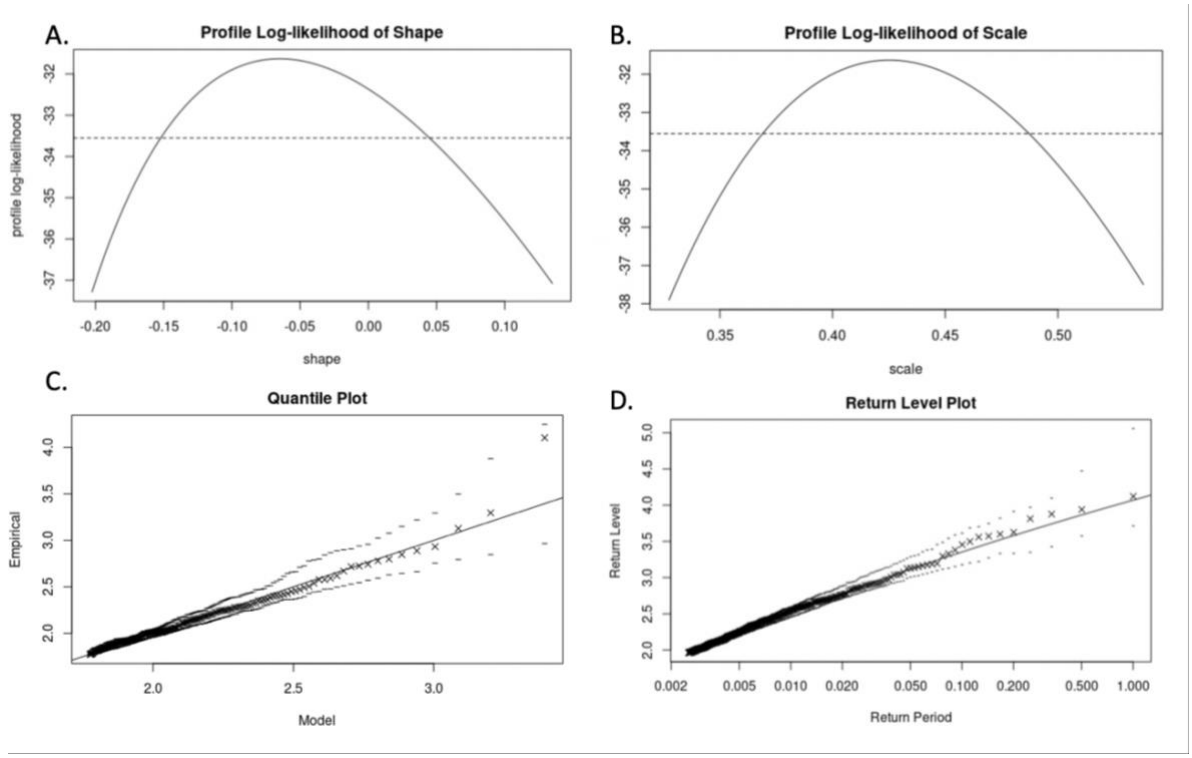


Figure 10 - A. and B. Demonstrating the profile likelihood from which the scale and shape parameters are decided. We can see the estimations are set on the maxima. The more peaked the profile likelihood the higher our confidence in the estimations. In C. and D., we can see the quantile plot and return level plot showing a relatively good fit.

3.2 Multivariate extreme value theory results

In this section, we want to estimate the multivariate tail dependence in the deviations or z-scores. We estimated the TPDM and applied an eigendecomposition to try to understand the extremal dependence.

First, to apply the method proposed by (Cooley and Thibaud, 2019), we have to transform the marginal distribution for each IDP, such that it fits the framework that requires $\boldsymbol{X}$ to be regularly varying with α=2. We estimate the marginal $\widehat{F_d}$ at IDP d using a rank transform. Then, we apply the nonparametric marginal transformation proposed by (Cooley and Thibaud, 2019):

$$G_d(x) = \sqrt{-log\widehat{F_d}(x)}$$

With this transformation, we ensure that the scale of X_d is 1 and α=2. Afterwards, we transform the variables into polar coordinates (r, w). Then we select $r_0 = 247$ or the 0.95 quantile threshold, giving $n_{exc} = 1960$. This threshold was selected after looking at several univariate diagnostic plots as described in section 3.2. In Figure 11, we show the data from two IDPs in the original scale and after the marginal transformations have been applied. Furthermore, we show the histogram of r exceeding the 0.95 quantile threshold. It can be noted that the variables fully lie on their respective axis after the transformation.

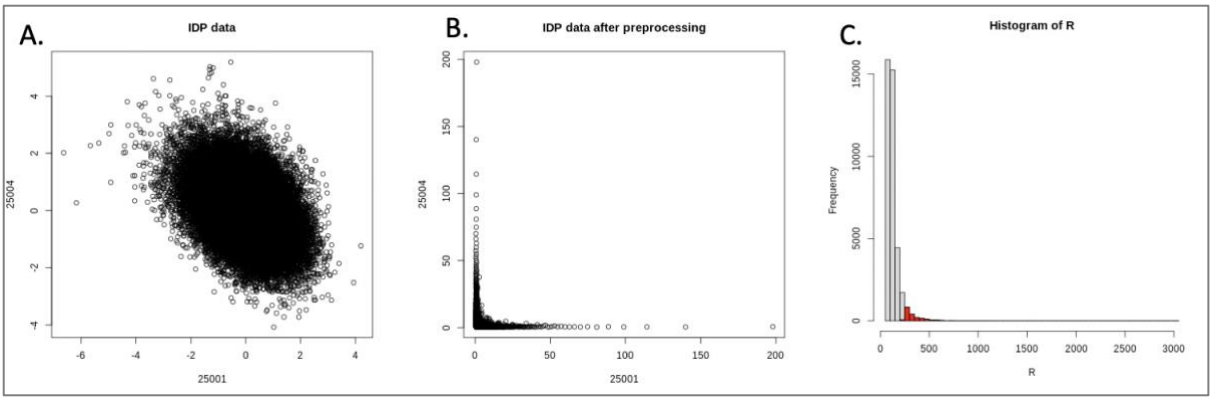
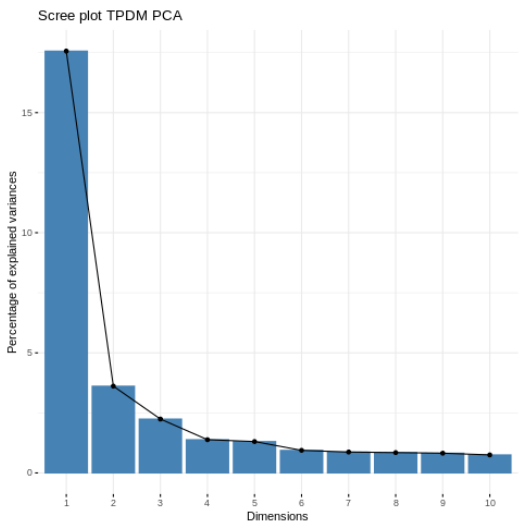


Figure 11 - Illustration of the applied marginal transformation to the IDP data. A. Scatterplot of the IDP data in its original dimensions. B. Scatterplot of the IDP data after transforming the marginal, such that the marginal is regularly varying with α=2. C. Histogram of R and the 0.95 largest components.

The transformed z-score data is then used to estimate the TPDM $\hat{\Sigma}$ according to section 2.5.1. In Figure 19 in the appendix, we give an example of the found TPDM separated per IDP modality. It can be noted that the matrixes, in general, are quite sparse, which is to be expected for extremal dependence and that every modality does show several IDPs with scale dependence in the extremes. In the next section, the extremal dependence will be further explored using an eigenvalue decomposition.

3.2.1 Interpretation of extreme principal components

We performed a standard eigendecomposition of $\hat{\Sigma}_X$ to obtain the eigenvectors (i.e. extreme PCs). Similar to correlation-PCA, we can look at the amount of scale that the first number of eigenvalues explain. In Figure 12, we plotted the scree plot of the scale, which shows that most of the scale is explained by the first two eigenvectors.



	eigenvalue	scale.percent	cumulative.scale.percent
Dim.1	359.49	17.56	17.56
Dim.2	73.95	3.61	21.17
Dim.3	45.93	2.24	23.42
Dim.4	28.34	1.38	24.8
Dim.5	26.78	1.31	26.11
Dim.6	19.26	0.94	27.05
Dim.7	17.82	0.87	27.92
Dim.8	17.36	0.85	28.77
Dim.9	16.84	0.82	29.59
Dim.10	15.4	0.75	30.35
Dim.11	12.85	0.63	30.97
Dim.12	11.85	0.58	31.55
Dim.13	10.99	0.54	32.09
Dim.14	10.5	0.51	32.6
Dim.15	10.17	0.5	33.1
Dim.16	9.65	0.47	33.57
Dim.17	9.48	0.46	34.03
Dim.18	9.31	0.45	34.49
Dim.19	8.88	0.43	34.92
Dim.20	8.78	0.43	35.35

Figure 12 - Showing the scree plot and table of scale values of the eigenvalue decomposition for the TPDM matrix.

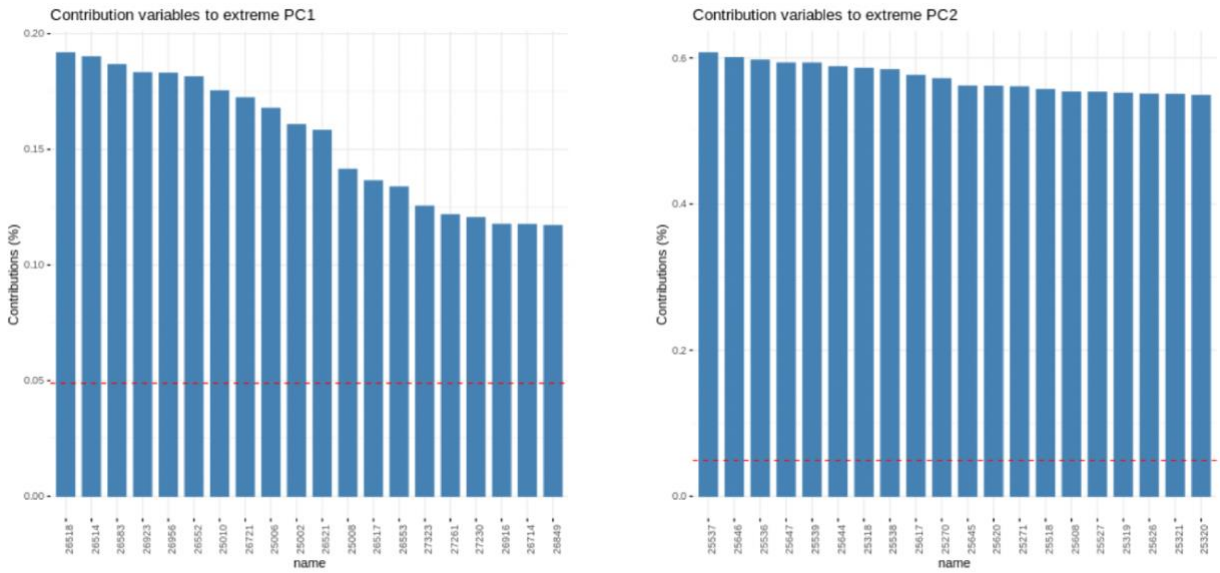
Two of the goals of standard PCA are to extract the most important information and to analyse the structure of this information in terms of the variables (Abdi and Williams, 2010). We selected the first two extreme PCs, as our informative basis. Afterwards, we try to understand this basis in terms of its original variables by estimating the individual contributions. We define the loadings for each component l:

$$f_l = u_l * \sqrt{\lambda_l}$$

With f the loading, u the eigenvector and λ the eigenvalue. The contribution of each IDP d to the component l is then calculated by:

$$ctr_{d,l} = \frac{f_{d,l}^2}{\lambda_l}$$

With λ_l the eigenvalue of the l-th component (Abdi and Williams, 2010). The higher the contribution value the more that particular IDP contributes to the component.
In Figure 13, we plotted the highest contributions of several IDPs to the first two extreme PCs. A line is added to denote the average contribution $(1/l)$. We can see that the first extreme PC is dominated by regional and tissue volume IDPs and cortical area IDPs and the second extreme PC is dominated by white matter tract diffusivity IDPs, showing that in the extreme tails these variables vary the most together and explain the largest amount of scale for all the IDP data.



	Row.names	IDP.short.name	Category.name	Category	contrib.1.top.
7	26518	aseg_global_volume_TotalGray	regional and tissue volume	0	0.19
5	26514	aseg_global_volume_BrainSeg	regional and tissue volume	0	0.19
11	26583	aseg_rh_volume_Cortex	regional and tissue volume	0	0.19
16	26923	aparc-pial_lh_area_TotalSurface	cortical area	1	0.18
17	26956	aparc-pial_rh_area_TotalSurface	cortical area	1	0.18
9	26552	aseg_lh_volume_Cortex	regional and tissue volume	0	0.18
4	25010	IDP_T1_SIENAX_brain-unnormalised_volume	regional and tissue volume	0	0.18
13	26721	aparc-Desikan_lh_area_TotalSurface	cortical area	1	0.17
2	25006	IDP_T1_SIENAX_grey_unnormalised_volume	regional and tissue volume	0	0.17
1	25002	IDP_T1_SIENAX_peripheral_grey_unnormalised_volume	regional and tissue volume	0	0.16
8	26521	aseg_global_volume_EstimatedTotalIntraCranial	regional and tissue volume	0	0.16
3	25008	IDP_T1_SIENAX_white_unnormalised_volume	regional and tissue volume	0	0.14
6	26517	aseg_global_volume_SubCortGray	regional and tissue volume	0	0.14
10	26553	aseg_lh_volume_CerebralWhiteMatter	regional and tissue volume	0	0.13
20	27323	aparc-DKTatlas_rh_volume_superiorfrontal	regional and tissue volume	0	0.13
19	27261	aparc-DKTatlas_rh_area_superiorfrontal	cortical area	1	0.12
18	27230	aparc-DKTatlas_lh_volume_superiorfrontal	regional and tissue volume	0	0.12
15	26916	aparc-Desikan_rh_volume_superiorfrontal	regional and tissue volume	0	0.12
12	26714	ThalamNuclei_lh_volume_Whole-thalamus	regional and tissue volume	0	0.12
14	26849	aparc-Desikan_rh_area_superiorfrontal	cortical area	1	0.12

	Row.names	IDP.short.name	Category.name	Category	contrib.1.top.
10	25537	IDP_dMRI_ProbtrackX_MD_slf_r	WM tract diffusivity	9	0.61
19	25646	IDP_dMRI_ProbtrackX_L3_str_l	WM tract diffusivity	9	0.60
9	25536	IDP_dMRI_ProbtrackX_MD_slf_l	WM tract diffusivity	9	0.60
20	25647	IDP_dMRI_ProbtrackX_L3_str_r	WM tract diffusivity	9	0.59
12	25539	IDP_dMRI_ProbtrackX_MD_str_r	WM tract diffusivity	9	0.59
17	25644	IDP_dMRI_ProbtrackX_L3_slf_l	WM tract diffusivity	9	0.59
3	25318	IDP_dMRI_TBSS_L3_Anterior_corona_radiata_R	WM tract diffusivity	9	0.59
11	25538	IDP_dMRI_ProbtrackX_MD_str_l	WM tract diffusivity	9	0.58
14	25617	IDP_dMRI_ProbtrackX_L2_slf_l	WM tract diffusivity	9	0.58
1	25270	IDP_dMRI_TBSS_L2_Anterior_corona_radiata_R	WM tract diffusivity	9	0.57
18	25645	IDP_dMRI_ProbtrackX_L3_slf_r	WM tract diffusivity	9	0.56
15	25620	IDP_dMRI_ProbtrackX_L2_str_r	WM tract diffusivity	9	0.56
2	25271	IDP_dMRI_TBSS_L2_Anterior_corona_radiata_L	WM tract diffusivity	9	0.56
7	25518	IDP_dMRI_ProbtrackX_MD_atr_r	WM tract diffusivity	9	0.56
13	25608	IDP_dMRI_ProbtrackX_L2_ifo_l	WM tract diffusivity	9	0.55
8	25527	IDP_dMRI_ProbtrackX_MD_ifo_l	WM tract diffusivity	9	0.55
4	25319	IDP_dMRI_TBSS_L3_Anterior_corona_radiata_L	WM tract diffusivity	9	0.55
16	25626	IDP_dMRI_ProbtrackX_L3_atr_r	WM tract diffusivity	9	0.55
6	25321	IDP_dMRI_TBSS_L3_Superior_corona_radiata_L	WM tract diffusivity	9	0.55
5	25320	IDP_dMRI_TBSS_L3_Superior_corona_radiata_R	WM tract diffusivity	9	0.55

Figure 13 - List of the top 20 contributions of the different IDPs to the first and second extreme PCs.

3.3 Extreme PCs correlation with behaviour

We wanted to correlate the extreme PCs to behavioural variables to ensure that the model can be used for behavioural predictions. In Figure 14 and Figure 15, we present the Manhattan plots of the p-values for the univariate Spearman correlation between the nIDPs and the first two extreme PCs. In general, we detected 43 significant Bonferroni corrected p-values for the first PC, 130 for the second PC and 35 for the third PC. We see that the extreme value space links quite strongly to behaviour. Furthermore, we expect that this will be even more apparent for extreme behaviour, for example, present during a psychosis. In the future, we consider that this method will be extremely sensitive to detecting brain-behaviour mappings for psychiatric disorders and we are planning to apply the extreme value framework in a clinical application.

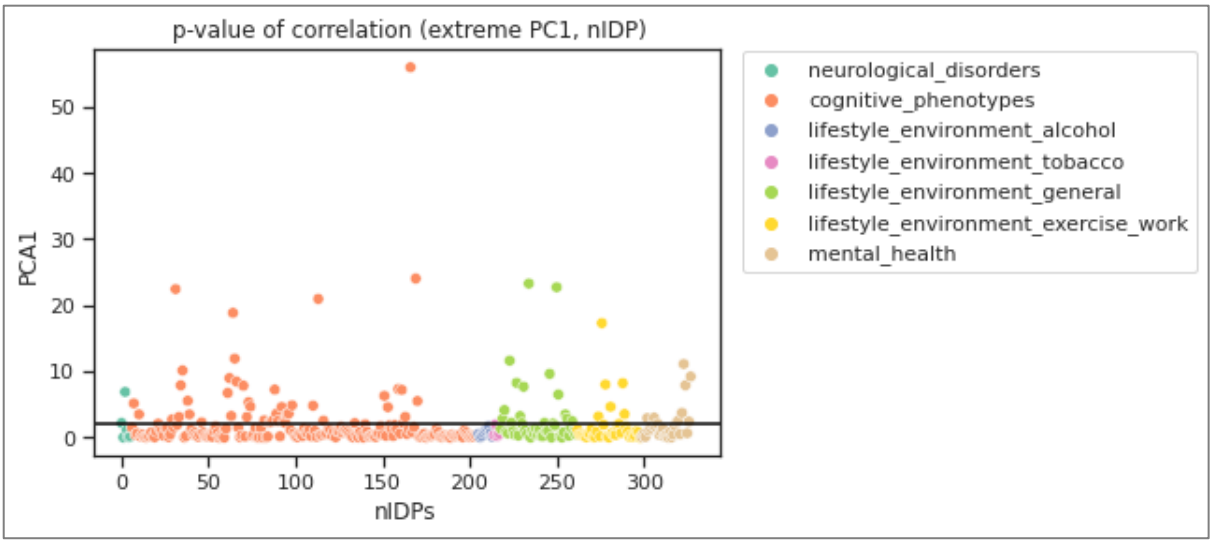


Figure 14 - shows a Manhattan plot of the p-values for the Spearman correlation between the nIDPs and the first extreme PC. The black line demonstrates the Bonferroni-corrected p-value threshold. Thus, values passing this line indicate a significant PC-nIDP correlation.

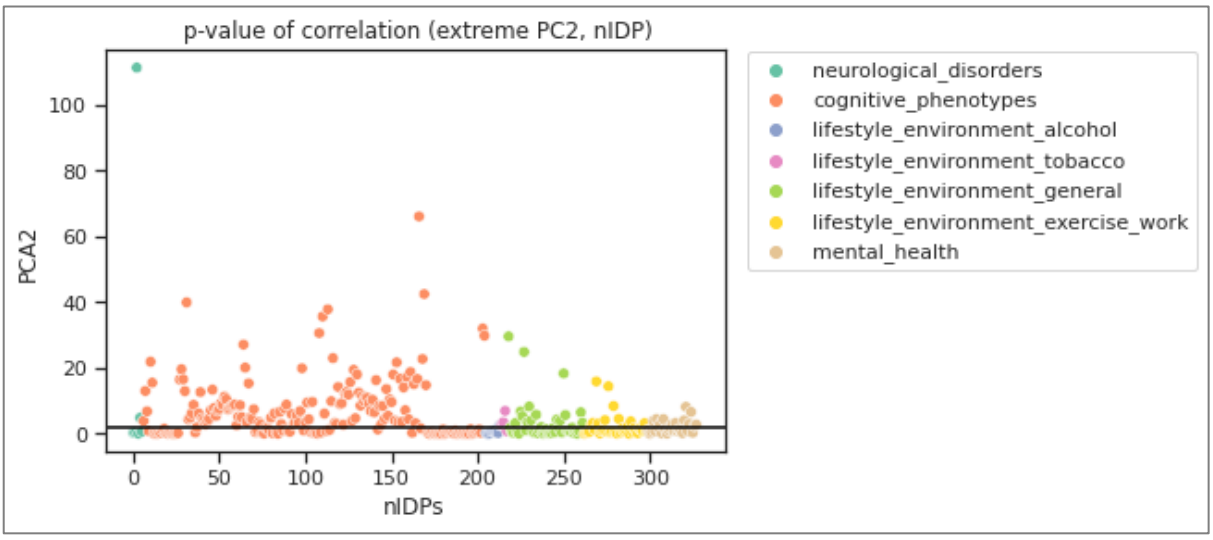


Figure 15 - shows a Manhattan plot of the p-values for the Spearman correlation between the nIDPs and the second extreme PC. The black line demonstrates the Bonferroni-corrected p-value threshold. Thus, values passing this line indicate a significant PC-nIDP correlation.

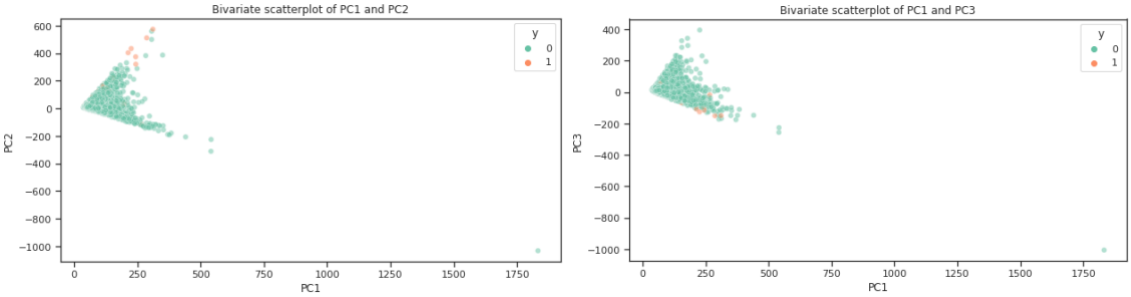


Figure 16 - Scatterplots of pairs of the first three extreme principal components, with the colours indicating neurodegenerative disorders for those particular participants.

Discussion

General

In this paper, we present one of the first neuroimaging studies that tries to characterize multivariate extreme tail behaviour of brain imaging data. We aim to explore the structure that can be found in the tails and view this through an eigendecomposition that can provide a good basis to further understand the tail dependence.

In general, normative models are used to capture and understand the heterogeneity in the population by modelling individual deviations from a reference cohort. In this study, we extended the normative method and presented a new method for exploring the extreme tail behaviour by summarizing it into a matrix of pairwise extremal dependence. We showed that: (i) the extremal dependence structure in the tail can be captured by a tail pairwise dependence matrix (TPDM); (ii) that the scale of the TPDM can be summarized by the first two extreme principal components; (iii) that the tail dependency in the first two extreme principal components was mostly explained by whole-brain structure or white matter tracts; (iv) that a strong link exists between extreme principal components and behavioural factors; and (iv) we observed a correlation between neurodegenerative disorders and the extreme PCs. Furthermore, we demonstrated how to implement the multivariate extreme value analysis on a large dataset from the UK Biobank, using the image-derived phenotypes (IDPs). The extreme PCA method presented in this paper could offer an alternative to methods mostly focusing on mean deviation scores. We think that concatenated multivariate extremes in brain data have a large effect on behaviour. Therefore, the extremes should be mapped more carefully if we want to understand complex human behaviour in, for example, psychiatric disorders.

Univariate extreme value methods

First, we set up a study for univariate extreme value analysis. We showed the framework for the peaks-over-threshold method for one IDP and analysed different methods for setting a threshold. We discussed several diagnostic tools and showed that a suitable threshold could be established for the univariate extremes. In general, selecting a threshold is quite a delicate task as we try to balance the bias from the sparsity in the data and the asymptotic tail approximation. Furthermore, one can argue that using graphical tools is subjective to the eye of the perceiver. However, as discussed in (Scarrott and MacDonald, 2012) graphical tools are still a preferred method for selecting the optimal threshold. Afterwards, we fitted a GPD and estimated the shape and scale parameters with a likelihood optimization.

Multivariate extreme value methods

Second, we wanted to understand the extremal dependence between the z-scores of the different IDPs and summarized this using the TPDM. There have been different attempts to understand and measure the tail dependence between different variables (Engelke and Ivanovs, 2020), but we thought the TPDM method was most appropriate for neuroimaging data. The TPDM has been proven to be similar to a covariance matrix (Cooley and Thibaud, 2019). Covariance matrixes have long been used in neuroimaging data to understand the underlying relationships between different brain areas and brain modalities. However, the covariance matrix is specified for the mean which might not correspond with tail dependence. Therefore, using the TPDM is a natural choice as it is similar to the covariance matrix, but specified for the extremes. Furthermore, just like with covariance PCA an eigendecomposition of the TPDM can be applied. This allows us to create an eigenbasis from which we can begin to interpret the most important variables quantifying the dependence structure.

Understanding the extreme PCs and applying it to UKbiobank

Here, we have employed an exploratory approach in this study to validate our novel method. We decided to apply it to a bigger cohort dataset such as the UK biobank and included the IDP data as our primary resource, as it has all the brain imaging modalities included and several behavioural phenotypes. We found evidence for extreme tail dependence between several IDPs. Afterwards, the extreme PCs showed that the tail dependence can mostly be explained by several imaging modalities, which is to be expected. The IDPs that explained the highest scale were related to regional and tissue volume and white matter tracts, showing that the structural brain measures have a high tail dependency.

Summary

We presented a clear framework for understanding and visualizing the dependency structure in the tails of different neuroimaging variable distributions. The method presented here is general and can be applied to several types of neuroimaging data. We showed that the method can capture general features such as age, but is also correlated to higher-order behavioural features. We wanted to demonstrate the brain-behaviour associations, as one of the main goals of neuroimaging is to understand human behaviour in more detail. In the future, we want to implement the extreme multivariate framework for understanding several psychiatric disorders. We expect that this method would be an excellent choice in a clinical setting, where extreme brain deviations and their underlying dependence could have a significant effect on extreme behaviour taking us further in the quest for individualized and precision medicine (Insel and Cuthbert, 2015).

Conclusion

The aim of this study was to understand and explore the extreme value tail dependence structure embedded in brain imaging data. We presented a framework for analysing extreme multivariate brain data and interpreting the found extreme eigenbasis. We showed that this framework can be used to find variables that vary together in the tails, through a careful implementation of the tail pairwise dependence matrix (TPDM). We showed that the

extreme value structure could be captured with several components, by applying an eigenvalue decomposition on the TPDM. Finally, we connected the extreme principal components with behaviour, demonstrating the potential for linking multivariate extreme brain deviations to neurological or psychiatric problems.

Data and code availability statement

The UK Biobank data used in this paper are available through a procedure described at http://www.ukbiobank.ac.uk/using-the-resource/.

The methods developed here are available in the PCNtoolkit package at: https://github.com/amarquand/PCNtoolkit

And the EVT code at: https://github.com/CharFraza/extremes_normative_modelling

CrediT authorship contribution statement

Charlotte J. Fraza: Conceptualization, Methodology, Software, Writing – original draft. Mariam Zabihi: methodology, writing – review & editing. Christian F. Beckmann: Supervision, Writing – review & editing. Andre F. Marquand: Supervision, Conceptualization, Methodology, Writing – review & editing.

Acknowledgements

This research was supported by grants from the European Research Council (ERC, grant "MENTALPRECISION" 10100118), the Wellcome Trust under an Innovator award ("BRAINCHART", 215698/Z/19/Z) and a Strategic Award (098369/Z/12/Z) and the Dutch Organisation for Scientific Research (VIDI grant 016.156.415). This research has been conducted using the UK Biobank resource under application number 23668.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis: Principal component analysis. WIREs Comp Stat 2, 433–459. https://doi.org/10.1002/wics.101

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L.R., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. NeuroImage 166, 400–424. https://doi.org/10.1016/j.neuroimage.2017.10.034

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. L., 2004. Statistics of extremes: theory and applications. John Wiley & Sons.

Cannon, T.D., Keller, M.C., 2006. Endophenotypes in the genetic analyses of mental disorders. Annu Rev Clin Psychol 2, 267–290. https://doi.org/10.1146/annurev.clinpsy.2.022305.095232

Chiapino, M., Sabourin, A., 2017. Feature Clustering for Extreme Events Analysis, with Application to Extreme Stream-Flow Data, in: Appice, A., Ceci, M., Loglisci, C., Masciari, E., Raś, Z.W. (Eds.), New Frontiers in Mining Complex Patterns, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 132–147. https://doi.org/10.1007/978-3-319-61461-8_9

Chiapino, M., Sabourin, A., Segers, J., 2018. Identifying groups of variables with the potential of being large simultaneously.

Coles, S., Bawa, J., Trenner, L., & Dorazio, P., 2001. An Introduction to Statistical Modeling of Extreme Values. Londen: Springer.

Cooley, D., Thibaud, E., 2019. Decompositions of dependence for high-dimensional extremes. Biometrika 106, 587–604. https://doi.org/10.1093/biomet/asz028

Davison, A.C., Smith, R.L., 1990. Models for Exceedances over High Thresholds. Journal of the Royal Statistical Society. Series B (Methodological) 52, 393–442.

De Haan, L., Ferreira, A., & Ferreira, A., 2006. Extreme value theory: an introduction. New York: Springer.

Dinga, R., Fraza, C.J., Bayer, J.M.M., Kia, S.M., Beckmann, C.F., Marquand, A.F., 2021. Normative modeling of neuroimaging data using generalized additive models of location scale and shape. https://doi.org/10.1101/2021.06.14.448106

Engelke, S., Ivanovs, J., 2020. Sparse Structures for Multivariate Extremes. arXiv:2004.12182 [math, stat].

Fisher, R.A., Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Mathematical Proceedings of the Cambridge Philosophical Society 24, 180–190. https://doi.org/10.1017/S0305004100015681

Fraza, C.J., Dinga, R., Beckmann, C.F., Marquand, A.F., 2021. Warped Bayesian linear regression for normative modelling of big data. Neuroimage 245, 118715. https://doi.org/10.1016/j.neuroimage.2021.118715

Gnecco, N., Meinshausen, N., Peters, J., Engelke, S., 2020. Causal discovery in heavy-tailed models. arXiv:1908.05097 [stat].

Gnedenko, B., 1943. Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. Annals of Mathematics 44, 423–453. https://doi.org/10.2307/1968974

Goix, N., Sabourin, A., Clémençon, S., 2017. Sparse representation of multivariate extremes with applications to anomaly detection. Journal of Multivariate Analysis 161, 12–31. https://doi.org/10.1016/j.jmva.2017.06.010

Insel, T.R., 2014. Mental disorders in childhood: shifting the focus from behavioral symptoms to neurodevelopmental trajectories. JAMA 311, 1727–1728. https://doi.org/10.1001/jama.2014.1193

Insel, T.R., Cuthbert, B.N., 2015. Brain disorders? Precisely: Precision medicine comes to psychiatry. Science 348, 499–500. https://doi.org/10.1126/science.aab2358

Jiang, Y., Cooley, D., Wehner, M.F., 2020. Principal Component Analysis for Extremes and Application to U.S. Precipitation. Journal of Climate 33, 6441–6451. https://doi.org/10.1175/JCLI-D-19-0413.1

Kia, S.M., Huijsdens, H., Rutherford, S., Dinga, R., Wolfers, T., Mennes, M., Andreassen, O.A., Westlye, L.T., Beckmann, C.F., Marquand, A.F., 2021. Federated Multi-Site Normative Modeling using Hierarchical Bayesian Regression. bioRxiv 2021.05.28.446120. https://doi.org/10.1101/2021.05.28.446120

Kievit, R.A., Davis, S.W., Griffiths, J., Correia, M.M., Cam-CAN, Henson, R.N., 2016. A watershed model of individual differences in fluid intelligence. Neuropsychologia 91, 186–198. https://doi.org/10.1016/j.neuropsychologia.2016.08.008

Marquand, A.F., Kia, S.M., Zabihi, M., Wolfers, T., Buitelaar, J.K., Beckmann, C.F., 2019. Conceptualizing mental disorders as deviations from normative functioning. Mol Psychiatry 24, 1415–1424. https://doi.org/10.1038/s41380-019-0441-1

Marquand, A.F., Rezek, I., Buitelaar, J., Beckmann, C.F., 2016a. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. Biol Psychiatry 80, 552–561. https://doi.org/10.1016/j.biopsych.2015.12.023

Marquand, A.F., Wolfers, T., Mennes, M., Buitelaar, J., Beckmann, C.F., 2016b. Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. Biol Psychiatry Cogn Neurosci Neuroimaging 1, 433–447. https://doi.org/10.1016/j.bpsc.2016.04.002

Marsh, T.J., Dale, M., 2002. The UK Floods of 2000–2001: A Hydrometeorological Appraisal. Water and Environment Journal 16, 180–188. https://doi.org/10.1111/j.1747-6593.2002.tb00392.x

McCarthy, P., 2020. Funpack. https://doi.org/10.5281/zenodo.3761702

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

National Academies of Sciences, E., 2016. Attribution of Extreme Weather Events in the Context of Climate Change. https://doi.org/10.17226/21852

Poon, S.-H., Rockinger, M., Tawn, J., 2004. Extreme Value Dependence in Financial Markets: Diagnostics, Models, and Financial Implications. The Review of Financial Studies 17, 581–610.

Resnick, S. I., 2007. Heavy-tail phenomena: probabilistic and statistical modeling. Springer Science & Business Media.

Rutherford, S., Fraza, C., Dinga, R., Kia, S.M., Wolfers, T., Zabihi, M., Berthet, P., Worker, A., Verdi, S., Andrews, D., Han, L.K., Bayer, J.M., Dazzan, P., McGuire, P., Mocking, R.T., Schene, A., Sripada, C., Tso, I.F., Duval, E.R., Chang, S.-E., Penninx, B.W., Heitzeg, M.M., Burt, S.A., Hyde, L.W., Amaral, D., Wu Nordahl, C., Andreasssen, O.A., Westlye, L.T., Zahn, R., Ruhe, H.G., Beckmann, C., Marquand, A.F., 2022. Charting brain growth and aging at high spatial precision. eLife 11, e72904. https://doi.org/10.7554/eLife.72904

Scarrott, C., MacDonald, A., 2012. A review of extreme value threshold estimation and uncertainty quantification. Revstat Statistical Journal 10, 33–60.

Smith, S., Fidel Alfro, A., Miller, K., 2020. UK Biobank Brain Imaging Documentation.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779. https://doi.org/10.1371/journal.pmed.1001779

Wolfers, T., Beckmann, C.F., Hoogman, M., Buitelaar, J.K., Franke, B., Marquand, A.F., 2020. Individual differences v. the average patient: mapping the heterogeneity in ADHD using normative models. Psychol Med 50, 314–323. https://doi.org/10.1017/S0033291719000084

Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F., 2015. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neuroscience & Biobehavioral Reviews 57, 328–349. https://doi.org/10.1016/j.neubiorev.2015.08.001

Wolfers, T., Doan, N.T., Kaufmann, T., Alnæs, D., Moberget, T., Agartz, I., Buitelaar, J.K., Ueland, T., Melle, I., Franke, B., Andreassen, O.A., Beckmann, C.F., Westlye, L.T., Marquand, A.F., 2018. Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. JAMA Psychiatry 75, 1146–1155. https://doi.org/10.1001/jamapsychiatry.2018.2467

Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., Charman, T., Tillmann, J., Banaschewski, T., Dumas, G., Holt, R., Baron-Cohen, S., Durston, S., Bölte, S., Murphy, D., Ecker, C., Buitelaar, J.K., Beckmann, C.F., Marquand, A.F., 2019. Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models. Biol Psychiatry Cogn Neurosci Neuroimaging 4, 567–578. https://doi.org/10.1016/j.bpsc.2018.11.013

Zhao, X., Scarrott, C., Oxley, L., Reale, M., 2010. Extreme value modelling for forecasting market crisis impacts. Applied Financial Economics 20, 63–72. https://doi.org/10.1080/09603100903262947

Supplement A.

Figure 17, shows the age and distribution of the participants currently included in our study.

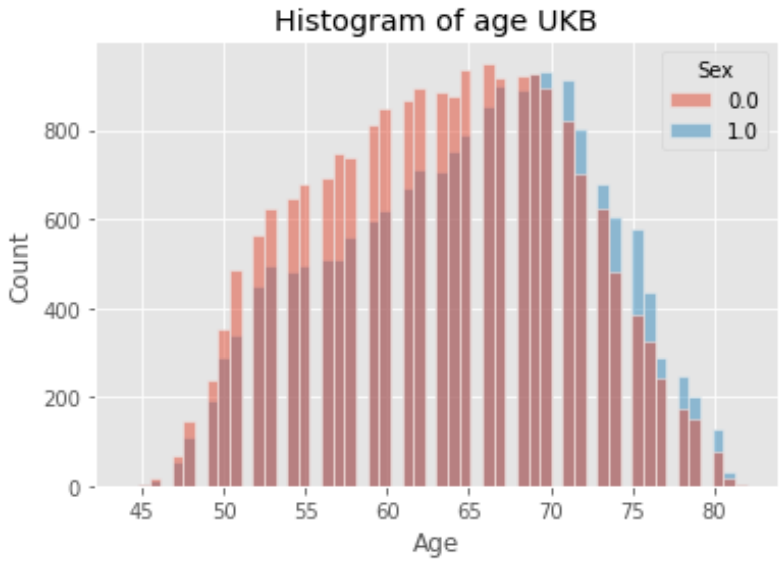


Figure 17 - Showing the age and gender distribution of the IDPs used in this study from the UK biobank.

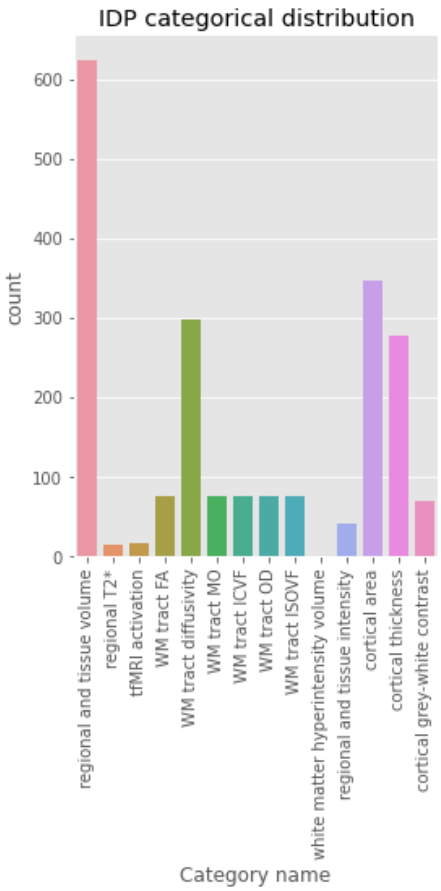


Figure 18 - Showing the distribution of the different IDP categories present in the UK Biobank dataset.

Figure 19 – showing the TPDM matrix for each IDP modality. The TPDM is calculated with the 95-quantile thresholded z-scores in polar coordinates. The images are made with the heatmap function with the hierarchical clustering setting.

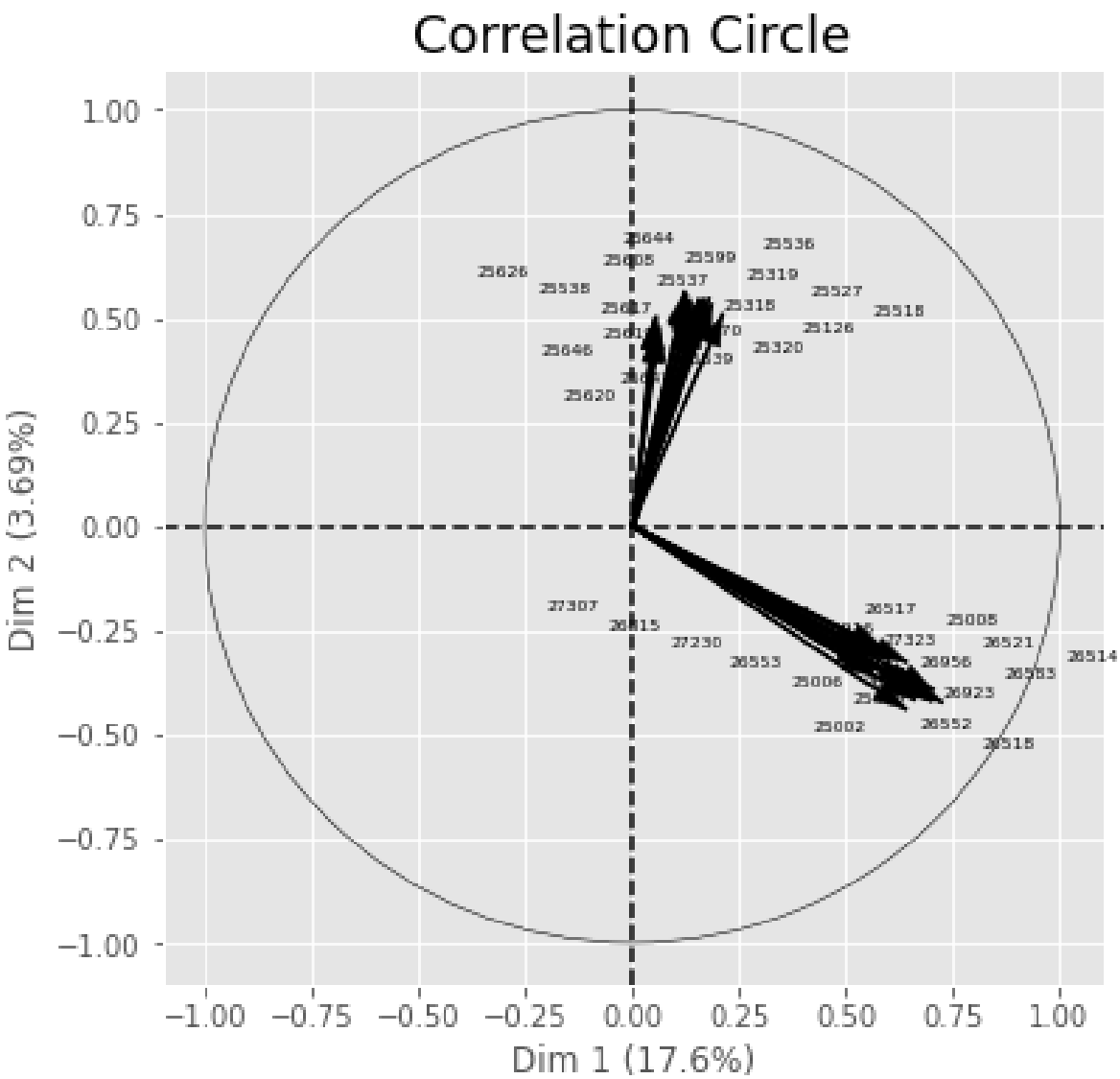


Figure 20 – showing the correlation circle with the loadings of the top 40 contributing variables on the extreme principal components 1 and 2.