Multivariate distance matrix regression for a manifold-valued response variable

Matthew Ryan  Gary Glonek  Melissa Humphries  Jono Tuke

February 14, 2022

Abstract

In this paper, we propose the use of geodesic distances in conjunction with multivariate distance matrix regression, called geometric-MDMR, as a powerful first step analysis method for manifold-valued data. Manifold-valued data is appearing more frequently in the literature from analyses of earthquake to analysing brain patterns. Accounting for the structure of this data increases the complexity of your analysis, but allows for much more interpretable results in terms of the data. To test geometric-MDMR, we develop a method to simulate functional connectivity matrices for fMRI data to perform a simulation study, which shows that our method outperforms the current standards in fMRI analysis.

Keywords: MDMR, Manifold, Geodesic, fMRI, Affine invariant, Simulation

1 Introduction

The process of finding a relationship between a variable of interest $y$ and a set of possible explanatory variables $x_1, x_2, \ldots, x_p$ is a fundamental notion in statistics. When exploring this relationship, accounting for any structure one may find inherently in the data allows for more accurate and directly interpretable results. For instance, if one of our explanatory variables $x_i$ is an ordinal categorical variable this should be accounted for. This is equally true when there is structure found in our response variable $y$, such as interesting geometrical properties when $y$ is manifold valued, that is, $y$ can naturally be viewed as a point on a Riemannian manifold.

The reason you would account for this geometric structure is not always immediately obvious in higher dimensions, but the idea can be highlighted in 2-dimensions. Consider the situation in Figure 1. Here you see 2-dimensional response data that belongs to a natural horseshoe-like shape. If you consider the Euclidean geometry between these points, it is difficult to detect any differences
between the two groups as they are interspersed and clustered. However, if you account for the geometry by travelling along the horseshoe, you see a clear differentiation between the groups as they suddenly become very far apart.

Accounting for the geometrical structure in response variables has appeared in a wide variety of applications, from studying patterns of earthquakes [Cohen & M (2015)] to analysing trends in neuroimaging data [Pennec et al. (2006); Venkatesh et al. (2020)]. This is usually done by either constructing statistical methods on a Riemannian manifold to be applied in a general framework, such as geodesic regression [Fletcher (2013)], or exploiting the Riemannian structure inherently found in the data of interest, such as Venkatesh et. al.’s work on participant identification [Venkatesh et al. (2020)]. A key reason to generalise these Euclidean methods to Riemannian manifolds is that it allows us to model complex non-linear relationships in the data in a more interpretable manner [Fletcher (2013)]. This notion of modelling complex non-linear relationships in an interpretable manner ties in nicely with the intention of multivariate distance matrix regression (MDMR) [Anderson (2001); McArdle & Anderson (2001); Zapala & Schork (2006)].

MDMR, otherwise known as PERMANOVA, is a subject-oriented analysis method which aims to associate observed differences in the response variables of subjects (as defined by a pairwise dissimilarity matrix $D$) to a given set of predictors. First developed by Anderson and McArdle [Anderson (2001); McArdle & Anderson (2001)] for use in ecological data, MDMR has been used in areas ranging from bioinformatics [Zapala & Schork (2006)] to neuroimaging [Ponsoda et al. (2017); Shehzad et al. (2014)]. The theory of MDMR has been well developed since its inception [Anderson & Walsh (2013); Anderson & Robinson (2003); Zapala & Schork (2012)], and it has proven itself as an effective method for determining associations in data with a large set of response variables such as gene expression data [Zapala & Schork (2006)] or functional magnetic resonance imaging (fMRI) [Ponsoda et al. (2017); Shehzad et al. (2014)]. This makes MDMR an attractive alternative to multivariate ANOVA because the results are valid when there are more response variables than there are subjects in the study, since the object under consideration is the dissimilarity matrix.

In this paper we propose a novel approach to MDMR for manifold-valued response data, which we call geometric-MDMR. Geometric-MDMR accounts for the geometry of the data through the use of a geodesic distance for the dissimilarity matrix $D$. When you have manifold-valued data, you can consider how far apart these data are as elements on the manifold by considering paths of minimal local length (geodesics) between the points. Since the geodesics on the manifold are precisely determined by the chosen geometry, this will respect the geometrical properties of the data. We show that our method has more power to detect group differences than simply using Euclidean distances through a simulation study. We also argue that Geometric-MDMR is intuitively more interpretable in terms of the data.
Figure 1: An example where considering the natural geometry of the data is the right thing to do. Here we have two sets of data points, the circles and the crosses. If considered as points in $\mathbb{R}^2$, you would find it difficult for distances between the points to distinguish the groups (the short, dashed line). However, when considered as points on the horseshoe shape this distinction in distance becomes clear (the long, bold line).
2 Method

Throughout this paper, the following notation is used:

- $N$ denotes the number of subjects under consideration. The letters $q$ and $p$ will denote the number of response and predictor variables respectively.
- Bold, lower case letters such as $\mathbf{x}$ and $\mathbf{y}$ denote vectors.
- $d$ will denote a distance or dissimilarity measure between the response variables.
- Upper case letters such as $A$ and $G$ denote matrices.
- A superscript $T$ such as $A^T$ will denote the transpose of a matrix.
- A matrix $A$ can be defined by its $ij^{th}$ element with the notation $A = [a_{ij}]_{ij}$.

2.1 MDMR

Multivariate distance matrix regression (MDMR) [Anderson (2001); McArdle & Anderson (2001); Zapala & Schork (2006)] is an alternative approach to multivariate ANOVA for testing hypotheses on high-dimensional data. MDMR provides a permutation-based test for ANOVA-like hypotheses through the calculation of a pseudo-$F$ statistic. The key difference between MDMR and multivariate ANOVA is that MDMR focuses on the pairwise dissimilarity matrix of the response variables and the relationship this has with the predictors.

Consider data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, where $y_i \in \mathbb{R}^q$ is a $q$-variate response variable, and $x_i = (1, x_{1i}, x_{2i}, \ldots, x_{pi})$ is a vector of $p$-predictors and an intercept. Let $d$ denote a dissimilarity or distance measure on the $y_i$ and let $D = [d(y_i, y_j)]_{ij}$ be the pairwise dissimilarity matrix. Denote by $X$ the matrix whose $i^{th}$ row is $x_i$. Consider the double-centred Gower matrix Gower (1966) $G$ given by

$$G = \left( I - \frac{1}{N} \mathbf{1}\mathbf{1}^T \right) A \left( I - \frac{1}{N} \mathbf{1}\mathbf{1}^T \right),$$

where $I$ is the $N \times N$ identity matrix, $\mathbf{1} \in \mathbb{R}^N$ is a vector of ones, and $A = [-\frac{1}{2}d(y_i, y_j)^2]_{ij}$. Then the pseudo-$F$ statistic is given by

$$\hat{F} = \frac{\text{tr}(HG)/((N-p-1))}{\text{tr}((I-H)G(I-H))/(N-1)},$$

where $H = X(X^TX)^{-1}X^T$ is the usual projection matrix. The motivation behind the pseudo-$F$ statistic is that when $d$ is the Euclidean distance and $q = 1$, then $\hat{F}$ corresponds with the usual $F$ statistic from ANOVA. Thus $\hat{F}$ is a natural extension of the $F$ statistic to an arbitrary dimension and distance measure.

As mentioned in the introduction, one of the main strengths of MDMR over ANOVA is that it is valid for $q > N$ since it relies solely on the distance
measure between the data. This makes MDMR a valuable association test for analysing gene expression data or neuroimaging data, which are generally very high dimensional.

2.2 A geometric point of view

Statistics and differential geometry are not new acquaintances, since manifolds appear very naturally in many fields of study [Fletcher (2013); Prechtl (1948); Pennec (1999)]. Geometrically inspired analysis has found applications all over statistics, from detecting humans in photographs [Tuzel et al. (2007)], to analyse the change of shape for parts of the brain [Fletcher (2013)], to modelling the spatial distribution of earthquakes [Cohen & M (2015)].

A key idea in the above works is analysing geodesics on the respective geometries. Geodesics are commonly described on manifolds and relate to the notions of distance and angles on the manifold, so they naturally depend on the geometry under study.

A manifold \( M \) is a space that locally looks like Euclidean space and patches together in a nice way. Generally speaking, a manifold is a space on which we can perform calculus. We may consider tangent vectors on a manifold, which are encapsulated in the tangent bundle \( TM \). To measure angles between tangent vectors and distances between points on the manifold we need to choose a geometry by way of a Riemannian metric \( g \) on the tangent space. A Riemannian metric is a smooth inner product on the tangent space, which makes sense of the ideas of angles and distances.

Let \( u, v \in M \) and \( \gamma : [0, 1] \to M \) be a path from \( u \) to \( v \). Using the Riemannian metric, we can measure the length of \( \gamma \) on \( M \) as

\[
L(\gamma) = \int_0^1 \sqrt{g(\dot{\gamma}(t), \dot{\gamma}(t))} \, dt,
\]

where \( \dot{\gamma} \) is the derivative in time of \( \gamma \). The path \( \gamma \) is a geodesic if

\[
L(\gamma) = \min_{\gamma' \in \Gamma} \int_0^1 \sqrt{g(\dot{\gamma}'(t), \dot{\gamma}'(t))} \, dt,
\]

where \( \Gamma \) is the set of all paths (defined on \([0, 1]\)) from \( u \) to \( v \). In fact, we define the Riemannian distance between \( u \) and \( v \) as

\[
d(u, v) = \min_{\gamma' \in \Gamma} \int_0^1 \sqrt{g(\dot{\gamma}'(t), \dot{\gamma}'(t))} \, dt.
\]

This notion of distance is the key idea to geometric-MDMR. It allows us to perform MDMR on manifold valued data while accounting for this Riemannian geometry.

2.2.1 Example: \( \mathbb{R}^n \)

The simplest example of a manifold would be normal Euclidean space \( \mathbb{R}^n \). This space is trivially a manifold with tangent space \( \mathbb{R}^n \). The canonical Riemannian
metric on this space is given by the dot product between vectors, which induces the usual Euclidean geometry and distance \( d(u, v) = \|u - v\|_2 \). The geodesics that achieve this distance are straight lines through \( \mathbb{R}^n \).

### 2.2.2 Example: \( S^2 \)

The 2-sphere \( S^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\} \) is a 2-dimensional manifold, which can be shown through the use of stereographic projection. The tangent space for a point \( v \in S^2 \) is the plane that touches \( S^2 \) at exactly the point \( v \). We can define the standard Riemannian metric on \( S^2 \) as the restriction of the Riemannian metric from \( \mathbb{R}^3 \) to \( S^2 \). Under this metric, the geodesics on \( S^2 \) are given by the arcs on the great circles, and the distance between two points is the length of these arcs.

### 2.2.3 Example: \( S_n^+ \)

Consider the space of positive definite symmetric \( n \times n \) matrices 

\[
S_n^+ = \{ A \in \text{GL}_n \mathbb{R} : A = A^T, \alpha^T A \alpha > 0 \text{ for all } \alpha \in \mathbb{R}^n \setminus \{0\} \},
\]

which is the natural home of covariance and correlation matrices. \( S_n^+ \) is an \( \frac{n(n + 1)}{2} \)-dimensional manifold with tangent space at \( A \in S_n^+ \) given by \( T_A S_n^+ = \{ B \in \text{GL}_n \mathbb{R} : B = B^T \} \). To define a Riemannian metric on this manifold is non-trivial. Following Förstner & Moonen (2003); Pennec et al. (2006), we consider the affine-invariant geometry. Let \( B_1, B_2 \in T_A S_n^+ \), then the affine-invariant Riemannian metric is given by 

\[
g_A(B_1, B_2) = \text{tr} \left( A^{-\frac{1}{2}} B_1 A^{-1} B_2 A^{-\frac{1}{2}} \right).
\]

The geodesic distance between \( B_1, B_2 \in S_n^+ \) is given by:

\[
d(B_1, B_2) = \| \log(B_1^{-\frac{1}{2}} B_2 B_1^{-\frac{1}{2}}) \|_2 = \sqrt{\sum_{i=1}^{n} \log(\sigma_i)^2},
\]

where \( \sigma_1, \sigma_2, \ldots, \sigma_n \) are the eigenvalues of \( B_1^{-\frac{1}{2}} B_2 B_1^{-\frac{1}{2}} \).

### 3 Simulation Study

Resting-state fMRI involves subjects lying inside an MRI scanner set to detect changes in the blood-oxygen-level-dependent (BOLD) contrast periodically over the scan time Ogawa et al. (1990). This results in hundreds of thousands of volumetric pixels (voxels) representing spatial regions in the brain, with each voxel having a time series of possibly hundreds of observations. Voxels are typically divided into regions of interest (ROI) which involves aggregating the
information of many voxels into similar regions across the brain, defined either functionally or anatomically. One then analyses the functional connectivity matrix of the ROIs as defined by the Pearson correlation matrix between the respective time series.

A typical resting-state fMRI study aims to detect differences in the functional connectivity of the brain between study groups, that is, they aim to detect if the brains of healthy control subjects function differently to patient groups. To explore the effectiveness of our method in such a study, we create a simulation designed to replicate group differences in the functional connectivity between subjects. In this sense, we can consider the response variable for each subject from a resting-state fMRI study as a correlation matrix (or functional connectivity matrix) that naturally lives on the space $S^+_R$, where $R$ is the number of regions in the ROI decomposition.

The process of simulating fMRI is non-trivial. There are packages to simulate fMRI data in python [Ellis et al. (2020)], MATLAB [Erhardt et al. (2012)], and R [Welvaert et al. (2011)]; each of these packages focus on generating the functional time series for a given voxel in the brain. Here, we are interested in the analysis of the functional connectivity matrix, and so propose a novel simulation method by constructing underlying functional connectivity matrices from real data and then perturbing them with known, implanted signals. This is done by randomly sampling a functional connectivity matrix from a cohort of real subjects and implanting a signal into the functional connectivity matrix based on the simulated subjects group (either “Patient” or “Control”). This matrix is then used as the scale matrix in a Wishart distribution to simulate from the Wishart distribution. We then normalise this simulation into a correlation matrix, and repeat this process for every subject. This method is summarised in Figure 2.

For the real cohort of subjects, we chose to look at the COBRE dataset [Aine et al. (2017)], which aims to explore differences in the functional connectivity between healthy controls ($n = 74$) and schizophrenic patients ($n = 72$). The correlation matrices we consider are the $39 \times 39$ matrices defined by the MSDL atlas [Varoquaux et al. (2011)]. This data comes with an array of phenotypic data to use as predictors, and for this simulation we focus solely on the subject group (“Patient” or “Control”). More information on this dataset is found in Appendix A.

If our simulated subject is a patient, we implant a signal of the form

$$B = \begin{bmatrix}
1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{b-1} \\
\rho & 1 & \rho & \rho^2 & \cdots & \rho^{b-2} \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\rho^{b-1} & \rho^{b-2} & \rho^{b-3} & \rho^{b-4} & \cdots & 1
\end{bmatrix}$$

into the default mode network (DMN). The DMN is a collection of four particular ROIs in the MSDL atlas, and is a well studied functional network in the brain that is believed to be most active when a subject is awake and at rest [Bijsterbosch et al. (2017)]. This functional network was chosen as it was
Figure 2: A visual representation of how to simulate the functional connectivity matrix for $b = 4$, $m = -0.55$, and $s = 0.267$. Figure A shows the randomly sampled subject from the real cohort of data (COBRE dataset). Figure B shows how the signal has been implanted into the DMN (upper left region). Figure C shows a simulation from the Wishart distribution using this matrix as the scale matrix. Figure D show the Wishart simulation normalised to a correlation matrix.
found to have no significant association with the patient group in the COBRE dataset. Note that the parameter $\rho$ controls the strength of the signal being implanted and the parameter $b$ controls the size of the signal, that is, how many consecutive ROIs we are implanting signal into.

For these simulations, we consider $b = 2, 3, \text{or } 4$, ranging from the smallest possible signal of interest ($b = 2$) to the size of the actual DMN in the MSDL atlas ($b = 4$). We also consider $\rho = \tanh(p)$ where $p \sim N(m, s^2)$ for $m = -1.83, -0.55, 0, 0.55, \text{or } 1.83$, and $s = 0.267$. The values for $m$ were chosen to provide a progression of correlations from $-0.95$ to $0.95$ on the atanh scale. The value $s = 0.267$ was chosen as this is the standard deviation observed in the DMN for the COBRE dataset, on the atanh scale.

The process of implanting this signal gives us a matrix $\tilde{A}(r)$ for $r \in [0,1]$. This matrix is such that $\tilde{A}(0)$ is the original correlation matrix, $\tilde{A}(1)$ has the full signal $B$ implanted, and $\tilde{A}(r) \in S^+_R$ for all $r \in [0,1]$. The specifics of the simulation method are explained in Appendix B.

We test the power of geometric-MDMR against the current standards of MDMR used in neuroimaging [Ponsoda et al. (2017); Shehzad et al. (2014)], which convert the functional connectivity matrices into vectors by taking the upper triangle, and use either Euclidean distance or a correlation-based distance on the derived vectors. We use the group as the predictor in our MDMR.

The results of the simulation study are seen in Figure 3, where you can clearly see that the geometric-MDMR outperforms the current standards. By using the geometric extension, we find that the MDMR results become more sensitive to subtle changes in the data, likely because this method considers the data in its natural geometry rather than forcing a Euclidean structure. In this example, the affine-invariant geometry pushes the matrices with zero or infinite determinants out to infinity, which is ideal as these would not be considered functional connectivity matrices. This results in distances between functional connectivity matrices being stretched along a curved path, and we can think of the distance between them as being the distance between valid functional connectivity matrices.

4 Discussion

With the melding of a differential point of view (through the use of geodesics) and multivariate distance matrix regression, we have provided a powerful tool for an a priori analysis on manifold-valued data. Through simulations, we have shown that our method has increased power to detect significant differences in manifold-valued data over the current standards of embedding this data into Euclidean space. We have also argued that geometric-MDMR is more interpretable in terms of the data than the current standards.

Geometric-MDMR would make a valuable contribution to any manifold driven analysis method as an a priori test of association in the data. This could allow researchers to easily reduce their predictor space by removing the predictors that are shown to have insignificant relationships as determined by
Figure 3: The results of the simulation power study comparing geometric-MDMR with the current standards in fMRI analysis. For each combination of the parameters, we repeated the simulation 999 times and calculated the p-value from an MDMR for each of the three methods. The y-axis represents the power from the simulation study as given by the proportion of simulations that returned a p-value of less than 0.05. The x-axis represents the transition from no signal implanted ($r = 0$) to a full signal implanted ($r = 1$) for each set of parameters. The columns represent different sizes of the signal as represented by $b$. The rows are different strengths of the signal as represented by the mean of the normal distribution we simulate from $m$. 

10
geometric-MDMR, leading to simplified and stronger post-hoc analysis.

It should be noted that multivariate distance matrix regression can only be used to determine if a relationship between the predictor variables and response variables exists, but not the nature of that relationship. This makes geometric-MDMR an excellent first step for an analysis, but should be used in conjunction with other methods to strengthen the results.

A drawback specific to considering geodesic distances is that it makes problems more mathematically complicated to formulate. Take for instance the example of fMRI analysis. The current standard when using MDMR [Ponsoda et al. (2017); Shehzad et al. (2014)] is to vectorise the upper triangle of the functional connectivity matrices, which is both simple to do and simple to conceptualise. Comparing this to the formula in Equation (1), it is clear how much more mathematically complicated geometric-MDMR can be. However, we believe that this increase in difficulty is greatly outweighed by the stronger interpretability and power of geometric-MDMR.

Geometric-MDMR provides an excellent first step to determine if a relationship is present in our data. So far, post-hoc analysis of MDMR tends to be done on a situational basis, so post-hoc analysis for geometric-MDMR has yet to be explored. A plausible candidate would be something like geodesic regression [Fletcher (2013)], so the conjunction of these two methods opens an avenue for further research.

References

Aine, C. J., Bockholt, H. J., Bustillo, J. R., Cañive, J. M., Caprihan, A., Gasparovic, C., Hanlon, F. M., Houck, J. M., Jung, R. E., Lauriello, J., Liu, J., Mayer, A. R., Perrone-Bizzozero, N. I., Posse, S., Stephen, J. M., Turner, J. A., Clark, V. P., & Calhoun, Vince D. 2017. Multimodal Neuroimaging in Schizophrenia: Description and Dissemination. *Neuroinformatics, 15*(Oct), 343–364.

Anderson, Marti J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*(Feb), 32–46.

Anderson, Marti J., & Robinson, John. 2003. Generalized discriminant analysis based on distances. *Australian & New Zealand Journal of Statistics, 45*(Sept), 301–318.

Anderson, Marti J., & Walsh, Daniel C. I. 2013. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs, 83*(Nov), 557–574.

Behzadi, Yashar, Restom, Khaled, Liu, Joy, & Liu, Thomas T. 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage, 37*(Aug), 90–101.
Bijsterbosch, Janine, Smith, Stephen, & Beckmann, Christian. 2017. *Introduction to Resting State fMRI Functional Connectivity*. Oxford University Press.

Cohen, Taco S, & M, Welling. 2015. Harmonic Exponential Families on Manifolds. Pages 1757–1765 of: Proceedings of the 32nd International Conference on Machine Learning (ICML).

Ellis, Cameron T., Baldassano, Christopher, Schapiro, Anna C., Cai, Ming Bo, & Cohen, Jonathan D. 2020. Facilitating open-science with realistic fMRI simulation: validation and application. *PeerJ*, 8(Feb), e8564.

Erhardt, Erik B., Allen, Elena A., Wei, Yonghua, Eichele, Tom, & Calhoun, Vince D. 2012. SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability. *NeuroImage*, 59(Feb), 4160–4167.

Fletcher, P. T. 2013. Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision*, 105(Nov), 171–185.

Fréchet, Maurice. 1948. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l’institut Henri Poincaré*, 10, 215–310.

Förstner, Wolfgang, & Moonen, Boudewijn. 2003. *A Metric for Covariance Matrices*.

Gower, J. C. 1966. Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika*, 53(Dec), 325.

McArdle, Brian H, & Anderson, Marti J. 2001. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, 82, 290–297.

Ogawa, Seiji, et al. 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. USA*, 87, 9868–9872. The first publication on BOLD I believe.

Pennec, Xavier. 1999. Probabilities and Statistics on Riemannian Manifolds: Basic Tools for Geometric measurements.

Pennec, Xavier, Fillard, Pierre, & Ayache, Nicholas. 2006. A Riemannian Framework for Tensor Computing. *International Journal of Computer Vision*, 66, 41–66.

Ponsoda, Vicente, Martínez, Kenia, Pineda-Pardo, José A., Abad, Francisco J., Olea, Julio, Román, Francisco J., Barbey, Aron K., & Colom, Roberto. 2017. Structural brain connectivity and cognitive ability differences: A multivariate distance matrix regression analysis. *Human Brain Mapping*, 38(Feb), 803–816.
A COBRE and the MSDL atlas

The data we consider in this paper is the COBRE dataset [Aine et al. (2017)], which was downloaded using the Python package nilearn v 0.6.2. This is an open source fMRI study that provides anatomical and functional MRI images for 72 patients with Schizophrenia and 74 healthy control patients. The data was preprocessed using NIAK 0.17 under CentOS version 6.3 with Octave version 4.0.2 and the Minc toolkit version 0.3.18. The data was also subjected to confound regression where they removed six motion parameters, the frame-wise displacement, five slow drift parameters, average parameters for white matter, lateral ventricles, and global signal, as well as 5 estimates for component based noise correction [Behzadi et al. (2007)].

The ROI atlas we consider for this data is the multi-subject dictionary learning (MSDL) atlas [Varoquaux et al. (2011)]. This is a functional brain atlas,
meaning that voxels are grouped together based on similar brain function instead of anatomical location. This atlas partitions the brain into 39 functional nodes belonging to 17 distinct brain networks. By brain network, we mean a collection of nodes that have been shown to work cohesively together.

B Simulating functional connectivity matrices

This section will describe the simulation process from Section 3 in more detail. Recall that to simulate a single subject we do the following:

1. Randomly select a correlation matrix $A$ from our chosen dataset (the COBRE data).

2. If the subject is to be in the patient group, generate a signal matrix $B$ as described in Section 3. Implant this signal into $A$.

3. Use the resulting matrix as the scale matrix for a Wishart distribution to produce a simulation for the subject.

4. Normalise this matrix into a correlation matrix.

Here, we describe how the signal matrix $B$ is implanted into the original matrix. In the following, fix values for $b$ and $m$ from Section 3.

Suppose you have a correlation matrix $A$ (an $R \times R$ matrix) and wish to implant a signal $B$ (a $b \times b$ matrix). Without loss of generality, suppose you are implanting $B$ into the top left corner of $A$ (you may rearrange the columns and rows of $A$ such that this is always true). Write

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{11}$ is of dimension $b \times b$. Define

$$B(r) = (1 - r)A_{11} + rB$$

for $r \in [0, 1]$, and

$$C(r) = \begin{bmatrix} \text{chol}(B(r))^T(\text{chol}(A_{11})^{-1})^T & 0 \\ 0 & I_{R-b} \end{bmatrix}$$

where chol stands for the Cholesky square root, and $I_{R-b}$ is the $(R-b) \times (R-b)$ identity matrix. Then

$$\tilde{A}(r) = C(r)AC(r)^T$$

has the property that $\tilde{A}(0) = A$, $\tilde{A}(1)_{11} = B$, and $\tilde{A}(r) \in S_R^+$ for all $r \in [0, 1]$. This is the process by which we implant the signal matrix $B$ into $A$. If the simulated subject is not in the patient group, we take $\tilde{A}(r) = A$ for all $r \in [0, 1]$.

Now that you have a valid correlation matrix $\tilde{A}$ for a subject, you can add variational noise via the Wishart distribution. That is, the simulation you
consider for the subject is a random observation from a \text{Wishart}(\hat{A}, k)$, where the degrees of freedom $k$ is randomly selected from integers between 39 and 150. The lower bound 39 is chosen as this is the dimension of $\hat{A}$ in the COBRE dataset, and the upper bound of 150 is chosen as this is the length of the time series observed for each subject in the COBRE dataset. The resulting matrix is then normalised so that the diagonal entries are all one and it is a valid correlation matrix.