Novel metrics and application of nearest-neighbor feature selection for comparing resting-state fMRI brain atlases

Bryan A. Dawkins[1], Rayus Kuplicki[2], Trang T. Le[3], Alejandro A. Hernandez[1], and Brett A. McKinney[1,4,*]
[1]Department of Mathematics, University of Tulsa, Tulsa, OK 74104, USA
[2]Laureate Institute for Brain Research, Tulsa, OK 74136, USA
[3]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104
[4]Tandy School of Computer Science, University of Tulsa, Tulsa, OK 74104, USA.

Abstract

Resting-state functional connectivity MRI (rs-fMRI) data consists of correlation matrices, where correlations are computed between the time series from brain Regions of Interest (ROIs). There are many different parcellations of the human brain into collections of ROIs. These parcellations, or atlases, can be used in case-control studies in order to understand and accurately classify subject phenotypes. We present new metrics for nearest-neighbor distance-based feature selection at the ROI level. Using our new metrics, we apply a novel nearest-neighbor feature selection algorithm to calculate relative importance of ROIs in two existing brain atlases. We use integer programming to derive a mapping between brain atlases to determine spatially similar ROIs. With ROI importance scores and spatial similarity between atlases, we create a new brain parcellation that combines aspects of both brain atlases.

1 Background

Resting-state fMRI data exists in high dimensions and has many sources of noise, such as physiological or motion related [1]. Feature selection is typically done with the purpose of determining brain regions of interest (ROIs) that accurately discriminate between cases and controls in order to understand a particular phenotype. The data consists of pairwise ROI-ROI correlations, where each ROI is a time series measuring brain activity in a particular region or regions of the brain while a subject is not performing a task. A typical data set consists of m subject-specific correlation matrices of dimension $p \times p$, where the pairwise correlations are computed between p ROIs from a brain atlas. Nearest-neighbor distance-based feature selection in rs-fMRI data has been performed using the private evaporative cooling method, which used pairwise ROI-ROI correlations as predictors of a particular phenotype. However, nearest-neighbor feature selection algorithms have not been applied at the ROI level to assess the relative importance of ROIs for a given phenotype. To address this, we have previously proposed a new distance metric that allows us to compute the importance of individual ROIs using a nearest-neighbor distance-based approach. We use this new distance metric with a novel nearest-neighbor feature selection algorithm called Nearest-neighbor Projected Distance Regression (NPDR) in order to compute ROI importance and the corresponding pseudo P values [2]. Our analysis is done on subject rs-fMRI correlation matrices generated by two well known brain atlases with spherical [3] and anatomically shaped [4] ROIs. Cross sections through each atlas were visualized (Fig. 1) using the Analysis of Functional NeuroImages (AFNI) software [5].

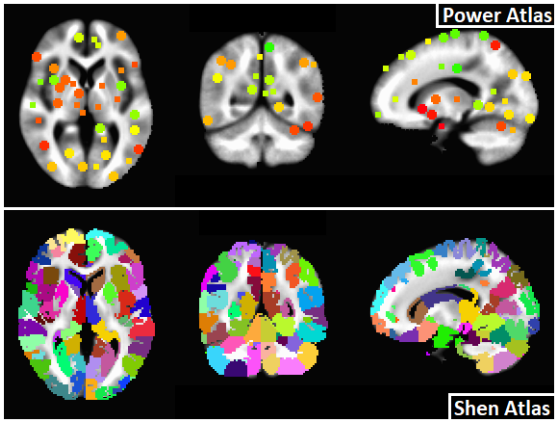

Fig 1. Two-dimensional slices trough brain atlases with spherical [3] and anatomically shaped [4] ROIs. Slices are shown at the same locations in each atlas for the three different orientiations. Atlases are in MNI space [6] so that voxel and ROI locations are comparable between them. Each atlas is labeled by the first author of the manuscripts in which they were introduced [3, 4].

In order to make spatial comparisons between any pair of brain atlases, we first compute a distance matrix containing all pairwise distances between the different collections of atlas ROIs. Distances are defined based on a set dissimilarity metric that accounts for differences in voxel collections between pairs of ROIs. In a particular coordinate system, voxels have well defined three-dimensional locations in a given brain atlas. As long as two different atlases are in the same coordinate system, we can compare voxel membership between opposing atlas ROIs. We use an integer program that defines the standard Assignment Problem (AP) to find the one-to-one mapping between the two sets of atlas ROIs [7]. The collection of all mapped ROIs gives the closest spatial analogy between the two atlases, which tells us the closest relationship between the two sets of ROIs from different atlases. The collection of unmapped ROIs gives an indication of spatial uniqueness in the two atlases, respectively. All ROIs can be further mapped to a well defined anatomical region of the brain, which allows us to point out potential targets for better understanding the phenotype of interest.

Our spatial mapping between atlases and relative importance scores for ROIs in each respective atlas provides a way to combine relevant and distinct aspects of each brain atlas into a new parcellation. This new atlas includes important ROIs that are in the optimal one-to-one mapping from the solution to the assignment problem and any important unmapped ROIs from each atlas. Spatial overlap and attribute importance can serve as a useful tool for other researchers to compare, contrast, and combine two atlases. In particular, our results show how one might choose either of the two atlases to study the phenotype of interest we are considering in this work.

2 Methods

In this section, we first describe real rs-fMRI data generated from healthy controls (HC) and subjects with major depressive disorder (MDD), eating disorder (ED), substance abuse (SA), or anxiety disorder (AD). Using integer programming, we then derive a one-to-one mapping between the ROIs in two brain atlases used to generate the real data mentioned previously. Finally, we use our new distance metric for rs-fMRI data, along with NPDR, to compute importance scores for ROIs in each atlas from the real data.

2.1 Real rs-fMRI data

This is where we describe the LIBR data.

2.2 Spatial overlap between brain atlases

Let R_A and R_B represent regions of interest (ROIs) in atlases A and B, respectively. We assume that atlases A and B are in the same coordinate space. Since R_A and R_B are just collections of voxels that have well defined three-dimensional coordinates within an atlas, the spatial overlap between R_A and R_B can be defined as the set intersection between the two ROIs. Spatial dissimilarity between R_A and R_B can be computed with the Jaccard metric, which is given by the following

$$\mathrm{d}^{\mathrm{J}}(R_A, R_B) = \frac{|R_A \cup R_B - R_A \cap R_B|}{|R_A \cup R_B|}, \tag{1}$$

where the $(-)$ sign denotes set complement and $|\cdot|$ represents set cardinality. If the intersection $R_A \cap R_B$ is empty, then the two ROIs do not share any voxels and the Jaccard distance (Eq. 1) between them is 1. On the other hand, the Jaccard distance is 0 if the union $R_A \cup R_B$ and intersection $R_A \cap R_B$ are the same sets, which means the two ROIs have exactly the same voxels. All other possible Jaccard distances between R_A and R_B are strictly within $(0, 1)$. Hence, the Jaccard metric is contained within $[0, 1]$. The reason for division by $|R_A \cup R_B|$ in the denominator of the Jaccard metric (Eq. 1) is specifically to normalized the distance to be within $[0, 1]$. Otherwise, this distance between two ROIs would be affected by the cardinalities of R_A and R_B, respectively. The Jaccard metric is intuitive in this context because ROIs are not just points in space, but rather they can have irregular three-dimensional shapes. Therefore, a Euclidean metric that gives the straight-line distance between two points does not necessarily indicate 'closeness' between two ROIs. It is possible to compute the Euclidean distance between the centroids of two ROIs, but the ROIs may not share many voxels due to their potentially irregular shapes. Therefore, it is more informative to use a distance metric that uses set operations like the Jaccard metric (Eq. 1). We show an example (Fig. 2) of the Jaccard distance between ROIs R_A and R_B that contain n_1 and n_2 voxels, respectively.

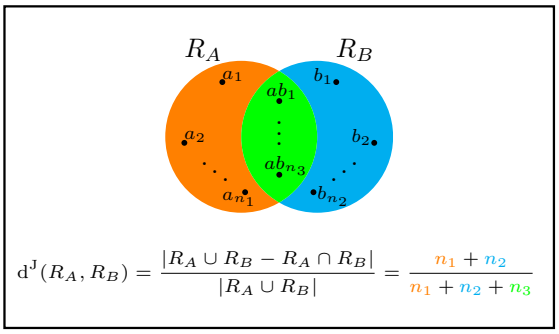

Fig 2. Example computation of Jaccard distance between ROIs R_A and R_B from two atlases A and B, respectively. There are n_1, n_2, n_3 voxels in R_A only, R_B only, and both R_A and R_B, respectively. The numerator gives the number of voxels unique to R_A (n_1) plus the number of voxels unique to R_B (n_2). The denominator contains the total number of voxels in R_A or R_B.

Each ROI in atlas A may overlap many different ROIs in atlas B. On the other hand, some ROIs in A may not overlap any ROIs in B. Furthermore, it is likely that A and B contain different numbers of ROIs. If we want to compute a minimum distance one-to-one mapping between the atlases, it is possible that some ROIs in A will not have a mapped partner in B. In order to efficiently compute this atlas-atlas mapping, we formulate this task as a standard Assignment Problem [7], which has a very concise definition (Fig. 3).

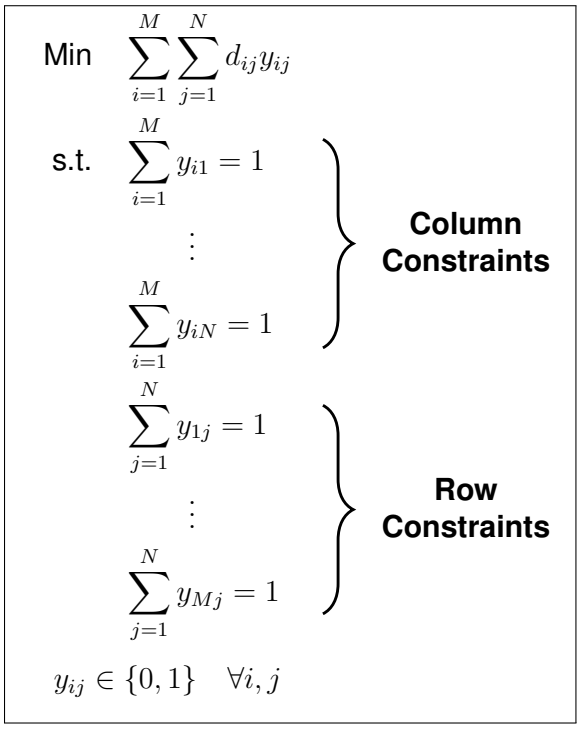

Fig 3. Assignment problem mathematical definition. The assignment matrix Y is binary, where $y_{ij} = 0$ if nodes i and j are assigned to each other and 0 otherwise. The distance matrix D between all nodes is computed to assign costs to arcs (or edges) between nodes. The distance between nodes i and j is denoted by d_{ij}. Therefore, the objective function is the sum of all pairwise distances in the collection of assigned arcs. the column and row constraints dictate that each node is connected to exactly one and only one other node.

The objective function is the sum over all pairwise distances between nodes included in the mapping, where inclusion is determined by the binary solution matrix Y that has the following definition

$$y_{ij} = \begin{cases} 1 & \text{nodes } i \text{ and } j \text{ connected,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Each row and each column of Y has a sum-to-one constraint, which means that each node in one collection is connected to exactly one other node in another disjoint collection. This problem assumes that the order (or size) of each collection is equal ($M = N$ in Fig. 3), so that a one-to-one assignment is possible. In the context of brain atlases, we will satisfy this requirement by adding artificial variables to our solution matrix Y. Pairwise distances between actual ROIs in a brain atlas and an artificial variable will be given a large constant value, so that ROIs in atlas A will preferentially map to another ROI in B if a mapping is possible. Absent the possibility of a mapping, an ROI will map to an artificial variable, which implies that this particular ROI goes unmapped in our unconstrained solution. We show a diagram of atlas-atlas mapping (Fig. 4) that depicts ROIs as nodes and possible mappings as dashed lines between nodes in atlases A and B. Each possible mapping has a distance (d_{ij}) associated with its inclusion in the solution matrix Y. The solution is ultimately found by choosing connections (solid lines) between atlas ROIs so that each ROI in A is connected to exactly one ROI in B, which simultaneously minimizes the sum of all distances in the mapping. We also show how this mapping might look with respect to the actual brain atlases we consider in this manuscript (Fig. 5). We have taken the unlabeled two-dimensional slices through each atlas (Fig. 1) and assigned ROIs to each other based on similar location. Each of the atlases are in the Montreal Neurological Institute (MNI) coordinate space [6], which allows us to compare the two atlases.

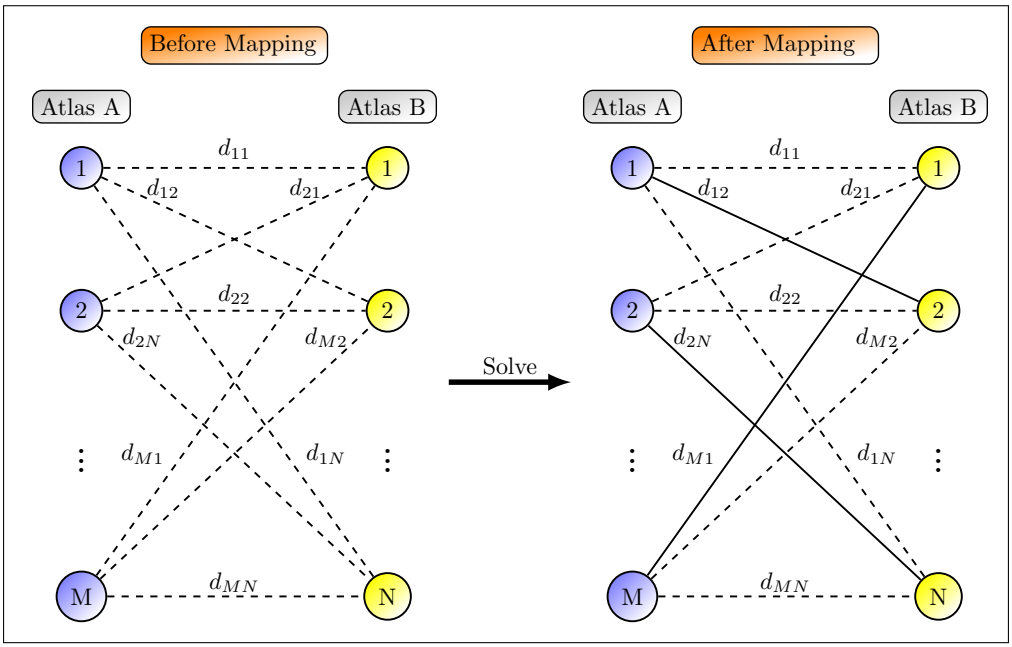

Fig 4. Assignment problem for brain atlas ROI-ROI mapping. Before the minimum distance pairwise mapping is found, all possible pairwise mappings are possible (dashed lines). Each ROI in atlas A must have exactly one partner in atlas B. After mapping, the minimum distance pairs are selected to give a one-to-one correspondence between ROIs (solid lines).

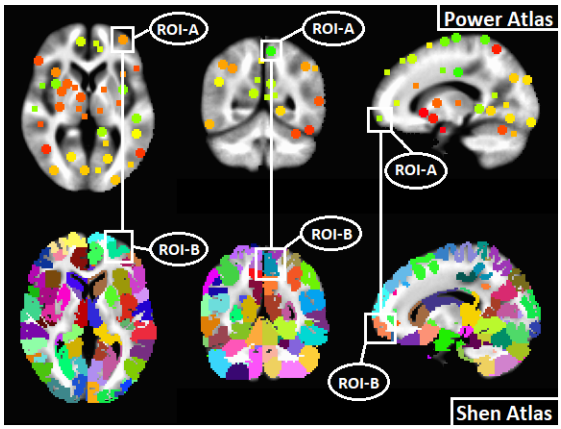

Fig 5. Visualization of mapping between spherical and anatomically shaped ROIs in two brain atlases [3, 4]. ROIs that are in similar locations are potential partners in the resulting mapping from solving the assignment problem (Fig. 3). The atlas with spherical ROIs [3] is labeled as atlas A, while the atlas with anatomically shapped ROIs is labeled as B.

Since there are 277 and 278 ROIs in the atlases with spherical and anatomically shaped ROIs, respectively, we first compute a distance matrix D that contains all pairwise distances between atlas ROIs. This gives us a 278×277 distance matrix to which we add a column of large numeric entries to represent distances from ROIs to an artificial variable. This makes the distance matrix square, which allows us to satisfy the one-to-one constraint of the Assignment Problem (Fig. 3). We used the open source statistical computing software **R** [8], which includes a package called lpSolve [9] for solving a variety of linear and integer programs. The lpSolve package contains a function called lp.assign(), which simply takes a square distance matrix as input and produces solution matrix Y and an objective function value as output. For distance matrices comparable in dimension to D, the function takes only a few seconds to run.

The solution matrix in the output from lp.assign() has a total of 230 actual mapped

pairs. Some of the entries in the solution matrix Y are extraneous, which we define as an ROI pair including an artificial variable, an ROI pair for which the Jaccard dissimilarity is maximal (e.g., $\mathrm{d}^{\mathrm{J}}(R_A, R_B) = 1$), or an ROI pair for which the Jaccard dissimilarity is simply larger than another pairwise dissimilarity. Therefore, there are 47 and 48 unmapped ROIs from the spherical and anatomically shaped atlases, respectively. We give a summary of both mapped and unmapped ROIs from each atlas that includes the mapping assignment solution, corresponding brain region for each ROI, and corresponding functional network for each ROI (cite supplementary spreadsheets/tables). We also show a visualization of one atlas overlaid onto another (Fig. 6). Mapped ROIs from each atlas are highlighted to show relative spatial similarity (top of Fig. 6), which shows validation of the Assignment Problem solution. Unmapped ROIs from each atlas are highlighted to show relative spatial dissimilarity (bottom of Fig. 6), which similarly validates their categorization as extraneous mappings.

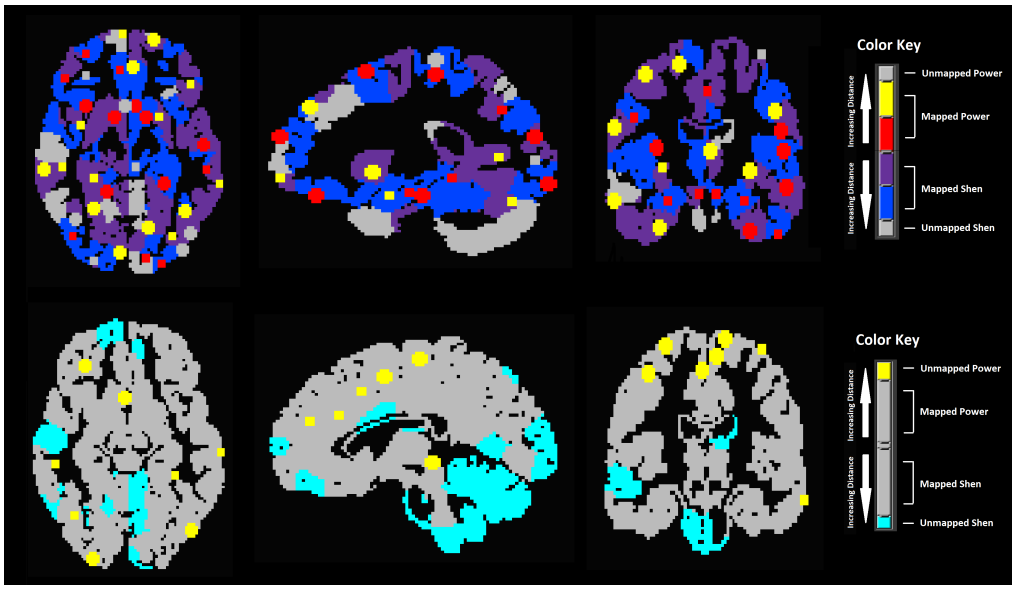

Fig 6. Visualization of mapped and unmapped ROIs from atlases with spherical and anatomically shaped ROIs, respectively, from Assignment Problem solution. Two-dimensional slices from each atlas are overlaid on top of each other. The upper slices highlight the mapped ROIs from each atlas. The lower slices highlight the unmapped ROIs from each atlas.

Mapped ROIs are an indication of informational and spatial similarity shared by both atlases. Subject data that is derived from the mapped ROIs are highly correlated because each ROI pair shares voxels from which ROI time series are created. Unmapped ROIs in a single atlas may or may not intersect one of the mapped ROIs in another atlas. These unmapped ROIs indicate the most unique aspects of each atlas relative to one another. Hence, unmapped ROIs should be considered for inclusion in a customized atlas if determined to be important for predicting the outcome.

2.3 Relative importance of ROIs

We load all subject-level resting-state fMRI (rs-fMRI) correlation matrices into a single matrix for our analysis (Fig. 7). Each column represents a single instance $i \in \mathcal{I}$ and each consecutive collection of $p - 1$ rows represents a single ROI $a \in \mathcal{A}$. We apply a Fisher r-to-z transform to all pairwise correlations and then standardize so that each column is approximately standard normally distributed ($\mathcal{N}(0, 1)$).

$$
\begin{array}{c}
 \\
\mathsf{ROI}_1(\boldsymbol{a=1}) \\
 \\
 \\
\mathsf{ROI}_2(\boldsymbol{a=2}) \\
 \\
 \\
\vdots \\
 \\
\mathsf{ROI}_p(\boldsymbol{a=p}) \\
 \\
 \\
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \boldsymbol{I_1} & \boldsymbol{I_2} & \boldsymbol{I_3} & \cdots & \boldsymbol{I_m} \end{array} \\
\left[\begin{array}{ccccc}
\hat{\boldsymbol{A}}_{12}^{(1)} & \hat{\boldsymbol{A}}_{12}^{(2)} & \hat{\boldsymbol{A}}_{12}^{(3)} & \cdots & \hat{\boldsymbol{A}}_{12}^{(m)} \\
\hat{\boldsymbol{A}}_{13}^{(1)} & \hat{\boldsymbol{A}}_{13}^{(2)} & \hat{\boldsymbol{A}}_{13}^{(3)} & \cdots & \hat{\boldsymbol{A}}_{13}^{(m)} \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{\boldsymbol{A}}_{1p}^{(1)} & \hat{\boldsymbol{A}}_{1p}^{(2)} & \hat{\boldsymbol{A}}_{1p}^{(3)} & \cdots & \hat{\boldsymbol{A}}_{1p}^{(m)} \\
\hline
\hat{\boldsymbol{A}}_{21}^{(1)} & \hat{\boldsymbol{A}}_{21}^{(2)} & \hat{\boldsymbol{A}}_{21}^{(3)} & \cdots & \hat{\boldsymbol{A}}_{21}^{(m)} \\
\hat{\boldsymbol{A}}_{23}^{(1)} & \hat{\boldsymbol{A}}_{23}^{(2)} & \hat{\boldsymbol{A}}_{23}^{(3)} & \cdots & \hat{\boldsymbol{A}}_{23}^{(m)} \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{\boldsymbol{A}}_{2p}^{(1)} & \hat{\boldsymbol{A}}_{2p}^{(2)} & \hat{\boldsymbol{A}}_{2p}^{(3)} & \cdots & \hat{\boldsymbol{A}}_{2p}^{(m)} \\
\hline
\vdots & \vdots & \vdots & \cdots & \vdots \\
\hline
\hat{\boldsymbol{A}}_{p1}^{(1)} & \hat{\boldsymbol{A}}_{p1}^{(2)} & \hat{\boldsymbol{A}}_{p1}^{(3)} & \cdots & \hat{\boldsymbol{A}}_{p1}^{(m)} \\
\hat{\boldsymbol{A}}_{p2}^{(1)} & \hat{\boldsymbol{A}}_{p2}^{(2)} & \hat{\boldsymbol{A}}_{p2}^{(3)} & \cdots & \hat{\boldsymbol{A}}_{p2}^{(m)} \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{\boldsymbol{A}}_{p,p-1}^{(1)} & \hat{\boldsymbol{A}}_{p,p-1}^{(2)} & \hat{\boldsymbol{A}}_{p,p-1}^{(3)} & \cdots & \hat{\boldsymbol{A}}_{p,p-1}^{(m)}
\end{array}\right]
\end{array}
= \boldsymbol{X}
$$

Fig 7. Organization based on brain regions of interest (ROIs) of resting-state fMRI correlation dataset consisting of transformed correlation matrices for m subjects. Each column corresponds to an instance (or subject) I_j and each subset of rows corresponds to the correlations for an ROI attribute (p sets). The notation $\hat{A}_{ak}^{(j)}$ represents the r-to-z transformed correlation between attributes (ROIs) a and $k \neq a$ for instance j.

We determine attribute importance using Nearest-neighbor Projected Distance Regression (NPDR) [2], which is a novel nearest-neighbor feature selection method that builds upon Relief-Based Algorithms (RBAs) [10, 11]. In order to compute distances, we previously introduced a metric for time series-correlation based (ts-corr) data like rs-fMRI (cite BoD theoretical). The one-dimensional projection (diff) onto a single ROI is define as follows

$$\mathrm{d}_{ij}^{\mathrm{ROI}}(a) = \sum_{k \neq a} \left| A_{ka}^{(i)} - A_{ka}^{(j)} \right|, \tag{3}$$

where $A_{ak}^{(i)}$ and $A_{ak}^{(j)}$ are the correlations between ROI a and ROI k for instances $i, j \in \mathcal{I}$, respectively. With this rs-fMRI diff, we define the pairwise distance between two instances $i, j \in \mathcal{I}$ as follows

$$D_{ij}^{\mathrm{fMRI}} = \sum_{a \in \mathcal{A}} \mathrm{d}_{ij}^{\mathrm{ROI}}(a). \tag{4}$$

3 Results

3.1 New brain atlas

4 Discussion

References

1. César Caballero Gaudes and Richard C. Reynolds. Methods for cleaning the BOLD fMRI signal. *NeuroImage*, 154:128–149, December 2017.

2. Trang T. Le, Bryan A. Dawkins, and Brett A. McKinney. Nearest-neighbor Projected-Distance Regression (NPDR) detects network interactions and controls for confounding and multiple testing. *Under Review*, 2019.

3. Jonathan D Power, Alexander L Cohen, Stephen M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann,

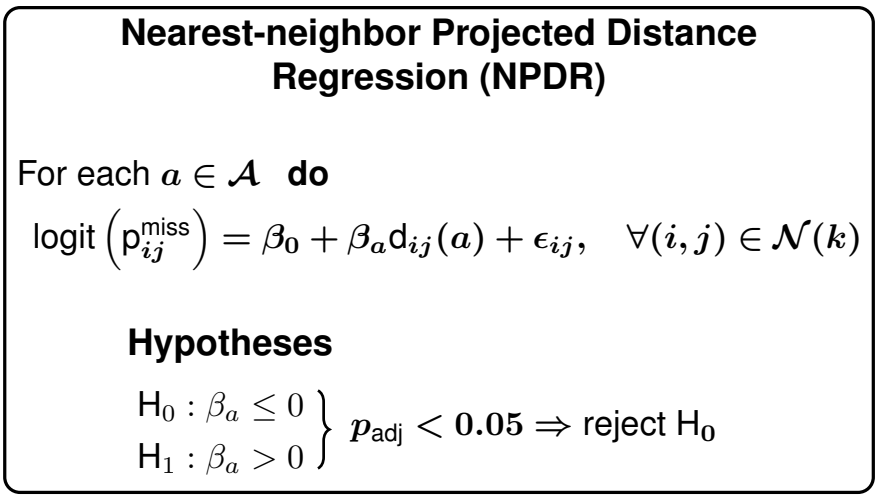

Fig 8. Nearest-neighbor Projected Distance Regression for binary response (case-control). For each attribute $a \in \mathcal{A}$, we estimate beta coefficients for a generalized linear model, where the predictors ($\text{d}_{ij}(a) = |X_{ia} - X_{ja}|$) are one-dimensional projected distances (diffs) with respect to a particular attribute. The argument $\text{p}_{ij}^{\text{miss}}$ is the probability that instances $i, j \in \mathcal{I}$ are in a different phenotype classes.

Fran M Miezin, Bradley L Schlaggar, and Steven E Peterson. Functional network organization of the human brain. *Neuron*, 72(4):665–678, November 2011.

4. X. Shen, F. Tokoglu, X. Papademetris, and R. T. Constable. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage*, (0):403–415, November 2013.

5. RW Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computational Biomedical Research*, 29(3):162–173, June 1996.

6. DL Collins, P Neelin, and TM Peters. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of Computer Assisted Tomography*, 18(2):192–205, April 1994.

7. David W. Pentico. Assignment problems: A golden anniversary survey. *European Journal of Operational Research*, 176:774–793, November 2007.

8. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

9. Michel Berkelaar and others. *lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs*, 2015. R package version 5.6.13.

10. Marko Robnik Šikonja and Igor Kononenko. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, 53:23 – 69, February 2003.

11. Ryan J. Urbanowicz, Randal S. Olson, Peter Schmitt, Melissa Meeker, and Jason H. Moore. Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics*, 85:168–188, 2018.