scientific reports

OPEN

Capturing functional connectomics using Riemannian partial least squares

Matthew Ryan, Gary Glonek, Jono Tuke & Melissa Humphries

For neurological disorders and diseases, functional and anatomical connectomes of the human brain can be used to better inform targeted interventions and treatment strategies. Functional magnetic resonance imaging (fMRI) is a non-invasive neuroimaging technique that captures spatio-temporal brain function through change in blood-oxygen-level-dependent (BOLD) signals over time. FMRI can be used to study the functional connectome through the functional connectivity matrix; that is, Pearson's correlation matrix between time series from the regions of interest of an fMRI image. One approach to analysing functional connectivity is using partial least squares (PLS), a multivariate regression technique designed for high-dimensional predictor data. However, analysing functional connectivity with PLS ignores a key property of the functional connectivity matrix; namely, these matrices are positive definite. To account for this, we introduce a generalisation of PLS to Riemannian manifolds, called R-PLS, and apply it to symmetric positive definite matrices with the affine invariant geometry. We apply R-PLS to two functional imaging datasets: COBRE, which investigates functional differences between schizophrenic patients and healthy controls, and; ABIDE, which compares people with autism spectrum disorder and neurotypical controls. Using the variable importance in the projection statistic on the results of R-PLS, we identify key functional connections in each dataset that are well represented in the literature. Given the generality of R-PLS, this method has the potential to investigate new functional connectomes in the brain, and with future application to structural data can open up further avenues of research in multi-modal imaging analysis.

The functional and anatomical connections of the human brain form complex networks that link the infrastructure of our minds. Understanding these connectomes has the potential to provide insight into the effect of neurological diseases which can be used to better inform targeted interventions and treatment strategies[1,2]. In particular, the functional connectome can shed new light onto psychiatric and neurological conditions such as schizophrenia and autism spectrum disorder (ASD), two conditions that alter brain function from healthy, neurotypical controls[3,4].

A popular approach used to investigate brain function is functional magnetic resonance imaging (fMRI), a non-invasive neuroimaging technique that measures the blood-oxygenation-level-dependent (BOLD) signal over time as a correlate of brain activity[5]. An fMRI image is a complex spatio-temporal picture of the brain with voxels (volumetric pixels) describing the spatial location and a time series for each voxel describing the BOLD signal. To reduce the spatial complexity, voxels can be collated into user-specified regions of interest (ROIs). Functional connectomes can then be investigated through Pearson's correlation matrix between ROIs, known as the functional connectivity matrix.

One approach to investigating functional connectivity is using the partial least squares (PLS) regression method. Introduced by Wold[6] for use in chemometrics, PLS is an extension of multivariate multiple regression to high-dimensional data that predicts the response data from a set of lower-dimensional latent variables (that is, unobserved variables) constructed from the predictor data. Popularised for fMRI by McIntosh et. al.[7], PLS has been used to explore the relationships between fMRI data and either behavioural data, experimental designs, or seed region activation[8]. However, standard PLS ignores the structure of functional connectivity data—functional connectivity matrices are correlation matrices and hence positive definite, that is, they have non-negative eigenvalues. By ignoring the positive definite criteria, standard PLS on functional connectivity data has the potential to ignore intricate non-linearities in the data and provide invalid predictions from the model.

School of Computer and Mathematical Sciences, The University of Adelaide, Adelaide 5005, Australia. email: matthew.ryan@adelaide.edu.au

For a given number of ROIs R, the space of $R \times R$ symmetric positive definite matrices—which includes functional connectivity matrices—forms a convex cone in $R(R+1)/2$-dimensional Euclidean space, that is, they form a shape much like an ice-cream cone in high dimensional Euclidean space that has a singularity at the origin. However, when considered with the affine invariant geometry[9], the space of symmetric positive definite matrices becomes a complete Riemannian manifold—a general mathematical space where we can perform calculus—with non-positive curvature, that is, it curves in on itself in interesting and complex ways. By considering this non-linear geometry on symmetric positive definite matrices we can glean interesting new insights into functional connectivity (see Pennec et. al.[10] and citations therein).

The process of generalising statistical models to Riemannian manifolds has the potential to investigate non-linearities in the data[11], and has been gaining significant interest in neuroimaging in recent years[10]. Fletcher[12] proposed principal geodesic analysis as a generalisation of principal components analysis to Riemannian manifolds for applications in shape analysis for medically-defined anatomical shapes. Later, Fletcher[11] generalised simple linear regression to Riemannian manifolds, called geodesic regression, to investigate how medically-defined anatomical shapes can change with age. Following this vein of research, Kim[13] generalised canonical correlations analysis to Riemannian manifolds to explore multi-modal imaging relationships between diffusion tensor images and structural MRI in Alzheimer's patients. However, PLS, which is closely related to all of these methods, has not yet been generalised to Riemannian manifolds.

Here we propose an extension of the PLS model to allow Riemannian manifold response and predictor data, which we call Riemannian partial least squares (R-PLS). The R-PLS model then allows us to predict from functional connectivity data while accounting for the intricate relationships enforced by the positive definite criteria. To fit the R-PLS model, we propose the tangent non-linear iterative partial least squares (tNIPALS) algorithm, which is related to previously proposed applications of PLS for functional connectivity data in the literature[14–17]. We determine the optimal number of latent variables using cross validation. To aid in interpretability of the high-dimensional functional connectivity data, we determine significant functional connections identified by R-PLS using permutation tests on the variable importance in the projection (VIP) statistic[18], a popular measure of variable importance from standard PLS.

We apply R-PLS to two datasets and two different ROI atlases to demonstrate its versatility in predicting phenotype data from functional connectivity. First is the COBRE dataset[19] which investigates differences in functional connectivity between healthy controls ($n = 74$) and patients with schizophrenia ($n = 72$). We consider two separate atlases on the COBRE dataset to test the generalisability of R-PLS across atlases; the multi-subject dictionary learning (MSDL) atlas[20] to look at a low-dimensional (39 ROIs), data-driven atlas, as well as the automated anatomic labelling (AAL) atlas, a higher-dimensional (116 ROIs) anatomical atlas. The second dataset is the ABIDE dataset from the New York University imaging site[21] which investigates differences in functional connectivity between typical healthy controls ($n = 98$) and subjects with ASD ($n = 75$). We consider the ABIDE data in the AAL atlas[22] to investigate the generalisability of R-PLS across datasets. Thus, when predicting using the MSDL atlas we are considering 780 unique functional connections (since $R = 39$), and when predicting using the AAL atlas we are considering 6786 unique functional connections (since $R = 116$).

Results

For each dataset and atlas we predict the multivariate phenotype information (age and group for the COBRE dataset, as well as sex and eye status for ABIDE) from the functional connectivity data using the R-PLS model. The categorical variables group, sex, and eye status were represented by binary values, and all phenotype information was standardised to have mean zero and standard deviation one. When analysing functional connectivity matrices in the AAL atlas there was one matrix in the COBRE dataset and 24 matrices in the ABIDE dataset which had low-rank, and hence were not positive definite. To deal with these low-rank functional connectivity matrices, we consider regularised functional connectivity matrices $\tilde{F} = F + I$ following Venkatesh et. al.[23], where I is the 116×116 identity matrix. We compare R-PLS to the standard PLS model using the upper triangle of the functional connectivity matrices as the predictors (raw correlations), as well as their Fisher transformed values (Fisher correlations).

Model fitting

We determine the optimal number of latent variables K in the PLS model through ten-fold cross validation using the "within one standard error" rule[24] when minimising the root mean square error (RMSE) on the multivariate phenotype information. Due to the interest in the COBRE and ABIDE datasets in investigating the differences between healthy controls and patients, we also present the group classification metrics of accuracy, sensitivity, specificity, and area under the operator receiver curve (AUC). Since we have represented group as a binary value, we classify subjects in the patient group (schizophrenia or ASD) if their predicted group score is greater than zero, and in the control group otherwise. Graphs of the cross validation results can be found in the supplementary material (Fig. S1).

For the COBRE dataset with the MSDL atlas, ten-fold cross validation showed that $K = 2$ latent variables was the most parsimonious, within one standard error of the minimum RMSE ($K = 3$). When compared with Euclidean PLS using raw and Fisher-transformed correlations, R-PLS outperformed both methods across all metrics except for specificity in group prediction (Table 1) . However, all three methods produced similar results for every metric.

When considering the COBRE dataset in the AAL atlas, ten-fold cross validation showed that $K = 3$ latent variables was the most parsimonious, within one standard error of the minimum RMSE ($K = 3$). Similar to the results from the MSDL atlas, we found that R-PLS outperformed the Euclidean PLS methods across all metrics,

	Riemannian	Raw correlations	Fisher correlations
COBRE-MSDL			
K	2	3	3
Full model metrics (SE)			
R^2	**0.25 (0.035)**	0.23 (0.033)	0.23 (0.036)
RMSE	**1.20 (0.036)**	1.21 (0.025)	1.21 (0.026)
Group classification (SE)			
Accuracy	**0.75 (0.045)**	0.73 (0.032)	0.74 (0.032)
Sensitivity	**0.81 (0.035)**	0.70 (0.057)	0.72 (0.055)
Specificity	0.69 (0.071)	**0.76 (0.048)**	**0.76 (0.048)**
AUC	**0.81 (0.039)**	0.78 (0.027)	0.79 (0.024)
COBRE-AAL			
K	3	3	3
Full model metrics (SE)			
R^2	**0.43 (0.034)**	0.38 (0.043)	0.38 (0.043)
RMSE	**1.04 (0.042)**	1.08 (0.047)	1.08 (0.047)
Group classification (SE)			
Accuracy	**0.79 (0.034)**	0.76 (0.038)	0.76 (0.038)
Sensitivity	**0.80 (0.049)**	0.75 (0.063)	0.75 (0.063)
Specificity	**0.78 (0.030)**	0.76 (0.030)	0.76 (0.030)
AUC	**0.86 (0.031)**	0.83 (0.039)	0.83 (0.040)
ABIDE			
K	3	3	3
Full model metrics (SE)			
R^2	**0.15 (0.015)**	0.07 (0.016)	0.07 (0.016)
RMSE	**1.80 (0.051)**	1.89 (0.059)	1.89 (0.059)
Group classification (SE)			
Accuracy	**0.58 (0.027)**	0.55 (0.032)	0.54 (0.032)
Sensitivity	**0.61 (0.058)**	0.52 (0.064)	0.51 (0.063)
Specificity	0.53 (0.063)	**0.58 (0.065)**	**0.58 (0.065)**
AUC	**0.64 (0.016)**	0.61 (0.047)	0.60 (0.046)

Table 1. Mean (SE) 10-fold cross validation results for Riemannian partial least squares (R-PLS) on the COBRE and ABIDE datasets, and Euclidean PLS using the raw and Fisher transformed correlations. The value K represents the optimal number of latent variables for each model when minimising the root mean square error (RMSE) using the within one standard error rule. The full model metrics are the multivariate R^2 and RMSE. The group classification metrics of accuracy, sensitivity, specificity, and area under the operator receiver curve (AUC) look at the classification for subject group only. R-PLS is the best model for both datasets and atlases over all model metrics, except for specificity (bold values).

although now with the inclusion of group specificity (Table 1). When using the AAL atlas on the COBRE dataset, we observe a substantial increase in the cross validated R^2 value over the Euclidean methods.

For the ABIDE dataset, ten-fold cross validation found $K = 3$ latent variables was the most parsimonious, within one standard error of the minimum RMSE ($K = 6$). When compared with Euclidean PLS using the raw and Fisher-transformed correlations, R-PLS outperformed both methods across all metrics except for specificity in group classification (Table 1). In particular, the R^2 value and AUC for R-PLS was substantially larger than the Euclidean methods.

Interpretation

To investigate the functional connectomes associated to each phenotype variable, we consider the regression coefficient matrix β_{PLS} (see Eq. 4 in the "Methods" section) where the i^{th} column represents the effect of the functional connectivity matrix on the i^{th} response variable (age, group, sex, or eye status). Much like the regression coefficients in ordinary least squares, the coefficient matrix β_{PLS} captures the multivariate association between functional connectivity and the phenotype data. We determine which functional connections are significantly associated with the phenotype variables through a permutation test of the VIP statistic (Eq. 5 in the "Methods" section) using 200 permutations at a significance level of $\alpha = 0.05$, as described in the "Methods" section. All analysis was performed using R[25].

We visualise the columns of the matrix β_{PLS} as symmetric matrices in the tangent space of the Fréchet mean for each dataset, and represent them as connectomes on standard brains images using the `nilearn` package in PYTHON. To assist in visualising patterns in the regression coefficients across the connectome, we average the coefficient values across all connections within and between predefined resting state networks similar to Wong et.

al.[14]. The within-network connectivity is then the average coefficient of all connections within a single resting state network, and the between-network connectivity is the average coefficient of all connections between two resting state networks. For the MSDL atlas this involves reducing the 39 ROIs to the 17 resting state networks associated to the atlas[26]. For the AAL atlas, we associate the 116 ROIs to the seven resting-state networks suggested by Parente and Colosimo[27] and an eighth containing the cerebellum and vermis, which we call the cerebellum network. The resting state networks for the MSDL and AAL atlases are visualised in the supplementary material (MSDL in Figs. S2–S7, AAL in Figs. S8–S10).

For the COBRE dataset with the MSDL atlas, a permutation test of the VIP statistic with 200 permutations found 45 significant functional connections between ROIs as being predictive of age and subject group (Fig. 1). When considered with the AAL atlas, a permutation test of the VIP statistic with 200 permutations found 249 significant functional connections between ROIs as being predictive of age and subject group (Fig. 2). For the ABIDE dataset, a permutation test of the VIP statistic with 200 permutations found 196 significant functional connections between ROIs as being predictive of age, subject group, sex and eye status (Figs. 3 and 4).

Across both atlases for the COBRE dataset, an increase in subject age tended towards a decrease of within-network connectivity (as measured by a mean decrease in functional connectivity within-networks) with particular emphasis on the auditory network, cingulate insula, and left and right ventral attention networks in the MSDL atlas, and the temporo-parietal, limbic, and basal ganglia networks in the AAL atlas (Figs. 1, 2a–c). Increased age was associated with an increase in between-network connectivity, particularly for the MSDL atlas which shows increased connectivity involving the cingulate insula and the motor network. Notably, an increase in age is associated with a decrease in between-network connectivity for the basal ganglia in the AAL atlas but not in the MSDL atlas. In the ABIDE dataset, increased age was associated to both increased and decreased functional connectivity within resting-state networks (Fig. 3a–c). Although we observed increased between-network connectivity for the thalamus and occipital networks, the cerebellum and default mode network exhibited decreased between-network connectivity with age. Note that the decreased within-network connectivity for the basal ganglia with age is also present in the ABIDE dataset.

For subjects in the schizophrenic group, the basal ganglia exhibited both increased and decreased connectivity with other networks across both atlases (Figs. 1 and 2d–f). In particular, in the MSDL atlas there was a decrease in connectivity between the basal ganglia and the cerebellum and salience networks, whereas we observed an increase in connectivity between the basal ganglia and auditory and language networks for the schizophrenic group. The AAL atlas similarly demonstrates the reduced connectivity between the basal ganglia and the cerebellum, and further shows reduced connectivity between the basal ganglia and the fronto-parietal network. In contrast to the analysis with the MSDL atlas, using the AAL atlas suggest there is an increase in connectivity between the basal ganglia and the default mode network for schizophrenic subjects. For both atlases, the default mode network was highly discriminatory for the schizophrenic group showing both increased and decreased between-network connectivity. Finally, we note the the AAL atlas highlights connectivity with the thalamus as being highly discriminatory for the schizophrenic group, which is not represented in the MSDL atlas since there is no thalamus ROI.

For subjects with ASD we observed increased within-network connectivity with the exception of the limbic network (Fig. 3d–f). We also observed decreased between-network connectivity particularly for connections involving the cerebellum or the limbic networks. We observed similar connectivity patterns for subject sex (Fig. 4a–c).

For subjects with their eyes closed, our model suggests there was decreased within-network connectivity (Fig. 4d–f). With the exception of the default mode network, the limbic network, and the basal ganglia, we saw decreased between-network connectivity with particular emphasis on the occipital network.

Discussion

The R-PLS model has identified many functional connections associated to age, ASD, schizophrenia, sex, and eye status that are well represented in the literature. Across both datasets and atlases, we identified the reduction of within-network connectivity with age that has been previously observed[28–30], with exceptions in the temporo-parietal, fronto-parietal, and limbic networks in the ABIDE dataset and the salience network in the COBRE dataset, which all show an increase in connectivity with age. Further, both datasets exhibit the decreased connectivity within the default mode network, consistent with existing literature[31,32]. We also note that the previously observed decrease of within-network connectivity for the basal ganglia[33] was prominent in the AAL atlas, but not the MSDL atlas. This is because there is only one region of interest for the basal ganglia in the MSDL atlas, so within-network connectivity is not defined.

For subjects with ASD, the decreased connectivity with the cerebellum[34] and the limbic[35] networks have been previously observed. However, the decreased between-network connectivity suggested by R-PLS is in contradiction with existing literature[14,36]; in particular, Wong et. al.[14] showed an increase in between-network connectivity associated to ASD on the full ABIDE dataset using logistic regression. Also, observe that the connectivity for subject sex is highly correlated with the connectivity for the ASD group. Although interactions between subject sex and ASD have been identified[37], we believe this highlights a possible limitation of R-PLS and requires further investigation in future research.

The role of the basal ganglia in schizophrenic patients has been previously observed, particularly the decrease in connectivity between the salience network and the basal ganglia[38,39] and the decreased connectivity between the cerebellum and basal ganglia[40]. Similarly, the importance of the thalamus in schizophrenia, identified when using the AAL atlas, is well-known[41,42]. Further, the connectivity patterns involving the default mode network have been previously reported in schizophrenic patients[43–47].

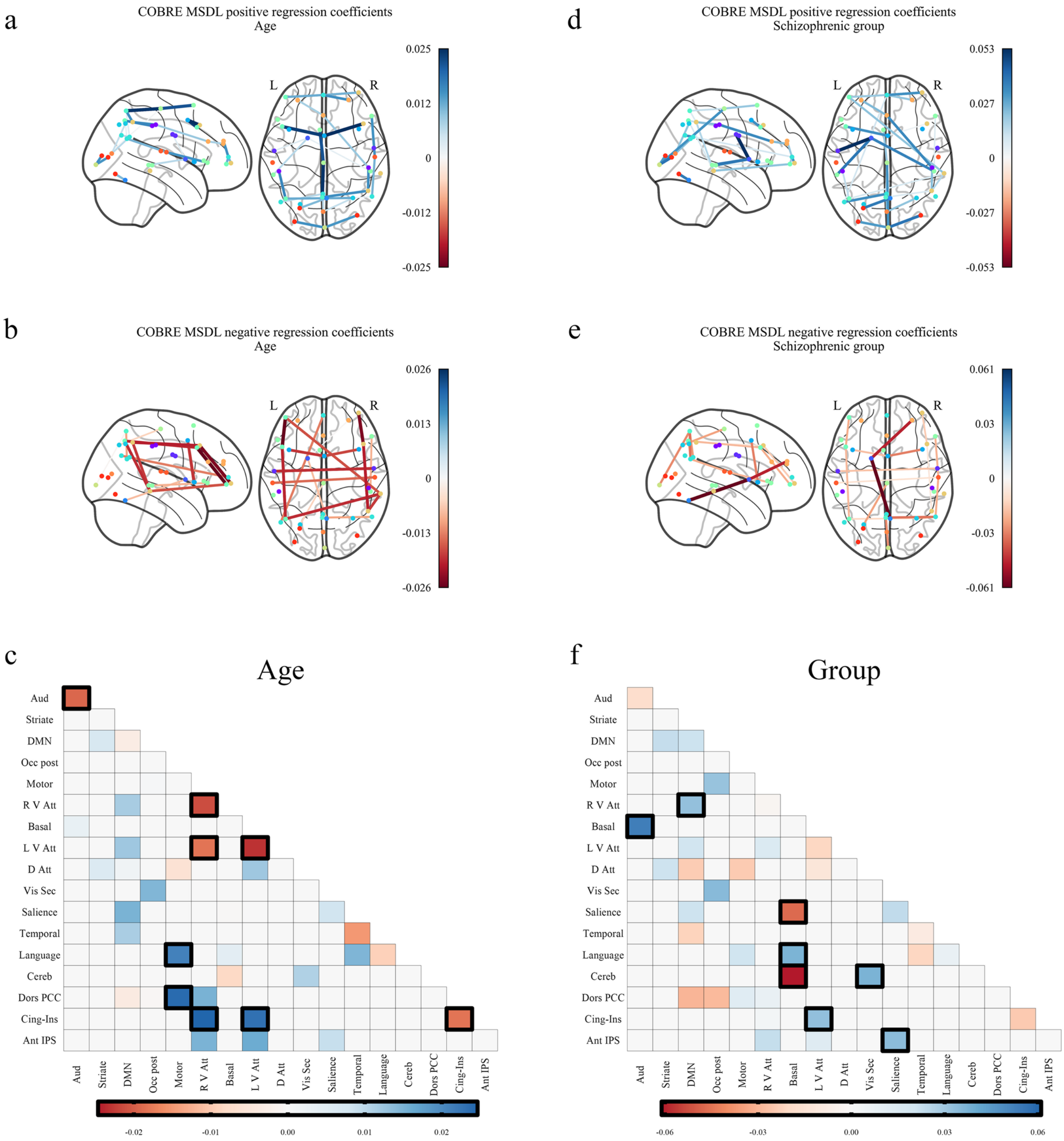

Figure 1. Significant regression coefficients for predicting age and schizophrenia as measured by variable importance in the projection (VIP) for the Riemannian partial least squares (R-PLS) model on the COBRE dataset and the multi-subject dictionary learning (MSDL) atlas with $K = 2$ latent variables, visualised as connectomes and symmetric matrices. Blue values represent connections that are positively associated with the phenotype, that is, an increase in connectivity between two regions with a blue edge would indicate an increase in the phenotype variable. Conversely, red values are connections that are negatively associated with the phenotype, that is, an increase in connectivity between two regions with a red edge would indicate a decrease in the phenotype variable. (**a**) Shows the connections that increase with age, (**b**) shows the connections that decrease with age, and (**c**) shows the average coefficient values for age between the 17 resting state networks of the MSDL atlas[26] (Figs. S2–S7). (**d**) Shows the connections that increase for patients with schizophrenia, (**e**) shows the connections that decrease for patients with schizophrenia, and (**f**) shows the average coefficient values for schizophrenia between the 17 resting state networks of the MSDL atlas[26] (Figs. S2–S7). The darker outlined boxes in (**c**) and (**f**) show the top 25% influential regions as measured by the absolute coefficient value within and between each network. The network abbreviations in (**c**) and (**f**) are: *Aud* auditory, *Striate* striate, *DMN* default model network, *Occ Post* occipital posterior, *Motor* motor network, *R V Att* right ventral attention network, *Basal* Basal Ganglia, *L V Att* left ventral attention network, *D Att* dorsal attention network, *Vis Sec* secondary visual cortex, *Salience* salience network, *Temporal* temporal network, *Language* language network, *Cereb* cerebellum, *Dors PCC* dorsal posterior cingulate cortex, *Cing-Ins* cingulate-insula network, *Ant IPS* anterior intraparietal sulcus.

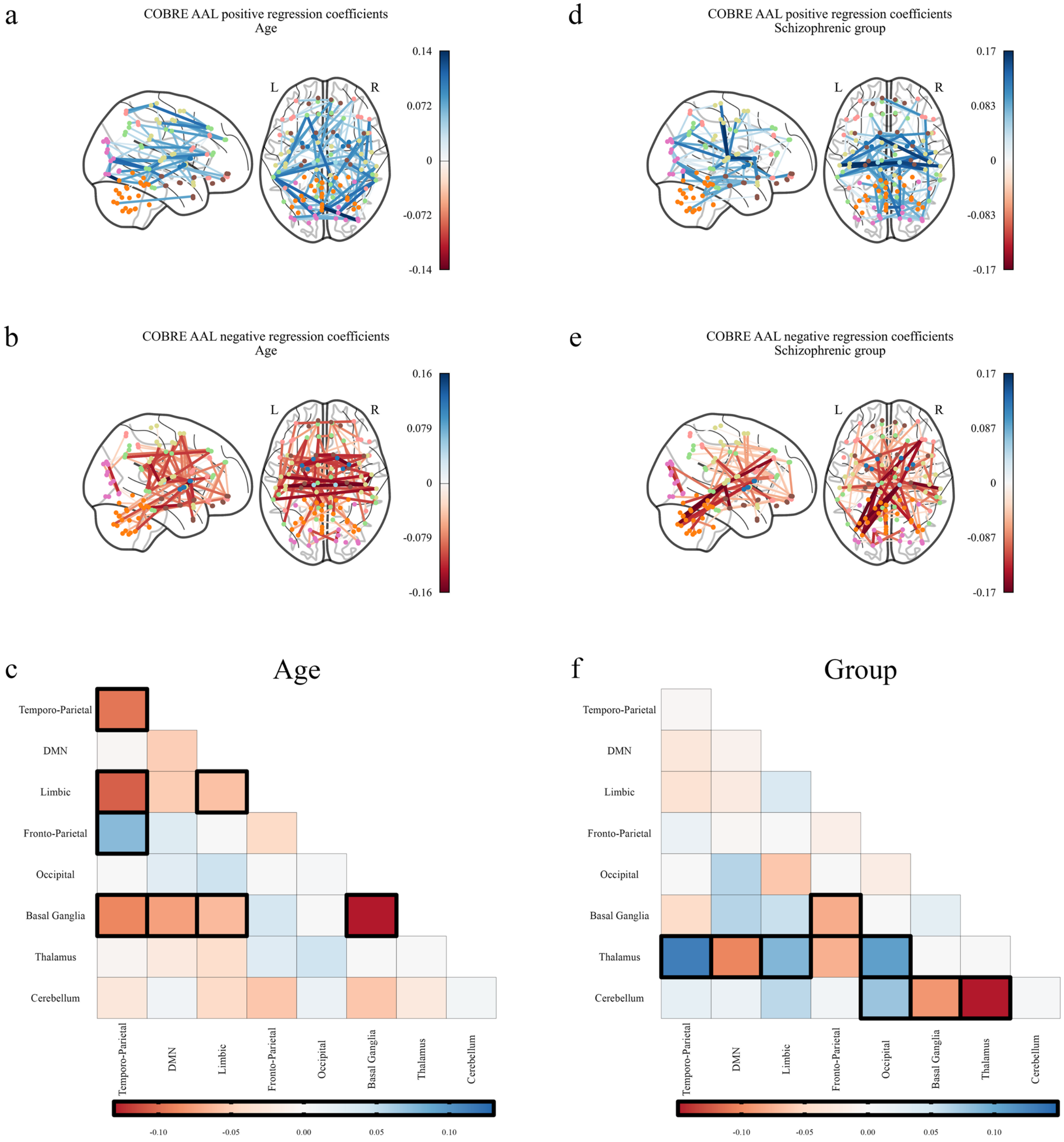

Figure 2. Significant regression coefficients for predicting age and schizophrenia as measured by variable importance in the projection (VIP) for the Riemannian partial least squares (R-PLS) model on the COBRE dataset and the automated anatomic labelling (AAL) atlas with $K = 3$ latent variables, visualised as connectomes and symmetric matrices. Blue values represent connections that are positively associated with the phenotype, that is, an increase in connectivity between two regions with a blue edge would indicate an increase in the phenotype variable. Conversely, red values are connections that are negatively associated with the phenotype, that is, an increase in connectivity between two regions with a red edge would indicate a decrease in the phenotype variable. (**a**) Shows the connections that increase with age, (**b**) shows the connections that decrease with age, and (**c**) shows the average coefficient values for age between the 7 resting state networks identified by Parente and Colosimo[27] and the cerebellum (Figs. S8–S10). (**d**) Shows the connections that increase for patients with schizophrenia, (**e**) shows the connections that decrease for patients with schizophrenia, and (**f**) shows the average coefficient values for schizophrenia between the 7 resting state networks identified by Parente and Colosimo[27] and the cerebellum (Figs. S8–S10). The darker outlined boxes in (**c**) and (**f**) show the top 25% influential regions as measured by the absolute coefficient value within and between each network. In (**c**) and (**f**), *DMN* default mode network.

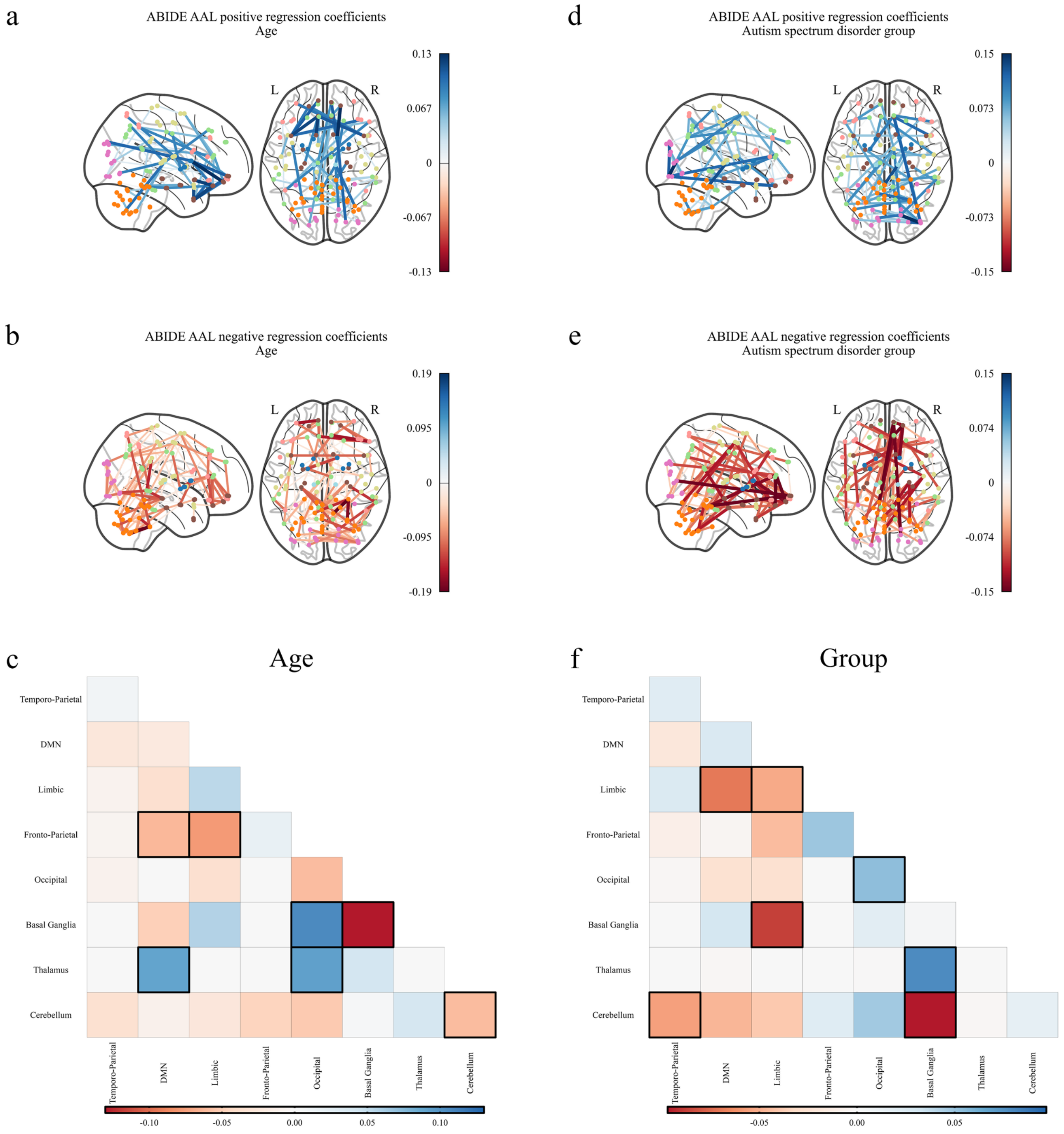
a
ABIDE AAL positive regression coefficients
Age
L
R
0.13
0.067
0
-0.067
-0.13
b
ABIDE AAL negative regression coefficients
Age
0.19
0.095
0
-0.095
-0.19
c
Age
Temporo-Parietal
DMN
Limbic
Fronto-Parietal
Occipital
Basal Ganglia
Thalamus
Cerebellum
-0.10
-0.05
0.00
0.05
0.10
d
ABIDE AAL positive regression coefficients
Autism spectrum disorder group
0.15
0.073
0
-0.073
-0.15
e
ABIDE AAL negative regression coefficients
Autism spectrum disorder group
0.15
0.074
0
-0.074
-0.15
f
Group
-0.05
0.00
0.05

Figure 3. Significant regression coefficients for predicting age and autism spectrum disorder (ASD) as measured by variable importance in the projection (VIP) for the Riemannian partial least squares (R-PLS) model on the ABIDE dataset and the automated anatomic labelling (AAL) atlas with $K = 3$ latent variables, visualised as connectomes and symmetric matrices. Blue values represent connections that are positively associated with the phenotype, that is, an increase in connectivity between two regions with a blue edge would indicate an increase in the phenotype variable. Conversely, red values are connections that are negatively associated with the phenotype, that is, an increase in connectivity between two regions with a red edge would indicate a decrease in the phenotype variable. (**a**) Shows the connections that increase with age, (**b**) shows the connections that decrease with age, and (**c**) shows the average coefficient values for age between the 7 resting state networks identified by Parente and Colosimo[27] and the cerebellum (Figs. S8–S10). (**d**) Shows the connections that increase for patients with ASD, (**e**) shows the connections that decrease for patients with ASD, and (**f**) shows the average coefficient values for ASD between the 7 resting state networks identified by Parente and Colosimo[27] and the cerebellum (Figs. S8–S10). The darker outlined boxes in (**c**) and (**f**) show the top 25% influential regions as measured by the absolute coefficient value within and between each network. In (**c**) and (**f**), *DMN* default mode network.

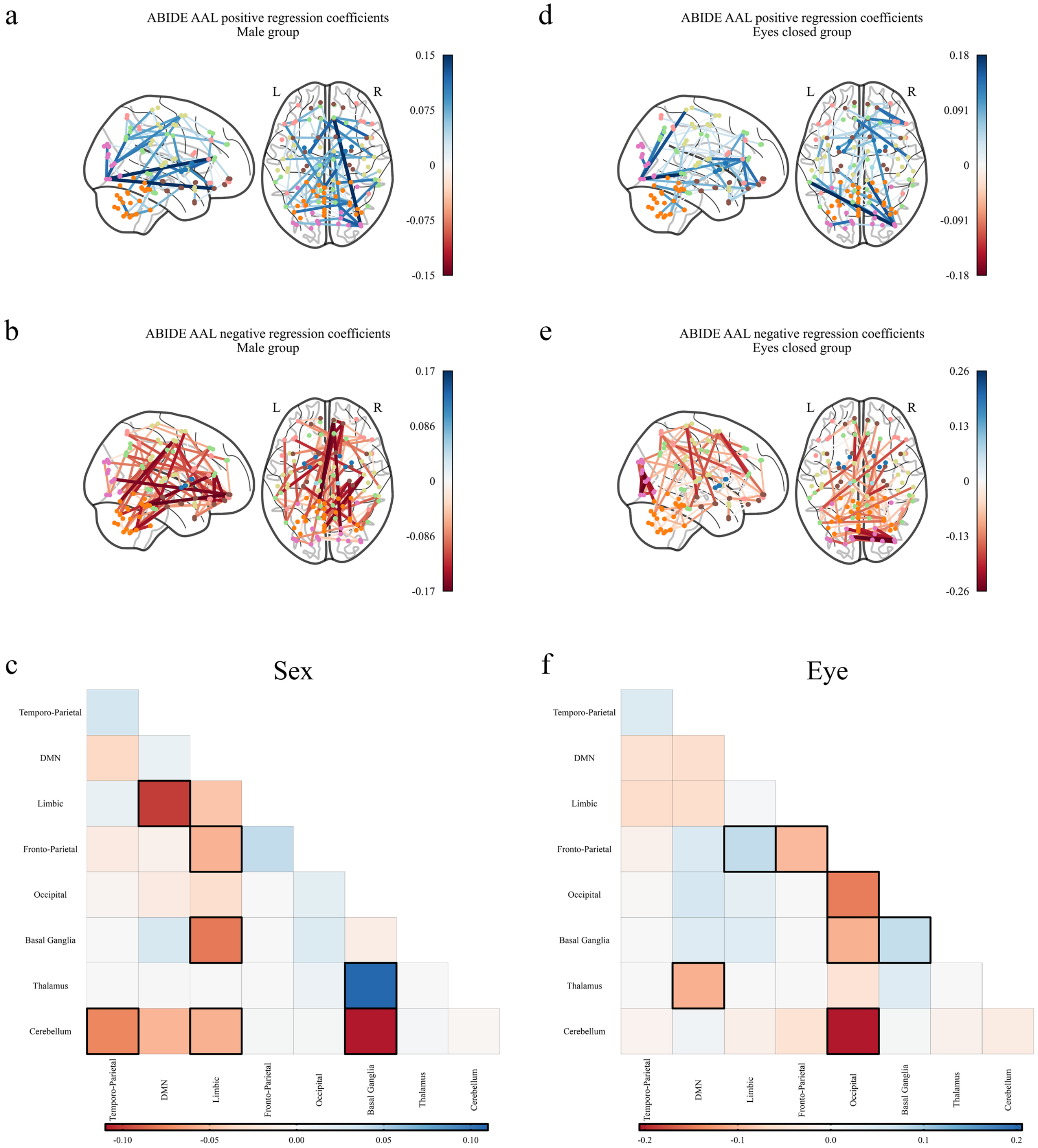

Figure 4. Significant regression coefficients for predicting sex and eye status as measured by variable importance in the projection (VIP) for the Riemannian partial least squares (R-PLS) model on the ABIDE dataset and the automated anatomic labelling (AAL) atlas with $K = 3$ latent variables, visualised as connectomes and symmetric matrices. Blue values represent connections that are positively associated with the phenotype, that is, an increase in connectivity between two regions with a blue edge would indicate an increase in the phenotype variable. Conversely, red values are connections that are negatively associated with the phenotype, that is, an increase in connectivity between two regions with a red edge would indicate a decrease in the phenotype variable. (**a**) Shows the connections that increase for males, (**b**) shows the connections that decrease for males, and (**c**) shows the average coefficient values for males between the 7 resting state networks identified by Parente and Colosimo[27] and the cerebellum (Figs. S8–S10). (**d**) Shows the connections that increase for patients with eyes open, (**e**) shows the connections that decrease for patients with eyes open, and (**f**) shows the average coefficient values for the eyes open group between the 7 resting state networks identified by Parente and Colosimo[27] and the cerebellum (Figs. S8–S10). The darker outlined boxes in (**c**) and (**f**) show the top 25% influential regions as measured by the absolute coefficient value within and between each network. In (**c**) and (**f**), *DMN* default mode network.

The results for eye status during scan are also well represented in the literature. The decreased within-network connectivity for the default mode network for patients with closed eyes has been previously reported by Yan et. al.[48], and the increased between-network connectivity for the default mode network has recently been discussed by Han et. al.[49]. Further, the observed decrease in connectivity for the occipital network agrees with Agcaoglu et. al.[50].

The use of the VIP statistic to identify significant connections in functional connectivity has not been previously studied. We have demonstrated that this statistic can identify many functional connections that have been addressed previously in the literature, but it is not without its limitations. First, with our focus on generalising partial least squares to Riemannian manifolds, the VIP statistic does not take into account the Riemannian geometry we are considering. This is mitigated by the tangent space approximation we are performing, which directly accounts for the geometry of the data, but further research could help better generalise the VIP statistic for R-PLS. Further, the VIP statistic associates the effects of a single predictor on the full multivariate response. In situations like we consider here, this makes it difficult to determine which functional connections are associated to which outcome variable. For example, the connectivity within the default mode network is deemed significant by the VIP statistic in the ABIDE dataset, but it is unclear whether this connectivity is significance for every outcome variable or a subset of them. Work has been done to generalise the VIP statistic when the outcome variable is multivariate[51], but further research is needed to investigate this generalisation.

The R-PLS method has shown to be generalisable over different atlases and datasets, but with a few notable differences. When analysing the COBRE dataset, we observe similar results between functional connectivity and age and subject group, but due to the different granularity of the MSDL and AAL atlases (39 ROI for MSDL and 116 for AAL) we are able to find further relationships. For example, the relationship between the within-network connectivity of the basal ganglia and age, or the role of the thalamus in schizophrenia. Comparing the AAL atlas across the COBRE and ABIDE datasets, we find that R-PLS is still the preferred model over Euclidean methods. Looking at the relationship between functional connectivity and age across both datasets, we see some similarities in the results with the differences explained by the different ages considered in each cohort[52] (mean age for COBRE = 37 years old, mean age for ABIDE = 15.2 years old, Tables S1 and S2).

However, further work is needed to verify R-PLS in a clinical context. First, we have not been able to investigate the test-retest reliability of this method due to the cross-sectional nature of the COBRE and ABIDE studies. A thorough study of the test-retest reliability of R-PLS would be invaluable to the method and would increase its versatility for clinical studies. Second, we have not investigated the effects of different preprocessing choices for the fMRI data. It would be beneficial to the generalisability of the results from R-PLS to determine how robust they are across preprocessing choices. The different preprocessing pipelines for the ABIDE study offer one avenue to investigate this, and is a clear area of future research.

These results suggest that R-PLS can provide insight into the functional connectome and how it relates to subject phenotype data. Further, due to the specification and generality of the R-PLS model, this method is readily applicable to other imaging modalities, and in particular to multimodal imaging studies. The application of R-PLS to multimodal imaging studies is an area of future research that may help to us to understand the functional networks that make up the human connectome.

Methods

Data

The International Neuroimaging Data-Sharing Initiative (INDI) is an initiative set to encourage free open access to neuroimaging datasets from around the world. We consider two datasets that are accessible as a part of the INDI.

COBRE

The Center for Biomedical Research Excellence (COBRE) have contributed structural and functional MRI images to the INDI that compare schizophrenic patients with healthy controls[19]. The data were collected with single-shot full k-space echo-planar imaging with a TR of 2000 ms, matrix size of 64×64 and 32 slices (giving a voxel size of $3 \times 3 \times 4$ mm^3). These data were downloaded using the PYTHON package `nilearn v 0.6.2`, and contains 146 subjects (Control = 74), each with phenotype information on subject group and age; further information is available in Table S1 of the supplementary material.

The fMRI data were preprocessed using NIAK 0.17 under CentOS version 6.3 with Octave version 4.0.2 and the Minc toolkit version 0.3.18[53]. The data were subjected to band pass filtering and nuisance regression where we 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[54].

For the COBRE dataset, we consider each fMRI in the MSDL atlas and the AAL atlas[22]. The MSDL atlas is a functional ROI decomposition of 39 nodes across 17 resting state networks[26]. The AAL atlas is an anatomical atlas of 116 nodes across the brain. Time series for each atlas were extracted for each ROI by taking the mean time series across the voxels in each region.

ABIDE

The Autism Brain Imaging Data Exchange (ABIDE) is part of the Preprocessed Connectomes Project in INDI[21]. The ABIDE data is a collection of preprocessed fMRI images from 16 international imaging sites with 539 individuals diagnosed with ASD and 573 neurotypical controls (NTC). The ABIDE initiative provides data preprocessed under four separate standard pipelines, as well as options for band-pass filtering and global signal regression.

Here we consider the 172 subjects (NTC = 98) of the New York University imaging site. We restrict to this site to reduce inter-site variation in imaging and because it is the largest individual imaging site. The data were collected with a 3 Tesla Allegra MRI using echo-planar imaging with a TR of 2000 ms, matrix size of 64 × 64 and 33 slices (giving a voxel size of $3 \times 3 \times 4$ mm ^{3}). The fMRI data were downloaded using the PYTHON package `nilearn v 0.6.2` preprocessed using the NIAK 0.7.1 pipeline[53]. The data were subjected to: motion realignment; non-uniformity correction using the median volume; motion scrubbing; nuisance regression which removed the first principal component of 6 motion parameters, their squares, mean white matter and cerebrospinal fluid signals, and low frequency drifts measured by a discrete cosine basis with a 0.01 Hz high-pass cut-off; band-pass filtering and; global signal regression. We consider the subjects preprocessed fMRI as well as subject group, age, sex, and eye status during scan (open or closed); further information is available in Table S2 of the supplementary material.

For the ABIDE dataset, we consider each fMRI in the AAL atlas[22], with time series were extracted by taking the mean time series across the voxels in each ROI.

Partial least squares in Euclidean space

PLS is a predictive modelling technique that predicts a response matrix $Y_{n\times q}$ from a set of predictors $X_{n\times p}$. Originally introduced in the chemometrics literature by Wold[6], PLS has found application in bioinformatics[55], social sciences[56], and neuroimaging[8,57,58]; see Rosipal and Krämer[59] and citations therein for further examples. As an extension of multivariate multiple regression, PLS has been shown to have better predictive accuracy than multivariate multiple regression when the standard regression assumptions are met[60]. A further advantage of PLS is that it is effective when $q > n$ or $p > n$ since it performs prediction from lower dimensional latent variables, that is, PLS constructs a new set of predictor variables from X to predict Y[60].

Let $X_{n\times p}$ and $Y_{n\times q}$ be predictor and response matrices respectively. Suppose X and Y are column centred, that is, suppose the means of each column of X and Y are 0. PLS proposes the existence of $L \leq \min\{p, n\}$ latent variables such that X and Y decompose into a set of *scores matrices* $T_{n\times L}$ and $U_{n\times L}$, and *loadings matrices* $P_{p\times L}$ and $Q_{q\times L}$ with

$$X = TP^T + E, \tag{1}$$

$$Y = UQ^T + F, \tag{2}$$

where $E_{n\times p}$ and $F_{n\times q}$ are error matrices, assumed to be a small as possible[61], and the superscript T denotes the matrix transpose. Further, PLS assumes that there is a diagonal matrix $B_{L\times L}$ with

$$U = TB + H_{n\times L}, \tag{3}$$

where H is a matrix of residuals. Equations (1) and (2) are called the *outer relationships* while Eq. (3) defines the *inner relationship* that connects X and Y. Combining the inner relationship and the outer relationship for Y gives

$$Y = TBQ^T + (HQ^T + F),$$

which highlights that Y is a regression on the latent scores T. Further, notice that the error in Y is given by $HQ^T + F$, that is, error in Y is a combination of error inherent to the response data (F) and error from the estimation of the inner relationship (HQ^T). The inclusion of the residual matrix H can complicate discussion of the PLS method, so it is common to consider the estimated inner relationship $\hat{U} \approx TB$ instead[61,62].

Estimation of the PLS model (Eqs. 1–3) is commonly done through the non-linear iterative partial least squares (NIPALS) algorithm (Algorithm S1 in the supplementary material). The inputs for the NIPALS algorithm are the data matrices X and Y and the pre-specified number of latent variables K; noting that the true number of latent variables L is unknown, the value K can be chosen with methods such as cross validation. The NIPALS algorithm outputs estimates of the scores, loadings, and regression coefficients as well as matrices $W_{p\times K}$ and $C_{q\times K}$ known as the weights. The weight matrices W and C are linear transformations of P and Q that more efficiently fit the PLS model and are defined within the NIPALS algorithm; see the supplementary material S1 for further information. Using the results of the NIPALS algorithm and Eqs. (1)–(3), we can write

$$\hat{Y} = X\hat{\beta}_{PLS}$$

where

$$\hat{\beta}_{PLS} = W(P^T W)^{-1} BC^T \tag{4}$$

is the matrix of regression coefficients. Using $\hat{\beta}_{PLS}$ we see that PLS is a linear regression technique similar to ordinary least squares and ridge regression.

Cross validation

We choose the optimal number of latent variables K for each PLS model through ten-fold cross validation[24]. To do this, we split each dataset into ten equal subsets $C_1, C_2 \ldots, C_{10}$ stratified by subject group (schizophrenia or ASD). For each subset $C_i, i = 1, 2, \ldots, 10$, we train the PLS models on the remaining nine subsets for each value of $\hat{K} = 1, 2, \ldots, 50$, using the phenotype data as the response variables and the functional connectivity as the predictors. We then predict on the subset C_i to calculate the test RMSE. By taking the average RMSE over all cross validation folds, we get an estimate of the test RMSE for the model. If K^* is the value of $\hat{K}$ that returns the minimum cross validated RMSE, the optimal K for our model is $K \leq K^*$ such that the cross validated RMSE for K is within one standard error of the cross validated RMSE for K^*.

The VIP statistic

To determine significant predictors of the response variables in the PLS model, we use the VIP statistic[18]. Suppose there are p predictor variables, q response variables, and K latent variables extracted using NIPALS. Following Tennenhaus[63], the VIP statistic for the jth predictor variable is

$$\mathrm{VIP_j} = \sqrt{\frac{\mathrm{p}}{\mathrm{Rd(Y,T)}} \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{Rd(Y,t_k)}\left(\mathrm{w_{jk}}\right)^2}, \tag{5}$$

where t_k is the k^{th} column of the score matrix T, w_{jk} is the kth weight for the jth predictor, $\mathrm{Rd(Y,t_k)} = \frac{1}{\mathrm{q}}\sum_{\mathrm{i=1}}^{\mathrm{q}} \mathrm{cor(Y_i,t_k)^2}$, and $\mathrm{Rd(Y,T)} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{Rd(Y,t_k)}$. The coefficient $\mathrm{cor(Y_i,t_k)^2}$ is the squared correlation between the jth response variable and the kth score. The denominator $\mathrm{Rd(Y,T)}$ in Eq. (5) measures the proportion of variance in Y explained by T, and the numerator $\mathrm{Rd(Y,t_k)(w_{jk})^2}$ measures the proportion of variance in Y described by the kth latent variable that is explained by the jth predictor[64]. Thus the VIP statistic measures the influence of each predictor on the explained variation in the model[65].

Commonly, the "greater than one" rule is used to find predictors significantly associated with the response. However, this rule is motivated by the mathematical properties of $\mathrm{VIP_j}$ rather than statistical properties[64]. Thus, we use a permutation test to determine significance of $\mathrm{VIP_j}$. This is an alternative to Afanador et. al.[66] who used 95% jackknife confidence intervals to determine significance of VIP .

Specifically, for each predictor variable j we permute the values H times. For each permutation $h = 1, 2, \ldots, H$ we refit the PLS model and calculate $\mathrm{VIP_{j,h}}$. The P-value for the jth VIP score is then

$$\text{P-value}_j = \frac{\#\left\{\mathrm{VIP_{j,h}} > \mathrm{VIP_j}\right\}}{H}. \tag{6}$$

For our data, the predictors are functional connectivity matrices. Thus, we know a priori that the diagonal elements are uninformative since they are identically one. Hence, if predictor j describes a diagonal element we set $P\text{-value}_j = 1$ for all i. To account for the multiple comparisons problem, we adjust all P-values using the false discovery rate[67] and determine significance at a significance level of $\alpha = 0.05$.

Mathematical preliminaries

Riemannian manifolds

Intuitively speaking, a Riemannian manifold M is a space where we can perform calculus, measure distances, and measure angles between tangent vectors. More specifically, a smooth d-dimensional manifold M is a connected, Hausdorff, second countable topological space that is covered by a set of coordinate charts $\{(U_i, \varphi_i : U_i \to \mathbb{R}^d)\}_{i\in I}$, defined by some indexing set I, such that every point in M belongs to a U_i for some $i \in I$ and the intersection maps $\varphi_i \circ \varphi_j^{-1}$ are smooth as maps $\mathbb{R}^d \to \mathbb{R}^d$ for every $i, j \in I$. These coordinate charts make the space M "locally Euclidean" in the sense that every point has a neighbourhood that looks like Euclidean space. Since concepts from differential calculus are local in nature, the construction of a smooth manifold allows us to perform calculus on these more general spaces.

An important concept in the study of manifolds is the tangent bundle $TM = \bigsqcup_{a\in M} T_aM$, where T_aM is the tangent space at a. The space T_aM is defined as the set of equivalence classes of curves through a such that γ_1 and γ_2 are equivalent if $\gamma_1'(0) = \gamma_2'(0)$, where the prime denotes the derivative. Then T_aM is a vector space that generalises the notion of vectors tangent to a surface to arbitrary smooth manifolds.

A *Riemannian* manifold is a manifold M together with a smooth map $g : M \times TM \times TM \to \mathbb{R}$ such that $g(a, \cdot, \cdot) = g_a : T_aM \times T_aM \to \mathbb{R}$ is an inner product for every $a \in M$. The Riemannian metric g allows us to measure angles between tangent vectors and measure distances between points on the manifold M. Further, g is used to define geodesics (locally length minimising curves) $\gamma : [t_0, t_1] \to M$ between two points $a, b \in M$. We only consider complete Riemannian manifolds here, which are spaces where every geodesic γ has domain $\mathbb{R}$.

Through geodesics we get the concepts of the Riemannian exponential and logarithm maps which allow us to smoothly move between the manifold and the tangent space. The Riemannian exponential at a point $a \in M$ is a map $\mathrm{Exp}_a : T_aM \to M$ defined by $\mathrm{Exp}(a, \cdot)(\gamma) = \mathrm{Exp}_a(\gamma) = \gamma(1)$, where γ is a geodesic such that $\gamma(0) = a$. The Riemannian exponential is a smooth map that is locally diffeomorphic and hence has a local inverse denoted $\mathrm{Log}(a, \cdot) = \mathrm{Log}_a : M \to T_aM$ defined by $\mathrm{Log}_a(b) = \gamma'(0)$ where $\gamma(t)$ is a geodesic from a to b. For a point $b \in M$ close to a, we think of $\mathrm{Log}_a(b)$ as the shortest initial velocity vector based at a pointing in the direction of b. Further information on Riemannian manifolds can be found in the books by Lee[68–70] or do Carmo (1992)[71]. An accessible introduction for medical imaging can be found in the book edited by Pennec et. al.[10].

Fréchet mean

To capture the centre of data on a manifold we consider the Fréchet (or intrinsic) mean of data $X_1, X_2, \ldots, X_n \in M$. First, consider the Riemannian distance between two close points $X_1, X_2 \in M$ defined by

$$d_g(X_1, X_2) = \left\| \mathrm{Log}_{X_1}(X_2) \right\|,$$

where $\|\cdot\|$ is the norm in $T_{X_1}M$ induced by the Riemannian metric. By generalising the sum of squared distances definition of the arithmetic mean, the Fréchet mean[72] is given by

$$\mu_X = \operatorname{argmin} \sum_{i=1}^{n} d_g(X_i, \mu_X)^2 \,.$$

We solve for μ_X using gradient decent[10]; see Algorithm S2 in the supplementary material for further information.

The affine invariant geometry for symmetric positive definite matrices

Let $GL_R\mathbb{R}$ be the set of $R \times R$ real invertible matrices. The set of symmetric positive definite matrices is defined by

$$S_R^+ = \left\{ A \in GL_R\mathbb{R} : A^T = A \text{ and } v^T A v > 0 \text{ for all } v \in \mathbb{R}^R \backslash \{0\} \right\},$$

where superscript T denotes matrix transpose. The set S_R^+ is a smooth manifold, which can be easily seen by embedding S_R^+ into $\mathbb{R}^{R(R+1)/2}$ as a convex cone. This construction shows that the tangent space at each $A \in S_R^+$ is given by the set of symmetric $R \times R$ matrices.

However, S_R^+ has an interesting intrinsic geometry known as the affine-invariant geometry[9]. Under the affine invariant geometry S_R^+ becomes a complete Hadamard manifold—a Riemannian manifold of non-positive curvature where Exp_A is a diffeomorphism for every $A \in S_R^+$.

The affine-invariant metric g is defined by

$$g_A(U, V) = \operatorname{Tr}\left(UA^{-1}VA^{-1}\right),$$

where $A \in S_R^+$, $U, V \in T_A S_R^+$, and Tr denotes the trace operator. Using g, we can calculate the Riemannian distance between $A, B \in S_R^+$ as

$$d_g(A, B)^2 = \sum_{r=1}^{R} \left(\log\left(\sigma_r\left(A^{-1/2}BA^{-1/2}\right)\right)\right)^2,$$

where $\sigma_r\left(A^{-1/2}BA^{-1/2}\right)$ are the eigenvalues of $A^{-1/2}BA^{-1/2}$, $r = 1, 2, \ldots, R$. Further, letting $A, B \in S_R^+$ and $U \in T_A S_R^+$, we get

$$\operatorname{Exp}_A(U) = A^{1/2} \operatorname{Exp}\left(A^{-1/2}UA^{-1/2}\right)A^{1/2}$$

and

$$\operatorname{Log}_A(B) = A^{1/2} \operatorname{Log}\left(A^{-1/2}BA^{-1/2}\right)A^{1/2},$$

where Exp and Log are the matrix exponential and logarithm respectively. The Riemannian distance, exponential, and logarithm are essential in the definition and fitting of the R-PLS model defined below.

Riemannian PLS

Let M and N be complete Riemannian manifolds. Let $X_1, X_2, \ldots, X_n \in M$ and $Y_1, Y_2, \ldots, Y_n \in N$, and let μ_X and μ_Y denote the respective Fréchet means. Let $L \leq \min\{\dim(M), n\}$. The R-PLS model proposes the existence of loadings $p_1, p_2, \ldots, p_L \in T_{\mu_X}M$ and $q_1, q_2, \ldots, q_L \in T_{\mu_Y}N$ such that, for each subject $i = 1, 2, \ldots, n$, there are scores $t_{i1}, t_{i2}, \ldots, t_{iL} \in \mathbb{R}$ and $u_{i1}, u_{i2}, \ldots, u_{iL} \in \mathbb{R}$ with

$$X_i = \operatorname{Exp}\left(\operatorname{Exp}_{\mu_X}\left(\sum_{l=1}^{L} t_{il} p_l \right), e_i \right), \tag{7}$$

$$Y_i = \operatorname{Exp}\left(\operatorname{Exp}_{\mu_Y}\left(\sum_{l=1}^{L} u_{il} q_l \right), f_i \right), \text{ and} \tag{8}$$

$$\hat{u}_{il} = \hat{\beta}_{0l} + \hat{\beta}_{1l} t_{il} \text{ for all } l = 1, 2, \ldots, L \text{ and } i = 1, 2, \ldots, n, \tag{9}$$

where $e_i \in T_{\operatorname{Exp}_{\mu_X}\left(\sum_{l=1}^{L} t_{il} p_l\right)} M$ and $f_i \in T_{\operatorname{Exp}_{\mu_Y}\left(\sum_{l=1}^{L} u_{il} q_l\right)} M$ are error vectors with $\|e_i\|, \|f_i\|$ small. Equations (7) and (8) are the *outer relationships* for Riemannian data, and Eq. (9) is the *inner relationship* connecting our response and predictor. Note that, since the Riemannian exponential map on Euclidean space is vector addition, if $M = \mathbb{R}^p$ and $N = \mathbb{R}^q$ the R-PLS model (Eqs. 7–9) reduce to the standard PLS model (Eqs. 1–3).

One approach to fitting R-PLS is by directly generalising NIPALS (Algorithm S1) to Riemannian manifolds, but this becomes computationally intensive and fails to converge for sample sizes above 20 (see Ryan[51] for more details). Instead, we propose a tangent space approximation to fitting R-PLS when our data is close to the Fréchet mean, similar to methods such as Riemannian canonical correlations analysis[13] and principal geodesic analysis[11].

The tNIPALS algorithm (Algorithm 1) works by first linearising the manifold data in a neighbourhood of the Fréchet mean using the Riemannian logarithm (see supplementary material S1 for further information), and then applying the Euclidean NIPALS algorithm to the linearised data which is now vector-valued. Thus, tNIPALS provides a combination of the simplicity and efficiency of Euclidean NIPALS with the geometry of the Riemannian manifold.

```
Input: Data $X_1, X_2, \dots, X_n$, $Y_1, Y_2, \dots, Y_n$, Desired number of components $K$.
Output: PLS weights $\{w_k\}_{k=1}^K, \{c_k\}_{k=1}^K$, Scores $\{t_k\}_{k=1}^K, \{u_k\}_{k=1}^K$, Loadings $\{p_k\}_{k=1}^K$, and Regression coefficients
        $\{\hat{\beta}_{1k}\}_{k=1}^K$.
Calculate Fréchet means $\mu_X, \mu_Y$ (Algorithm S2*);
Linearise the data by;
    $x_i \leftarrow \mathrm{Log}_{\mu_X} X_i$;
    $y_i \leftarrow \mathrm{Log}_{\mu_Y} Y_i$;
Map $x_i$, $y_i$ to Euclidean space via coordinates $\phi$ on $T_{\mu_X} M$ and $\psi$ on $T_{\mu_Y} M$;
Perform NIPALS (Algorithm S1*) on $\{(x_i, y_i)\}$ to get weights $\{w_k\}_{k=1}^K, \{c_k\}_{k=1}^K$, scores $\{t_k\}_{k=1}^K, \{u_k\}_{k=1}^K$, loadings
  $\{p_k\}_{k=1}^K$, and regression coefficients $\{\hat{\beta}_{1k}\}_{k=1}^K$;
Map $w_k, c_k$ and $p_k$ back to their appropriate tangent spaces using $\phi^{-1}$ and $\psi^{-1}$.
```

*Found in the supplementary material.

Algorithm 1: Tangent non-linear iterative partial least squares.

The tNIPALS algorithm provides a more general approach to Wong et. al.'s[14] method for constructing predictors from functional connectivity matrices to predict ASD using PLS and logistic regression by considering a Euclidean response and symmetric positive definite predictor. Similarly, Zhang and Liu[16] and Chu et al.[15] also proposed PLS methods using the affine-invariant geometry for symmetric positive definite matrices that is generalised by tNIPALS. Further, the tNIPALS algorithm for R-PLS is closely related to the PLS method for symmetric positive definite matrices offered by Perez and Gonzalez-Farias[17], where they also propose linearising symmetric positive definite matrices in the affine-invariant geometry to fit the PLS model.

Data availibility

The data and R package (`spdMatrices`) used to complete this work are available on GitHub (Matthew-Ryan1995/Riemannian-statistical-techniques-with-applications-in-fMRI). The code to perform the analyses and generate the figures is also found on GitHub (Matthew-Ryan1995/R-PLS-for-functional-connectivity).

Received: 11 July 2023; Accepted: 10 October 2023
Published online: 13 October 2023

References

1. Contreras, J. A., Goñi, J., Risacher, S. L., Sporns, O. & Saykin, A. J. The structural and functional connectome and prediction of risk for cognitive impairment in older adults. *Curr. Behav. Neurosci. Rep.* **2**, 234–245 (2015).
2. Yang, F. N., Liu, T. T. & Wang, Z. Functional connectome mediates the association between sleep disturbance and mental health in preadolescence: a longitudinal mediation study. *Hum. Brain Mapp.* **43**, 2041–2050 (2022).
3. Woodward, N. D. & Cascio, C. J. Resting-state functional connectivity in psychiatric disorders. *J. Am. Med. Assoc. Psychiatry* **72**, 743–744. https://doi.org/10.1001/JAMAPSYCHIATRY.2015.0484 (2015).
4. Shi, Y. & Toga, A. W. Connectome imaging for mapping human brain pathways. *Mol. Psychiatry* **22**, 1230–1240 (2017).
5. Ogawa, S., Lee, T. M., Kay, A. R. & Tank, D. W. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 9868–9872 (1990).
6. Wold, H. Soft modelling by latent Variables: The non-linear iterative partial least squares (NIPALS) approach. *J. Appl. Probab.* **12**, 117–142. https://doi.org/10.1017/S0021900200047604 (1975).
7. McIntosh, A. R., Bookstein, F. L., Haxby, J. V. & Grady, C. L. Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage* **3**, 143–157. https://doi.org/10.1006/NIMG.1996.0016 (1996).
8. Krishnan, A., Williams, L. J., McIntosh, A. R. & Abdi, H. Partial least squares (PLS) methods for neuroimaging: A tutorial and review. *Neuroimage* **56**, 455–475. https://doi.org/10.1016/j.neuroimage.2010.07.034 (2011).
9. Pennec, X., Fillard, P. & Ayache, N. A Riemannian framework for tensor computing. *Int. J. Comput. Vis.* **66**, 41–66. https://doi.org/10.1007/s11263-005-3222-z (2006).
10. Pennec, X., Sommer, S. & Fletcher, T. *Riemannian Geometric Statistics in Medical Image Analysis* (Elsevier, 2019).
11. Fletcher, P. T. Geodesic regression and the theory of least squares on Riemannian manifolds. *Int. J. Comput. Vis.* **105**, 171–185. https://doi.org/10.1007/s11263-012-0591-y (2013).
12. Fletcher, P. T., Lu, C. & Joshi, S. *Statistics of shape via principal geodesic analysis on lie groups.* **1**, 95–101. https://doi.org/10.1109/cvpr.2003.1211342 (2003).
13. Kim, H. J. *et al.* Canonical correlation analysis on Riemannian manifolds and its applications. 251–267, https://doi.org/10.1007/978-3-319-10605-2_17 (Springer, 2014).
14. Wong, E., Anderson, J. S., Zielinski, B. A. & Fletcher, P. T. Riemannian Regression and Classification Models of Brain Networks Applied to autism. In *Connectomics in neuroImaging: second international workshop, CNI 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018 : proceedings. CNI (Workshop) (2nd : 2018 : Granada, Spain)***11083**, 78. https://doi.org/10.1007/978-3-030-00755-3_9 (2018).
15. Chu, Y. *et al.* Decoding multiclass motor imagery EEG from the same upper limb by combining Riemannian geometry features and partial least squares regression. *J. Neural Eng.* **17**, 046029. https://doi.org/10.1088/1741-2552/ABA7CD (2020).
16. Zhang, C. & Liu, Q. Region Constraint Person Re-Identification via Partial Least Square on Riemannian Manifold. *IEEE Access* **6**, 17060–17066. https://doi.org/10.1109/ACCESS.2018.2808602 (2018).
17. Perez, R. A. & Gonzalez-Farias, G. Partial Least Squares Regression on Symmetric Positive-Definite Matrices. *Revista Colombiana de Estadistica* **36**, 177–192 (2013).
18. Wold, S., Johansson, E. & Cocchi, M. PLS: partial least squares projections to latent structures. *3D QSAR Drug Des.* 523–550 (1993).
19. Aine, C. J. *et al.* Multimodal neuroimaging in Schizophrenia: Description and dissemination. *Neuroinformatics* **15**, 343–364. https://doi.org/10.1007/s12021-017-9338-9 (2017).

20. Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P. & Thirion, B. LNCS 6361—Detection of Brain Functional-Connectivity Difference in Post-stroke Patients Using Group-Level Covariance Modeling (2010).
21. Craddock, C. *et al.* The Neuro Bureau Preprocessing Initiative: open sharing of preprocessed neuroimaging data and derivatives. *Front. Neuroinf.* **7**. https://doi.org/10.3389/CONF.FNINF.2013.09.00041/EVENT_ABSTRACT (2013).
22. Tzourio-Mazoyer, N. *et al.* Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**, 273–289. https://doi.org/10.1006/nimg.2001.0978 (2002).
23. Venkatesh, M., Jaja, J. & Pessoa, L. Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification. *Neuroimage* **207**, 116398. https://doi.org/10.1016/j.neuroimage.2019.116398 (2020).
24. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, New York, 2009).
25. R Core Team. R: A Language and Environment for Statistical Computing (2022).
26. Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V. & Thirion, B. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. vol. 6801 LNCS, 562–573, https://doi.org/10.1007/978-3-642-22092-0_46 (Springer, 2011).
27. Parente, F. & Colosimo, A. Functional connections between and within brain subnetworks under resting-state. *Sci. Rep.* **10**, 1–13. https://doi.org/10.1038/s41598-020-60406-7 (2020).
28. Varangis, E., Habeck, C. G., Razlighi, Q. R. & Stern, Y. The effect of aging on resting state connectivity of predefined networks in the brain. *Front. Aging Neurosci.* **11**, 234. https://doi.org/10.3389/FNAGI.2019.00234/BIBTEX (2019).
29. Edde, M., Leroux, G., Altena, E. & Chanraud, S. Functional brain connectivity changes across the human life span: From fetal development to old age. *J. Neurosci. Res.* **99**, 236–262. https://doi.org/10.1002/JNR.24669 (2021).
30. Ferreira, R., Xavier, J., Costeira, J. P. & Barroso, V. Newton method for riemannian centroid computation in naturally reductive homogeneous spaces. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* **3**. https://doi.org/10.1109/ICASSP.2006.1660751 (2006).
31. Tomasi, D. & Volkow, N. D. Aging and functional brain networks. *Mol. Psychiatry* **17**, 549–558. https://doi.org/10.1038/mp.2011.81 (2011).
32. Vidal-Piñiro, D. *et al.* Decreased Default Mode Network connectivity correlates with age-associated structural and cognitive changes. *Front. Aging Neurosci.* **6**, 256. https://doi.org/10.3389/FNAGI.2014.00256/BIBTEX (2014).
33. Rodriguez-Sabate, C., Morales, I. & Rodriguez, M. The influence of aging on the functional connectivity of the human basal ganglia. *Front. Aging Neurosci.* **13**, 785666 (2022).
34. Ramos, T. C., Balardin, J. B., Sato, J. R. & Fujita, A. Abnormal cortico-cerebellar functional connectivity in autism spectrum disorder. *Front. Syst. Neurosci.* **12**, 74 (2019).
35. Pascual-Belda, A., Díaz-Parra, A. & Moratal, D. Evaluating functional connectivity alterations in autism spectrum disorder using network-based statistics. *Diagnostics* **8**, 51 (2018).
36. Assaf, M. *et al.* Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients. *Neuroimage* **53**, 247–256. https://doi.org/10.1016/J.NEUROIMAGE.2010.05.067 (2010).
37. Smith, R. E. *et al.* Sex differences in resting-state functional connectivity of the cerebellum in autism spectrum disorder. *Front. Hum. Neurosci.* **13**, 104. https://doi.org/10.3389/FNHUM.2019.00104/BIBTEX (2019).
38. Zhang, B. *et al.* Altered Functional Connectivity of Striatum Based on the Integrated Connectivity Model in First-Episode Schizophrenia. *Front. Psych.* **10**, 756. https://doi.org/10.3389/FPSYT.2019.00756/BIBTEX (2019).
39. Orliac, F. *et al.* Links among resting-state default-mode network, salience network, and symptomatology in schizophrenia. *Schizophr. Res.* **148**, 74–80. https://doi.org/10.1016/J.SCHRES.2013.05.007 (2013).
40. Duan, M. *et al.* Altered basal ganglia network integration in schizophrenia. *Front. Hum. Neurosci.* **9**, 561. https://doi.org/10.3389/FNHUM.2015.00561/BIBTEX (2015).
41. Andreasen, N. C. The role of the thalamus in schizophrenia. *Can. J. Psychiatry* **42**, 27–33 (1997).
42. Pergola, G., Selvaggi, P., Trizio, S., Bertolino, A. & Blasi, G. The role of the thalamus in schizophrenia from a neuroimaging perspective. *Neurosci. Biobehav. Rev.* **54**, 57–75 (2015).
43. Karbasforoushan, H. & Woodward, N. Resting-state networks in schizophrenia. *Curr. Top. Med. Chem.* **12**, 2404–2414. https://doi.org/10.2174/156802612805289863 (2012).
44. Woodward, N. D., Rogers, B. & Heckers, S. Functional resting-state networks are differentially affected in schizophrenia. *Schizophr. Res.* **130**, 86–93. https://doi.org/10.1016/J.SCHRES.2011.03.010 (2011).
45. Dong, D., Wang, Y., Chang, X., Luo, C. & Yao, D. Dysfunction of large-scale brain networks in schizophrenia: A meta-analysis of resting-state functional connectivity. *Schizophr. Bull.* **44**, 168–181. https://doi.org/10.1093/SCHBUL/SBX034 (2018).
46. Yu, Q. *et al.* Brain connectivity networks in schizophrenia underlying resting state functional magnetic resonance imaging. *Curr. Top. Med. Chem.* **12**, 2415. https://doi.org/10.2174/156802612805289890 (2012).
47. Wang, H. *et al.* Evidence of a dissociation pattern in default mode subnetwork functional connectivity in schizophrenia. *Sci. Rep.* **5**, 14655 (2015).
48. Yan, C. *et al.* Spontaneous brain activity in the default mode network is sensitive to different resting-state conditions with limited cognitive load. *PLoS ONE* **4**, e5743. https://doi.org/10.1371/JOURNAL.PONE.0005743 (2009).
49. Han, J. *et al.* Eyes-open and eyes-closed resting state network connectivity differences. *Brain Sci.* **13**, 122 (2023).
50. Agcaoglu, O., Wilson, T. W., Wang, Y. P., Stephen, J. & Calhoun, V. D. Resting state connectivity differences in eyes open versus eyes closed conditions. *Hum. Brain Mapp.* **40**, 2488. https://doi.org/10.1002/HBM.24539 (2019).
51. Ryan, M. *Riemannian statistical techniques with applications in fMRI*. Ph.D. thesis, The University of Adelaide (2023).
52. Grady, C., Sarraf, S., Saverino, C. & Campbell, K. Age differences in the functional interactions among the default, frontoparietal control, and dorsal attention networks. *Neurobiol. Aging* **41**, 159–172 (2016).
53. Bellec, P. *et al.* A neuroimaging analyses kit for Matlab and Octave. 1–5 (Organization on Human Brain Mapping, 2011).
54. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**, 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042 (2007).
55. Nguyen, D. V. & Rocke, D. M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**, 39–50. https://doi.org/10.1093/BIOINFORMATICS/18.1.39 (2002).
56. Hulland, J. Use of partial least squares (PLS) in strategic management research: A review of four recent studies. *Strateg. Manag. J.* **20**, 195–204. https://doi.org/10.1002/(SICI)1097-0266(199902)20:2 (1999).
57. McIntosh, A. R. & Lobaugh, N. J. Partial least squares analysis of neuroimaging data: Applications and advances. *Neuroimage* **23**, S250–S263. https://doi.org/10.1016/J.NEUROIMAGE.2004.07.020 (2004).
58. Lin, F. H. *et al.* Multivariate analysis of neuronal interactions in the generalized partial least squares framework: simulations and empirical studies. *Neuroimage* **20**, 625–642. https://doi.org/10.1016/S1053-8119(03)00333-1 (2003).
59. Rosipal, R. & Krämer, N. Overview and Recent Advances in Partial Least Squares. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)***3940 LNCS**, 34–51, https://doi.org/10.1007/11752790_2 (2006).
60. Garthwaite, P. H. An interpretation of partial least squares. *J. Am. Stat. Assoc.* **89**, 122–127. https://doi.org/10.1080/01621459.1994.10476452 (1994).
61. Geladi, P. & Kowalski, B. R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **185**, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9 (1986).
62. Höskuldsson, A. PLS regression methods. *J. Chemom.* **2**, 211–228. https://doi.org/10.1002/CEM.1180020306 (1988).

63. Tenenhaus, M. *La régression PLS: Théorie et pratique* (Technip, 1998).
64. Tran, T. N., Afanador, N. L., Buydens, L. M. & Blanchet, L. Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). *Chemom. Intell. Lab. Syst.* **138**, 153–160. https://doi.org/10.1016/J.CHEMOLAB.2014.08.005 (2014).
65. Galindo-Prieto, B., Eriksson, L. & Trygg, J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *J. Chemom.* **28**, 623–632. https://doi.org/10.1002/cem.2627 (2014).
66. Afanador, N. L., Tran, T. N. & Buydens, L. M. Use of the bootstrap and permutation methods for a more robust variable importance in the projection metric for partial least squares regression. *Anal. Chim. Acta* **768**, 49–56. https://doi.org/10.1016/J.ACA.2013.01.004 (2013).
67. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).
68. Lee, J. M. *Introduction to Topological Manifolds* Vol. 202 (Springer, New York, 2011).
69. Lee, J. M. *Introduction to Smooth Manifolds* Vol. 218 (Springer, New York, 2012).
70. Lee, J. M. *Introduction to Riemannian Manifolds*, vol. 176 (Springer International Publishing, 2018).
71. do Carmo, M. P. *Riemannian Geometry* (Birkhauser Boston Inc, 1992).
72. Fréchet, M. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré* **10**, 215–310 (1948).

Acknowledgements

M.R. acknowledges the Australian Government Research Training Program Scholarship that funded this research.

Author contributions

M.R. developed the methods and analysed the data with consultation from G.G., J.T. and M.H. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-44687-2.

Correspondence and requests for materials should be addressed to M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023