The effect of outliers and their exclusion on resting-state connectivity-based parcellation

Niels Reuter[1,2], Sarah Genon[2], Shahrzad Kharabian[2], Tobias Kalenscher[3], Felix Hoffstaedter[1,2], Rainer Goebel[4,5,6], Simon Eickhoff[1,2] & Kaustubh Patil[1,2]

[1]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany;
[2]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany
[3]Department of Comparative Psychology, Institute of Experimental Psychology, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany
[4]Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6201BC Maastricht, The Netherlands
[5]Maastricht Brain imaging Centre, Faculty of Psychology and Neuroscience, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands
[6]Department of Neuroimaging and Neuromodeling, Netherlands Institute for Neuroscience, an Institute of the Royal Netherlands Academy of Arts and Sciences (KNAW), 1105BA Amsterdam, The Netherlands

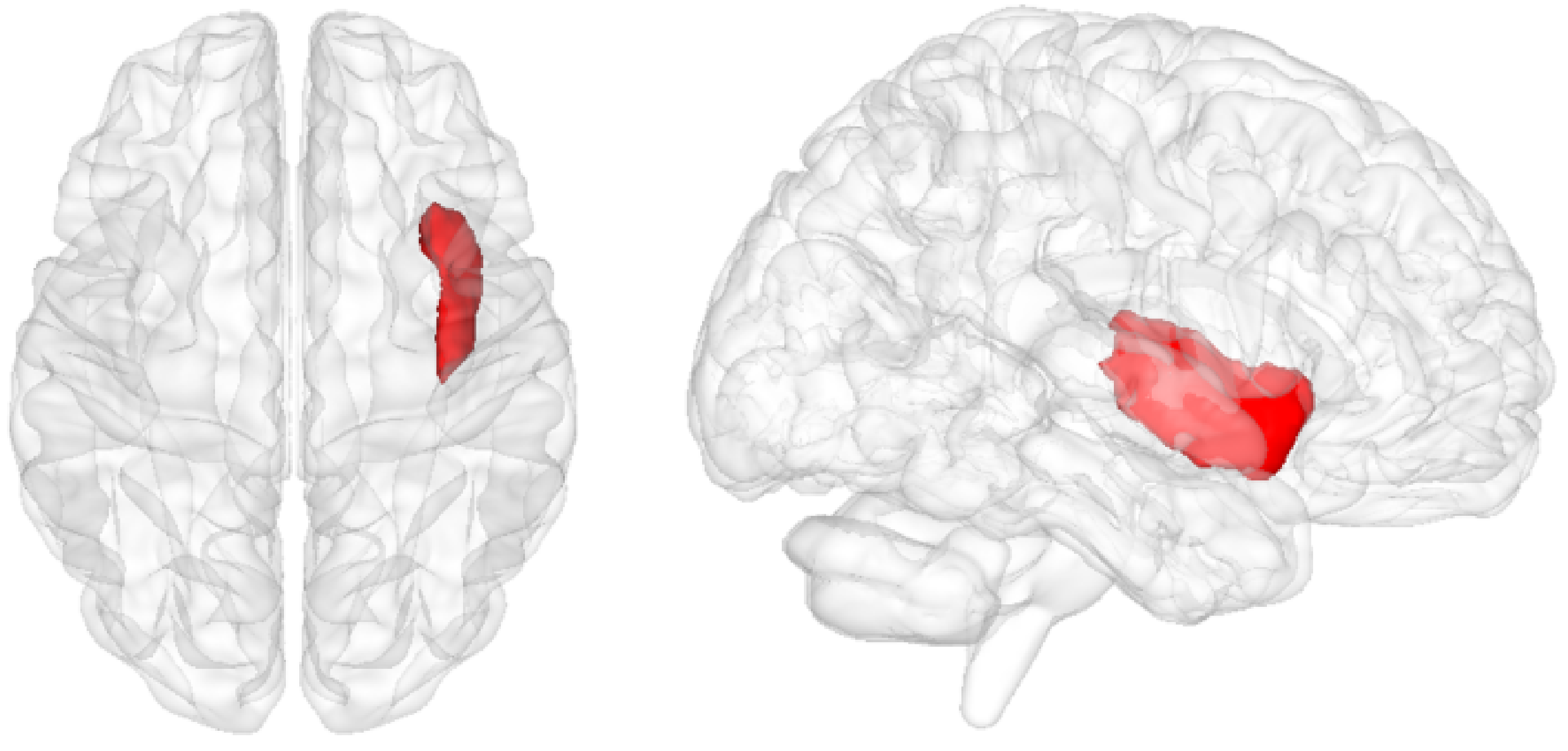
n.reuter@fz-juelich.de

Introduction

Regional connectivity-based parcellation (CBP) aims to find biologically meaningful subregions by clustering voxels of a region of interest (ROI).

Using a **large resting-state fMRI (rs-fMRI) sample**, we show that deviant connectivity profiles substantially influence group-based clustering results on the well researched [1] right (R) insula ROI (***Fig. 1***) as defined by the Harvard-Oxford Atlas [2].

1 Region-of-Interest for parcellation



R Insula, Superior View

R Insula, Right Sagittal View (used in Fig. 6)

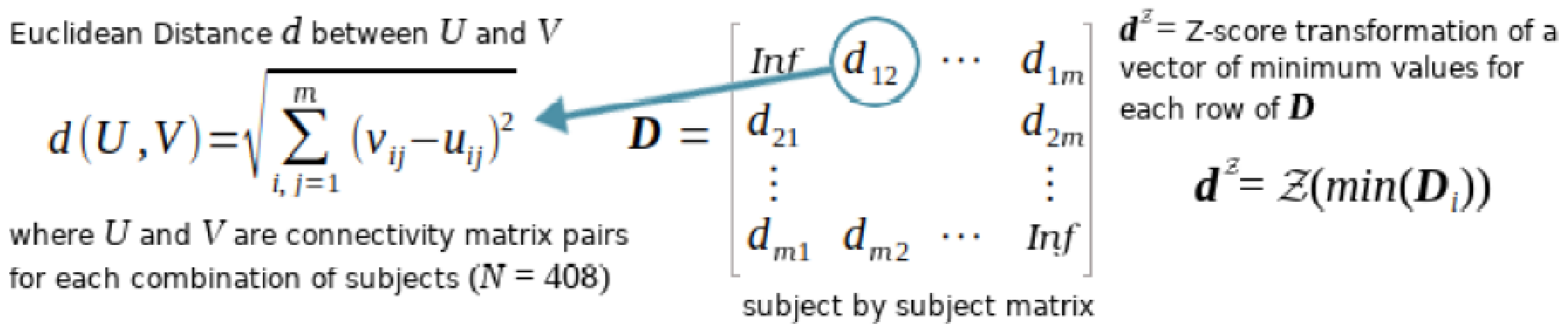
Methods

Sample: rs-fMRI data of 408 healthy unrelated subjects from the *Human Connectome Project* [3]

Connectivity: Correlations between time-series of each ROI voxel and all brain gray-matter voxels

Deviant Detection: Identify nearest-neighbor subjects based on Euclidean distance (***Fig. 2***). Three detection thresholds defined as (1) a *conservative* k-means (k = 2) cluster-defined threshold (***Fig. 4***), (2) a *standard* 1.69 (.95 left tail area on standard normal distribution), and (3) a *liberal* 2.5 on Z-scored distances

2



Euclidean Distance d between U and V

$$d(U,V)=\sqrt{\sum_{i,j=1}^{m}(v_{ij}-u_{ij})^2}$$

where U and V are connectivity matrix pairs for each combination of subjects ($N = 408$)

$$\mathbf{D} = \begin{bmatrix} Inf & d_{12} & \cdots & d_{1m} \\ d_{21} & & & d_{2m} \\ \vdots & & & \vdots \\ d_{m1} & d_{m2} & \cdots & Inf \end{bmatrix}$$

subject by subject matrix

$\mathbf{d}^z$ = Z-score transformation of a vector of minimum values for each row of $\mathbf{D}$

$$\mathbf{d}^z = \mathcal{Z}(min(\mathbf{D}_i))$$

Clustering: subject-wise k-means (k = 2 to 5) on each connectivity matrix; hierarchical clustering with average linkage and Hamming distance for group clustering

Analysis: Adjusted rand index (ARI) between all subject k-means cluster results retaining highest values per subject (***Fig. 3***). Principal component analysis on connectivity matrices noting principal component numbers (PC_n) retaining 95% variance

3

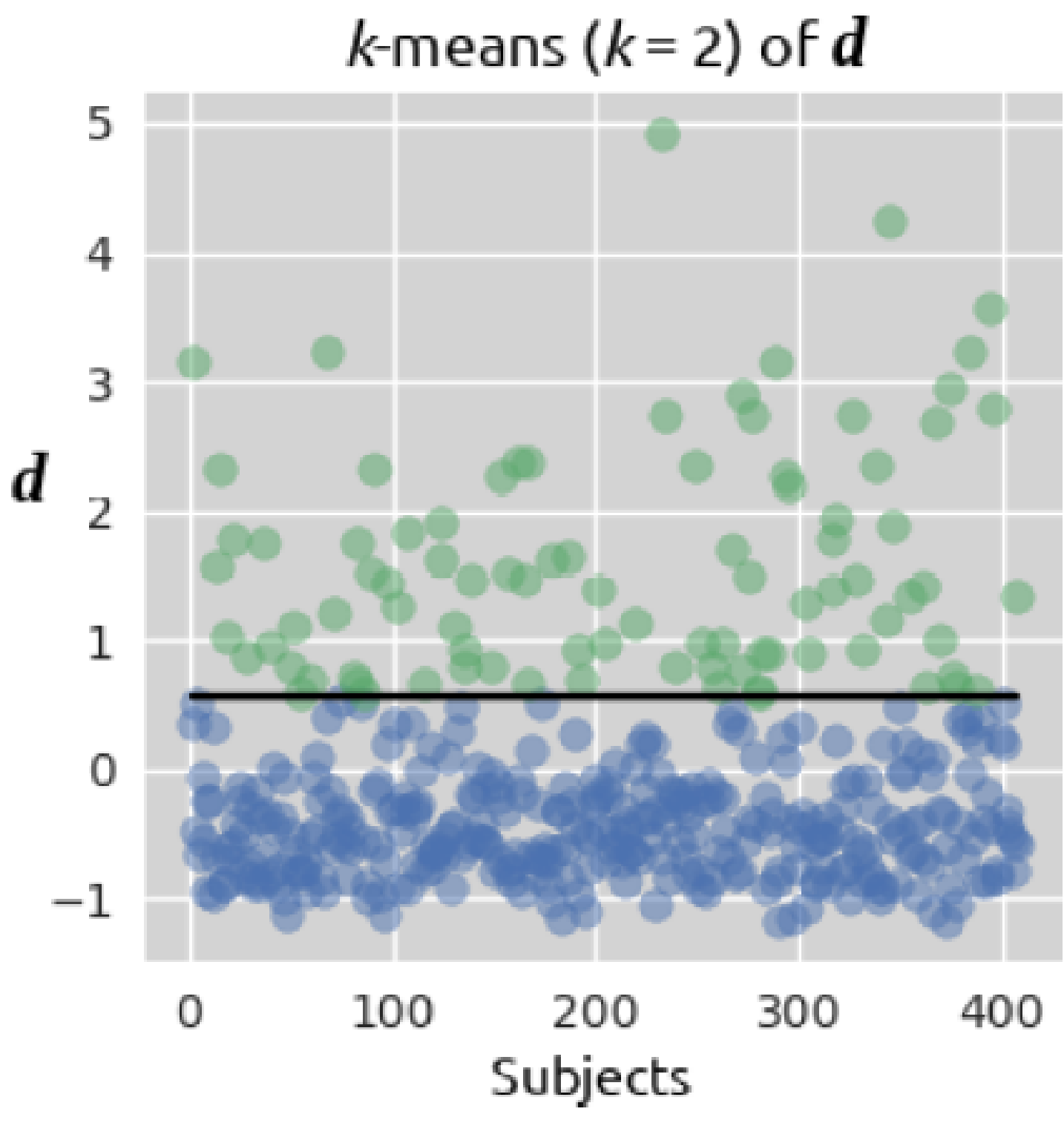
$\mathbf{A}$ = Adjusted Rand Index between pairs of clusterings for each combination of subjects ($N = 408$)

$$\mathbf{A} = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & & & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix} \qquad \mathbf{a} = \mathcal{Z}(max(A_i))$$

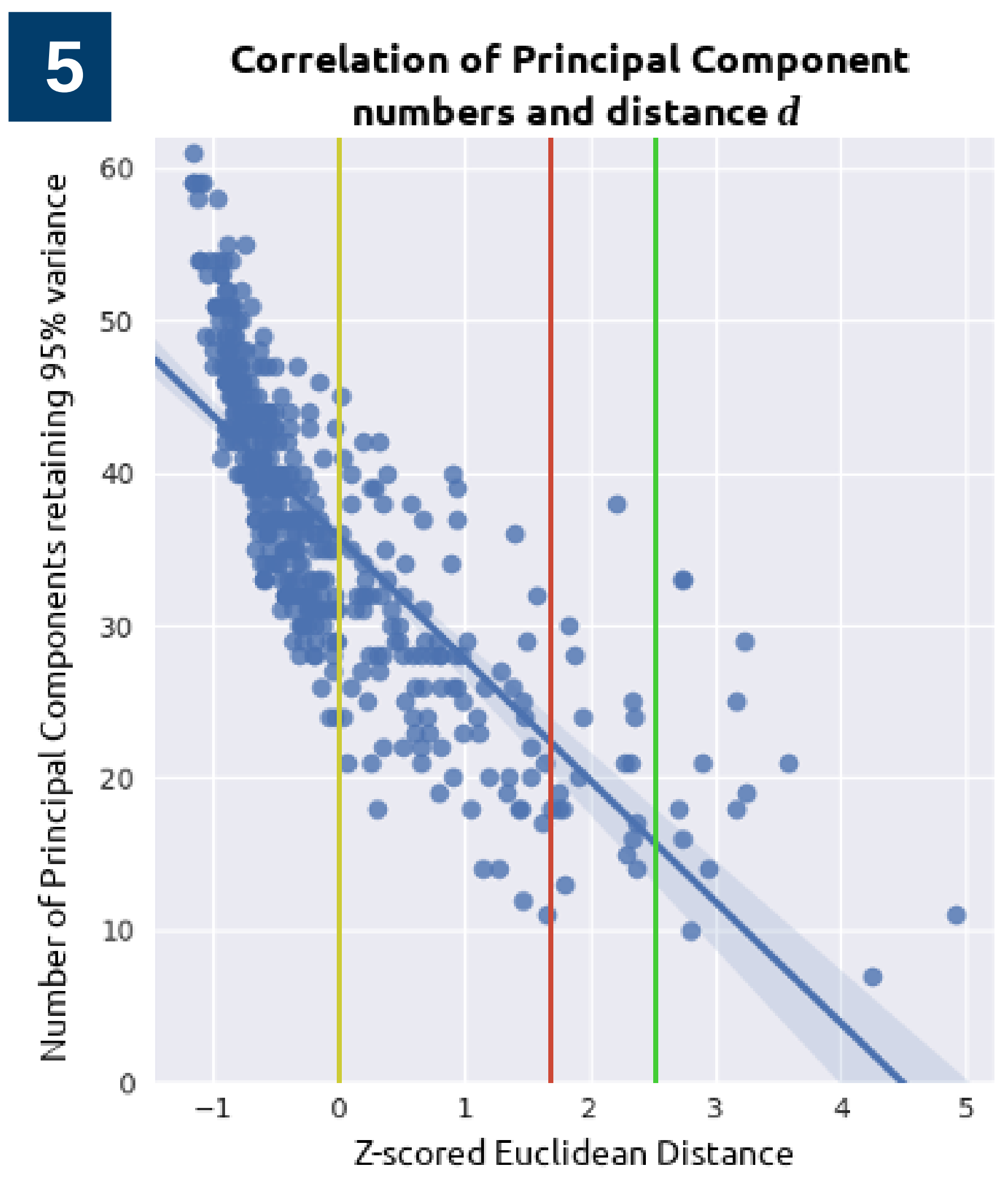
Results

4 Derivation of an estimate for the conservative threshold using k-means (k = 2) to cluster deviant and non-deviant connectivity matrices by their distances $\mathbf{d}^z$.

Black horizontal line shows separation, which was rounded to 0 and defined as threshold value for group clustering. Green circles represent outlier subjects as defined by this conservative threshold.



5

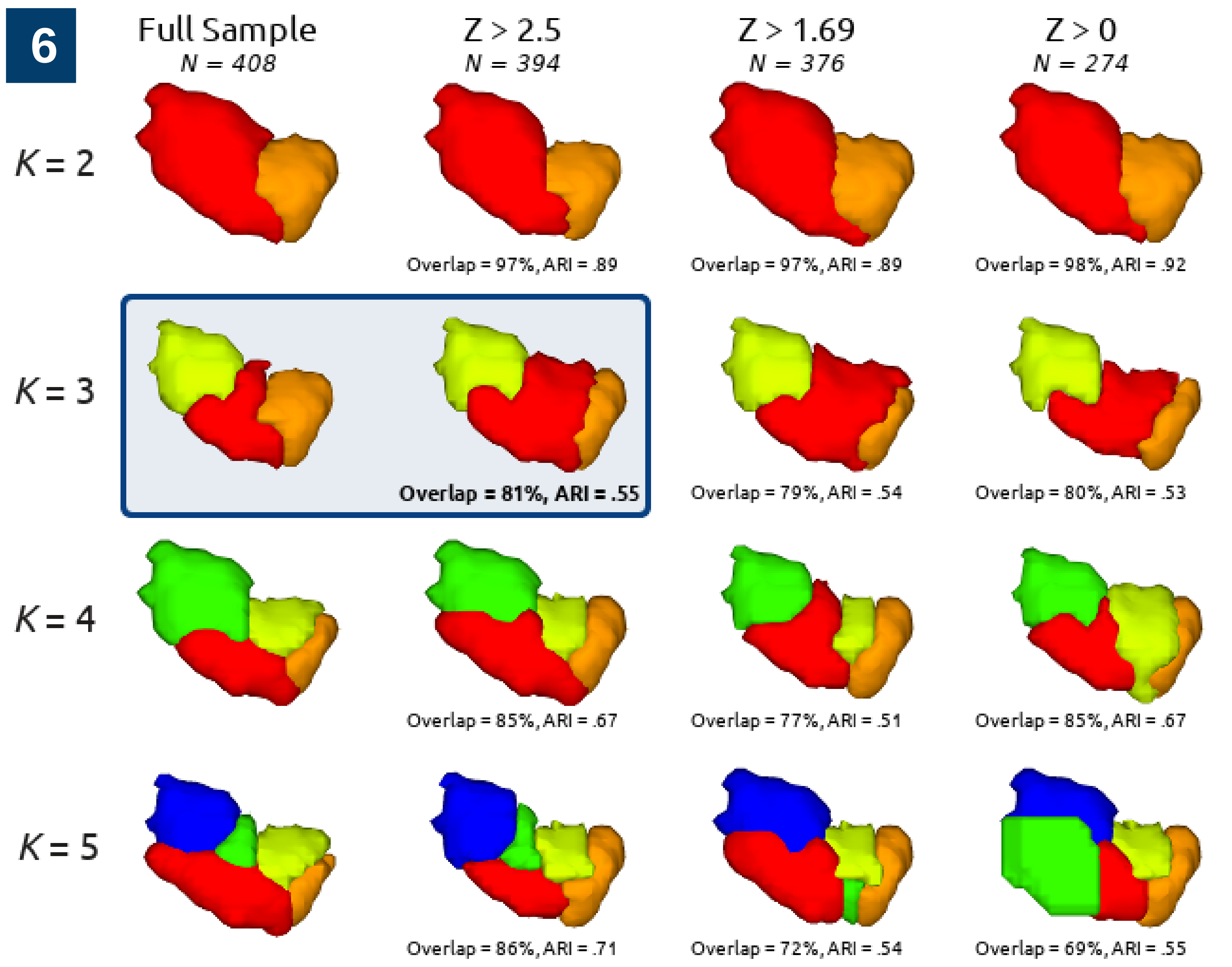


Correlation between distance vector $\mathbf{d}^z$ (***Fig. 2***) and PC_n. Vertical lines represent outlier thresholds for 0, 1.69, and 2.5 as yellow, red, and green, and applying them removes 134, 32, and 14 subjects respectively. Correlation value is -.79. Thus, outlier time-series seem to have lower intrinsic dimensionality.

Distance $\mathbf{d}^z$ and ARI (***Fig. 3***) correlate as -.38, -.41, -.49, and -.53 for k = 2, 3, 4, and 5 accordingly.

Results suggest outliers cluster differently, thus including them into a group-level consensus might be detrimental.

6



Group-level clustering of R-insula with and without outlier-removal. Clusterings ordered by k clusters (vertical) and outlier threshold (horizontal). Overlap and ARI values show similarity of clustering to clustering without outlier-removal. This figure was visualized with the BrainNet Viewer [4].

Differences can be found between these group-level parcellations. For instance, comparing the *liberal* 2.5 threshold-removed group parcellation for k = 3 with a group parcellation without outlier-removal (see highlight) shows only an 81% overlap.

Discussion

The differences in clusterings highlight the influence of outliers. A negative correlation between PC_n and distance $\mathbf{d}^z$ implies low intrinsic dimensionality comes paired with connectivity that is more distant from the sample (***Fig. 5***). While assessment of group-level parcellations reveals that clustering results were only relatively stable across thresholds for k = 2 (***Fig. 6***), ample evidence suggests more than 2 clusters in the R-insula [5,6,7]. Thus, differences due to outliers in k > 2 clusterings are problematic. As linkage algorithms in hierarchical clustering as well as k-means clustering are sensitive to outliers [8], it is important to remove them by using a proper identification threshold. In the future we will focus on automatic identification of parameters that lead to biologically meaningful parcellations.

References:
[1] Eickhoff, S.B. Thirion, B., Varoquaux, G., & Bzdok, D. (2015), ‘Connectivity-based parcellation: Critique and implications’, *Human Brain Mapping*, **vol. 36**, no. 12, pp. 4771-4792;
[2] Harvard Oxford Atlas distributed with FSL (http://www.fmrib.ox.ac.uk/fsl/)
[3] Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., for the WU-Minn HCP Consortium (2013), ‘The WU-Minn Human Connectome Project: An overview’, *NeuroImage*, **vol. 80**, pp. 62-79
[4] Xia, M., Wang, J., & He, Y. (2013), ‘BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics’, *PloS ONE*, **vol. 8**, no. 7
[5] Cauda, F., D’Agata, F., Sacco, K., Duca, S., Geminiani, G., & Vercelli, A. (2011), ‘Functional connectivity of the insula in the resting brain’, *NeuroImage*, **vol. 55**, no. 1, pp. 8–23
[6] Deen, B., Pitskel, N. B., & Pelphrey, K. A. (2011), ‘Three systems of insular functional connectivity identified with cluster analysis’, *Cerebral Cortex*, **vol. 21**, no. 7, pp. 1498–1506
[7] Kurth, F., Zilles, K., Fox, P. T., Laird, A. R., & Eickhoff, S. B. (2010), ‘A link between the systems: functional differentiation and integration within the human insula revealed by meta-analysis’, *Brain Structure and Function*, **vol. 214**, no. 5–6, pp. 519–534
[8] Duda, R. O., Hart, P. E., & Stork, D. G. (2000) *Pattern Classification, 2nd Edition*. Hoboken, New Jersey: Wiley-Interscience.

Acknowledgments: This study was supported by the Deutsche Forschungsgemeinschaft (DFG, EI 816/11-1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain" and and the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement 785907 (HBP SGA2)