Biostatistics (2016), **0**, 0, *pp.* 1–6
doi:10.1093/biostatistics/output

PCA leverage: outlier detection for high-dimensional functional magnetic resonance imaging data
Supplementary Materials

AMANDA F. MEJIA[a], MARY BETH NEBEL[b], ANI ELOYAN[c], BRIAN CAFFO[a] and

MARTIN A. LINDQUIST[a*]

[a]*Department of Biostatistics, Johns Hopkins University, USA*

[b]*Center for Neurodevelopmental and Imaging Research, Kennedy Krieger Institute, USA*

[c]*Department of Biostatistics, Brown University, USA*

Appendix A: Functional MRI data, pre-processing and quality control

Table 1. *For each dataset, the total number of subjects (N), number of subjects that passed quality inspection (N^+), number of subjects used to perform GICA (N_{ICA}), and the number of signal GICA networks identified (Q^+).*

Dataset	N	N^+	N_{ICA}	Q^+
California Institute of Technology (Caltech)	38	19	19	8
Carnegie Mellon University (CMU)	27	18	18	5
Kennedy Krieger Institute (KKI)	146	140	50	8
University of Leuven: Sample 1 (Leuven 1)	29	23	23	6
University of Leuven: Sample 2 (Leuven 2)	35	31	31	10
Ludwig Maximilians University Munich (LMU)	57	55	50	10
NYU Langone Medical Center (NYU)	184	108	50	13
Oregon Health and Science University (OHSU)	28	28	28	8
Olin Institute of Living at Hartford Hospital (Olin)	36	29	29	4
University of Pittsburgh School of Medicine (Pitt)	57	54	50	12
Social Brain Lab, Netherlands Institute for Neurosciences (SBL)	30	30	30	9
San Diego State University (SDSU)	36	32	32	14
Stanford University (Stanford)	40	35	35	5
Trinity Centre for Health Sciences (Trinity)	49	47	47	11
University of California, Los Angeles: Sample 1 (UCLA 1)	82	44	44	11
University of California, Los Angeles: Sample 2 (UCLA 2)	27	18	18	6
University of Michigan: Sample 1 (UM 1)	110	89	50	6
University of Michigan: Sample 2 (UM 2)	35	34	34	9
University of Utah School of Medicine (USM)	101	94	50	10
Yale Child Study Center (Yale)	56	46	46	11

Image pre-processing consisted of the following steps. SPM12b's segmentation tool was first used to correct for broad intensity variations across the MPRAGE volume; the bias-corrected MPRAGE was then registered to the first (stabilized) functional volume and normalized to Montreal Neuological Institute (MNI) space. Volumes corresponding to the first 10 seconds of the rs-fMRI scan were dropped to allow for magnetization stabilization. The remaining volumes were slice-time adjusted using the slice acquired at the middle of the repetition time (which varied by site). Rigid body realignment parameters were estimated with respect to the first (stabilized) functional volume of the rs-fMRI scan and used to calculate mean framewise displacement (FD), a summary measure of between-volume participant motion (Power *and others*, 2014). The non-linear spatial transformation estimated from the co-registered MPRAGE was then applied to the functional data along with the estimated rigid body realignment parameters and resulted in 2-mm isotropic voxels in MNI space. Each resting state scan was temporally detrended on a voxelwise basis and spatially smoothed using a 5-mm full width at half maximum (FWHM) Gaussian kernel (Smith *and others*, 2004).

After pre-processing, each rs-fMRI scan was quality inspected for motion and issues with registration and normalization using the following procedure. First, scans were flagged for quality if mean FD across the scan was greater than 2 standard deviations above the sample mean. We then calculated the Pearson spatial correlation between the first (stabilized) volume of each subject's MNI-registered data and SPM's EPI template (Allen *and others*, 2011). In total, 229 subjects were found to have major quality problems. Table 1 displays the number of subjects from each data collecting site that passed quality inspection. All scans were included in our analysis to assess the effect of outlier removal; however, only those scans that passed quality inspection were used to create group-level ICA maps.

Appendix B: Estimation of subject-level resting-state networks

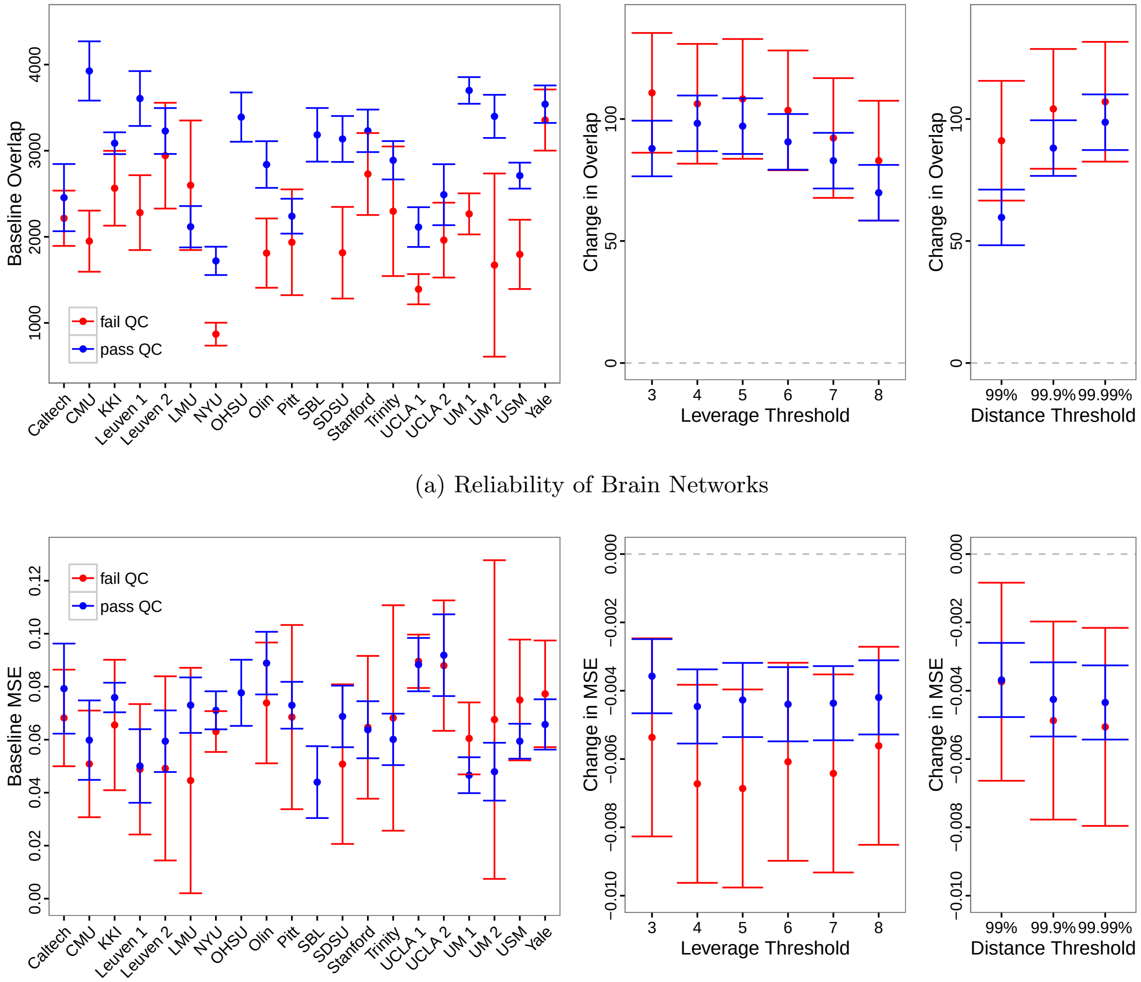
Let $k = 1, \dots, 20$ index datasets collectively forming the ABIDE. For each dataset, define a group-level brain mask as those voxels shared by at least 10% of subjects in the dataset. To define group-level brain networks through ICA, we use all subjects that passed quality inspection, downsampling to 50 subjects for those datasets containing more than 50 such subjects. Let $\mathcal{M}_k$ be the resulting set of subjects for dataset k. For each subject $i \in \mathcal{M}_k$, let $\mathbf{Y}_i$ be the $T_i \times V_k$ data matrix after centering each voxel across time, where T_i is the number of length of the rsfMRI scan of subject i and V_k is the number of voxels in the group-level brain mask. For each subject $i \in \mathcal{M}_k$, we perform PCA and retain 50 PCs to obtain $\mathbf{Y}_i = \mathbf{U}_i\mathbf{D}_i\mathbf{V}_i^t + \mathbf{E}_i$, resulting in the reduced $50 \times V_k$ subject-level data $\tilde{\mathbf{Y}}_i = \mathbf{D}_i\mathbf{V}_i^t$. (Note that for ICA we use PCA to reduce dimensionality along the temporal dimension, whereas for outlier detection we reduce along the spatial dimension.) Next, we temporally concatenate all subjects to form the $50|\mathcal{M}_k| \times V_k$ matrix $\mathbf{Y}_k$. We then perform

PCA again at the group level with $Q = 30$ components to obtain $\mathbf{Y}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^t + \mathbf{E}_k$, resulting in the reduced $Q \times V_k$ group-level data $\tilde{\mathbf{Y}}_k = \mathbf{D}_k\mathbf{V}_k^t$. Finally, we apply the fastICA algorithm (Marchini *and others*, 2013) to obtain $\tilde{\mathbf{Y}}_k = \mathbf{A}_k\mathbf{S}_k$, where $\mathbf{S}_k$ is a $Q \times V_k$ matrix whose rows contain the group-level spatial ICs and $\mathbf{A}_k$ is the $Q \times Q$ mixing matrix.

To identify those ICs corresponding to resting-state networks, we first standardize each spatial IC by its mean and standard deviation. We threshold the result at ± 2 to identify approximately the top 5% of voxels, a standard approach for identification of resting-state networks (Eloyan *and others*, 2013). We then visually inspect each spatial IC and label those corresponding to known resting-state brain networks. This results in 4-14 signal ICs per dataset (Table 1). While the number of signal ICs identified for some datasets is quite small, this is not surprising given the widely varying quality and quantity of data in each dataset. We observe a positive association between sample size and number of signal ICs: based on a simple no-intercept linear model, we estimate that for every additional subject included in analysis, on average 0.23 (95% CI: [0.20, 0.26]) additional signal ICs are identified through GICA. As more subjects are included in GICA, more resting-state brain networks can be clearly identified. Let $\tilde{\mathbf{S}}_k$ denote the $Q_k \times V_k$ matrix containing the Q_k resting-state networks identified for dataset k.

To obtain subject-level ICs, we perform dual regression (Beckmann *and others*, 2009) as follows. Note that $\tilde{\mathbf{S}}_k$ has been centered and scaled across voxels and $\mathbf{Y}_i$ has been robustly centered and scaled across time, as described in Section 2.1. Let $\mathbf{Y}_i$ also be centered across voxels. In the first regression, temporal ICs for subject $i \in \mathcal{M}_k$ are obtained by regressing $\mathbf{Y}_i^t$ against $\tilde{\mathbf{S}}_k^t$ to obtain $\mathbf{A}_i^t = \left(\tilde{\mathbf{S}}_k\tilde{\mathbf{S}}_k^t\right)^{-1}\tilde{\mathbf{S}}_k\mathbf{Y}_i^t$. In the second regression, subject-level spatial ICs for subject $i \in \mathcal{M}_k$ are obtained by regressing $\mathbf{Y}_i$ against $\mathbf{A}_i$ to obtain $\mathbf{S}_i = \left(\mathbf{A}_i^t\mathbf{A}_i\right)^{-1}\mathbf{A}_i^t\mathbf{Y}_i$. This results in the ICA decomposition $\mathbf{Y}_i \approx \mathbf{A}_i\mathbf{S}_i$, where $\mathbf{A}_i$ is $T_i \times Q_k$ and $\mathbf{S}_i$ is $Q_k \times V_k$. We are interested in $\mathbf{S}_i$, whose rows contain the resting-state brain networks for subject i.

Appendix C: Effect of outlier removal, stratified by quality control results



(a) Reliability of Brain Networks

(b) Reliability of Between-Network Connectivity

Fig. 1: Estimates and 95% confidence intervals for the model coefficients after stratifying by quality inspection results. The coefficients for each outlier removal method (α_m) show that while both groups of subjects benefit from outlier removal, those who failed quality inspection tend to improve slightly more, as we might expect.

References

Allen, Elena A, Erhardt, Erik B, Damaraju, Eswar, Gruner, William, Segall, Judith M, Silva, Rogers F, Havlicek, Martin *and others*. (2011). A baseline for the multivariate comparison of resting-state networks. *Frontiers in Systems Neuroscience* **5**.

Beckmann, Christian F, Mackay, Clare E, Filippini, Nicola and Smith, Stephen M. (2009). Group comparison of resting-state fMRI data using multi-subject ICA and dual regression. *NeuroImage* **47**(Suppl 1), S148.

Eloyan, Ani, Crainiceanu, Ciprian M and Caffo, Brian S. (2013). Likelihood-based population independent component analysis. *Biostatistics* **14**(3), 514–527.

Marchini, J L, Heaton, C and Ripley, B D. (2013). *fastICA: FastICA algorithms to perform ICA and projection pursuit*. R package version 1.2-0.

Power, Jonathan D, Mitra, Anish, Laumann, Timothy O, Snyder, Abraham Z, Schlaggar, Bradley L and Petersen, Steven E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341.

Smith, Stephen M, Jenkinson, Mark, Woolrich, Mark W, Beckmann, Christian F, Behrens, Timothy EJ, Johansen-Berg, Heidi *and others*. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23**, S208–S219.