Supporting information for "Distributional Independent Component Analysis for Diverse Neuroimaging Modalities" by Ben Wu[1,2], Subhadip Pal[3], Jian Kang[4,*] and Ying Guo[5,**]

[1]Center for Applied Statistics, Renmin University of China, Beijing, 100872, CN
[2]School of Statistics, Renmin University of China, Beijing, 100872, CN
[3]Department of Biostatistics and Bioinformatics, University of Louisville, Louisville, KY, 40292, USA
[4]Department of Biostatistics, University of Michigan, Ann Arbor, MI, 48109, USA
[5]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, 30322, USA
[*]email:jiankang@umich.edu
[**]email:yguo2@emory.edu

1. Web Appendix A: Additional Real Data Results

We show in this section real data results for randomly selected additional subjects from the PNC study. We specified the same number of ICs as that in the Section 3 of the main text. The functional/structural networks identified for the additional subjects are fairly consistent with the findings reported in the main text of the paper. See Web Figures 1 and 2 for details.

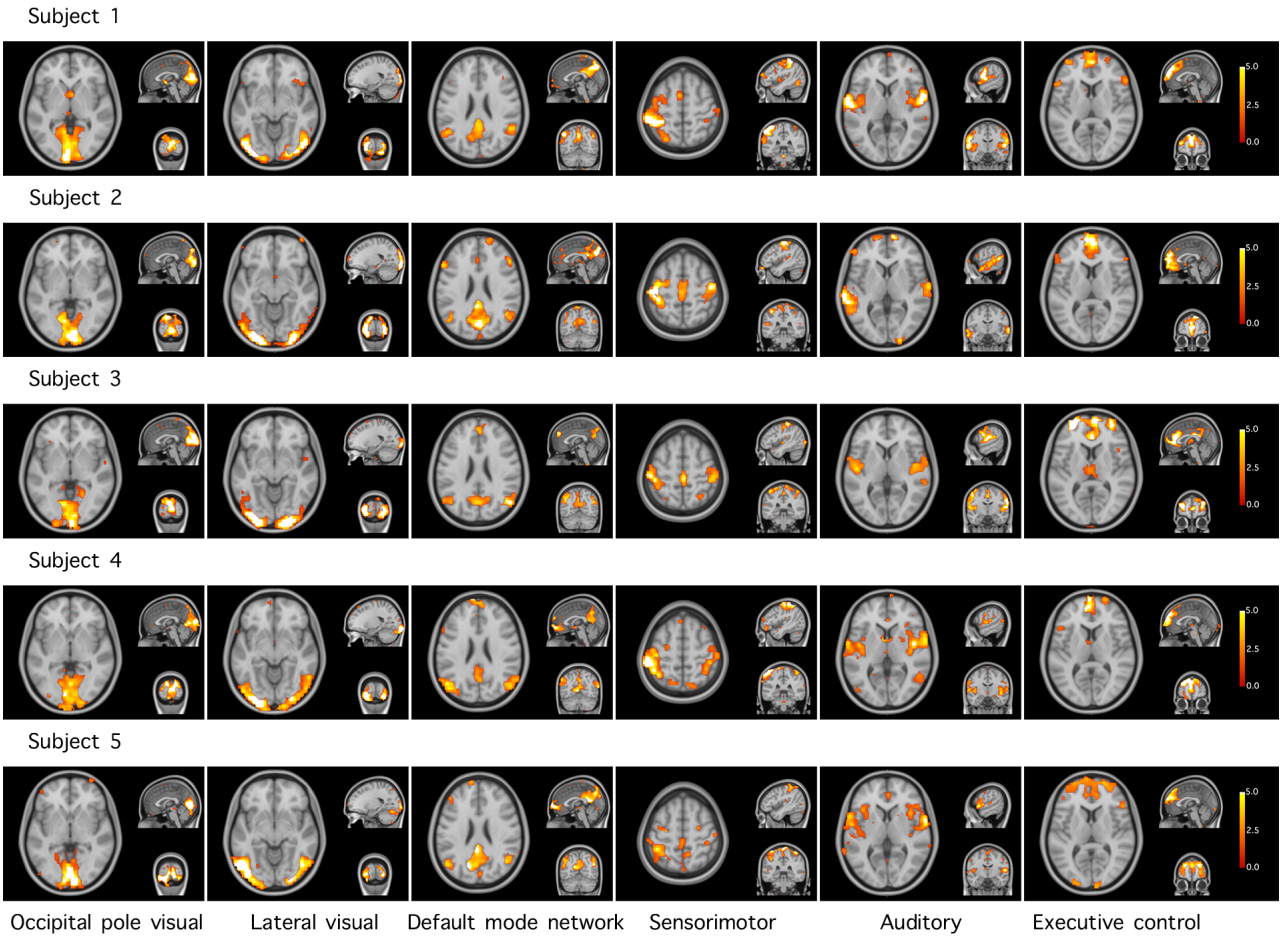
2. Web Appendix B: Complexity and Computational Time

DICA is not computationally demanding in both stages, where the stage 1 is fitting a mixture model using the EM algorithm. The computational complexity within each iteration is $O(JKT)$ where T is the dimension of imaging signals at each voxel, J is number of voxels and K is the number of mixture components. The EM algorithm usually only takes a couple of iterations to get converged. In stage 2, DICA applies the classical ICA method to the voxel-wise posterior probability weights. When we choose the Infomax ICA which is also an iterative method, the computational complexity is $O(JKL)$ within each iteration. In contrast, when we directly apply the Infomax ICA to the raw imaging data, the complexity within each iteration is $O(JTL)$. In the analysis of fMRI data, K is typically less than T. Thus the stage 2 of DICA can be faster than the traditional ICA. The actual computational costs of DICA for the analysis fMRI and DTI data were just about hundreds of seconds. For rs-fMRI data, we use the R package "mclust" to estimate the MoG model at stage one, and R package "ica" to perform infomax ICA at stage two. The user CPU time was around 150 seconds on a MacBook Pro with 3.1 GHz Dual-Core Intel Core i5 processor and 8 GB memory. For DTI data, we fit the mixture of Wishart distributions with R code at stage one, and use R package "ica" to perform infomax ICA at stage two. The user CPU time was around 585 seconds on the same computer.

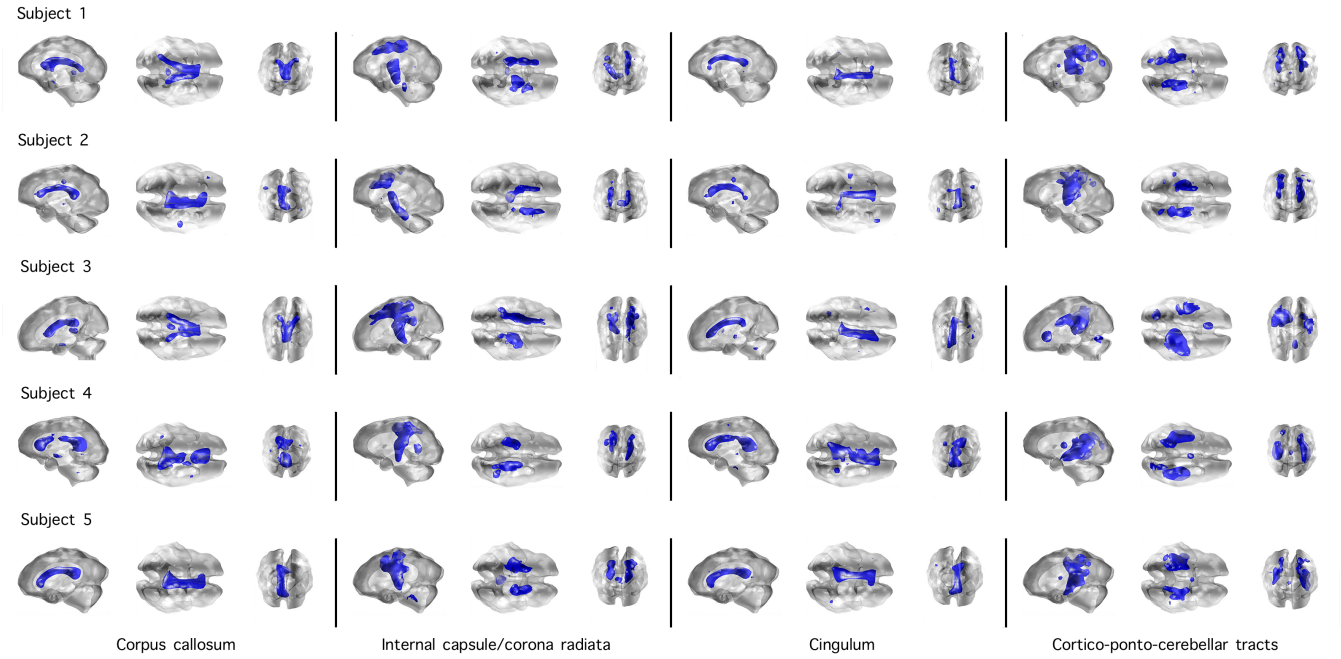
3. Web Appendix C: Sensitivity Analysis

3.1. Simulation Study IV

We conducted a simulation study in this section to evaluate different selections of K and L, i.e., the number of components at stage one and two respectively. The data were generated as in the simulation study I. Of note, the true number of mixture components $K = 6$ and the true number of source signals $L = 3$. At stage one, we estimated a mixture model with $K = 4, 5, \ldots, 11$. Web Table 1 shows the means of BICs for each selection of K across 100 simulation repetitions. We can easily find from the table that $K = 6$ has



Web Figure 1: Brain functional networks estimated from additional subjects' resting state fMRI data from the PNC study using DICA.



Web Figure 2: Brain structural networks (the estimated IC spatial maps) estimated from additional subjects' DTI data from the PNC study using DICA. DICA discovered components corresponding to major white fiber pathways in the brain.

the smallest averaged BIC, which means BIC leads to the correct number. In fact, BIC has successfully found the true K for all the 100 repetitions. In Web Table 2, we show the correlations between the true and the estimated source signals with different selection of K and L. When the L was overspecified, we show the results with the top three best matched estimated source signals. Results in Web Table 2 show that different choices of K and L yield pretty similar findings as the case when the true K and L are specified. Overspecification of K seems to lead to better results as compared to underspecification of K. This suggests that we should choose a relative large K for the mixture distribution in applications. Underspecification of L inevitably leads to reduced signal recovery for some of the source(s), but the estimated source signals were still highly correlated with the true signals for most of the sources. It is worth noting that even when K and L are misspecified the estimates from DICA are still more accurate as compared to those from FastICA or Infomax ICA.

Web Table 1: Mean of the BICs at stage one with 100 simulation repetitions. The true $K = 6$.

K	4	5	**6**	7
BIC	114398.13	93492.49	58071.05	58157.62
K	8	9	10	11
BIC	58239.51	58321.54	58403.20	58484.77

3.2. Selection of K and L for real data

In this section, we evaluate how robust the results are with respect to the selection of K and L for real data analysis. The data we used were the same as in the section 3 Application to Real Imaging Data. We compared the source signals estimated with the model we used in the main text ($K = 20$ and $L = 14$) and the signals estimated with other selection of K and L. We evaluated K from 16 to 28 and L from 10 to 26. Similar as the simulation studies, the results remain fairly consistent with the various values

Web Table 2: Mean(standard deviation) of the correlations between the true and estimated source signals based on different selections of K and L with 100 simulation repetitions. The true parameters are K=6, L=3.

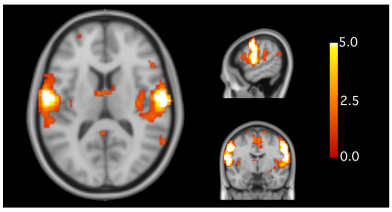
	K=4, L=2	K=5, L=2	K=6, L=2	K=7, L=2
Source 1	0.828(0.033)	0.889(0.114)	0.989(0.000)	0.989(0.002)
Source 2	0.772(0.085)	0.914(0.105)	0.974(0.000)	0.973(0.004)
Source 3	0.181(0.091)	0.145(0.013)	0.133(0.002)	0.132(0.021)
	K=5, L=3	**K=6, L=3**	K=7, L=3	K=8, L=3
Source 1	0.882(0.099)	0.993(0.000)	0.990(0.010)	0.990(0.006)
Source 2	0.894(0.083)	0.991(0.000)	0.988(0.028)	0.988(0.023)
Source 3	0.732(0.165)	0.950(0.001)	0.940(0.030)	0.939(0.015)
	K=6, L=5	K=7, L=5	K=8, L=5	K=8, L=7
Source 1	0.941(0.002)	0.941(0.001)	0.933(0.037)	0.919(0.045)
Source 2	0.870(0.002)	0.871(0.003)	0.870(0.015)	0.843(0.059)
Source 3	0.913(0.001)	0.912(0.001)	0.912(0.002)	0.893(0.057)

of K and L. Specifically, as shown in the Web Table 3, the correlations with the original results largely range between 0.7 to 0.95 in the majority of cases. The correlations are relatively higher when K or L are specified greater than the original settings than when they are specified smaller than the original values. The findings are consistent for both fMRI and DTI data. It is worth noting that some of the networks, such as the sensorimotor and auditory, that were extracted as separate ICs under the original setting are merged into a single IC when L is specified considerably lower than the original setting ($L = 10$) (Web Figure 3). When L is specified much larger than the original setting, an IC identified under the original setting may split into multiple ICs representing subnetworks of the original network. For example, we observe the IC corresponding to the default mode network (DMN) under the original setting are extracted as two ICs (with $L = 26$) which represent the medial prefrontal cortex subnetwork of DMN and the posterior cingulate and temporoparietal subnetwork of DMN (Web Figure 3). Our results are consistent with the findings from previous work regarding the effects of the number of ICs on the extracted brain networks (Smith et al., 2009; Hyatt et al., 2015).

Web Table 3: The sensitivity analysis of the DICA results with respect to the selection of K and L for the real data analysis. Presented values are the average correlations between the estimated ICs under the the original parameter setting (i.e., $K = 20$, $L = 14$) and their matching ICs estimated under other selections of K and L (Note that L is less than K based on the ICA assumption). The matching ICs were the components with the highest correlations with the original ICs.

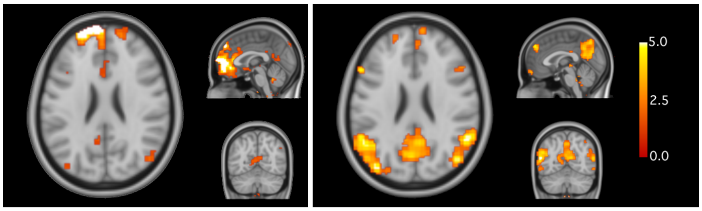
	fMRI				DTI			
L \ K	16	20	24	28	16	20	24	28
10	0.653	0.788	0.667	0.544	0.559	0.598	0.551	0.297
14	0.779	1.000	0.678	0.686	0.877	1.000	0.804	0.756
18	-	0.936	0.819	0.875	-	0.886	0.931	0.812
22	-	-	0.799	0.888	-	-	0.918	0.865
26	-	-	-	0.889	-	-	-	0.895

(a) L=10



An IC includes both Sensorimotor and Auditory

(b) L=26



Default mode IC 1 Default mode IC 2

Web Figure 3: The sensitivity analysis of the number of ICs, i.e. L, for real imaging data analysis. Brain functional networks based on a subject's resting state fMRI data from the PNC study using DICA with $L = 10$ and $L = 26$. With $L = 10$ which is smaller than the original setting of $L = 14$, the sensorimotor and auditory ICs under the original setting are merged into one IC. With $L = 26$ which is greater than the original setting, the default mode network IC is split into two ICs representing two subnetworks of DMN.

References

Hyatt, C. J., Calhoun, V. D., Pearlson, G. D., and Assaf, M. (2015). Specific default mode subnetworks support mentalizing as revealed through opposing network recruitment by social and semantic fmri tasks. *Human Brain Mapping* **36,** 3047–3063.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences* **106,** 13040–13045.