REGRESSION-ASSISTED INDEPENDENT VECTOR ANALYSIS: A SOLUTION TO LARGE-SCALE FMRI DATA ANALYSIS

H. Yang[1,*], *B. Gabrielson*[1,*], *V. D. Calhoun*[**], *T. Adali*[*]

[*] Dept. of CSEE, University of Maryland Baltimore County, Baltimore, USA
[**] Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, and Emory University, Atlanta, USA
[1] These authors contributed equally to this work

1. ABSTRACT

Multi-subject fMRI data is instrumental in understanding the brain function and studying different brain disorders. It is desirable to analyze fMRI datasets jointly to leverage the cross information that exists across multiple datasets. Independent Vector Analysis (IVA) is a powerful solution that can effectively leverage statistical dependence across multiple datasets, and is an attractive solution for fMRI data analysis. However, the computational costs of IVA can be intractable when dealing with a large number of datasets. In this paper, we propose an efficient method for large-scale fMRI analysis with the assistance of multilinear regression. As we demonstrate with results from resting state fMRI data, the proposed method achieves similar estimation performance to IVA on all available datasets, but with significantly improved efficiency and reliable performance.

2. INTRODUCTION

Functional magnetic resonance imaging (fMRI), reflecting neural activity changes in the brain by measuring the blood-oxygenation-level-dependent (BOLD) signal, has been widely used for understanding brain function and studying different brain disorders. Joint analysis of data from a large number of subjects provides a more complete picture of understanding subjects across different diseases and groups. Joint blind source separation (JBSS) is able to leverage statistical dependencies across the datasets, which makes it a powerful tool for jointly extracting interpretable sources from multiple datasets [1–5]. Among the JBSS methods, independent vector analysis (IVA) has been an attractive solution for multi-subject analyses [6].

IVA generalizes independent component analysis (ICA) to multiple datasets by defining multidimensional sources, each called a source component vector (SCV): a group of statistically dependent sources composed of one source from each dataset. IVA has been demonstrated effective at preserving subject variability [2, 7–9], identifying biomarkers [2], and identifying homogeneous subgroups [5, 10]. Despite the many strengths of IVA, its high computational cost makes it infeasible for a large number of datasets [11]. In such situations, practical methods such as group-ICA [12] become the primary alternative to overcoming complexity issues. Group-ICA operates by reducing all datasets to "shared dimensions" via a two-step principal component analysis on the concatenation of datasets, and then estimating components from an ICA on these shared dimensions. However, because group-ICA's approach necessarily assumes all sources are shared across the datasets, it may have a limited ability to preserve subject (dataset) variability compared to IVA. Thus, there exists a need for methods better exploiting potential variability in the data like IVA, while also being computationally tractable like group-ICA.

To provide a practical solution to these challenges, we propose a computationally efficient, multilinear regression-assisted approach to IVA. Under the assumption that a base model, a representative subset of the datasets, can capture the variability across all datasets, sources can be estimated by using sources from a base model as regressors for the remaining datasets. This enables regression IVA (regIVA) to efficiently estimate sources from additional datasets that are consistent with the base model. To enable all datasets to fully interact post-regression, the output from regression IVA is used as initialization for IVA, which we call regression-assisted IVA (RegAssist-IVA). The regression strategy utilized in our proposed approach leverages linear dependence, specifically correlation between source estimates, thereby, providing an extension of IVA with an assumed multivariate Gaussian distribution (IVA-G) [3].

To assess the performance of the proposed approach, we employ regression-assisted IVAG (RegAssist-IVAG) and IVA-G on a cohort of 98 individuals, comprising 49 healthy controls (HC) and 49 schizophrenia patients (SZ), sourced from the Baltimore site of B-SNIP data [13]. The evaluation encompasses three key aspects: spatial maps of resting-state

This work was supported in part by NIH R01MH118695, NIH R01MH123610, NIH R01AG073949, NSF 2316420. The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF): `hpcf.umbc.edu`.

networks (RSNs), power ratio, and cross-joint-ISI. The findings indicate that RegAssist-IVAG demonstrates comparable performance to IVA-G in terms of RSNs and power ratios of the estimated components while exhibiting enhanced reproducibility compared to IVA-G. Notably, there is a significant improvement in CPU time with RegAssist-IVAG compared to the standard IVA-G.

The rest of this paper is organized as follows: the background for IVA is presented in Section 3. The details of the proposed method RegAssist-IVA are in Section 4. Experimental data and results are introduced in Section 5 followed by the discussion in Section 6.

3. INDEPENDENT VECTOR ANALYSIS

For a given K datasets (subjects), IVA models each dataset as a mixture of N independent sources. The generative model of IVA of the k^{th} dataset can be written as:

$$\mathbf{x}^{[k]}(v) = \mathbf{A}^{[k]}\,\mathbf{s}^{[k]}(v)\,, \qquad 1 \le k \le K, \tag{1}$$

where $\mathbf{x}^{[k]}(v) = [x_1^{[k]}(v), \ldots, x_N^{[k]}(v)]^\top$ represents an observation vector at sample index v (superscript $\top$ represents transpose); $\mathbf{s}^{[k]}(v) = [s_1^{[k]}(v), \ldots, s_N^{[k]}(v)]^\top$ comprises N statistically independent, zero mean, and unit variance latent sources, while $\mathbf{A} \in \mathbb{R}^{N\times N}$ is an unknown invertible mixing matrix. The goal of IVA is to estimate demixing matrices $\mathbf{W}^{[k]} \in \mathbb{R}^{N\times N}$ such that the estimates $\mathbf{y}^{[k]}(v) = [y_1^{[k]}(v), \ldots, y_N^{[k]}(v)]^\top$, where $\mathbf{y}^{[k]}(v) = \mathbf{W}^{[k]}\mathbf{x}^{[k]}(v)$, are maximally independent within a dataset. Simultaneously, IVA maximizes dependency across datasets through the definition of source component vector (SCV), which is denoted as $\mathbf{s}_n(v) = [s_n^{[1]}(v), \ldots, s_n^{[K]}(v)]^\top \in \mathbb{R}^K, 1 \le n \le N$, which includes the n^{th} source component $s_n^{[k]}(v)$ from each of the K datasets. The estimation of the n^{th} SCV can be expressed as $\mathbf{y}_n(v) = [y_n^{[1]}(v), \ldots, y_n^{[K]}(v)]^\top$. For simplicity, we drop the sample index v for the remainder of the paper.

Estimates of K demixing matrices can be achieved by minimizing the mutual information among the SCVs, which can be written as the following cost function

$$\mathcal{J}_{\text{IVA}}\,(\boldsymbol{\mathcal{W}}) = \sum_{n=1}^{N} \mathcal{H}(\mathbf{y}_n) - \sum_{k=1}^{K} \log|\det(\mathbf{W}^{[k]})|, \tag{2}$$

where $\boldsymbol{\mathcal{W}} = \{\mathbf{W}^{[1]}, \mathbf{W}^{[2]}, \ldots, \mathbf{W}^{[K]}\}$ represents the K demixing matrices for K datasets, $\mathbf{y}_n$ is the estimated SCV, and $\mathcal{H}(\mathbf{y}_n) = -\mathbb{E}\{\log p_n(\mathbf{y}_n)\}$ denotes the (differential) entropy of $\mathbf{y}_n$ and $p_n(\cdot)$ is the multivariate probability density function (pdf) of the n^{th} SCV. Note that the mutual information within a SCV can be written as $\mathcal{I} = \{\mathbf{y}_n\} = \sum_{k=1}^{K} \mathcal{H}\{y_n^{[k]}\}$ - $\mathcal{H}\{\mathbf{y}_n\}$, based on which (2) can be rewritten as

$$\begin{aligned}\mathcal{J}_{\text{IVA}}\,(\boldsymbol{\mathcal{W}}) \;=\; &\sum_{n=1}^{N}\left(\sum_{k=1}^{K}\mathcal{H}\Big\{y_n^{[k]}\Big\} - \mathcal{I}\Big\{\mathbf{y}_n\Big\}\right) \\ &- \sum_{k=1}^{K}\log\Big|\det\Big(\mathbf{W}^{[k]}\Big)\Big|,\end{aligned} \tag{3}$$

which shows that minimizing $\mathcal{J}_{\text{IVA}}\,(\boldsymbol{\mathcal{W}})$ is equivalent to minimizing individual source entropies $\mathcal{H}\{y_n^{[k]}\}$ while simultaneously maximizing the mutual information within each SCV $\mathcal{I}\mathbf{y}_n$. Therefore, (3) illustrates that IVA maximizes independence across SCVs and maximizes dependence within SCVs simultaneously.

IVA with multivariate Gaussian distribution (IVA-G) [3] modeled that the sources in an SCV are multivariate Gaussian distribution with i.i.d samples. Substituting into (2) the entropy of a K-dimensional multivariate Gaussian vector, the cost function of IVA-G can be written as

$$\begin{aligned}\mathcal{J}_{\text{IVA-G}}\,(\boldsymbol{\mathcal{W}}) = \frac{NK\,\log(2\pi e)}{2} \;+\; &\frac{1}{2}\sum_{n=1}^{N}\log\Big|\det\Big(\hat{\boldsymbol{\Sigma}}_n\Big)\Big| \\ &- \sum_{k=1}^{K}\log\Big|\det\Big(\mathbf{W}^{[k]}\Big)\Big|\end{aligned} \tag{4}$$

where $\hat{\boldsymbol{\Sigma}}_n \in \mathbb{R}^{K\times K}$ is the sample covariance of the n^{th} SCV, $\mathbf{y}_n$. Similar to (3), minimizing (4) is equivalent to minimizing correlations across N SCVs while simultaneously maximizing correlations within each SCV. A similar strategy is applied to formulate the cost function of regressing a new dataset onto a previously learned IVA-G model.

4. REGRESSION ASSISTED-IVA

4.1. Multilinear Regression Formulation

For a IVA-G model derived from K_b datasets, the learned SCVs can be written as $\mathbf{Y}_n = [\mathbf{y}_n^{[1]}, \mathbf{y}_n^{[2]}, \ldots, \mathbf{y}_n^{[K]}]^\top \in \mathbb{R}^{K\times V}$, where $\mathbf{Y}^{[k]} = \mathbf{W}^{[k]}\mathbf{X}^{[k]}$, with $\mathbf{Y}^{[k]}, \mathbf{X}^{[k]} \in \mathbb{R}^{N\times V}$ denote the estimated sources and observed samples, respectively. For a new dataset, $\mathbf{X}^{[i]}$, the corresponding N estimates can be achieved by making it maximally correlated to its corresponding SCV, while also maximally uncorrelated to the $N-1$ other SCVs [14]. The process of regressing a IVA model to a new dataset is referred as regression IVA. The cost function of regression IVA can be expressed as

$$\begin{aligned}\mathcal{J}_{\text{regIVA}}\,(\hat{\mathbf{w}}_n^{[i]}) \;=\; &\hat{\mathbf{w}}_n^{[i]\top}\left[\hat{\mathbf{R}}_n^{[i]} \;-\; \sum_{\substack{m=1\\ m\neq n}}^{N}\hat{\mathbf{R}}_m^{[i]}\right]\hat{\mathbf{w}}_n^{[i]} \\ &\text{s.t.}\;\; \Big\|\mathbf{w}_n^{[i]}\Big\|_2 = 1\end{aligned} \tag{5}$$

where $\mathbf{w}_n^{[i]}$ is the demixing vector estimating the n^{th} source for the provided dataset $\mathbf{X}^{[i]}$, and $\hat{\mathbf{R}}_n^{[i]} = (\frac{1}{T-1})^2\,\mathbf{X}^{[i]}\,\mathbf{Y}_n^\top\,\mathbf{Y}_n\,\mathbf{X}^{[i]\top} \in \mathbb{R}^{N\times N}$.

The solution that maximizes the quadratic form $\big[\hat{\mathbf{R}}_n^{[i]} - \sum_{\substack{m=1\\ m\neq n}}^{N}\hat{\mathbf{R}}_m^{[i]}\big] \in \mathbb{R}^{N\times N}$ also maximizes (5), leading to the analytical solution obtained by the eigenvector corresponding to

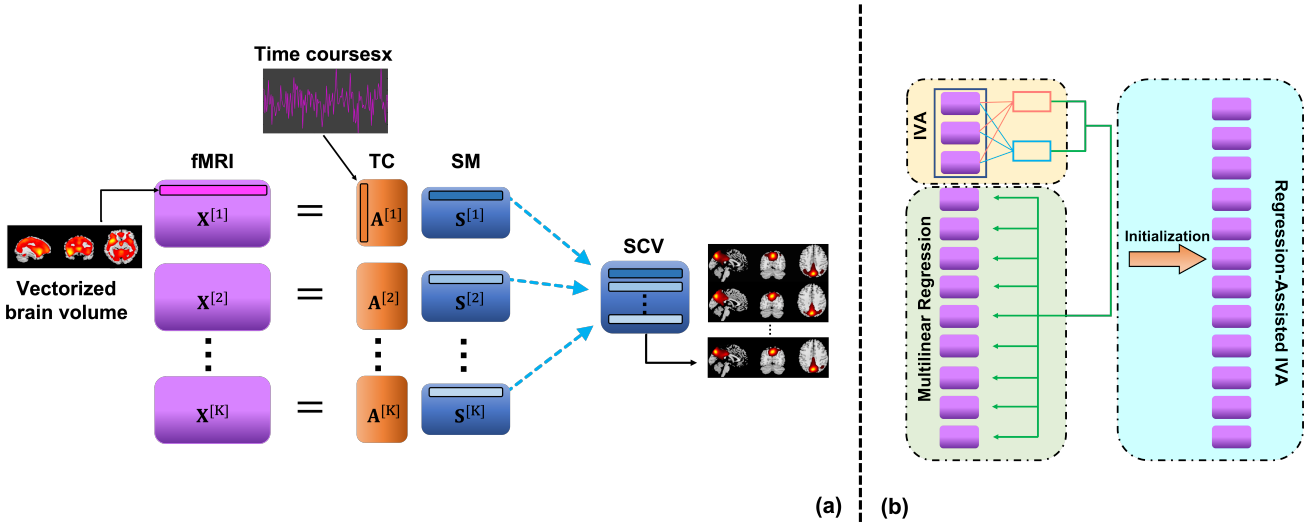

Fig. 1: Flowchart of IVA and regression assisted-IVA. (a): Apply IVA to fMRI data where each row of the estimated source matrix (SM) corresponds to a functional network and each column of the mixing matrix is the corresponding time courses (TC). (b): In the case of RegAssist-IVA, the standard IVA (referred to as IVA-G in this context) is initially applied to the base model, followed by a multilinear regression process applied to the remaining datasets. The outcomes of this regression step serve as an initialization for a standard IVA run on the entire dataset, enabling comprehensive interaction among datasets.

the largest eigenvalue of the quadratic form. The aforementioned process can be repeated for every source in a given additional dataset, facilitating IVA-G model to be easily scaled to any arbitrarily large number of datasets.

4.2. Regression Assisted-IVA

Multilinear regression process provides an efficient and closed-form source separation solution for a given dataset. However, due to the limitation of regression, the newly arrived dataset is not able to interact with the existing dataset. This may cast a shadow on using IVA which allows datasets to fully interact with each other. To enable all datasets to fully interact post-regression, the output from regression IVA is used as initialization for IVA, providing a good start point for the algorithm, which we call regression-assisted IVA (RegAssist-IVA). The outline of performing regression-assisted IVA-G on K datasets can be summarized as:

1. With K total datasets $\mathbf{X}^{[k]}$, divide the datasets into two groups: K_b datasets of the "base model" estimated by IVA-G, and K_a additional datasets that will be regressed onto the base model, with $K = K_b + K_a$.

2. Perform IVA-G on the K_b datasets of the base model, estimating demixing matrices $\mathbf{W}^{[k]}$ for these datasets. Use these $\mathbf{W}^{[k]}$ to obtain source estimates for each dataset $\mathbf{Y}^{[k]}$, and from these obtain the N SCVs of the base model $\mathbf{Y}_n$.

3. For each source of the K_a additional datasets $\mathbf{X}^{[i]}$, estimate the source's corresponding demixing vector $\mathbf{w}_n^{[i]}$ by the principal eigenvector of the corresponding quadratic form $\left[\hat{\mathbf{R}}_n^{[i]} - \sum_{\substack{m=1 \\ m \neq n}}^{N} \hat{\mathbf{R}}_m^{[i]}\right]$. As done by IVA-G, this regression step estimates sources that are maximally correlated with one SCV, while maximally uncorrelated to all other SCVs.

4. Initialize IVA-G on K datasets with $\mathbf{W}^{[k]}, k = 1, 2, \ldots, K$ achieved by step 3. This initialization allows the base model to fully interact with the regressed datasets.

5. EXPERIMENTAL RESULT

The proposed method, RegAssist-IVAG, is applied to 98 subjects, including 49 healthy control (HC) and 49 schizophrenia patients (SZ), from the Baltimore site of B-SNIP [13]. All subjects underwent a single 5-minute run of resting-state fMRI on a 3-T scanner. Subjects were instructed to keep their eyes open, focus on a crosshair displayed on a monitor, and remain still during the entire scan. We removed the first three time points and performed head motion correction followed by the slice-timing correction. The corrected fMRI data were then warped into the standard Montreal Neurological Institute (MNI) space through an echo-planar imaging template and then were resampled to $3 \times 3 \times 3$ mm^3 isotropic voxels. The resampled fMRI data were further smoothed using a Gaussian kernel with a full width at half maximum (FWHM) equal to 6 mm.

The base model of RegAssist-IVAG comprises a selection of 30 subjects (15 HC and 15 SZ) chosen randomly. For comparison, IVA-G on the full dataset and on the base model with

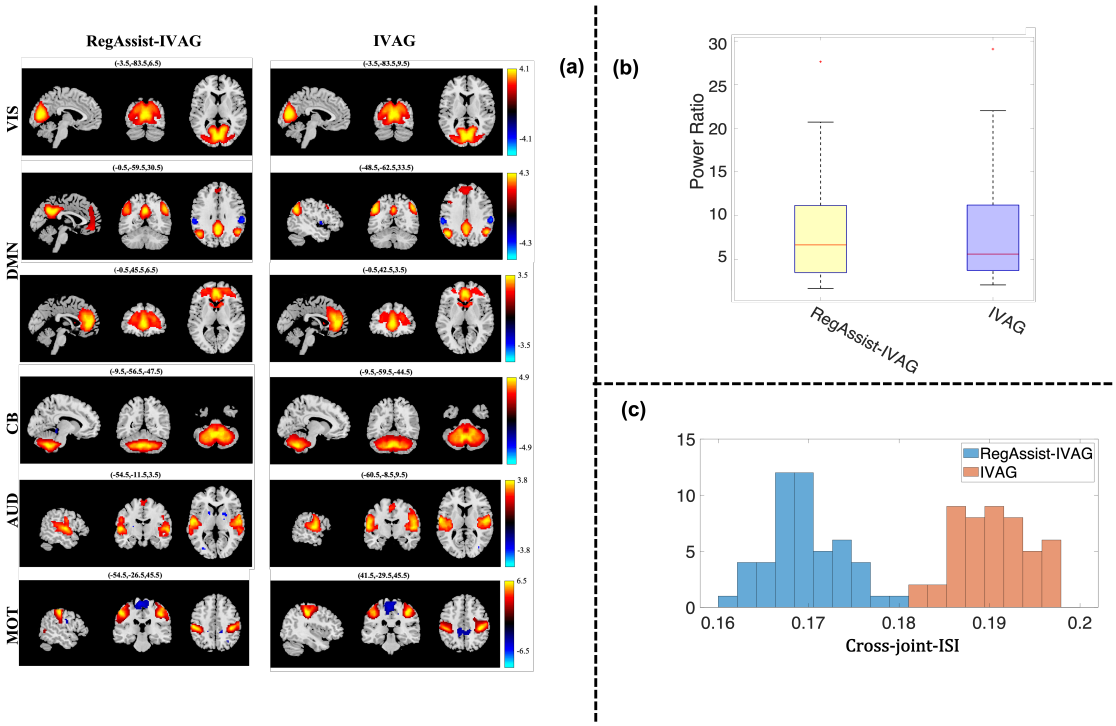

Fig. 2: Experimental results. (a): Spatial maps of resting state functional networks. The results from RegAssist-IVA and IVA-G are very close to each other. Five resting state functional networks are displayed, including visual network (VIS), default-mode network (DMN), cerebellar network (CB), auditory network (AUD), sensorimotor network (MOT). (b): Power ratio comparison. A higher power ratio usually indicates BOLD-related activity. Both IVA-G and RegAssist-IVAG show a similar range of power ratios, but the results from RegAssist-IVAG have a higher median value. (c): Cross-joint-ISI comparison. Results are generated from 50 runs with random initialization for IVA-G and the base model of RegAssist-IVAG. The RegAssist-IVAG provides more consistent results *i,e,* lower cross-joint-ISI, across runs than IVA-G.

random initialization are implemented for 50 different runs, and the best run, *i.e.,* the most consistent run, is selected using cross-joint inter-symbol-interference (cross-joint-ISI) [15]. The performance of IVA-G and RegAssist-IVAG are evaluated from three aspects: spatial maps of resting-state networks (RSNs), power ratio, and cross-joint-ISI. The assessment of RSN spatial maps visulizes the performance of the source separation algorithms, with a focus on the focal nature and strength of activation areas. Another metric useful for assessing the quality of fMRI component estimation is the power spectra of RSN time courses, including the power ratio between low-frequency ($< 0.1, \text{Hz}$) and high-frequency ($> 0.15, \text{Hz}$) bands. Given that neural-activity-related BOLD signals typically operate below 0.15 Hz, low power ratio values are generally associated with cardiac and respiratory noise, while high values suggest BOLD activity [16].

RegAssist-IVAG requires a CPU time of 1.91h, compared to 119.29h for IVA-G with 50 runs to achieve comparable performance. In Figure 2 (a), the spatial maps from 6 resting-state functional networks RSNs generated from the two algorithms are displayed. The similarity in RSN spatial maps from both methods suggests that RegAssist-IVAG can perform comparably to IVA-G's decomposition in a fraction of the time. The comparison of the power ratio between the two algorithms is included in Figure 2 (b), revealing a similar range of power ratios for IVA-G and RegAssist-IVAG. Figure 2 (c) demonstrates that RegAssist-IVAG yields lower cross-joint-ISI, indicating more consistent results across runs compared to IVA-G.

6. CONCLUSION

We proposed a regression assisted-IVA framework tailored for large-scale data analysis. This method leverages multilinear regression to scale IVA for extensive datasets without compromising performance. Our results illustrate that RegAssist-IVA produces fully interpretable network estimates and provides more consistent performance than standard IVA. The proposed method demonstrates significantly enhanced computational efficiency compared with standard IVA. Acknowledging the influence of the base model on RegAssist-IVA's performance, future endeavors will focus on optimizing base model selection and generalizing the methodology.

7. REFERENCES

[1] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Intl. Conf. Indep. Comp. Anal. Signal Sep.* Springer, 2006, pp. 165–172.

[2] T. Adali, M. Anderson, and G. S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Process. Mag*, vol. 31, no. 3, pp. 18–33, 2014.

[3] M. Anderson, T. Adali, and X.-L. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Trans. Signal Process*, vol. 60, no. 4, pp. 1672–1683, 2011.

[4] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, 1971.

[5] H. Yang, M. Akhonda, F. Ghayem, et al., "Independent vector analysis based subgroup identification from multisubject fMRI data," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2022, pp. 1471–1475.

[6] A. M. Engberg, K. W. Andersen, M. Mørup, et al., "Independent vector analysis for capturing common components in fMRI group analysis," in *Int. Workshop Pattern Recognit. Neuroimaging*. IEEE, 2016, pp. 1–4.

[7] J. Laney, K. P. Westlake, S. Ma, et al., "Capturing subject variability in fMRI data: A graph-theoretical analysis of GICA vs. IVA," *J. Neurosci. Methods*, vol. 247, pp. 32–40, 2015.

[8] J. T. Dea, M. Anderson, E. Allen, et al., "IVA for multi-subject fMRI analysis: A comparative study using a new simulation toolbox," in *IEEE Int. Workshop Mach. Learn. Signal. Process.* IEEE, 2011, pp. 1–6.

[9] S. Ma, R. Phlypo, V. D. Calhoun, et al., "Capturing group variability using IVA: a simulation study and graph-theoretical analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2013, pp. 3128–3132.

[10] Q. Long, S. Bhinge, V. D. Calhoun, et al., "Independent vector analysis for common subspace analysis: Application to multi-subject fMRI data yields meaningful subgroups of schizophrenia," *NeuroImage*, vol. 216, pp. 116872, 2020.

[11] H. Yang, T. Vu, Q. Long, et al., "Identification of homogeneous subgroups from resting-state fMRI data," *Sensors*, vol. 23, no. 6, pp. 3264, 2023.

[12] V. D. Calhoun, T. Adali, G. D. Pearlson, et al., "A method for making group inferences from functional MRI data using independent component analysis," *Hum. Brain Mapp.*, vol. 14, no. 3, pp. 140–151, 2001.

[13] C. A. Tamminga, E. I. Ivleva, M. S. Keshavan, et al., "Clinical phenotypes of psychosis in the bipolar-schizophrenia network on intermediate phenotypes (b-snip)," *Am. J. Psychiatry*, vol. 170, no. 11, pp. 1263–1274, 2013.

[14] B. Gabrielson, M. Sun, M. A. B. S. Akhonda, et al., "Independent vector analysis with multivariate gaussian model: a scalable method by multilinear regression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2023, pp. 1–5.

[15] Q. Long, C. Jia, Z. Boukouvalas, et al., "Consistent run selection for independent component analysis: Application to fMRI analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2018, pp. 2581–2585.

[16] E. A. Allen, E. B. Erhardt, E. Damaraju, et al., "A baseline for the multivariate comparison of resting-state networks," *Front. Syst. Neurosci.*, vol. 5, pp. 2, 2011.