DECENTRALIZED SPATIALLY CONSTRAINED SOURCE-BASED MORPHOMETRY

Debbrata K. Saha, Rogers F. Silva, Bradley T. Baker, Vince D. Calhoun

Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS),
Georgia State University, Georgia Institute of Technology, and Emory University, Atlanta, GA 30303

2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761419

ABSTRACT

There is growing interest in extracting multivariate patterns (covarying networks) from structural magnetic resonance imaging (sMRI) data to analyze brain morphometry. Constrained source-based morphometry (constrained SBM) is a hybrid approach which provides a fully automated strategy for extracting subject-specific parameters characterizing gray matter networks. In constrained SBM, constrained independent component analysis (ICA) is used to compute maximally independent sources and statistical analysis is used to identify sources significantly associated with variables of interest. However, constrained SBM is built on the assumption that the data are locally accessible. As such, it cannot take advantage of decentralized (i.e., federated) data. While open data repositories have grown in recent years, there are various reasons (e.g., privacy concerns for rare disease data, institutional or IRB policies, etc.) that restrict a large amount of existing data to local access only. To overcome this limitation, we introduce a novel approach: decentralized constrained source-based morphometry (dcSBM). In our approach, data samples are located at different sites and each site operates the constrained ICA in a distributed manner. Finally, a master node simply aggregates result estimates from each local site and runs the statistical analysis centrally. We apply our method to UK Biobank sMRI data and validate our results by comparing to centralized constrained SBM results.

Index Terms— sMRI, SBM, Federated System, UK Biobank

1. INTRODUCTION

Structural magnetic resonance imaging (sMRI) is a powerful tool to analyze brain morphometry. A common approach to capture the brain structure changes is to divide the brain into regions of interest (ROI) and estimate the differences between groups. Voxel based morphometry (VBM) is another approach to identify the voxel-wise differences across the whole brain [1]. VBM, as a massive univariate approach, does not utilize the information of the relationship across voxels. Source based morphometry (SBM), a multivaiate data-driven approach was introduced to detect the whole brain structure automatically by utilizing the information across voxels [2]. SBM computes the spatially independent sources by using the combined techniques of ICA and VBM; and performs statistical analysis to identify the dominant sources to distinguish patients from healthy controls. SBM has been utilized to successfully examine different disorders [3] such as autism spectrum disorders [4], and Parkinson's [5].

Constrained source-based morphometry (constrained SBM) was proposed [6] as a hybrid approach that possesses the inherent advantages of SBM while also allowing for correspondence among datasets and automation. Unlike SBM, constrained SBM uses constrained-independent component analysis [ICA] [7]. In constrained SBM, a whole brain component template is incorporated as a prior constraint and the algorithm jointly optimizes independence to update the template as well as for similarity to the template. Constrained SBM can thus be used to compute structural networks across the whole brain in a fully automated manner.

Both SBM and constrained SBM approaches are built on the assumption that data are locally accessible. But collecting neuroimaging data is expensive and time consuming [8]. While open data have offered great benefits, in many cases there are challenges associated with data sharing and pooling related to regulatory concerns or to de-identification. Recent studies have shown that it is possible to identify specific subjects from a dataset consisting of patients with rare diseases [9, 10]. In the past few years, there has been extensive research to leverage data across multiple sites [11, 12, 13, 14, 15]. In this paper, we introduce a novel method: decentralized constrained source-based morphometry (dcSBM). In our approach, we perform decentralized constrained ICA across multiple sites. In dcSBM, we compute the most significant loading parameters from a linear regression model. Our results show one source in the somatomotor domain and another in the cognitive control domain. To the best of our knowledge, our proposed method is the first approach to implement constrained SBM in a decentralized manner.

2. METHODS

Constrained SBM is a multivariate alternative to the voxel-based morphometry (VBM) approach to investigate gray matter differences between patients and healthy controls. It utilizes a set of prior/reference maps to guide the source estimation. To deploy such approach on data located at different sites, we propose decentralized constrained SBM (dcSBM).

In the centralized case, we are tasked with finding gray matter differences using constrained SBM from a dataset $X = [x_1 \ldots, x_N]^\top$, where $x_i \in \mathbb{R}^V$ is the i-th subject's V-dimensional vector of real-valued features, with N subjects. In the decentralized setting with L sites, each site ℓ has dataset $D_\ell = \{(x_i, y_i) : i \in \{1, 2, ..., N_\ell\}\}$, where $y_i \in \mathbb{R}$ is the subject age. First, each local site ℓ runs constrained ICA separately on their local data X and obtains mixing and source matrices, A_ℓ and S_ℓ, respectively. A master node then aggregates all A_ℓ centrally for further analysis.

2.1. Data Acquisition and Preprocessing

T1 structural MRI data is acquired using straight sagittal orientation. By using the results of population brain size and shape from [16], the imaging matrix is automatically angled such a way that the front of the brain is tilted down by $16°$, with respect to the anterior commissure - posterior commissure line. The T1-weighted structural image consists of following parameters: Resolution: $1 \times 1 \times 1mm$, Field-of-view: $208 \times 256 \times 256$ matrix, Duration: 5 minutes, 3D MPRAGE, sagittal, in-plane acceleration iPAT=2, prescan-normalise. To include reasonable amounts of neck/mouth, the superior-inferior field-of-view is defined as $256mm$.

In our analysis, 3000 unaffected subjects are selected from the UK Biobank study. For subject-level preprocessing of T1-weighted sMRI data, the modulated gray matter probabilistic segmentation maps were generated from T1-weighted images using SPM12 [1]. These were smoothed using a Gaussian kernel with FWHM = $10mm$.

2.2. Spatially Constrained ICA

Constrained ICA is an enhanced ICA model that incorporates prior information into the decomposition process and extracts one or several desired independent sources S. A reference R is chosen to carry the prior information of the desired sources. Using a fast fixed-point algorithm, constrained ICA was performed on the subject-volume matrix X [7], which is embedded in the group ICA toolbox GIFT [2]. Based on the reference vector R, the source matrix S was extracted from the matrix X. The mixing matrix was also computed during this process. Thirty replicable spatial references (R) were identified based on separate group ICA results from two independent sMRI datasets in: 1) the human connectome project (HCP) and 2) the genomics superstruct project (GSP), each containing about 3500 unaffected subjects [17]. We then used sMRI data from 3000 consented subjects participating in the UK Biobank study. In our decentralized setting, we assigned $N_\ell = 1000$ subjects to each local site. The data on each site were processed by spatially constrained ICA. A one-dimensional vector of V in-brain gray matter voxels from each subject formed a $N_\ell \times V$ matrix X_ℓ.

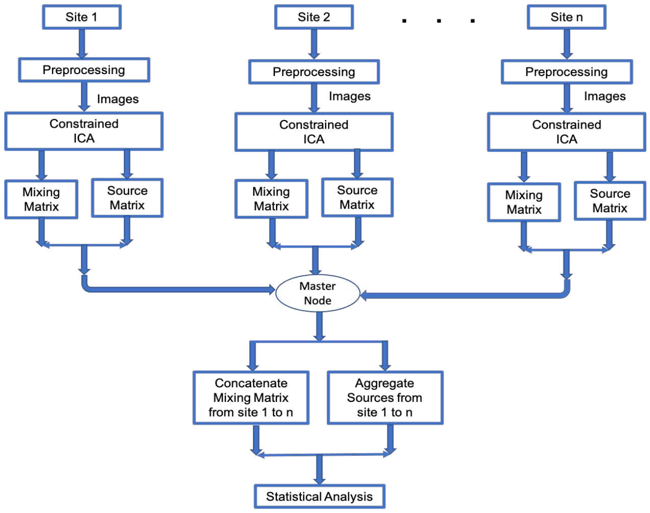

Fig. 1. Flow diagram of decentralized constrained SBM

Local processing by ICA resulted in local mixing and source matrices. In the mixing matrix, the scores in each column (also called loading parameters) represent how each component contributes to each subject. In the source matrix, the scores in each row represent statistically independent spatial configurations that highlight areas of coherent variability across subjects. We perform the same analysis to each local site and, finally, concatenate the mixing matrix from each site by means of simple aggregation to simplify the ensuing linear regression step (Section 2.4). We refer the reader to [18] for a fully decentralized regression approach that produces identical results as our simple aggregator. Strictly for the purposes of quality assessment of the proposed approach, we also aggregated the source matrices from each site. The procedure is shown in Figure 1.

To compare with the decentralized statistical results, we also run constrained ICA in a centralized setting. We pool all datasets together and run spatially constrained ICA on 3000 subjects. We collect the decomposed mixing matrix and source matrix and compare with the decentralized outputs.

2.3. Pairwise Correlation

In our statistical analysis, we computed the pairwise correlation coefficients between columns in the centralized and decentralized loading parameters. We also computed the pairwise correlation coefficients between rows of the centralized and average decentralized source matrices.

[1] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

[2] http://trendscenter.org/software/gift

2.4. Linear Regression

We run linear regression on both centralized and decentralized SBM models. We setup the predictor matrix Z with dimensions 3000×33, where variables 1 to 30 are the loading parameters from the mixing matrix. Column 31 is gender and the remaining columns contain site information. We used subject age as the response variable. Finally, we fit the data matrix Z with a linear regression model. From the analysis, we extracted the p value for each variable. We also created an effects plot of the predictors from our regression model using the **plotEffects** function in Matlab. The effect plot demonstrates the estimated main effect on the response variable by changing each predictor value. We also created the same analysis for the *centralized* constrained SBM to compare with our decentralized results.

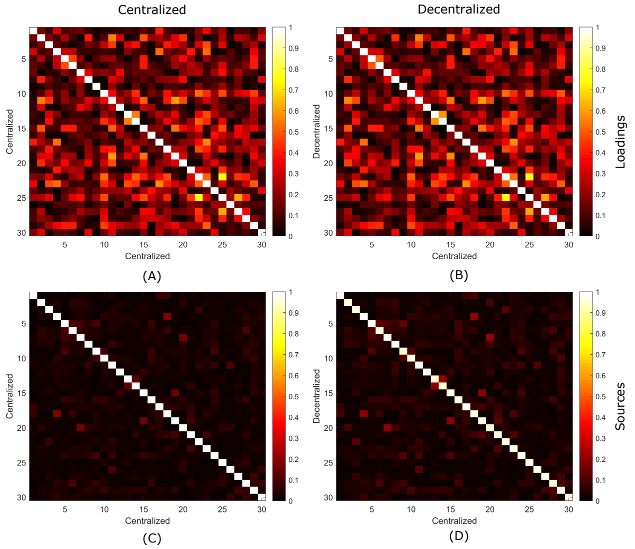

Fig. 2. Correlation plots of loading parameters and sources. (A) Correlation among centralized loadings only, (B) correlation between centralized vs decentralized loadings. (C) Correlation among centralized sources only, (D) correlation between centralized vs decentralized sources. (A,C) the *expected* similarity structure among loadings and sources, respectively, for the centralized case; (B,D) the *recovered* similarity structure between centralized and decentralized estimates, with near-1 correlations along the main diagonal.

3. RESULTS

Using spatially constrained ICA, we estimated thirty independent components in decentralized SBM. To compare with the decentralized case, we also estimated thirty components from the pooled dataset, and used centralized SBM.

We computed the correlation between the centralized and decentralized loading parameters. We present the correlation plot in Figure 2. In subplot (A), we present the correlations between columns of the loading matrix in centralized SBM. In subplot (B), we demonstrate the correlation between the centralized and decentralized loadings. From the centralized vs decentralized correlation plot, we observe the highest correlation across the diagonals. This observation demonstrates that the scores in loading parameter i in centralized SBM and loading parameter j (where $i == j$) in decentralized SBM are highly similar (c.f. the high correlation across the diagonal). We also evaluated the correlations between the centralized and average decentralized source matrix and presented the results in subplot (C) and (D). We also observe similar behaviour between the centralized and aggregate decentralized sources where we get the high correlations across the diagonal. To check the reliability of our algorithm, we repeated this experiment 10 times while randomly shuffling subjects across sites each time. We present the experimental results in Figure 3, where the correlation, Mean Square Error (MSE), Max Absolute Error (MaxAE), and Median Absolute Error (MedianAE) are between the correlation matrices in Fig. 2(A) and Fig. 2(B) for each of the 10 shuffled runs (similarly for Figs. 2(C) and (D)). Each boxplot contains 10 points from 10 shuffled runs, and we observe that all 10 values are very similar, indicating high reliability.

We then computed the p value for all the variables from the linear regression model for both centralized and decentralized experiments. Finally, we visualized the scatter plot of centralized versus decentralized p values. The results are presented in Figure 4. We only plotted the p values of loading parameters. Note that centralized and decentralized p values for the same number of loading parameters are very similar.

We also generated an effect plot to show the main effect on the response variable (age) while changing each predictor value (30 loading parameters, sex, and site). The experimental results of the effect plots for the centralized and decentralized cases are presented in Figure 5. In the figure, a horizontal line across the effect value implies the 95% confidence interval for the effect value. In the plot, we observe the similar effect of each variable in both centralized and decentralized cases. From visual inspection, we found the loading parameters 12 and 22 to have the largest effects on the response variable. We show the adjusted response plot using the Matlab function **plotAdjustedResponse** for variables 12 and 22 in the linear regression model. An adjusted response function describes the relationship between the fitted response and a single predictor, with the other predictors averaged out by averaging the fitted values over the data used in the fit. The results are presented in Figure 7. Panels (A) and (B) represent the relationship between the fitted response and predictor variables 12 and 22, respectively. Finally. we show the spatial maps of components 12 and 22 in Figure 6. Notice that one source is found within the somatomotor domain and the other one highlights the cognitive control domain. The expression level of these sources across subjects is associated with age.

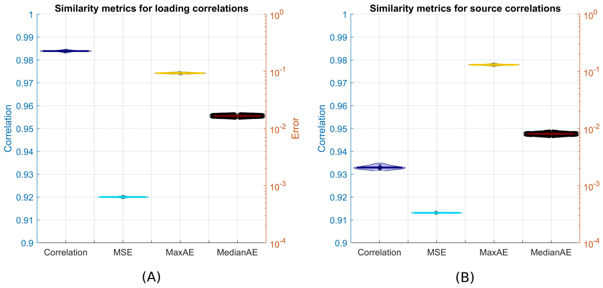

Fig. 3. Similarity metrics for (A) loading parameters and (B) sources. The left Yaxis is used for correlation. The right Yaxis is used for the remaining three metrics.

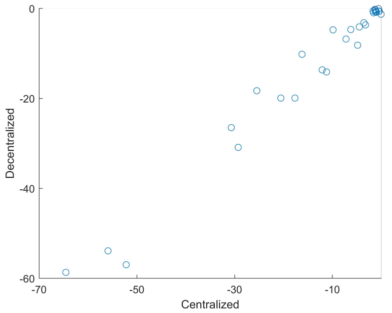

Fig. 4. Scatter plot of centralized and decentralized p values (log-log scale) indicating high consistency and overall agreement between centralized and decentralized estimates.

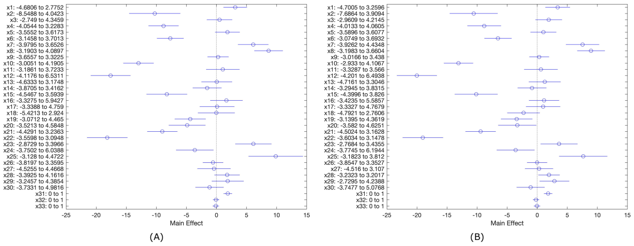

Fig. 5. Results for centralized and decentralized regression analysis. Panel (A) depicts the effect sizes for each loading estimated with the centralized approach, and (B) depicts the effect sizes for each loading estimated with the decentralized approach. Here, we observe very similar effect sizes in both centralized and decentralized analysis, suggesting consistency between the two approaches.

4. CONCLUSION

In this work, we have proposed a novel approach to perform constrained SBM in a distributed manner. We have identified and visualized two gray matter sources that are significantly associated with subject age. We compared our decentralized approach with the traditional (centralized) constrained SBM and showed that our results very closely approximate the centralized estimates but without any raw data sharing. The use of decentralized or federated approaches provides a powerful way to 1) leverage existing data which is required to stay locally private, and 2) integrate such data with openly available datasets. Decentralized approaches can also democratize computational resources, i.e., a single research group may not have the computational resources to analyze all datasets at a centralized location, but a larger federated consortium permits such analyses to be carried out and advance large-scale scientific studies. Although our demonstration artificially emulates sites, we fully expect that future work on real multi-site data will achieve the same performance, despite potential presence of site effects. We will also investigate the performance of dcSBM for group comparison in decentralized patient data.

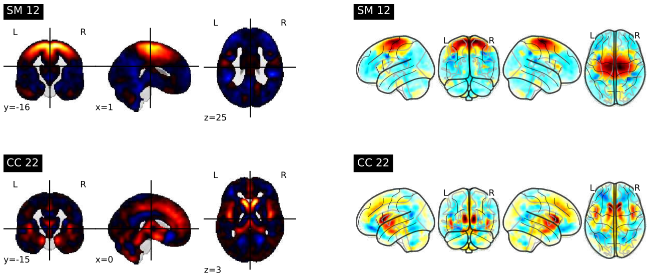

Fig. 6. Visual summary of sources 12 (supplementary motor area) and 22 (insula + caudate), which are typical of somatomotor and cognitive control domains, respectively. The panels to the left depict cross-sectional views while the panels to the right depict "glass brain" views.

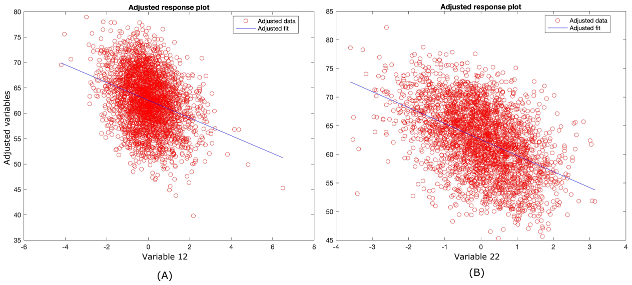

Fig. 7. Adjusted response plot for loadings corresponding to sources 12 (supplementary motor area) and 22 (insula + caudate). The plots depict the association between loading values (subject-specific expression levels) and age (adjusted variable).

5. ACKNOWLEDGMENTS

This work was funded by the NIH (R01DA040487) and the NSF (2112455). This research has been conducted using the UK Biobank Resource under Application Number 34175.

6. REFERENCES

[1] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N.A. Henson, K. J. Friston, and R. S.J. Frackowiak, "A voxel-based morphometric study of ageing in 465 normal adult human brains," *NeuroImage*, vol. 14, no. 1, pp. 21–36, 2001.

[2] L. Xu, K. Groth, G. Pearlson, D. Schretlen, and V. Calhoun, "Source-based morphometry: The use of independent component analysis to identify gray matter differences with application to schizophrenia," *Human brain mapping*, vol. 30, pp. 711–24, 03 2009.

[3] C. Gupta, J. Turner, and V. Calhoun, "Source-based morphometry: a decade of covarying structural brain patterns," *Brain Structure and Function*, vol. 224, pp. 1–14, 12 2019.

[4] A. Grecucci, D. rubicondo, R. Siugzdaite, L. Surian, and R. job, "Uncovering the social deficits in the autistic brain. a source-based morphometric study," *Frontiers in Neuroscience*, vol. 10, 08 2016.

[5] E. Premi, V. Calhoun, V. Garibotto, R. Turrone, A. Alberici, E. Cottini, A. Pilotto, S. Gazzina, M. Magoni, B. Paghera, B. Borroni, and A. Padovani, "Source-based morphometry multivariate approach to analyze [123i]fp-cit spect imaging," *Molecular Imaging and Biology*, vol. 19, 02 2017.

[6] L. Luo, L. Xu, R. Jung, G. Pearlson, T. Adali, and V. Calhoun, "Constrained source-based morphometry identifies structural networks associated with default mode network," *Brain connectivity*, vol. 2, pp. 33–43, 04 2012.

[7] Q. Lin, J. Liu, Y. Zheng, H. Liang, and V. D. Calhoun, "Semiblind spatial ica of fmri using spatial constraints," *Human Brain Mapping*, vol. 31, no. 7, pp. 1076–1088, 2010.

[8] D. Landis, W. Courtney, C. Dieringer, R. Kelly, M. King, B. Miller, R. Wang, D. Wood, J. A. Turner, and V. D. Calhoun, "Coins data exchange: An open platform for compiling, curating, and disseminating neuroimaging data," *NeuroImage*, vol. 124, pp. 1084–1088, 2016, Sharing the wealth: Brain Imaging Repositories in 2015.

[9] L. Sweeney, "k-anonymity: A model for protecting privacy1," *Int J Uncertain Fuzziness Knowl-Based Syst*, 2013.

[10] L. Sweeney, M. Crosas, and M. Bar-Sinai, "Sharing sensitive data with confidence: The datatags system," *Technol. Sci.*, 10 2015.

[11] S. M. Plis, A. D. Sarwate, D. Wood, C. Dieringer, D. Landis, C. Reed, S. R. Panta, J. A. Turner, J. M. Shoemaker, K. W. Carter, P. Thompson, K. Hutchison, and V. D. Calhoun, "Coinstac: A privacy enabled model and prototype for leveraging and processing decentralized brain imaging data," *Frontiers in Neuroscience*, vol. 10, pp. 365, 2016.

[12] J. Rissman and A. D. Wagner, "Distributed representations in memory: Insights from functional brain imaging," *Annual Review of Psychology*, vol. 63, no. 1, pp. 101–128, 2012, PMID: 21943171.

[13] D. K. Saha, V. D. Calhoun, Y. Du, Z. Fu, S. R. Panta, S. Kwon, A. D. Sarwate, and S. M. Plis, "Privacy-preserving quality control of neuroimaging datasets in federated environment," *bioRxiv*, 2021.

[14] H. Gazula, B. Holla, Z. Zhang, J. Xu, E. Verner, R. Kelly, S. Jain, R. Bharath, G. Barker, D. Basu, A. Chakrabarti, K. Kalyanram, K. Kumaran, L. Singh, R. Kuriyan, P. Murthy, V. Benega, S. Plis, A. Sarwate, and V. Calhoun, "Decentralized multisite vbm analysis during adolescence shows structural changes linked to age, body mass index, and smoking: a coinstac analysis," *Neuroinformatics*, 01 2021.

[15] D. K. Saha, V. D. Calhoun, S. R. Panta, and S. M. Plis, "See without looking: joint visualization of sensitive multi-site datasets," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2672–2678.

[16] M. Mennes, M. Jenkinson, R. Valabregue, J. K. Buitelaar, C. Beckmann, and S. Smith, "Optimizing full-brain coverage in human brain mri through population distributions of brain size," *NeuroImage*, vol. 98, pp. 513–520, 2014.

[17] Y. Du, Z. Fu, J. Sui, S. Gao, Y. Xing, D. Lin, M. Salman, A. Abrol, M. A. Rahaman, J. Chen, L. E. Hong, P. Kochunov, E. A. Osuch, and V. D. Calhoun, "Neuromark: An automated and adaptive ica based pipeline to identify reproducible fmri markers of brain disorders," *NeuroImage: Clinical*, vol. 28, pp. 102375, 2020.

[18] H. Gazula, B. T. Baker, E. Damaraju, S. M. Plis, S. R. Panta, R. F. Silva, and V. D. Calhoun, "Decentralized analysis of brain imaging data: Voxel-based morphometry and dynamic functional network connectivity," *Frontiers in Neuroinformatics*, vol. 12, 2018.