

Received 12 June 2024, accepted 20 July 2024, date of publication 29 July 2024, date of current version 16 August 2024. Digital Object Identifier 10.1109/ACCESS.2024.3435526

RESEARCH ARTICLE

Identifying the Relationship Structure Among Multiple Datasets Using Independent Vector Analysis: Application to Multi-Task fMRI Data

ISABELL LEHMANN^{®1}, (Graduate Student Member, IEEE), TANUJ HASIJA^{®1}, (Member, IEEE), BEN GABRIELSON^{®2}, MOHAMMAD A. B. S. AKHONDA², VINCE D. CALHOUN^{®3}, (Fellow, IEEE), AND TÜLAY ADALI^{®2}, (Fellow, IEEE)

¹Signal and System Theory Group, Paderborn University, 33098 Paderborn, Germany

²Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County (UMBC), Baltimore, MD 21250, USA ³Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, and

Emory University, Atlanta, GA 30303, USA

Corresponding author: Isabell Lehmann (isabell.lehmann@sst.upb.de)

This work was supported in part by German Research Foundation (DFG) under Grant SCHR 1384/3-2; in part by the NIH under Grant R01 MH18695, Grant R01 MH123610, and Grant R01 AG073949; and in part by NSF under Grant 2316420.

ABSTRACT Identifying relationships among multiple datasets is an effective way to summarize information and has been growing in importance. In this paper, we propose a robust 3-step method for identifying the relationship structure among multiple datasets based on Independent Vector Analysis (IVA) and bootstrap-based hypothesis testing. Unlike previous approaches, our theory-backed method eliminates the need for user-defined thresholds and can effectively handle non-Gaussian data. It achieves this by incorporating higher-order statistics through IVA and employing an eigenvalue decomposition-based feature extraction approach without distributional assumptions. This way, our method estimates more interpretable components and effectively identifies the relationship structure using hierarchical clustering. Simulation results demonstrate the effectiveness of our method, as it achieves perfect Adjusted Mutual Information (AMI) for different values of the correlation between the components. When applied to multi-task fMRI data from patients with schizophrenia and healthy controls, our method successfully reveals activated brain regions associated with the disorder, and identifies the relationship structure of task datasets that matches our prior knowledge of the experiment. Moreover, our proposed method extends beyond task datasets, offering broad applicability in subgroup identification in neuroimaging and other domains.

INDEX TERMS Blind source separation, bootstrap, data-driven, fMRI, independent vector analysis, relationship structure.

I. INTRODUCTION

In recent years, identifying relationships among multiple datasets has received growing attention in medical applications, such as for making group inferences when estimating brain activations [1] or for better localization of brain activity [2]. By defining a *grouping* as a subset of datasets that

The associate editor coordinating the review of this manuscript and approving it for publication was Cristian A. Linte.

share some latent characteristics, analyzing how the datasets are distributed in these groupings and how these groupings are related reveals the structure of these relationships. Identifying this *relationship structure* provides an even finer understanding of the complete dataset and therefore goes beyond only simple identification of the groupings.

One potential application of this is subgroup identification [3]. In this context, each dataset corresponds to a subject, and a grouping of subject datasets is called a subject grouping or *subgroup*. The identified subgroups can be used in multiple applications. For example, in Electronic Health Records (EHR) data, identifying subgroups can help uncover previously unknown connections between illnesses [4], while in precision medicine, estimating the dose of medication for a patient can be achieved based on the known dose of other patients in the same subgroup [5], [6]. Furthermore, by identifying the relationship structure among the datasets in a mixed cohort, for example, the associations among data of patients with schizophrenia, bipolar disorder, and their subtypes can be found.

There are multiple approaches for subgroup identification. For example, in [7], a subgroup identification method is proposed that allows the inclusion of any set of covariates for classification, while in [8], model-based partitioning is introduced for estimating a treatment effect function to identify subgroups that are linked to predictive factors through a decision tree. Many of these subgroup identification methods are either model-based [8] or heuristic approaches [7], and thus have strong assumptions or can be prone to subjective thresholds. Furthermore, they all are only identifying the subgroups, but not the relationship among them. These limitations highlight the need for further development and refinement of methods that can effectively learn the relationship structure among multiple datasets data without subjective choices. Here, data-driven techniques provide an alternative approach to the problem by extracting lower-dimensional features linked across multiple datasets. Joint Blind Source Separation (JBSS) techniques, especially those based on matrix decompositions such as Independent Component Analysis (ICA), summarize multiple datasets through latent variables, called sources/components, that are directly interpretable, without imposing strong constraints [9]. Based on these latent components, *subgroups* can be identified as subsets of datasets where the components within a subgroup are more similar to each other than to those outside the subgroup.

The well-known matrix decomposition method ICA is based on a linear mixture model, where a dataset is decomposed into a mixing matrix and sources, which are assumed to be statistically independent. ICA has been found to be successful for decomposing neuroimaging datasets, e.g., functional Magnetic Resonance Imaging (fMRI) datasets, where usually little is known about the underlying sources [9], [10]. ICA has been extended to analyze multi-set data using either the joint ICA [11] or the group ICA [10] models, which are limited in performance, though, as they assume a common mixing matrix or source vector for all datasets. Independent Vector Analysis (IVA) is a more flexible extension of ICA to multiple datasets that makes use of the statistical dependence of the source components across datasets [9], [12]. Over the years, there has been an increase in the utilization of IVA for fMRI data fusion [13], [14] because IVA 1) naturally aligns the sources across datasets, 2) effectively retains subject variability in multi-subject fMRI data [15], and 3) can identify common and distinct brain activations between groups of patients and controls [13], [14], [16].

In [13], subgroup identification is performed using IVA, where subgroups of patients with schizophrenia are found with a semi-heuristic procedure with the goal of better understanding the underlying heterogeneity of schizophrenia. An approach based on IVA and Gershgorin discs is used for subgroup identification in [16], but their method identifies subgroups within each Source Component Vector (SCV) separately, which is less robust compared with taking the information from several SCVs together into account. A more comprehensive review of data-driven subgroup identification methods can be found in [17]. Furthermore, these approaches only identify the subgroups, but do not reveal their full relationship structure. For example, in [18], hierarchical clustering is used for identifying groupings of representative rules that are used for subgroup discovery instead of directly revealing the relationship among different datasets. Finally, the authors of [14] show that information about the relationship of the datasets can be visually seen in the covariance matrices of the SCVs estimated by IVA, but they do not provide an objective method for fusing the information of all SCVs together to discover the full relationship structure. Herein lies the novelty of our paper: we do not only discover groupings of the datasets but also identify their relationship structure. Moreover, we leverage information from all SCVs collectively, compared with considering each SCV individually, which enhances the robustness of our method.

More precisely, in this paper, we propose a powerful method for identifying the relationship structure among multiple datasets by leveraging key properties of IVA. To the best of our knowledge, no other method exists to achieve this goal. In our 3-step method, 1) a latent representation of the datasets is found using IVA, 2) the SCVs are identified that contain information about the relationship of the datasets, and 3) features are extracted from these SCVs as the input of a hierarchical clustering algorithm. By finding how close these datasets are to each other, our method does not only identify groupings of the datasets but also fully discovers their relationship structure in a resulting dendrogram. We summarize the contributions of this work as follows:

- developing a method for identifying the relationship structure among multiple datasets,
- verifying the success of our method by identifying a meaningful relationship structure in multi-task fMRI datasets, as the method groups together the task datasets from similar tasks, and
- revealing interpretable latent components in the fMRI data and observing significantly stronger deactivation of the Default Mode Network (DMN) areas in patients with schizophrenia compared with healthy controls.

With simulated datasets, we experimentally demonstrate the superior performance of our proposed method compared with the competing techniques. Our paper is organized as follows: We formulate the problem in section II and explain our method for solving it in section III. Then, we evaluate our method in simulations in section IV, apply it to real multi-task fMRI data in section V, and finally summarize and discuss the main findings of our paper in section VI.

II. PROBLEM FORMULATION

Let $\mathbf{x}^{[k]} \in \mathbb{R}^R$ denote the k^{th} dataset, where R denotes the dimension of the dataset.

Given K datasets, identify their relationship structure, i.e., group datasets based on the similarity (high statistical dependence) between their latent variables. Note that this cannot be achieved through simple clustering approaches, as revealing the relationship structure relies on the dependence of the latent variables across datasets.

The datasets are assumed to be generated according to the following model:

$$\mathbf{x}^{[k]} = \mathbf{A}^{[k]} \mathbf{s}^{[k]}, \quad k = 1, \dots, K,$$
(1)

where $\mathbf{s}^{[k]} \in \mathbb{R}^R$ is the latent source vector containing *R* source components, and $\mathbf{A}^{[k]} \in \mathbb{R}^{R \times R}$ is the unknown mixing matrix. For example, in fMRI data, *R* can correspond to the number of subjects, and the sources are spatial maps.

Let the r^{th} Source Component Vector (SCV) be defined as the concatenation of the r^{th} source component of all *K* datasets [9]:

$$\mathbf{s}_r = \begin{bmatrix} s_r^{[1]} & \dots & s_r^{[K]} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^K, \tag{2}$$

where $s_r^{[k]} \in \mathbb{R}$ denotes the r^{th} source component of the k^{th} dataset. Without loss of generality, the source components are assumed to be zero-mean and unit-variance. Furthermore, we assume the source components to be independent among SCVs, which is a common assumption for identifying source components by only observing datasets [19]. This way, the definition of an SCV allows to capture the dependence information across multiple datasets within an SCV. As the source components are zero-mean and unit-variance, the covariance matrix of the r^{th} SCV coincides with its correlation matrix and can be defined as

$$\mathbf{C}_r = \mathbf{E}\left[\mathbf{s}_r \mathbf{s}_r^{\mathsf{T}}\right] \in \mathbb{R}^{K \times K}.$$
(3)

In [13] and [14], it is shown that information about the relationship of the datasets can be inferred from the linear dependence (correlation) of the source components across datasets. This information is revealed in the covariance matrices of the SCVs. An SCV covariance matrix with all non-zero off-diagonal values implies that all components within that SCV are dependent, and hence describes a *common* SCV, i.e., with the components common across all datasets [13]. Similarly, an SCV covariance matrix exhibiting a block-diagonal structure (with off-block elements equal to 0) implies that all components within each block are dependent. We call SCVs with these covariance matrices *structured*. Note that this also holds for covariance matrices that can be transformed into block-diagonal matrices using

III. METHOD

In this paper, we propose a method for identifying the relationship structure among multiple datasets. Figure 1 shows the three steps of which our method consists: 1) estimation of latent sources, 2) identification of common and structured SCVs, and 3) identification of the relationship structure using structured SCVs. The following sections explain the details of each step.

A. STEP 1: ESTIMATION OF SCVS

To be able to infer information about the relationship structure from the SCV covariance matrices, the unknown source components in the SCVs must be estimated from the observed datasets. For this, we use IVA.

1) INDEPENDENT VECTOR ANALYSIS

The generative model for IVA is given by (1). The goal of IVA is to jointly estimate the *K* source vectors [12]

$$\widehat{\mathbf{s}}^{[k]} = \mathbf{W}^{[k]} \mathbf{x}^{[k]},\tag{4}$$

where $\hat{\mathbf{s}}^{[k]} \in \mathbb{R}^R$ denotes the estimate of $\mathbf{s}^{[k]}$, and $\mathbf{W}^{[k]} \in \mathbb{R}^{R \times R}$ is the demixing matrix for the *k*th observed dataset. The degree of independence is measured by mutual information among the estimated SCVs, defined as [9]

$$I(\widehat{\mathbf{s}}_{1}; \ldots; \widehat{\mathbf{s}}_{R}) = \sum_{r=1}^{R} \left(\sum_{k=1}^{K} H\left(\widehat{\mathbf{s}}_{r}^{[k]}\right) - I\left(\widehat{\mathbf{s}}_{r}\right) \right)$$
$$- \sum_{k=1}^{K} \log \left| \det\left(\mathbf{W}^{[k]}\right) \right| - C, \quad (5)$$

where $I(\widehat{s}_r)$ is the mutual information of the K source components in the r^{th} estimated SCV $\hat{\mathbf{s}}_r$, $H\left(\hat{\mathbf{s}}_r^{[k]}\right)$ denotes the entropy of $\hat{s}_r^{[k]}$, and C is a constant term. By minimizing the mutual information among the estimated SCVs, IVA maximizes independence among the SCVs, while simultaneously maximizing the mutual information within each SCV. Through the selection of an appropriate multivariate density model, IVA can take either second-order or all-order statistics into account [9]. IVA can correctly estimate the SCVs if and only if there exist no subsets of source components within two SCVs that meet the following conditions: 1) the source components in the subsets are Gaussian distributed and independent of the other source components within the same SCV, and 2) the source components in the subsets of the two SCVs have proportional covariance matrices [9]. Hence, IVA is able to identify a very broad class of signals.







FIGURE 1. Visualization of our proposed method for identifying the relationship structure among multiple datasets. We have K observed datasets of dimension $R \times V$, where V is the number of voxels and R the (reduced) dimension of the data. In step 1, the latent sources in the SCVs \hat{S}_r are estimated by applying IVA on the observed datasets. In step 2, the SCVs are identified as structured or common by applying an eigenvalue decomposition (EVD) on their covariance matrices and using the proposed bootstrap technique (BT) to estimate \hat{d}_r , i.e., the number of eigenvalues greater than 1. If $\hat{d}_r = 1$, the r^{th} SCV is identified as common, and if $\hat{d}_r > 1$, as structured. We denote the indices of the structured SCVs as $S = \{i_1, \dots, i_{|S|}\}$. In Step 3, the \hat{d}_r leading eigenvectors of the structured SCVs are concatenated to form a feature matrix, which is the input of hierarchical clustering. The resulting clusters are the identified groupings (in this example, there are two groupings: orange and green), and the dendrogram reveals the relationship structure among the datasets.

IVA-L-SOS [15] assumes that the SCVs underlay a multivariate Laplacian distribution while allowing for a non-identity covariance matrix. This way, dependence is not only measured by higher-order statistics, but correlations, i.e., second-order statistics, are also taken into account. As it has been shown that the modeling assumptions of IVA-L-SOS are a good match for the properties of real fMRI data [13], [15], we will use it in our method. However, if our method is applied to data with different source distributions, Step 1 can be replaced by another source separation method with appropriate assumptions for that type of data.

2) SAMPLES

In practice, we observe V samples of each dataset $\mathbf{x}^{[k]}$, which form the observed datasets $\mathbf{X}^{[k]} \in \mathbb{R}^{R \times V}$. Then, the estimated source matrices are

$$\widehat{\mathbf{S}}^{[k]} = \mathbf{W}^{[k]} \mathbf{X}^{[k]} \in \mathbb{R}^{R \times V}, \tag{6}$$

and the r^{th} estimated SCV is defined as

$$\widehat{\mathbf{S}}_{r} = \begin{bmatrix} \left(\widehat{\mathbf{s}}_{r:}^{[1]} \right)^{\mathsf{T}} & \dots & \left(\widehat{\mathbf{s}}_{r:}^{[K]} \right)^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{K \times V}, \qquad (7)$$

where $\widehat{\mathbf{s}}_{r:}^{[k]} \in \mathbb{R}^{1 \times V}$ denotes the r^{th} row of $\widehat{\mathbf{S}}^{[k]}$, i.e., the r^{th} estimated source component of the k^{th} dataset. The estimation of the SCVs with samples is shown in Step 1 in Figure 1.

B. STEP 2: IDENTIFICATION OF COMMON AND STRUCTURED SCVS

In Step 2, we must determine which of the estimated SCVs are common and which are structured. Only the structured SCVs contain information about the relationship between the datasets, and thus, only these structured SCVs are utilized in Step 3 to determine features for identifying the relationship structure.

The SCVs estimated by IVA-L-SOS can be identified as either common or structured based on the eigenvalues of their covariance matrices. Let the Eigenvalue Decomposition (EVD) of the true covariance matrix of the r^{th} SCV C_r (defined in (3)) be

$$\mathbf{C}_r = \mathbf{U}_r \operatorname{diag}\left(\mathbf{\lambda}_r\right) \mathbf{U}_r^{\mathsf{T}} \tag{8}$$

where $\mathbf{U}_r \in \mathbb{R}^{K \times K}$ contains the eigenvectors of \mathbf{C}_r as columns, and $\lambda_r \in \mathbb{R}^K$ contains the corresponding eigenvalues. When $\mathbf{C}_r = \mathbf{I}$, where \mathbf{I} denotes the identity matrix, all its eigenvalues are equal to 1. However, when $\mathbf{C}_r \neq \mathbf{I}$, some eigenvalues are different from 1. Let d_r be defined as the number of eigenvalues of \mathbf{C}_r that are greater than 1. We assume that the SCV covariance matrices have only one of the following two structures:

- For *common* SCVs, where all the source components within an SCV are correlated with each other, the covariance matrix C_r has 1s on the diagonal, and all off-diagonal values (corresponding to the correlation coefficients) are non-zero. In [20], it is shown that, under certain conditions on correlation coefficients, C_r has exactly one eigenvalue greater than 1, i.e., $d_r = 1$.
- For *structured* SCVs, the covariance matrix C_r is block-diagonal with 1s on the diagonal and zero off-block values (or can be permuted into a block-diagonal matrix $\tilde{C}_r = PC_rP^T$, where $P \in \mathbb{R}^{K \times K}$ is an orthogonal permutation matrix). We assume that the covariance matrices of structured SCVs have at least two blocks. To determine the number of eigenvalues greater than 1 for such SCV covariance matrices, we present the following corollary:

Corollary 1: Let $\mathbf{C}_r \in \mathbb{R}^{K \times K}$ be a block-diagonal matrix consisting of d_r blocks $\mathbf{B}_i \in \mathbb{R}^{K_i \times K_i}$ such that

$$\mathbf{C}_r = \begin{bmatrix} \mathbf{B}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \mathbf{B}_{d_r} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{I} \end{bmatrix},$$

and \mathbf{B}_i contains 1s on the diagonal and positive entries on its off-diagonal elements. Then, \mathbf{C}_r has exactly d_r eigenvalues greater than one.

As it is proven in [20] that $|\{\lambda(\mathbf{B}_i) > 1\}| = 1$ (and the remaining eigenvalues are less than or equal to 1), it follows naturally for d_r blocks in \mathbf{C}_r that $|\{\lambda(\mathbf{C}_r) > 1\}| = d_r$. Note that Corollary 1 also holds if $\widetilde{\mathbf{C}}_r$ (and not \mathbf{C}_r) is block-diagonal because $\widetilde{\mathbf{C}}_r$ has the same eigenvalues as \mathbf{C}_r .

Under the assumption that a structured SCV consists of at least two blocks, we can differentiate between a common SCV $(d_r = 1)$ and a structured SCV $(d_r > 1)$ by counting the number of eigenvalues greater than 1. We define $C = \{r : d_r = 1\}$ as the set of indices of the common SCVs and $S = \{r : d_r > 1\}$ as the set of indices of the structured SCVs. In the following, we denote the indices in S as

$$\mathcal{S} = \{i_1, \dots, i_{|\mathcal{S}|}\}.$$
(9)

Note that we do not consider SCVs with identity covariance matrices, which would correspond to completely uncorrelated datasets, because we assume that in real data some correlations typically exist in each SCV.

Now, when we estimate covariance matrices over finite samples V,

$$\widehat{\mathbf{C}}_r = \frac{1}{V} \widehat{\mathbf{S}}_r \widehat{\mathbf{S}}_r^\mathsf{T},\tag{10}$$

the estimated correlation coefficients corresponding to the uncorrelated datasets will not be exactly zero, and thus, more than d_r eigenvalues for a structured SCV will be greater than 1. Thus, by just counting the eigenvalues greater than 1, d_r would be overestimated. Consequently, it is necessary to estimate d_r . Estimating the number of significant eigenvalues is commonly addressed as a model-order selection problem in the literature [21]. However, these model-order selection techniques assume certain asymptotic properties (as $V \to \infty$) on the non-significant eigenvalues, for example, assuming they all or a subset of them are equal to each other (in [20] and [21]). This is not applicable for the SCV covariance matrices. For example, if all the source components in an SCV covariance matrix would belong to one of the blocks, and thus there would be no uncorrelated source components, then for arbitrary correlation coefficients, none of the non-significant eigenvalues would be equal to each other.

To estimate d_r for the r^{th} SCV covariance matrix \mathbf{C}_r , we perform a binary hypothesis test for each $k = 1, \ldots, K$ with the null hypothesis

$$H_0: \quad \lambda_r^{[k]} \le 1, \tag{11}$$

and the alternative

$$H_1: \quad \lambda_r^{[k]} > 1. \tag{12}$$

Here, $\lambda_r^{[k]}$ is the k^{th} eigenvalue of \mathbb{C}_r . As in practice we only estimate $\widehat{\mathbb{C}}_r$, we only know the estimated eigenvalues $\widehat{\lambda}_r^{[k]}$. We define a test statistic $\widehat{T} = \widehat{\lambda}_r^{[k]} - 1$, and to perform the hypothesis test, we must know the distribution of the statistic under H_0 [22]. Neither the sample nor the asymptotic distribution of the test statistic is known. We, therefore, propose a bootstrap-based hypothesis test [23] to estimate this distribution. Under certain conditions, the distribution estimated by bootstrap converges to the true distribution if the number of samples goes to infinity $(V \to \infty)$ [24], and thereby we can estimate \widehat{d}_r for each SCV.

The pseudocode for our method is described in Algorithm 1. In the following, we describe the steps of the algorithm.

1) The sample correlation matrix $\widehat{\mathbf{C}}_r = \left| \frac{1}{V} \widehat{\mathbf{S}}_r \widehat{\mathbf{S}}_r^{\mathsf{T}} \right|$ is calculated (line 1). The absolute value is necessary because of a possible sign ambiguity in the sources. Then, an EVD is applied on $\widehat{\mathbf{C}}_r$ to get the eigenvalues $\widehat{\boldsymbol{\lambda}}_r = \left[\widehat{\lambda}_r^{[1]}, \dots, \widehat{\lambda}_r^{[K]} \right]$, sorted in descending order (line 2). Here, $\widehat{\boldsymbol{\lambda}}_r^{[K]}$ denotes the estimate of $\lambda_r^{[K]}$.

Algorithm 1 Bootstrap Algorithm for Estimating \hat{d}_r **Input:** $\widehat{\mathbf{S}}_r \in \mathbb{R}^{K \times V}, B, P_{\mathrm{fa}}$ 1: $\widehat{\mathbf{C}}_r \leftarrow \operatorname{abs}\left(\frac{1}{V}\widehat{\mathbf{S}}_r\widehat{\mathbf{S}}_r^{\mathsf{T}}\right)$ \triangleright s.t. $\widehat{\lambda}_r^{[1]} > \cdots > \widehat{\lambda}_r^{[K]}$ 2: $\widehat{\lambda}_r, \widehat{\mathbf{U}}_r \leftarrow \mathrm{EVD}(\widehat{\mathbf{C}}_r)$ 3: for b = 1, ..., B do $\triangleright V$ integers $\sim \mathcal{U}_{int}(1, V)$ $_{b}\mathbf{j} \leftarrow \text{randint}(1, V, V)$ 4: $b\widehat{\mathbf{S}}_{r} \leftarrow \widehat{\mathbf{S}}_{r}(:, b\mathbf{j}) \\ b\widehat{\mathbf{C}}_{r} \leftarrow abs\left(\frac{1}{V}b\widehat{\mathbf{S}}_{r}b\widehat{\mathbf{S}}_{r}^{\mathsf{T}}\right) \\ b\widehat{\mathbf{\lambda}}_{r}, \ b\widehat{\mathbf{U}}_{r} \leftarrow EVD(b\widehat{\mathbf{C}}_{r})$ 5: 6: 7: 8: end for for k = 1, ..., K do $\widehat{T} \leftarrow \widehat{\lambda}_r^{[k]} - 1$ 9: 10: for $b = 1, \dots, B$ do ${}_{b}\widehat{T}^{*} \leftarrow {}_{b}\widehat{\lambda}_{r}^{[k]} - \widehat{\lambda}_{r}^{[k]}$ 11: 12: end for 13: \triangleright s.t. $_1\widehat{T}^* \leq \cdots \leq _B\widehat{T}^*$ $\widehat{\mathbf{T}}^* \leftarrow \operatorname{sort}(\widehat{\mathbf{T}}^*)$ 14: $\eta = \lfloor (B+1) \cdot (1-P_{\rm fa}) \rfloor$ 15: if $\widehat{T} < {}_{\eta}T^*$ then 16: $\triangleright \lambda_r^{[k]} < 1$ $\epsilon^{[k]} \leftarrow 0$ 17: else 18: $\epsilon^{[k]} \leftarrow 1$ \triangleright H_0 is rejected. $\lambda_r^{[k]} > 1$ 19: end if 20: 21: end for 22: $d_r = \text{count nonzero}(\epsilon) \triangleright \text{Number of } \epsilon \text{ entries equal 1}$ **Output:** dr

- 2) The SCVs are resampled with replacement for *B* times, where for each resampling, V indices are drawn from $\mathcal{U}_{int}(1, V)$ (uniform distribution of integers between 1 and V) (line 4), and the SCVs are resampled on those indices (line 5). The resampled SCVs $_{b}\widehat{\mathbf{S}}_{r}$ thus also have V samples. The prescript b denotes the b^{th} bootstrap resample.
- 3) The eigenvalues ${}_{b}\widehat{\boldsymbol{\lambda}}_{r}$ from the covariance matrices ${}_{b}\widehat{\boldsymbol{C}}_{r} = \left|\frac{1}{V}{}_{b}\widehat{\boldsymbol{S}}_{r}{}_{b}\widehat{\boldsymbol{S}}_{r}\right|$ are calculated (lines 6–7).
- (line 12).
- 6) The values of ${}_{b}\widehat{T}^{*}$ are sorted in ascending order (line 14). Using a given false alarm probability P_{fa} , a threshold $\eta \widehat{T}^*$ is found, where $\eta = \lfloor (B+1) \cdot (1-P_{fa}) \rfloor$ (line 15), and $_{\eta}\widehat{T}^*$ is $_b\widehat{T}^*$ for $b = \eta$. If $\widehat{T} > _{\eta}\widehat{T}^*$, then H_0 is rejected, i.e., $\lambda_r^{[k]}$ is greater than 1 with a significance of $1 - P_{fa}$, and the k^{th} element in a vector ϵ is set to 1 (lines 16–20).
- 7) In the end, the number of 1s in ϵ is counted, which equals to \hat{d}_r , the estimated number of eigenvalues greater than 1 in the r^{th} SCV (line 22).

C. STEP 3: IDENTIFICATION OF THE RELATIONSHIP STRUCTURE USING STRUCTURED SCVS

In Step 3, the eigenvectors of the covariance matrices of the structured SCVs are used as features for the hierarchical clustering. In [20], it is shown that the eigenvector corresponding to an eigenvalue greater than 1 characterizes the correlated datasets for the corresponding block. More specifically, for each eigenvalue greater than 1, the eigenvector element corresponding to a dataset that is not part of the correlated datasets is 0, while the eigenvector elements of the correlated datasets are greater than 0. This means, the d_r leading eigenvectors contain information about the relationship of the datasets within the r^{th} SCV. As stated in Step 2, the covariance matrices are estimated from finite samples, and d_r is also estimating using the proposed bootstrap-based hypothesis test. We will thus use this estimated \hat{d}_r from Step 2. As different SCVs provide complementary information about the relationship of the datasets, the \hat{d}_r leading eigenvectors ($r \in S$) from all structured SCV covariance matrices are horizontally concatenated to form a feature matrix, **F**, of $\sum_{r \in S} \hat{d}_r$ eigenvector columns, which is then fed into the hierarchical clustering:

$$\mathbf{F} = \operatorname{concat}\left(\widehat{\mathbf{U}}_{i_{1}}\left(:, 1:\widehat{d}_{i_{1}}\right), \dots, \widehat{\mathbf{U}}_{i_{|\mathcal{S}|}}\left(:, 1:\widehat{d}_{i_{|\mathcal{S}|}}\right)\right) \\ \in \mathbb{R}^{K \times \sum_{r \in \mathcal{S}} \widehat{d}_{r}}.$$
 (13)

This way, our method leverages the knowledge of all SCVs together instead of performing an analysis separately on each SCV. The advantage of using hierarchical clustering, compared with, e.g., K-means clustering, is that no prior knowledge or estimation of the number of clusters is necessary. Additionally, while K-means clustering would only identify the groupings of datasets, which correspond to the resulting clusters, hierarchical clustering also identifies the relationship structure among the datasets, which is revealed in the dendrogram.

D. COMPUTATIONAL COMPLEXITY

We compute the big- \mathcal{O} complexity for each step of the proposed method. In Step 1, the complexity is dominated by the multiplication of $\widehat{\mathbf{C}}_r^{-1} \in \mathbb{R}^{K \times K}$ with $\widehat{\mathbf{S}}_r \in \mathbb{R}^{K \times V}$ in the main loop of IVA-L-SOS. In each iteration, $\widehat{\mathbf{C}}_{r}^{-1}$ is updated K times per SCV (and thus this multiplication is performed Ktimes per SCV), which results for all R SCVs in a complexity of $\mathcal{O}(IRK^3V)$ for I iterations. In Step 2, the dominant cost is the calculation of the covariance matrices of the resampled SCVs $_{b}\widehat{\mathbf{S}}_{r} \in \mathbb{R}^{K \times V}$, which for all *B* bootstrap resamples in all R SCVs has complexity $\mathcal{O}(BRK^2V)$. In Step 3, the cost is dominated by the hierarchical clustering of the K datasets, which is of complexity $\mathcal{O}(K^3)$ [25]. Thus, our method has a big- \mathcal{O} complexity of $\mathcal{O}(RK^2V(IK+B))$.

IV. SIMULATIONS

To demonstrate the performance of our proposed bootstrap technique, we simulate common and structured SCVs.¹ We generate the entries of the SCV covariance matrices C_r

¹Python code available at https://github.com/SSTGroup/relationship_ structure_identification



FIGURE 2. R = 6 simulated SCV covariance matrices (of dimension 10 × 10) for $\rho = 0.2$. We have $d_1 = 1$, $d_2 = d_3 = d_4 = 2$, $d_5 = 3$, and $d_6 = 4$. Using these covariance matrices, 6 SCVs with Laplacian-distributed sources are generated.

according to the following model:

$$\mathbf{C}_{r}(k,l) = \begin{cases} 1, & k = l \quad \text{(diagonal elements)} \\ 0, & k \neq l, \mathbf{C}_{r}(k,l) \notin \mathbf{B}_{i}, 1 \leq i \leq d_{r} \\ \rho + n_{r}(k,l) & k \neq l, \mathbf{C}_{r}(k,l) \in \mathbf{B}_{i}, 1 \leq i \leq d_{r} \end{cases}$$
(14)

where \mathbf{B}_i is the *i*th block, ρ is the correlation coefficient of the correlated sources, and $n_r(k, l) = n_r(l, k) \sim \mathcal{N}(0, 0.0025)$ is added variability to fulfill the identifiability conditions of IVA [9]. Figure 2 shows the structure of the six 10×10 SCV covariance matrices in our simulations for $\rho = 0.2$. C_1 corresponds to a common SCV, i.e., $d_1 = 1$, while the other SCVs are structured. C_2 , C_3 and C_4 all have two blocks, i.e., $d_2 = d_3 = d_4 = 2$, C₅ has 3 blocks, i.e., $d_5 = 3$, and C_6 has 4 blocks, i.e., $d_6 = 4$. Note that C_3 and C_4 contain 1 and 4 uncorrelated source components, respectively. Using these SCV covariance matrices, we generate R = 6 SCVs $\mathbf{S}_r \in \mathbb{R}^{K \times V}$, each with K = 10 source components and V = 1000 samples. The samples are drawn from a Laplacian distribution as described in [26] (in section 6.4), with zero mean and with covariance matrices specified as shown in Figure 2.

We perform two experiments. In the first (section IV-A), we compare the performance of our proposed bootstrap technique and two competing methods for estimating d_r from the true generated SCVs. We use the true sources here because the identified sources from IVA may be estimated correctly but not aligned correctly among SCVs, i.e., the correlated blocks in the covariance matrices may be permuted among SCVs. Due to this, \hat{d}_r of the output of IVA may not match \hat{d}_r of the generated sources. In a second experiment (section IV-B), we apply the complete method, including the IVA step and the clustering, on simulated observed datasets and investigate its robustness for different correlation coefficients. In both of these scenarios, we did not consider noise as we assumed that for real data, noise is effectively removed during a PCA-based dimension reduction step in the preprocessing. This is based on the assumption that the problem is essentially overdetermined, i.e., there are more



FIGURE 3. Boxplot of $P(\hat{d}_r = d_r)$, r = 1, ..., 6, for different values of ρ for the BT, EV and GD technique. On average, BT is always superior to EV and GD.



FIGURE 4. Average value of \hat{d}_r for SCVs 4 and 5 for different values of ρ for the BT, EV and GD technique. The true values are $d_4 = 2$ and $d_5 = 3$. The BT technique estimates both values close to the ground truth.

observations than underlying source components of interest, which is the common scenario in most applications including ours.

A. ESTIMATING \widehat{d}_r

We evaluate the performance of three techniques for estimating \hat{d}_r from the true sources. The first method is our proposed bootstrap (BT) technique, described in Algorithm 1, with B = 1000 bootstrap resamples and $P_{fa} = 0.05$. The second method (EV) directly counts how many eigenvalues are greater than 1 in the r^{th} SCV, i.e., $\hat{d}_r = |\{k : \lambda_r^{[k]} > 1\}|$. The third method is the Gershgorin Disc (GD) technique from [16].

We simulate 50 Monte-Carlo runs. We investigate the two performance metrics

$$P(\widehat{d}_r = d_r) = \frac{\#(\widehat{d}_r = d_r)}{\#\text{runs}},$$
(15)

which estimates the probability that \hat{d}_r is estimated correctly, and $\mu_{\hat{d}_r}$, which is the average value of the estimated \hat{d}_r .

The boxplot of $P(\hat{d}_r = d_r)$ for the R = 6 SCVs is shown in Figure 3 for different values of ρ , with circles denoting outliers. Notably, our proposed BT technique demonstrates robust performance by accurately estimating \hat{d}_r even for very small correlation values in the underlying data, showcasing its effectiveness in handling the Laplacian (non-Gaussian) data distribution and varying correlation values ρ . In contrast, the EV technique only achieves high $P(\hat{d}_r = d_r)$ values with increasing ρ , while the GD technique does not perform well for all values of ρ .

To investigate the reason for small values of $P(\hat{d}_r = d_r)$ in the EV and GD technique, we show $\mu_{\hat{d}_4}$ and $\mu_{\hat{d}_5}$ in Figure 4, estimated for the BT, EV and GD techniques. We see (blue triangles) that the EV technique overestimates d_4 (also d_3 , not shown here), with an improvement for higher ρ values, while the BT (circles) and GD (squares) techniques estimate the value close to the ground truth for all ρ values. In orange squares, we see that the GD technique strongly underestimates d_5 (also d_2 and d_6 , not shown) compared with the BT (circles) and EV (triangles) techniques for all ρ values. The over- and underestimation of \hat{d}_r in the EV and GD techniques is the reason for the decreased values of $P(\hat{d}_r = d_r)$ compared with the BT technique. Thus, the BT technique is superior to EV and GD in terms of estimating \hat{d}_r .

B. IDENTIFYING THE RELATIONSHIP STRUCTURE

As we have shown that the BT technique is superior to EV and GD in terms of estimating \hat{d}_r , we evaluate our complete method for identifying the relationship structure only using BT. We are not aware of other methods to identify the relationship structure, and not only groupings of the datasets, with whom we could compare our method.

We generate the SCVs as in section IV-A and the datasets as $\mathbf{X}^{[k]} = \mathbf{A}^{[k]}\mathbf{S}^{[k]}$, where the elements of $\mathbf{A}^{[k]} \in \mathbb{R}^{6\times 6}$ are drawn from $\mathcal{N}(0, 1)$. After performing IVA-L-SOS with 20 random initializations and choosing the estimated sources of the most consistent run² as in [27], the BT technique is applied to the estimated SCVs to estimate \hat{d}_r . The \hat{d}_r leading eigenvectors of the structured SCVs are concatenated in a feature matrix **F**, which we feed into the hierarchical clustering. Hierarchical clustering is performed using the linkage function from scikit-learn [28] with the 'ward' linkage (minimizing the variance of the clusters) and visualized using the dendrogram function.

Figure 5 shows the ground truth of the dendrogram for clustering the datasets using the SCV covariance matrices of this experiment. Ground truth means that we set d_r manually and use the d_r eigenvectors of the true (instead of sample) covariance matrix for clustering. In the figure, the labels of the datasets are denoted on the x-axis, and the y-axis shows the distance between the clusters. A small distance means that the corresponding clusters are very similar. We see that three groupings exist in the datasets, i.e., datasets 5–7 form one grouping (orange), datasets 8-10 form a second grouping (green), and datasets 1-4 form a third grouping (red). We also see that the groupings of datasets 5-7 and 8-10 are closer to each other than to that of datasets 1-4. The labels that result from the clustering are the same within each cluster but with arbitrary ordering. By denoting the true label for the k^{th} dataset with c_k , we choose $c_1 = c_2 = c_3 = c_4 = 1$, $c_5 = c_6 = c_7 = 2$, and $c_8 = c_9 = c_{10} = 3$.

We are not aware of a performance metric that captures how well the relationship structure of the datasets is estimated; instead, we use the Adjusted Mutual Information (AMI) [29] between the true and estimated clusters, which evaluates if the groupings are identified correctly. The AMI



FIGURE 5. The ground truth dendrogram for our simulations shows the relationship structure of the 10 datasets. There exist 3 groupings: consisting of datasets 5–7 (orange), 8–10 (green), and 1–4 (red).

is a normalized metric based on the mutual information of the true and estimated clustering, i.e., the AMI is equal to 1 if the true and the estimated clusters are equal, and equal to 0 if the mutual information between the true and estimated clusters equals the expected value of the mutual information between the true and a random clustering. Furthermore, the AMI corrects for the permutation ambiguity between the true and the estimated clusters. Our method has an AMI of 1 for $\rho = 0.2, 0.5, 0.8, i.e.$, it correctly identifies the groupings in 100% of the runs for all the correlation values.

V. REAL FMRI MULTI-TASK DATA

Finally, we apply our method on real multi-task fMRI data. Data from multiple tasks provide complementary information about the brain [30], [31] because different tasks involve cognitive functions that are either task-specific or common across all tasks [32]. By analyzing multiple datasets jointly, the function of and relationship between brain networks can be identified [32], which helps to understand the brain organization [33]. Furthermore, jointly analyzing data from multiple cognitive tasks may also help to understand complex disorders like schizophrenia, which is a neuropsychiatric disorder associated with cognitive deficits [30], [34] and altered connections between brain regions [35]. These altered connections might not be captured when analyzing only the data from a single task [30]. By analyzing data from multiple tasks jointly, latent neural patterns, i.e., biomarkers, may be revealed, which help yield new features like a common network for the tasks that captures differences between patients with schizophrenia and controls [11], [35].

A. DATASETS AND PREPROCESSING

We apply our proposed method on 10 fMRI datasets from the MIND Clinical Imaging Consortium (MCIC) collection [36], which are collected from 271 subjects (121 patients with schizophrenia and 150 healthy controls) that perform three different tasks: Auditory Oddball (AOD), Sensory Motor (SM), and Sternberg Item Recognition Paradigm (SIRP). In the AOD task, three stimuli are played: a frequent standard stimulus (1 kHz tone), an infrequent target stimulus (1.2 kHz tone), and an infrequent novel stimulus (computer-generated,

²Python code available at https://github.com/SSTGroup/independent_ vector_analysis [14]

complex sound). After hearing the target stimulus, the subject must press a button with the right index finger. During the SM task, tones are played in increasing order until the highest pitch is reached, then in decreasing order. The subject must press a button with the right thumb every time a new tone occurs. The SIRP task consists of two phases, encoding (SIRP-E) and probe (SIRP-P). In the encoding phase, a set of digits is presented on a screen, and the subject needs to learn this set. In the following probe (SIRP-P), digits are presented subsequently in a pseudo-random order. The subject must press a button with the right thumb if the digit was in the set and with the left thumb if not.

We expect task-specific activations of the auditory brain regions for the auditory tasks (AOD and SM) and of the visual brain regions for the visual tasks (SIRP), along with activations of the Default Mode Network (DMN) for all tasks. This prior knowledge about the relationship of the task datasets provides us with the opportunity to directly assess the success of the proposed method with real data, which is typically not easily achievable, as for real data one often does not know the true relationship structure. Furthermore, as the data is collected from patients and controls, we can also use these datasets to evaluate which tasks show a clearer discrimination of patients with schizophrenia.

For each of the subjects, multiple three-dimensional brain scans are collected over time during each task. The recorded scans are then preprocessed as in [35]: Using the Statistical Parametric Mapping (SPM) MATLAB toolbox [37], a simple voxelwise linear regression is applied to the data to eliminate the temporal dimension. The regressors are created by convolving the hemodynamic response function in SPM with the desired predictors for each task, which will be described in the next paragraph. For each subject and task, the resulting regression coefficient maps, also called "contrast images", are flattened and used as one-dimensional features that capture the variations across subjects. The flattened feature vectors of length J = 48,546 voxels are concatenated for I = 271 subjects to create the k^{th} task dataset $\mathbf{X}^{[k]} \in \mathbb{R}^{I \times J}$, k = 1, ..., K.

In the AOD task, the occurrences of the novel stimuli (AOD-N), the target stimuli (AOD-T), and the target with standard stimuli (AOD-TS) are each modeled as delta functions and used as predictors. Thus, there are three task datasets for the AOD task. In the SM task, the whole block is used as the predictor; thus, there is one SM task dataset. For both phases of SIRP (E and P), also the whole block is used as the predictor. This way, the data allows us to analyze the learning and retrieving phases of this task separately. As the SIRP task is repeated with 1, 3, and 5 digits in the set, there are six datasets for the SIRP task. Thus, in total, we have K = 10 datasets.

B. IMPLEMENTATION DETAILS

As typically the dimension of the datasets is much higher than the number of latent sources, a dimension reduction via Principal Component Analysis (PCA) is performed as preprocessing. This way, each observed dataset is transformed separately in a lower-dimensional subspace of dimension *R*. Then, IVA is performed on the dimension-reduced datasets $\mathbf{X}^{[k]} \in \mathbb{R}^{R \times V}$ to estimate the source matrices $\widehat{\mathbf{S}}^{[k]} \in \mathbb{R}^{R \times V}$. From these, the SCVs $\widehat{\mathbf{S}}_r$, r = 1, ..., R are formed.

As it is mostly the case for real data, also in the MCIC data, we do not know the ground truth dimension R of the latent sources. Thus, selecting an appropriate value for R is important to get meaningful results. Our method gives robust clustering results for a wide range of values for $R = 20, 21, \ldots, 30$. In this paper, we present the results for R = 25 because at this order, the estimated fMRI activation maps are 1) stable, i.e., not split (as in higher orders) or merged (as in lower orders), and 2) meaningful, i.e., physically interpretable.

Our approach is primarily data-driven, with the exception of two user-selected parameters, the probability of false alarm P_{fa} and the number of bootstrap resamples *B*. With P_{fa} , we can directly control the risk of overestimating d_r , i.e., the number of eigenvalues greater than 1 in the r^{th} SCV covariance matrix. The higher the value for *B* is chosen, the better the distribution is estimated, but for a too high value of *B*, there will not be a better estimate at some point. We choose B = 1000 bootstrap resamples and $P_{fa} = 0.05$ to estimate \hat{d}_r , as these are typical values for these parameters and achieve good results in general.

As we are also interested in identifying the components that discriminate between patients and controls, we apply a two-sample *t*-test on the first 150 (controls) and the following 121 (patients) values of each column of the estimated mixing matrix $\widehat{\mathbf{A}}^{[k]}$. We consider *p*-values smaller than 0.05 to be significant, indicating that the corresponding activated brain areas are different in patients with schizophrenia and healthy controls. Using this *t*-test, we corrected the signs of the estimated sources $\widehat{\mathbf{S}}^{[k]}$ to overcome the sign ambiguity of IVA: We made sure that the *t*-values of the datasets that show a significant difference between patients and controls (p < 0.05) are positive or made positive by multiplying the estimated sources and corresponding subject profiles by -1 (if the *t*-value is negative). This way, positive values of the (zero-mean) sources indicate higher activations in controls, and negative values indicate higher activation in patients.

C. RESULTS WITH THE fMRI DATA

1) IDENTIFICATION OF COMMON AND STRUCTURED SCVs In Figure 6, a subset of the covariance matrices of the SCVs estimated by IVA-L-SOS is shown along with the estimated values for \hat{d}_r . Light values correspond to high correlations. For SCVs 3 and 13, $\hat{d}_r = 1$, thus, these SCVs are identified as common. SCVs 15–17, 20, 22–23, and 25 are identified as structured because $\hat{d}_r > 1$. The visible block structure in each SCV already reveals information about the relationship of the datasets [14]. For example, the covariance matrix of SCV16 clearly shows high correlations within the AOD and SM tasks



FIGURE 6. Subset of the estimated SCV covariance matrices. SCVs 3 and 13 are identified as common, and SCVs 15–17, 20, 22–23, and 25 are identified as structured.



FIGURE 7. Estimated fMRI activation maps corresponding to SCV3. The default mode network and the visual regions are activated higher in controls (shown by red/yellow voxels), while the sensorimotor areas are activated higher in patients (shown by blue voxels).



FIGURE 8. Estimated fMRI activation maps corresponding to SCV16. The auditory regions are activated higher in controls in the AOD and SM datasets and not activated in the SIRP datasets.



FIGURE 9. Estimated fMRI activation maps corresponding to SCV17. The visual regions are activated higher in controls in the SIRP-E datasets, higher in patients in the SIRP-P datasets, and not activated in the AOD and SM datasets.

and within the SIRP tasks, but small correlations across those tasks.

2) IVA-L-SOS

The estimated source components (fMRI activation maps) provide information about which brain regions are activated in which tasks. In the following, we present SCV3, SCV16, and SCV17 as examples for common and structured SCVs, because they either show activations in brain areas that are common in all tasks or correspond to the specific tasks. All fMRI activation maps are thresholded with |z| = 2 before visualization and plotted above the structural anatomical

images. Significant p-values are displayed in magenta with the superscript *. Because of the sign correction, red or yellow voxels indicate that a brain area is activated higher in controls, and blue voxels indicate higher activation in patients. Then, we corrected the sign of the non-significant datasets manually by matching the color (red/blue) of the activated areas to the significant datasets of the same task.

In SCV 3 in Figure 7, we see activations of the DMN (red areas in slices 4–6). The DMN is known to have a decreased activation when a task is performed [32], [38]. The higher activation of the DMN in controls means that the

deactivation is stronger in patients. This can be interpreted as patients needing to focus more on a task to perform it well. The sensorimotor regions (blue areas in slices 2-3) are activated higher in patients and therefore support this interpretation. There are minor activations in the visual regions (red areas in slices 7-9), which are expected because the subjects had their eyes open during all tasks. The very small *p*-values in the AOD-T and AOD-TS datasets indicate that especially when the target stimulus occurs, the patients are significantly stronger engaged with the task, i.e., have significantly smaller activation of the DMN. These lower p-values in the AOD datasets are expected since the AOD task has been shown to be important in discriminating patients with schizophrenia and healthy controls, as patients have a smaller oddball response [39]. In the SIRP-P datasets, there is no significant difference between patients and controls for the DMN. In contrast, in the SIRP-E datasets, the p-values become smaller with increasing task difficulty, i.e., the deactivations of the DMN become stronger for the patients. This coincides with the literature, as with an increasing level of difficulty of a task, the deactivation of the DMN becomes stronger [32].

The fMRI activation maps corresponding to SCV16 are shown in Figure 8. The auditory regions (red areas in slices 5–7) are activated higher in the controls in the AOD and SM datasets. The *p*-values are also significant for the AOD and SM datasets, supporting the literature that activations in the auditory regions may be a biomarker for differentiating patients with schizophrenia and healthy controls [40].

In contrast, the fMRI activation maps corresponding to SCV17, shown in Figure 9, show strong activations in the visual areas (red/blue areas in slice 7) for the SIRP datasets. The significant difference between patients and controls here is found for SIRP-E5 (in accordance with SCV3) and SIRP-P1. What is most surprising here is that visual regions are activated higher in controls for the SIRP encoding but higher in patients for the SIRP probe. An explanation might be that patients need to focus more on a digit to remember it, while the controls just briefly see the digit, and their memory can be accessed faster.

3) IDENTIFICATION OF THE RELATIONSHIP STRUCTURE

The information of all SCVs is fused to identify the relationship structure among the datasets. As described in section III-C, the \hat{d}_r leading eigenvectors of the structured SCVs are concatenated to form a feature matrix, which is the input of the clustering algorithm. The dendrogram for the hierarchical clustering of the feature matrix is shown in Figure 10. Here, a grouping refers to a group of task datasets (in contrast to a group of subjects, as in subgroup identification). The SIRP tasks form one grouping (orange), and the AOD and SM tasks form a second grouping (green). As the SIRP task involves a visual stimulus and the AOD and SM tasks both involve auditory stimuli, these resulting groupings are meaningful. Within the SIRP grouping, there



FIGURE 10. The dendrogram for the real data reveals the relationship structure among the datasets. Two groupings are found, consisting of the SIRP datasets (orange) and of the auditory datasets, AOD and SM (green). Within the SIRP datasets, the SIRP-E and SIRP-P datasets form two groupings. Within the auditory datasets, there is one grouping consisting of the AOD datasets.

are two finer groupings visible: one consisting of the encoding datasets and one the other of the probe datasets. This makes sense because they refer to two different phases of the SIRP task. Within the auditory grouping, the AOD datasets form another grouping.

VI. DISCUSSION

In this paper, we have proposed a method to identify the relationship structure among multiple datasets. Our method consists of three steps: 1) estimating latent sources from observed datasets using IVA, 2) identifying structured SCVs (SCVs whose covariance matrices have more than one eigenvalue greater than 1), and 3) extracting features from the structured SCVs, which are then used by hierarchical clustering to identify the relationship structure among the datasets. Compared with previous studies, our proposed method alleviates the need to assume Gaussianity in the data by 1) including higher-order statistics through the use of IVA-L-SOS for source estimation, which leads to more interpretable components, and 2) not relying on any distributional assumptions, which is achieved by estimating the number of eigenvalues greater than 1 using their theoretical properties and a bootstrap-based hypothesis testing approach.

Our simulations demonstrate the success of our method in terms of 1) correctly estimating the number of blocks per SCV against competing methods and 2) identifying the relationship structure among multiple datasets. Applying our method to real multi-task fMRI data has revealed activated brain areas that are known to be affected by schizophrenia: We see significantly stronger deactivations of the DMN in patients and significantly stronger activations of the auditory brain regions in controls. The identified relationship structure of the task datasets is consistent with, but extends, existing work. While we were able to draw conclusions from the here presented estimated fMRI maps, it is important to note that interpretation is not always straightforward.

We have carefully selected and presented SCVs that show activations in meaningful brain regions, which we identified according to Brodmann areas [41]. Despite the simplicity of the task design, many of the unrepresented components defy straightforward explanation. However, interpreting a subset of the most meaningful components can already help in understanding how the brain functions. Thus, a good guideline is to find a range of R values that lead to stable results after the dimension reduction and then go through the estimated components and compare them with established brain regions such as the Brodmann areas [41] to facilitate interpretation. It is important to remember that relying solely on comparisons with known brain atlas components might cause us to overlook brain regions not included in the userdefined atlas. Nevertheless, comparing the brain regions with an atlas provides an initial reference point before conducting further investigations.

Importantly, the proposed method is not limited to task datasets but is applicable to more general problems, e.g., for identifying subgroups of subjects in neuroimaging, and other fields. After having provided the confirmation of the success of our method on these clinically well-understood datasets, the direct implications for clinical significance and treatment strategies could be further explored in the future by applying our method to diverse datasets, with a specific focus on precision medicine and other relevant applications. For example, by identifying the relationship structure among patient datasets, our method allows for a detailed analysis of associations among data of, e.g., patients with schizophrenia, bipolar disorder, and their subtypes. However, a limitation of our proposed method is that it does not identify SCVs with identity covariance matrices, i.e., SCVs that consist of completely uncorrelated sources. A possible way to overcome this limitation in the future may be to adapt the bootstrap test.

DATA AVAILABILITY

The IRB has determined the data cannot be shared outside the study team, and any ongoing use must be approved by the fBIRN PI.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

ACKNOWLEDGMENT

The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF). The facility is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS-0821258, CNS-1228778, OAC-1726023, and CNS-1920079) and the SCREMS program (grant no. DMS-0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See hpcf.umbc.edu for more information on HPCF and the projects using its resources.

REFERENCES

- N. M. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE Signal Process. Mag.*, vol. 27, no. 4, pp. 39–50, 2010.
- [2] V. D. Calhoun and T. Adali, "Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Rev. Biomed. Eng.*, vol. 5, pp. 60–73, 2012.
- [3] I. Lipkovich, A. Dmitrienko, and R. B. D'Agostino, "Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials," *Statist. Med.*, vol. 36, no. 1, pp. 136–196, Jan. 2017.
- [4] Y. Wang, Y. Zhao, T. M. Therneau, E. J. Atkinson, A. P. Tafti, N. Zhang, S. Amin, A. H. Limper, S. Khosla, and H. Liu, "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *J. Biomed. Informat.*, vol. 102, Feb. 2020, Art. no. 103364.
- [5] W. Loh, L. Cao, and P. Zhou, "Subgroup identification for precision medicine: A comparative review of 13 methods," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 5, p. e1326, Sep. 2019.
- [6] National Research Council (U.S.) Committee on a Framework for Developing a New Taxonomy of Disease, *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease.* New York, NY, USA: Academic, 2011.
- [7] J. C. Foster, J. M. G. Taylor, and S. J. Ruberg, "Subgroup identification from randomized clinical trial data," *Statist. Med.*, vol. 30, no. 24, pp. 2867–2880, Oct. 2011.
- [8] H. Seibold, A. Zeileis, and T. Hothorn, "Model-based recursive partitioning for subgroup analyses," *Int. J. Biostatistics*, vol. 12, no. 1, pp. 45–63, May 2016.
- [9] 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, May 2014.
- [10] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar, "A method for making group inferences from functional MRI data using independent component analysis," *Human Brain Mapping*, vol. 14, no. 3, pp. 140–151, Nov. 2001.
- [11] V. D. Calhoun, T. Adalı, K. A. Kiehl, R. Astur, J. J. Pekar, and G. D. Pearlson, "A method for multitask fMRI data fusion applied to schizophrenia," *Human Brain Mapping*, vol. 27, no. 7, pp. 598–610, Jul. 2006.
- [12] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. Int. Conf. Independ. Compon. Anal. Signal Separat.*, 2006, pp. 165–172.
- [13] Q. Long, S. Bhinge, V. D. Calhoun, and T. Adali, "Independent vector analysis for common subspace analysis: Application to multi-subject fMRI data yields meaningful subgroups of schizophrenia," *NeuroImage*, vol. 216, Aug. 2020, Art. no. 116872.
- [14] I. Lehmann, E. Acar, T. Hasija, M. A. B. S. Akhonda, V. D. Calhoun, P. J. Schreier, and T. Adali, "Multi-task fMRI data fusion using IVA and PARAFAC2," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* (ICASSP), May 2022, pp. 1466–1470.
- [15] S. Bhinge, R. Mowakeaa, V. D. Calhoun, and T. Adali, "Extraction of timevarying spatiotemporal networks using parameter-tuned constrained IVA," *IEEE Trans. Med. Imag.*, vol. 38, no. 7, pp. 1715–1725, Jul. 2019.
- [16] H. Yang, M. A. B. S. Akhonda, F. Ghayem, Q. Long, V. D. Calhoun, and T. Adali, "Independent vector analysis based subgroup identification from multisubject fMRI data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1471–1475.
- [17] P. Bunouf, M. Groc, A. Dmitrienko, and I. Lipkovich, "Data-driven subgroup identification in confirmatory clinical trials," *Therapeutic Innov. Regulatory Sci.*, vol. 56, no. 1, pp. 65–75, Jan. 2022.
- [18] U. Niemann, M. Spiliopoulou, B. Preim, T. Ittermann, and H. Völzke, "Combining subgroup discovery and clustering to identify diverse subpopulations in cohort study data," in *Proc. IEEE 30th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2017, pp. 582–587.
- [19] X. Chen, Z. J. Wang, and M. McKeown, "Joint blind source separation for neurophysiological data analysis: Multiset and multimodal methods," *IEEE Signal Process. Mag.*, vol. 33, no. 3, pp. 86–107, May 2016.
- [20] T. Hasija, T. Marrinan, C. Lameiro, and P. J. Schreier, "Determining the dimension and structure of the subspace correlated across multiple data sets," *Signal Process.*, vol. 176, Nov. 2020, Art. no. 107613.
- [21] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 387–392, Apr. 1985.

- [22] S. M. Kay, Fundamentals of Statistical Signal Processing: Detection Theory. Upper Saddle River, NJ, USA: Prentice-Hall, 1998.
- [23] A. M. Zoubir and D. R. Iskander, Bootstrap Techniques for Signal Processing. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [24] R. Davidson and J. G. MacKinnon, "The power of bootstrap and asymptotic tests," *J. Econometrics*, vol. 133, no. 2, pp. 421–441, Aug. 2006.
- [25] K. Sasirekha and P. Baby, "Agglomerative hierarchical clustering algorithm—A review," *Int. J. Sci. Res. Publications*, vol. 3, no. 3, p. 83, 2013.
- [26] S. Kotz, T. Kozubowski, and K. Podgórski, The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Economics, Engineering, and Finance. Springer, 2001.
- [27] Q. Long, C. Jia, Z. Boukouvalas, B. Gabrielson, D. Emge, and T. Adali, "Consistent run selection for independent component analysis: Application to fMRI analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2581–2585.
- [28] F. Pedregosa, G. Varoquaux, and A. Gramfort, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, no. 85, pp. 2825–2830, 2011.
- [29] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" in *Proc.* 26th Annu. Int. Conf. Mach. Learn., Jun. 2009, pp. 1073–1080.
- [30] M. Ramezani, K. Marble, H. Trang, I. S. Johnsrude, and P. Abolmaesumi, "Joint sparse representation of brain activity patterns in multi-task fMRI data," *IEEE Trans. Med. Imag.*, vol. 34, no. 1, pp. 2–12, Jan. 2015.
- [31] B. Mijović, K. Vanderperren, and N. Novitskiy, "The 'why' and 'how' of JointICA: Results from a visual detection task," *NeuroImage*, vol. 60, no. 2, pp. 1171–1185, 2012.
- [32] N. Sanford, J. C. Whitman, and T. S. Woodward, "Task-merging for finer separation of functional brain networks in working memory," *Cortex*, vol. 125, pp. 246–271, Apr. 2020.
- [33] Connectome Coordination Facility. (2022). Components of the Human Connectome Project: Task FMRI. [Online]. Available: https://www. humanconnectome.org/study/hcp-young-adult/project-protocol/task-fmri
- [34] T. E. Mwansisya, A. Hu, Y. Li, X. Chen, G. Wu, X. Huang, D. Lv, Z. Li, C. Liu, Z. Xue, J. Feng, and Z. Liu, "Task and resting-state fMRI studies in first-episode schizophrenia: A systematic review," *Schizophrenia Res.*, vol. 189, pp. 9–18, Nov. 2017.
- [35] A. M. Michael, S. A. Baum, J. F. Fries, B. Ho, R. K. Pierson, N. C. Andreasen, and V. D. Calhoun, "A method to fuse fMRI tasks through spatial correlations: Applied to schizophrenia," *Human Brain Mapping*, vol. 30, no. 8, pp. 2512–2529, Aug. 2009.
- [36] R. L. Gollub et al., "The MCIC collection: A shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia," *Neuroinformatics*, vol. 11, no. 3, pp. 367–388, Jul. 2013.
- [37] Statistical Parametric Mapping Toolbox: SPM12, Members Collaborators Wellcome Centre for Human Neuroimaging, London, U.K., 2020.
- [38] M.-L. Hu, X.-F. Zong, J. J. Mann, J.-J. Zheng, Y.-H. Liao, Z.-C. Li, Y. He, X.-G. Chen, and J.-S. Tang, "A review of the functional and anatomical default mode network in schizophrenia," *Neurosci. Bull.*, vol. 33, no. 1, pp. 73–84, Feb. 2017.
- [39] D. I. Kim, D. H. Mathalon, J. M. Ford, M. Mannell, J. A. Turner, G. G. Brown, A. Belger, R. Gollub, J. Lauriello, C. Wible, D. O'Leary, K. Lim, A. Toga, S. G. Potkin, F. Birn, and V. D. Calhoun, "Auditory oddball deficits in schizophrenia: An independent component analysis of the fMRI multisite function BIRN study," *Schizophrenia Bull.*, vol. 35, no. 1, pp. 67–81, Jan. 2009.
- [40] V. D. Calhoun, K. A. Kiehl, P. F. Liddle, and G. D. Pearlson, "Aberrant localization of synchronous hemodynamic activity in auditory cortex reliably characterizes schizophrenia," *Biol. Psychiatry*, vol. 55, no. 8, pp. 842–849, Apr. 2004.
- [41] C. Lacadie, R. Fulbright, J. Arora, R. Constable, and X. Papademetris, "Brodmann areas defined in MNI space using a new tracing tool in bioimage suite," in *Proc. 14th Annu. Meeting Org. Human Brain Mapping*, vol. 771, 2008. [Online]. Available: https://scholar.google.com/ citations?view_op=view_citation&hl=de&user=4YP7AQYAAAAJ&cstar t=100&pagesize=100&sortby=pubdate&citation_for_view=4YP7AQYA AAAJ:anf4URPfarAC



ISABELL LEHMANN (Graduate Student Member, IEEE) was born in Germany, in 1994. She received the B.Sc. and M.Sc. degrees in electrical engineering from Paderborn University, Paderborn, Germany, in 2017 and 2019, respectively, where she is currently pursuing the Ph.D. degree in electrical engineering.

Her research interests include signal processing techniques, especially matrix and tensor factorizations, for medical data.



TANUJ HASIJA (Member, IEEE) received the Ph.D. degree (summa cum laude) in electrical engineering from Paderborn University, Germany, in 2021.

He is currently a tenured Postdoctoral Faculty Member with the Signal and System Theory Group, Institute of Electrical Engineering and Information Technology, Paderborn University, and a Research Fellow with the Department of Neurology, Boston Children's Hospital affiliated

with Harvard Medical School. His research interests include multimodal machine learning and statistical signal processing with applications in epilepsy, brain imaging, and neuroscience.

Dr. Hasija is currently serving as an elected member of the Technical Area Committee (TMTSP) of the European Association for Signal Processing and an Associate Member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society.



BEN GABRIELSON received the B.A. degree in physics from Franklin and Marshall College, Lancaster, PA, USA, in 2013, and the M.S. degree in electrical engineering from the University of Maryland, Baltimore County (UMBC), Baltimore, MD, USA, in 2020, where he is currently pursuing the Ph.D. degree in electrical engineering, under the supervision of Dr. Tülay Adali.

His research interests include matrix and tensor factorizations, with a particular emphasis on blind source separation and joint blind source separation.



MOHAMMAD A. B. S. AKHONDA received the B.S. degree in electronics and communication engineering from Khulna University of Engineering and Technology, Bangladesh, in 2013, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, Baltimore County, Baltimore, MD, USA, in 2019 and 2022, respectively, under the supervision of Dr. Tülay Adali.

Before starting his graduate studies, he was a Software Engineer with the Samsung Research and Development Institute. His research interests include joint blind source separation, multimodal and multiset data fusion, common and distinct subspace analysis, and joint model order selection problems.



VINCE D. CALHOUN (Fellow, IEEE) received the bachelor's degree in electrical engineering from The University of Kansas, Lawrence, KS, USA, in 1991, the dual master's degrees in biomedical engineering and information systems from Johns Hopkins University, Baltimore, MD, USA, in 1993 and 1996, respectively, and the Ph.D. degree in electrical engineering from the University of Maryland, Baltimore County, Baltimore, MD, USA, in 2002.

He is currently the founding Director of the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS) where he holds appointments at Georgia State, Georgia Tech, and Emory University. He is the author of more than 1000 full journal articles. His work includes the development of flexible methods to analyze neuroimaging data, including blind source separation, deep learning, multimodal fusion and genomics, and neuroinformatics tools.

Dr. Calhoun is a fellow of the Institute of Electrical and Electronic Engineers, the American Association for the Advancement of Science, the American Institute of Biomedical and Medical Engineers, the American College of Neuropsychopharmacology, the Organization for Human Brain Mapping (OHBM), and the International Society of Magnetic Resonance in Medicine. He currently serves on the IEEE BISP Technical Committee. He is also a member of the IEEE Data Science Initiative Steering Committee and the IEEE Brain Technical Committee.



TÜLAY ADALI (Fellow, IEEE) received the Ph.D. degree in electrical engineering from North Carolina State University, Raleigh, NC, USA, in 1992.

In 1992, she joined as a Faculty Member with the University of Maryland Baltimore County (UMBC), Baltimore, MD, USA, where she is currently a Distinguished University Professor.

Prof. Adali is a fellow of the AIMBE and AAIA, a Fulbright Scholar, and an IEEE SPS

Distinguished Lecturer. She was a recipient of the SPS Meritorious Service Award, the Humboldt Research Award, the IEEE SPS Best Paper Award, the SPIE Unsupervised Learning and ICA Pioneer Award, the University System of Maryland Regents' Award for Research, and the NSF CAREER Award. She served as the Chair of the Brain Technical Community and the Signal Processing Society (SPS) Vice President for Technical Directions 2019–2022. Over the years, she has served the IEEE and the IEEE Signal Processing Society (SPS) in numerous capacities. She is currently the Editorin-Chief of IEEE Signal Processing Magazine. She has been part of the organizing committees of many conferences and workshops, including the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), the Technical Chair (2017 and 2026), the Special Sessions Chair (2018 and 2024), the Publicity Chair (2000 and 2005), and the Publications (2008). She was the General/Technical Chair of the IEEE Machine Learning for Signal Processing (MLSP) and Neural Networks for Signal Processing (NNSP) Workshops 2001-2009, 2014, and 2023. She served multiple terms in three technical committees of the SPS (NNSP/MLSP, Bio Imaging and Signal Processing, and Signal Processing Theory and Methods) and chaired the NNSP/MLSP Technical Committee, from 2003 to 2005 and from 2011 to 2013. She served or is currently serving on the editorial board of multiple journals, including IEEE TRANSACTIONS ON SIGNAL PROCESSING, PROCEEDINGS OF THE IEEE, and IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING.