COMPLETE MODEL IDENTIFICATION USING INDEPENDENT VECTOR ANALYSIS: APPLICATION TO THE FUSION OF TASK FMRI DATA

M. A. B. S. Akhonda[1]*, Ben Gabrielson*[1]*, Vince D. Calhoun*[2]*, and Tülay Adali*[1]

[1]Dept. of CSEE, University of Maryland, Baltimore County, Baltimore, MD 21250
[2]Tri-institutional Center for Translational Research in Neuroimaging and Data Science, Atlanta, GA 30303

ABSTRACT

Linear latent variable models have proven effective for data fusion using joint decomposition of multiple matrices. Identification of the dependence structure of the latent variables—*components*—across multiple datasets is key to this success as this helps explain the underlying relationship across the datasets, and allows the design of a joint decomposition model that best fits the properties of the problem. However, identification of the complete dependence structure across more than two datasets is a difficult problem due to large number of possible dependence scenarios. In this paper, we address this problem, i.e., the estimation of not only the number of components that are dependent across $N \geq 2$ datasets but also their *complete* dependence structure, i.e., the index of datasets across which they are dependent. The method, complete model identification using IVA (CMI-IVA), builds on the well-structured formulation of independent vector analysis (IVA), which generalizes multiset canonical correlation analysis, and provides a key step in facilitating this difficult problem. Properties of CMI-IVA are established and its performance is first verified using simulations. We then apply the method to real functional magnetic resonance (fMRI) data and demonstrate that CMI-IVA provides meaningful interpretation of the data in terms of number of components dependent across datasets and the associated components.

***Index Terms*—** Model identification, Data fusion, FMRI, Common and distinct components

1. INTRODUCTION

Methods based on latent variable analysis have proven useful for joint data analysis and data fusion across various disciplines such as medical imaging, remote sensing, metabolomics, and chemometrics, among others [1–4]. They allow explanation of the unique as well as the common or dependent factors across multiple datasets through latent variables, i.e., components. Hence, it is of particular interest to leverage these variables' dependence structure to perform exploratory analysis, resulting in a more unified picture and global view of the system of interest [5, 6]. Numerous studies, such as those in medical imaging, have focused on this very aspect, either for fusion of different brain imaging modalities such as functional magnetic resonance imaging (fMRI), electroencephalograph (EEG), and structural MRI (sMRI) [7, 8], diffusion tensor imaging (DTI), and magnetoencephalography (MEG) [9] or when using multiple datasets from the same imaging modality data such as fMRI data collected for different experimental setups, conditions, tasks or subjects [10, 11].

This work was supported in part by NSF-CCF 1618551, NSF-NCS 1631838, and NIH R01 MH118695.

Identification of the number of dependent components—*or the model order*—and their dependence structure, i.e., datasets across which they are dependent, across multiple datasets can be posed as a *model identification* problem. While methods for identifying the number of components within a single dataset are well developed, see, e.g., [12–14], methods for identifying those across multiple datasets are quite limited. The latter is a more challenging problem, and the complexity of the problem increases significantly due to the increase in possible dependence scenarios for more than two datasets. The underlying components might be dependent across all datasets, subsets of datasets, or none of the datasets. Moreover, the components that are dependent across datasets might not be the ones with the highest variance. Hence, a traditional principal component analysis (PCA) based approach applied to individual datasets might eliminate these components at a first step thus preventing their discovery in subsequent steps. Current methods to solve the model identification problem are either limited to two datasets [15, 16] or focused on identifying the number of dependent components across all datasets [17]. One important method, principal component analysis prior to canonical correlation analysis (PCA-CCA), exploits CCA's strength and is suitable for sample rich and sample poor scenarios, but it is limited to only two datasets [16]. Another method, multiset-CCA followed by knee point detection (MCCA-KPD), concentrates on discovering the model order only for the components dependent across all datasets, disregarding the dependence present across subsets of datasets [17]. On top of that, determining only the model order without the knowledge of dependence structure, i.e., identifying datasets across which components are dependent, is often not sufficient to describe the complete underlying relationships among the datasets. A recently proposed method, complete model selection (CMS) based on eigen-analysis of a joint coherence matrix [18], provides the required flexibility, but it is computationally costly and requires large memory allocation for most applications, particularly for those in medical imaging, which we discuss in this paper. One way to alleviate this issue is to first transform the problem into a more convenient space that enables working with a smaller dimensionality and where underlying assumptions for the analysis can be more readily satisfied. This motivates us to propose a new method, which we introduce next.

We introduce complete model identification using independent vector analysis (CMI-IVA), to estimate not only 1) the model order, which includes identifying the number of components dependent across all subset of datasets, but also 2) the complete dependence structure of the components available in the datasets. CMI-IVA exploits the strength of IVA framework to transform multiple datasets into a space where we have an effective decoupling across the subspaces such that by simply counting the number of eigenvalues greater than 1, we can identify the order of the dependent components. The identity of the datasets that result in highly dependent

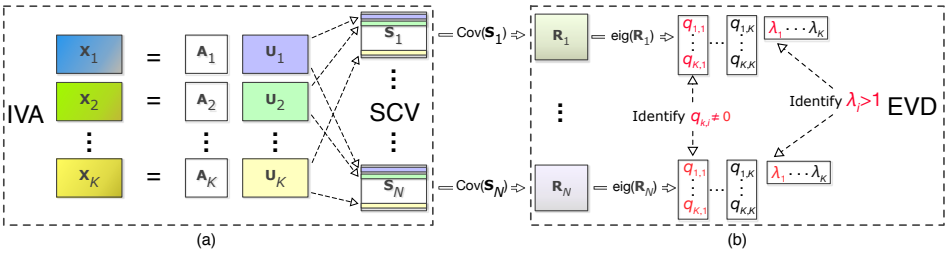

Fig. 1: Generative model of CMI-IVA. (a) IVA step estimates the SCVs and (b) EVD step identifies the covariance matrices of the SCVs with eigenvalues greater than one to estimate the model order and dependece structure.

components can be found by evaluating the corresponding eigenvectors. We validate the performance of the method with simulated data as well as real fMRI data collected from healthy subjects and patients with schizophrenia doing an auditory oddball task (AOD). First using simulations, we show that CMI-IVA outperforms other methods with respect to the estimation results. Then with real fMRI data, we demonstrate that the number of dependent brain maps and their dependence structure identified by CMI-IVA provides meaningful results and better interpretation of the task data given the knowledge of the datasets used for the analysis.

2. METHODOLOGY

2.1. Problem formulation using IVA

Independent component analysis (ICA) is a data-driven blind source separation (BSS) technique that decomposes a dataset into a set of components based on the assumption of independence. ICA has proven powerful in recovering interpretable, i.e., physically meaningful, features in many studies, see e.g., [5]. However, ICA is limited to analyze a single dataset at a time. IVA generalizes ICA to multiple datasets by additionally taking the dependence of the datasets into account.

Consider K datasets, each containing T samples, formed from a linear mixture of N independent components as

$$\mathbf{x}_k(t) = \mathbf{A}_k \mathbf{s}_k(t), k = 1, 2, \ldots, K, t = 1, 2, \ldots, T, \quad (1)$$

where $\mathbf{A}_k \in \mathbb{R}^{N \times N}, k = 1, 2, \ldots, K$ are invertible mixing matrices. Given this model, IVA solution finds K demixing matrices $\mathbf{W}_k$, $k = 1, 2, \ldots, K$ such that source components from each dataset can be estimated through $\mathbf{u}_k(t) = \mathbf{W}_k \mathbf{x}_k(t)$. For a given set of observations $\mathbf{X}_k$, the above equation can be written as, $\mathbf{U}_k = \mathbf{W}_k \mathbf{X}_k$, where $\mathbf{X}_k, \mathbf{U}_k \in \mathbb{R}^{(N \times T)}$ and $\mathbf{U}_k = [\mathbf{u}_k^{[1]}, \mathbf{u}_k^{[2]}, \ldots, \mathbf{u}_k^{[N]}]^\mathsf{T}$. The estimated components are independent within a dataset while maximally dependent on corresponding components across the datasets. This way, IVA takes the dependence among the corresponding sources across multiple datasets into account, to obtain decompositions that fully leverage the commonalities across the datasets. This is done by modeling the source component vector (SCV), where nth SCV can be defined as

$$\mathbf{s}_n(t) = [s_1^{[n]}(t), s_2^{[n]}(t), \ldots, s_K^{[n]}(t)]^\mathsf{T} \in \mathbb{R}^K, \quad n = 1, 2, \ldots, N, \quad (2)$$

i.e., by concatenating the nth source components from each of the K dataset, where $\mathbf{s}_k^{[n]} \in \mathbb{R}^T$ is the nth source from the kth dataset. Since SCVs are defined using corresponding components across all K datasets, their covariance matrices preserve the dependence structure of the components across the data sets and can be used to solve the model identification problem. For simplicity, we do not take sample dependency into account in the rest of the article and consider simple independent and identically distributed samples, thus dropping index t in (1). We assume all components are zero mean and unit variance so that the covariance and the correlation matrix coincide and are written for the nth SCV as,

$$\mathbf{R}_n = \begin{bmatrix} 1 & \rho_{1,2}^{[n]} & \ldots & \rho_{1,K}^{[n]} \\ \rho_{2,1}^{[n]} & 1 & \ldots & \rho_{2,K}^{[n]} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{K,1}^{[n]} & \rho_{K,2}^{[n]} & \ldots & 1 \end{bmatrix} \in \mathbb{R}^{K \times K}, \quad (3)$$

where, $\rho_{k_1,k_2}^{[n]}$ represents unknown correlation coefficient between the nth component across datasets k_1 and k_2. While $\rho_{k_1,k_2}^{[n]} = 0$ represents no dependence, $\rho_{k_1,k_2}^{[n]} \neq 0$ indicates dependence among the nth components across datasets k_1 and k_2. Thus, one can simply identify the SCVs with non-zero $\rho_{k_1,k_2}^{[n]}$ to determine the dimensionality and the structure of the dependent components. Depending on the density function used to model the SCVs, IVA can take both second and higher-order statistics (SOS and HOS) into account [19]. When a multivariate Gaussian distribution is used to model the SCV, thus taking only the SOS into account, the method is called IVA with multivariate Gaussian distribution (IVA-G) [20]. We make use of IVA-G to solve the model identification problem by estimating

1. the model order d, which in our case represents the total number of SCV covariance matrices with non-zero $\rho_{k_1,k_2}^{[n]}$, and
2. the corresponding correlation structures, meaning the indices of the datasets k_1 and k_2 for which $\rho_{k_1,k_2}^{[n]} \neq 0$.

We use IVA-G since it generalizes MCCA without the constraint of an orthogonal demixing matrix [19, 20]. As such, IVA-G is readily applicable to any problem for which MCCA has proved useful. Moreover, strong identifiability condition of IVA-G—i.e., the ability to uniquely identify a wide range of signals under very general conditions—allows IVA-G to identify components as long as any subset of the components inside an SCV do not share a unique dependence structure [21]. This property allows IVA-G to preserve dependence structure more efficiently compared with other methods such as CCA and MCCA.

PCA-CCA, developed for $K = 2$, defines model order as $d_{k_1,k_2} = \{n : k_1, k_2 = 1, 2$ for which $\rho_{k_1,k_2}^{[n]} \neq 0\}$, while MCCA-KPD, developed for $K > 2$, defines model order, $d_{\text{all}} = \{n : \rho_{k_1,k_2}^{[n]} \neq 0 \forall k_1, k_2\}$, dependent across all datasets. Since model order only represents the number of components correlated across a fixed number of datasets, estimating it without the proper knowledge of correlation structure does not provide complete relationship across multiple datasets. The method proposed in [18], CMS, estimates both the model order and the correlation structure across

multiple datasets. It concatenates all the datasets into a large matrix to estimate an $NK \times NK$ block-diagonal coherence matrix, and uses the eigenvalue decomposition to solve the model identification problem. Given the fact that number of samples N can be quite large, as in our fMRI example, the memory allocation requirement is significant. As we discuss next, when the problem is first transformed into a decoupled domain through the use of IVA-G as the first step, we can then work with smaller $K \times K$ SCV covariance matrices by building on the development in [18] as we describe next. The individual SCV covariance matrices estimated by IVA-G compactly summarize the dependence structure across datasets while providing an effective decoupling across the SCVs through uncorrelatedness requirement built into the IVA-G estimation.

2.2. Complete model identification using IVA (CMI-IVA)

We propose a novel method, named CMI-IVA, to solve the complete model identification problem. CMI-IVA uses the robust IVA-G framework to transform multiple datasets jointly into a space where there is an effective decoupling among the subspaces and use that information in subsequent steps to identify and estimate the model order and the corresponding correlation structure. We perform CMI-IVA in two significant steps: an IVA-G step followed by an eigenvalue decomposition step. We start from the SCV definition of IVA-G and show that eigenvalues and eigenvectors of SCV covariance matrix $\mathbf{R}_n$ in (3) can be used to completely characterize the underlying relationship across the datasets. This result builds on results from Theorem 1 and 2 in [18] and the Perron-Frobenius theorem [22], and is established in our case—with highly correlated components—to identify the eigenvalues greater than one and their corresponding eigenvectors. We decompose $\mathbf{R}_n$ using eigenvalue decomposition as

$$\mathbf{R_n} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\mathsf{T}}, \tag{4}$$

where $\mathbf{Q}$ is the $K \times K$ matrix whose ith column $\mathrm{q}_i \in \mathbb{R}^K$ is the ith eigenvector of $\mathbf{R}_n$, and $\mathbf{D}$ is a diagonal matrix whose diagonal elements are the corresponding eigenvalues $\lambda_i, \quad i = 1, 2, \ldots K$. We start by showing that the largest eigenvalues of $\mathbf{R}_n$, where $n = 1, 2, \ldots, N$, can be used to identify the model order and then use the corresponding eigenvectors to identify the index of datasets across which the correlation structure is defined. This requires that components are uncorrelated within each transformed dataset and correlated only among the corresponding components across datasets. This is an important assumption made in the development of CMS in [18] as it works directly with observation matrices. When IVA-G is used as a first step as we propose to do, this condition is automatically satisfied.

2.2.1. Model order estimation

In this section, we show that if nth component is correlated across any subset of K datasets with $\rho^{[n]}_{k_1,k_2} \neq 0$ for $k_1 \neq k_2$ and $k_1, k_2 = 1, 2, \ldots, K$, then $\mathbf{R}_n$ has at least one eigenvalue greater than one. We start with the case where the nth component at each dataset share no correlation across datasets.

Case 1 ($\rho^{[n]}_{k_1,k_2} = 0 \; \forall \; k_1, k_2$): If nth components across all K datasets are all uncorrelated to each other, $\mathbf{R}_n$ is an $K \times K$ identity matrix and has eigenvalues all equal to one, i.e., $\lambda_i = 1, \quad i = 1, 2, \ldots K$.

Case 2 ($\rho^{[n]}_{k_1,k_2} = 1 \forall \, k_1, k_2$): If nth components are fully correlated across all K datasets with $\rho^{[n]}_{k_1,k_2} = 1$, then $\mathbf{R}_n$ is a matrix of ones. The characteristic polynomial of $\mathbf{R}_n$ is $(\lambda - K)\lambda^{K-1}$. The rank of the matrix is 1 and eigenvalues are equal to K with multiplicity 1 and 0 with multiplicity $K - 1$ [23]. If nth components are fully correlated across any subset of K datasets with dimension L, $\mathbf{R}_n$ has one eigenvalue equal to or greater than L.

Case 3 ($\rho^{[n]}_{k_1,k_2} \neq 0 : \exists \, k_1, k_2$): If nth components are correlated across any subset of K with non-zero correlation values, $\mathbf{R}_n$ can be written as

$$\mathbf{R}_n = \begin{bmatrix} \mathbf{F}_n & 0 \\ 0 & \mathbf{I} \end{bmatrix}, \tag{5}$$

where $\mathbf{F}_n$ is an $L \times L$ matrix similar to $\mathbf{R}_n$, where $L \leq K$. We assume that all correlation values are transitive within the $\mathbf{F}_n$ matrix, meaning if the nth component is correlated across datasets l_1 and l_2, and across datasets l_2 and l_3, it is also correlated across datasets l_1 and l_3, and the values of $\rho^{[n]}_{l_1,l_2}$ are either 0, or greater than $((L-1)/L)^2$ for $L \geq 4$ datasets. Under these assumptions, we have the results from Theorem 1 in [18] to show that $\mathbf{F}_n$ has at least one positive eigenvalue greater than one as long as correlation exists among nth components across any L datasets. Since $\mathbf{F}_n$ is similar to $\mathbf{R}_n$, then $\mathbf{R}_n$ also has at least one eigenvalue greater than one.

The above three cases indicate that the largest eigenvalue of $\mathbf{R}_n$ is bounded by 1 and K, and it is greater than one if correlation exist across any subset of datasets. That means if d components are correlated across multiple datasets we will have d SCV covarinace matrices with largest eigenvalue greater than one. One can use this property of the SCV covariance matrix to identify the model order or number of components correlated across datasets.

2.2.2. Correlation structure identification

We show that we can identify the model order d using the eigenvalues of $\mathbf{R}_n$'s. However, eigenvalues alone are not enough to identify the datasets across which these d components are correlated. Eigenvectors of $\mathbf{R}_n$, on the other hand, especially the ones corresponding to the eigenvalues greater than one, preserves the identity of the datasets in their non-zero elements. To show that we start from the case where $\mathbf{R}_n$ is an $K \times K$ identity matrix and has no eigenvalues greater than one. Since diagonal elements of $\mathbf{R}_n$ are all zeros, no correlation structure exists across any datasets. When $\mathbf{R}_n$ is a matrix of ones and has one eigenvalue equal to K and rest of the eigenvalues as zeros, the eigenvector corresponding to the eigenvalue K is the vector of all ones [23], meaning all datasets are contributing equally and nth component is correlated across all K datasets. When $\mathbf{R}_n$ has the structure in (5) and has at least one eigenvalue greater than one, we have the results of Theorem 2 in [18] to show that non-zero elements of the eigenvectors corresponding to the eigenvalues greater than one preserves the indices of the datasets $l_j, \; j = 1, \ldots L$, across which the nth component is correlated. This implies that the non-zero elements of the eigenvectors of $\mathbf{R}_n$ associated with the eigenvalues greater than one preserves the identity of the datasets across which the components are correlated. Based on this information, one can identify components that are correlated across all, subset of datasets and no dataset and form the correlated and distinct subspaces.

3. IMPLEMENTATION AND RESULTS

3.1. Simulation example

This section generates simulation examples for $K = 3$ datasets to compare the relative performance of PCA-CCA, MCCA-KPD, CMS and CMI-IVA. We compare the relative performance of all four methods to estimate the model order and evaluate the performance of

CMS and CMI-IVA only for the estimation of the correlation structure. This is because PCA-CCA and MCCA-KPD are limited to estimate only the model order. For each dataset, we generate $N = 10$ components from Laplacian distribution with $T = 1000$ independent and identically distributed (i.i.d.) samples. We select Laplacian distribution since it is a good match to fMRI data [24]. We introduce correlation to three components from each dataset with correlation values $0.9, 0.7$, and 0.5. These are the components correlated across all three datasets. In addition to that, we make a single component correlated across datasets 1 and 2 and another component across datasets 2 and 3, both using correlation value 0.5. Hence, there are $d = 5$ components correlated across datasets—3 across all datasets and 2 across pairwise datasets—and 5 are distinct to each dataset. These latent sources are then linearly mixed with mixing matrices, $\mathbf{A}_k \in \mathbb{R}^{N\times N}$, $k = 1, 2, 3$, with elements from a standard Gaussian distribution $N(0, 1)$ resulting datasets $\mathbf{X}_k \in \mathbb{R}^{N\times T}$ for $k = 1, 2, 3$. Finally, the Gaussian noise of dimension $N \times T$ with variance v is added to each dataset.

We evaluate each method's performance by changing the number of samples, correlation values, or signal to noise ratio (SNR) while keeping other parameters fixed. In the first case, we change the number of samples in each dataset from 200 to 2000. In the second case, we change the correlation values of only the pairwise correlated components from 0.1 to 0.9, where a small correlation value indicates less association across pairwise datasets, and a higher correlation value indicates the opposite. Finally, we adjust the SNR of the datasets by varying noise variance v to test each method's robustness under different noise levels. The performance of all three methods are averaged across 50 runs and shown in Figure 2.

In Figure 2, the first column shows the model order estimation performance of all four methods, and the second column shows CMS and CMI-IVA's performance for the identification of the corresponding correlation structure. We use Frobenius norm distance between true and estimated structures to determine the estimation error. Since PCA-CCA is limited to two datasets, we run PCA-CCA for all pairwise combinations of datasets and show the average order. We note that CMI-IVA and CMS estimate model order closer to the true order $d = 5$ compared with the other two methods in all three simulation cases. MCCA-KPD and PCA-CCA, due to their limitations, estimate model order lower than the actual value $d = 5$. In Figure 2(a), CMI-IVA and CMS's performance to estimate both model order and correlation structure improves with the increase in the number of samples. CMS provides better performance for small sample number compared with CMI-IVA, however CMI-IVA results lower structure estimation error than CMS as the the number of sample increases. Performance of MCCA-KPD for model order estimation also improves with the increase in sample number since both IVA and MCCA require large sample support for efficient estimation of SCVs. However, the performance of PCA-CCA remains unchanged due to its robustness for both large and small sample size scenarios. In Figure 2(b), model order estimated by the CMI-IVA and CMS switch from $d = 3$ to $d = 5$ and estimation error reduces as the correlation value of pair-wisely correlated components increases. It shows that as the association between the pairwise datasets grow strong, these two methods identify those associations and estimates model order accordingly. The performance of MCCA-KPD and PCA-CCA remain unchanged since both methods are unable to identify the components correlated across pairwise datasets. Finally, all four methods' performance improves as the SNR increases in Figure 2(c). Estimation error of CMI-IVA and CMS for the correlation structure also decreases with the increase in SNR values, where CMI-IVA provides more stable results with smaller error bars compared with CMS. Overall, the model order estimated by the CMI-IVA and CMS represents all the correlated components across the datasets and along with the estimated correlation structure, represents the true underlying relationship among the datasets better than MCCA-KPD and PCA-CCA. CMI-IVA, because of initial IVA-G step, provides more accurate estimation of correlation structure than CMS, specially for higher sample size, correlation values and lower SNRs. Given the large memory requirement for CMS, we use only CMI-IVA in the application to real data discussed next.

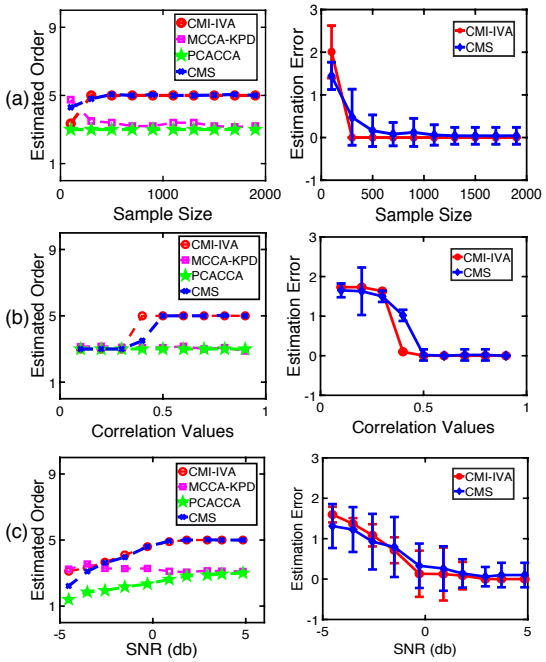

Fig. 2: Performance of PCA-CCA, MCCA-KPD, CMS and CMI-IVA for different (a) samples, (b) correlation values and (c) SNR. The first column shows the model order estimation performance of all four methods, and the second column shows the correlation structure estimation performances of only CMS and CMI-IVA since PCA-CCA and MCCA-KPD estimate only the model order and do not provide information about the correlation structure.

3.2. FMRI data and extracted features

We use multiset fMRI data collected from 150 healthy controls and 121 patients with schizophrenia performing auditory odd ball task (AOD). The fMRI datasets are from the Mind Research Network Clinical Imaging Consortium Collection [25] (publicly available at http://coins.mrn.org). During the collection process, subjects are listening to three different types of auditory stimuli; frequent low-tone stimuli (standard), infrequent task-irrelevant stimuli (novel), and infrequent task-relevant stimuli (target) requiring a button-press response. Infrequent tones are allocated in a pseudo-random manner to ensure randomness in the process. In this study, we use regressors created by modeling target+standard, novel, and only target stimuli as delta functions and convolving the functions with the default hemodynamic response function (HRF) in statistical parametric mapping (SPM) toolbox [26]. Subject averaged contrast images between the three stimuli tones are then used as multivariate features. Thus the feature datasets are formed for target+standard (target+std), novel, and target tones and have dimension, $\mathbf{X}_k \in \mathbb{R}^{271\times 48546}$, $k = 1, 2, 3$. All three datasets are then reduced to dimension 25, signal

subspace order estimated using minimum description length criterion and by taking sample dependency into account [14]. We note that 'target+std' and 'target' are the task-relevant datasets since they include responses related to the button-pressing task, while 'novel' is the task-irrelevant dataset. This particular selection of feature datasets allows us to test our expectations in terms of what can be expected to be shared and unshared across these three datasets.

3.3. Results and discussion

We use CMI-IVA to identify the number of correlated components and their corresponding correlation structure across three AOD feature datasets. We also use PCA-CCA and MCCA-KPD, but only to estimate the model order since they do not estimate the corresponding components or correlation structure. We do not able to use CMS here due to its limitation with large sample datasets. To enable better reproducibility of the results, we run CMI-IVA 25 times and select the most consistent run using cross intersymbol interference (Cross-ISI) measure [27]. In addition to that, we use $c = 0.2$ as a threshold to remove the insignificant correlation values from the SCV covariance matrices. In total, CMI-IVA identifies 24 components correlated across datasets, where 20 components are correlated across all three datasets and 4 components are correlated across the datasets 'target+std' and 'target'. To compare with the existing methods, PCA-CCA and MCCA-KPD estimate 18 and 16 as the model order. Since the data is from healthy controls and patients with schizophrenia, we perform a two-sample t-test on the columns of the estimated mixing matrices to identify the columns that show group difference ($p < 0.05$) between healthy controls and patients. We refer to the components associated with these columns as discriminative components. Estimated components identified as correlated across datasets and showing group differences are then thresholded at $Z = 2$ and shown in Figure 3.

Figure 3(a) shows the components correlated across all three datasets, while Figure 3(b) shows the components correlated across 'target+std' and 'target' datasets. Figure 3(c) shows the correlation matrices of components in Figure 3(a) and 3(b) across three feature datasets. In figure 3(a) and 3(b), the color red, orange, and yellow mean higher activation in healthy controls over the patient, and blue means the opposite. Here, estimated components show higher activation in dorsal default mode network (DMN) and visual areas for healthy controls, while in auditory, motor, and sensory-motor areas for patients with schizophrenia. These are the areas known to differentiate between healthy controls and patients with schizophrenia in many prior studies, e.g., [8, 28, 29], thus increasing our confidence in the method. Looking at the components in Figure 3(a), we note that components show group differences in DMN, auditory, and visual areas for all three datasets. These components are not directly related to the button-pressing task and have similar p-values across all three datasets. On the other hand, components in the first row of Figure 3(b) identified as correlated across two task-relevant datasets show activations in motor and sensory-motor areas, strongly related to the button-pressing task. Moreover, the components show lower p-values, i.e., higher differentiation between healthy subjects and patients, in task-relevant datasets. Overall, components identified as correlated across all datasets have similar p-values and show activation in areas not related to the task, while components identified as correlated across task related datasets have lower p-values and show activation in areas strongly related to the task. This explains the relationship among the datasets, where task-related datasets share more commonality between them compared with the non-task ones.

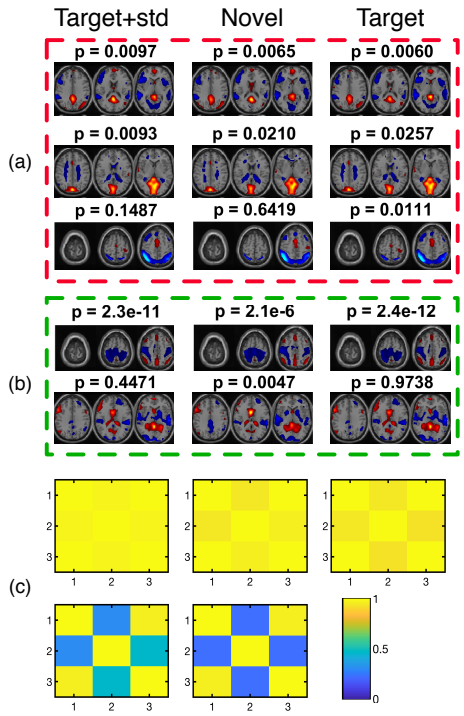

Fig. 3: Estimated brain components identified by CMI-IVA as linked across (a) all datasets and (b) 'target+std' and 'target' datasets, and (c) their corresponding correlation matrices. We are only showing the components that are correlated across datasets and at least one associated component showing group difference between healthy controls and patients. In (a) and (b), the color red, orange and yellow means higher activation in controls and blue means higher activation in patients. Here, components in (a) show activations in dorsal DMN, auditory and visual areas, while components in (b) show activations in motor and sensory motor areas.

4. CONCLUSION

In this paper, we introduce CMI-IVA for solving the complete model identification problem. We compare the method's performance in simulations with PCA-CCA, MCCA-KPD and CMS and then apply our method to real fMRI data. We find that CMI-IVA can identify the number of correlated components as well as the corresponding correlation structure across multiple datasets. When applied to real fMRI data, we show that task-related components estimated by CMI-IVA are correlated across task-related datasets, and non-task related components are correlated across all datasets, thus explaining the nature of the datasets' relationship in a meaningful way. One practical aspect of the implementation is that the exact threshold for eigenvalues will not be at exactly zero. In our implementation, we found out that simple zero-imputation of insignificant correlation values in the SCV covariance matrices provided satisfactory performance. However, it is desirable to make this threshold data driven as well. Possible ways to achieve this include use of information-theoretic criterion (ITC) or bootstrap based hypotheses tests [18]. Although we use multiset data in this work, CMI-IVA can also be used with multimodal datasets. In addition, the success of the method using IVA-G in this context inspires us to use other IVA algorithms such as those that take HOS of the data into account and apply for the fusion of more challenging problems such as subgroup identification in multi-subject data, common and distinct subspace identification in medical and behavioral datasets, remote sensing applications, among others.

5. REFERENCES

[1] D. Lahat, T. Adalı, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[2] E. Acar, R. Bro, and A. K. Smilde, "Data fusion in metabolomics using coupled matrix and tensor factorizations," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1602–1620, 2015.

[3] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28 – 44, 2013.

[4] Y. Zhang, Z. Dong, Sh. Wang, X. Yao X. Yu, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, Ja. Ramirez, F. J. Martinez, and J. M. Gorriz, "Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation," *Information Fusion*, vol. 64, pp. 149–187, 2020.

[5] T. Adalı, M. A. B. S. Akhonda, and V. D. Calhoun, "ICA and IVA for data fusion: An overview and a new approach based on disjoint subspaces," *IEEE Sensors Letters*, pp. 1–1, 2018.

[6] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Information Fusion*, vol. 19, pp. 4–19, 2014.

[7] J. Sui and V. D. Calhoun, *Multimodal Fusion of Structural and Functional Brain Imaging Data*, pp. 853–869, Springer New York, New York, NY, 2016.

[8] C. Jia, M. A. B. S. Akhonda, Q. Long, V. D. Calhoun, S. Waldstein, and T. Adalı, "C-ICT for discovery of multiple associations in multimodal imaging data: Application to fusion of fMRI and DTI data," in *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, March 2019, pp. 1–5.

[9] J. A. Pineda-Pardo, R. Bruña, M. Woolrich, A. Marcos, A. C. Nobre, F. Maestú, and D. Vidaurre, "Guiding functional connectivity estimation by structural connectivity in MEG: an application to discrimination of conditions of mild cognitive impairment," *NeuroImage*, vol. 101, pp. 765 – 777, 2014.

[10] Q. Yu, B. B. Risk, K. Zhang, and J. S. Marron, "JIVE integration of imaging and behavioral data," *NeuroImage*, vol. 152, pp. 38 – 49, 2017.

[11] J. S. Turek, C. T. Ellis, L. J. Skalaban, N. B. Turk-Browne, and T. L. Willke, "Capturing shared and individual information in fmri data," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 826–830.

[12] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 33, no. 2, pp. 387–392, April 1985.

[13] Yi-Ou Li, Tülay Adalı, and Vince D. Calhoun, "Estimating the number of independent components for functional magnetic resonance imaging data," *Human Brain Mapping*, vol. 28, no. 11, pp. 1251–1266, 2007.

[14] G. S. Fu, M. Anderson, and T. Adalı, "Likelihood estimators for dependent samples and their application to order detection," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4237–4244, Aug. 2014.

[15] Y. Wu, K. W. Tam, and F. Li, "Determination of number of sources with multiple arrays in correlated noise fields," *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1257–1260, June 2002.

[16] Y. Song, P. J. Schreier, and N. J. Roseveare, "Determining the number of correlated signals between two data sets using PCA-CCA when sample support is extremely small," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 3452 – 3456.

[17] S. Bhinge, Y. Levin-Schwartz, and T. Adalı, "Estimation of common subspace order across multiple datasets: Application to multi-subject fMRI data," in *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, 2017, pp. 1–5.

[18] T. Hasija, C. Lameiro, T. Marrinan, and P. J. Schreier, "Determining the dimension and structure of the subspace correlated across multiple data sets," *CoRR*, vol. abs/1901.11366, 2019.

[19] T. Adalı, M. Anderson, and G. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 18–33, May 2014.

[20] M. Anderson, T. Adalı, and X. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1672–1683, April 2012.

[21] M. Anderson, Geng-Shen Fu, R. Phlypo, and T. Adali, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4399–4410, Sep. 2014.

[22] R. Bellman, *Introduction to Matrix Analysis*, Society for Industrial and Applied Mathematics, 2nd edition, 1997.

[23] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge university press, 2012.

[24] Q. Long, S. Bhinge, Y. Levin-Schwartz, Z. Boukouvalas, V. D. Calhoun, and T. Adalı, "The role of diversity in data-driven analysis of multi-subject fMRI data: Comparison of approaches based on independence and sparsity using global performance metrics," *Human brain mapping*, vol. 40, no. 2, pp. 489–504, 2019.

[25] R. L. Gollub, J. M. Shoemaker, M. D. King, T. White, S. Ehrlich, S. R. Sponheim, V. P. Clark, J. A. Turner, B. A. Mueller, V. Magnotta, D. O'Leary, B. C. Ho, S. Brauns, D. S. Manoach, L. Seidman, J. R. Bustillo, J. Lauriello, J. Bockholt, K. O. Lim, B. R. Rosen, S. C. Schulz, V. D. Calhoun, and N. C. Andreasen, "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, 2013.

[26] SPM5, "Statistical Parametric Mapping," http://www.fil.ion.ucl.ac.uk/spm/software/spm5, 2011.

[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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 2581–2585.

[28] D. Öngür, M. Lundy, I. Greenhouse, A. K. Shinn, V. Menon, B. M. Cohen, and P. F. Renshaw, "Default mode network abnormalities in bipolar disorder and schizophrenia," *Psychiatry Research: Neuroimaging*, vol. 183, no. 1, pp. 59 – 68, 2010.

[29] M. A. B. S. Akhonda, Q. Long, S. Bhinge, V. D. Calhoun, and T. Adali, "Disjoint subspaces for common and distinct component analysis: Application to task FMRI data," in *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, March 2019, pp. 1–6.