

## RESEARCH ARTICLE

# Sparse Group Bases for Multisubject fMRI Data

MUHAMMAD USMAN KHALID 

School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia

e-mail: mukhalid@unimelb.edu.au

**ABSTRACT** Considering that functional magnetic resonance imaging (fMRI) signals from multiple subjects (MS) can be represented together as a sum of common and a sum of distinct rank-1 matrices, a new MS dictionary learning (DL) algorithm named sparse group (common + distinct) bases (sgBACES) is proposed. Unlike existing MS-DL algorithms that ignore fMRI data's prior information, it is formulated as a penalized plus constrained rank-1 matrix approximation, where  $l_1$  norm-based adaptive sparse penalty,  $l_0$  norm-based dictionary regularization, and lag-1 based autocorrelation maximization have been introduced in the minimization problem. Besides, spatial dependence among neighbouring voxels has been exploited for fine-tuning the sparsity parameters. To my best knowledge, the sgBACES algorithm is the first to effectively take temporal and spatial prior information into account for an MS-fMRI-DL framework. It also has the advantage of not requiring a separate sparse coding stage. Studies based on synthetic and experimental fMRI datasets are used to compare the performance of sgBACES with the state-of-the-art algorithms in terms of correlation strength and computation time. It emerged that the proposed sgBACES algorithm enhanced the signal-to-noise ratio (SNR) of the recovered time courses (TCs) and the precision of the recovered spatial maps (SMs). A 10.2% increase in the mean correlation value over the ShSSDL algorithm is observed for motor-task based fMRI data.


**INDEX TERMS** Sparse representation, dictionary learning,  $l_0$  constraint,  $l_1$  penalization, multi-subject analysis, fMRI, lag-1 autocorrelation, spatial dependencies, adaptive penalty.

## I. INTRODUCTION

Since its formation [1], fMRI has emerged as a powerful neuroimaging technique to investigate brain activity. Due to its high spatial resolution, it has been particularly precise in mapping neural activities during task-based (TB), or resting-state (RS) experiments [2]. For TB-fMRI data analysis, the general linear model (GLM) [3], which has been implemented in the statistical parametric mapping (SPM) toolbox [4] is commonly used. It employs regressors based on the canonical hemodynamic response function (HRF) and its derivatives [5]. An obvious disadvantage of this approach is its inability to account for any unpredictable experimental variance, such as HRF variability across subjects [6]. In contrast, the flexibility of data-driven methods such as independent component analysis (ICA) [7], principal component

analysis (PCA) [8], canonical correlation analysis (CCA) [9], and their variants [10]–[15] to adapt to individual hemodynamics across subjects and different functional networks by learning underlying trends from the data make them applicable to both TB activation detection and RS functional connectivity analysis. Therefore, these methods have been extensively adapted to fMRI data over the last three decades. In this period, ICA, in particular, has dominated the fMRI literature with applications to both TB and RS fMRI data.

Regarding MS analysis, spatial ICA (sICA) [16] has enjoyed more success than temporal ICA owing to fMRI data's lower spatial variations than temporal [17]. On the other hand, due to the sparse nature of brain networks, the sparse assumption has gained more popularity in the last decade, and this led authors in [18] to question ICA's ability to handle independence among components, which was refuted in [19]. They concluded that ICA does select for maximal independence and emphasized that, like

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang .

independence, sparsity is a reasonable assumption for fMRI analysis. Nevertheless, ICA's assumption of independence was again challenged in [20] where ICA, compared to sparse DL, faced difficulty retrieving neural dynamics when moderate to significant overlaps among functional networks were present. A similar trend was encountered in [21] where ICA could not reveal the activation maps that DL discovered. The sparse assumption has also been supported by biological evidence of sparse coding in the brain [22], and highlighted in an earlier ICA study [23].

In the last two decades, sparse representation [24] has been extensively utilized to tackle numerous signal and image processing problems such as denoising [25], inpainting [26], super-resolution [27], classification [28], and many others. Its usefulness becomes evident for fMRI when combined with DL [29]. It allows representing the blood-oxygen-level-dependent (BOLD) signal by a linear combination of a few atoms from the trained dictionary, thus allowing retrieval of underlying neural dynamics. This motivated the development of a sparse GLM framework for fMRI by authors in [30]. Since then many DL algorithms have been developed specifically for single-subject fMRI data, for instance, basis expansion dictionary [31], consistent adaptive sequential dictionary [32], adaptive complex-valued dictionary [33], and information assisted dictionary [34].

In the context of MS studies, there are some [35], [36] that directly apply existing DL algorithms such as ODL [26] to obtain group-level brain networks. Whereas, authors in other publications developed DL algorithms explicitly for MS-fMRI studies [37]–[40]. However, they could not inter-relate spatiotemporal (ST) dynamics across subjects, because their models could not retain the data decomposition format like the latter approaches presented in [41], [42]. Algorithms in [41], [42] were inspired by the approach used in image classification [43] to decompose the whole-brain (WB) MS datasets into common and subject-specific TCs and SMs while using incoherence penalty to transfer shared features to the common dictionary. These algorithms might suffer from convergence issues because they are based on alternating minimization (AM) approach that updates dictionary and sparse code separately and performance issues because they ignore fMRI data's prior information such as temporal smoothness and spatial dependence among neighbouring voxels. Moreover, recently, an MS-DL algorithm was applied to spatial features to reveal SMs that were common among healthy controls and schizophrenic subjects and those that were specific to each group [21], and more recently, MS-fMRI data were decomposed using the Tucker-2 model into shared, and individual TCs/SMs [44].

This paper proposes a new DL algorithm that decomposes the WB-fMRI dataset from multiple subjects into a sum of common and distinct rank-1 matrices. It is entirely different from existing MS algorithms in terms of the objective function, and both dictionary and sparse code (SC) update stage. In addition to the adaptive sparse penalty on coefficient rows, the proposed algorithm introduced smoothing and

autocorrelation maximization constraints on atoms in the rank-1 minimization problem. Its solution is based on a power method variant, and is named sparse group bases (sgBACES). A second algorithm named rgBACESW that takes a much simpler perspective on MS learning is also presented. The sgBACES algorithm yields superior dynamics compared to sICA [16], CODL [40], and ShSSDL [42] in terms of correlation strength of retrieved signals with the ground truth. In this paper, the main contributions are

- 1) A new rank-1 matrix approximation problem that takes into account the prior information about lag-1 autocorrelations, and the smoothness of fMRI time series and imposes an adaptive penalty on sparse coefficients to facilitate multi-subject fMRI learning.
- 2) For the proposed approximation problem, a computationally efficient algorithm, specifically when the number of data variables is from small to medium that learns sparse representation matrix (RM) or mixing matrix and sparse coefficient matrix as a pair and does not alternate between DL and SC stage.
- 3) A greedy approach that estimates adaptive sparse penalty parameters based on each voxel's two-dimensional (2D) neighbourhood within each slice.
- 4) A thresholding correlation-based RM estimation approach for the sparse basis expansion that particularly helps reduce the computational burden.

The rest of the paper is organized into six sections. Section 2 discusses the related work, section 3 provides background on multi-subject DL, section 4 describes the proposed algorithm, section 5 presents experimental studies, and the paper ends with section 6 containing concluding remarks.

## II. RELATED WORK

In essence, the sgBACES algorithm is motivated by the fact that when interesting source signals are mixed, their autocorrelation is lower, which can be maximized using exploratory techniques such as CCA [9]. In this work, the autocorrelation becomes even weaker due to the risk of overfitting through the mixing of DCT bases. Moreover, the effect of nuisance factors such as scanner-induced drift, which have influenced each voxel's time series by causing a lag-1 autocorrelation structure, might still exist even after the conventional temporal pre-processing [45]. This can be eliminated from the signals of interest while maximizing the autocorrelation [9].

As it was discussed in [31] that the BOLD signal at any voxel can be assumed smooth since the linear convolution model that is frequently used in fMRI leads to a neural response that is smoothed by the HRF. The notion of smoothness was included as prior information in the DL model where instead of imposing it through the roughness penalty matrix, it was enforced through sparse basis expansion via the  $l_1$  norm. This has motivated me to incorporate the DCT-based dictionaries into the proposed MS-DL model. The main drawback of using  $l_1$  norm-based penalization for sparse basis expansion is that it is hard to relate its penalty value to the number of selected DCT bases that represent the dictionary

**TABLE 1.** For fMRI group analysis, a summary of existing state-of-the-art data-driven algorithms, the proposed algorithm (sgBACES), and its variant (rgBACESW).

Algorithm	Data decomposition type	Methodology	Advantages and limitations
[16] sICA	PCA and sICA	Draws group inferences using three main steps i) PCA that reduces the temporal dimension of each subject's and all subject's concatenated dataset, ii) spatial ICA that extracts group ICs, and iii) back-reconstruction that obtains individual ICA maps	i) Only one tuning parameter, ii) unable to learn individual TCs, iii) retrieves TCs and SMs with low accuracy precisely when there are significant overlaps among the functional networks
[40] CODL	PCA and sparse DL	Draws group inferences using two main steps i) PCA that reduces the temporal dimension of each subject's dataset, and ii) sparse DL that extracts group SMs	i) Few tuning parameters, ii) unable to learn individual TCs and SMs, iii) retrieves TCs and SMs with medium accuracy
[42] ShSSDL	Sparse DL	Extracts common and subject-specific TCs and SMs using dictionary learning with single sparsity, and block dictionary update while imposing incoherence penalization among dictionaries	i) Few tuning parameters, ii) requires dual AM approach where dictionary and SC are updated separately that may result in convergence issues, iii) retrieves TCs and SMs with medium accuracy
Proposed (rgBACESW)	Sparse DL with sparsity promotion	Extracts common and subject-specific TCs and SMs using dictionary learning with single sparsity, adaptive sparse penalty, and sequential dictionary update	i) Few tuning parameters, ii) converges consistently due to single AM approach, iii) retrieves TCs and SMs with medium accuracy
Proposed (sgBACES)	Double sparse DL with sparsity promotion and ST autocorrelation maximization	Extracts common and subject-specific TCs and SMs using dictionary learning with double sparsity, adaptive sparse penalty, sequential dictionary update, and fMRI data's prior information	i) Several tuning parameters, ii) converges consistently due to single AM approach, iii) retrieves superior TCs and SMs due to utilization of prior information and DCT bases in a sparse manner

atom being estimated using the error matrix. This can be resolved by using  $l_0$  based constraint, because it counts for the number of non-zeros and naturally seems to be a better option for basis selection.

To solve the sparse signal recovery problem given by equation (1), an AM-based DL algorithm that rotates between the SC stage and the DL stage is usually considered, where the DL stage consists of sequential estimates of dictionary atoms and respective SC using singular value decomposition (SVD) on the reduced error matrix [46]. However, this solution is not always convergent over the iterates of a sequential learning [47], its alternative [32] on the other hand, has shown superior convergence. This algorithm is a variant of the power method that solves a regularized rank-1 matrix approximation for the full error matrix by promoting adaptive sparse penalty in the minimization problem resulting in consistent estimates of dictionary/SC.

Based on the discussion above and the ideas mentioned in the following (background) section, a novel algorithm named sgBACES and its minor variant named rgBACESW have been presented in this paper. As the proposed algorithm and its variant are quantitatively compared with other data-driven algorithms in this paper's experimental section, a theoretical comparison among them in terms of their decomposition type, methodology, advantages, and limitations is given in Table 1.

### III. BACKGROUND

As shown in Fig. 1, consider a single subject's whole-brain BOLD time courses consisting of  $N$  scanned volumes and

$V$  number of voxels arranged along the column direction in  $Y = [y_1, y_2 \dots y_V] \in \mathbb{R}^{N \times V}$ . Assuming, sparseness along its row direction each signal in  $Y$  can be represented as a linear combination of a few dictionary atoms from  $D \in \mathbb{R}^{N \times K}$  according to the sparse coefficient strength in each column of the sparse code matrix  $X \in \mathbb{R}^{K \times V}$  given as  $y_v \approx Dx_v, v = 1, \dots, V$ . To achieve this decomposition, a sparsity constraint is imposed on coefficient columns, and a normalization constraint on atoms as

$$\min_{D, X} \|Y - DX\|_F^2 \quad \text{sub.to. } \|x_v\|_0 \leq \zeta, \quad \|d_k\|_2 = 1, \quad (1)$$

where  $\|\cdot\|_F$ ,  $\|\cdot\|_2$ , and  $\|\cdot\|_0$  is the Frobenius,  $l_2$ , and  $l_0$  norm, respectively,  $\zeta \ll K$  is the hyperparameter that controls sparsity, and each atom is normalized to a unit norm to avoid scaling ambiguity. Due to the non-convexity of (1), it is solved conventionally using the AM approach, which comprises SC and DL stages. During the SC stage, sparse coefficient matrix is estimated using either coordinate descent with soft thresholding [48] in case of  $l_1$  norm on  $X$  or a greedy strategy such as orthogonal matching pursuit (OMP) [49] in case of  $l_0$  norm. On the other hand, a block update is performed when a computationally efficient dictionary is desired [50], or a sequential update is carried out to obtain a dictionary with superior performance [26], [46].

In order to extend (1) so that it can simultaneously analyze multiple subjects such that the dictionary which is common to all subjects, as well as a set of dictionaries that are specific to each subject can be learned, the ShSSDL algorithm was proposed in [42]. Wherein authors while assumed incoherence among the trained dictionaries  $D' = [D_c, D_1, D_2, \dots, D_M]$ ,

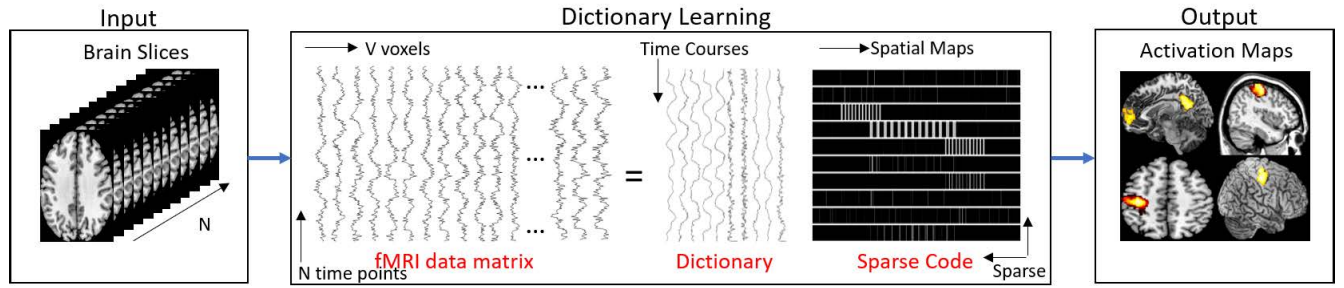


FIGURE 1. A data-driven fMRI analysis flowchart showing single-subject fMRI data matrix decomposition through dictionary learning.

proposed the following model to learn shared  $D_c$  and subject-specific dictionaries  $D_m$ ,  $\{m = 1, \dots, M\}$  for  $M$  number of subjects as

$$\min_{D', X'} \sum_{m=1}^M \|Y_m - D_c X_c - D_m X_m\|_F^2 + \eta \|D_m^T A_m\|_F^2, \quad (2)$$

sub.to.  $\|x_{c,v}\|_0 \leq \zeta_c, \quad \|x_{m,v}\|_0 \leq \zeta_m, \quad \|d'_k\|_2 = 1,$

where  $Y_m$  is the dataset from  $m$ -th subject,  $X' = [X_c^T, X_1^T, X_2^T, \dots, X_M^T]^T$ ,  $X_c \in \mathbb{R}^{K_c \times V}$ ,  $X_m \in \mathbb{R}^{K_m \times V}$ ,  $D_c \in \mathbb{R}^{N \times K_c}$ ,  $D_m \in \mathbb{R}^{N \times K_m}$ ,  $d'_k$  is the  $k$ -th column of  $D'$ ,  $\eta$  is the tuning parameter for controlling coherence,  $\zeta_c/\zeta_m$  are the tuning parameters that control sparsity for common/subject-specific sparse coefficient matrices  $X_c/X_s$  while  $x_{c,v}/x_{m,v}$  are their  $v$ -th columns, and  $A_m = [D_c, D_1, D_{m-1}, D_{m+1}, D_M]$  contains all dictionaries except the one that is currently being updated.

The strategy adopted to solve (2) consisted of AM between SC and DL stage, where  $D_c$  and  $D_m$  were kept fixed while updating  $X_c$  and  $X_m$  and then  $D_c$  and  $D_m$  were updated while keeping  $X_c$  and  $X_m$  fixed. Besides, the algorithm had an internal optimization routine between  $X_c$  and  $X_m$ , where two were updated in an alternating manner while keeping  $D_c$  and  $D_m$  fixed, although no such optimization was used for  $D_c$  and  $D_m$ . From experimental fMRI data, I observed that the convergence of this internal optimization had a considerable impact on this algorithm's overall performance. Moreover, when all other variables were kept fixed,  $X_c$  was updated via OMP using the mean of common-level residuals  $R_c = 1/M \sum_{m=1}^M Y_m - D_m X_m$  as

$$X_c = \arg \min_{X_c} \|R_c - D_c X_c\|_F^2 \quad \text{sub.to.} \quad \|x_{c,v}\|_0 \leq \zeta_c,$$

and on the other hand  $X_m$  was updated using the subject-level residuals  $R_m = Y_m - D_c X_c$  as

$$X_m = \arg \min_{X_m} \|R_m - D_m X_m\|_F^2 \quad \text{sub.to.} \quad \|x_{m,v}\|_0 \leq \zeta_m,$$

and dictionary  $D_c$  was updated in a block-wise manner using alternating directions method of multipliers (ADMM) that solves the augmented Lagrangian function given as

$$L(D_c, Z, W) = \|R_c - D_c X_c\|_F^2 + \eta \|Z^T A_c\|_F^2 + \mu \|D_c - Z\|_F^2 + \text{tr}[W^T (D_c - Z)].$$

where  $A_c$  contains only subject-level dictionaries,  $Z$  is the relaxation variable,  $W$  is the Lagrangian multiplier, and  $\mu$  is the tuning parameter. Similarly, an update for  $D_m$  can be obtained by solving an identical function with respect to  $D_m$ ,  $Z$ , and  $W$  by using  $R_m$ ,  $A_m$ ,  $D_m$ , and  $X_m$ . The use of the dual AM approach to solving (2) is motivated by the strategy used in conventional single-subject DL algorithms, that is, to update atoms while preserving the sparsity patterns in  $X_c$  and  $X_m$  [46]. Alternatively, as described in the next section, a regularized rank-1 approximation where sparsity is promoted can allow to update  $x_m^k/d_{m,k}$  as one pair and  $x_c^k/d_{c,k}$  as another.

#### IV. METHODS

Throughout this paper, scalar values are denoted by small italic or capital italic letters, vectors are represented by small, and matrices by capital letters. Vectors attached with subscripts and superscripts indicate the specific column and row of the matrix, respectively. Subscript signified by the letter  $c$  or  $m$  attached to any scalar value, vector, or matrix corresponds to a common or subject-level feature. Subscript  $p$  or  $i$  attached to the dictionary matrix indicate the DCT bases dictionary or partial dictionary whose accessible columns are based on the indices vector  $i$ . The error matrix attached with a subscript  $k$  indicates the  $k$ -th error matrix for the corresponding dictionary atom and SC row.

The dataset used in this paper is an open access data that is publicly available [51], and for this reason, ethics approval was not required from the Human Research Ethics of the University of Melbourne.

For fMRI group analysis, a dataset  $Y_m$  from  $m$ -th subject that can be represented as a linear combination of a few atoms from the common dictionary and the distinct (subject-specific) dictionary is considered to form an oversimplified model as

$$\min_{D', X'} \sum_{m=1}^M \|Y_m - D_c X_c - D_m X_m\|_F^2 + \zeta_c \|X_c\|_1 + \zeta_m \|X_m\|_1, \quad (3)$$

sub.to.  $\|d'_k\|_2 = 1$

where  $\|X\|_1$  is the  $l_1$  norm of  $X$  given as  $\sum_{k=1}^K \sum_{v=1}^V |x_v^k|$ . According to the conventional dictionary update approach [46],  $k$ -th dictionary atom/sparse code update for subject-specific dictionary is based on rank-1 approximation

## Algorithm 1 (A1) for Multi-Subject Dictionary Learning

**Given:**

Training set ( $Y_m \in \mathbb{R}^{N \times V}$ ,  $m = 1, \dots, M$ )  
 Tuning parameters ( $\zeta_c, \zeta_s, \lambda, \mu, \alpha, K_{pc}, K_{pm}$ )  
 Constants ( $K_c, K_m, T, \tau = 0.01$ )

1. **Initialize:**

$D_c, D_{pc}, D_m, D_{pm}$  with DCT bases  
 $B_c, B_m$  with identity matrix I  
 $X_c, X_s, D_{c,t}$  with matrix of zeros O

2. **while**  $\|D_c - D_{c,t}\|_F / \|D_{c,t}\|_F > \tau$   
 3. **for**  $t \leftarrow 1$  to  $T$

4.  $D_{c,t} \leftarrow D_c$

5. Subject-specific dynamics estimation:

**for**  $m \leftarrow 1$  to  $M$

    Compute:  $R_m \leftarrow Y_m - D_c X_c$

    Estimate:  $B_m$  &  $X_m$  using A2 for sgBACES **or**

    Estimate:  $B_m$  &  $X_m$  using A4 for rgBACESW

    Compute:  $D_m \leftarrow D_{pm} B_m$

    Compute:  $R_{cm} \leftarrow Y_m - D_m X_m$

**end for**

6. Common dynamics estimation:

    Compute:  $R_c \leftarrow 1/M \sum_{m=1}^M R_{cm}$

    Estimate:  $B_c$  &  $X_c$  using A2 for sgBACES **or**

    Estimate:  $B_c$  &  $X_c$  using A4 for rgBACESW

    Compute:  $D_c \leftarrow D_{pc} B_c$

7. **end for**

8. **end while**

**Output:**  $D_c, D_m, X_c$  and  $X_m$

of the error matrix of all signals  $E_k = R_m - \sum_{i=1, i \neq k}^{K_m} d_{m,i} x_m^i$  when  $k$ -th atom/sparse code has been removed

$$\{d_{m,k}, x_m^k\} = \left\| E_k - d_{m,k} x_m^k \right\|_F^2$$

and for common dictionary/sparse code update when error matrix is  $E_k = R_c - \sum_{i=1, i \neq k}^{K_c} d_{c,i} x_c^i$  it is

$$\{d_{c,k}, x_c^k\} = \left\| E_k - d_{c,k} x_c^k \right\|_F^2$$

**A. sgBACES**

This subsection discusses the sgBACES algorithm, which deduces its name from **s**parse **g**roup **b**ases using **a**daptive **c**onsistent **s**equential dictionary learning that can maximize autocorrelation via quadratic error minimization.

A rank-1 matrix approximation problem for atom/SC update has been formulated as a fusion of  $l_1$  norm adaptive

## Algorithm 2 (A2) for Solving the Minimization Problem (4)

**Given:**

Subject or common level residual matrix ( $R \in \mathbb{R}^{N \times V}$ )  
 Tuning parameters ( $\zeta, \lambda, \mu, \alpha$ )  
 Dictionary, mixing, and sparse code matrix ( $D_p, B, X$ )  
 Constant ( $K$ )

1. **Initialize:**

$D \leftarrow D_p B$

2. **for**  $k \leftarrow 1$  to  $K$

3.  $x^k \leftarrow 0$

4.  $f_i^1 \leftarrow 0, f_i^n \leftarrow d_i^{n-1}, i = 1, \dots, K, n = 2, \dots, N$

5.  $Z \leftarrow \frac{1}{\mu+2} (F^T F)^{-1} F^T D X$

6.  $E_k \leftarrow R - D X + F Z$

7.  $\tilde{x}^k \leftarrow d_k^T E_k$

8. Estimate  $\hat{x}^k$  using A3

9. Estimate  $\zeta_j^k$  for  $j = 1, \dots, V$ :

**if**  $|x_j^k| > \alpha |\tilde{x}_j^k|$

$\zeta_j^k \leftarrow \frac{\zeta}{|x_j^k|}$

**else**

$\zeta_j^k \leftarrow \frac{\zeta}{|\tilde{x}_j^k|} (1 + |x_j^k|)$

**end if**

10.  $x^k \leftarrow \text{sgn}(\tilde{x}^k) \circ \left( |\tilde{x}^k| - \frac{\zeta^k}{2} \right)_+$

11.  $c \leftarrow \text{sort} \left( \left| D_p^T (d_k + \eta E_k x^k) \right|, \text{descend} \right)$

12.  $i_s \leftarrow \{c\}_{l=1}^\lambda$

13.  $b_{k,i_s} \leftarrow (D_{p,i_s}^T D_{p,i_s})^{-1} D_{p,i_s}^T (d_k + \eta E_{k,i_r} x_{i_r}^k)$

14.  $b_k \leftarrow b_k / \|D_p b_k\|_2$

15.  $d_k \leftarrow D_p b_k$

16. **end for**

**Output:** B and X

penalization on the coefficient matrix (X),  $l_0$  constraint on the representation matrix (B), and lag-1 error minimization for the dictionary matrix (D). The  $l_1$  penalization results in a consistent estimate of the atom/SC as a pair for common/distinct updates. The  $l_0$  constraint provides the regularization of dictionary atoms through sparse basis expansion. The lag-1 error minimization allows producing atoms with maximum autocorrelation. Whereas, the adaptive threshold for each entry of the sparse code allows to control the shrinkage amount in a multisample data-driven manner. Overall, this strategy enables processing of the WB-MS-fMRI data within a reasonable amount of computation time to acquire common and distinct TCs and SMs that closely resemble the ground truth.

The lag-1 autocorrelations are considered by introducing variables  $F \in \mathbb{R}^{N \times K}$  and  $Z \in \mathbb{R}^{K \times V}$  where each column of F is the lag-1 version of its corresponding column in the original dictionary, and its respective SC estimate is stored in Z. Assuming that the currently updated dictionary atom

Algorithm 3 (A3) for 2D Convolution Operation

**Given:**

Row vector  $\tilde{x}$  of size  $V = \{P \times Q \times R\}$ , where  $p = \{1, \dots, P\}$ ,  $q = \{1, \dots, Q\}$ ,  $r = \{1, \dots, R\}$ ,  $v = \{1, \dots, V_r\}$  with  $V_r$  number of voxels within  $r$ -th slice, and  $P \times Q$  represent the size of each of the  $R$  slices

1. Reshape  $\tilde{x} \in \mathbb{R}^V$  to produce 3D arrays H, and  $W \in \mathbb{R}^{P \times Q \times R}$
2. Initialize  $H_r \in \mathbb{R}^{P \times Q}$  with zeros and zero pad the corners of W
3. For  $v$ -th voxel:  $h_r(p, q) = \sum_{u=-1}^1 \sum_{w=-1}^1 |W_r(p-u, q-w)|$
4. Repeat step 3 for all  $V_r$  number of voxels
5. Repeat steps 2, 3, and 4 for all  $R$  slices
6. Reshape 3D array H to a row vector and store it in  $\hat{x}$

**Output:**  $\hat{x}$

has lower autocorrelation due to nuisance factors and overfitting by DCT bases and given that the transformed variates are uncorrelated  $Z^T F^T F Z = I$ , the aim is to minimize the quadratic error between the currently updated atom/SC row and whole dictionary/SC matrix to produce an atom that has maximum autocorrelation [14]. By considering these propositions, the penalized plus constrained minimization problem for the rank-1 approximation of the full error matrix  $E_k$  is

$$\begin{aligned} \{b_k, x^k, Z\} = \arg \min_{b_k, x^k, Z} & \left\| 2E_k - D_p b_k x^k \right\|_F^2 \\ & + \left\| D_p b_k x^k - 2FZ \right\|_F^2 + 2 \sum_{j=1}^V \zeta_j^k |x_j^k|, \\ \text{sub.to. } & \|b_k\|_0 \leq \lambda, \quad Z^T F^T F Z = I, \\ & \|D_p b_k\|_2 = 1 \end{aligned} \quad (4)$$

where  $E_k$  is either based on a common level residual as  $R_c - \sum_{i=1, i \neq k}^{K_c} d_{c,i} x_c^i$  or a subject level residual as  $R_m - \sum_{i=1, i \neq k}^{K_m} d_{m,i} x_m^i$ ,  $D_p \in \mathbb{R}^{N \times K_p}$  is the base dictionary [52], which can be built using DCT basis or spline basis [53],  $B \in \mathbb{R}^{K_p \times K}$  is the sparse RM [31] and  $b_k$  is its  $k$ -th column,  $K < K_p < N$ ,  $\|\cdot\|_0$  counts the number of non-zero elements,  $\zeta_j^k$  is the sparsity parameter controlling every coefficient value in  $k$ -th row,  $\zeta^k = \zeta_1^k, \dots, \zeta_V^k$ ,  $\zeta$  is a scalar value provided to the algorithm, and  $I$  is the identity matrix. As  $l_0$  norm is associated with a constrained least square problem, it is posed and solved as a separate problem, and not made part of the regularized problem while deriving  $x^k$  [54]. Therefore, sequential update for  $d_k = D_p b_k$  and  $x^k$  and a block update for  $Z$  can be obtained by solving the following for (4)

$$\begin{aligned} L(b_k, x^k, Z) = & 4E_k^T E_k + 2x^{kT} b_k^T D_p^T D_p b_k x^k \\ & - 4E_k^T D_p b_k x^k - 4Z^T F^T D_p b_k x^k \\ & + 2(\mu + 2)Z^T F^T F Z + 2 \sum_{j=1}^V \zeta_j^k |x_j^k| \end{aligned} \quad (5)$$

Algorithm 4 (A4) for Solving the Minimization Problem (13)

**Given:**

Subject or common level residual matrix ( $R \in \mathbb{R}^{N \times V}$ )  
 Tuning parameter ( $\zeta$ )  
 Dictionary, mixing, and sparse code matrix ( $D_p, B, X$ )  
 Constants ( $\epsilon = 10^{-5}, K$ )

1. **Initialize:**

$$D \leftarrow D_p B$$

2. **for**  $k \leftarrow 1$  to  $K$
3.  $b_o \leftarrow 0, x^k \leftarrow 0$
4.  $E_k \leftarrow R - DX$
5.  $\tilde{x}^k \leftarrow d_k^T E_k$
6.  $\zeta_j^k \leftarrow \zeta / |\tilde{x}_j^k|$  for  $j = 1, \dots, V$
7. *Representation vector and sparse code estimation:*  
**while**  $\|b_k - b_o\|_2 / \|b_o\|_2 > \epsilon$   
 $b_o \leftarrow b_k$   
 $x^k \leftarrow \text{sgn}(b_k^T D_p^T E_k) \circ \left( |b_k^T D_p^T E_k| - \frac{\zeta^k}{2} \right)_+$   
 $b_k \leftarrow (D_p^T D_p)^{-1} D_p^T (d_k + \eta E_k x^{kT})$   
 $b_k \leftarrow b_k / \|D_p b_k\|_2$   
**end while**
8.  $d_k \leftarrow D_p b_k$
9. **end for**

**Output:** B and X

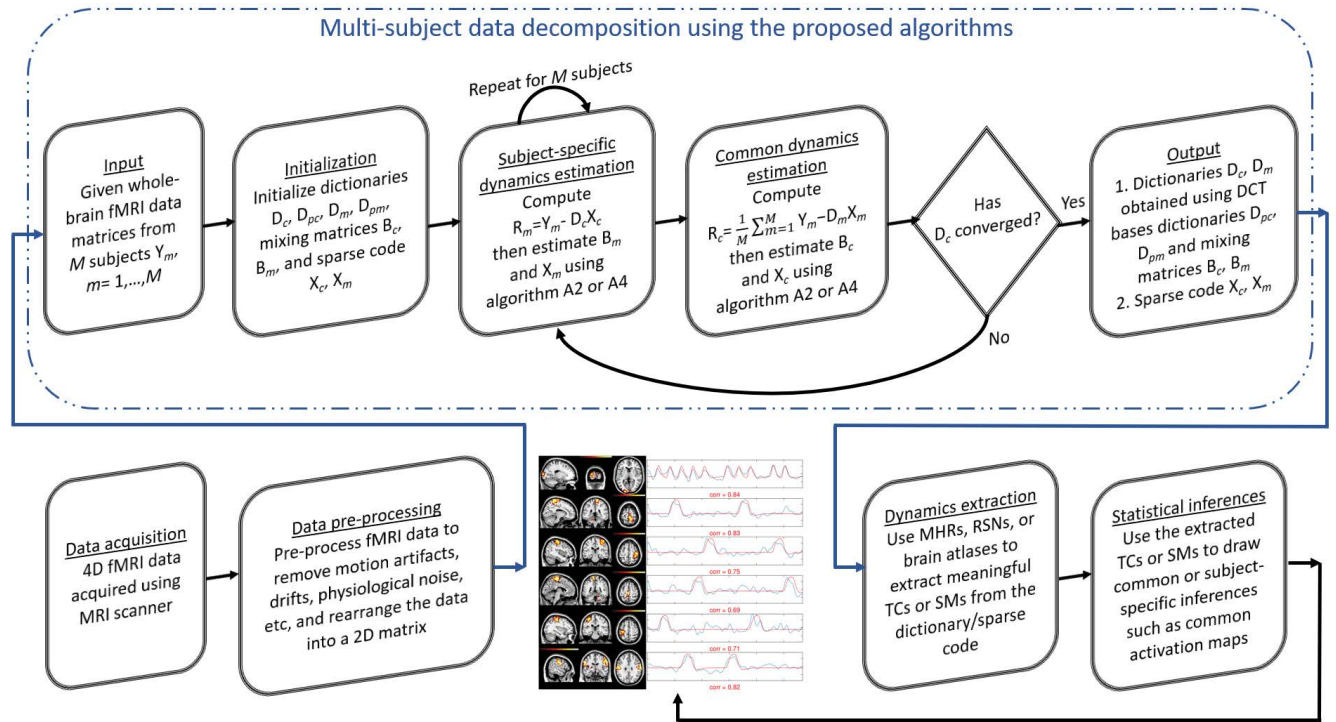
where  $\mu$  is the Lagrangian multiplier. By solving (5) with respect to  $x^k$ , a closed-form solution for the sparse code row is obtained as

$$\begin{aligned} 4b_k^T D_p^T D_p b_k x^k - 4b_k^T D_p^T E_k - 4b_k^T D_p^T FZ \\ + 2 \sum_{j=1}^V \zeta_j^k \frac{d|x_j^k|}{dx^k} = 0 \\ \implies x^k = \text{sgn}(b_k^T D_p^T \acute{E}_k) \circ \left( |b_k^T D_p^T \acute{E}_k| - \frac{\zeta^k}{2} \right)_+ \end{aligned} \quad (6)$$

where  $\acute{E}_k = E_k + FZ$ , vector  $\zeta^k$  holds  $V$  different values,  $(x)_+$ ,  $\text{sgn}(\cdot)$ , and  $\circ$  define the component-wise max between  $(0, x)$ , the component-wise sign, and the Hadamard product, respectively [55]. At this point, it is necessary to consider (4) to solve for  $x^k$  without penalization on coefficient matrix and constraint on sparse representation matrix and denote it as  $\tilde{x}^k$  (the non-penalized estimate of the  $k$ -th coefficient row). Hence, by reconsidering (5) and solving it with respect to  $\tilde{x}^k$ , following solution is obtained

$$\begin{aligned} 4b_k^T D_p^T D_p b_k \tilde{x}^k - 4b_k^T D_p^T E_k - 4b_k^T D_p^T FZ = 0 \\ \implies \tilde{x}^k = b_k^T D_p^T (E_k + FZ) \end{aligned} \quad (7)$$

Equation (7) becomes meaningful when the shrinkage amount for each entry of the corresponding penalized sparse code row  $x^k$  is estimated. This is because, it was discussed in [32] that simultaneously tuning all entries of  $\zeta^k$  can be



**FIGURE 2.** A flowchart describing the proposed multi-subject dictionary learning framework.

a daunting task, and therefore it was considered a reliable approach to compute it by utilizing  $\tilde{x}^k$  as  $\zeta^k = \zeta 1_V / |\tilde{x}^k|$ , where  $1_V$  is a vector of ones of length  $V$ . However, this might be an inferior approach because it utilizes only one data sample to estimate each threshold value, and due to this reason, another optimization problem is imposed for finding an optimal estimate of  $\zeta^k$ , which is given as

$$\zeta^k = \arg \min_{\zeta^k} \zeta 1_V / |\tilde{x}^k| \quad \text{sub.to.} \quad \zeta_j^k \lll \zeta / |\tilde{x}_j^k| \quad (8)$$

where  $\lll$  indicates much much smaller and much much greater. Finding a closed-form solution from (8) is hard, and a greedy approach must be adopted. An intuitive strategy would be to rely on quadratic or cubic power of  $|\tilde{x}^k|$ , but it may produce an estimate that is not the function of any additional data samples and thus may not lead to an estimator with minimum variance [56]. Computation-wise, exploiting statistical dependencies that exist in MR images among neighbouring voxels is a viable approach to obtain a small amount of shrinkage to significant entries with strong neighbourhood and vice-versa. To elaborate, a convolution operation can be used to sum up all the values in the  $3 \times 3$  grid neighbourhood of each entry of  $x^k$  as  $\zeta_j^k = \zeta / (\sum_{l=1}^8 |\tilde{x}_l^k| + |\tilde{x}_j^k|)$ , where  $\sum_{l=1}^8 |\tilde{x}_l^k|$  is the sum of absolute coefficient values of 8 voxels in the neighbourhood of  $|\tilde{x}_j^k|$  [9], and  $|\tilde{x}_j^k| = \sum_{l=1}^8 |\tilde{x}_l^k| + |\tilde{x}_j^k|$ . Moreover, this approach may generally reduce the shrinkage amount of all coefficient entries, even those with a weak entry and a weak neighbourhood, which can be resolved by imposing this condition  $|\tilde{x}_j^k| > \alpha |\tilde{x}_j^k|$ , where  $\alpha$  is a tuning parameter. So the neighbourhood values of  $\tilde{x}_j^k$  are used in a

different way when it itself and its neighbourhood is weak as  $\zeta_j^k = \frac{\zeta}{|\tilde{x}_j^k|} (1 + |\tilde{x}_j^k|)$ . After a few mathematical manipulations, it can be shown that a solution from (5) can be obtained for  $b_k$  as

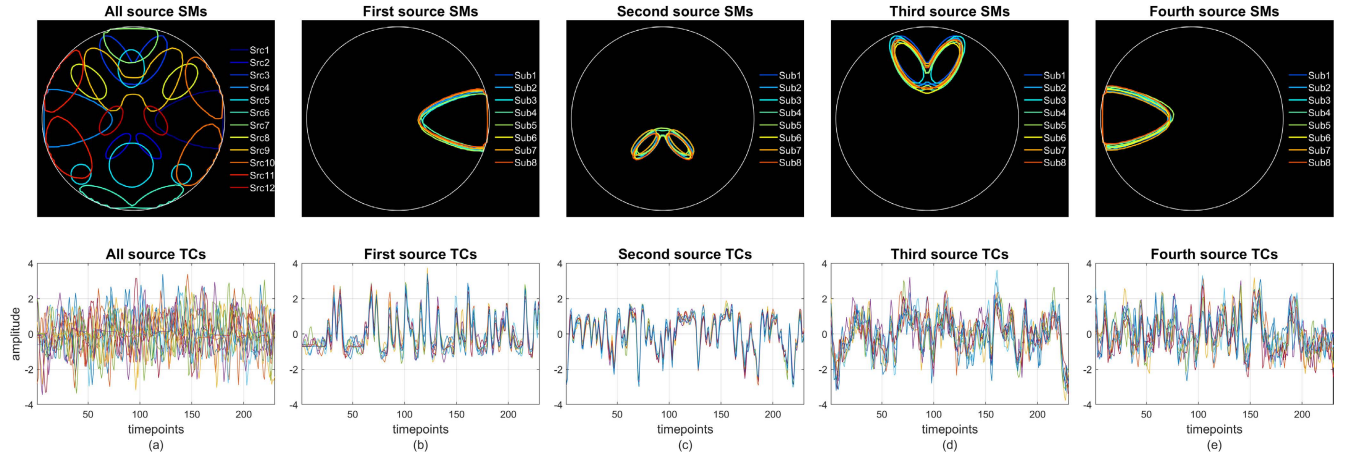
$$4D_p b_k x^k x^{k\top} - 4E_k x^k x^{k\top} - 4FZ x^k x^{k\top} = 0 \\ \implies b_k = (D_p^\top D_p)^{-1} D_p^\top (d_k + \eta \dot{E}_k x^k x^{k\top}) \quad (9)$$

where  $\eta = 1/x^k x^{k\top}$ . However, after taking into account the  $l_0$  norm constraint that was imposed on  $b^k$  in problem (4), its solution can be reformulated as a constrained problem

$$b_k = \arg \min_{b_k} \|d_k + \eta \dot{E}_k x^k x^{k\top} - D_p b_k\|_2^2, \\ \text{sub.to.} \quad \|b_k\|_0 \leq \lambda \quad (10)$$

This is a well-known NP-hard minimization problem, and finding its optimal solution is impossible. However, greedy approaches such as OMP can be used to find an approximate solution. In this case, a solution based on thresholding the correlation values [57] was opted for due to this approach's computational efficiency. For this purpose, finding the indices set  $i_s$ , which contains the indices of  $\lambda$  number of bases in  $D_p$  that are most correlated with  $d_k + \eta \dot{E}_k x^k x^{k\top}$  is the central idea. Once the indices vector  $c$  is found by sorting  $|D_p^\top (d_k + \eta \dot{E}_k x^k x^{k\top})|$  in descending order, the indices corresponding to lambda-largest values can be extracted as  $i_s = \{c_j\}_{j=1}^\lambda$ , and (9) can be rewritten as

$$b_{k,i_s} = (D_{p,i_s}^\top D_{p,i_s})^{-1} D_{p,i_s}^\top (d_k + \eta \dot{E}_k x^k x^{k\top}) \quad (11)$$



**FIGURE 3.** The top row in column (a) highlights the location and shape of 12 different spatial sources, whereas its bottom row shows the corresponding temporal sources. Columns (b-e) highlight the first four spatiotemporal sources that are common to all subjects but have spatial and temporal variability across subjects.

The computational burden can be further reduced by considering the non-zero entries of the coefficient row, which leads to a reduced error matrix. This is because the promotion of sparsity through the use of full error matrix  $\hat{E}_k$  only affects the sparse code, and preservation of sparsity by deploying reduced error matrix  $\hat{E}_{k,i_r}$  is enough to obtain consistent estimates of  $b^k$ . Particularly, in order to obtain relevant (non-zero) indices  $i_r$  “any non-zero” operator is defined as ( $\mathbb{R}_{\neq 0}$ ) that selects for indices corresponding to non-zero entries in  $x^k$  as  $i_r = \mathbb{R}_{\neq 0}x^k$  to retain only its non-zero entries as  $x_{i_r}^k = x_{\mathbb{R}_{\neq 0}x^k}^k$ , and corresponding columns of the error matrix as  $\hat{E}_{k,i_r} = \hat{E}_{k,\mathbb{R}_{\neq 0}x^k}$ . Thus, the computationally efficient update for  $b_k$  is obtained as

$$b_{k,i_s} = (D_{p,i_s}^\top D_{p,i_s})^{-1} D_{p,i_s}^\top (d_k + \eta \hat{E}_{k,i_r} x_{i_r}^{k\top}) \quad (12)$$

Either one can pursue a different route and obtain the block update for  $Z$  according to the approach that was carried out in [14] using

$$Z = \arg \min_Z \|2Y_m - D_p B X\|_F^2 + \|D_p B X - 2FZ\|_F^2 + 2 \sum_{j=1}^V \zeta_j^k |x_j^k|,$$

$$\text{sub.to. } \|b_k\|_0 \leq \lambda, \quad Z^\top F^\top F Z = I, \quad \|D_p b_k\|_2 = 1$$

or one can get the block update of  $Z$  from (5) by employing full dictionary and coefficient matrices instead of just currently updated atom/sparse code given as

$$4(\mu + 2)F^\top F Z - 4F^\top D_p B X = 0 \\ \implies Z = \frac{1}{\mu + 2} \left( (F^\top F)^{-1} F^\top D_p B X \right)$$

By taking everything into account, the multi-subject dictionary learning framework that utilizes the proposed algorithms is described in Algorithm 1. Here the main idea had been to update the dictionary atom and its respective sparse

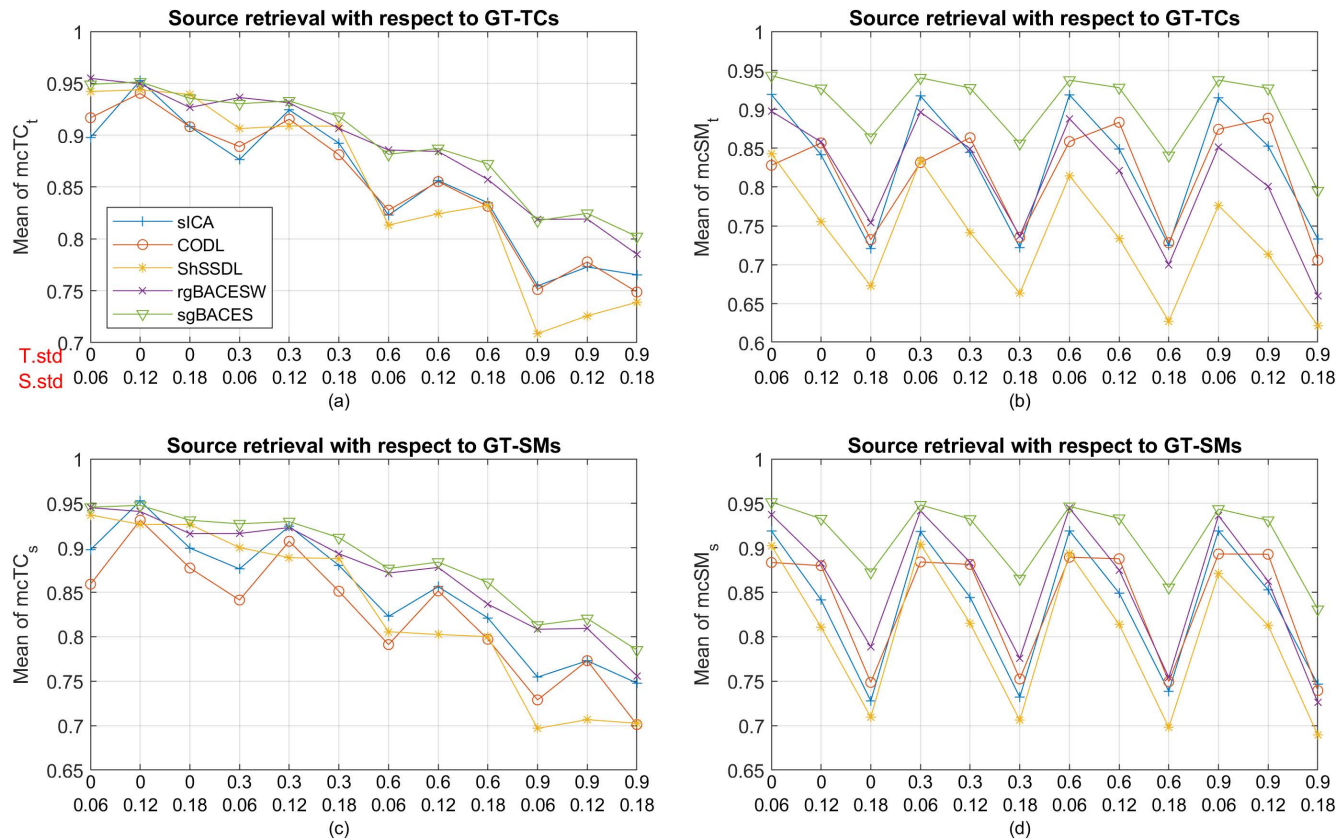
code row together,  $D_{c,t}$  represents the common dictionary at  $t$ -th iteration of the algorithm,  $R_{cm}$  is the common-level residual from  $m$ -th subject,  $I$  represents the identity matrix that has been split according to the values  $K_c/K_m$ ,  $O$  stands for a matrix of zeros, and  $T$  indicates the total number of iterations. Unlike the dual AM approach used in ShSSDL, the proposed multi-subject DL routine given in Algorithm 1 consists of a single AM approach between common dictionary/sparse code ( $x_c^k/d_{c,k}$ ) update stage and subject-specific dictionary/sparse code ( $x_m^k/d_{m,k}$ ) update stage, however, each of these dictionary/sparse code set is updated as a pair. They are updated using the estimates obtained from the minimization problem (4) or (13), and their algorithms are given in Algorithm 2 and 4, respectively. Both algorithms sequentially iterate over all dictionary columns and corresponding sparse code rows, where  $R$  is the generic term used for the matrix of residuals. In order to adaptively estimate each voxel’s sparse penalty, Algorithm 2 deploys the 2D convolution operation via Algorithm 3.

### B. rgBACESW

In contrast, a rather simplistic approach was adopted in the rgBACESW algorithm where regular basis expansion is used, the autocorrelation of dictionary atoms is not maximized, and neighbouring voxel values are not utilized while estimating the adaptive sparse penalty. Its algorithm is presented as Algorithm 4. Here, dictionary and sparse code are still updated as a pair, but there is no sparsity constraint on the mixing matrix, and the prior spatiotemporal information about fMRI data is not taken into account. It is based on the following formulation [31], [32].

$$\{b_k, x^k\} = \arg \min_{b_k, x^k} \|E_k - D_p b_k x^k\|_F^2 + \sum_{j=1}^V \zeta_j^k |x_j^k|, \\ \text{sub.to. } \|D_p b_k\|_2 = 1 \quad (13)$$





**FIGURE 4.** a) and b) show the mean values of mcTC and mcSM, respectively over 100 trials for the first recovery scheme, i.e., when sources are recovered with respect to GT-TCs, whereas c) and d) show the mean values of mcTC and mcSM, respectively over 100 trials when sources are retrieved with respect to GT-SMs. T.std and S.std labels along the x-axis indicate the standard deviation for temporal and spatial noise, respectively.

Update for  $d_k = D_p b_k$  and  $x^k$  as a pair obtained by solving the Lagrangian expression for (13) is given by

$$x^k = \text{sgn}\left(b_k^T D_p^T E_k\right) \circ \left(|b_k^T D_p^T E_k| - \frac{\zeta^k}{2}\right)_+$$

$$b_k = (D_p^T D_p)^{-1} D_p^T \left(d_k + \eta E_k x^{kT}\right)$$

The dictionary learning process is presented in Fig. 2 showing how to utilize sgBACES or its variant rgBACESW for multi-subject fMRI analysis.

## V. EXPERIMENTS

In this section, the performance of the proposed algorithm is evaluated and compared to existing MS-DL algorithms using two different multi-subject datasets. The comparative study involved the sICA, CODL, ShSSDL, rgBACESW, and sgBACES algorithms. The first was a synthetic fMRI dataset for 8 subjects generated using Simtb toolbox [58]. The second dataset was an experimental (motor task-based 3T MRI unprocessed dataset) fMRI dataset acquired by randomly selecting 24 subjects without taking into account their gender and age from the quarter 3 release of the Human Connectome Project (HCP) [59]. The resulting datasets consisted of young adults (male and female) aged between 22 and 35 years, except one subject whose age was over 36 years. Using both

datasets, the performance of all algorithms is compared in terms of their ability to retrieve the common and subject-specific spatiotemporal patterns.

### A. MATLAB CODES

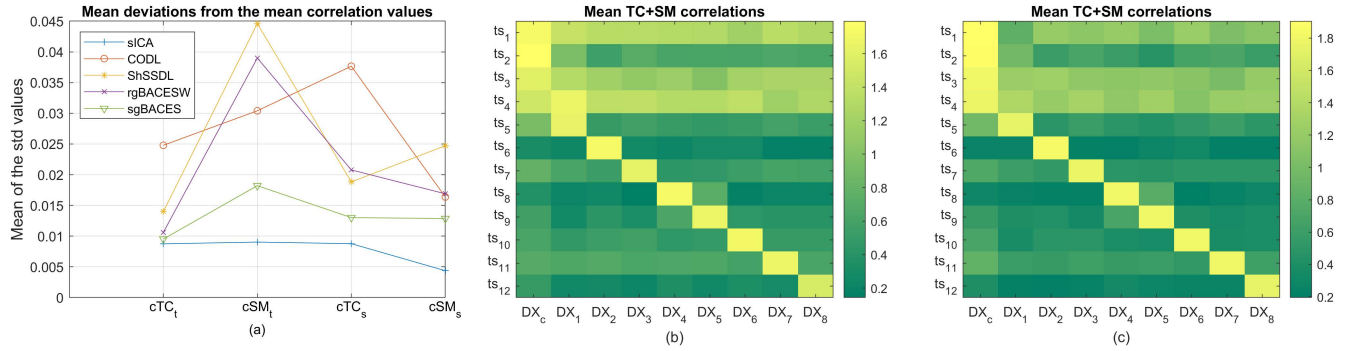
All algorithms are implemented in Matlab 2021b. The Matlab implementation of sICA was based on the derivations described in the spatial ICA paper [16] using fastICA [60] package.<sup>1</sup> The Matlab implementation of the dictionary part of the CODL algorithm was performed according to the derivations in the paper [26] whereas the Matlab code for  $l_1$  norm-based sparse coding was obtained from the spams toolbox (mexLasso).<sup>2</sup> The Matlab code for ShSSDL was taken from the Github link provided by the authors,<sup>3</sup> however, the spams toolbox's mexOMP was used for its sparse coding instead of OMP-Box of [61]. The simulation code (in the form of mat files) of sgBACES and its variant rgBACESW has been made available online at the time of this paper's submission.<sup>4</sup> All results, including the computation time reported in this

<sup>1</sup><https://research.ics.aalto.fi/ica/fastica/code/dlcode.shtml>

<sup>2</sup><http://thoth.inrialpes.fr/people/mairal/spams/downloads.html>

<sup>3</sup>[https://github.com/AsifIqbal8739/ShSSDL\\_2017](https://github.com/AsifIqbal8739/ShSSDL_2017)

<sup>4</sup><https://codeocean.com/capsule/0579804/tree>



**FIGURE 5.** a) Over 100 trials and 12 noise instances mean of the std values for four correlation cases, and over 100 trials and 12 noise instances, the correlation matrix between the ground truth TCs and SMs (TS) and the trained atoms and sparse code (DX) respectively obtained using atoms/SC from b) ShSSDL and, c) sgBACES.

**TABLE 2.** Table of tested values for parameter selection for the synthetic dataset.

Algorithm	Tested Values
CODL	$b = \{200, 600, 1200, 3600\}, \beta = \{.25 : .25 : 3\}$
ShSSDL	$\zeta_c = \{1, 2, 3\}, \zeta_m = \{1, 2, 3\}, \eta = \{0, 0.5, 1, 2, 5, 20\}$
rgBACESW	$\zeta_c = \{.25 : .25 : 4\}, \zeta_m = \{2 : 2 : 30\}, K_{pc}/K_{pm} = \{60, 90, 120, 150\}$
sgBACES	$\zeta_c = \{0.5 : 0.5 : 15\}, \zeta_m = \{5 : 15 : 170\}, \alpha = \{0.5 : 0.5 : 10\}, \mu = \{1, 2, 3, 4, 8, 18, 28, 38, 48, 98, 198\}, \lambda = \{15 : 15 : 150\}, K_{pc}/K_{pm} = \{60, 90, 120, 150\}$

paper for rgBACESW and sgBACES, are based on the mex version of the mat files.

### B. SYNTHETIC DATASET GENERATION

This section used the Simtb toolbox to generate a realistic fMRI dataset for eight different subjects. Twelve distinct temporal and spatial sources were used to obtain these eight datasets, consisting of 230 timepoints with a repetition time (TR = 2 secs) and  $60 \times 60$  voxels, respectively. Out of these twelve, five spatiotemporal sources were used for each subject’s data generation, where four sources were common to all subjects with some variability across subjects as shown in Fig. 3, whereas one of the sources was unique to each subject’s dataset. To elaborate, there were 12 sources, four common to all subjects, while each of the remaining eight sources is unique to each of the 8 subjects. Thus, temporal sources  $TC \in \mathbb{R}^{230 \times 5}$  and reshaped spatial sources  $SM \in \mathbb{R}^{5 \times 3600}$  for  $M = 8$  subjects have been created. For the common spatial maps, the intersubject variability has been introduced using parameters of the Gaussian distribution (mean ( $\omega$ ) and standard deviation (std) ( $\sigma$ )) via random translation in x and y direction ( $\omega = 0, \sigma = 0.6$ ), random rotation ( $\omega = 0, \sigma = 0.9$ ), and random scaling ( $\omega = 3, \sigma = 0.03$ ). Similarly, variability in temporal sources was added by varying the HRF parameters according to the suggestions given in the Simtb manual. The resulting common spatial and temporal sources are shown in Fig. 3b-e. The first subject’s common

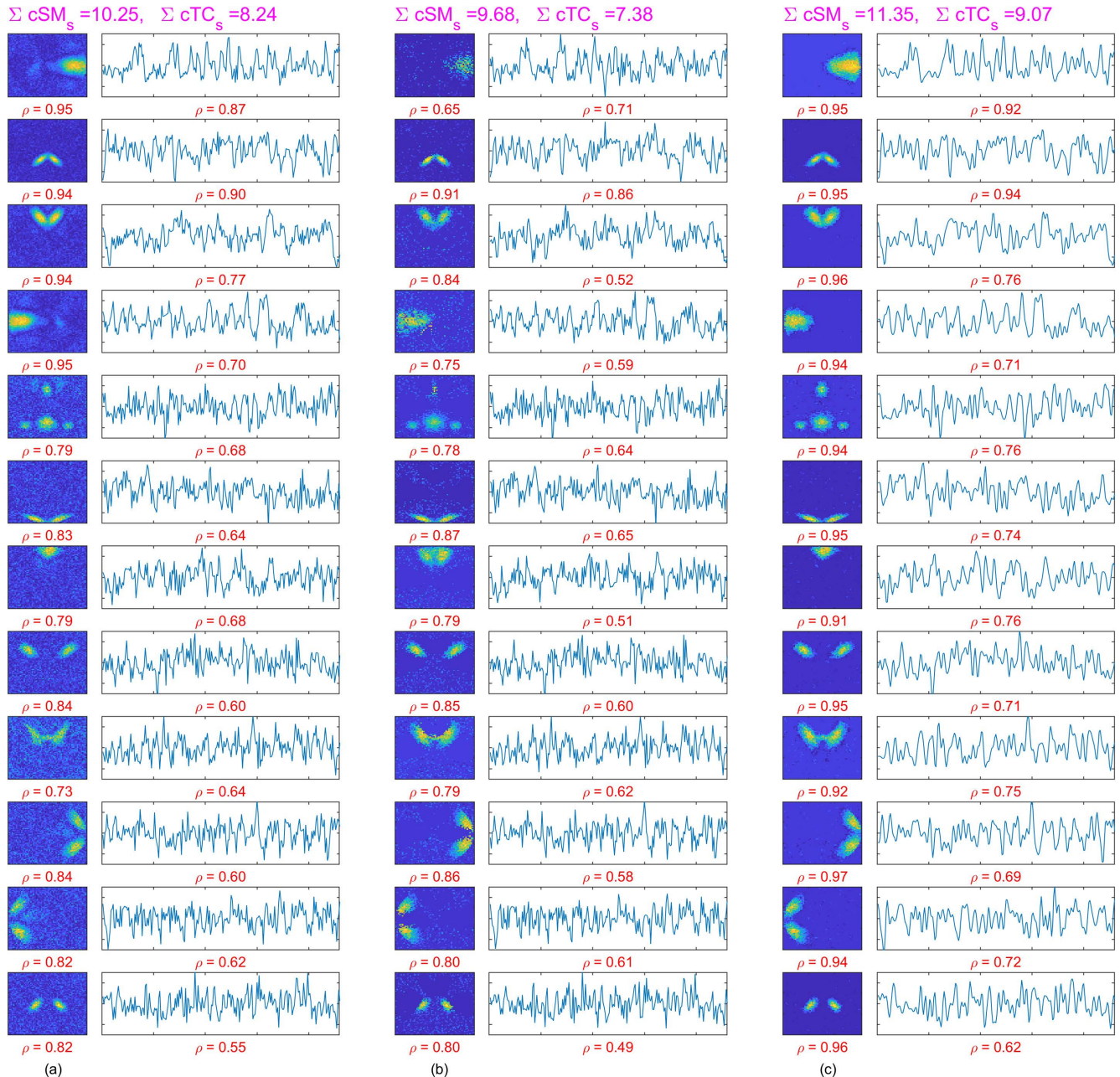
**TABLE 3.** Learning time (LT) on a ASUS RoG system with 48 GB 2400 MHz RAM, Intel i7 -7700 HQ 8 CPUs, and Nvidia GTX1070 display.

Average LT for synthetic data in seconds				
sICA	CODL	ShSSDL	rgBACESW	sgBACES
0.89	6.49	10.79	4.61	5.67
Total LT for 24 subject’s fMRI data in minutes				
sICA	CODL	ShSSDL	rgBACESW	sgBACES
11	43	55	115	68

and unique and all other unique spatiotemporal sources are shown in Fig. 3a, which are also treated as the ground truth TCs and SMs. By using these sources each subject’s dataset was generated using a linear mixture model given as  $Y_m = (TC_m + \Gamma_T)(SM_m + \Gamma_S)$ . Here random matrices  $\Gamma_T \in \mathbb{R}^{230 \times 5}$ , and  $\Gamma_S \in \mathbb{R}^{5 \times 3600}$  were produced using Gaussian distribution with ( $\omega = 0, \sigma = n_t$ ), and ( $\omega = 0, \sigma = n_s$ ), respectively, where  $n_t$  and  $n_s$  are the std values for temporal and spatial noise. The generated datasets  $\{Y_m\}_{m=1}^M$  were passed on to all participating algorithms for source retrieval.

### C. SYNTHETIC DATASET DICTIONARY LEARNING

For an unbiased comparison, the parameter settings were kept the same across all algorithms wherever possible. As the exact number of sources for real fMRI data is not known, instead of learning the same number of components as the number of generating sources, more components were allowed to be trained for the simulated dataset. In this regard, the total number of components to be learned was set to 16 for sICA, and CODL, and accordingly, the common dictionary size was set to  $K_c = 12$  and subject-specific to  $K_m = 4$  in the case of ShSSDL, rgBACESW, and sgBACES. All dictionary learning algorithms were iterated for 20 iterations, whereas CODL was iterated for 60 iterations. After testing the initialization scenario among data, random, and DCT bases for each algorithm individually, each algorithm’s best-performing initialization was provided. Concatenated data was used for CODL, random initialization (as recommended in their paper) was used for ShSSDL, and DCT bases were used for rgBACESW



**FIGURE 6.** The recovered TCs and SMs by three different algorithms, a) sICA, b) ShSSDL, and c) sgBACES, along with the temporal and spatial correlation values ( $\rho$ ) for each source, and the sum of these correlation values shown at the top.

and sgBACES dictionary initialization. Sixteen components were kept after both subject-wise and group-wise temporal reduction using PCA in the case of sICA. For a fair comparison with other dictionary learning algorithms, temporal reduction of the datasets was not performed for CODL, its batch size was set to  $b = 3600$ , and its sparsity parameter was set to  $\beta = 1.5$ . For ShSSDL, the sparsity parameter was set to  $\zeta_c = 2$  and  $\zeta_m = 1$ , internal iteration for sparse coding between  $X_c$  and  $X_m$  was set to 3, incoherence penalty was set to  $\eta = 2$ , and rest of its parameters that were provided with the publicly shared algorithm were left untouched. For

rgBACESW, the sparsity parameter was set to  $\zeta_c = 2$  and  $\zeta_m = 14$ , tolerance parameter for mixing matrix convergence was set to  $\epsilon = 10^{-5}$ , and number of DCT bases  $K_{pc}/K_{pm}$  were set to 120. In the case of sgBACES, the sparsity parameter was set to  $\zeta_c = 11$  and  $\zeta_m = 125$ , tuning parameter for adaptive penalization was set to  $\alpha = 1.5$ , Lagrangian multiplier was set to  $\mu = 3$ , sparsity parameter for sparse basis expansion was set to  $\lambda = 105$ , and number of DCT bases  $K_{pc}/K_{pm}$  were set to 120. Although,  $\alpha = 5$  produced better performance, the results showed more deviation from the mean than those obtained with  $\alpha = 1.5$ . Multiple parameter

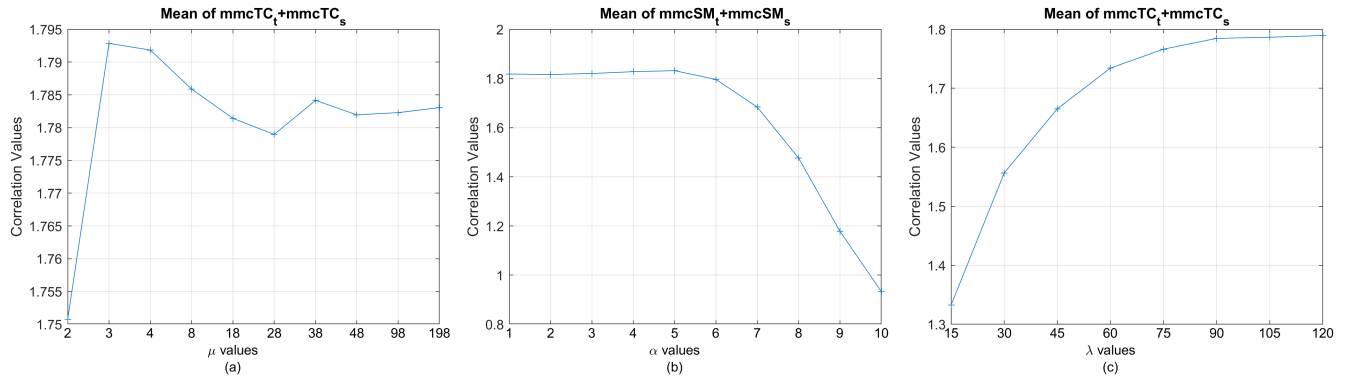


FIGURE 7. The effect of hyperparameters a)  $\mu$ , b)  $\alpha$ , and c)  $\lambda$  on the performance of sgBACES.

values were tried for all these tuning parameters, and the ones that produced the best performance in terms of correlation value between the retrieved and the ground truth TCs and SMs were kept. The tested values of tuning parameters for all algorithms are given in Table 2 for the dataset used in this section, where the colon operator indicates a Matlab increment by the given factor.

**D. SYNTHETIC DATASET RESULTS**

To establish that the proposed algorithm is robust and consistent, multiple multi-subject datasets (each dataset consisting of 8 subjects) of different noise intensities were generated. This was realized by varying the std values of spatial and temporal noise i.e.  $n_t = \{0, 0.3, 0.6, 0.9\}$  and  $n_s = \{0.06, 0.12, 0.18\}$ , and using their different combinations. Twelve different combinations of these std values generated 12 different noise instances. In addition, the learning process was repeated 100 times for each of the 12 noise instances. Recovered TCs/SMs were obtained by correlating every algorithm’s trained dictionary/sparse code pair with the ground truth (GT) TCs/SMs pair and using the highest absolute correlation coefficient values indices. The corresponding highest correlation values were also saved as  $cTC/cSM$ . Specifically, the correlation values were computed with respect to GT-TCs or GT-SMs, thus producing two different recovery schemes. The first recovery scheme computes  $cTC_t/cSM_t$  that consists of finding those recovered TCs that are maximally correlated with the GT-TCs and then obtaining their corresponding recovered SMs, which is generally the convention in real fMRI data analysis. Whereas the second scheme computes  $cTC_s/cSM_s$  that are obtained by finding those recovered SMs that are maximally correlated with the GT-SMs and then obtaining their corresponding temporal components, which in regards to real fMRI data analysis requires the knowledge of functional brain networks that can be obtained from brain atlases [34].

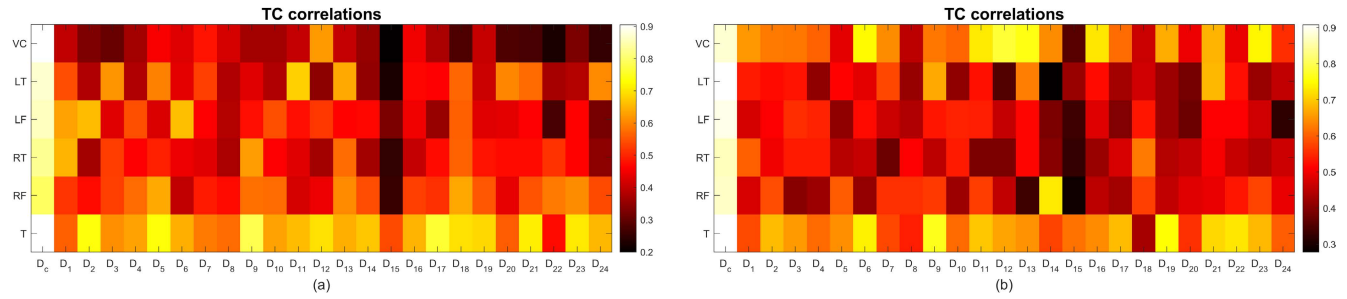
Moreover, for sICA and CODL, correlation values were computed between GT-TCs/GT-SMs and the group-level retrieved components. In the case of other algorithms, common GT-TCs and GT-SMs were correlated with

common dictionary/sparse code, and unique GT-TCs and GT-SMs were correlated with subject-specific dictionary/sparse code. For each trial, the mean of the  $cTC_t/cTC_s$  values ( $mcTC_t/mcTC_s$ ) and the mean of the  $cSM_t/cSM_s$  values ( $mcSM_t/mcSM_s$ ) over all sources were saved and their mean over all trials is plotted in Fig. 4 for all 12 noise instances. For all algorithms, the std values of  $mcTC_t/mcTC_s/mcSM_t/mcSM_s$  over all trials was computed and the mean of these std values over all noise instances is shown in Fig. 5a. The sum of the highest correlation between atoms of each dictionary ( $D_c-D_9$ ) and GT-TCs ( $t_1-t_{12}$ ) and the highest correlation between a coefficient row of each sparse code matrix ( $X_c-X_9$ ) and GT-SMs ( $s_1-s_{12}$ ) are saved, and their mean is taken over all trials and all noise instances to produce a correlation matrix in Fig. 5b and 5c for two different algorithms.

From Fig. 4, it can be concluded that sgBACES performed consistently better in terms of recovered TCs correlation with the GT-TCs over all trials and appeared less sensitive to high noise levels as it held to the highest correlation values almost over all noise instances and for both recovery schemes, although higher TC correlations are observed for the second recovery scheme. In contrast, SM correlations for sgBACES were consistently higher for all noise instances and both schemes, whereas, sICA was a runner-up in the first recovery scheme and rgBACESW emerged as a runner-up in the second scheme.

Fig. 5a could have been presented differently by using error bars to show the deviations from the mean correlation values across trails, however, visually, it did not lead to any conclusive results. Therefore, the mean values were plotted that revealed sICA across trials is more consistent with its source recovery results than dictionary learning algorithms. Among DL algorithms, however, sgBACES has shown the lowest variance.

As described earlier in this section, Fig. 5b shows the correlation matrix between the ground truth TCs and SMs ( $t_s$ ) and the trained atoms and sparse code ( $DX$ ), respectively, by ShSSDL. Here it can be seen that the most correlated components with first four common sources ( $ts_1, ts_2, ts_3, ts_4$ ) were



**FIGURE 8.** Correlation matrix between the MHRs and the trained dictionaries by a) ShSSDL, and b) sgBACES.

found in common dictionary/sparse code ( $DX_c$ ) whereas the most correlated components with rest of the unique sources ( $ts_5$ - $ts_{12}$ ) were found in subject-specific dictionary/sparse code ( $DX_1$ - $DX_8$ ). This phenomenon is more highlighted in the case of the proposed algorithm as shown in Fig. 5c, where common dictionary/sparse code seems to be more uncorrelated with the subject-specific dictionaries/sparse codes than Fig. 5b.

For component-wise visual inspection, a dataset was generated using  $n_t = 1.2$ ,  $n_s = 0.12$ , and Matlab's seed generator  $\text{rng}(1, \text{'philox'})$ . Both recovered TCs, and unthresholded SMs, and their correlation values are shown in Fig. 6 for sICA, ShSSDL, and sgBACES. The sum of  $cSM_s$  and  $cTC_s$  is shown at the top for each of these three algorithms, where the proposed algorithm, whose results are given in column (c), has performed better in terms of SMs and TCs correlation with the ground truth. The recovered spatiotemporal dynamics by the proposed algorithm are less noisy while, at the same time, TCs are smooth, and SMs are more specific compared to the other two algorithms. The results from CODL and rgBACESW are not shown due to lack of space. It is also noteworthy that the best value of  $\alpha = 5$  was selected for sgBACES in this case. Furthermore, for all algorithms, the average computation time in seconds over all trials and noise instances is given in Table 3.

### E. EFFECT OF HYPERPARAMETERS

This section discusses the impact that sgBACES' tuning parameters  $\mu$ ,  $\alpha$ , and  $\lambda$  have on dictionary performance. The effect of these parameters has been highlighted in Fig. 7, wherein all three sub-figures, the results are based on 5 trials over all 12 noise instances, and for different values of  $\mu$ ,  $\alpha$ , and  $\lambda$ . The mean of  $mcTC_t/mcTC_s/mcSM_t/mcSM_s$  values over 5 trials is saved as  $mmcTC_t/mmcTC_s/mmcSM_t/mmcSM_s$ . In Fig. 7a the mean of  $mmcTC_t + mmcTC_s$  over all noise instances is plotted against different values of  $\mu$  when values of  $\alpha$  and  $\lambda$  were fixed at 5 and 105, respectively. It is obvious here that the autocorrelation of recovered TCs is maximized when  $\mu = 3$ . Similarly, mean of  $mmcSM_t + mmcSM_s$  over all noise instances is plotted against different values of  $\alpha$  in Fig. 7b when  $\mu$  is fixed at its best value 3, and  $\lambda = 105$ . In this case, the

performance in terms of correlation increases slightly with the increasing values of  $\alpha$  before it falls when  $\alpha$  exceeds 5. In contrast, at  $\mu = 3$ , and  $\alpha = 5$ , the performance of sgBACES in terms of mean of  $mmcTC_t + mmcTC_s$  continues to rise dramatically as more number of bases are incorporated until their effect becomes negligible after  $\lambda = 90$  as shown in Fig. 7c. Although, best correlation value is attained at  $\lambda = 120$ , this would lead to regular basis expansion and a case of overfitting. A phenomenon where reconstructed time sources attempt to capture all temporal variability in the training data resulting in low SNR and lower autocorrelations at all lags.

### F. MULTI-SUBJECT fMRI DATASET

In order to recover common and individual spatiotemporal dynamics from multi-subject fMRI data, an unprocessed motor task dataset of 24 randomly selected subjects was obtained from the Q3 release of the HCP [59]. Functional MR images of all 24 datasets were acquired using Siemens 3 Tesla scanner equipped for gradient echo planar imaging (EPI) with parameters: TR = 0.72 s, echo time (TE) = 33.1 ms, field of view (FOV) = 208 × 180 mm, flip angle (FA) = 52°, matrix size = 104 × 90, slice thickness = 2 mm with 72 contiguous slices, and 2 mm isotropic voxels, echo spacing = 0.58 ms, BW = 2290 Hz/Px, and 284 EPI volumes were collected for each scan. More details on this data acquisition can be found in [59].

This task was motivated by the experiments in [62] during which visual cues were presented to the participants asking them to tap their right or left fingers, or squeeze their right or left toes, or move their tongue in order to map motor areas of the brain. In each of the two runs, all subjects were presented a 3 second visual cue followed by one of the 10 movement types (hand movement 2 left 2 right, foot movement 2 left 2 right, and 2 tongue movements) that lasted for 12 seconds. Besides, there were 3 fixation blocks per run that lasted 15 seconds each. Six different modeled HRF (MHR) were generated by convolving the six task stimuli related to the aforementioned movement types (left toe (LT), left finger (LF), right toe (RT), right finger (RF), tongue (T), visual cue (VC)) with canonical HRF to obtain GT-TCs.

**TABLE 4.** Table of tested values for parameter selection for fMRI data.

Algorithm	Tested Values
CODL	$b = \{2, 5, 10, 236.115\} \times 1000, \beta = \{1 : 1 : 10\}$
ShSSDL	$\zeta_c = \{1, 2, 3, 4, 5\}, \zeta_m = \{1, 2, 3, 4, 5\}, \eta = \{0.5, 1, 5, 20, 100, 200, 500, 1000\}$
rgBACESW	$\zeta_c = \{25 : 25 : 100\}, \zeta_m = \{25 : 25 : 300\}, K_{pc}/K_{pm} = \{60, 90, 120, 150\}$
sgBACES	$\zeta_c = \{25 : 25 : 100\}, \zeta_m = \{25 : 25 : 300\}, \alpha = \{0.5 : 0.5 : 4\}, \mu = \{1, 2, 3, 4, 8, 18, 28, 38, 48, 98, 198, 398\}, \lambda = \{30, 40, 50, 60, 90, 120\}, K_{pc}/K_{pm} = \{60, 90, 120, 150\}$

### G. fMRI DATASET PREPROCESSING

The image pre-processing of every subject's fMRI dataset was carried out in Matlab 2021b using SPM-12 [4] whose pre-processing pipeline consisted of realignment, normalization, spatial smoothing, masking, and temporal filtering. All functional images were realigned to the first image in order to correct for any head movements that may have occurred during the course of the experiment. In the next step, all images were spatially normalized to a standard Talarach template, resampled to  $2 \times 2 \times 2 \text{ mm}^3$  voxels, and spatially smoothed using a  $6 \times 6 \times 6 \text{ mm}^3$  full-width at half-maximum (FWHM) Gaussian kernel. Images were masked to remove any data outside the scalp, and only those voxels were kept, which exceeded a masking threshold. The 4-dimensional datasets collected from the masked images were rearranged and stored into a 2-dimensional matrix  $Y_m$  to be used as a whole brain's dataset for the  $m$ -th subject. After carrying out the steps mentioned above, the size of  $Y$  for each subject was found to be  $284 \times 236115$ . From all datasets, low-frequency trends were removed using a DCT basis set with a cutoff of 1/150 Hz, and high-frequency fluctuations were removed using FWHM of 2 s. Each column of  $Y$  obtained after pre-processing was normalized to have zero mean and unit variance for all subjects.

### H. fMRI DATASET DICTIONARY LEARNING

After trying different numbers between 50 and 150, the total number of components to be learned was set to 120 for sICA, whereas, 70 for CODL, and accordingly  $40 + 30$  in the case of ShSSDL, rgBACESW, and sgBACES, where  $K_c = 40$  is the common, and  $K_m = 30$  is the subject-specific dictionary size. All dictionary learning algorithms were iterated for 15 iterations, whereas CODL was iterated for 30 iterations. For dictionary initialization, the same approach as in the case of the simulated dataset was followed. Concatenated data was used for CODL, the common dictionary was initialized from the first subject's dataset, and subject-specific dictionaries were initialized from their respective datasets (as recommended in their paper) for ShSSDL, and DCT bases were used for rgBACESW and sgBACES dictionary initialization. For sICA, 120 components were kept after both subject-wise, and group-wise temporal reduction had been performed using PCA. For a fair comparison with other dictionary learning algorithms, temporal reduction of the datasets was not

performed in the case of CODL, and its batch size was set to  $b = 236115$ , and its sparsity parameter was set to  $\beta = 6$ . For ShSSDL, the sparsity parameter was set to  $\zeta_c = 1$  and  $\zeta_m = 2$ , internal iteration was set to 5, incoherence penalty was set to  $\eta = 100$ , and the rest of the parameters that were provided with the algorithm were left untouched. For rgBACESW, the sparsity parameter was set to  $\zeta_c = 25$  and  $\zeta_m = 100$ , tolerance parameter was set to  $\epsilon = 10^{-5}$ , and number of DCT bases  $K_{pc}/K_{pm}$  were set to 60. For sgBACES, the sparsity parameter was set to  $\zeta_c = 75$  and  $\zeta_m = 250$ , tuning parameter for adaptive penalization was set to  $\alpha = 2$ , Lagrangian multiplier was set to  $\mu = 3$ , sparsity parameter for sparse basis expansion was set to  $\lambda = 50$ , and number of DCT bases  $K_{pc}/K_{pm}$  were set to 60. Multiple values were tried for all these tuning parameters, and the ones that produced the best performance in terms of correlation value between the retrieved and the ground truth TCs/SMs were kept. The tested values of tuning parameters for all algorithms are given in Table 4.

### I. fMRI DATASET RESULTS

In the absence of ground truth for SMs, the first recovery approach was adopted by considering the six MHRs. According to the initial model (3), the common spatiotemporal dynamics are expected to be contained in  $D_c/X_c$  and subject-specific dynamics in subject-level dictionary/sparse codes i.e.  $D_1/X_1 - D_{24}/X_{24}$ . Therefore, it was anticipated to find the components that are most correlated to six MHRs in the common dictionary  $D_c$  and its corresponding spatial maps in  $X_c$ . This was confirmed by correlating the six MHRs with atoms from dictionary  $D_0$  as well as subject-specific dictionaries  $D_1 - D_{24}$  and saving the highest correlation values as shown in Fig. 8. This figure shows that the most correlated atoms with MHRs are found in  $D_c$  for both a) ShSSDL and b) sgBACES. It can be further observed from this correlation matrix that even by selecting the best incoherence value for ShSSDL, the tongue component was still found in subject-specific dictionaries, whereas its leakage to other dictionaries was to a lesser extent in the case of sgBACES. However, some sgBACES's components that were substantially correlated with visual cue MHR were found in subject-specific dictionaries, yet it cannot be considered a complication because sgBACES did not impose the strict condition of incoherence among dictionaries.

As the first recovery approach was adopted, the recovered TCs that were most correlated with MHRs were considered, and their values were saved. In the case of sICA and CODL, MHRs were correlated with the group-level retrieved TCs, whereas in the case of the other three algorithms, the atoms from the common dictionary were correlated with MHRs. These correlation values are specified in Table 5 in rows categorized as cTC, and their corresponding spatial maps are shown in Fig. 9. This figure shows that sgBACES produced activation maps that are very specific and localized to the motor and visual area of the brain compared to maps recovered by the other three algorithms. Here, the SMs from



**FIGURE 9.** Recovered spatial maps thresholded at  $p < 0.001$  for the six MHRs by a) sICA, b) CODL, c) ShSSDL, and d) sgBACES. All spatial maps were converted to Z-score, then absolute values of the Z-score were used to calculate the right-tailed p-values. Using these p-values, original spatial maps were thresholded at  $p < 0.001$ , and their absolute value is plotted here.

rgBACESW are not shown due to lack of space. These results are comparable to those reported in other papers for the same dataset [36], [41]. Also, from Table 5 the highest correlation values of recovered TCs with MHRs were observed for sgBACES.

A similar strategy is adopted to analyze subject-specific dictionaries as presented in [42]. They found ten well-established resting-state networks (R1-R10) [63] that may be present in every subject’s subject-specific sparse code matrices irrespective of the similarity of the corresponding temporal dynamics. By considering the resting state network (RSN) templates from [63] all subject-specific sparse codes  $X_1$ - $X_{24}$  were correlated with them, and the most correlated RSNs were saved. By taking an absolute mean of these most correlated RSNs from every subject, ten averaged SMs (R1-R10) were generated, which are presented in Fig. 10. The ten averaged SMs were then correlated with the RSN templates, and the most correlated values were saved and given in Table 5 in the rows highlighted by cSM. This table shows that the SMs recovered by sgBACES produced the highest correlation with the RSN templates.

Overall, the mean of all correlation values between the recovered spatiotemporal dynamics and the ground truth is highest for sgBACES. It is also noticeable from Fig. 10 that

all spatial maps retrieved by sgBACES show better similarity with RSN templates, specifically the default mode network (R4), where all its activated areas are more prominent than those recovered by other algorithms.

Furthermore, the convergence rate of dictionary learning as a function of algorithm iterations is presented in Fig. 11a, which shows all algorithms converged around the fifteenth iteration with some fluctuations exhibited by ShSSDL and rgBACESW. Over the course of learning, the six values obtained by keeping the maximum correlation values between the recovered TCs from the common dictionary and MHRs ( $cTC_c$ ) were summed and plotted in Fig. 11b as a function of algorithm iterations. In this plot, sgBACES is shown to perform consistently better than all other algorithms except for some initial iterations when sICA’s correlation is superior. Similarly, the ten values obtained by keeping the maximum correlation values between the averaged recovered SMs from subject-specific sparse code matrices, and RSN templates ( $cSM_s$ ) were summed and plotted in Fig. 11c as a function of algorithm iterations. Once more, sgBACES performed consistently better than the rest of the algorithms, other than some initial iterations.

In Fig. 12a, the effect of taking the atom’s autocorrelation into account is shown as a function of  $\mu$ . Here, the mean of



FIGURE 10. Thresholded spatial maps at  $p < 0.001$  obtained from a) RSN templates, b) CODL, c) ShSSDL, and d) sgBACES.

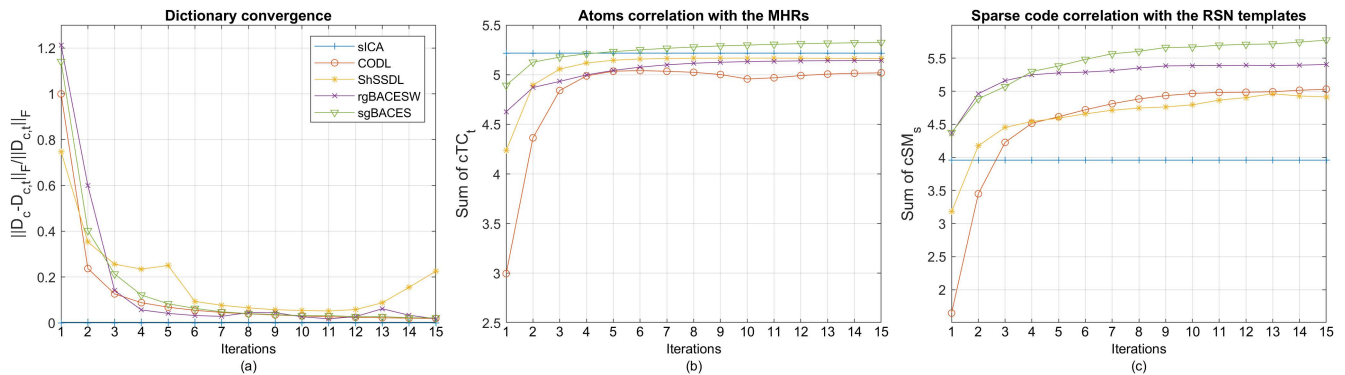


FIGURE 11. As a function of algorithm iterations, a) dictionary convergence, b) sum of highest correlations between common dictionary atoms and MHRs, and c) sum of highest correlations between averaged subject-specific sparse codes and RSN templates.

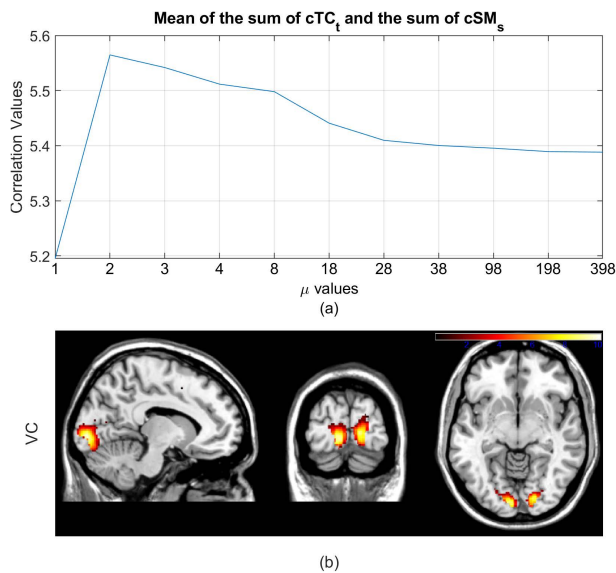
the sum of highest correlations between common dictionary atoms and MHRs and the sum of highest correlations between averaged subject-specific sparse codes and RSN templates

is plotted against the different values of  $\mu$ . It shows that, at  $\mu = 2$ , the maximum autocorrelation of atoms is achieved. However,  $\mu = 3$  was selected because cTC values for all



**TABLE 5.** cTC) Highest absolute correlation values between the MHRs and atoms, a) from the group-level dictionary in the case of sICA and CODL, and b) from the common dictionary in the case of the other three algorithms. cSM) Highest absolute correlation values between the RSN templates and a) SMs from the group-level coefficient matrix in the case of sICA and CODL, b) averaged RSNs obtained using the subject-specific SC of the other three algorithms.

Correlation type	Ground truth	Algorithms				
		a) sICA	a) CODL	b) ShSSDL	b) rgBACESW	b) sgBACES
cTC	VC	0.8425	0.8556	<b>0.9042</b>	0.9030	0.8800
	LT	0.8876	0.8394	0.8717	0.8656	<b>0.9080</b>
	LF	0.8904	0.8510	0.8601	0.8631	<b>0.8924</b>
	RT	0.8357	0.7562	0.8292	0.7884	<b>0.8651</b>
	RF	0.8590	0.8248	0.7928	0.8271	<b>0.8722</b>
	T	0.9025	0.8934	0.9056	0.8971	<b>0.9077</b>
cSM	R1	0.5782	0.6489	0.6244	0.7544	<b>0.7678</b>
	R2	0.6411	<b>0.6850</b>	0.5622	0.6658	0.6304
	R3	0.5372	0.5133	0.5572	0.5667	<b>0.6010</b>
	R4	0.3362	<b>0.6727</b>	0.6175	0.5968	0.6151
	R5	0.2756	0.3420	0.3504	0.4035	<b>0.4219</b>
	R6	0.2495	0.3063	0.4265	0.4578	<b>0.4907</b>
	R7	0.3498	0.4657	0.4196	0.4502	<b>0.5619</b>
	R8	0.2865	0.4342	0.3438	0.4428	<b>0.5252</b>
	R9	0.3913	0.4512	0.4627	0.4928	<b>0.5225</b>
	R10	0.3135	0.5238	0.5504	0.5733	<b>0.6370</b>
Mean		0.5735	0.6290	0.6296	0.6593	<b>0.6937</b>



**FIGURE 12.** a) Effect of  $\mu$  values on the mean of the sum of TC and SM correlations, b) Spatial map recovered by rgBACESW thresholded at  $p < 0.001$  for visual cue MHR.

6 MHRs were found to be above 0.80 at this particular value of  $\mu$ . After trying many different combinations of sparsity parameters for rgBACESW, in addition to what is mentioned in Table 4, the best spatial map recovered by rgBACESW for visual cue MHR is shown in Fig. 12b. It is clear from this figure that rgBACESW's spatial map is not comparable to other algorithms.

The computation time consumed by all algorithms to process 24 whole-brain fMRI datasets to learn group-level dynamics is given in Table 3.

## VI. CONCLUSION

A new dictionary learning algorithm for multi-subject fMRI data analysis is presented in this paper, which outperformed spatial ICA and existing MS dictionary learning algorithms. The effectiveness of sgBACES is demonstrated using synthetic and motor-task based fMRI datasets. Its performance was found to be consistent across trials and datasets, and robust against high noise levels. Although, it is slightly expensive in terms of computation time compared to CODL and ShSSDL, it produced spatial maps and temporal dynamics that were less noisy and closer to the ground truth. Instead of the columns, it enforced sparsity on the rows of the sparse code matrix, and column sparsity was still found to be recoverable. Thus, it can simultaneously learn common and distinct neural activity without requiring a separate sparse code stage. The improvements exhibited by the sgBACES algorithm for MS-fMRI analysis can be attributed mainly to i) the incorporation of fMRI data's prior information in the minimization problem, ii) the estimation of dictionary/sparse code as a pair, and iii) utilization of the DCT bases through sparse mixing matrix.

It is essential to highlight some concerns regarding the proposed MS-DL algorithm, such as tuning parameter selection, memory limitations, and numerical cost. The memory constraints limited the number of subjects I could incorporate in the experimental fMRI study. The analysis on the ASUS RoG system with 48 GB RAM could only be performed on 24 subjects while 95% memory was consumed. In contrast, there was no bound in the simulation study, and many more subjects than just 8 could have been considered, but this would have resulted in increased simulation cost in terms of time, because the experiments had to be repeated 100 times over 12 noise instances. The selection of seven tuning parameters

can be a formidable task, but some guidance and direction have been provided in this paper with the help of a simulation study and experimental fMRI data.

The computational cost of the proposed algorithm can be further reduced in the future by considering online learning that is scalable in both matrix dimensions [64]. One can also consider using a correlation-based DL algorithm with a weighted least square approach that during each iteration updates only a subset of dictionary atoms resulting in a computationally efficient online learning [65].

## ACKNOWLEDGMENT

The fMRI data was provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

## REFERENCES

- [1] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proc. Nat. Acad. Sci. USA*, vol. 87, no. 24, pp. 9868–9872, 1990.
- [2] K. J. Friston, "Modalities, modes, and models in functional neuroimaging," *Science*, vol. 326, no. 5951, pp. 399–403, Oct. 2009.
- [3] K. J. Friston, P. Jezzard, and R. Turner, "Analysis of functional MRI time-series," *Hum. Brain Mapping*, vol. 1, no. 2, pp. 153–171, 1994.
- [4] W. Penny, K. Friston, J. Ashburner, S. Kiebel, and T. Nichols, *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Cambridge, MA, USA: Academic Press, 2006.
- [5] M. A. Lindquist, "The statistical analysis of fMRI data," *Statist. Sci.*, vol. 23, no. 4, pp. 439–464, Nov. 2008.
- [6] G. K. Aguirre, E. Zarahn, and M. D'Esposito, "The variability of human, BOLD hemodynamic responses," *NeuroImage*, vol. 8, no. 4, pp. 360–369, Nov. 1998.
- [7] M. J. McKeown, T.-P. Jung, S. Makeig, G. Brown, S. S. Kindermann, T.-W. Lee, and T. J. Sejnowski, "Spatially independent activity patterns in functional MRI data during the stroop color-naming task," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 3, pp. 803–810, Feb. 1998.
- [8] A. H. Andersen, D. M. Gash, and M. J. Avison, "Principal component analysis of the dynamic response measured by fMRI: A generalized linear systems framework," *Magn. Reson. Imag.*, vol. 17, no. 6, pp. 795–815, Jul. 1999.
- [9] O. Friman, M. Borga, P. Lundberg, and H. Knutsson, "Exploratory fMRI analysis by autocorrelation maximization," *NeuroImage*, vol. 16, no. 2, pp. 454–464, 2002.
- [10] O. Friman, M. Borga, P. Lundberg, and H. Knutsson, "Detection and detrending in fMRI data analysis," *NeuroImage*, vol. 22, no. 2, pp. 645–655, Jun. 2004.
- [11] K. Worsley, J.-I. Chen, J. Lerch, and A. C. Evans, "Comparing connectivity via thresholding correlations and SVD," *Philos. Trans. Roy. Soc. B Biol. Sci.*, vol. 360, pp. 913–920, May 2005.
- [12] Y. Zhong, H. Wang, G. Lu, Z. Zhang, Q. Jiao, and Y. Liu, "Detecting functional connectivity in fMRI using PCA and regression analysis," *Brain Topography*, vol. 22, no. 2, pp. 134–144, Sep. 2009.
- [13] M. U. Khalid and A.-K. Seghouane, "Improving functional connectivity detection in FMRI by combining sparse dictionary learning and canonical correlation analysis," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, Apr. 2013, pp. 286–289.
- [14] M. U. Khalid, A. Shah, and A.-K. Seghouane, "Sparse dictionary learning for fMRI analysis using autocorrelation maximization," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 4286–4289.
- [15] M. A. Qadar and A.-K. Seghouane, "A projection CCA method for effective fMRI data analysis," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 11, pp. 3247–3256, Nov. 2019.
- [16] V. Calhoun, T. Adali, G. Pearson, and J. Pekar, "A method for making group inferences from fMRI data using independent component analysis," *Hum. Brain Mapping*, vol. 14, no. 3, pp. 51–140, 2001.
- [17] 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.
- [18] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D'ardenne, W. Richter, J. D. Cohen, and J. Haxby, "Independent component analysis for brain fMRI does not select for independence," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 26, pp. 10415–10422, 2009.
- [19] V. D. Calhoun, V. K. Potluru, R. Phlypo, R. F. Silva, B. A. Pearlmuter, A. Caprihan, S. M. Plis, and T. Adali, "Independent component analysis for brain fMRI does indeed select for maximal independence," *PLoS ONE*, vol. 8, no. 10, pp. 1–8, 2013.
- [20] W. Zhang, J. Lv, X. Li, D. Zhu, X. Jiang, S. Zhang, Y. Zhao, L. Guo, J. Ye, D. Hu, and T. Liu, "Experimental comparisons of sparse dictionary learning and independent component analysis for brain network inference from fMRI data," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 1, pp. 289–299, Jan. 2019.
- [21] R. Jin, K. K. Dontaraju, S.-J. Kim, M. A. B. S. Akhonda, and T. Adali, "Dictionary learning-based fMRI data analysis for capturing common and individual neural activation maps," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1265–1279, Oct. 2020.
- [22] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, Jun. 1996.
- [23] M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI data by blind separation into independent spatial components," *Hum. Brain Mapping*, vol. 6, no. 3, pp. 160–188, 1998.
- [24] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2015.
- [25] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.
- [26] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 689–696.
- [27] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces* (Lecture Notes in Computer Science), vol. 6920. Berlin, Germany: Springer, 2012, pp. 711–730.
- [28] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.
- [29] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.
- [30] K. Lee, S. Tak, and J. C. Ye, "A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion," *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1076–1089, May 2011.
- [31] A.-K. Seghouane and A. Iqbal, "Basis expansion approaches for regularized sequential dictionary learning algorithms with enforced sparsity for fMRI data analysis," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1796–1807, Sep. 2017.
- [32] A.-K. Seghouane and A. Iqbal, "Consistent adaptive sequential dictionary learning," *Signal Process.*, vol. 153, pp. 300–310, Dec. 2018.
- [33] A. Iqbal, M. Nait-Meziane, A.-K. Seghouane, and K. Abed-Meraim, "Adaptive complex-valued dictionary learning: Application to fMRI data analysis," *Signal Process.*, vol. 166, Jan. 2020, Art. no. 107263.
- [34] M. Morante, Y. Kopsinis, S. Theodoridis, and A. Protopapas, "Information assisted dictionary learning for fMRI data analysis," *IEEE Access*, vol. 8, pp. 90052–90068, 2020.
- [35] J. Lv, X. Jiang, X. Li, D. Zhu, H. Chen, T. Zhang, S. Zhang, X. Hu, J. Han, H. Huang, and J. Zhang, "Sparse representation of whole-brain fMRI signals for identification of functional networks," *Med. Image Anal.*, vol. 20, no. 1, pp. 112–134, 2014.
- [36] S. Zhao, J. Han, J. Lv, X. Jiang, X. Hu, Y. Zhao, B. Ge, L. Guo, and T. Liu, "Supervised dictionary learning for inferring concurrent brain networks," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 2036–2045, Oct. 2015.
- [37] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion, "Multi-subject dictionary learning to segment an atlas of brain spontaneous activity," in *Proc. Inf. Process. Med. Imag.*, vol. 22, 2011, pp. 562–573.

- [38] M. U. Khalid and A.-K. Seghouane, "Multi-subject fMRI connectivity analysis using sparse dictionary learning and multiset canonical correlation analysis," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2015, pp. 683–686.
- [39] J. Lv, B. Lin, Q. Li, W. Zhang, Y. Zhao, X. Jiang, L. Guo, J. Han, X. Hu, C. Guo, and J. Ye, "Task fMRI data analysis based on supervised stochastic coordinate coding," *Med. Image Anal.*, vol. 38, pp. 1–16, May 2017.
- [40] A. Mensch, G. Varoquaux, and B. Thirion, "Compressed online dictionary learning for fast resting-state fMRI decomposition," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1282–1285.
- [41] A. Iqbal and A.-K. Seghouane, "A dictionary learning algorithm for multi-subject fMRI analysis based on a hybrid concatenation scheme," *Digit. Signal Process.*, vol. 83, pp. 249–260, Dec. 2018.
- [42] A. Iqbal, A.-K. Seghouane, and T. Adali, "Shared and subject-specific dictionary learning (ShSSDL) algorithm for multisubject fMRI data analysis," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 11, pp. 2519–2528, Nov. 2018.
- [43] T. H. Vu and V. Monga, "Fast low-rank shared dictionary learning for image classification," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5160–5175, Nov. 2017.
- [44] Y. Han, Q.-H. Lin, L.-D. Kuang, X.-F. Gong, F. Cong, Y.-P. Wang, and V. D. Calhoun, "Low-rank tucker-2 model for multi-subject fMRI data decomposition with spatial sparsity constraint," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 667–679, Mar. 2022.
- [45] R. E. Kelly, G. S. Alexopoulos, Z. Wang, F. M. Gunning, C. F. Murphy, S. S. Morimoto, D. Kanellopoulos, Z. Jia, K. O. Lim, and M. J. Hoptman, "Visual inspection of independent components: Defining a procedure for artifact removal from fMRI data," *J. Neurosci. Methods*, vol. 189, no. 2, pp. 233–245, Jun. 2010.
- [46] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 4311–4322, Nov. 2006.
- [47] C. Bao, H. Ji, Y. Quan, and Z. Shen, "Dictionary learning for sparse coding: Algorithms and convergence analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1356–1369, Jul. 2016.
- [48] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Ann. Appl. Statist.*, vol. 1, no. 2, pp. 302–332, Dec. 2007.
- [49] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Conf. Rec. 27th Asilomar Conf. Signals, Syst. Comput.*, vol. 1, Nov. 1993, pp. 40–44.
- [50] K. Egan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, Mar. 1999, pp. 2443–2446.
- [51] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, and S. D. Penna, "The human connectome project: A data acquisition perspective," *Neuroimage*, vol. 62, no. 4, pp. 2222–2231, 2012. [Online]. Available: <https://www.humanconnectome.org/>
- [52] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1553–1564, Mar. 2010.
- [53] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
- [54] M. Nikolova, "Relationship between the optimal solutions of least squares regularized with  $\ell_0$ -norm and constrained by k-sparsity," *Appl. Comput. Harmon. Anal.*, vol. 41, no. 1, pp. 237–265, Jul. 2016.
- [55] W. J. Fu, "Penalized regressions: The bridge versus the lasso," *J. Comput. Graph. Statist.*, vol. 7, no. 3, pp. 397–416, 1998.
- [56] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Hoboken, NJ, USA: Prentice-Hall, 1997.
- [57] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms," *J. Fourier Anal. Appl.*, vol. 14, nos. 5–6, pp. 655–687, Dec. 2008.
- [58] E. B. Erhard, E. A. Allen, Y. Wei, T. Eichele, and V. D. Calhoun, "SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability," *NeuroImage*, vol. 59, no. 4, pp. 4160–4167, 2012.
- [59] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, and D. Nolan, "Function in the human connectome: Task-fMRI and individual differences in behavior," *NeuroImage*, vol. 80, pp. 169–189, Oct. 2013.
- [60] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, Jun. 2000.
- [61] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," Dept. Comput. Sci., Univ. Haifa, Israel, Tech. Rep., 2008.
- [62] R. L. Buckner, F. M. Krienen, A. Castellanos, J. C. Diaz, and B. T. Yeo, "The organization of the human cerebellum estimated by intrinsic functional connectivity," *J. Neurophysiol.*, vol. 106, no. 5, pp. 2322–2345, 2011.
- [63] S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, and C. F. Beckmann, "Correspondence of the brain's functional architecture during activation and rest," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 31, pp. 13040–13045, Aug. 2009.
- [64] A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux, "Dictionary learning for massive matrix factorization," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1737–1746.
- [65] Y. Naderahmadian, S. Beheshti, and M. A. Tinati, "Correlation based online dictionary learning algorithm," *IEEE Trans. Signal Process.*, vol. 64, no. 3, pp. 592–602, Feb. 2016.



**MUHAMMAD USMAN KHALID** received the M.Sc. degree in electrical engineering from the Royal Institute of Technology, Stockholm, in 2010, and the Ph.D. degree in engineering and computer science from Australian National University, Canberra, in 2015. He has previously worked with the Department of Electrical and Electronic Engineering, The University of Melbourne, where he is currently a Research Fellow with the School of Mathematics and Statistics. His research interests include sparse representation, signal processing, and image processing.

...