# Shared and Subject-Specific Dictionary Learning Algorithm for Multi-Subject fMRI Data Analysis (ShSSDL)

Asif Iqbal, Abd-Krim Seghouane, *Senior Member, IEEE* and Tülay Adalı, *Fellow, IEEE*

***Abstract*— *Objective:* Analysis of functional magnetic resonance imaging (fMRI) data from multiple subjects is at the heart of many medical imaging studies, and approaches based on dictionary learning (DL) are recently noted as promising solutions to the problem. However, the DL-based methods for fMRI analysis proposed to date do not naturally extend to multi-subject analysis. In this paper, we propose a dictionary learning algorithm for multi-subject fMRI data analysis. *Methods:* The proposed algorithm (named ShSSDL) is derived based on a temporal concatenation, which is particularly attractive for the analysis of multi-subject task-related fMRI data sets. It differs from existing dictionary learning algorithms in both its sparse coding and dictionary update stages and has the advantage of learning a dictionary shared by all subjects as well as a set of subject-specific dictionaries. *Results:* Performance of the proposed dictionary learning algorithm is illustrated using simulated and real fMRI datasets. The results show that it can successfully extract shared as well as subject-specific latent components. *Conclusion:* In addition to offering a new dictionary learning approach, when applied on multi-subject fMRI data analysis, the proposed algorithm generates a group level as well as a set of subject-specific spatial maps. *Significance:* The proposed algorithm has the advantage of learning simultaneously multiple dictionaries providing us with a shared as well discriminative source of information about the analyzed fMRI data sets.**

***Index Terms*—Dictionary learning, functional magnetic resonance imaging (fMRI), multi-subject analysis, sparse decomposition, temporal concatenation.**

## I. Introduction

Blood-oxygen-level dependent (BOLD) based functional magnetic resonance imaging (fMRI) has been very useful for the identification of functional networks related to a particular task [1], [2] leading to better understanding of the functional localizations within the brain. For task-related data, the general linear model (GLM) has been extensively used, see e.g. [3]–[5], however a key drawback of this approach is that it assumes that the canonical hemodynamic response function (HRF) [6] is known, and is same across different functional networks, which are both limiting assumptions. In addition, it can only extract information (spatial maps) related to task-related activity and ignores the intrinsic brain function, which might not be directly linked to the external stimuli [7]. Hence, data-driven methods such as principal component analysis (PCA) [8], independent component analysis (ICA) [9], [10], canonical correlation analysis (CCA) [11], and multiset CCA [12] have been increasingly used for fMRI analysis. In particular ICA has been widely used in the analysis of both task-related and resting state fMRI datasets [13]. More recently sparse nature of spatial networks has been noted, and a healthy debate has started over the role of different starting points for the analysis of fMRI data, independence vs sparsity [14], [15]. In [15], it is noted that interpretations should be made carefully regarding the role of independence vs sparsity for the final decomposition clarifying issues in [14], but it is also emphasized that sparsity is a useful starting point in fMRI analysis, which has been also noted in [16], the pioneering work that started the activity in data-driven fMRI analysis. These have been among the driving forces for the increasing interest in solutions to fMRI analysis that are based on dictionary learning (DL).

There has been growing interest in the sparse representation of signals [17] and they have been successfully applied to a number of problems including signal representation [18], [19], image denoising [20], and recognition [21], [22] leading to very desirable performance. In these applications, the idea is to represent the given signal as a linear combination of a few columns (atoms) taken from an overcomplete basis set, called a *dictionary*. A key motivating result is that sparse linear codes for natural images (over the dictionary) develop similar receptive fields to the simple cells in the primary visual cortex [23]. Hence, various dictionary learning and sparse representation methods [24] have been proposed for the analysis of fMRI data. In [25], the authors formulate a sparse general linear model framework, in [26], DL and CCA are used to extract meaningful temporal dynamics from the fMRI data which are then used to generate activation maps using regression analysis. [27] provides a fast incoherent $K$ singular value decomposition (K-SVD) method tuned for fMRI data analysis. Authors in [28] include the correlation structure and in [29], the temporal smoothness prior in the DL formulation. In [30]–[32], based on assumption that the underlying dynamics of each voxels' fMRI time courses are sparse with linear neural integration, authors use DL methods to identify functional brain networks. All these algorithms fac-

This work was supported in part by the Australian Research Council through Grant FT. 130101394 to Seghouane and Iqbal and in part by the grants NSF-NCS-FO 1631838 and NSF-CCF 1618551 to Adalı.

Abd-Krim Seghouane and Asif Iqbal are with the Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, Australia. e-mails: abd-krim.seghouane@unimelb.edu.au and aiqbal1@student.unimelb.edu.au.

Tülay Adalı is with the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore, USA, email: adali@umbc.edu.

torize the whole-brain fMRI dataset into *typically* an overcomplete basis dictionary and the corresponding coefficient matrix via dictionary learning method [19]. In these formulations, atoms of the dictionary correspond to time courses (TCs) and the corresponding coefficient vector (row of the coefficient matrix) represents the spatial activity map corresponding to the TC. Furthermore, in [33], authors analyze resting state fMRI (rsfMRI) datasets by decomposing subject datasets into sub-specific TCs and spatial maps (SMs) while modeling the sub-specific SMs as noisy versions of shared population-level brain maps. In [34], authors combine DL and multi-set CCA to extract population-level activation maps. Finally, authors in [35] learn a dictionary from temporally concatenated multi-subject rsfMRI datasets reduced in the time dimension, leading to reduction in computational load without loosing much reliability in terms of activation maps recovery.

Though the methods that promote sparsity have been shown to successfully estimate functional networks, a major limitation in their formulation is that, for each single subject dataset, the network components (atoms) are learned in an unsupervised way, making brain activity analysis across multiple subjects difficult. In [25], [30], [31], such comparisons are performed by sorting and visually checking individual activation patterns (which is time consuming) and then averaging their results. Such an approach becomes computationally prohibitive as the number of datasets and the size of dictionary increases, but more importantly, it fails to take advantage of the additional shared information across the multiple datasets while performing the analysis.

In this paper, our main focus is on task-fMRI (tfMRI) datasets where all subjects perform an exact similar task (study specific). We propose a joint learning framework to decompose the multi-subject tfMRI datasets into a shared dictionary/sparse code pair as well as sub-specific ones. We named this framework as Shared and Subject-Specific Dictionary Learning (ShSSDL). Our aim is to represent each voxels' TC from every subject dataset by a linear combination of a few atoms from a shared dictionary and a few atoms from a sub-specific one. By using a formulation based on temporal concatenation of the fMRI datasets, we separate the shared dictionary (containing similar neural dynamics) and their corresponding shared sparse coefficient matrix (spatial maps (SMs)) from their sub-specific counterparts. As a result, by locating a neural dynamic of interest from the shared dictionary, we can directly analyze the associated population-level spatial map corresponding to the entire multi-subject tfMRI dataset.

Rest of the paper is organized as follows. In Section II, we review briefly the dictionary learning and sparse coding problems. The proposed dictionary learning algorithm is described in Section III with the proposed solution derived in Section IV. The experimental validation is presented in Section V followed by the concluding remarks in Section VI.

## II. Background

Denote an fMRI dataset $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$, containing $N$ variables (brain voxels) with $n$ observations (time points), where $\mathbf{y}_i \in \mathbb{R}^n$ contains the observations for $i^{th}$ variable. According to the sparse representation theory, all variables in $\mathbf{Y}$ can be compactly represented by a dictionary as:

$$\{\mathbf{D}, \mathbf{X}\} = \arg\min_{\mathbf{D},\mathbf{X}} \parallel \mathbf{Y} - \mathbf{D}\,\mathbf{X} \parallel_F^2$$

where $\mathbf{D} \in \mathbb{R}^{n \times K}$ is the dictionary and $\mathbf{X} \in \mathbb{R}^{K \times N}$ is the coefficients matrix, that makes the total representation error as small as possible. This optimization problem is ill-posed unless extra constraints are imposed on the dictionary $\mathbf{D}$ and the sparse codes $\mathbf{X}$. The common constraint on $\mathbf{X}$, is that each column of $\mathbf{X}$ is sparse hence the name *sparse codes* and columns of $\mathbf{D}$ (atoms) are constrained to have unit $\ell_2$ norm. Let the sparse coefficient vectors $\mathbf{x}_i$, $i = 1, ..., N$ constitute the columns of the matrix $\mathbf{X}$, with these constraints, the above objective function can be re-stated as the minimization problem

$$\min_{\mathbf{D},\mathbf{X}} \parallel \mathbf{Y} - \mathbf{D}\,\mathbf{X} \parallel_F^2 \;\; \text{s.t.} \;\; \parallel \mathbf{x}_i \parallel_0 \leq s, \;\; , \parallel \mathbf{d}_k \parallel_2 = 1 \qquad (1)$$

where $\mathbf{d}_k$ is the $k^{th}$ column of $\mathbf{D}$, $\parallel . \parallel_F$ is the Frobenius norm, $\parallel . \parallel_0$ is the $\ell_0$ quasi-norm, which counts the number of nonzero coefficients, $\parallel . \parallel_2$ is the $\ell_2$ norm, and a constant $s \ll K$. The constraint on $\mathbf{D}$ keeps the atoms from getting arbitrarily large leading to small values of $\mathbf{x}_i$ [18]. The generally used optimization strategy, not necessarily leading to a global minimum consists in splitting the problem into two stages which are alternately solved within an iterative loop. These two stages are, first, the sparse coding stage, where $\mathbf{D}$ is fixed and the sparse coefficient vectors are found by solving

$$\begin{gathered} \hat{\mathbf{x}}_i = \arg\min_{\mathbf{x}_i} \parallel \mathbf{y}_i - \mathbf{D}\,\mathbf{x}_i \parallel_2^2; \\ \textit{subject to} \;\; \parallel \mathbf{x}_i \parallel_0 \leq s, \forall\, i = 1, ..., N. \end{gathered} \qquad (2)$$

In practice, the sparse coding stage is often approximately solved by using either a greedy pursuit or convex relaxation approaches [36]. Second, the dictionary update stage where $\mathbf{X}$ is fixed and $\mathbf{D}$ is derived by solving

$$\hat{\mathbf{D}} = \arg\min_{\mathbf{D}} \parallel \mathbf{Y} - \mathbf{D}\,\mathbf{X} \parallel_F^2 \;\; \text{s.t.} \;\; \parallel \mathbf{d}_k \parallel_2 = 1 \qquad (3)$$

This is where most of the DL algorithms differ. One approach is to update all atoms of the dictionary at the same time (*in parallel*) using least squares [37], [38] or maximum likelihood [39], [40], whereas, the other approach is to update the dictionary atom by atom by breaking the global minimization (3) into $K$ sequential minimization problems [18], [41].

## III. Shared and Subject-Specific Dictionary Learning (ShSSDL)

Direct application of the dictionary learning algorithm described above on each fMRI dataset separately, as done in [25], [30], [31], generates multiple dictionaries and avoids taking into consideration the shared information across the multiple datasets while performing the analysis. For development of the proposed dictionary learning algorithm to simultaneously analyze fMRI datasets from a group of $p$ subjects, we could consider the spatial concatenation of the fMRI datasets $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_p]$ of size $n \times pN$ as in [42] where $n$ is the

number of time points for each subject, $N$ the number of voxels per subject and $p$ the number of subjects with the model

$$[\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_p] = \mathbf{D}_0\,\mathbf{X} = \mathbf{D}_0\,[\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p] \tag{4}$$

where the dictionary $\mathbf{D}_0$ is a matrix of size $n \times K_0$ that contains the corresponding set of shared TCs across all the subjects and the sparse codes $\mathbf{X}$ is a matrix of size $K_0 \times pN$ that contains spatial activation maps of the $p$ subjects associated to each of the TCs. In rest of the paper, all dictionaries are constrained to have atoms with unit $\ell_2$ norm.

The model considered in (4) accounts only for the shared information and ignores the sub-specific information. In contrast, we propose to learn not only a dictionary that is common for all subjects but also $p$ sub-specific ones as well. Our proposed algorithm aims to decompose each subject tfMRI dataset in a structured way, i.e., instead of just learning a dictionary that is shared by all subjects, we propose to learn a structured dictionary $\acute{\mathbf{D}}_i = [\mathbf{D}_0, \mathbf{D}_i]$ of size $n \times (K_0 + K_i)$ where $\mathbf{D}_0$ and $\mathbf{D}_i$ are the shared and sub-specific dictionaries associated with subject $i$ respectively. Therefore, we not only require $\mathbf{D}_0$ to have the capability of characterizing the shared information across the multiple subjects fMRI datasets, but also require $\mathbf{D}_i$ to contain the sub-specific information as well. Thus, each subject dataset $\mathbf{Y}_i$ can be decomposed as

$$\mathbf{Y}_i \simeq \acute{\mathbf{D}}_i\,\acute{\mathbf{X}}_i = [\mathbf{D}_0, \mathbf{D}_i] \begin{bmatrix} \mathbf{X}_0 \\ \mathbf{X}_i \end{bmatrix} = \mathbf{D}_0\mathbf{X}_0 + \mathbf{D}_i\mathbf{X}_i \tag{5}$$

where the shared information dictionary $\mathbf{D}_0$ and sub-specific dictionary $\mathbf{D}_i$ are matrices of sizes $n \times K_0$ and $n \times K_i$ respectively, $\mathbf{X}_0$ is a matrix of size $K_0 \times N$ representing the sparse codes of $\mathbf{Y}_i$ over $\mathbf{D}_0$ and $\mathbf{X}_i$ is a matrix of size $K_i \times N$ representing the sparse coding of $\mathbf{Y}_i$ over $\mathbf{D}_i$. As a consequence of such formulation, besides having a shared dictionary $\mathbf{D}_0$, $\mathbf{X}_0$ (the sparse codes of $\mathbf{Y}_i$, $i = 1, ..., p$ over $\mathbf{D}_0$) should be the same for all $p$ subjects as well. This implies that the common characteristics underlying all $p$ fMRI datasets $\mathbf{Y}_i$, are shared between $p$ subjects via the shared information dictionary $\mathbf{D}_0$. In contrast, the unique sparse code matrix $\mathbf{X}_i$, expresses unique characteristics associated with the fMRI datasets $\mathbf{Y}_i$, $i = 1, ..., p$.

Based on the data model (5), the problem of dictionary learning for temporally concatenated multi-subject fMRI datasets can be cast as

$$\min_{\mathbf{\Theta},\mathbf{X}} \left\| \overbrace{\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_p \end{bmatrix}}^{\mathbf{Y}} - \overbrace{\begin{bmatrix} \mathbf{D}_0 & \mathbf{D}_1 & 0 & 0 & \cdots & 0 \\ \mathbf{D}_0 & 0 & \mathbf{D}_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{D}_0 & 0 & 0 & \cdots & 0 & \mathbf{D}_p \end{bmatrix}}^{\mathbf{\Theta}} \overbrace{\begin{bmatrix} \mathbf{X}_0 \\ \mathbf{X}_1 \\ \vdots \\ \vdots \\ \mathbf{X}_p \end{bmatrix}}^{\mathbf{X}} \right\|_F^2$$
$$\text{s.t.} \quad \| \mathbf{x}_i^j \|_0 \leq s_i, \quad \| \mathbf{x}_0^j \|_0 \leq s_0, \quad \| \mathbf{d}_k \|_2 = 1$$
$$\forall \quad i = 1, 2, \ldots, p \text{ and } j = 1, 2, \ldots, N \tag{6}$$

where $\mathbf{Y} = \left[\mathbf{Y}_1^T, \mathbf{Y}_2^T, \cdots, \mathbf{Y}_p^T\right]^T$ is the full dataset matrix, $\mathbf{Y}_i$ are the individual fMRI datasets, $\mathbf{x}_i^j$ is the $j$th column of $\mathbf{X}_i$, $s_i, s_0$ are the signal sparsity parameters, and $p$ is the number of subjects selected for analysis. Similarly, $\mathbf{X} = \left[\mathbf{X}_0^T, \mathbf{X}_1^T, \mathbf{X}_2^T, \cdots, \mathbf{X}_p^T\right]^T$ contains the corresponding sparse codes for each dataset. Our aim here is to represent each signal from the dataset $\mathbf{Y}_i$ as a linear combination of a few atoms from $\mathbf{D}_0$ (*shared dictionary*) and $\mathbf{D}_i$ (*sub-specific dictionary*). To achieve this goal, we propose the following Shared and Subject-Specific Dictionary Learning (ShSSDL) model:

$$\min_{\mathbf{\Theta},\mathbf{X}} \sum_{i=1}^{p} \left\{ \frac{1}{2} \| \mathbf{Y}_i - \mathbf{D}_0\mathbf{X}_0 - \mathbf{D}_i\mathbf{X}_i \|_F^2 + \eta \| \mathbf{D}_i^T\mathbf{A}_i \|_F^2 \right\}$$
$$\text{subject to } \| \mathbf{x}_i^j \|_0 \leq s_i, \| \mathbf{x}_0^j \|_0 \leq s_0; \| \mathbf{d}_k \|_2 = 1 \tag{7}$$

The minimization of the first term of (7) is equivalent to maximizing the representation power of $\mathbf{D}_0$ and $\mathbf{D}_i$ for each dataset. However, just minimizing the first term of (7) is not sufficient in separating the shared and sub-specific temporal dynamics into separate dictionaries, i.e., the shared patterns may appear in different sub-specific dictionaries as well and vice versa. To alleviate this problem we included an incoherence penalty term $\| \mathbf{D}_i^T\mathbf{A}_i \|_F$ into the objective function (7), where $\mathbf{A}_i = [\mathbf{D}_0, \mathbf{D}_1, \ldots, \mathbf{D}_{i-1}, \mathbf{D}_{i+1}, \ldots, \mathbf{D}_p]$ is the concatenation of all except the currently updating dictionary. The incoherence of the dictionaries is controlled by a positive parameter $\eta$. The authors in [43] have demonstrated that incoherent dictionaries improve the effectiveness of sparse representation. A similar penalty term has been used in [27] to force the learned dictionary atoms to be incoherent among themselves and in [44] to transfer common features from sub-specific dictionaries to the shared dictionary.

Due to the structure and constraints of the minimization problem of (7), the objective is non-convex and solving it in its current form is difficult. In the next section we propose an alternating optimization procedure named ShSSDL algorithm to solve the problem in (7) block by block.

## IV. Objective Optimization

As in the case of most dictionary learning algorithms, we follow an alternating optimization procedure to optimize (7) one block at a time, which is outlined below:

- Fix $\mathbf{D}_0, \mathbf{D}_i$ and find their sparse coefficient matrices $\mathbf{X}_0, \mathbf{X}_i$
- Fix $\mathbf{X}_0, \mathbf{X}_i$ and optimize for $\mathbf{D}_0, \mathbf{D}_i$

The details of these minimizations are given in their respective sections.

### A. Sparse coding stage

By fixing $\mathbf{D}_0, \mathbf{D}_i$, the minimization objective (7) reduces to

$$\hat{\mathbf{X}}_0, \hat{\mathbf{X}}_i = \min_{\mathbf{X}_0,\mathbf{X}_i} \sum_{i=1}^{p} \frac{1}{2} \| \mathbf{Y}_i - \mathbf{D}_0\mathbf{X}_0 - \mathbf{D}_i\mathbf{X}_i \|_F^2$$
$$\text{s.t.} \ \| \mathbf{x}_i^j \|_0 \leq s_i, \| \mathbf{x}_0^j \|_0 \leq s_0, \forall j = 1, 2, \ldots, N \tag{8}$$

To solve this problem, we proceed with a block by block update strategy, i.e. fixing $\mathbf{X}_i$, we try to find $\mathbf{X}_0$ by minimizing

$$\hat{\mathbf{X}}_0 = \min_{\mathbf{X}_0} \frac{1}{2} \| \mathbf{E} - \mathbf{D}_0\mathbf{X}_0 \|_F^2; \ \text{s.t.} \ \| \mathbf{x}_0^j \|_0 \leq s_0 \tag{9}$$

**Algorithm 1:** Sparse Coding Solution to (8)

**Input:** Complete dataset $\mathbf{Y},\ \mathbf{D}_0,\ \mathbf{D}_i,\ \mathbf{X}_0,\ \mathbf{X}_i,\ p$
**Parameters:** $s_o,\ s_i,\ \zeta$
**for** $v = 1 : \zeta$ **do**
  Finding $\mathbf{X}_0$:
    Compute $\mathbf{E} = 1/p \sum_{i=1}^{p} \{\mathbf{Y}_i - \mathbf{D}_i\mathbf{X}_i\}$
    Use OMP to solve:

$$\hat{\mathbf{X}}_0 = \min_{\mathbf{X}_0} \frac{1}{2} \parallel \mathbf{E} - \mathbf{D}_0\mathbf{X}_0 \parallel_F^2;\ \text{s.t.}\ \parallel \mathbf{x}_0^j \parallel_0 \leq s_0$$

  Finding $\mathbf{X}_i$:
  **for** $i = 1 : p$ **do**
    Compute $\mathbf{G}_i = \mathbf{Y}_i - \mathbf{D}_0\mathbf{X}_0$
    Use OMP to solve:

$$\hat{\mathbf{X}}_i = \min_{\mathbf{X}_i} \frac{1}{2} \parallel \mathbf{G}_i - \mathbf{D}_i\mathbf{X}_i \parallel_F^2;\ \text{s.t.}\ \parallel \mathbf{x}_i^j \parallel_0 \leq s_i$$

  **end**
**end**
**Output:** $\mathbf{X}_0,\ \mathbf{X}_i$

where $\mathbf{E} = 1/p \sum_{i=1}^{p} \{\mathbf{Y}_i - \mathbf{D}_i\mathbf{X}_i\}$ and $\mathbf{x}_0^j$ is the $j$th column of $\mathbf{X}_0$. Here we opt to include a scaling factor $1/p$ to keep the $\mathbf{X}_0$ matrix entries from getting very high. After learning $\mathbf{X}_0$, we fix it and find update for $\mathbf{X}_i$ by minimizing

$$\hat{\mathbf{X}}_i = \min_{\mathbf{X}_i} \frac{1}{2} \parallel \mathbf{G}_i - \mathbf{D}_i\mathbf{X}_i \parallel_F^2;\ \text{s.t.}\ \parallel \mathbf{x}_i^j \parallel_0 \leq s_i \tag{10}$$

where $\mathbf{G}_i = \mathbf{Y}_i - \mathbf{D}_0\mathbf{X}_0$ for all $p$ subject datasets. The problems (9) and (10) can be solved by the greedy pursuit algorithm Orthogonal Matching Pursuit (OMP) [45]. The sparse coding stage is then obtained by iterating (9) and (10) until convergence or by applying only a few iterations of these equations. Based on our experience, iterating them more than twice does not lead to significant improvement, thus in our experiments, we iterate these equations only twice. Details of the sparse coding stage are summarized in Algorithm 1.

### *B. Dictionary update stage*

After optimizing for $\mathbf{X}_0$ and $\mathbf{X}_i$, the shared and sub-specific dictionaries are found by minimizing (7) with fixing $\mathbf{X}_0, \mathbf{X}_i$, the resulting objective function becomes

$$\begin{aligned}\hat{\mathbf{D}}_0, \hat{\mathbf{D}}_i = \min_{\mathbf{D}_0, \mathbf{D}_i} \sum_{i=1}^{p} \left\{ \frac{1}{2} \parallel \mathbf{Y}_i - \mathbf{D}_0\mathbf{X}_0 - \mathbf{D}_i\mathbf{X}_i \parallel_F^2 \right\} \\ + \eta \sum_{i=0}^{p} \parallel \mathbf{D}_i^T\mathbf{A}_i \parallel_F^2\end{aligned} \tag{11}$$

Similar to the sparse coding stage, we divide the above problem into two separate problems and try to minimize them alternatively, i.e., we fix all $\mathbf{D}_i$ to minimize for $\mathbf{D}_0$ and vice versa.

*1) Solving for $\boldsymbol{D}_0$:* Fixing $\mathbf{D}_i$ in (11), the minimization of (11) is equivalent to minimization of

$$\hat{\mathbf{D}}_0 = \min_{\mathbf{D}_0} \frac{1}{2} \parallel \mathbf{E} - \mathbf{D}_0\mathbf{X}_0 \parallel_F^2 + \eta \parallel \mathbf{D}_0^T\mathbf{A}_0 \parallel_F^2 \tag{12}$$

where $\mathbf{E} = 1/p \sum_{i=1}^{p} \{\mathbf{Y}_i - \mathbf{D}_i\mathbf{X}_i\}$ is the error matrix and $\mathbf{A}_0 = [\mathbf{D}_1, \ldots, \mathbf{D}_p]$. To solve this, we start by introducing a relaxation variable $\mathbf{Z}$, such that (12) is rewritten as

$$\begin{aligned}\hat{\mathbf{D}}_0 = \min_{\mathbf{D}_0} \frac{1}{2} \parallel \mathbf{E} - \mathbf{D}_0\mathbf{X}_0 \parallel_F^2 + \eta \parallel \mathbf{Z}^T\mathbf{A}_0 \parallel_F^2 \\ \text{s.t.}\ \parallel \mathbf{D}_0 - \mathbf{Z} \parallel_F^2 = 0\end{aligned} \tag{13}$$

The above problem can then be solved by using the alternating directions method of multipliers (ADMM) [46] method. Choosing the ADMM method to solve this problem was particularly due to its usefulness when the optimization variables ($\mathbf{D}$ and $\mathbf{Z}$) admit closed form solutions. On the other hand, ADMM has only a single tuning parameter $\mu$ and with a few mild conditions, it can be shown that the algorithm does converge for all parameter values; see [47] and references therein. The augmented Lagrangian function of (13) is

$$\begin{aligned}L_\mu(\mathbf{D}_0, \mathbf{Z}, \mathbf{W}) = \frac{1}{2} \parallel \mathbf{E} - \mathbf{D}_0\mathbf{X}_0 \parallel_F^2 + \eta \parallel \mathbf{Z}^T\mathbf{A}_0 \parallel_F^2 \\ + \frac{\mu}{2} \parallel \mathbf{D}_0 - \mathbf{Z} \parallel_F^2 + \text{tr}\left[\mathbf{W}^T(\mathbf{D}_0 - \mathbf{Z})\right]\end{aligned} \tag{14}$$

where $\mathbf{W}$ is the Lagrange multiplier matrix and $\mu$ is a positive number. Starting with $\mathbf{Z} = \mathbf{W} = \mathbf{0}$, the solution to (14) minimization is found by alternatively solving each of the following subproblems until convergence:

$$\mathbf{D}_0 = \left(\mathbf{E}\mathbf{X}_0^T + \mu\mathbf{Z} - \mathbf{W}\right)\left(\mathbf{X}_0\mathbf{X}_0^T + \mu\mathbf{I}\right)^{-1} \tag{15}$$

$$\mathbf{Z} = \left(2\eta\mathbf{A}_0\mathbf{A}_0^T + \mu\mathbf{I}\right)^{-1}(\mathbf{W} + \mu\mathbf{D}_0) \tag{16}$$

$$\mathbf{W} = \mathbf{W} + \mu(\mathbf{D}_0 - \mathbf{Z}) \tag{17}$$

where $\mathbf{I}$ is the identity matrix whose dimensions should be taken from the context.

*2) Solving for sub-specific $\boldsymbol{D}_i$s:* After finding the shared dictionary $\mathbf{D}_0$, we find the individual sub-specific dictionaries by minimizing the following function for each subject dataset

$$\hat{\mathbf{D}}_i = \min_{\mathbf{D}_i} \frac{1}{2} \parallel \mathbf{G}_i - \mathbf{D}_i\mathbf{X}_i \parallel_F^2 + \eta \parallel \mathbf{D}_i^T\mathbf{A}_i \parallel_F^2 \ \forall i = 1, 2, \ldots, p \tag{18}$$

where $\mathbf{G}_i = \mathbf{Y}_i - \mathbf{D}_0\mathbf{X}_0$ for all $p$ subject datasets. The minimization of (18) is similar to (12) and can be obtained by solving the following subproblems till convergence for all the subject datasets:

$$\mathbf{D}_i = \left(\mathbf{G}_i\mathbf{X}_i^T + \mu\mathbf{Z} - \mathbf{W}\right)\left(\mathbf{X}_i\mathbf{X}_i^T + \mu\mathbf{I}\right)^{-1} \tag{19}$$

$$\mathbf{Z} = \left(2\eta\mathbf{A}_i\mathbf{A}_i^T + \mu\mathbf{I}\right)^{-1}(\mathbf{W} + \mu\mathbf{D}_i) \tag{20}$$

$$\mathbf{W} = \mathbf{W} + \mu(\mathbf{D}_i - \mathbf{Z}) \tag{21}$$

The step by step review of the dictionary update stage is summarized in Algorithm 2. Once the dictionaries have been initialized, we repeat steps outlined in Algorithm 1 and 2 multiple times or until a stopping criteria is reached. The summary of ShSSDL framework is shown in Fig. 1.

**Algorithm 2:** ADMM algorithm for (11)

**Input:** Complete dataset $\mathbf{Y}$, $\mathbf{D}_0, \mathbf{D}_i, \mathbf{X}_0, \mathbf{X}_i$, $p$ and $\eta$
**Initialize** : $\mathbf{Z} = 0$, $\mathbf{W} = 0$, $\mu = 10^{-4}$, $\max_\mu = 10^{10}$, $\rho = 2.5$, $\epsilon = 10^{-4}$

**Solving for $\mathbf{D}_0$:**
Formulate $\mathbf{A}_0 = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_p]$
Compute $\mathbf{E} = 1/p \sum_{i=1}^{p} \{\mathbf{Y}_i - \mathbf{D}_i \mathbf{X}_i\}$
**while** $\| \boldsymbol{D}_0 - \boldsymbol{Z} \|_F \geq \epsilon$ **do**
  Fix other variables and update $\mathbf{D}_0$ by:

$$\mathbf{D}_0 = \left(\mathbf{E}\mathbf{X}_0^T + \mu\mathbf{Z} - \mathbf{W}\right)\left(\mathbf{X}_0\mathbf{X}_0^T + \mu\mathbf{I}\right)^{-1}$$

  Normalize the columns of $\mathbf{D}_0$ to have $\ell_2$-norm = 1
  Fix other variables and update $\mathbf{Z}$ by:

$$\mathbf{Z} = \left(2\eta\mathbf{A}_0\mathbf{A}_0^T + \mu\mathbf{I}\right)^{-1}(\mathbf{W} + \mu\mathbf{D}_0)$$

  Normalize the columns of $\mathbf{Z}$ to have $\ell_2$-norm = 1
  Update the Lagrange multiplier $\mathbf{W}$ by:

$$\mathbf{W} = \mathbf{W} + \mu\left(\mathbf{D}_0 - \mathbf{Z}\right)$$

  Update $\mu$ as: $\mu = \min(\rho\mu, \max_\mu)$
**end**
**Solving for $\mathbf{D}_i$:**
**for** $i = 1 : p$ **do**
  Initialize $\mathbf{Z} = \mathbf{W} = 0$, and $\mu = 10^{-4}$
  Formulate $\mathbf{A}_i = [\mathbf{D}_0, \mathbf{D}_1, \ldots, \mathbf{D}_{i-1}, \mathbf{D}_{i+1}, \ldots, \mathbf{D}_p]$
  Compute $\mathbf{G}_i = \mathbf{Y}_i - \mathbf{D}_0\mathbf{X}_0$
  **while** $\| \boldsymbol{D}_i - \boldsymbol{Z} \|_F \geq \epsilon$ **do**
    Fix other variables and update $\mathbf{D}_i$ by:

$$\mathbf{D}_i = \left(\mathbf{G}_i\mathbf{X}_i^T + \mu\mathbf{Z} - \mathbf{W}\right)\left(\mathbf{X}_i\mathbf{X}_i^T + \mu\mathbf{I}\right)^{-1}$$

    Normalize columns of $\mathbf{D}_i$ to have $\ell_2$-norm = 1
    Fix other variables and update $\mathbf{Z}$ by:

$$\mathbf{Z} = \left(2\eta\mathbf{A}_i\mathbf{A}_i^T + \mu\mathbf{I}\right)^{-1}(\mathbf{W} + \mu\mathbf{D}_i)$$

    Normalize columns of $\mathbf{Z}$ to have $\ell_2$-norm = 1
    Update the Lagrange multiplier $\mathbf{W}$ by:

$$\mathbf{W} = \mathbf{W} + \mu\left(\mathbf{D}_i - \mathbf{Z}\right)$$

    Update $\mu$ as: $\mu = \min(\rho\mu, \max_\mu)$
  **end**
**end**
**Output:** $\mathbf{D}_0$, $\mathbf{D}_i$

## V. Experimental Evaluation

In this section, we present a performance comparison between the proposed ShSSDL algorithm and CODL [35] using simulated and real experimental multi-subject fMRI datasets. In simulation study the goal is to analyze ShSSDLs' ability in decomposing the datasets into group-level (shared) TC/SM pairs and the sub-specific TC/SM pairs. After establishing performance of the algorithm, we analyze Q1 release of the publicly available Human Connectome Project (HCP) motor task fMRI datasets [48] to validate our proposed algorithm.

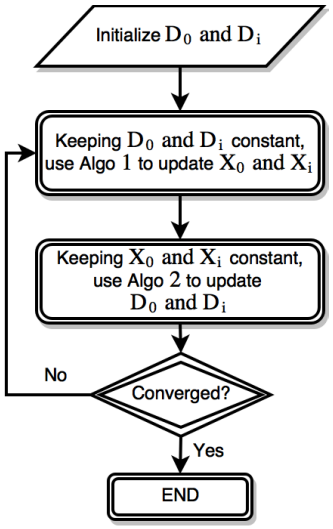



Fig. 1. Flowchart summarizing ShSSDL Algorithm.

The performance comparison with CODL for both experiments is also provided. The details of these experiments are given below.

### A. Simulation Study

We tested the algorithms in two different scenarios:

1) A basic multi-subject fMRI dataset: The subject brain scans are all perfectly aligned and have the same HRFs, thus the shared TCs/SMs are exactly the same across all subjects.
2) A more realistic multi-subject fMRI dataset: The subject brain scans have spatial variability and have different HRFs, generating small differences in the shared TCs/SMs across all subjects.

*1) Scenario 1:* To generate basic fMRI datasets for $p = 6$ subjects, we used the publicly available SimTB toolbox [49] to generate 9 source images of size $(100 \times 100)$ voxels and their corresponding TCs with 150 time points. Repetition time (TR) was set to 2 sec/sample. The required $p$ datasets $\mathbf{Y}_i \in \mathbb{R}^{150\times 10^4}$ were then generated by a linear mixture model and were corrupted with additive white Gaussian noise (AWGN) with $\mu = 0$ and $\sigma = 0.2$. Every subject dataset contained 4 TC/SM pairs in total, out of which 3 pairs were shared (attributed to 3 tasks) and 1 unique pair which was different for each subject. The simulated sources and their TCs (ground truth) are shown in Supplemental Fig. 1 a). Here sources/TCs $(1, 2, 3)$ are shared across all subjects, whereas, sources/TCs $(4, 5, \ldots, 9)$ are unique and are only present in subjects $(1, 2, \ldots, 6)$ respectively.

*2) Scenario 2:* In this scenario, we again used SimTB toolbox [49] to generate $p = 6$ fMRI datasets. Similar to the previous scenario, all subjects contained 3 shared TC/SM pairs and a unique one. To simulate inter-subject variability, we added spatial variability in shared SMs across all subjects by introducing random translations ($\mu = 0, \sigma = 2$ voxels) in x and y directions, rotations ($\mu = 0, \sigma = 2.5$ degrees), and spreads ($\mu = 1, \sigma = 0.03$), where $\mu$ and $\sigma$ represent mean and

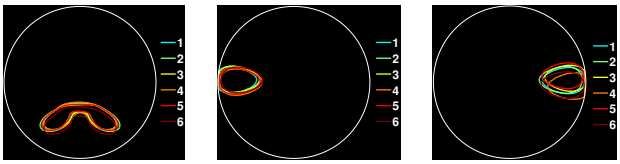


Fig. 2. The shared spatial maps showing the spatial variability of sources across subjects. Each color represents the source contours for a different subject.

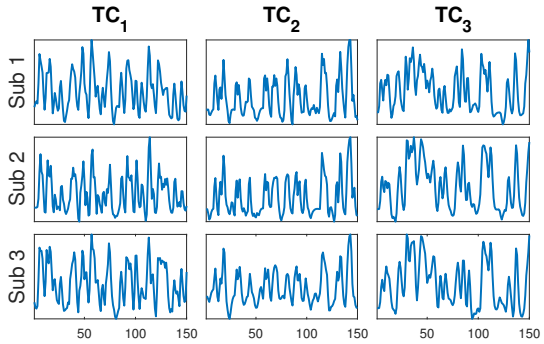



Fig. 3. The shared TCs for first 3 subjects simulated using different parameters for the canonical HRFs.

standard deviation respectively. The resulting SMs are shown in Fig. 2. Similarly, we introduced temporal variability in the simulated shared TCs as well that can be seen in Fig. 3 and the datasets were generated via the linear mixture model. For comparisons, the ground truth (GT) for shared SMs/TCs were generated by taking the mean of each SM/TC shared pair for all subjects. The resulting simulated sources and their TCs (GT) for all subjects are shown in Supplemental Fig. 2 a).

*3) Dictionary Learning:* For both simulation scenarios, the datasets were decomposed in a similar way. Starting with random initial dictionaries, the signal matrices $\mathbf{Y}_i$s' were decomposed by our algorithm into a shared dictionary $\mathbf{D}_0$ and 6 sub-specific dictionaries $\mathbf{D}_i$s along with their shared and sub-specific activation patterns in $\mathbf{X}_0$ and $\mathbf{X}_i$ respectively. In case of real fMRI datasets, we do not know the exact number of components present in the dataset [50]. Hence, rather than assuming same number of components as present in the generated datasets, we allow both shared and sub-specific dictionaries to learn more components from the datasets by setting the dictionary sizes as $(\mathbf{D}_0, \mathbf{D}_i) \in \mathbb{R}^{150\times 10}$ (20 components per subject). These dictionaries were learned using signal

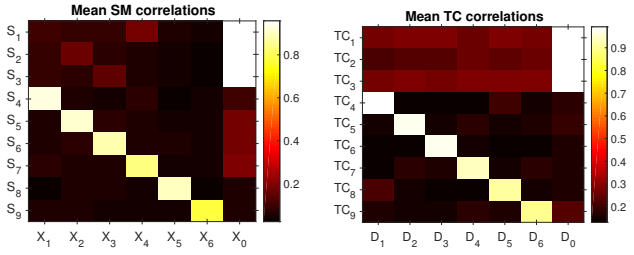



Fig. 4. The mean Ground Truth (GT) SM and TC correlation coefficients for scenario 1 over 100 trials with respect to all estimated sparse code and dictionary matrices, $\mathbf{X}_i$s and $\mathbf{D}_i$s.

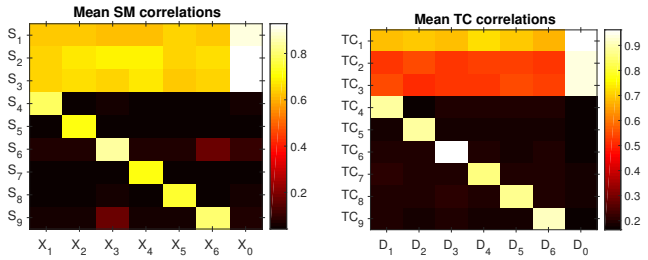



Fig. 5. The mean Ground Truth (GT) SM and TC correlation coefficients for scenario 2 over 100 trials (regenerating datasets for each trial) with respect to all estimated sparse code and dictionary matrices, $\mathbf{X}_i$s and $\mathbf{D}_i$s.

sparsity $s_0 = 2$, $s_i = 3$, and the algorithm was iterated 20 times. The incoherence penalty parameter $\eta$ and convergence parameter $\epsilon$ were set to 2.5 and $10^{-4}$ respectively, while other parameters were kept the same as outlined in Algorithm 2.

To compare CODL and ShSSDL on the common grounds, we opted not to reduce the temporal dimension of the simulated fMRI datasets. As a result, the CODL algorithm essentially reduces to ODL (online dictionary learning) algorithm [19]. Thus, by temporally concatenating the full subject datasets, we used ODL to learn a dictionary of size $900 \times 20$ with $\lambda = 0.15$, batchsize of 200, and 100 iterations. Keeping the dictionary size constant, we experimented with different combinations of parameter values for $(s_0, s_i, \eta$ and $\lambda)$ and selected those with best performance in terms of correlation of extracted SMs/TCs with respect to the GT for both algorithms to allow for a fair comparison.

TABLE I
MEAN, MEDIAN AND STANDARD DEVIATION OF MOST CORRELATED TCS AND SMS W.R.T. GROUND TRUTH AS RECOVERED BY THE SHSSDL AND CODL OVER 100 TRIALS

| | | TCs | | | SMs | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Median | STD | Mean | Median | STD |
| Scenario 1 | ShSSDL | **0.976** | **0.987** | 0.021 | 0.917 | 0.936 | 0.045 |
| | CODL | 0.973 | 0.977 | 0.024 | **0.919** | **0.940** | 0.075 |
| Scenario 2 | ShSSDL | **0.951** | 0.952 | 0.014 | **0.962** | **0.969** | 0.018 |
| | CODL | 0.948 | **0.979** | 0.080 | 0.894 | 0.875 | 0.104 |

*4) Results:* To demonstrate the average performance, we repeated the simulations 100 times with

- different noise ($\sigma = 0.2$) realizations under scenario 1,
- different datasets (using SimTB) under scenario 2.

To analyze the recovered dictionaries/sparse code pairs, we correlated the GT sources/TCs with all the sparse codes/dictionary atoms as recovered by CODL and ShSSDL algorithm. For the case of CODL, we extracted the sub-specific TCs corresponding to the shared SMs from the recovered dictionary, and used the mean TCs for comparison. For visual comparison, the most correlated sparse code/atom pairs as recovered by both algorithms over a single trial are shown in Supplemental Figs. 1 and 2 for scenarios 1 and 2 respectively. To demonstrate the effectiveness of ShSSDL algorithm in separating the shared info from the sub-specific info, we illustrate the mean ground truth (GT) SM/TC correlation coefficients for scenario 1 and 2 over 100 trials with respect to all recovered

sparse code/dictionary matrices $\mathbf{X}_i$ and $\mathbf{D}_i$ in Fig. 4 and Fig. 5 respectively. Here, each row represents the mean correlation coefficients of a particular GT SM and TC with every $\mathbf{X}_i$ and $\mathbf{D}_i$ matrix. It is clear from Fig. 4 and 5 that, for both scenarios, the shared sources $(S_1, S_2, S_3)$ and their respective TCs $(TC_1, TC_2, TC_3)$ have been identified and localized into the shared sparse code and dictionary matrices respectively. However, for scenario 1, the shared SM and TC pairs have been completely removed from the sub-specific ones, whereas, for scenario 2 (real case), the sub-specific pairs were able to capture the subject variability as well. This, however, did not affect the overall performance of the algorithm as shared SMs and TC pairs $(\mathbf{X}_i, \mathbf{D}_i)$ were still highly correlated with the GT. For completeness, the mean, median and standard deviation of the SM and TC - GT correlation coefficients recovered by both algorithms over 100 trials are presented in table I. Here it can be seen that, for scenario 1, the overall recovery results of both algorithms are very similar. However, for scenario 2, the ShSSDL's recovery results surpass that of CODL. Another key takeaway from the table is stability of the ShSSDL algorithms' results, i.e. for every case, variance of the achieved results is small when compared to CODLs'. For reproducibility, we have made our simulation code available at https://github.com/AsifIqbal8739/ShSSDL_2017.

In this section, we have shown that our proposed algorithm was successful in localizing the shared TCs/SMs and sub-specific pairs into their corresponding dictionary/sparse code matrices. After establishing its effectiveness, in the next section, we further validate our method on multi-subject experimental task fMRI datasets.

### *B. Multi-subject task fMRI Analysis*

In this section, we apply ShSSDL to fifteen $(p = 15)$ motor task fMRI datasets acquired from the HCP Q1 release [48] to separate the shared temporal hemodynamics and the corresponding activation patterns into a shared dictionary matrix and its sparse code matrix. The acquisition parameters for all datasets were: $90 \times 104$ matrix, 220mm FOV, 72 slices, TR = 0.72s, TE = 33.1ms, flip angle = $52^o$, BW = 2290 Hz/Px, in-plane FOV = $208 \times 180$ mm with 2.0 mm isotropic voxels. The obtained data was already preprocessed with the preprocessing pipeline consisting of motion correction, temporal pre-whitening, slice time correction, global drift removal, and the scans were spatially normalized to a standard MNI152 template and resampled to 2x2x2 mm$^3$ voxels. The reader is referred to [51] and [48] for more details regarding data acquisition and preprocessing.

This task is based on the development in [52] in which participants were presented with a visual cue, asking them to tap their left or right fingers, squeeze their left or right toes, or move their tongue to map the motor areas of the brain. Subjects were presented with a 3 sec visual cue followed by the cue for a specific task, with movement block length of 12 sec (10 movements). A total of 13 blocks with 4 foot movements (2 for each foot), 4 hand movements (2 for each hand), 2 tongue movements, and 3 15-second fixation blocks were carried out by each subject. We generated the paradigm time courses (PTCs) for six task stimuli corresponding to the movements of Left Toe (LT), Right Toe (RT), Left Finger (LF), Right Finger (RF), Tongue, and the visual cue (VC) by convolving the canonical hemodynamic response function [6] with the block task regressors. These PTCs are shown in Fig. 6 a).

*1) fMRI Data Preprocessing:* The tfMRI run duration was 3:34 (min:sec) with a total of $n = 284$ scans. The scans were spatially smoothed using a Gaussian kernel with 6x6x6 mm$^3$ FWHM. Low frequency drifts were removed by using a Discrete Cosine Transform (DCT) basis set with a cutt off frequency of 1/150 Hz, whereas, the high frequency temporal fluctuations were removed by smoothing each voxel TC using a 2.0 sec FWHM Gaussian kernel. Brain volumes from each subject were masked to remove the non-brain voxels resulting in $N = 283494$ voxels for each subject which were then vectorized and placed as rows of $\mathbf{Y}_i (i = 1, \ldots, p)$ yielding a data matrix with size $n \times N$ for each subject. Each column of the matrix $\mathbf{Y}_i$ was then normalized to have zero mean and unit variance as well.

*2) Dictionary Learning:* The shared dictionary $\mathbf{D}_0 \in \mathbb{R}^{n \times 40}$ was initialized with data from subject 1 and sub-dictionaries $\mathbf{D}_i \in \mathbb{R}^{n \times 20}$ were initialized from the corresponding subject datasets. The data matrices $\mathbf{Y}_i s$ were then decomposed using ShSSDL Algorithm into $\mathbf{D}_0/\mathbf{X}_0$ and $\mathbf{D}_i\mathbf{X}_i$ pairs with signal sparsity $s_0 = 1$, $s_i = 2$, and dictionary incoherence controlling parameter $\eta = 500$. The algorithm 1 and 2 were iterated 15 times to optimize the objective (7). Similar to the simulation section, using temporally concatenated full datasets, we used ODL to learn a dictionary $\mathbf{D}_{ODL} \in \mathbb{R}^{np \times 70}$ with $\lambda = 6$ [35], batch size of 2000, and 100 iterations.

*3) Results:* Once the learning process of the proposed ShSSDL model (7) is complete, based on our model, the shared information should have been recovered in the $\mathbf{D}_0/\mathbf{X}_0$ pair and the sub-specific information should be present in $\mathbf{D}_i/\mathbf{X}_i$. Thus, according to the (5) model, we expect those activation maps which have very similar temporal dynamics across all analyzed subjects to be localized in the $\mathbf{D}_0/\mathbf{X}_0$ pair. As we have used motor experimental task dataset with same six PTCs across all subjects, we expect to find atoms most correlated to them in $\mathbf{D}_0$ with their respective activation patterns in $\mathbf{X}_0$. To confirm this, we correlated all six PTCs with atoms from $\mathbf{D}_0$ as well as sub-specific $\mathbf{D}_i s$ and the highest atom-PTC correlation results are shown in Fig. 7. From the figure it is evident that the highest correlated atoms are indeed from the shared dictionary $\mathbf{D}_0$. For visualization, these most correlated atoms (from $\mathbf{D}_0$) w.r.t. the six PTCs as recovered by both algorithms are shown in Fig. 6 a) (ShSSDL) and b) (CODL). Due to the unavailability of the ground truth, we compared both algorithms in terms of their correlations w.r.t. the six PTCs. Fig. 6 a) shows that the extracted TCs by ShSSDL have high correlations as compared to the TCs recovered by CODL Fig. 6 b). In Fig. 6 c) (ShSSDL) and d) (CODL), we have used the corresponding (z-scored) sparse code rows to show the most informative population level activation maps. From the figure it can be seen that both algorithms were able to localize the activity in the motor-cortex region of the brain. Moreover, these results are also

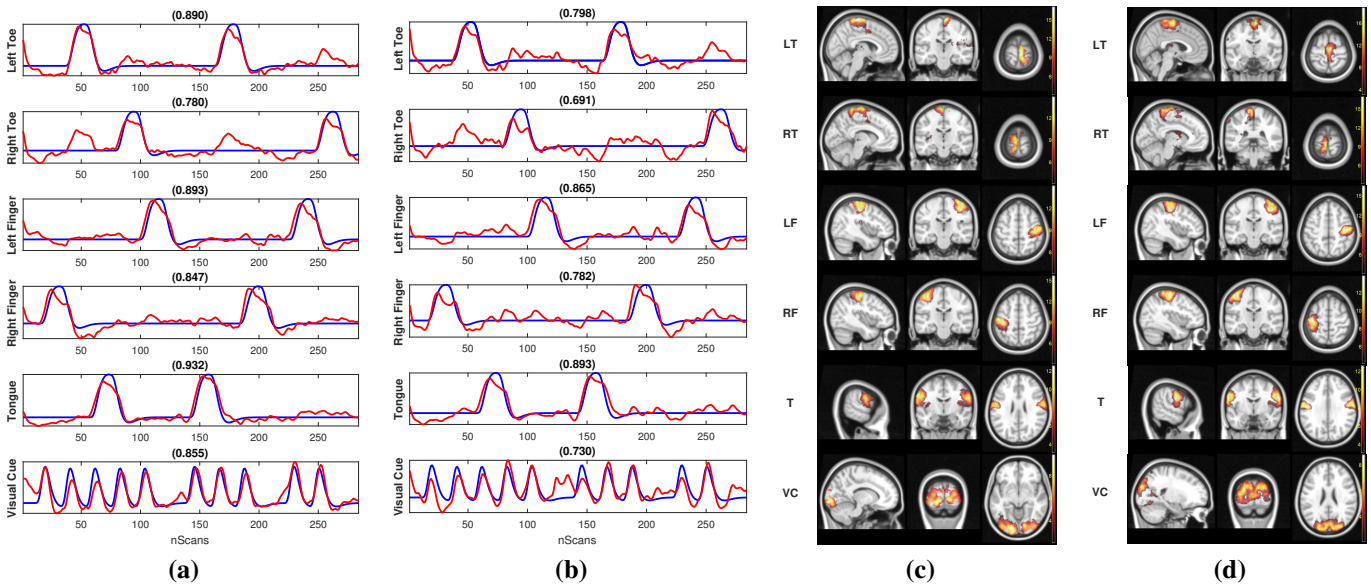



Fig. 6. Most correlated $\mathbf{D}_0$ atoms (*red*) with their respective PTCs (*blue*) recovered by a) ShSSDL, b) CODL. The corresponding correlation coefficients are also given inside parenthesis above each TC plot. Most informative population level z-scored ($p < 0.001$) activation maps for Left Toe, Right Toe, Left Finger, Right Finger, Tongue, and Visual Cue regressors recovered by c) ShSSDL and d) by CODL.

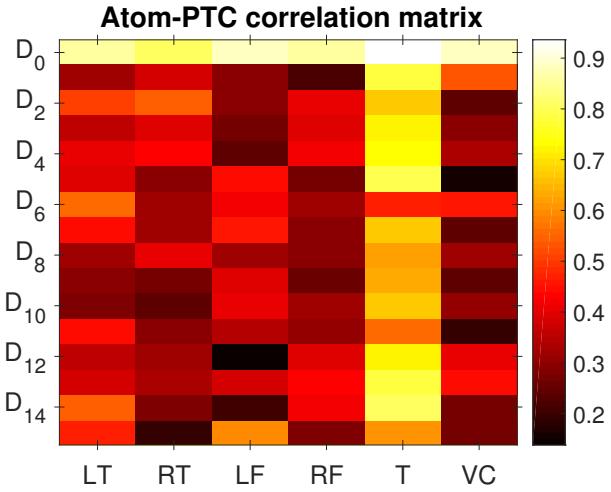



Fig. 7. Correlation matrix of the highest correlated dictionary atoms with six PTCs labeled as Left Toe (LT), Right Toe (RT), Left Finger (LF), Right Finger (RF), Tongue (T), and Visual Cue (VC).

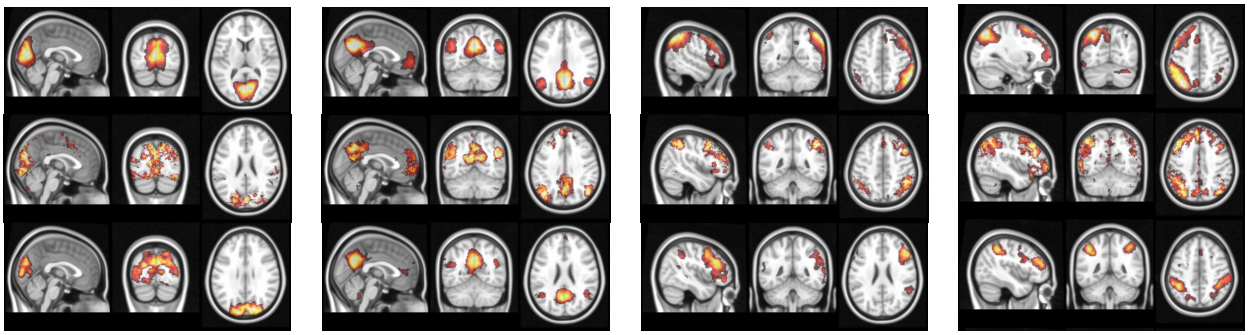
consistent with the results presented in [32] for the same datasets.

To analyze the sub-specific dictionaries, we take a slightly different approach. In [32], the authors have highlighted the existence of ten well-established resting state networks (RSNs) [50] in the experimental task datasets as well. Although these resting state networks could be present in all subject datasets, their respective temporal dynamics need not be similar. Based on this info, we expect to find these RSNs in the sub-specific sparse code matrices $\mathbf{X}_i$. Using the RSN templates from [50], we correlated them with every $\mathbf{X}_i$ and stored the most correlated ones for each subject. Using the most correlated RSNs from each subject, we then generated 10 average RSN maps. For comparison, the correlation of these averaged RSN maps and those recovered by CODL w.r.t. the RSN templates are given in table II. Here it can be seen that the RSN maps recovered by ShSSDL show better correlation with the RSN templates as compared to CODL. To visualize, the most correlated activation maps corresponding to RSNs $[1, 4, 9, 10]$ are shown in Fig. 8. Here the top row contains the template RSNs, middle row contains the maps recovered by ShSSDL, and the bottom row shows the maps recovered by CODL. Except RSN 1, all other maps extracted by ShSSDL show much better resemblance to the templates especially RSN 4 (Default Mode Network) where the activations in frontal lobe are prominent as well.

*4) Parameters for Dictionary Learning:* The most important parameters in our proposed method (and in any dictionary learning algorithm in general) are the dictionary sizes $K_0$, $K_i$, signal sparsity levels $s_0$, $s_i$, and the incoherence penalty parameter $\eta$. Similar to most existing dictionary learning algorithms [25], [31], [32], there is no predefined way to find/calculate these parameters beforehand. In our experiments, we tried different combinations for these parameters and used the ones which gave us best overall results in terms of sum of the 6 atom-PTC correlation coefficients (as seen in Fig. 6 (a)). For the selection of dictionary size and sparsity levels, our aim was to learn most important temporal dynamics from the data instead of getting a small overall representation error. So, instead of learning big overcomplete dictionaries, we tested with dictionary size combinations of $K_0 \in [40, 50, 60, 70]$ and $K_i \in [15, 20, 25, 30]$ with signal sparsity parameters of $s_0 \in \{1, 2\}$ and $s_i \in \{1, 2, 3, 4\}$. We also experimented with different values for the incoherence penalty parameter $\eta \in \{0.5, 1, 5, 20, 100, 200, 500, 1500\}$. The Parameter values for ($K_0$, $K_i$, $s_0$, $s_i$, and, $\eta$) leading to the best results were found to be $(40, 20, 1, 2, 500)$ respectively. Furthermore, to investigate the effect of $\eta$ on atom-PTC correlations (keeping other parameters constant), we correlated $\mathbf{D}_0$ with all $\mathbf{D}_i$s for different values of $\eta$, summed up their

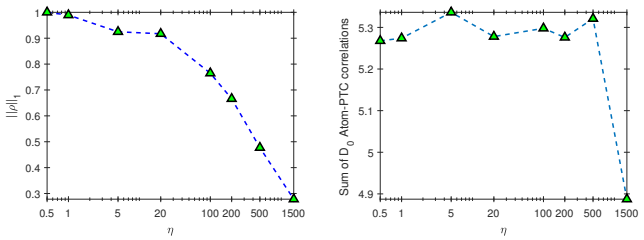
TABLE II
CORRELATION COEFFICIENTS OF MOST CORRELATED SPATIAL MAPS W.R.T. THE RSN TEMPLATES AS RECOVERED BY SHSSDL AND CODL

| RSN | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ShSSDL | 0.56 | 0.45 | **0.51** | 0.68 | **0.41** | **0.57** | **0.54** | **0.51** | **0.59** | **0.58** | **0.54** |
| CODL | **0.58** | **0.47** | 0.41 | **0.72** | 0.28 | 0.32 | 0.38 | 0.29 | 0.42 | 0.45 | 0.43 |



RSN 1 RSN 4 RSN 9 RSN 10

Fig. 8. Most informative population level z-scored ($p < 0.001$) activation maps. Top row: RSN Templates, Mid row: ShSSDL, Bottom row: CODL




(a) (b)

Fig. 9. Effects of different $\eta$ on (a) the overall normalized sum of $\mathbf{D}_0$ atom correlations with $\mathbf{D}_i$s, (b) sum of $\mathbf{D}_0$ atom-PTC correlation coefficients.

coefficients and have shown the normalized results in Fig. 9 a). We also summed up the $\mathbf{D}_0$ atom-PTC correlation coefficients for each $\eta$ and have shown the results in Fig. 9 b). Fig. 9 a) shows that for smaller values of $\eta$, sub-specific $\mathbf{D}_i s$ and $\mathbf{D}_0$ contain highly correlated atoms, i.e. all dictionaries try to explain same information from the data. Whereas our aim is to separate shared from sub-specific information. For higher values of $\eta$, the incoherence starts to increase while the atom-PTC correlation results show very slight variation. However, too much regularization ($\eta = 1500$) leads to too much bias, thus we chose to use $\eta = 500$ to achieve a balance between incoherence and correlation results. Another interesting observation was that for small values of $\eta$, we needed to learn dictionaries with much larger sizes to achieve good results. The reason could be that when we have smaller dictionaries with highly correlated atoms, most of the atoms try to explain the same information and ignore the remaining info. The naive solution to this problem will be to learn big dictionaries with more atoms to explain more trends present in the data. This problem, however, can be overcome by forcing the dictionaries to be incoherent. Thus, in our experiments, choosing larger values of $\eta$, the dictionary sizes had very small effect on the atom-PTC correlation results.

## VI. CONCLUSION

Dictionary learning algorithms have proven to be a successful alternative to conventional data driven methods such as ICA for fMRI data analysis. In this paper, we introduce a new dictionary learning algorithm named ShSSDL for multi-subject tfMRI data analysis. This model offers the advantage of accounting for both the shared as well as the distinct information among the group of subjects. Similar to existing algorithms, it is a two stages procedure with a sparse coding stage where both the shared and sub-specific sparse codes are estimated and a dictionary update stage where both the shared and the set of sub-specific dictionaries are updated. To illustrate its effectiveness, the proposed algorithm was tested on both simulated and real tfMRI datasets from Q1 release of the publicly available HCP motor task fMRI datasets [48]. Our experimental results demonstrate the proposed algorithms' proficiency in separating the shared TC/SM pairs and subject-specific ones into separate dictionary/sparse code pairs, thus providing an efficient way to analyze the shared and subject-specific TC/SMs for multi-subject tfMRI datasets.

## REFERENCES

[1] K. J. Friston, "Modalities, modes, and models in functional neuroimaging," *Science*, vol. 326, no. 5951, pp. 399–403, 2009.
[2] W. Du, V. D. Calhoun, H. Li *et al.*, "High classification accuracy for schizophrenia with rest and task fMRI data," *Frontiers in human neuroscience*, vol. 6, no. 145, 2012.
[3] K. J. Worsley, "An overview and some new developments in the statistical analysis of pet and fMRI data," *Human Brain Mapping*, vol. 5, no. 4, pp. 254–258, 1997.
[4] K. J. Friston, P. Fletcher, O. Josephs *et al.*, "Event-related fMRI: characterizing differential responses," *Neuroimage*, vol. 7, no. 1, pp. 30–40, 1998.
[5] C. F. Beckmann, M. Jenkinson, and S. M. Smith, "General multilevel linear modeling for group analysis in fMRI," *Neuroimage*, vol. 20, no. 2, pp. 1052–1063, 2003.
[6] M. A. Lindquist, "The statistical analysis of fMRI data," *Statistical Science*, vol. 23, no. 4, pp. 439 – 464, 2008.
[7] M. E. Raichle, "Two views of brain function," *Trends in cognitive sciences*, vol. 14, no. 4, pp. 180–190, 2010.

[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," *Magnetic Resonance Imaging*, vol. 17, no. 6, pp. 795–815, 1999.

[9] M. J. McKeown, T.-P. Jung, S. Makeig *et al.*, "Spatially independent activity patterns in functional mri data during the stroop color-naming task," *Proceedings of the National Academy of Sciences*, vol. 95, no. 3, pp. 803–810, 1998.

[10] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar, "A method for making group inferences from functional mri data using independent component analysis," *Human Brain Mapping*, vol. 14, no. 3, pp. 140–151, 2001.

[11] O. Friman, M. Borga, P. Lundberg, and H. Knutsson, "Exploratory fMRI analysis by autocorrelation maximization," *NeuroImage*, vol. 16, no. 2, pp. 454–464, 2002.

[12] Y. O. Li, T. Adali, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.

[13] V. D. Calhoun and T. Adali, "Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Reviews in Biomedical Engineering*, vol. 5, pp. 60–73, 2012.

[14] I. Daubechies, E. Roussos, S. Takerkart *et al.*, "Independent component analysis for brain fMRI does not select for independence," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 415–10 422, 2009.

[15] V. D. Calhoun, V. K. Potluru, R. Phlypo *et al.*, "Independent component analysis for brain fMRI does indeed select for maximal independence," *PloS one*, vol. 8, no. 8, p. e73309, 2013.

[16] M. J. McKeown, S. Makeig, G. G. Brown *et al.*, "Analysis of fMRI data by blind separation into independent spatial components," DTIC Document, Report, 1997.

[17] 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.

[18] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, *Online Learning for Matrix Factorization and Sparse Coding*. JMLR.org, 2010, pp. 19–60.

[20] A. K. Seghouane and M. Hanif, "A sequential dictionary learning algorithm with enforced sparsity," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, Conference Proceedings, pp. 3876–3880.

[21] L. Li, S. Li, and Y. Fu, "Learning low-rank and discriminative dictionary for image classification," *Image and Vision Computing*, vol. 32, no. 10, pp. 814–823, 2014.

[22] T. H. Vu and V. Monga, "Fast low-rank shared dictionary learning for image classification," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5160–5175, 2017.

[23] B. A. Olshausen, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[24] L. Yuanqing, Y. Zhu Liang, B. Ning *et al.*, "Sparse representation for brain signal processing: A tutorial on methods and applications," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 96–106, 2014.

[25] L. Kangjoo, T. Sungho, and Y. Jong Chul, "A data-driven sparse glm for fMRI analysis using sparse dictionary learning with mdl criterion," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 5, pp. 1076–1089, 2011.

[26] M. U. Khalid and A.-K. Seghouane, "Improving functional connectivity detection in fMRI by combining sparse dictionary learning and canonical correlation analysis," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*. IEEE, 2013, pp. 286–289.

[27] V. Abolghasemi, S. Ferdowsi, and S. Sanei, "Fast and incoherent dictionary learning algorithms with application to fMRI," *Signal, Image and Video Processing*, vol. 9, no. 1, pp. 147–158, 2015.

[28] A. K. Seghouane and A. Iqbal, "Sequential dictionary learning from correlated data: Application to fMRI data analysis," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 3002–3015, 2017.

[29] ——, "Basis expansions approaches for regularized sequential dictionary learning algorithms with enforced sparsity for fMRI data analysis," *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1796–1807, 2017.

[30] J. Lv, X. Jiang, X. Li *et al.*, "Sparse representation of whole-brain fMRI signals for identification of functional networks," *Medical Image Analysis*, vol. 20, no. 1, pp. 112–134, 2015.

[31] S. Zhang, X. Li, J. Lv *et al.*, "Sparse representation of higher-order functional interaction patterns in task-based fMRI data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 626–634.

[32] S. Zhao, J. Han, J. Lv *et al.*, "Supervised dictionary learning for inferring concurrent brain networks," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 2036–2045, 2015.

[33] G. Varoquaux, A. Gramfort, F. Pedregosa *et al.*, "Multi-subject dictionary learning to segment an atlas of brain spontaneous activity," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2011, pp. 562–573.

[34] M. U. Khalid and A. K. Seghouane, "Multi-subject fMRI connectivity analysis using sparse dictionary learning and multiset canonical correlation analysis," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2015, pp. 683–686.

[35] A. Mensch, G. Varoquaux, and B. Thirion, "Compressed online dictionary learning for fast resting-state fMRI decomposition," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 1282–1285.

[36] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 948–958, 2010.

[37] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 5. IEEE, 1999, pp. 2443–2446.

[38] K. Skretting and K. Engan, "Recursive least squares dictionary learning algorithm," *Signal Processing, IEEE Transactions on*, vol. 58, no. 4, pp. 2121–2130, 2010.

[39] K. K. Delgado, J. Murray, B. Rao *et al.*, "Dictionary learning algorithms for sparse representation," *Neural computation*, vol. 15, no. 2, pp. 349–396, 2003.

[40] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, pp. 337–365, 2000.

[41] S. K. Sahoo and A. Makur, "Dictionary training for sparse representation as generalization of k-means clustering," *Signal Processing Letters, IEEE*, vol. 20, no. 6, pp. 587–590, 2013.

[42] M. Svensen, F. Kruggel, and H. Benali, "ICA of fMRI group study data," *Neuroimage*, vol. 16, pp. 551–563, 2002.

[43] R. Gribonval and M. Nielsen, "Sparse decompositions in "incoherent" dictionaries," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 1. IEEE, 2003.

[44] S. Yubao, L. Qingshan, T. Jinhui, and T. Dacheng, "Learning discriminative dictionary for group sparse representation," *Image Processing, IEEE Transactions on*, vol. 23, no. 9, pp. 3816–3828, 2014.

[45] J. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4655–4666, 2007.

[46] S. Boyd, N. Parikh, E. Chu *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[47] J. Eckstein and W. Yao, "Augmented lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results," *RUTCOR Research Reports*, vol. 32, 2012.

[48] D. M. Barch, G. C. Burgess, M. P. Harms *et al.*, "Function in the human connectome: Task-fMRI and individual differences in behavior," *NeuroImage*, vol. 80, pp. 169–189, 2013.

[49] E. B. Erhardt, E. A. Allen, Y. Wei *et al.*, "SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability," *NeuroImage*, vol. 59, no. 4, pp. 4160–4167, 2012.

[50] S. M. Smith, P. T. Fox, K. L. Miller *et al.*, "Correspondence of the brain's functional architecture during activation and rest," *Proceedings of the National Academy of Sciences*, vol. 106, no. 31, pp. 13 040–13 045, 2009.

[51] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson *et al.*, "The minimal preprocessing pipelines for the human connectome project," *NeuroImage*, vol. 80, pp. 105–124, 2013.

[52] R. L. Buckner, F. M. Krienen, A. Castellanos *et al.*, "The organization of the human cerebellum estimated by intrinsic functional connectivity," *Journal of Neurophysiology*, vol. 106, no. 5, pp. 2322–2345, 2011.