Joint fMRI analysis and subject clustering using sparse dictionary learning

Seung-Jun Kim, Krishna K. Dontaraju

Seung-Jun Kim, Krishna K. Dontaraju, "Joint fMRI analysis and subject clustering using sparse dictionary learning," Proc. SPIE 10394, Wavelets and Sparsity XVII, 103940F (24 August 2017); doi: 10.1117/12.2273914

SPIE.

Event: SPIE Optical Engineering + Applications, 2017, San Diego, California, United States

Joint fMRI Analysis and Subject Clustering Using Sparse Dictionary Learning

Seung-Jun Kim and Krishna K. Dontaraju

Dept. of Computer Science and Electrical Engineering,
University of Maryland, Baltimore County, Baltimore, MD 21250, U.S.A.

ABSTRACT

Multi-subject fMRI data analysis methods based on sparse dictionary learning are proposed. In addition to identifying the component spatial maps by exploiting the sparsity of the maps, clusters of the subjects are learned by postulating that the fMRI volumes admit a subspace clustering structure. Furthermore, in order to tune the associated hyper-parameters systematically, a cross-validation strategy is developed based on entry-wise sampling of the fMRI dataset. Efficient algorithms for solving the proposed constrained dictionary learning formulations are developed. Numerical tests performed on synthetic fMRI data show promising results and provides insights into the proposed technique.

Keywords: Functional magnetic resonance imaging (fMRI), sparse dictionary learning, subspace clustering, multi-subject fMRI analysis.

1. INTRODUCTION

Functional magnetic resonance imaging (fMRI) provides a noninvasive tool for studying brain activities. The technique has found widespread uses in research and clinical settings since its inception. The most prevalent method acquires the blood oxygenation level dependent (BOLD) contrast, where the differences in the magnetic properties of oxygenated and deoxygenated hemoglobin are utilized to reveal the neural activities in the brain with a relatively high resolution.[1]

Analysis of fMRI data plays an important part in understanding the brain function. The most straightforward analysis methods are the region-of-interest (ROI)- or seed-based ones, where the fMRI time courses from predefined ROIs or voxels are correlated to obtain functional connectivity maps.[2] The alternative is the class of data-driven methods, of which achieving blind source separation through independent component analysis (ICA) is a major approach. These methods do not require prior knowledge on temporal structures of the data, and are fully multi-variate. More recently, a promising approach is to impose the notion of sparsity rather than independence to ensure identifiability, as in sparse dictionary learning.[3,4]

An important benefit of sparse dictionary learning approach is that the model allows an *overcomplete* representation, where more spatial components than the measurement dimension could be identified. Such a property is expected to be useful for large-scale studies, involving thousands of subjects, since it can uncover not only the components that are globally shared, but also fine-grained features that are unique to individuals and groups.[5–7]

In this work, a sparse dictionary learning-based multi-subject fMRI data analysis method is proposed. However, differently from the conventional dictionary learning analysis of fMRI data,[8,9] our goal is to jointly uncover the clusters in the subjects based on the different spatial activation patterns in the voxel space. For this, it is postulated that individual fMRI volumes per time point belong to subspace clusters.[10] Thus, appropriate regularizers motivated by subspace clustering objectives are imposed on the learned dictionary, in addition to the usual sparsity-promoting regularization of the spatial maps. Specifically, the low-rank and group-sparsity regularizers are employed.[11,12]

Further author information: (Send correspondence to S.J.K.)
S.J.K.: E-mail: sjkim@umbc.edu, Telephone: +1 410 455 3372
K.K.D.: E-mail: krishna5@umbc.edu
This work was supported in part by NSF grant 1631838.

Wavelets and Sparsity XVII, edited by Yue M. Lu, Dimitri Van De Ville, Manos Papadakis, Proc. of SPIE
Vol. 10394, 103940F · © 2017 SPIE · CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2273914

Dictionary learning was shown to be effective for multi-subject fMRI data analysis.[9,13,14] One approach introduced common spatial maps shared across the subjects.[9] Group sparse dictionary learning was used for fMRI-based group study of the healthy and Alzheimer's disease subjects,[15] where the given group information was used. A hierarchical structured sparsity regularizer was employed for a supervised learning task involving fMRI data in Ref. 16. Low-rank dictionary learning was shown to be helpful for a face recognition task in Ref. 17.

An important practical issue with dictionary learning is to tune the hyper-parameters. Unfortunately, it seems there is no golden criterion and the hyper-parameters are often adjusted heuristically to see meaningful outcomes from the analysis.[18] In order to mitigate this issue, we also develop a cross-validation-based strategy using the prediction error of hidden measurements within the proposed dictionary learning framework.

The rest of the paper is organized as follows. In Section 2, the joint dictionary learning and subspace clustering formulation is presented. Then an efficient solution method is derived in Section 3. A hyper-parameter tuning strategy based on prediction error of missing measurements is proposed in Section 4. The results from numerical tests using synthetic fMRI data are presented in Section 5. Conclusions and future directions are discussed in Section 6.

2. PROBLEM FORMULATION

Consider fMRI scans for M subjects denoted by $\mathbf{X}_m \in \mathbb{R}^{T\times V}$, $m = 1, 2, \ldots, M$, representing T time points and V voxels. Matrices $\{\mathbf{X}_m\}$ are then vertically stacked in $\mathbf{X} := [\mathbf{X}_1^\top, \mathbf{X}_2^\top, \ldots, \mathbf{X}_M^\top]^\top \in \mathbb{R}^{MT\times V}$, where $^\top$ denotes transposition. Sparse dictionary learning is employed to factorize $\mathbf{X}$ into $\mathbf{D} \in \mathbb{R}^{MT\times K}$ and $\mathbf{S} \in \mathbb{R}^{V\times K}$ such that

$$\mathbf{X} \approx \mathbf{D}\mathbf{S}^\top. \tag{1}$$

In particular, we want to enforce $\mathbf{S}$ to be sparse and $\mathbf{D}$ to have some structure that reveals subject groups. Encouraging sparsity in $\mathbf{S}$ has well-documented merits such as capturing union-of-subspace structures and allowing model identifiability. In the fMRI analysis context, sparsity in $\mathbf{S}$ exposes sparse brain activations in the voxel space, referred to as the spatial maps. Sparse dictionary learning has recently been adopted in fMRI data analysis for single- and multiple-subject datasets with promising results.[8,9,19]

A standard formulation for sparse dictionary learning is given as[20]

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{S}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{S}^\top\|_F^2 + \lambda\|\mathbf{S}\|_1 \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\|\mathbf{S}\|_1 := \sum_{v,k} |S_{vk}|$ with S_{vk} representing the (v,k)-entry of $\mathbf{S}$, which promotes sparsity in $\mathbf{S}$. The set $\mathcal{D}$ is defined as

$$\mathcal{D} := \{\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K] \in \mathbb{R}^{MT\times K} : \|\mathbf{d}_k\|_2 \le 1\}. \tag{3}$$

That is, the columns of $\mathbf{D} \in \mathcal{D}$ are normalized to have the Euclidean norms no larger than unity. Constraining $\mathbf{D}$ in this way ensures that the individual spatial maps have comparable support sizes, and makes the search range for λ more consistent in practice.

The goal of this work is to uncover the subject clusters at the same time as the component spatial maps are estimated. The underlying assumption is that the multi-subject fMRI data admit a subspace clustering structure. That is, the fMRI volumes (the rows of $\mathbf{X}$) are viewed as points in a union of subspaces, with the spatial maps (the columns of $\mathbf{S}$) being the (potentially redundant) bases for the subspaces, and the rows of $\mathbf{D}$ the coefficient vectors. There are various ways to encode such a structure, including those based on the low rank and sparsity.[12,21]

2.1 Dictionary Learning with Low Rank Dictionary

The subspace clustering based on the low rank encourages the coefficient matrix $\mathbf{D}$ to have a small rank. A formulation studied in Ref. 11 is

$$\min_{\mathbf{D}} \|\mathbf{D}\|_* \text{ subject to } \mathbf{X} = \mathbf{D}\mathbf{S}^\top \tag{4}$$

where $\|\mathbf{D}\|_*$ is the nuclear norm of $\mathbf{D}$ (sum of singular values of $\mathbf{D}$) for encouraging a low rank in $\mathbf{D}$. It can be shown that if the subspaces are independent and $\mathbf{S}$ collects the (redundant) bases of all the subspaces as its columns, then an optimal solution $\mathbf{D}^*$ to (4) possesses the property that the (t,k)-entry of $\mathbf{D}$ is zero if the t-th row of $\mathbf{X}$ and the k-th spatial map in $\mathbf{S}$ belong to different subspaces. Thus, this property can be used to perform clustering of the fMRI volumes and hence the subjects.

Another interpretation follows by noting that a low-rank $\mathbf{D}$ would allow $\mathbf{D}$ to be further factorized as $\mathbf{D} = \mathbf{AB}$, and appreciating that this model is related to matrix tri-factorization $\mathbf{X} \approx \mathbf{ABS}^\top$. Regularized matrix tri-factorizations can simultaneously cluster rows and columns (times and voxels in the present application) via $\mathbf{A}$ and $\mathbf{S}$, while revealing the connection between the clusters via $\mathbf{B}$.[22]

To perform the sparse dictionary learning with a low-rank dictionary, we use the following formulation.

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{S}} \frac{1}{2}\|\mathbf{X} - \mathbf{DS}^\top\|_F^2 + \lambda\|\mathbf{S}\|_1 + \mu\|\mathbf{D}\|_*. \tag{5}$$

2.2 Dictionary Learning with Group Sparse Dictionary

The subspace clustering structure can be encoded by promoting sparsity on $\mathbf{D}$ as well. The key idea is that each fMRI volume per time point is explained by a linear combination of a small number of spatial maps. Reducing the number of involved spatial maps by sparsity in the coefficient matrix $\mathbf{D}$ ultimately selects the correct set of bases for the subspace to which the fMRI volumes belong. The *group* sparsity is then motivated by the expectation that the set of bases selected are common across the volumes corresponding to a single subject.

Thus, the optimization problem for sparse dictionary learning with a group sparse dictionary can be formulated as

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{S}} \frac{1}{2}\|\mathbf{X} - \mathbf{DS}^\top\|_F^2 + \lambda\|\mathbf{S}\|_1 + \mu\|\mathbf{D}\|_{\mathcal{G}} \tag{6}$$

where $\|\mathbf{D}\|_{\mathcal{G}}$ is defined for $\mathbf{D} = [\mathbf{D}_1^\top, \mathbf{D}_2^\top, \ldots, \mathbf{D}_M^\top]^\top$ with $\mathbf{D}_m = [\mathbf{d}_{m,1}, \mathbf{d}_{m,2}, \ldots, \mathbf{d}_{m,K}] \in \mathbb{R}^{T\times K}$, $m = 1, 2, \ldots, M$, as

$$\|\mathbf{D}\|_{\mathcal{G}} := \sum_{m=1}^{M}\sum_{k=1}^{K} \|\mathbf{d}_{m,k}\|_2. \tag{7}$$

Thus, the penalty term $\mu\|\mathbf{D}\|_{\mathcal{G}}$ in (6) promotes that a column in the dictionary $\mathbf{D}_m$ corresponding to subject m is entirely zero. If $\mathbf{d}_{m,k}$ is entirely zero, subject m's fMRI volumes do not include the k-th spatial map. In other words, $\mathbf{X}_m$ belongs to the subspace spanned by the spatial maps corresponding to the nonzero columns in $\mathbf{D}_m$, revealing a subject-wise subspace clustering structure.

3. ALGORITHM DERIVATION

As (5) and (6) are nonconvex optimization problems, it is difficult to obtain the globally optimal solutions. However, with $\mathbf{D}$ fixed, the optimization w.r.t. $\mathbf{S}$ is the well-known Lasso problem. Likewise, with $\mathbf{S}$ fixed, the problem for $\mathbf{D}$ is convex. Thus, an alternating minimization algorithm can be developed.

Specifically, to update $\mathbf{S}$, one solves

$$\min_{\mathbf{S}} \left[\frac{1}{2}\|\mathbf{X} - \mathbf{DS}^\top\|_F^2 + \lambda\|\mathbf{S}\|_1 = \sum_{v=1}^{V}\left(\frac{1}{2}\|\mathbf{x}_v - \mathbf{D}\mathbf{s}_v\|_2^2 + \lambda\|\mathbf{s}_v\|_1\right)\right] \tag{8}$$

using existing Lasso problem solvers,[23–25] where $\mathbf{x}_v$ is the v-th column of $\mathbf{X}$ and $\mathbf{s}_v$ is the transpose of the v-th row of $\mathbf{S}$.

In order to update $\mathbf{D}$, one needs to solve a constrained optimization problem for $\mathbf{D}$. For example, the low-rank dictionary is obtained by

$$\min_{\mathbf{D}\in\mathcal{D}} \frac{1}{2}\|\mathbf{X} - \mathbf{DS}^\top\|_F^2 + \mu\|\mathbf{D}\|_*. \tag{9}$$

To solve this efficiently, one can employ the alternating direction method of multipliers (ADMM).[26] Introduce a copy $\mathbf{D}'$ of $\mathbf{D}$ to write (9) equivalently as

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{D}'} \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{S}^\top\|_F^2+\mu\|\mathbf{D}'\|_* \tag{10}$$

$$\text{subject to } \mathbf{D}=\mathbf{D}'. \tag{11}$$

Then, the augmented Lagrangian is formed as

$$L(\mathbf{D},\mathbf{D}';\mathbf{\Lambda}):=\frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{S}^\top\|_F^2+\mu\|\mathbf{D}'\|_*+\text{tr}\left(\mathbf{\Lambda}^\top(\mathbf{D}-\mathbf{D}')\right)+\frac{\rho}{2}\|\mathbf{D}-\mathbf{D}'\|_F^2 \tag{12}$$

with $\rho>0$. The strategies for adjusting ρ even on the fly are discussed in Ref. 26. The standard ADMM algorithm iterates the updates given by

$$\mathbf{D}^{(n+1)}=\arg\min_{\mathbf{D}\in\mathcal{D}} L(\mathbf{D},\mathbf{D}'^{(n)};\mathbf{\Lambda}^{(n)}) \tag{13}$$

$$\mathbf{D}'^{(n+1)}=\arg\min_{\mathbf{D}'} L(\mathbf{D}^{(n+1)},\mathbf{D}';\mathbf{\Lambda}^{(n)}) \tag{14}$$

$$\mathbf{\Lambda}^{(n+1)}=\mathbf{\Lambda}^{(n)}+\rho\left(\mathbf{D}^{(n+1)}-\mathbf{D}'^{(n+1)}\right) \tag{15}$$

where n is the iteration index. Defining $\mathbf{U}:=\mathbf{\Lambda}/\rho$, and plugging (12) into (13)–(15) yields

$$\mathbf{D}^{(n+1)}=\arg\min_{\mathbf{D}\in\mathcal{D}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{S}^\top\|_F^2+\frac{\rho}{2}\|\mathbf{D}-\mathbf{D}'^{(n)}+\mathbf{U}^{(n)}\|_F^2 \tag{16}$$

$$\mathbf{D}'^{(n+1)}=\arg\min_{\mathbf{D}'} \frac{\mu}{\rho}\|\mathbf{D}'\|_*+\frac{1}{2}\|\mathbf{D}^{(n+1)}-\mathbf{D}'+\mathbf{U}^{(n)}\|_F^2 \tag{17}$$

$$\mathbf{U}^{(n+1)}=\mathbf{U}^{(n)}+\mathbf{D}^{(n+1)}-\mathbf{D}'^{(n+1)}. \tag{18}$$

Problem in (16) can be solved via block coordinate descent (BCD) with the blocks being the columns in $\mathbf{D}$. It is first noted that upon defining

$$\mathbf{A}:=\mathbf{S}^\top\mathbf{S}+\rho\mathbf{I} \tag{19}$$

$$\mathbf{B}:=\mathbf{X}\mathbf{S}+\rho\left(\mathbf{D}'^{(n)}-\mathbf{U}^{(n)}\right) \tag{20}$$

update (16) is equivalent to

$$\mathbf{D}^{(n+1)}=\arg\min_{\mathbf{D}\in\mathcal{D}} \frac{1}{2}\text{tr}\left(\mathbf{D}^\top\mathbf{D}\mathbf{A}\right)-\text{tr}\left(\mathbf{D}^\top\mathbf{B}\right). \tag{21}$$

Thus, the BCD update for $\mathbf{d}_k\in\mathbb{R}^{MT}$ is done by iteratively (over ℓ) updating[20]

$$\tilde{\mathbf{d}}_k=\frac{1}{A_{kk}}\left(\mathbf{b}_k-\mathbf{D}_k^{(\ell)}\mathbf{a}_k\right)+\mathbf{d}_k^{(\ell)} \tag{22}$$

$$\mathbf{d}_k^{(\ell+1)}=\frac{1}{\max\{\|\tilde{\mathbf{d}}_k\|_2,1\}}\tilde{\mathbf{d}}_k \tag{23}$$

sequentially for $k=1,2,\ldots,K$, where A_{kk} is the (k,k)-entry of $\mathbf{A}$, $\mathbf{a}_k$ and $\mathbf{b}_k$ are the k-th columns of $\mathbf{A}$ and $\mathbf{B}$, respectively, and $\mathbf{D}_k^{(\ell)}:=[\mathbf{d}_1^{(\ell+1)},\ldots,\mathbf{d}_{k-1}^{(\ell+1)},\mathbf{d}_k^{(\ell)},\ldots,\mathbf{d}_K^{(\ell)}]$.

Problem in (17) is the simple proximal operator associated with the nuclear norm penalty, which has a closed-form solution.[27] Upon denoting the singular value decomposition (SVD) of $\tilde{\mathbf{D}}^{(n+1)}:=\mathbf{D}^{(n+1)}+\mathbf{U}^{(n)}=\mathbf{V}\mathbf{\Sigma}\mathbf{W}^\top$, where $\mathbf{V}$ and $\mathbf{W}$ are orthonormal and $\mathbf{\Sigma}$ is nonnegative, and defining $S_\tau^+(x):=\max\{0,x-\tau\}$,

$$\mathbf{D}'^{(n+1)}=\mathbf{V}S_{\mu/\rho}^+(\mathbf{\Sigma})\mathbf{W}^\top \tag{24}$$

Input: $\mathbf{X}$, $\lambda > 0$, $\mu \geq 0$, and $\rho > 0$	
Output: $\mathbf{D}$ and $\mathbf{S}$	
1:	Initialize $\mathbf{D}$
2:	Repeat
	Sparse coding:
3:	$\mathbf{S} = \arg\min_{\mathbf{S}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{S}^\top\|_F^2 + \lambda\|\mathbf{S}\|_1$
	Dictionary update:
4:	For $n = 1, 2, \ldots$
	Without hidden entries:
5:	Compute $\mathbf{D}^{(n+1)}$ by solving (16) via iterating (22)–(23)
	With hidden entries:
5:	Compute $\mathbf{D}^{(n+1)}$ by solving (31) via iterating (35)–(36)
	For a low-rank dictionary:
6:	$\mathbf{D}'^{(n+1)} = \mathbf{V} S^+_{\mu/\rho}(\mathbf{\Sigma})\mathbf{W}^\top$
	For a group-sparse dictionary:
6:	$\mathbf{d}'^{(n+1)}_{m,k} = \max\{\|\tilde{\mathbf{d}}_{m,k}\|_2 - \frac{\mu}{\rho}, 0\}\frac{\tilde{\mathbf{d}}_{m,k}}{\|\tilde{\mathbf{d}}_{m,k}\|_2},\ \forall m, k$
7:	$\mathbf{U}^{(n+1)} = \mathbf{U}^{(n)} + \mathbf{D}^{(n+1)} - \mathbf{D}'^{(n+1)}$
8:	Next n
9:	$\mathbf{D} = \mathbf{D}^{(\infty)}$
10:	Until convergence

Table 1: Overall algorithm.

where $S^+_{\mu/\rho}(\cdot)$ is applied entry-wise.

For the case of group sparse dictionary in (6), the update step in (17) must be replaced by

$$\mathbf{D}'^{(n+1)} = \arg\min_{\mathbf{D}'} \frac{\mu}{\rho}\|\mathbf{D}'\|_{\mathcal{G}} + \frac{1}{2}\|\mathbf{D}^{(n+1)} - \mathbf{D}' + \mathbf{U}^{(n)}\|_F^2 \tag{25}$$

which also admits a closed-form solution.[28] First define $[\tilde{\mathbf{D}}_1^\top, \ldots, \tilde{\mathbf{D}}_M^\top]^\top := \tilde{\mathbf{D}}^{(n+1)}$, where $\tilde{\mathbf{D}}_m = [\tilde{\mathbf{d}}_{m,1}, \ldots, \tilde{\mathbf{d}}_{m,K}] \in \mathbb{R}^{T\times K}$ for $m = 1, 2, \ldots, M$. Likewise, define $\mathbf{d}'^{(n+1)}_{m,k}$ from $\mathbf{D}'^{(n+1)}$. Then,

$$\mathbf{d}'^{(n+1)}_{m,k} = \tilde{\mathbf{d}}_{m,k} - \Pi_{\mu/\rho}\left(\tilde{\mathbf{d}}_{m,k}\right) = \begin{cases} \frac{\|\tilde{\mathbf{d}}_{m,k}\|_2 - \mu/\rho}{\|\tilde{\mathbf{d}}_{m,k}\|_2}\tilde{\mathbf{d}}_{m,k}, & \text{if } \|\tilde{\mathbf{d}}_{m,k}\|_2 > \frac{\mu}{\rho} \\ 0, & \text{otherwise} \end{cases} \tag{26}$$

where Π_τ represents the projection to an ℓ_2-ball of radius τ.

The overall algorithm is given in Table 1.

4. HYPER-PARAMETER TUNING VIA MISSING ENTRY PREDICTION

An important practical issue for dictionary learning-based fMRI data analysis is to tune the hyper-parameters like λ and μ. For supervised learning problems, hyper-parameters are tuned by cross-validation, where the classifiers are learned based on a training dataset and hyper-parameters are adjusted based on the classification error of a validation set. Since fMRI data analysis is an unsupervised learning task, instead of using the classification error, one could use the reconstruction error as the tuning signal. That is, the time courses and spatial maps learned from training data are employed for the reconstruction of unseen data.[29] However, it is not clear how to split the dataset into the training set and validation set.

In this section, we develop a systematic cross-validation strategy for fMRI data analysis using dictionary learning. The idea is to hide portion of the entries in the fMRI scans and estimate the dictionary and the spatial

maps using only the remaining entries. Then, the hidden entries are predicted based on the estimated dictionary and spatial maps, and the prediction error is used for tuning the hyper-parameters.

Let Ω be the set of indices for the entries in $\mathbf{X}$ that are observed (i.e., not hidden). Define the projection of $\mathbf{X} \in \mathbb{R}^{MT \times K}$, denoted as $\mathcal{P}_\Omega(\mathbf{X}) \in \mathbb{R}^{MT \times K}$, as

$$[\mathcal{P}_\Omega(\mathbf{X})]_{tk} := \begin{cases} X_{tk}, & \text{if } (t,k) \in \Omega \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

where $[\cdot]_{tk}$ represents the (t,k)-entry. Then dictionary learning with missing entries is formulated as

$$\min_{\mathbf{D} \in \mathcal{D}, \mathbf{S}} \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{X} - \mathbf{D}\mathbf{S}^\top)\|_F^2 + \lambda \|\mathbf{S}\|_1 + \mu \Psi(\mathbf{D}) \tag{28}$$

where $\Psi(\mathbf{D})$ is equal either to $\|\mathbf{D}\|_*$ or $\|\mathbf{D}\|_{\mathcal{G}}$. Once $\hat{\mathbf{D}}$ and $\hat{\mathbf{S}}$ are learned from (28), the root-mean-square prediction error (RMSE) for the hidden entries can be computed as

$$\text{RMSE} = \sqrt{\|\mathcal{P}_{\Omega^c}(\mathbf{X} - \hat{\mathbf{D}}\hat{\mathbf{S}}^\top)\|_F^2 / |\Omega^c|}. \tag{29}$$

Next, the algorithm for solving (28) is derived. Again employing an alternating minimization strategy, the sparse coding step is given by

$$\mathbf{s}_v := \arg\min_{\mathbf{s}} \frac{1}{2} \|\mathbf{x}_v|_{\Omega_v} - \mathbf{D}|_{\Omega_v} \mathbf{s}\|_2^2 + \lambda \|\mathbf{s}\|_1, \ v = 1, 2, \ldots, V \tag{30}$$

where $\Omega_v := \{t : (t,v) \in \Omega\}$ and $\mathbf{x}_v|_{\Omega_v} \in \mathbb{R}^{|\Omega_v|}$ is the vector collecting the entries of $\mathbf{x}_v$ whose indices belong to Ω_v. Likewise, $\mathbf{D}|_{\Omega_v}$ is the sub-dictionary formed by taking only the subset of rows of $\mathbf{D}$.

The derivation for the dictionary update rule is analogous to the derivation in Sec. 3. One only needs to replace (16) by

$$\mathbf{D}^{(n+1)} = \arg\min_{\mathbf{D} \in \mathcal{D}} \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{X} - \mathbf{D}\mathbf{S}^\top)\|_F^2 + \frac{\rho}{2} \|\mathbf{D} - \mathbf{D}'^{(n)} + \mathbf{U}^{(n)}\|_F^2. \tag{31}$$

This can be again solved using BCD. For this, define

$$\mathbf{A}_t := \sum_{v=1}^{V} \mathbb{1}_{\{t \in \Omega_v\}} \mathbf{s}_v \mathbf{s}_v^\top + \rho \mathbf{I}, \ t = 1, 2, \ldots, MT \tag{32}$$

$$\mathbf{B}' := \mathcal{P}_\Omega(\mathbf{X})\mathbf{S} + \rho \left(\mathbf{D}'^{(n)} - \mathbf{U}^{(n)} \right) \tag{33}$$

$$\mathbf{\Phi}_{kk'} := \text{diag}([A_{1,kk'}, A_{2,kk'}, \ldots, A_{MT,kk'}]) \tag{34}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator for the condition $\{\cdot\}$, yielding 1 if the condition is satisfied, and 0 if not. Then the solution to (31) can be obtained by iteratively performing the updates

$$\tilde{\mathbf{d}}_k = \mathbf{\Phi}_{kk}^{-1} \left(\mathbf{b}'_k - \sum_{k'=1}^{k-1} \mathbf{\Phi}_{kk'} \mathbf{d}_{k'}^{(\ell+1)} - \sum_{k'=k+1}^{K} \mathbf{\Phi}_{kk'} \mathbf{d}_{k'}^{(\ell)} \right) \tag{35}$$

$$\mathbf{d}_k^{(\ell+1)} = \frac{1}{\max\{\|\tilde{\mathbf{d}}_k\|_2, 1\}} \tilde{\mathbf{d}}_k \tag{36}$$

sequentially for $k = 1, 2, \ldots, K$, until convergence. The overall algorithm is again listed in Table 1.

5. NUMERICAL TESTS

In this section, the results from preliminary tests of the proposed methods on synthetic fMRI datasets are reported.

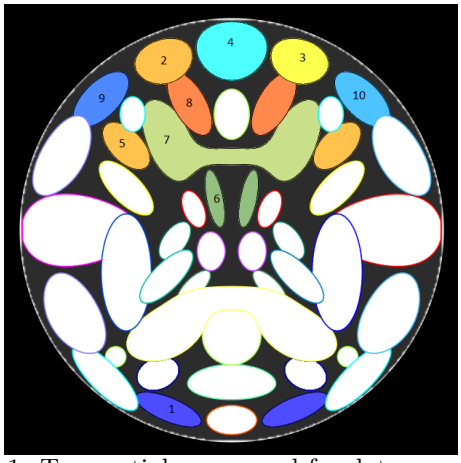
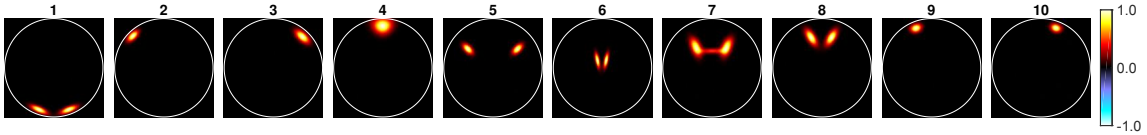


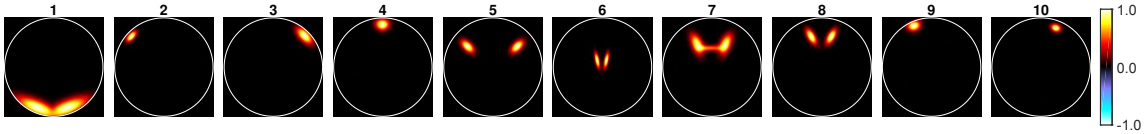
Figure 1: Ten spatial maps used for data generation.

Table 2: Parameter choice in SimTB.

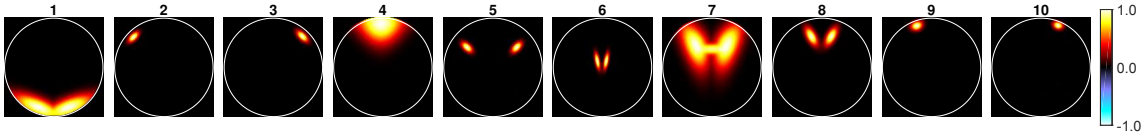
Dataset	Group	Component	Translation	Rotation	Size
1	1	All (1-10)	$\mathcal{N}(0,2)$	$\mathcal{N}(0,3)$	$\mathcal{U}(0.3,1)$
	2	1	$\mathcal{N}(0,2)$	$\mathcal{N}(0,3)$	$\mathcal{U}(0.3,8.3)$
		2-10	$\mathcal{N}(0,2)$	$\mathcal{N}(0,3)$	$\mathcal{U}(0.3,1)$
2	1	All (1-10)	$\mathcal{N}(0,2)$	$\mathcal{N}(0,3)$	$\mathcal{U}(0.3,1)$
	2	1,4	$\mathcal{N}(0,2)$	$\mathcal{N}(0,3)$	$\mathcal{U}(5.3,8.3)$
		7	$\mathcal{N}(0,2)$	$\mathcal{N}(0,3)$	$\mathcal{U}(0.3,8.3)$
		2,3,5,6,8-10	$\mathcal{N}(0,2)$	$\mathcal{N}(0,3)$	$\mathcal{U}(0.3,1)$



(a) Group 1 in Datasets 1 or 2.

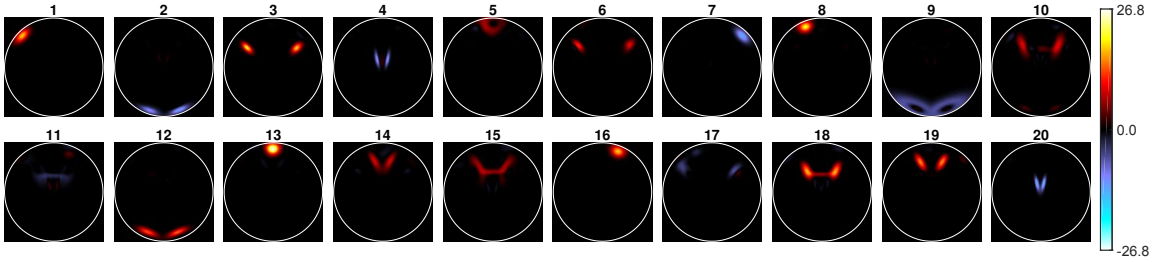


(b) Group 2 in Dataset 1.

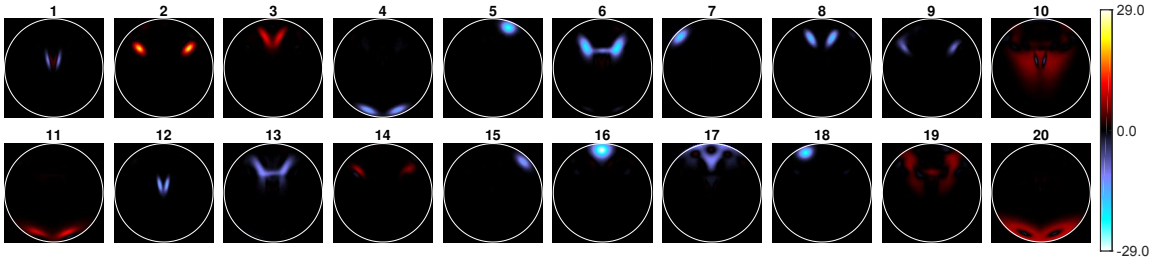


(c) Group 2 in Dataset 2.

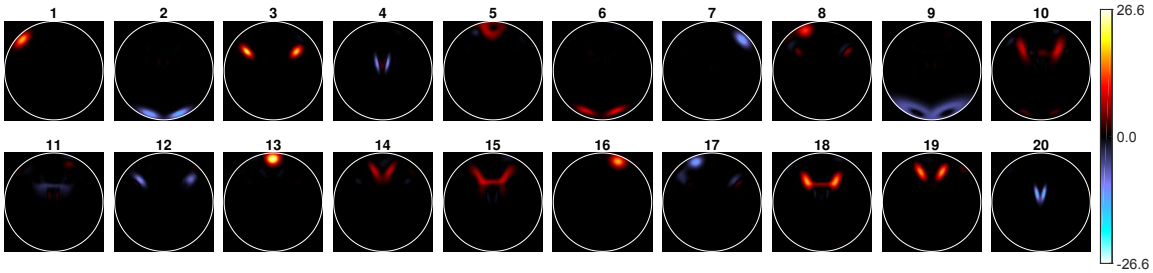
Figure 2: Representative spatial maps generated by SimTB.



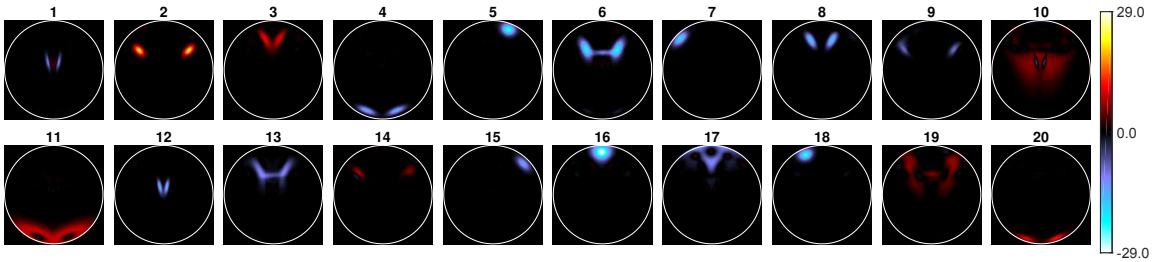
(a) Low-rank, Dataset 1



(b) Low-rank, Dataset 2



(c) Group-sparse, Dataset 1



(d) Group-sparse, Dataset 2

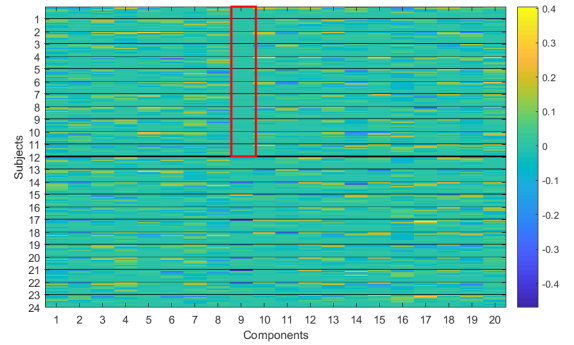
Figure 3: Spatial maps learned from dictionary learning.

Table 3: Tuned λ and μ values and RMSE

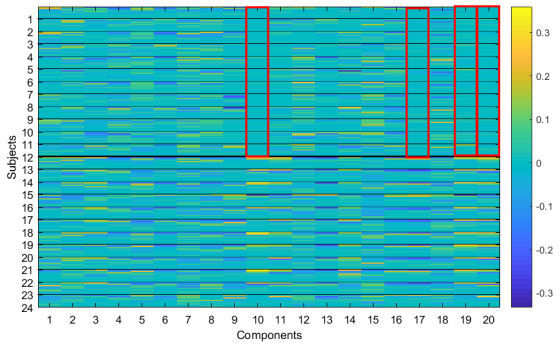
Algorithm	Dataset	λ^*	μ^*	RMSE
Low-rank	1	0.1	1	0.0816
	2	0.1	1	0.0833
Group-sparse	1	0.1	10	0.0816
	2	0.1	1	0.0742

5.1 Dataset Generation

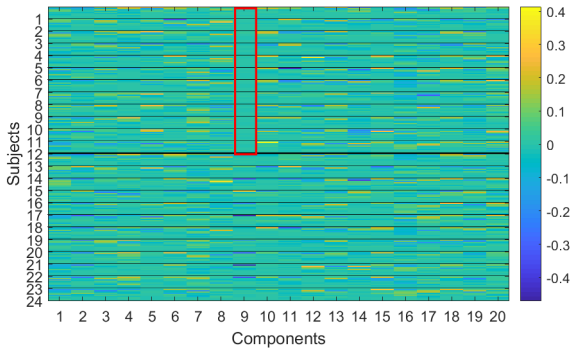
Two sets of fMRI data have been generated using fMRI simulator called SimTB.[30] The simulator creates 1000 fMRI volumes of 148×148 (2D) voxels per subject, by mixing synthetic spatial maps and time courses. For simplicity, we chose 10 spatial components out of 30 that SimTB is capable of generating, as shown in Fig. 1. The parameters for the spatial maps were manipulated to create two distinct groups in each dataset. The manipulated parameters included translation, rotation, and the size of the components. Compared to Group 1, Group 2 was generated with larger variations in the spatial map sizes. In Dataset 1, only one spatial map component was made different between the groups. In Dataset 2, there were three varied components. Each group consisted of 12 subjects. The precise parameters used for data generation is listed in Table 2, where $\mathcal{N}(a, b)$ represents Gaussian distribution with mean a and variance b, and $\mathcal{U}(a, b)$ represents the uniform distribution over interval $[a, b]$. The BOLD model for the time course generation was set as the canonical hemodynamic response function with the event trial experimental design. Five event trial types were used with the amplitude of events sampled from $\mathcal{U}(0.5, 1.5)$. Rician noise with the contrast-to-noise ratio of 2 was added. Fig. 2 depicts the representative component spatial maps for Group 1 and Group 2 in Datasets 1 and 2. It can be seen that component 1 in Fig. 2(b) and components 1, 4 and 7 in Fig. 2(c) have larger sizes than the corresponding components in Fig. 2(a).



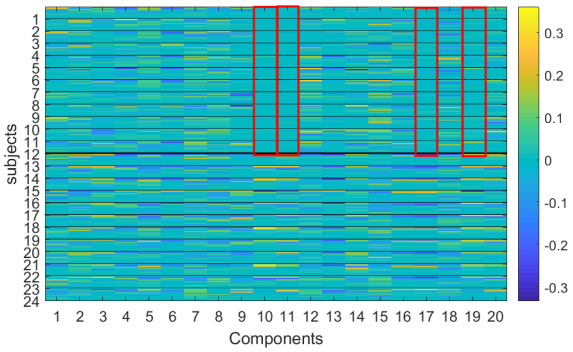
(a) Low-rank, Dataset 1



(b) Low-rank, Dataset 2



(c) Group-sparse, Dataset 1



(d) Group-sparse, Dataset 2

Figure 4: Learned dictionaries.

5.2 Test Results

The proposed dictionary learning algorithms were tested on the synthetic data generated. First, PCA was employed to reduce the temporal dimension to $T = 10$ per subject, and a spatial mask was applied to process only the actual brain voxels with size $V = 21,904$. Then, 24 subjects' datasets (12 per group) were concatenated to form a data matrix $\mathbf{X} \in \mathbb{R}^{240\times 21904}$. In order to tune the hyper-parameters λ and μ, 10% of the entries in $\mathbf{X}$ were uniformly sampled and hidden. Then, the hidden entries were predicted as explained in Section 4, using combinations of $\lambda \in \{0, 0.01, 0.1, 0.5, 1, 5\}$ and $\mu \in \{0, 0.1, 1, 10, 100, 1000\}$. The λ and μ values that yielded the minimum RMSE in (29) were then chosen, as shown in Table 3. The RMSE was averaged over 10 independent samplings. $K = 20$ components were learned in all experiments.

Fig. 3 depicts the learned component spatial maps using low-rank and group-sparse dictionary learning algorithms based on Datasets 1 and 2. Fig. 3(a) shows that the low-rank dictionary learning algorithm applied to Dataset 1 could identify all 10 ground truth components of Group 1 shown in Fig. 2(a) as components 1, 3, 7, 8, 12, 13, 16, and 18-20. In addition, various components that capture the individual variability have been identified, such as components 2, 4-6, 9-11, 14, 15 and 17. In particular, component 9 is clearly seen as the component that can discriminate Group 1 and Group 2 subjects. That is, the absence of component 9 can be used as an indication of Group 1 subjects. Very similar spatial maps were obtained by using the group-sparse dictionary learning algorithm as shown in Fig. 3(c). Also, a similar observation can be made for the spatial maps learned from Dataset 2, shown in Figs. 3(b) and (d) for low-rank and group-sparse dictionary learning, respectively. In particular, components 10, 17, 19, and 20 in Fig. 3(b) are seen to contribute the group discrimination, and components 10, 11, 17 and 19 in Fig. 3(d).

Such discriminative components can be readily identified by looking at the learned dictionary $\mathbf{D}$. Fig. 4(a) depicts the 240×20 matrix $\mathbf{D}$ learned from Dataset 1 using low-rank dictionary learning. The first 120 rows correspond to the subjects in Group 1 and the next 120 rows to Group 2. It can be seen that the entries in the 9-th column for Group 1 subjects (indicated by the red box) are identically zero. The 9-th column is the coefficient vector for the 9-th spatial map in Fig. 3(a), which discriminates Group 1 and Group 2. A similar dictionary was obtained using group-sparse dictionary learning as shown in Fig. 4(c). For Dataset 2, more discriminative components can be identified by looking for all-zero columns in the dictionaries for Group 1 subjects, as depicted as the red boxes in Figs. 4(b) and (d).

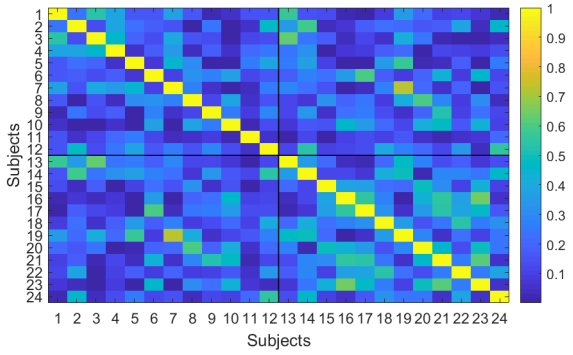
In order to see the feasibility of performing subject clustering based on the learned dictionary, an affinity matrix was constructed as the correlation coefficients of the dictionary columns corresponding to the individual subjects. That is, the (i, j)-entry of the affinity matrix $\mathbf{R}$ is computed as

$$r_{ij} := \frac{|\mathrm{tr}(\mathbf{D}_i^\top \mathbf{D}_j)|}{\|\mathbf{D}_i\|_F \|\mathbf{D}_j\|_F}. \tag{37}$$

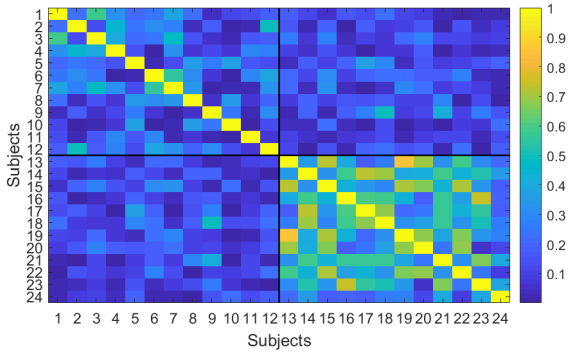
The affinity matrix can be input to a graph clustering algorithm to obtain the clusters. It can be seen from Figs. 5(b) and (d) that the affinity matrices obtained for Dataset 2 exhibits high connectivity among the subjects in Group 2. Somewhat weaker connectivity can be seen in Figs. 5(a) and (c) based on Dataset 1.

6. CONCLUSIONS AND FUTURE DIRECTIONS

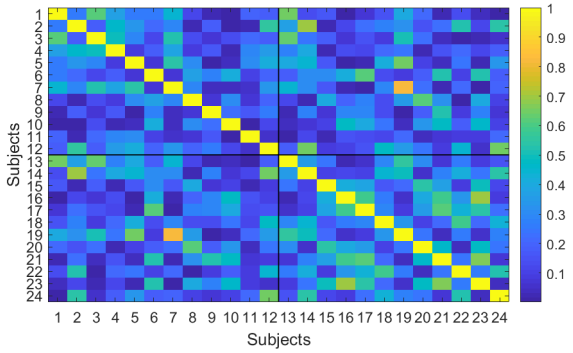
Sparse dictionary learning-based multi-subject fMRI data analysis methods have been developed. In addition to identifying the underlying functional connectivity maps based on the sparsity prior imposed on the spatial map factor, the subject clustering structure was learned by constraining the dictionary to be low-rank or group-sparse. In order to systematically fix the hyper-parameters via cross-validation, entry-wise random sampling of the fMRI dataset and prediction were performed. Efficient algorithms for solving the proposed dictionary learning formulations were also derived. Preliminary tests on synthetic fMRI datasets showed promising results. As the future works, we will test the proposed methods on real data. Furthermore, online and decentralized algorithms for fast analysis of large-scale datasets will be pursued.



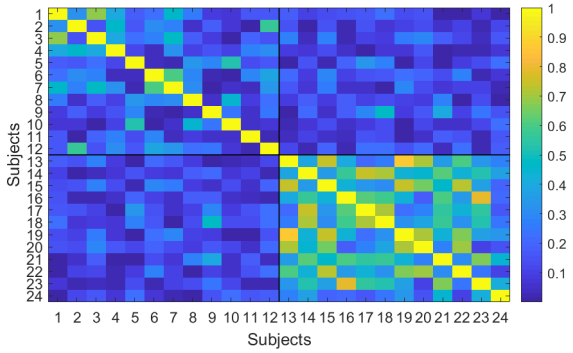
(a) Low-rank, Dataset 1



(b) Low-rank, Dataset 2



(c) Group-sparse, Dataset 1



(d) Group-sparse, Dataset 2

Figure 5: Affinity matrix.

REFERENCES

[1] Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., and Turner, R., "Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation," *Proc. Natl. Acad. Sci.* **89**, 5675–5679 (1992).

[2] Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Essen, D. C. V., and Raichle, M. E., "The human brain is intrinsically organized into dynamic," *Proc. Natl. Acad. Sci.* **102**(27), 9673–9678 (2005).

[3] Olshausen, B. A. and Field, D. J., "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vis. Res.* **37**(23), 3311–3325 (1997).

[4] Lewicki, M. and Sejnowski, T., "Learning overcomplete representations," *Neural Computation* **15**(2), 349–396 (2000).

[5] Essen, D. C. V., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K., "The WU-Minn Human Connectome Project: An overview," *NeuroImage* **80**, 62–79 (2013).

[6] Scott, A., Courtney, W., Wood, D., de la Garza, R., Lane, S., King, M., Wang, R., Roberts, J., Turner, J. A., and Calhoun, V. D., "COINS: An innovative informatics and neuroimaging tool suite built for large heterogeneous datasets," *Front. Neuroinform.* **5**, 1–15 (2011).

[7] Kim, S.-J., Calhoun, V. D., and Adalı, T., "Flexible large-scale fMRI analysis: A survey," in [*Proc. of the 42nd Intl. Conf. Acoust., Speech, Sig. Process.*], 6319–6323 (Mar. 2017).

[8] Lee, K., Tak, S., and Ye, J. C., "A data-drive sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion," *IEEE Trans. Med. Imag.* **30**, 1076–1089 (May 2011).

[9] Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B., "Multi-subject dictionary learning to segment an atlas of brain spontaneous activity," in [*Proc. Int'l. Conf. Info. Process. Med. Imag.*], 562–573 (2011).

[10] Vidal, R., "Subspace clustering," *IEEE Sig. Proc. Mag.* **28**(2), 52–68 (2011).

[11] Liu, G., Lin, Z., and Yu, Y., "Robust subspace segmentation by low-rank representation," in [*Proc. of the ACM KDD Conf.*], 126–135 (Aug. 2010).
[12] Elhamifar, E. and Vidal, R., "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 2765–2785 (Nov. 2013).
[13] Lv, J., Li, X., Zhu, D., Jiang, X., Zhang, X., Hu, X., Zhang, T., Guo, L., and Liu, T., "Sparse representation of group-wise fMRI signals," in [*Proc. of the 16th Intl. Conf. Med. Imag. Computing Computer Assist. Intervention*], 608–616 (Sep. 2013).
[14] Varoquaux, G. and Craddock, R. C., "Learning and comparing functional connectomes across subjects," *NeuroImage* **80**, 405–415 (2013).
[15] Lee, J., Jeong, Y., and Ye, J. C., "Group sparse dictionary learning and inference for resting-state fMRI analysis of Alzheimer's disease," in [*Proc. 10th Intl. Symp. Biomed. Imaging*], 540–543 (Apr. 2013).
[16] Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Eger, E., Bach, F., and Thirion, B., "Multiscale mining of fMRI data with hierarchical structured sparsity," *SIAM Imag. Sci.* **5**(3), 835–856 (2012).
[17] Ma, L., Wang, C., Xiao, B., and Zhou, W., "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in [*Proc. of the IEEE Conf. Comput. Vis. Pattern Recognition*], 2586–2593 (Jun. 2012).
[18] Lv, J. et al., "Sparse representation of whole-brain fMRI signals for identification of functional networks," *Medical Image Analysis* **20**, 112–134 (2015).
[19] Mensch, A., Varoquaux, G., and Thirion, B., "Compressed online dictionary learning for fast resting-state fMRI decomposition," in [*Proc. of the IEEE 13th Intl. Symp. Biomed. Imag.*], 1282–1285 (2016).
[20] Mairal, J., Bach, F., Ponce, J., and Sapiro, G., "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.* **11**, 19–60 (2010).
[21] Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y., "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 171–184 (Jan. 2013).
[22] Ding, C., Li, T., Peng, W., and Park, H., "Orthogonal nonnegative matrix tri-factorization for clustering," in [*Proc. of the ACM KDD Conf.*], 126–135 (Aug. 2006).
[23] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., "Least angle regression," *Ann. Statist.* **32**(2), 407–499 (2004).
[24] Beck, A. and Teboulle, M., "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.* **2**(1), 183–202 (2009).
[25] Friedman, J., Hastie, T., and Tibshirani, R., "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.* **33**(1), 1–22 (2010).
[26] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.* **3**(1), 1–122 (2011).
[27] Cai, J.-F., Candes, E. J., and Shen, Z., "A singular value thresholding algorithm for matrix completion," *SIAM J. Optimiz.* **20**, 1956–1982 (Mar. 2010).
[28] Jenatton, R., Mairal, J., Obozinski, G., and Bach, F., "Proximal methods for hierarchical sparse coding," *J. Mach. Learn. Res.* **12**, 2297–2334 (2011).
[29] Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., and Rottenberg, D., "The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework," *NeuroImage* **15**, 747–771 (2002).
[30] Erhardt, E. B., Allen, E. A., Wei, Y., Eichele, T., and Calhoun, V. D., "SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability," *NeuroImage* **59**(4), 4160–4167 (2012).