Flexible Multisubject Multiset FMRI Data Analysis Using Robust Discriminative Dictionary Learning

Rui Jin[†], Shuai Xu[†], Seung-Jun Kim[†*], Vince Calhoun[‡]
[†]*Dept. of Computer Science & Electrical Engineering, University of Maryland, Baltimore County, Baltimore, MD*
[‡]*Tri-institutional Center for Translational Research in Neuroimaging & Data Science (TReNDS), Atlanta, GA*

***Abstract*—A novel dictionary learning method is proposed for analyzing multi-subject multiset functional magnetic resonance imaging (fMRI) data. It is assumed that the subjects are grouped based on subject-specific attributes, and multiple fMRI data sets are available for each subject. The proposed algorithm can identify four types of neural activation maps from such data sets. First, the maps shared across groups and those that are discriminative of group differences are estimated by incorporating the group attribute labels. Furthermore, the maps that are common across data sets and those unique to a data set are extracted using group sparsity constraints. Notably, the map types can be flexibly determined without pre-specifying the numbers of the maps of different types. Numerical tests on synthetic and real data verify the benefit of the proposed method.**

I. Introduction

Functional magnetic resonance imaging (fMRI) studies are instrumental for discovering underlying functional networks and potential biomarkers for brain diseases. Various fMRI data analysis methods have been developed. Model-driven methods such as the general linear model (GLM) obtain spatial activation maps by fitting the data to predefined temporal stimuli [1]. Data-driven methods, including independent component analysis (ICA) or dictionary learning (DL), make minimal modeling assumptions such as statistical independence or sparsity of component maps [2]. More recently, deep learning techniques have been employed as well [3].

As large-scale fMRI studies that involve more than hundreds of subjects become increasingly common, powerful data-driven methods for multisubject fMRI data are actively pursued. In particular, subject-specific phenotypes such as diagnosis, gender, and handedness may be incorporated. The group analysis allows estimation of features that are discriminative across subject groups. A DL-based method was developed to estimate the activation maps that are group-discriminative using Fisher's discriminant cost [4]. An ICA-based method was extended to extract shared and group-specific activation maps through orthogonality constraints [5].

It is useful to analyze multiple fMRI data sets jointly. For instance, task fMRI data sets acquired performing different tasks can be processed through multiset analysis to provide complementary views to brain functions. A number of ICA-based methods were developed for multiset fMRI data analysis, and it was shown that multiset analysis can differentiate the subject groups better than single-set analysis [6], [7]. However, the methods did not incorporate group attributes directly for the analysis.

In this work, we propose a novel DL-based method for multisubject multiset fMRI data analysis. The group attribute labels are incorporated through Fisher discriminant cost, thus estimating maps that are indicative of group differences, in addition to the maps that are shared across subject groups. Furthermore, maps that are common across multiple data sets and those unique to a data set (or a subset of data sets) are also identified by incorporating group-sparsity constraints. However, rather than rigidly allocating these four map types a priori, robust techniques are employed to flexibly determine the map types in a data-driven fashion. The algorithm is evaluated on synthetic and real data sets to verify the validity and effectiveness of the approach.

The rest of the paper is organized as follows. The proposed formulation is presented in Sec. II. The algorithm is derived in Sec. III. The method is evaluated in Sec. IV. Conclusions are provided in Sec. V.

This work was supported in part by NSF grants 1631838 and 2242412.
*Corresponding author e-mail:sjkim@umbc.edu.

II. Problem Formulation

A. Dictionary Learning for FMRI Data Analysis

Let us first briefly recap the DL-based analysis method developed for a single fMRI data set [4], which is extended to multiset analysis in this work. Let $\mathbf{X} \in \mathbb{R}^{M\times V}$ be a fMRI data set with M subjects and V voxels. DL amounts to factorizing $\mathbf{X}$ into a dictionary matrix $\mathbf{D} \in \mathbb{R}^{M\times K}$ and a sparse matrix $\mathbf{Z} \in \mathbb{R}^{K\times V}$, such that $\mathbf{X} \approx \mathbf{DZ}$. In the fMRI data analysis context, the factorization yields K sparse neural activation maps in the rows of $\mathbf{Z}$ and corresponding subject-wise weights for individual maps in the columns of $\mathbf{D}$.

If group attributes for subjects are available, discriminative DL can be employed to obtain maps that are characteristic of the differences between groups, in addition to maps that are shared across groups. Discriminative DL can be formulated using the Fisher discriminant cost [8]. Suppose that there are G groups. For $g \in \{1,\ldots,G\}$, let $\mathcal{N}_g$ with cardinality N_g be the set of subject indices belonging to group g. Discriminative DL learns feature vectors $\mathbf{Q} := [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_M]$ for M subjects. The Fisher discriminant cost function $f(\mathbf{Q})$ promotes features $\{\mathbf{q}_m\}$ to be clustered within the same group and dispersed across different groups. Define $\mathbf{m}_g := \frac{1}{N_g}\sum_{m\in\mathcal{N}_g}\mathbf{q}_m$, and $\mathbf{m} := \frac{1}{M}\sum_{m=1}^{M}\mathbf{q}_m$. Also define the within-class scatter matrix $\mathbf{S}_\mathrm{w}$ and the between-class scatter matrix $\mathbf{S}_\mathrm{b}$ as

$$\mathbf{S}_\mathrm{w}(\mathbf{Q}) := \sum_{g=1}^{G}\sum_{m\in\mathcal{N}_g}(\mathbf{q}_m-\mathbf{m}_g)(\mathbf{q}_m-\mathbf{m}_g)^\top \quad (1)$$

$$\mathbf{S}_\mathrm{b}(\mathbf{Q}) := \sum_{g=1}^{G} N_g(\mathbf{m}_g-\mathbf{m})(\mathbf{m}_g-\mathbf{m})^\top. \quad (2)$$

where $^\top$ denotes transposition. Then $f(\mathbf{Q})$ is defined as

$$f(\mathbf{Q}) := \mathrm{tr}\{\mathbf{S}_\mathrm{w}(\mathbf{Q})\} - \mathrm{tr}\{\mathbf{S}_\mathrm{b}(\mathbf{Q})\} + \|\mathbf{Q}\|_F^2 \quad (3)$$

where the last term ensures that $f(\mathbf{Q})$ is convex [9].

Now split $\mathbf{Z}$ into $\mathbf{Z} := [\overline{\mathbf{Z}}^\top, \widetilde{\mathbf{Z}}^\top]^\top$, where $\widetilde{\mathbf{Z}} \in \mathbb{R}^{\widetilde{K}\times V}$ contains $\widetilde{K}$ discriminative maps and $\overline{\mathbf{Z}} \in \mathbb{R}^{\overline{K}\times V}$ contains $\overline{K}$ shared ones, with $\overline{K}+\widetilde{K}=K$. Likewise, $\mathbf{D}$ is partitioned as $\mathbf{D} := [\overline{\mathbf{D}}, \widetilde{\mathbf{D}}]$, where the rows of $\widetilde{\mathbf{D}} \in \mathbb{R}^{M\times\widetilde{K}}$ are the discriminative features for individual subjects and the rows of $\overline{\mathbf{D}} \in \mathbb{R}^{M\times\overline{K}}$ are non-discriminative ones. Hence, $\widetilde{\mathbf{D}}^\top$ can be used as the feature matrix $\mathbf{Q}$. The DL formulation is given by [4]

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{Z}} \frac{1}{2}\|\mathbf{X}-\mathbf{DZ}\|_F^2+\lambda\|\mathbf{Z}\|_1+\frac{\mu}{2}f(\widetilde{\mathbf{D}}^\top) \quad (4)$$

where the constraint set $\mathcal{D} := \{[\mathbf{d}_1, \ldots, \mathbf{d}_K] : \|\mathbf{d}_k\|_2 \leq 1, k = 1, \ldots, K\}$ for $\mathbf{D}$ is employed to avoid scaling ambiguity of bi-factorization, $\|\mathbf{Z}\|_1$ is a regularizer promoting sparsity in $\mathbf{Z}$, and λ and μ are pre-specified positive regularization parameters. It is worth noting that one must also pre-specify the values of $\overline{K}$ and $\widetilde{K}$.

B. Flexible Dictionary Learning for Multiset FMRI Data

Formulation (4) learns shared and discriminative spatial maps $\overline{\mathbf{Z}}$ and $\widetilde{\mathbf{Z}}$ from a single multisubject data set $\mathbf{X}$ and group attributes $\{\mathcal{N}_g\}$. The method is now extended to the case of analyzing multiple data sets acquired for the same set of subjects. Let $\{\mathbf{X}^s \in \mathbb{R}^{M \times V}\}_{s=1}^S$ be a collection of S fMRI data sets from M subjects. For example, the data sets can be captured from the fMRI scans taken for S different tasks performed by the subjects.

We again aim at obtaining K sparse spatial maps $\mathbf{Z}^s \in \mathbb{R}^{K \times V}$ and corresponding weights $\mathbf{D}^s \in \mathbb{R}^{M \times K}$ such that $\mathbf{X}^s \approx \mathbf{D}^s \mathbf{Z}^s$ for $s = 1, 2, \ldots, S$. However, it is expected that the S sets of maps $\{\mathbf{Z}^s\}$ are related since they are due to the same set of subjects. In particular, there may be maps that show up commonly in all data sets and maps that appear uniquely only in one data set (or a subset of data sets). Furthermore, the group attribute labels $\{\mathcal{N}_g\}$ should also be utilized to extract maps that are group-discriminative and shared across groups. In summary, four types of maps are conceived. The maps can be either common or distinct across the data sets. Also, the maps can be either group-discriminative or shared depending on whether or not they are characteristic of the group differences.

To identify the four types of maps, rather than splitting $\mathbf{Z}$ and $\mathbf{D}$ with pre-specified dimensions as in Sec. II-A, here we propose to flexibly determine the map types by employing robust methods. To estimate common and distinct maps across data sets, first collect the k-th maps extracted from S data sets as

$$\mathbf{Z}_k := \begin{bmatrix} \mathbf{Z}^1(k,:) \\ \vdots \\ \mathbf{Z}^S(k,:) \end{bmatrix} \in \mathbb{R}^{S \times V} \tag{5}$$

for $k = 1, \ldots, K$, where $(k,:)$ and $(:,k)$ extract the k-th row and the k-th column, respectively. Inspired by [10], we decompose $\mathbf{Z}_k$ as

$$\mathbf{Z}_k = \widehat{\mathbf{Z}}_k + \check{\mathbf{Z}}_k \tag{6}$$

where $\widehat{\mathbf{Z}}_k$ is constrained to be column-wise group-sparse and $\check{\mathbf{Z}}_k$ is constrained to be both element-wise sparse and row-wise group-sparse. The column-sparsity of $\widehat{\mathbf{Z}}_k$ encodes the prior information that common maps from different data sets tend to have shared supports, meaning that similar neural activation patterns are exhibited across all data sets. However, not all maps are expected to show such commonality, which is handled by $\check{\mathbf{Z}}_k$. The row-sparsity imposed on $\check{\mathbf{Z}}_k$ allows a small subset of rows in $\mathbf{Z}_k$ to deviate from the coupling across all S maps effected by $\widehat{\mathbf{Z}}_k$. The element-wise sparsity on $\check{\mathbf{Z}}_k$ again encourages that the maps obtained in $\mathbf{Z}_k$ are spatially sparse, which is expected for neural activations.

A similar idea is employed to extract discriminative and non-discriminative maps for predicting group attributes. Collect the weight matrices from S data sets and decompose it as

$$\boldsymbol{\Delta} := [\mathbf{D}^1, \mathbf{D}^2, \ldots, \mathbf{D}^S] = \widetilde{\boldsymbol{\Delta}} + \widehat{\boldsymbol{\Delta}} \in \mathbb{R}^{M \times SK}. \tag{7}$$

Each row of $\widetilde{\boldsymbol{\Delta}}$ is constrained to be discriminative by imposing the Fisher discriminant constraint. If $\widehat{\boldsymbol{\Delta}}$ were entirely zero, this would render all SK maps in $\{\mathbf{Z}^s\}$ to be group-discriminative. However, not all maps are expected to be group-discriminative. Suppose that the k-th map of the s-th data set is not group-discriminative. Then, the corresponding column (the $[(s-1)K + k]$-th column) in $\widehat{\boldsymbol{\Delta}}$ can assume nonzero values, allowing the same column in $\boldsymbol{\Delta}$ to be exempt from the discriminative constraint imposed by the Fisher cost. Thus, column-sparsity is imposed on $\widehat{\boldsymbol{\Delta}}$. The nonzero columns in $\widehat{\boldsymbol{\Delta}}$ indicate non-discriminative maps (in the corresponding rows of $[(\mathbf{Z}^1)^\top, (\mathbf{Z}^2)^\top, \ldots, (\mathbf{Z}^S)^\top]^\top$), and the rest discriminative.

Based on these ideas, our DL problem is formulated as

$$\begin{aligned} \min_{\{\mathbf{D}^s \in \mathcal{D}\}, \{\mathbf{Z}^s\}} \frac{1}{2} \sum_{s=1}^S \|\mathbf{X}^s - \mathbf{D}^s \mathbf{Z}^s\|_F^2 + \sum_{k=1}^K \Big\{ \lambda_1 \|\widehat{\mathbf{Z}}_k\|_{2,1} \\ + \lambda_2 \Big[\alpha \|\check{\mathbf{Z}}_k\|_{1,2} + (1-\alpha) \|\check{\mathbf{Z}}_k\|_1 \Big] \Big\} + \frac{\lambda_3}{2} f(\widetilde{\boldsymbol{\Delta}}^\top) + \lambda_4 \|\widehat{\boldsymbol{\Delta}}\|_{2,1} \end{aligned} \tag{8}$$

where $\{\lambda_i\}$ and α are nonnegative weights associated with various regularizers. The regularizers based on the $\ell_{2,1}$-mixed norm $\|\cdot\|_{2,1}$ levied on $\widehat{\mathbf{Z}}_k$ and $\widehat{\boldsymbol{\Delta}}$ promote column-sparsity. Similarly, the regularizer $\|\check{\mathbf{Z}}_k\|_{1,2}$ encourages row-sparsity. Additional ℓ_1-norm-based regularization is imposed on $\check{\mathbf{Z}}_k$ to ensure that the obtained maps in $\mathbf{Z}_k$ are sparse.

It must be emphasized that, unlike (4), formulation (8) does not require rigid splitting of $\mathbf{Z}^s$ into common and unique parts or $\mathbf{D}^s$ into discriminative and non-discriminative parts. Instead, each map can be determined flexibly which of the four types it belongs to in a data-driven fashion. Moreover, when a map is found to be shared across a subset of data sets, this subset can be identified easily from the group-sparsity structure. The price to pay is that (8) needs more regularization parameters to be specified than (4).

III. Algorithm Derivation

In this section, an algorithm to solve (8) is derived in the proximal alternating linearized minimization (PALM) framework, which allows to find a stationary point to a class of nonconvex nonsmooth problems by iteratively optimizing over blocks of variables [11]. For (8), we choose the following $(2K + 2)$ blocks: $\{\widehat{\mathbf{Z}}_k\}_{k=1}^K$, $\{\check{\mathbf{Z}}_k\}_{k=1}^K$, $\widehat{\boldsymbol{\Delta}}$, and $\boldsymbol{\Delta}$. Furthermore, let us define $\mathbf{X} := [(\mathbf{X}^1)^\top, \ldots, (\mathbf{X}^S)^\top]^\top \in \mathbb{R}^{SM \times V}$ and $\mathbf{D}_k := \mathrm{bdiag}\{\mathbf{D}^1(:,k), \ldots, \mathbf{D}^S(:,k)\} \in \mathbb{R}^{SM \times S}$. Then, the continuously differentiable part of the objective function in (8) is given by

$$\begin{aligned} & h(\{\widehat{\mathbf{Z}}_k\}_{k=1}^K, \{\check{\mathbf{Z}}_k\}_{k=1}^K, \widehat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}) \\ & := \frac{1}{2} \left\| \mathbf{X} - \sum_{k=1}^K \mathbf{D}_k (\widehat{\mathbf{Z}}_k + \check{\mathbf{Z}}_k) \right\|_F^2 + \frac{\lambda_3}{2} f\left(\left(\boldsymbol{\Delta} - \widehat{\boldsymbol{\Delta}} \right)^\top \right). \end{aligned} \tag{9}$$

Note that all entries of $\{\mathbf{D}_k\}$ in (9) are included in variable $\boldsymbol{\Delta}$. The algorithm proceeds by iteratively updating the blocks sequentially via appropriate proximal operators, with other blocks fixed at the current iterates [11], [12].

First, with $\{\widehat{\mathbf{Z}}_{k'}\}_{k' \neq k}$, $\{\check{\mathbf{Z}}_{k'}\}$, $\widehat{\boldsymbol{\Delta}}$, and $\boldsymbol{\Delta}$ fixed at the current iterates, $\widehat{\mathbf{Z}}_k$ is updated for $k = 1, 2, \ldots, K$. Denote the current (t-th) iterate of $\widehat{\mathbf{Z}}_k$ by $\widehat{\mathbf{Z}}_k^{(t)}$. Let $\hat{\rho}_k := \hat{\gamma}_k \widehat{L}_k$, where $\hat{\gamma}_k > 1$ and $\widehat{L}_k$ is the Lipschitz constant of $\nabla_{\widehat{\mathbf{Z}}_k} h(\cdot)$ w.r.t. $\widehat{\mathbf{Z}}_k$. Let $\nabla_{\widehat{\mathbf{Z}}_k} h^{(t)}$ denote the gradient of $h(\cdot)$ w.r.t. $\widehat{\mathbf{Z}}_k$ at the current iterate, i.e., $\nabla_{\widehat{\mathbf{Z}}_k} h^{(t)} := \nabla_{\widehat{\mathbf{Z}}_k} h(\widehat{\mathbf{Z}}_1^{(t+1)}, \ldots, \widehat{\mathbf{Z}}_{k-1}^{(t+1)}, \widehat{\mathbf{Z}}_k^{(t)}, \ldots, \widehat{\mathbf{Z}}_K^{(t)}, \{\check{\mathbf{Z}}_{k'}^{(t)}\}, \widehat{\boldsymbol{\Delta}}^{(t)}, \boldsymbol{\Delta}^{(t)})$. Then, upon defining $\widehat{\mathbf{U}}_k := \widehat{\mathbf{Z}}_k^{(t)} - \frac{1}{\hat{\rho}_k} \nabla_{\widehat{\mathbf{Z}}_k} h^{(t)}$, the update for $\widehat{\mathbf{Z}}_k$ is done through the proximal operation

$$\widehat{\mathbf{Z}}_k^{(t+1)} = \arg\min_{\widehat{\mathbf{Z}}_k} \lambda_1 \|\widehat{\mathbf{Z}}_k\|_{2,1} + \frac{\hat{\rho}_k}{2} \left\| \widehat{\mathbf{Z}}_k - \widehat{\mathbf{U}}_k \right\|_F^2 \tag{10}$$

which has a closed-form solution [12]

$$\widehat{\mathbf{Z}}_k^{(t+1)}(:,v) = \max\left\{0, 1 - \frac{\lambda_1}{\hat{\rho}_k \|\widehat{\mathbf{U}}_k(:,v)\|_2}\right\} \widehat{\mathbf{U}}_k(:,v) \ \forall v. \quad (11)$$

Next, for updating $\check{\mathbf{Z}}_k$, $k = 1, 2, \ldots, K$, first define

$$\check{r}(\check{\mathbf{Z}}_k) := \lambda_2 \left[\alpha \|\check{\mathbf{Z}}_k\|_{1,2} + (1-\alpha)\|\check{\mathbf{Z}}_k\|_1\right]. \quad (12)$$

Define also $\check{\mathbf{U}}_k := \check{\mathbf{Z}}_k^{(t)} - \frac{1}{\check{\rho}_k} \nabla_{\check{\mathbf{Z}}_k} h^{(t)}$, and introduce $\check{\rho}_k := \check{\gamma}_k \check{L}_k$ with $\check{\gamma}_k > 1$ and the Lipschitz constant $\check{L}_k$ of $\nabla_{\check{\mathbf{Z}}_k} h(\cdot)$. Then, $\check{\mathbf{Z}}_k$ can be updated via

$$\check{\mathbf{Z}}_k^{(t+1)} = \arg\min_{\check{\mathbf{Z}}_k} \check{r}(\check{\mathbf{Z}}_k) + \frac{\check{\rho}_k}{2} \left\|\check{\mathbf{Z}}_k - \check{\mathbf{U}}_k\right\|_F^2. \quad (13)$$

The proximal operator for the combination of the $\ell_{1,2}$-norm and the ℓ_1-norm in $\check{r}$ can be computed in sequence [12]. The proximal operation due to the ℓ_1-norm is first performed to get $\check{\mathbf{U}}'_k$ as

$$\check{U}'_k(s,v) := \max\left\{0, 1 - \frac{\lambda_2(1-\alpha)}{\check{\rho}_k |\check{U}_k(s,v)|}\right\} \check{U}_k(s,v) \ \forall s, v. \quad (14)$$

Then, the proximal operation due to the $\ell_{1,2}$-norm is carried out as

$$\check{\mathbf{Z}}_k^{(t+1)}(s,:) := \max\left\{0, 1 - \frac{\lambda_2 \alpha}{\check{\rho}_k \|\check{\mathbf{U}}'_k(s,:)\|_2}\right\} \check{\mathbf{U}}'_k(s,:) \ \forall s. \quad (15)$$

The update for $\widehat{\mathbf{\Delta}}$ is done with $\widehat{\mathbf{W}} := \widehat{\mathbf{\Delta}}^{(t)} - \frac{1}{\rho_1} \nabla_{\widehat{\mathbf{\Delta}}} h^{(t)}$ as

$$\widehat{\mathbf{\Delta}}^{(t+1)} = \arg\min_{\widehat{\mathbf{\Delta}}} \lambda_4 \|\widehat{\mathbf{\Delta}}\|_{2,1} + \frac{\rho_1}{2} \left\|\widehat{\mathbf{\Delta}} - \widehat{\mathbf{W}}\right\|_F^2 \quad (16)$$

where $\rho_1 := \gamma_1 L_{\widehat{\Delta}}$ with $\gamma_1 > 1$ and the Lipschitz constant $L_{\widehat{\Delta}}$ of $\nabla_{\widehat{\mathbf{\Delta}}} h(\cdot)$. The update (16) can be shown to be equivalent to

$$\widehat{\mathbf{\Delta}}^{(t+1)}(:,j) = \max\left\{0, 1 - \frac{\lambda_4}{\rho_1 \|\widehat{\mathbf{W}}(:,j)\|_2}\right\} \widehat{\mathbf{W}}(:,j) \ \forall j. \quad (17)$$

Finally, to update $\mathbf{\Delta}$, define $r_{\mathcal{D}}(\mathbf{\Delta})$ as $r_{\mathcal{D}}(\mathbf{\Delta}) := 0$ if $\mathbf{D}^s \in \mathcal{D}$, for all s, and $r_{\mathcal{D}}(\mathbf{\Delta}) := \infty$, otherwise. Then, upon defining $\mathbf{W} := \mathbf{\Delta}^{(t)} - \frac{1}{\rho_2} \nabla_{\mathbf{\Delta}} h^{(t)}$, the relevant proximation operation is

$$\mathbf{\Delta}^{(t+1)} = \arg\min_{\mathbf{\Delta}} r_{\mathcal{D}}(\mathbf{\Delta}) + \frac{\rho_2}{2} \|\mathbf{\Delta} - \mathbf{W}\|_F^2 \quad (18)$$

where $\rho_2 := \gamma_2 L_{\Delta}$ with $\gamma_2 > 1$. This is equivalent to

$$\mathbf{\Delta}^{(t+1)}(:,j) = \frac{1}{\max\{1, \|\mathbf{W}(:,j)\|_2\}} \mathbf{W}(:,j) \ \forall j. \quad (19)$$

The overall algorithm is presented in Table I.

TABLE I: Algorithm for solving (8).

Input: $\{\mathbf{X}^s\}, K, \{\lambda_l\}_{l=1}^4, \alpha, \{\widehat{\gamma}_k, \check{\gamma}_k\}, \gamma_1, \gamma_2$, MAXITER
Output: $\mathbf{\Delta}^{(MAXITER)}, \{\mathbf{Z}_k^{(MAXITER)}\}$

Initialize $\{\widehat{\mathbf{Z}}_k^{(1)}\}, \{\check{\mathbf{Z}}_k^{(1)}\}, \widehat{\mathbf{\Delta}}^{(1)}$, and $\mathbf{\Delta}^{(1)}$ randomly.
For $t = 1, \ldots, MAXITER - 1$
  / Update $\{\widehat{\mathbf{Z}}_k\}$ */*
  For $k = 1, \ldots, K$
    $\widehat{\mathbf{U}}_k = \widehat{\mathbf{Z}}_k^{(t)} - \frac{1}{\hat{\rho}_k} \nabla_{\widehat{\mathbf{Z}}_k} h^{(t)}$
    $\widehat{\mathbf{Z}}_k^{(t+1)}(:,v) = \max\left\{0, 1 - \frac{\lambda_1}{\hat{\rho}_k \|\widehat{\mathbf{U}}_k(:,v)\|_2}\right\} \widehat{\mathbf{U}}_k(:,v) \ \forall v$
  End For
  / Update $\{\check{\mathbf{Z}}_k\}$ */*
  For $k = 1, \ldots, K$
    $\check{\mathbf{U}}_k = \check{\mathbf{Z}}_k^{(t)} - \frac{1}{\check{\rho}_k} \nabla_{\check{\mathbf{Z}}_k} h^{(t)}$
    $\check{U}'_k(s,v) = \max\left\{0, 1 - \frac{\lambda_2(1-\alpha)}{\check{\rho}_k |\check{U}_k(s,v)|}\right\} \check{U}_k(s,v) \ \forall s, v$
    $\check{\mathbf{Z}}_k^{(t+1)}(s,:) = \max\left\{0, 1 - \frac{\lambda_2 \alpha}{\check{\rho}_k \|\check{\mathbf{U}}'_k(s,:)\|_2}\right\} \check{\mathbf{U}}'_k(s,:) \ \forall s$
  End For
  / Update $\widehat{\mathbf{\Delta}}$ */*
  $\widehat{\mathbf{W}} = \widehat{\mathbf{\Delta}}^{(t)} - \frac{1}{\rho_1} \nabla_{\widehat{\mathbf{\Delta}}} h^{(t)}$
  $\widehat{\mathbf{\Delta}}^{(t+1)}(:,j) = \max\left\{0, 1 - \frac{\lambda_4}{\rho_1 \|\widehat{\mathbf{W}}(:,j)\|_2}\right\} \widehat{\mathbf{W}}(:,j) \ \forall j$
  / Update $\mathbf{\Delta}$ */*
  $\mathbf{W} = \mathbf{\Delta}^{(t)} - \frac{1}{\rho_2} \nabla_{\mathbf{\Delta}} h^{(t)}$
  $\mathbf{\Delta}^{(t+1)}(:,j) = \frac{1}{\max\{1, \|\mathbf{W}(:,j)\|_2\}} \mathbf{W}(:,j), \ \forall j$
End For
Set $\mathbf{Z}_k^{(MAXITER)} = \widehat{\mathbf{Z}}_k^{(MAXITER)} + \check{\mathbf{Z}}_k^{(MAXITER)}, \ \forall k$

IV. Evaluation

A. Test with Synthetic Data Sets

The proposed method was first tested on $S = 3$ synthetic data sets. For each data set, $K = 10$ sparse maps with $V = 48,546$ were generated from a Bernoulli ($p = 0.5$)-Gaussian distribution. Each pair of common maps were made correlated with a correlation coefficient 0.95. There are six common maps in each data set. The dictionaries for $M = 247$ subjects were generated from $\mathcal{N}(0,1)$. To simulate two groups, a step signal was added to the entries of 109 subjects in the atoms associated with discriminative maps. Five discriminative maps were made for each data set. The data sets were then obtained as the products of the dictionaries and the sparse maps, plus Gaussian noise. All map types of the synthetic data sets are shown in Fig. 1, where the same pattern/color in a column signifies common maps, and the stars indicate discriminative maps.

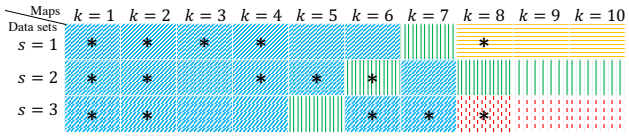


Fig. 1: Map types in the synthetic data sets.

A grid search for $\{\lambda_l\}_{l=1}^4$ was performed by splitting the data into training and validation sets and choosing the parameter set with best classification accuracy. The atoms indicated by the column-sparsity of $\widehat{\mathbf{\Delta}}$ were used to perform classification. The value of α was set to 0.99. With the parameters fixed, (8) was solved using 50 different initializers and the most stable maps were selected [4].

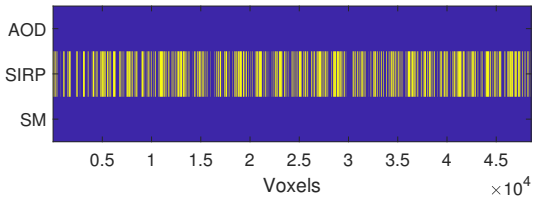
In addition to our proposed algorithm, existing methods were also tested for comparison. The benchmark methods are the multiset canonical correlation analysis + joint ICA (MCCA+jICA) [13], disjoint subspace analysis using ICA (DS-ICA) [7], joint sparse representation analysis (jSRA) [14], and single-set discriminative DL [4] methods. The DS-ICA algorithm was run with various common orders $K_c = 4, 5, 6, 7$ (denoted as DS-ICA-K_c). To assess the map estimation performance, the correlation coefficients between the matching ground-truth and estimated maps were computed and averaged. Also, to see whether the (non-)discriminative maps are correctly estimated as (non-)discriminative, the p-values of the estimated atoms were computed using two-sample t-tests. Then, the fraction of the discriminative maps with the p-values less than 0.05 (as well as the non-discriminative maps with the p-values larger than 0.05) was recorded as discriminability estimation accuracy. In Table II, it is seen that our method achieves the best performance in terms of both metrics. Discriminative DL, which can incorporate the labels, also achieved 100% discriminability estimation accuracy.

	Proposed	DS-ICA-4	DS-ICA-5	DS-ICA-6	DS-ICA-7	MCCA+jICA	discr. DL	jSRA
Avg. corr. coeff.	**0.98**	0.91	0.93	0.87	0.86	0.90	0.94	0.61
Discr. est. acc.	**100%**	37%	53%	57%	53%	53%	**100%**	70%

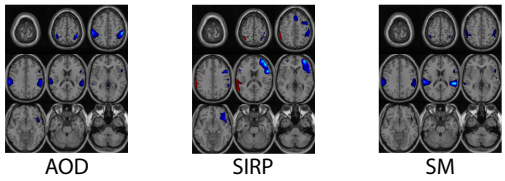
TABLE II: Performance comparison.

s \ k	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
1	(100,100)	(100,100)	(100,100)	(100,100)	(100,0)	(100,0)	(76,0)	(75,100)	(77,0)	(100,0)
2	(100,100)	(100,100)	(100,0)	(100,100)	(100,100)	(76,100)	(100,0)	(77,0)	(77,0)	(78,0)
3	(100,100)	(100,100)	(100,0)	(100,0)	(76,0)	(100,100)	(100,100)	(100,100)	(77,0)	(77,0)

TABLE III: Percentages of zeros in the entries of $\breve{\mathbf{Z}}_k(s,:)$ and $\widehat{\mathbf{D}}^s(:,k)$.



(a) Zero and nonzero entries in $\breve{\mathbf{Z}}_k$.



(b) Maps estimated in $\mathbf{Z}_k$.

Fig. 2: An example of the estimated $\breve{\mathbf{Z}}_k$ and $\mathbf{Z}_k$ for some k. As expected, the activating areas in the AOD and SM maps are similar while those in the SIRP map are not.

The map types are also indicated by our method through the sparsity of $\breve{\mathbf{Z}}_k$ and $\widehat{\mathbf{D}}^s$. All-zero and non-zero rows of $\breve{\mathbf{Z}}_k$ correspond to common and distinct maps, respectively. All-zero and non-zero columns of $\widehat{\mathbf{\Delta}}$ indicate discriminative and non-discriminative maps, respectively. To check the correctness of these indications, Table III shows the fraction of zeros in $\breve{\mathbf{Z}}_k(s,:)$ and the fraction of zeros in $\widehat{\mathbf{D}}^s(:,k)$ in ordered pairs. Comparing Table III with the ground-truth map types in Fig. 1 reveals that all map types are corrected estimated. For example, map #6 in set #3 is a common and discriminative map, which is correctly captured through the all-zero row $\breve{\mathbf{Z}}_6(3,:)$ and the all-zero column $\widehat{\mathbf{D}}^3(:,6)$.

B. Test with Real FMRI Data Sets

The method was also evaluated using real fMRI data from MIND Clinical Imaging Consortium (MCIC) [15]. Three ($S=3$) tasks are due to subjects performing auditory oddball (AOD), Sternberg item recognition paradigm (SIRP), and sensory motor (SM) tasks. Data from 138 healthy control (HC) subjects and 109 schizophrenia (SZ) patients are used. The spatial features for each subject were extracted through regressing the time series on to a design matrix, which was the convolution of stimulus onset functions and the hemodynamic response function. Finally, the contrast images between the target and the standard stimuli were used as the input matrices $\{\mathbf{X}^s \in \mathbb{R}^{M\times V}\}$ with $M=247$ subjects and $V=48,546$ voxels. The grid search resulted in $\lambda_1=0.008$, $\lambda_2=0.35$, $\lambda_3=0.14$, $\lambda_4=4.1$, $K=30$ with fixed $\alpha=0.99$. The classification accuracy was 80%. The most stable maps were obtained for further analysis.

The common and distinct maps across data sets were then identified by checking the row-sparsity of $\breve{\mathbf{Z}}_k$. The average of the correlation coefficients between all pairs of common maps was computed to be 0.81, whereas the same between distinct maps was 0.40. This indicates that the group-sparsity priors work as intended. One representative set of maps is shown in Fig. 2 for a chosen k. In Fig. 2(a), the entries of $\breve{\mathbf{Z}}_k$ are depicted, where the blue color represents zeros, and the yellow non-zeros. It can be seen that the AOD and the SM maps are identified as common, whereas the SIRP map is determined to be distinct. The maps estimated in $\mathbf{Z}_k$ are shown in Fig. 2(b) for the three data sets. The correlation coefficients between (AOD, SM), (AOD, SIRP), and (SIRP, SM) are 0.69, 0.081, and 0.041, respectively. We also looked into discriminative and non-discriminative maps as identified from the column-sparsity of $\widehat{\mathbf{\Delta}}$. The fraction of the maps that were identified as discriminative, which also exhibited p-values less than 0.05, was 63%. Thus, our method shows the right tendency in extracting the discriminative/non-discriminative maps based on the Fisher's criterion and the group-sparsity priors.

The estimated discriminative maps are shown in Fig. 3. Most maps turn out to be common across all data sets. Maps for $k=4$ are included because the p-value of the map from the AOD data set is very close to 0.05, and the estimated sparsity pattern indicates discriminative maps. The estimated maps show activations in visual, auditory, default mode network (DMN), sensory-motor and motor areas. These areas are closely related to the performed tasks. In the AOD task, 14 maps are showing significant group differences. Among these results, maps for $k=4,5,7,13,20$ capture motor, auditory, anterior DMN and sensory-motor areas. The map type estimation results are often related to the corresponding tasks. For example, maps for $k=4$ reflecting motor areas are estimated as common maps only in AOD and SM, while the one in SIRP is distinct. This makes sense as the AOD and SM tasks are related to auditory and motor functions, while the SIRP task focuses more on the visual function. Maps for $k=5$ are common maps related to the auditory areas but show much higher group differences in AOD and SM than in the SIRP task. This is reasonable given the similarity between the AOD and SM tasks. A similar observation can be made for the common maps $k=3$ showing activations in the visual areas. The map in SIRP shows much higher group differences than the other two in AOD and SM. This is because that the SIRP task is designed to be a visual task. There are 7 and 14 maps showing significant group differences in SIRP and SM, respectively, and the task-related areas like visual, DMN, sensory-motor, motor areas are captured.

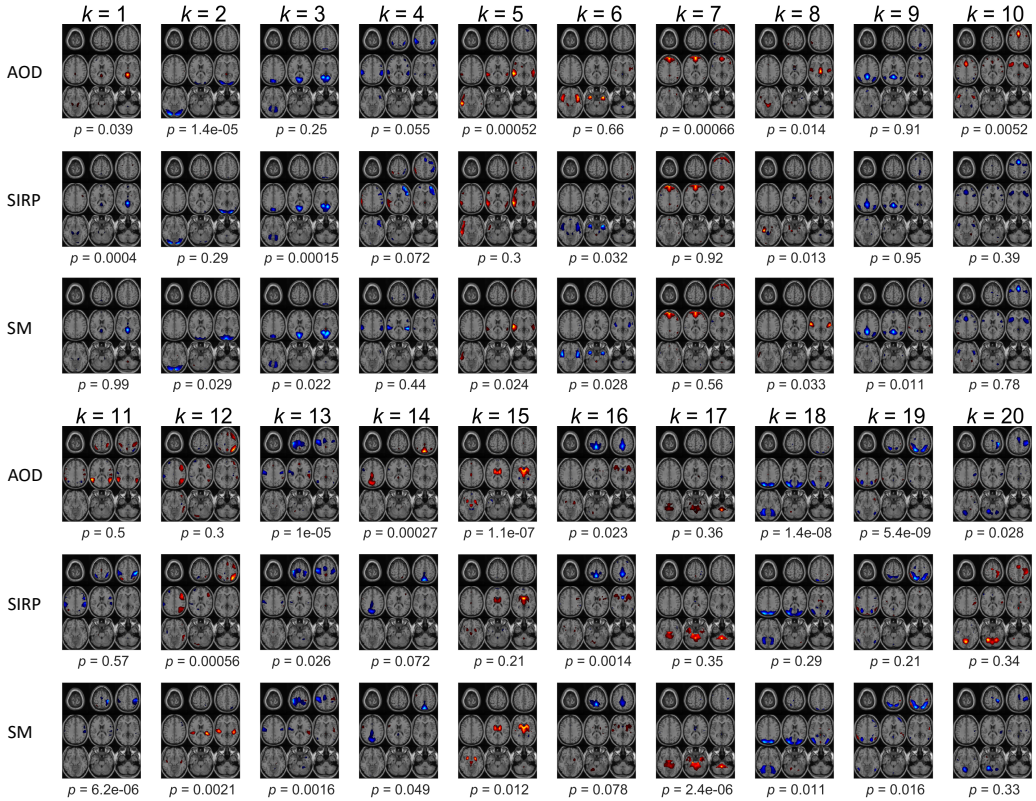


Fig. 3: Estimated discriminative maps.

V. Conclusion

A novel DL method has been proposed for analyzing multisubject multiset fMRI data. The group attributes of the subjects are exploited to extract a set of neural activation maps characteristic of the group differences. Furthermore, multiple data sets are analyzed jointly, accounting for possible commonality in the neural activations across the data sets. Four map types—discriminative or shared across groups and common or distinct across data sets—are flexibly determined through robust methods using suitable group-sparsity priors. An iterative algorithm has been derived and tested on synthetic and real fMRI data sets. The test results show that the method is effective in finding different types of maps.

References

[1] S. M. Smith, "Overview of fMRI analysis," *British J. Radiology*, vol. 77, no. suppl_2, pp. S167–S175, Jan. 2004.

[2] V. D. Calhoun and T. Adali, "Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Rev. Biomed. Eng.*, vol. 5, pp. 60–73, Aug. 2012.

[3] W. Yan, G. Qu, W. Hu, A. Abrol, B. Cai, C. Qiao, S. M. Plis, Y.-P. Wang, J. Sui, and V. D. Calhoun, "Deep learning in neuroimaging: Promises and challenges," *IEEE Signal Process. Mag.*, vol. 39, no. 2, pp. 87–98, 2022.

[4] R. Jin, K. K. Dontaraju, S.-J. Kim, M. A. B. S. Akhonda, and T. Adali, "Dictionary learning-based fMRI data analysis for capturing common and individual neural activation maps," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1265–1279, Oct. 2020.

[5] M. Maneshi, S. Vahdat, J. Gotman, and C. Grova, "Validation of shared and specific independent component analysis (SSICA) for between-group comparisons in fMRI," *Front. Neurosci.*, vol. 10, pp. 417–435, Sep. 2016.

[6] Y. Levin-Schwartz, V. D. Calhoun, and T. Adalı, "Quantifying the interaction and contribution of multiple datasets in fusion: Application to the detection of schizophrenia," *IEEE Trans. Med. Imaging*, vol. 36, no. 7, pp. 1385–1395, Jul. 2017.

[7] M. A. B. S. Akhonda, B. Gabrielson, S. Bhinge, V. D. Calhoun, and T. Adali, "Disjoint subspaces for common and distinct component analysis: Application to the fusion of multi-task fMRI data," *J. Neurosci. Methods*, vol. 358, Jul. 2021, Art. no. 109214.

[8] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. Int. Conf. Computer Vis.*, Barcelona, Spain, Nov. 2011, pp. 543–550.

[9] S. Li and Y. Fu, "Learning robust and discriminative subspace with low-rank constraints," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 27, no. 11, pp. 21602173, Nov. 2016.

[10] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proc. ACM SIGKDD Int. Conf. Knowl.*, Beijing, China, Aug. 2012, pp. 895–903.

[11] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, no. 1, pp. 459–494, Aug. 2014.

[12] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Found. Trends Mach. Learn.*, vol. 4, no. 1, pp. 1–106, Jan. 2012.

[13] J. Sui, T. Adali, G. Pearlson, H. Yang, S. R. Sponheim, T. White, and V. D. Calhoun, "A CCA+ICA based model for multi-task brain imaging data fusion and its application to schizophrenia," *Neuroimage*, vol. 51, no. 1, pp. 123–134, 2010.

[14] M. Ramezani, K. Marble, H. Trang, I. S. Johnsrude, and P. Abolmaesumi, "Joint sparse representation of brain activity patterns in multi-task fMRI data," *IEEE Trans. Med. Imaging*, vol. 34, no. 1, pp. 2–12, 2014.

[15] R. L. Gollub, J. M. Shoemaker, M. D. King, T. White, S. Ehrlich, S. R. Sponheim, V. P. Clark, J. A. Turner, B. A. Mueller, V. Magnotta, D. OLeary, B. C. Ho, S. Brauns, D. S. Manoach, L. Seidman, J. R. Bustillo, J. Lauriello, J. Bockholt, K. O. Lim, B. R. Rosen, S. C. Schulz, V. D. Calhoun, and N. C. Andreasen, "The MCIC collection: A shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia," *Neuroinformatics*, vol. 11, no. 3, pp. 367388, Jul. 2013.