A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion

Kangjoo Lee, Sungho Tak, and Jong Chul Ye

Abstract— **We propose a novel statistical analysis method for functional MRI to overcome the drawbacks of conventional data-driven methods such as the independent component analysis (ICA). Although ICA has been broadly applied to functional MRI due to its capacity to separate spatially or temporally independent components, the assumption of independence has been challenged by recent studies showing that ICA does not guarantee independence of simultaneously occurring distinct activity patterns in the brain. Instead, sparsity of the signal has been shown to be more promising. This coincides with biological findings such as sparse coding in V1 simple cells, electrophysiological experiment results in the human medial temporal lobe, and etc. The main contribution of this paper is, therefore, a new data driven fMRI analysis that is derived solely based upon the sparsity of the signals. A compressed sensing based data-driven sparse generalized linear model is proposed that enables estimation of spatially adaptive design matrix as well as sparse signal components that represent synchronous, functionally organized and integrated neural hemodynamics. Furthermore, an MDL based model order selection rule is shown to be essential in selecting unknown sparsity level for sparse dictionary learning. Using simulation and real fMRI experiments, we show that the proposed method can adapt individual variation better compared to the conventional ICA methods.**

Index Terms—**Sparse generalized linear model, sparse dictionary learning, K-SVD, statistical parametric mapping, data-driven fMRI analysis, compressed sensing, MDL principle**

I. INTRODUCTION

STATISTICAL parametric mapping (SPM) is a widely accepted mass-univariate approach for voxel-wise statistical analysis of brain activity using functional magnetic resonance imaging (fMRI) [1–5]. It uses general linear model (GLM) and random field theory to analyze and make inferences about regional brain activities. This hypothesis-driven method employs a canonical hemodynamic response function (HRF) and its various derivatives to construct regressors in the design matrix for the general linear model by convolving them with the stimulus function. The canonical HRF is the basis of a parametric model that estimates changes in the fMRI blood oxygen level-dependent (BOLD) signal evoked by an instantaneous burst of activation. The major problem in the aforementioned hypothesis-driven method is, however, the non-adaptivity of the canonical HRF [5]. Specifically, the canonical HRF does not fully consider individual and experimental variance or unpredicted phenomena during the task period, thereby reducing the sensitivity of detection. Furthermore, commonly used forms of the canonical HRF

including initial dip [6–8] and post undershoot [9–11] are still controversial within the neuroscience community [12–17].

To overcome these drawbacks, a variety of data-driven methods have been suggested, including principal component analysis (PCA) [18, 19] and independent component analysis (ICA) [20–22]. Although the time-series of BOLD are measured at each voxel, the signals related to the experimental paradigm are usually localized on a small set of regions, and are mixed with other simultaneous time-varying effects [23]. Accordingly, these approaches isolate functional spatial patterns that contain spatially localized neural dynamics. Moreover, since they do not require any prior knowledge about the paradigm, these methods can be applied to a resting state analysis of functional connectivity MRI (fcMRI) [24, 25], which does not have a predefined paradigm. Note that PCA finds components that are uncorrelated, while ICA finds components that are spatially- (spatial ICA) or temporally-independent (temporal ICA) [26, 27].

Currently, the ICA has become the main tool for data-driven fMRI analysis. More specifically, let the observed vector $\mathbf{y} = [y_1, y_2, \ldots, y_m]^T$ be a mixture of the source vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, as follows:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \qquad (1)$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a mixing matrix. The ICA then aims to find an unmixing matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$ such that output vector $\mathbf{s} = [s_1, s_2, \ldots, s_n]^T = \mathbf{W}\mathbf{y}$ provides estimates of all n spatially or temporally independent source signals. Currently, there are two ICA approaches for fMRI: temporal ICA (tICA) [28] and spatial ICA (sICA) [29, 30]. In sICA, a measurement matrix $\mathbf{Y} \in \mathbb{R}^{m \times N}$ is formed by collecting temporal time courses of length m across the voxels. [1] The resulting matrix $\mathbf{X} = \mathbf{W}\mathbf{Y} \in \mathbb{R}^{n \times N}$ of ICs then contains n-spatially independent components. Here, the $\mathbf{W}^{-1}$ matrix contains n-task related time series corresponding to the n-spatial ICs (see Figure 1(a)). On the other hand, the measurement matrix for tICA is collected as $\mathbf{Y} \in \mathbb{R}^{N \times m}$ where N is the number of voxel and m is the length of the time series. In this case, the estimated IC matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ contains n-temporally independent components as shown in Figure 1(b). Using the extracted independent time series, the 'HYBICA' approach [29] and the unified 'SPM-ICA' method [28] were recently proposed to combine ICA with a parametric approach such as SPM. The major difference between these methods is that 'HYBICA' uses spatial ICA while the unified 'SPM-

---

[1] Rather than using whole voxel N, major principle components are often used to reduce the complexity.
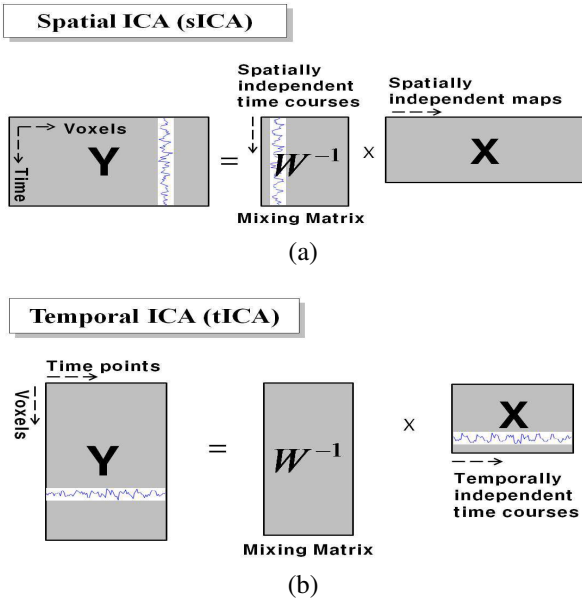
Fig. 1. Pictorial description of (a) sICA (spatial ICA), and (b) tICA (temporal ICA)

ICA' uses temporal ICA to separate the task-related temporal components. As data-driven fMRI is attracting greater interest recently, many ICA algorithms such as Infomax [31] and FastICA [32] are being implemented for group analysis of fMRI, and are now available in group ICA of fMRI Toolbox (GIFT) software package [30] or in Multivariate Exploratory Linear Optimized Decomposition into Independent Components (MELODIC) equipped in the FSL package [33].

However, the popularity of ICAs has been challenged recently by a number of studies showing that independence is not adaptive for blind source separation in fMRI [26, 34]. Also, it was shown that the most influential factor for the success rate of the ICA algorithm is sparsity of the components, rather than independence. Furthermore, many hemodynamics are rarely independent from each other due to interconnections between biological neural networks as well as preprocessing steps such as smoothing, normalization, and realignment. Due to the assumption of a low level additive noise signal, the performance of ICA is very sensitive to noise. Thus, a more effective decomposition approach that overcomes the drawbacks of the ICA is required for a data-driven fMRI analysis.

The observation that sparsity is more effective than independency in determining neural activity [26, 34] is supported by biological findings of *sparse coding* in the brain. For example, for simple- cells in the primary visual cortex (V1), Olshausen et al [35] showed that a set of receptive fields learned by maximizing sparseness in the output of a neural network model is spatially localized, oriented, and selective to spatial structure at a specific scale, similar to cortical simple cells. This finding effectively models the inference on retinal images with signals coming from optic nerve fibers, which deliver sparsely distributed events from activated neurons. This is based on the observation of the singular property of neurons, which are activated when the input stimulus is similar to the receptive features of each neuron. Similarly, the medial temporal lobe (MTL) neuron fires selectively to visual stimuli. Analyzing the neural responses of neurons from the hippocampus, amygdata, entorhinal cortex, and parahippocampal gyrus using implanted depth electrodes in the human MTL, it was shown that a single unit in the right anterior hippocampus fired with a frequency up to $20Hz$ by pictures of the actor Steve Carell while there were no statistically significant responses during presentation of other faces or at the baseline level [36]. A similar result was obtained in a study where a single neuron in the left posterior hippocampus was activated by different views of the actress Jennifer Aniston, but not by other pictures [37]. More interestingly, the authors reported that MTL neurons selectively respond to pictures of different views or drawings (pencil sketches, caricatures, coloured photographs with different backgrounds) of individuals, even to letter strings with their names, as well as to landmarks or objects. These results suggest that a sparse set of neurons encode specific concepts rather than responding to every input. These findings support the idea of sparsity of the neural response, which coincides with numerical findings using data-driven fMRI analysis methods.

Motivated by ICA analyses and biological findings, we develop a new data driven fMRI analysis method solely based on the sparsity of underlying hemodynamic signals. In this method, the BOLD signal at a specific voxel may be regarded as a combination of a sparse set of dynamic components, where each component has different time-series signal patterns. Assuming that the components for each voxel are sparse and the neural integration of the dynamics is linear, applying the sparse dictionary learning algorithm [38–40] would be reasonable to identify each component. However, the problem of applying simple sparse dictionary learning technique in fMRI is that the dictionary size is usually too big to be used as a design matrix. The main contribution of the present article is, therefore, a novel *data-driven sparse* GLM framework for a maximum likelihood (ML) estimation of spatially adaptive design matrices and sparse response signals. More specifically, a maximum likelihood framework is formulated based on the observation of sparse coding in the brain. This formulation results in spatially adaptive design matrices as a subset of atoms acquired from a learned global dictionary using a sparse dictionary learning algorithm. However, the sparse dictionary learning algorithm is usually sensitive to assumed sparsity level. Therefore, another important contribution of this paper is to show that the unknown sparsity level can be automatically estimated by minimum description length principle (MDL) [41]. The MDL principle is known to balance the trade-offs between goodness-of-fit on the data and the complexity of the model. Our results show that MDL can adapt sparsity level for each individual data effectively.

Additional features of the proposed algorithm are: (i) it is individually adaptive since the global dictionary is obtained by a fully data-driven decomposition; (ii) it does not require any knowledge of a paradigm and is appropriate for event-related or resting state functional connectivity MRI (fcMRI); and (iii) similar to HYBICA, the algorithm can be easily

incorporated within a SPM framework, and thus a statistically rich analysis is feasible using hypothesis testing, random field theory, etc [5]. Using extensive simulation and experiments with block and event-related paradigms, we show that the proposed hybrid method can adapt individual variation better and the activation maps are tightly localized in the target areas of the brain more sensitively compared to the conventional spatial and temporal ICA methods such as Infomax and FastICA algorithms, respectively.

The remaining parts of the paper are organized as follows. The theory of the data-driven sparse GLM model and the resulting sparse dictionary learning is presented in Section II, followed by a description of the methods used in this paper in Section III. Section IV provides experimental results, and Section V provides a conclusion and discussion.

II. THEORY

A. Notation

Throughout the paper, $\mathbf{x}^i$ and $\mathbf{x}_j$ correspond to the i-th row and the j-th column of matrix $\mathbf{X}$, respectively. When S is an index set, $\mathbf{X}^S$ and $\mathbf{A}_S$ correspond to a submatrix collecting corresponding rows of $\mathbf{X}$ and columns of $\mathbf{A}$, respectively; $\mathbf{x}_S$ denotes a subvector collecting the corresponding elements of $\mathbf{X}$.

B. Data-driven Sparse GLM

Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{m \times N}$ and $\mathbf{E} = [\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N] \in \mathbb{R}^{m \times N}$, where $\boldsymbol{y}_i \in \mathbb{R}^m$ and $\boldsymbol{\varepsilon}_i \in \mathbb{R}^m$ represent samples of a BOLD signal and the corresponding noise at the i-th voxel, respectively. SPM assumes the following generalized linear model (GLM) [5]:

$$\mathbf{y}_i = \mathbf{D}\mathbf{x}_i + \boldsymbol{\varepsilon}_i, \ i = 1, \ldots, N, \tag{2}$$

where $\mathbf{D} \in \mathbb{R}^{m \times n}$ denotes the regressors, and $\mathbf{x}_i \in \mathbb{R}^n$ denotes the corresponding response signal strength at the i-th voxel. The noise covariance matrix from the fMRI signal is usually modeled as being independent to each voxel [5]:

$$\mathbf{C} = \mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j^T] = \sigma_i{}^2 \boldsymbol{\Lambda} \delta[i-j], \tag{3}$$

where $\boldsymbol{\Lambda}$ denotes the common temporal correlation matrix for all voxels [5], $\sigma_i{}^2$ denotes the unknown variance at the i-th voxel and $\delta[\cdot]$ denotes a discrete delta function. Since the correlation structure $\boldsymbol{\Lambda}$ is common for all voxels, we can apply the same whitening filter $\boldsymbol{\Lambda}^{-1/2}$ for all $\mathbf{y}_i$ [5]; hence, without losing generality, $\boldsymbol{\Lambda}$ can be assumed as identity matrix $\mathbf{I}_m \in \mathbb{R}^{m \times m}$.

Now, for the noise vector $\boldsymbol{\varepsilon}_i \sim N(0, \sigma_i^2 \mathbf{I}_m)$, the pdf of the measurement $\mathbf{y}_i$ is given by

$$f(\mathbf{y}_i) = \frac{1}{(2\pi)^{\frac{m}{2}} \sigma_i^m} \exp\left(-\frac{1}{2\sigma_i^2}(\mathbf{y}_i - \mathbf{D}\mathbf{x}_i)^T(\mathbf{y}_i - \mathbf{D}\mathbf{x}_i)\right). \tag{4}$$

Assuming that noise at each voxel is independent of each other,

the joint probability density function can be represented as:

$$
\begin{aligned}
L(\mathbf{y}_1, \ldots, \mathbf{y}_N) & \\
&= \prod_{i=1}^{N} f(\mathbf{y}_i) \tag{5} \\
&= (2\pi)^{-\frac{mN}{2}} \prod_{i=1}^{N} \sigma_i^{-m} \\
&\quad \cdot \exp\left(-\sum_{i=1}^{N} \frac{1}{2\sigma_i^2}(\mathbf{y}_i - \mathbf{D}\mathbf{x}_i)^T(\mathbf{y}_i - \mathbf{D}\mathbf{x}_i)\right). \tag{6}
\end{aligned}
$$

Then, the log-likelihood function in terms of unknown parameters is given by

$$
\begin{aligned}
l(\mathbf{D}, \mathbf{X}, \boldsymbol{\Lambda}) = & -\frac{m}{2} \sum_{i=1}^{N} \log(2\pi\sigma_i^2) \\
& -\frac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma_i^2}(\mathbf{y}_i - \mathbf{D}\mathbf{x}_i)^T(\mathbf{y}_i - \mathbf{D}\mathbf{x}_i). \tag{7}
\end{aligned}
$$

Unlike the standard GLM model that uses a pre-defined matrix $\mathbf{D}$ as a design matrix [5], the proposed data-driven sparse GLM model assumes $\mathbf{D}$ as an unknown *global* dictionary, of which atom is assumed to indicate a principally dominant neural response in a small set of synchronous neural dynamics. Accordingly, to describe the BOLD signal at each voxel as a sparse combination from the global dictionary, our data-driven sparse GLM model assumes that the signal contribution is sparse, i.e. $\|\mathbf{x}_i\|_0 \leq k$. Then, the maximization of the log-likelihood under the sparsity constraint can be formulated by introducing an unknown support set $\{I_i\}_{i=1}^N$ into the log-likelihood function:

$$
\begin{aligned}
l\left(\mathbf{D}, \mathbf{X}, \boldsymbol{\Lambda}, \{I_i\}_{i=1}^N\right) = & -\frac{m}{2} \sum_{i=1}^{N} \log(2\pi\sigma_i^2) \\
& -\frac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma_i^2}(\mathbf{y}_i - \mathbf{D}_{I_i}\mathbf{x}_{I_i,i})^T(\mathbf{y}_i - \mathbf{D}_{I_i}\mathbf{x}_{I_i,i}) \tag{8}
\end{aligned}
$$

where $\mathbf{x}_{I_i,i}$ denotes a subvector of $\mathbf{x}_i$ collected from elements in the index set I_i. Now, the maximum likelihood formulation is given by

$$\max_{\mathbf{D}, \mathbf{X}, \boldsymbol{\Lambda}, \{I_i\}_{i=1}^N} l\left(\mathbf{D}, \mathbf{X}, \boldsymbol{\Lambda}, \{I_i\}_{i=1}^N\right). \tag{9}$$

Since Eq. (9) is a complicated non-convex nonlinear optimization problem, we address the problem by using alternative maximization. Specifically, we first assume that $\boldsymbol{\Lambda}$ is known and start estimating $\mathbf{D}$ and $\mathbf{X}$. Then, $\boldsymbol{\Lambda}$ can be updated using the new estimates of $\mathbf{D}$ and $\mathbf{X}$. This procedure is repeated until convergence. More specifically, if $\boldsymbol{\Lambda}$ is known, we represent an equivalent optimization problem for the maximum likelihood estimation problem Eq. (9) with respect to $\mathbf{D}, \mathbf{X}$:

$$
\begin{aligned}
\min_{\mathbf{D}, \mathbf{X}} \quad & \left\|\mathbf{Y}\boldsymbol{\Lambda}^{-\frac{1}{2}} - \mathbf{D}\mathbf{X}\boldsymbol{\Lambda}^{-\frac{1}{2}}\right\|_F \\
\text{subject to} \quad & \|\mathbf{x}_i\|_0 \leq k, \tag{10}
\end{aligned}
$$

where $\|\cdot\|_F$ denotes the Frobenius norm, the pseudo-norm

$\|\mathbf{x}\|_0$ denotes the number of nonzero-elements, and

$$\mathbf{\Lambda}^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \sigma_2 & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \sigma_N \end{bmatrix}. \tag{11}$$

Note that the estimation problem Eq. (10) can be addressed by using sparse dictionary learning algorithm [38–40]. Sparse learning algorithms recently have been extensively investigated within the context of compressed sensing, which asserts accurate reconstruction from a limited number of measurements if the underlying signal is sparse and if the sensing matrix is sufficiently incoherent [42–44]. Among the various methods, the sparse decomposition method known as the K-SVD algorithm [38] is widely used in image processing fields such as image compression [45], denoising [46], etc. This is a generalized K-means clustering process, which effectively decomposes signals into a sparse linear combination of dictionary atoms. Specifically, given a set of m-by-N training signals, we search the best possible dictionary $\mathbf{D}$ for the sparse representation of the measurement using Eq. (10). The K-SVD algorithm includes a two-step process per iteration: (i) sparse coding, where we find the best coefficient matrix $\mathbf{X}$ with a fixed $\mathbf{D}$, and (ii) codebook update stage, wherein we change the columns of $\mathbf{D}$ sequentially and the corresponding coefficients. More specifically, for a dictionary estimate $\hat{\mathbf{D}}$, the sparse coding step solves the following for $i = 1, \cdots, N$:

$$\min_{\mathbf{x}_i} \left\| \frac{\mathbf{y}_i}{\sigma_i} - \hat{\mathbf{D}} \frac{\mathbf{x}_i}{\sigma_i} \right\|_2^2, \quad \text{subject to } \| \mathbf{x}_i \|_0 \le k. \tag{12}$$

The sparse coding stage can then be solved using basis persuit (BP) [47, 48], or orthogonal matching pursuit (OMP) [49, 50], and etc. As it will be shown later, simple thresholding based on correlation [51] is computationally efficient, and more robust in finding F-contrast activation map. Hence, we calculate the square of correlation between the measurement vector $\mathbf{y}_i$ and the atom $\mathbf{d}_j$;

$$C_{\mathbf{y}_i}(j) = \frac{\|\mathbf{y}_i^T \mathbf{d}_j\|_2^2}{\sigma_i^2 \|\mathbf{d}_j\|_2^2}, \quad j = 1, \cdots, n . \tag{13}$$

Then, the active index set I_i can be estimated by collecting indices that correspond to the k-largest coefficients from $\{C_{\mathbf{y}_i}(j)\}_{j=1}^n$. The corresponding sparse coding for the signal estimate $\hat{\mathbf{x}}_{I_i, i}$ is then given by

$$\hat{\mathbf{x}}_{I_i, i} = \left(\mathbf{D}_{\hat{I}_i}^T \mathbf{D}_{\hat{I}_i} \right)^{-1} \mathbf{D}_{\hat{I}_i}^T \mathbf{y}_i . \tag{14}$$

Note that in Eq (13), σ_i^2 does not influence the active set estimate I_i since it is common to every $\mathbf{d}_j$.

With estimated $\mathbf{X}$ and $\mathbf{\Lambda}$, K-SVD puts in question only one column in the dictionary, $\mathbf{d}_j$, and the corresponding coefficient $\mathbf{x}^j$, the j-th row of $\mathbf{X}$. This can be solved using singular value decomposition (SVD) with sparsity constraint. Specifically, we first define new matrices

$$\tilde{\mathbf{Y}} = \mathbf{Y} \mathbf{\Lambda}^{-\frac{1}{2}} \tag{15}$$

and

$$\tilde{\mathbf{X}} = \mathbf{X} \mathbf{\Lambda}^{-\frac{1}{2}}. \tag{16}$$

Then, for each $p = 1, 2, \cdots, n$, the K-SVD does the following: (i) define the index set ω^p corresponding to non-zero indices of $\tilde{\mathbf{x}}^p$, (ii) compute $\mathbf{E}_p = \tilde{\mathbf{Y}} - \sum_{j \ne p} \mathbf{d}_j \tilde{\mathbf{x}}^j$, (iii) define $\mathbf{\Omega}_p$ as a diagonal matrix with ones for the indices corresponding to ω^p and zeros elsewhere, (iv) choose a subset $\mathbf{E}_p^{\mathcal{R}} = \mathbf{E}_p \mathbf{\Omega}_p$, (v) take SVD to the restricted $\mathbf{E}_p^{\mathcal{R}}$,

$$\mathbf{E}_p^{\mathcal{R}} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T = \sum_{j=1}^J \sigma_j \mathbf{u}_j \mathbf{v}_j^T \tag{17}$$

and (vi) update $\hat{\mathbf{d}}_p = \mathbf{u}_1$, $\hat{\mathbf{x}}_{\mathcal{R}}^p = \sigma_1 \mathbf{v}_1^T$. Even though all of the dictionary elements can be estimated by data, a fixed atom needs be included to account for measurement error. Notably, the temporal variation of BOLD signal usually drifts due to vaso-motion, breathing, etc. Hence, a constant dictionary atom $\mathbf{d}_1 = [1, \ldots, 1]^T \in \mathbb{R}^m$ is used to account for DC-bias and drift. In this case, the dictionary update step is modified in order not to update the fixed atom.

C. Minimum Description Length Principle For Sparsity Level

The K-SVD algorithm is often sensitive to the choice of sparsity level k. Since the sparsity k affects the whole performance of dictionary learning, a criteria to determine the optimal k needs to be investigated. This problem corresponds to a model order selection problem that finds the optimal k having the best trade-off between filedity and model complexity. Examples of effective model order selection include minimum description length (MDL) suggested by Rissanen [41], Akaike information criterion (AIC) [52] and Schwartz information criterion (SIC) [53]. Among these, we chose the MDL method due to its asymptotic consistency [41]. According to MDL principle, the best choice of model order n_0 to explain the data $\mathbf{y}$ is the one which minimizes

$$\mathbf{MDL}(n_0) = \mathbf{L}(\mathbf{y}|n_0) + \mathbf{L}(n_0), \tag{18}$$

where $\mathbf{L}(\mathbf{y}|n_0)$ is the goodness-of-fit of $\mathbf{y}$ when encoded with n_0, and $\mathbf{L}(n_0)$ is the code length in bits to encode the model itself. In our problem, k number of regressors are distinctly selected for each voxel, so the total model order becomes $n_0 = kN$.

In order to apply MDL, we first need to calculate the value of log likelihood. This can be done by applying ML estimate of $\mathbf{D}, \mathbf{X}, \mathbf{\Lambda}, \{I_i\}_{i=1}^N$. Since $\mathbf{D}, \mathbf{X}, \{I_i\}$ are estimated using K-SVD, the remaining unknown parameter is $\mathbf{\Lambda}$. This value can be calculated by

$$\frac{\partial}{\partial \sigma_i^2} l(\mathbf{D}, \mathbf{X}, \mathbf{\Lambda}, \{I_i\}) = 0, \tag{19}$$

resulting in

$$\sigma_i^2 = \frac{1}{m} (\mathbf{y}_i - \mathbf{D}_{I_i} \mathbf{x}_{I_i, i})^T (\mathbf{y}_i - \mathbf{D}_{I_i} \mathbf{x}_{I_i, i}). \tag{20}$$

By plugging the ML estimate Eq (14) into Eq (20), we have

$$\sigma_i^2 = \frac{1}{m} \mathbf{y}_i^T \mathbf{P}_{\mathbf{D}_{I_i}}^\perp \mathbf{y}_i \tag{21}$$

where $\mathbf{P}_{\mathbf{D}_{I_i}}^{\perp}$ denotes the projector associated with the orthogonal complement of the range space of $\mathbf{D}_{I_i}$. Therefore, the code length for the goodness-of-fit becomes

$$
\begin{aligned}
\mathbf{L}(\mathbf{y}|n_0) &= -\log_2(P(\mathbf{y}|n_0)) \\
&= \frac{m}{2}\sum_{i=1}^{N}\log_2(2\pi\hat{\sigma}_i^2) \\
&= \frac{m}{2}\sum_{i=1}^{N}\log_2\left(\frac{2\pi}{m}\mathbf{y}_i^T\mathbf{P}_{\mathbf{D}_{\hat{I}_i}}^{\perp}\mathbf{y}_i\right) \quad \text{(bits).} \quad (22)
\end{aligned}
$$

Now, the code length $\mathbf{L}(n_0)$ for MDL criterion can be described by the code length for the location of nonzero coefficient and magnitude as:

$$
\mathbf{L}(n_0) = \frac{1}{2}n_0\log_2 n + n_0\log_2 n = \frac{3}{2}n_0\log_2 n, \quad (23)
$$

where n is the number of dictionary atoms. The MDL prior given in Eq. (23) is often called Saito's MDL [54].

Note that for a small n_0, $\mathbf{L}(\mathbf{y}|n_0)$ becomes large whereas $\mathbf{L}(n_0)$ becomes small, and for a large n_0, $\mathbf{L}(\mathbf{y}|n_0)$ becomes small whereas $\mathbf{L}(n_0)$ becomes large. Hence, the sum of the two description code lengths can be minimized at an appropriate n_0 value, exhibiting a trade-off between goodness-of-fit and complexity of the models. Thus, we can solve the optimization problem Eq. (18) for various n_0 values, and choose the n_0 value that gives the minimum cost. We are aware that MDL criterion was also used to estimate the number of independent components from the aggregate dataset in the group ICA analysis [30]. They calculated both AIC and MDL estimates and used the average of the two as the number of components to utilize the property of statistical consistency of MDL criterion and lower signal-to-noise ratio advantages of AIC. However, to our best knowledge, we are not aware of any prior work that uses MDL for sparsity level selection in sparse dictionary learning.

D. Activation Detection

So far, we presented a maximum likelihood approach for estimating parameters for the proposed data-driven sparse GLM model. This section presents a statistical test for detecting activated pixels. In a classical inference, a binary hypothesis test is performed in which the null hypothesis is tested against the alternative hypothesis. Ardekani et al [55] describes three properties that are often required for hypothesis testing in fMRI. First, it is desirable that a hypothesis test is invariant to a scale factor for the measurement. Second, the test needs invariance to a rotation of the response signal in the signal subspace. Third, the test has to be invariant to an unknown bias in the nuisance subspace [55].

For example, consider a given GLM model

$$
\mathbf{y}_i = \mathbf{D}\mathbf{x}_i + \varepsilon_i = [\mathbf{A}\ \mathbf{B}]\mathbf{x}_i + \varepsilon_i \quad (24)
$$

where $\mathbf{A}$ and $\mathbf{B}$ denotes submatrices of response signal subspace (to test) and nuisance subspace, respectively. In case the response and nuisance subspace are orthogonal, Ardekani et al shows that the only test that satisfies aforementioned

properties is in the following form [55]

$$
F_i = \frac{\mathbf{y}_i^T\mathbf{P}_{\mathbf{A}}\mathbf{y}_i}{\mathbf{y}_i(\mathbf{I}-\mathbf{P}_{\mathbf{A}}-\mathbf{P}_{\mathbf{B}})\mathbf{y}_i}. \quad (25)
$$

If the response and nuisance subspace are not orthogonal, (i.e. $\mathbf{A}^T\mathbf{B}\neq 0$), the corresponding test is given by [55],

$$
F_i = \frac{\mathbf{y}_i^T\mathbf{P}_{\mathbf{A}'}\mathbf{y}_i}{\mathbf{y}_i(\mathbf{I}-\mathbf{P}_{\mathbf{A}'}-\mathbf{P}_{\mathbf{B}})\mathbf{y}_i}, \quad (26)
$$

and $\mathbf{A}'$ is a projection of $\mathbf{A}$ onto orthogonal complement of $\mathbf{B}$:

$$
\mathbf{A}' = \mathbf{P}_{\mathbf{B}}^{\perp}\mathbf{A}. \quad (27)
$$

Note that we are interested in testing whether a specific temporal dynamics from a global dictionary $\mathbf{D}$ is present in a specific voxel. In most of the cases, we do not know the sign of the dynamics, so t-test is not appropriate. Usually, the neural dynamics we want to test is a specific atom from a global dictionary that have learned all the temporal dynamics from all voxels using sparse dictionary learning. Hence, at each voxel, a binary hypothesis test is performed in which the null hypothesis $H_0 : \theta_i = 0$ is tested against the alternative hypothesis $H_1 : \theta_i \neq 0$, where θ_i is the response signal from

$$
\mathbf{y}_i = \mathbf{z}\theta_i + \mathbf{D}_{I_i\backslash\mathbf{z}}\mathbf{x}_{I_i,i} + \varepsilon_i \quad (28)
$$

where $\mathbf{z}$ denotes an atom from the global dictionary $\mathbf{D}$ that contains a neural dynamics of interest, and $\mathbf{D}_{I_i\backslash\mathbf{z}}$ denotes a reduced size local design matrix made by removing atom $\mathbf{z}$ from $\mathbf{D}_{I_i}$ (If $\mathbf{z}$ is not an atom of the local design matrix $\mathbf{D}_{I_i}$, then $\mathbf{D}_{I_i\backslash\mathbf{z}} = \mathbf{D}_{I_i}$). We can easily see when $\mathbf{z}$ does not belong to the set of atoms from $\mathbf{D}_{I_i}$, $\theta_i = 0$. Hence, H_0 holds. In other cases, using Eq (26), we have

$$
\begin{aligned}
F_i &= \frac{\mathbf{y}_i^T\mathbf{P}_{\mathbf{P}_{\mathbf{D}_{I_i\backslash\mathbf{z}}}^{\perp}\mathbf{z}}\mathbf{y}_i}{\mathbf{y}_i^T\mathbf{P}_{\mathbf{D}_{I_i}}^{\perp}\mathbf{y}_i}\frac{m-k}{q_1} \\
&= \frac{\mathbf{y}_i^T(\mathbf{P}_{\mathbf{D}_{I_i\backslash\mathbf{z}}}^{\perp}-\mathbf{P}_{\mathbf{D}_{I_i}}^{\perp})\mathbf{y}_i}{\mathbf{y}_i^T\mathbf{P}_{\mathbf{D}_{I_i}}^{\perp}\mathbf{y}_i}(m-k) \quad (29)
\end{aligned}
$$

where $\mathbf{P}_{\mathbf{D}_{I_i}}^{\perp}, \mathbf{P}_{\mathbf{D}_{I_i\backslash\mathbf{z}}}^{\perp}$ denote the projection on the orthogonal complement on the range space of $\mathbf{D}_i$ and $\mathbf{D}_{I_i\backslash\mathbf{z}}$, respectively. We use the projector update rule $\mathbf{P}_{\mathbf{D}_{I_i}} = \mathbf{P}_{\mathbf{P}_{\mathbf{D}_{I_i\backslash\mathbf{z}}}^{\perp}\mathbf{z}} + \mathbf{P}_{\mathbf{D}_{I_i\backslash\mathbf{z}}}$, and the constants come from the degree of freedom, i.e. $m-rank(D) = m-k$ and $q_1 = rank(\mathbf{P}_{\mathbf{D}_{I_i\backslash\mathbf{z}}}^{\perp}\mathbf{z}) = 1$. Since the main purpose of the proposed model is to extract time-series components which represent synchronous, functionally related neural hemodynamics, we can test different atoms from a global dictionary $\mathbf{D}$ (which are trained by sparse learning algorithm) to identify where the specific temporal dynamic signal, represented by the atom $\mathbf{z}$, is originated from. For example, in the case of the analysis of block-paradigm or event-related task experiments, we can select the atom $\mathbf{z}$ by choosing the atom which has the highest correlation with the original stimulus function or predefined HRF. In case of resting state analysis, the original stimulus is not known, and each atom from the trained global dictionary $\mathbf{D}$ needs to be tested to identify activated regions that have similar temporal dynamics.

These identified and separated regions can be used as network nodes for functional connectivity MRI [56].

Recall that in K-SVD sparse coding stage, any pursuit algorithm can be used to obtain the k-sparse vector $\mathbf{x}_i$. However, we found that simple thresholding detection was better than OMP for sparse dictionary learning from our experimental results. This can be explained as follows. If $\mathbf{z}$ is an atom of the design matrix $\mathbf{D}_{\hat{I}_i}$ and $\mathbf{z} \cong \mathbf{P}_{\mathbf{D}_{I_i \setminus \mathbf{z}}}^{\perp}$, the square of correlation of the i-th column vector $\mathbf{y}_i$ with an atom $\mathbf{z}$ in Eq. (13) can be rewritten as

$$
\begin{aligned}
C_{\mathbf{y}_i}(j) &= \frac{\|\mathbf{y}_i^T \mathbf{z}\|^2}{\hat{\sigma}_i^2 \|\mathbf{z}\|^2} = \frac{\mathbf{y}_i^T \mathbf{P_z} \mathbf{y}_i}{\mathbf{y}_i \mathbf{P}_{\mathbf{D}_{\hat{I}_i}}^{\perp} \mathbf{y}_i} \\
&\simeq \frac{\mathbf{y}_i^T (\mathbf{P}_{\mathbf{D}_{I_i \setminus \mathbf{z}}}^{\perp} - \mathbf{P}_{\mathbf{D}_{I_i}}^{\perp}) \mathbf{y}_i}{\mathbf{y}_i^T \mathbf{P}_{\mathbf{D}_{I_i}}^{\perp} \mathbf{y}_i}.
\end{aligned} \tag{30}
$$

due to the projector update rule. Hence, it selects atoms that have the highest F-values under this situation. The orthogonality is often observed when the dynamic signal varies with an alternating pattern whereas the slow varying background and/or DC components make up the nuisance space. In addition, as shown in the following section, the complexity of K-SVD with simple correlation is much lower than that of K-SVD with OMP. Therefore, we prefer to use a simple thresholding scheme for dictionary learning.

E. Complexity analysis

Note that the most time consuming part of the algorithm is the K-SVD dictionary learning step. Hence, we conducted a complexity analysis of the K-SVD step with a simple thresholding algorithm and OMP using the number of required multiplication [57] [2]. For the analysis, we consider a signal $\mathbf{y}_i \in \mathbb{R}^m$ and a dictionary $\mathbf{D} \in \mathbb{R}^{m \times n}$ with sparsity level k. In OMP implementation, the operation at the k-th iteration includes multiplication of $\mathbf{D}^T$ with residual ($T_{\mathbf{D}^T}$ number of multiplications), update of the coefficient vector, and a back-substitution to update the residual. Hence, according to [57], the total number of multiplication of the OMP summing over all k iterations is therefore

$$
T_{OMP} \approx k^3 + k \cdot T_{\mathbf{D}^T} = k^3 + 2kmn. \tag{31}
$$

The main difference of thresholding compared to OMP is that the support set is estimated once using the square of correlation of the residual with the global dictionary $\mathbf{D}$, and then the coefficients are calculated using matrix inversion. Hence, the number of multiplication is

$$
T_{Thr} \approx k^3 + T_{\mathbf{D}^T} = k^3 + 2mn. \tag{32}
$$

The remaining part, the dictionary update step, is then applied as follows. The dominant operations in a single K-SVD iteration include sparse-coding, atom updates, and coefficient updates. The main difficulty in calculating the complexity is that the atom update complexity depends on the number of signals using it [57]. However, with an elegant cumulative

[2] We assume that the complexity of finding the k-largest coefficient for the case of thresholding is negligible compared to the multiplication since $k \ll n$.
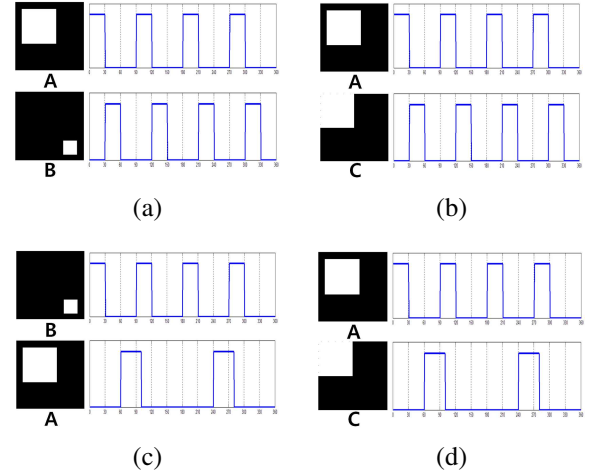

Fig. 2. Simulated activation patterns for (a) temporally and spatially uncorrelated events, (b) spatially dependent events, (c) temporally dependent events, and (d) temporally and spatially dependent events. The visual patterns used in each temporal combination are shown on left side of the temporal dynamics.

analysis, Rubinstein et al [57] show that the total computational complexity of K-SVD with OMP is given by

$$
\begin{aligned}
&T_{K-SVD(OMP)} \\
&= N \cdot T_{OMP} + mn^2 + 4mNk + 4Nkn + 4mn^2 \\
&= N \cdot \left(k^3 + 2kmn + 4mk + 4kn \right) + 5mn^2. \tag{33}
\end{aligned}
$$

Similarly, we can calculate the K-SVD with thresholding as:

$$
\begin{aligned}
&T_{K-SVD(Thr)} \\
&= N \cdot T_{Thr} + mn^2 + 4mNk + 4Nkn + 4mn^2 \\
&= N \cdot \left(k^3 + 2mn + 4mk + 4kn \right) + 5mn^2. \tag{34}
\end{aligned}
$$

Hence, the computational complexity is proportional to the voxel size and the complexity with thresholding is much smaller than the one with OMP.

In addition, due to computational complexity of K-SVD for large number of voxels ($N \gg 1$), we made the following two approximations. First, rather than using all N voxels for K-SVD algorithm, we downsampled the number of voxels. The down-sampling factor we used for the experiment was 64. Second, rather than successively applying K-SVD using Eq. (10) with newly updated $\boldsymbol{\Lambda}$, we assumed a constant $\boldsymbol{\Lambda}$ value and performed K-SVD algorithm for sparse dictionary learning only once. Even though the proposed simplification makes the algorithm sub-optimal in theory, we found that the overall performance of the algorithm was basically similar according to our experiments.

III. METHOD

We used the proposed method on four different fMRI datasets: (i) simulation data, (ii) block-paradigm auditory stimulus task dataset of a single subject (SPM open-dataset: http://www.fil.ion.ucl.ac.uk/spm/), (iii) block-paradigm RFT task dataset of four subjects, and (iv) event-related RFT task dataset of four subjects.
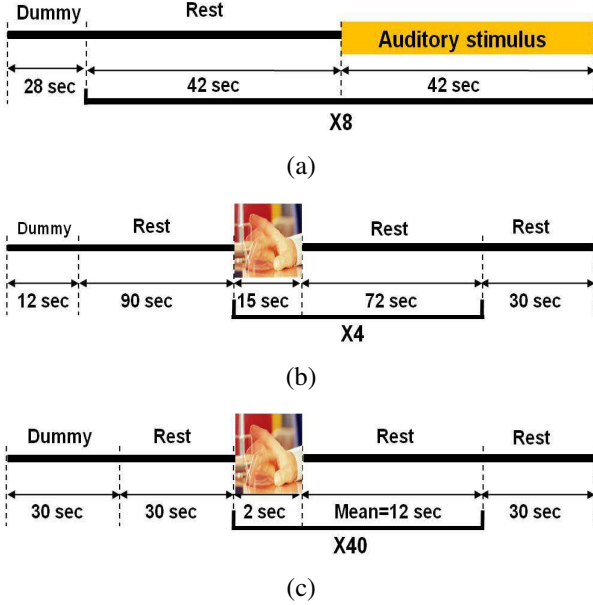
(a)

(b)

(c)

Fig. 3. Experimental paradigms for (a) auditory stimulus tasks, (b) block paradigm right finger tapping tasks, and (c) event-related right finger tapping tasks.

A. Simulation method

We generated simulated data to test the validity of the proposed method in comparison to sICA, tICA, and PCA. Two pairs of temporal waveforms were created containing similar paradigms as used in the simulation by Calhaun et al [58]. The temporal paradigms were each $360\ sec$ and visual patterns comprised of box signals were repeated as shown in Figure 2. Three different visual patterns of 10×10 voxels were created, with amplitudes of 1 in $\{2, \ldots, 6\} \times \{2, \ldots, 6\}$ for pattern A, $\{8, 9\} \times \{8, 9\}$ for pattern B, $\{1, \ldots, 5\} \times \{1, \ldots, 5\}$ for pattern C, and 0 elsewhere. Additionally, random white Gaussian noise with $\mathcal{N}(0, 0.11)$ was added. The spatial patterns and the corresponding time series in Figure 2 represent four different scenarios: temporally and spatially uncorrelated events (Figure 2(a)), spatially dependent events (Figure 2(b)), temporally dependent events (Figure 2(c)), and temporally and spatially dependent events (Figure 2(d)). These four datasets were analyzed using the proposed method, sICA, tICA and PCA. For the proposed method, we used 1-sparsity $(k = 1)$ and the dictionary size of 3 $(n = 3)$, which was learned using K-SVD algorithm with 20 iterations. During the sparse coding stage, simple thresholding algorithm based on the square of correlation was employed. We used FastICA software toolbox for sICA and tICA [32]. Prewhitening and dimension reduction into 3 components were implemented simultaneously using a principle component analysis (PCA). Three independent components were then learned for both ICA methods with 'pow3' nonlinearity function.

B. Behavior protocol and data acquisition

1) Auditory stimulus task: Auditory stimulus task datasets of a single subject were used to compare the results of the

proposed method with the results of conventional methods. As shown in Figure 3(a), a total of 96 acquisitions (TR= $7\ sec$), giving sixteen $42\ sec$ blocks, were used for the analysis except for the first 'dummy' 4 scans. The condition for successive blocks alternated between rest and auditory stimulation, starting with rest. Auditory stimulation was with bi-syllabic words presented binaurally at a rate of 60 per minute. Whole brain BOLD/EPI images were acquired on a modified 2T Siemens MAGNETOM Vision system. Each acquisition consisted of 64 contiguous slices, with a matrix size of 64×64 and in-plane resolution of $3\ mm \times 3\ mm \times 3\ mm$. The data set was obtained from an open-dataset in the SPM5 software package (Wellcome Department of Cognitive Neurology, London, UK)(http://www.fil.ion.ucl.ac.uk/spm/).

2) Block-paradigm right finger tapping task: The proposed method was also applied to a block-paradigm right finger tapping (RFT) task to evaluate the performance. A $15\ sec$ task period alternated with a $72\ sec$ resting period was repeated 4 times for each subject followed by an additional $30\ sec$ of rest as illustrated in Figure 3(b). The total recording time was $480\ sec$. During the task period, subjects were instructed to perform right finger flexion, and to focus on a fixed point in the resting time to minimize eye movement, thinking, and so on. A total of 4 healthy right-handed subjects were examined (mean age = 25.4 ± 2.3 years). A 3.0T funtional MRI system (ISOL, Republic of Korea) was used to measure the BOLD response. During the experiment with the blocked task paradigm, the echo planar imaging (EPI) sequence was used with TR/TE = $3000/35\ ms$, flip angle = $80°, 35$ slices, and $4\ mm$ slice thickness. Each acquisition consisted of 35 continuous slices, with a matrix size of 64×64 and in-plane resolution of $3.44\ mm \times 3.44\ mm \times 4\ mm$. In the subsequent anatomical scanning session, T1-weighted structural images were acquired. Four dummy scans were also discarded.

3) Event-related right finger tapping task: We also used an event-related RFT task. The BOLD responses and T1-weighted structural images were acquired using the same scanner. A total of 4 healthy subjects were examined (mean age = 25.7 ± 2.2 years). The EPI sequence was used with TR/TE = $2000/35\ ms$, flip angle = $80°$, and $4\ mm$ slice thickness. Each acquisition consisted of 24 contiguous slices, with a matrix size of 64×64 and in-plane resolution of $3.44\ mm \times 3.44\ mm \times 3.44\ mm$. The total recording time was $650\ sec$. After a dummy scan of $30\ sec$, the right finger tapping task period and resting period were repeated 40 times followed by an additional $30\ sec$ of rest (see Figure 3(c)). For the resting period immediately after the task, the interstimulus interval (ISI) ranged between 4 and $20\ sec$ with an average ISI period of $12\ sec$. Fifteen dummy scans were also discarded.

C. Signal Processing

1) Preprocessing for SPM: The images were first spatially realigned to correct changes in signal intensity over time, which can arise from within-subject head motion during the scanning session. The maps were then spatially normalized to a standard Talairach space and resampled to $2\ mm \times 2\ mm \times 2\ mm$ voxels. The preprocessed data were used as
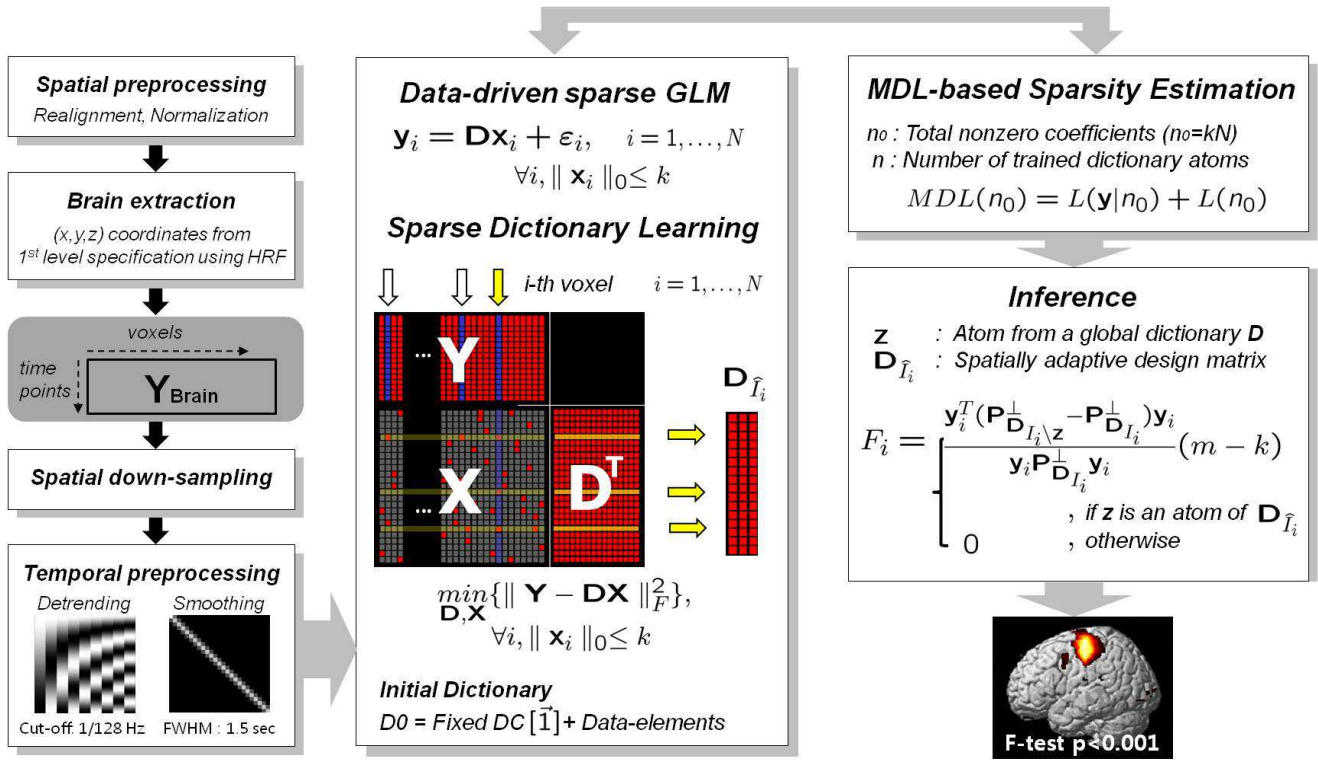
Fig. 4. Block diagram of the proposed method.

a measurement in the GLM model. Spatial smoothing was then applied with a $8\,mm \times 8\,mm \times 8\,mm$ full-width at half maximum (FWHM) Gaussian kernel.

2) Data-driven sparse GLM: The BOLD time-series data was arranged as a matrix $\mathbf{Y} \in \mathbb{R}^{m \times N}$, where m is the number of time points and N is the number of the voxels. The data was then down-sampled 8 times along the spatial direction to reduce the computation time in K-SVD learning. We used a discrete cosine transform (DCT) basis set with a cutoff frequency of $1/128\,Hz$ to eliminate low frequency drifts, and thereby to improve the signal-to-noise ratio. After detrending, the data were temporally smoothed using $1.5\,sec$ full-width at half maximum of the Gaussian kernel to remove high frequency noise. The data were then decomposed using the K-SVD algorithm, where correlation based thresholding method was used in the sparse coding stage. We first initialized the dictionary by the data elements, and the first atom in the dictionary was set to be the DC component in order to capture the remaining drift signal. A total of 30 iterations of K-SVD learning were performed for every dataset. The maximum number of sparsity (k) was determined using MDL criterion by varying k from 1 to 10, whereas the number of dictionary elements to train (n) was set to 40. After the dictionary learning with the optimal sparsity k at each voxel, the nonzero k atoms are used as a local design matrix. Then, F-map was calculated using Eq. (29) offline, and the resulting F-map and degree of freedom were imported to SPM5 to obtain the activation map for a given p-value. Random field correction was used. The above process is illustrated in Figure 4.

3) Independent component analysis: We used the GIFT software package [30] to separate spatially independent sources from the data. Prewhitening and dimension reduction to 40 components were simultaneously conducted using principle components analysis (PCA). The output of the PCA was used to estimate 40 independent components, which is the same as the size of the dictionary yielded by K-SVD learning. This number provided a fairly reasonable separation within the estimates for previous results [30, 58] when considering data size. We used Infomax [31] and FastICA [32] with a pow3 nonlinearity function to compare the results with the proposed method. Spatially independent images and the corresponding time courses were obtained to construct the regressors in a GLM analysis. In addition, we conducted tICA using the FastICA algorithm [58]. In tICA, we downsampled the data 4 times in the slice direction, and implemented the FastICA algorithm for each slice. In this case, 10 temporal ICs were learned for each slice and collected. After we captured time courses from sICA or tICA, we computed the correlation coefficient of learned ICs with canonical HRF. We selected the most correlated time-series components as a design matrix to implement a GLM analysis through the SPM5 software package. We also calculated the F-map to show the activation maps.

IV. EXPERIMENTAL RESULTS

A. Simulation results

Figure 5 illustrates the results from the analysis of simulated experiments. The time courses are given by a trained dictionary (data-driven sparse GLM), pseudo-inverse of the unmixing matrix $\mathbf{W}^{-1}$ (sICA), temporally independent components (tICA), and principal component (PCA), respectively.
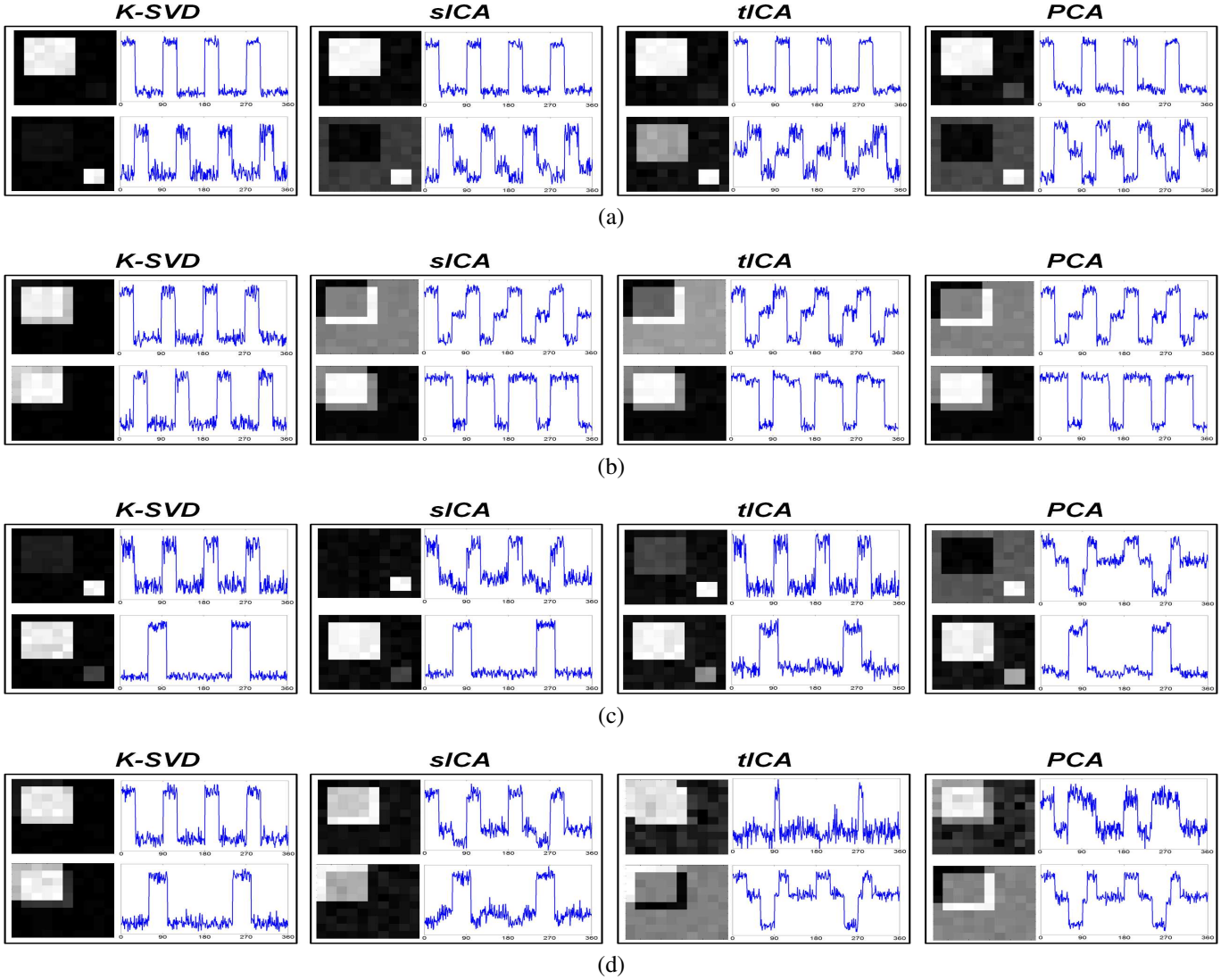
Fig. 5. Simulation results for the activation scenario given in Figure 2. Extracted time series and the corresponding z-maps are shown according to the following order: (a) temporally and spatially uncorrelated events, (b) spatially dependent events, (c) temporally dependent events and (d) temporally and spatially dependent events.

Note that PCA failed to resolve the two activations for all the simulated scenarios. When two activation areas are not spatially overlapped as shown in Figs. 5(a)(c), both sICA and tICA successfully decomposed the two times series which are uncorrelated (Fig. 5(a)) and correlated (Fig. 5(c)). When activation areas are significantly overlapped as in Figs. 5(b)(d), neither sICA nor tICA are successful and the resulting z-maps were erroneous.

However, in all the simulation scenarios, the time course extracted using the proposed method closely followed the original temporal responses, confirming that the proposed method can effectively extract the sparse components. Moreover, the activation map by z-contrast follow the ground-truth accurately. This coincides with the box simulation result in [34], which showed that the success of separation using sparsity is less sensitive to overlapped region than using independency.

B. Auditory stimulus task results

Figure 6 illustrates the surface projection of F-statistics maps from a fMRI individual analysis of the auditory stimulus task at a random field correction $p < 0.001$. The regressors in the design matrix for a GLM analysis are constructed by: (a) canonical HRF, (b) data-driven sparse GLM, (c) sICA (Infomax), (d) sICA (FastICA), (e) tICA (FastICA), and (f) PCA, as shown in the right side of the maps. The maximum number of sparsity (k) was chosen as 2 according to our MDL criterion. The correlation coefficients of the most correlated atom in each dictionary with the canonical HRF are: (b) 0.8214, (c) 0.8072, (d) 0.6045, (e) 0.7430, and (f) 0.5778, which showed that the extracted time course by the proposed algorithm is the most correlated. When we use the signal components that are the most correlated with canonical HRF, the results show that the neural activation in auditory area are correctly identified, except for sICA using FastICA.
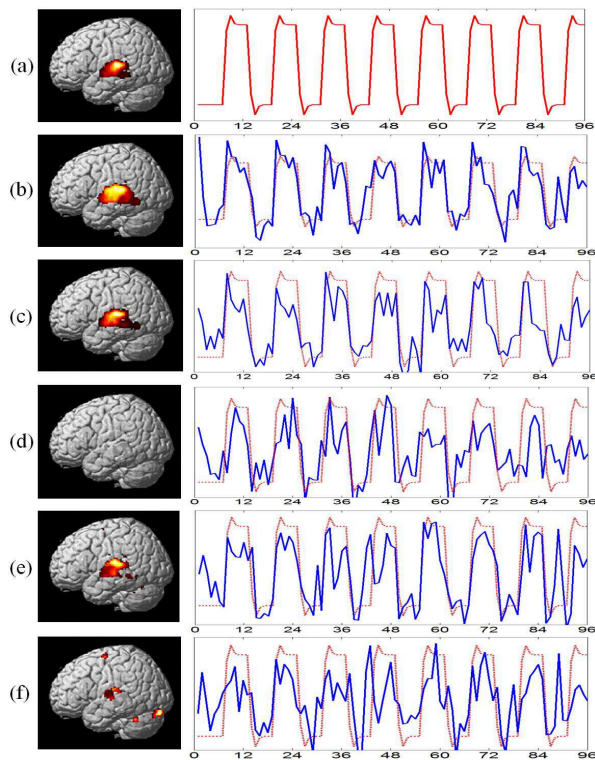
Fig. 6. F-statistics activation maps for auditory stimulus dataset of single subject at a random field correction $p < 0.001$ using design matrices constructed by (a) a canonical HRF, (b) data-driven sparse GLM, (c) sICA using Infomax algorithm, (d) sICA using FastICA algorithm, (e) tICA using FastICA algorithm, and (f) PCA, respectively. The dotted lines correspond to the canonical HRF convolved with experimental paradigm, and the solid lines correspond to the individual task-related components extracted using each data-driven decomposition methods.
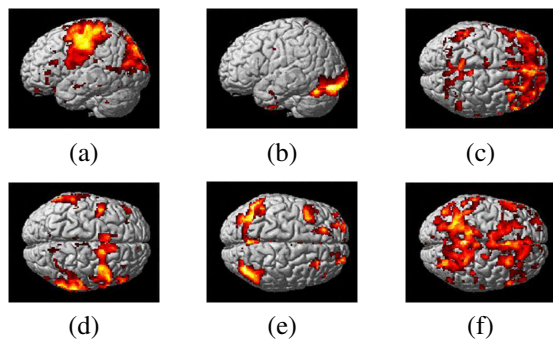
Fig. 7. F-statistics activation maps with a random field correction $p < 0.001$ for auditory task dataset with respect to various non-task related signal components.

Even though a ground truth of activation is not available to quantitatively compare the spatial response pattern, the data set we used is a standard dataset for validation of hypothesis-driven methods, and thus a close match with the canonical HRF confirms the validity of the algorithm. In Figure 7, we also show the F-statistics maps corresponding to other non-task related signal atoms at a random field correction $p < 0.001$. In this case, the regressor $\mathbf{z}$ is determined as $\mathbf{z} = \mathbf{Dc}$ where $\mathbf{c}$ is a given contrast. We used the elementary contrast vector, i.e. $\mathbf{c} = \mathbf{e}_i \in \mathbb{R}^n$ that has zeros except

for 1 element at the i-th location, where the i-index varies depending on the neural dynamics of interest. As discussed before, the voxels corresponding to each global dictionary atom are assumed as neural dynamics which are activated and localized in a small set of area. Hence, by varying $\mathbf{c}$, we can test the hypothesis that a specific neural dynamics $\mathbf{z}$ is originated from a specific voxel. These maps indicate sparsely distributed patterns of task-related signals, nontask-related signals and etc. The activations from motor cortex and somatosensory cortex are shown with secondary visual cortex in Figure 7(a). Similarly, we can observe activation at the visual cortex in Figure 7(b), prefrontal area in Figure 7(c), parietal lobe in Figures 7(d),(e) and (f), and etc, which are obtained by changing the contrast vector.

C. Block-paradigm right finger tapping task

We employed our approach to block-paradigm RFT task data. Figure 8 illustrates the surface projections of F-statistics maps from four fMRI individual data analysis at a random field correction $p < 0.001$. The regressors in the design matrix for the GLM analysis are constructed by canonical HRF, data-driven sparse GLM with K-SVD, sICA using Infomax, sICA using FastICA, and PCA. The proposed MDL criteria selected the optimal sparsity as $k = 2, 2, 2$ and 3 for each subjects, respectively. The most correlated atom with the canonical HRF and DC components are used as regressors in the design matrix for sICA, tICA and PCA, as suggested in [26, 58]. In the proposed method, we used a contrast vector for F-map to identify an atom that was the most correlated with canonical HRF for activation detection. Our method clearly localized activations in the target region related to the task.

In Figure 9, the task-related components of 4 subjects are superimposed to compare the reliability of the algorithms. The mean time course extracted by the proposed method resembles that of the canonical HRF, showing consistent results of task-related activity across the 4 task cycles. This shows that the proposed method reliably detects neural activation.

D. Event-related right finger tapping task

Event-related RFT task fMRI data were also studied. Figure 10 illustrates the surface projection of F-statistics maps from 4 individual analysis at a random field correction $p < 0.001$. The regressors in the design matrix for the GLM analysis are constructed by the canonical HRF, data-driven sparse GLM with K-SVD, sICA using Infomax, sICA using FastICA, and PCA. Our MDL criterion selected the optimal sparsity level $k = 6, 6, 9$ and 7 for each subject, respectively. The contrast vector for F-map was again selected to identify the most correlated atom with canonical HRF. The trained dictionary using our method extracts clearly localized activation in the target region related to the task, outperforming the conventional data-driven methods with event-related fMRI. Unlike the ICA, our algorithm was able to separate the motor area from other areas such as the auditory cortex for all of the individuals. In Figure 11, we also show the F-statistics maps using contrast vector to select atoms that correspond to non-task related atoms of subject 2 at a random field
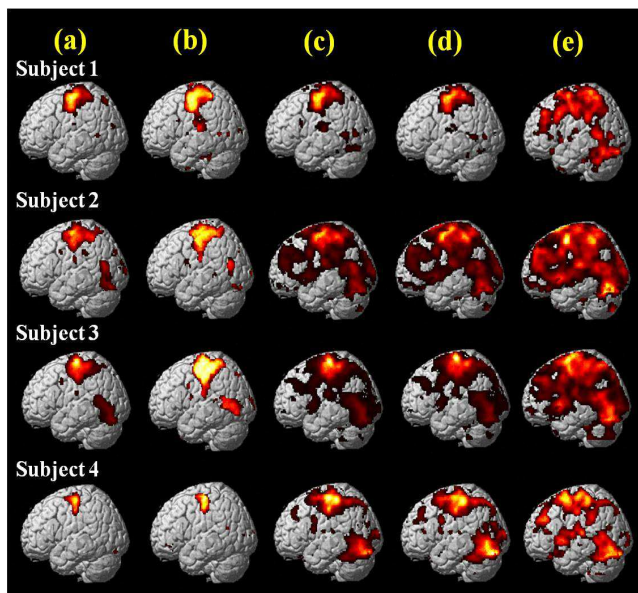
Fig. 8. *F*-statistics activation maps for block-paradigm right finger tapping task dataset of 4 subjects at a random field correction $p < 0.001$ using the design matrices constructed by (a) canonical HRF, (b) data-driven sparse GLM, (c) sICA using Infomax, (d) sICA using FastICA, and (e) PCA.
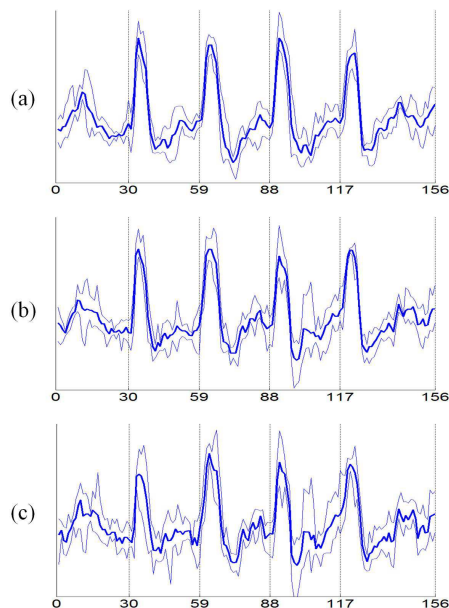

Fig. 10. *F*-statistics activation maps for event-related right finger tapping task dataset of 4 subjects at a random field correction $p < 0.001$ using the design matrices constructed by (a) canonical HRF, (b) data-driven sparse GLM, (c) sICA using Infomax, (d) sICA using FastICA, and (e) PCA.

Fig. 9. Reliability of task-related temporal trace of 4 subjects using: (a) data-driven sparse GLM, (b) sICA using Infomax, and (c) sICA using FastICA. The bold line is the mean time course of each task-related component used in individual analysis for 4 subjects. The maximum and minimum values at each time points are shown above and below.

Fig. 11. *F*-statistics activation maps with a random field correction $p < 0.001$ for event-related right finger tapping task dataset of subject 2 with respect to various non-task related signals.

correction $p < 0.001$. The proposed method extracted parietal lobe and secondary frontal area in Figures 11(a) and (b). Furthermore, the results showed activations from the visual cortex in Figure 11(c), right motor cortex in Figure 11(d), prefrontal area in Figure 11(e), and etc. The maps in Figure 11(f) may be due to subject head movement, since similar motion artifacts were obtained from previous ICA results [26, 59]. Considering complex interconnection of the brain network during the experiments, it is expected that the areas
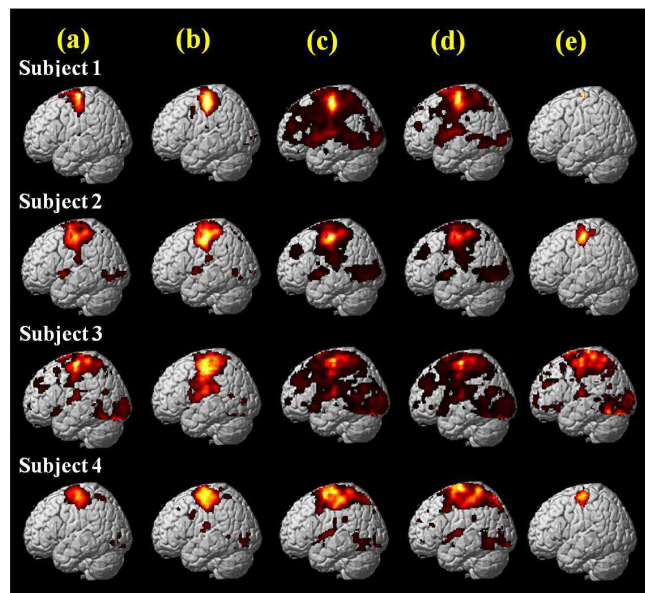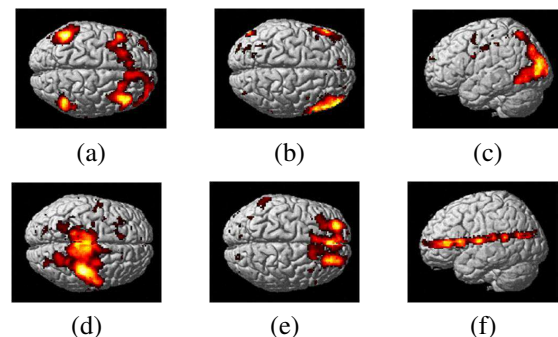
may be activated during information processing with distinct temporal dynamics. Our results clearly identify such dynamics by changing the contrast vector for *F*-map.

E. Choice of k

In order to show the effectiveness of our MDL criteria for sparsity level selection, we illustrated the results by varying sparsity level in Figure 12 to show the sensitivity of the K-SVD sparse dictionary learning with respect to sparsity. Note that the results are dependent upon the sparsity level *k*. The optimal *k* decided by MDL were (a) 2, (b) 3, and (c) 6, respectively, which clearly show the best trade-off between sensitivity and specificity. Extensive experiments with other dataset also confirm the accuracy of the MDL based sparsity selection.

F. OMP vs Simple thresholding

We also performed the comparative study of OMP and simple thresholding analysis in the sparse coding stage (See
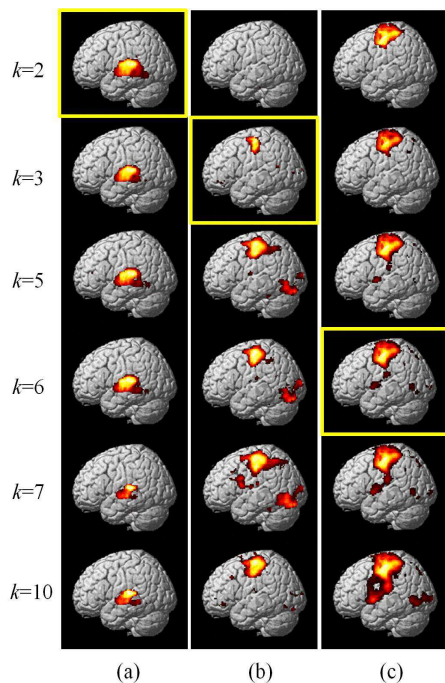
Fig. 12. Activation maps acquired by varying sparsity level: (a) auditory stimulus dataset, (b) block paradigm RFT task, and (c) event-related RFT task. The optimal k determined by MDL were 2,3 and 6, respectively. Box indicates the images corresponding to the optimal sparsity level.
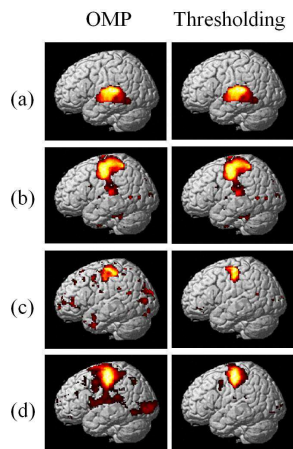

Fig. 13. Comparison of OMP and thresholding for the data-driven sparse GLM: the activation maps from auditory stimulus dataset in (a), block paradigm RFT task in (b) and (c), and event-related RFT task in (d) are shown. The optimal k determined by MDL were 2, 2, 3 and 6, respectively.

Figure 13). Although the optimal k determined by MDL were the same for both methods, the ability of separating the source signal from the data was somewhat different. In Figure 13(a) and (b), OMP as well as the simple thresholding method successfully extract the task-related dictionary atom. However, OMP failed in extracting correct activation map in Figure 13(c) and (d). Similar behaviors were obtained in other dataset. Therefore, we chose simple thresholding for our study.

V. CONCLUSION AND DISCUSSION

In order to decompose a BOLD time-series that represents distinct brain dynamics including task-related and non-task-related signal components, we presented a novel data-driven sparse GLM framework that combines statistical parametric mapping with sparse dictionary learning for a data-driven brain fMRI analysis. We showed that a maximum likelihood estimation framework with sparsity constraint provides spatially adaptive design matrices as a subset of learned dictionary acquired from sparse dictionary learning algorithm.

Various ICA studies such as 'SPM-ICA' consider the component time course showing greatest correlation with the experimental task reference function as a consistently task-related (CTR) component and note that exactly one CTR component have a time course that is highly correlated with the reference function [26, 28]. Considering the autocorrelations between areas, 'HYBICA' uses a metric based on the predicted sum of squares statistic (PRESS) to select the best number of spatially independent components, and sequentially combines them into one hybrid task-related component to utilize in a GLM framework. Even though this approach reduces collinearity among the task-related regressors by combining all task-related regressors, a more correlated component can possibly contain a time-course that is related to a distinct neural signal. Unlike the ICA, the proposed method operates under a sparse distribution of task-related activation and spatially adaptive design matrix, whose sparsity level is determined by MDL criterion. Hence, it obtains a finer decomposition and better adapts to individual and spatial variation.

In summary, we have proposed a new data-driven analysis for a brain fMRI analysis using data-driven sparse GLM, which decomposes BOLD signals into sparse dictionary atoms. This method overcomes the limitations of ICA by exploiting sparsity of the components instead of independence. Furthermore, the unknown sparsity level can be estimated by MDL criterion. We show that this approach extracts individually adaptive activation patterns more accurately than spatial and temporal ICA, by a simulation and task-evoked experimental results.

REFERENCES

[1] K. Friston, P. Jezzard, and R. Turner, "Analysis of functional MRI time-series," *Human Brain Mapping*, vol. 1, no. 2, pp. 153–171, 1994.
[2] K. Friston, C. Frith, R. Turner, and R. Frackowiak, "Characterizing evoked hemodynamics with fMRI," *NeuroImage*, vol. 2, no. 2PA, pp. 157–165, 1995.
[3] K. Friston, A. Holmes, J. Poline, P. Grasby, S. Williams, R. Frackowiak, and R. Turner, "Analysis of fMRI time-series revisited," *NeuroImage*, vol. 2, no. 1, pp. 45–53, 1995.
[4] K. Friston, A. Holmes, K. Worsley, J. Poline, C. Frith, R. Frackowiak *et al.*, "Statistical parametric maps in functional imaging: a general linear approach," *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1995.
[5] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. E. Penny, *Statistical Parametric Mapping: The analysis of functional brain images.* Academic Press, 2006.
[6] R. Buxton, "The elusive initial dip." *NeuroImage*, vol. 13, no. 6 Pt 1, p. 953, 2001.

[7] E. Yacoub, A. Shmuel, J. Pfeuffer, P. Van De Moortele, G. Adriany, K. Ugurbil, and X. Hu, "Investigation of the initial dip in fMRI at 7 Tesla," *NMR in Biomedicine*, vol. 14, no. 7-8, pp. 408–412, 2001.

[8] D. Heeger and D. Ress, "What does fMRI tell us about neuronal activity?" *Nature Reviews Neuroscience*, vol. 3, no. 2, pp. 142–151, 2002.

[9] E. Yacoub, K. Ugurbil, and N. Harel, "The spatial dependence of the poststimulus undershoot as revealed by high-resolution BOLD-and CBV-weighted fMRI," *Journal of Cerebral Blood Flow & Metabolism*, vol. 26, no. 5, pp. 634–644, 2005.

[10] M. Schroeter, T. Kupka, T. Mildner, K. Uludag, and D. Von Cramon, "Investigating the post-stimulus undershoot of the BOLD signal–a simultaneous fMRI and fNIRS study," *NeuroImage*, vol. 30, no. 2, pp. 349–358, 2006.

[11] J. Frahm, J. Baudewig, K. Kallenberg, A. Kastrup, K. Merboldt, and P. Dechent, "The post-stimulation undershoot in BOLD fMRI of human brain is not caused by elevated cerebral blood volume," *NeuroImage*, vol. 40, no. 2, pp. 473–481, 2008.

[12] A. Kleinschmidt, H. Obrig, M. Requardt, K. Merboldt, U. Dirnagl, A. Villringer, and J. Frahm, "Simultaneous recording of cerebral blood oxygenation changes during human brain activation by magnetic resonance imaging and near-infrared spectroscopy," *Journal of Cerebral Blood Flow & Metabolism*, vol. 16, no. 5, pp. 817–826, 1996.

[13] R. Buxton, E. Wong, and L. Frank, "Dynamics of blood flow and oxygenation changes during brain activation: the balloon model," *Magnetic Resonance in Medicine*, vol. 39, no. 6, pp. 855–864, 1998.

[14] E. Yacoub and X. Hu, "Detection of the early decrease in fMRI signal in the motor area," *Magnetic Resonance in Medicine*, vol. 45, no. 2, pp. 184–190, 2001.

[15] G. Strangman, J. Culver, J. Thompson, and D. Boas, "A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation," *NeuroImage*, vol. 17, no. 2, pp. 719–731, 2002.

[16] A. Devor, A. Dunn, M. Andermann, I. Ulbert, D. Boas, and A. Dale, "Coupling of total hemoglobin concentration, oxygenation, and neural activity in rat somatosensory cortex," *Neuron*, vol. 39, no. 2, pp. 353–359, 2003.

[17] D. Boas, G. Strangman, J. Culver, R. Hoge, G. Jasdzewski, R. Poldrack, B. Rosen, and J. Mandeville, "Can the cerebral metabolic rate of oxygen be estimated with near-infrared spectroscopy?" *Physics in Medicine and Biology*, vol. 48, pp. 2405–2418, 2003.

[18] A. Andersen, D. Gash, and M. 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.

[19] R. Viviani, G. Gron, and M. Spitzer, "Functional principal component analysis of fMRI data," *Human Brain Mapping*, vol. 24, no. 2, pp. 109–129, 2005.

[20] M. McKeown, T. Sejnowski *et al.*, "Independent component analysis of fMRI data: examining the assumptions," *Human Brain Mapping*, vol. 6, no. 5-6, pp. 368–372, 1998.

[21] J. Stone, J. Porrill, N. Porter, and I. Wilkinson, "Spatiotemporal independent component analysis of event-related fMRI data using skewed probability density functions," *NeuroImage*, vol. 15, no. 2, pp. 407–421, 2002.

[22] B. Biswal and J. Ulmer, "Blind source separation of multiple signal sources of fMRI data sets using independent component analysis," *Journal of Computer Assisted Tomography*, vol. 23, no. 2, p. 265, 1999.

[23] C. Phillips, S. Zeki, and H. Barlow, "Localization of function in the cerebral cortex: past, present and future," *Brain*, vol. 107, no. 1, p. 328, 1984.

[24] M. Greicius, B. Krasnow, A. Reiss, and V. Menon, "Functional connectivity in the resting brain: a network analysis of the default mode hypothesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 1, p. 253, 2003.

[25] P. Fransson, "Spontaneous low-frequency BOLD signal fluctuations: an fMRI investigation of the resting-state default mode of brain function hypothesis," *Human Brain Mapping*, vol. 26, no. 1, pp. 15–29, 2005.

[26] M. McKeown, S. Makeig, G. Brown, T. Jung, S. Kindermann, A. Bell, and T. Sejnowski, "Analysis of fMRI data by blind separation into independent spatial components," *Human Brain Mapping*, vol. 6, no. 3, pp. 160–188, 1998.

[27] F. Esposito, E. Formisano, E. Seifritz, R. Goebel, R. Morrone, G. Tedeschi, and F. Di Salle, "Spatial independent component analysis of functional MRI time-series: To what extent do results depend on the algorithm used?" *Human Brain Mapping*, vol. 16, no. 3, pp. 146–157, 2002.

[28] D. Hu, L. Yan, Y. Liu, Z. Zhou, K. Friston, C. Tan, and D. Wu, "Unified SPM-ICA for fMRI analysis," *NeuroImage*, vol. 25, no. 3, pp. 746–755, 2005.

[29] M. McKeown, "Detection of consistently task-related activations in fMRI data with hybrid independent component analysis," *NeuroImage*, vol. 11, no. 1, pp. 24–35, 2000.

[30] V. Calhoun, T. Adali, G. Pearlson, and 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.

[31] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[32] A. Hyvarinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[33] S. Smith, M. Jenkinson, M. Woolrich, C. Beckmann, T. Behrens, H. Johansen-Berg, P. Bannister, M. De Luca, I. Drobnjak, D. Flitney *et al.*, "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, pp. S208–S219, 2004.

[34] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D'Ardenne, W. Richter, J. Cohen, and J. Haxby, "Independent component analysis for brain fMRI does not select for independence," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, p. 10415, 2009.

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

[36] R. Quiroga, G. Kreiman, C. Koch, and I. Fried, "Sparse but not "Grandmother-cell" coding in the medial temporal lobe," *Trends in Cognitive Sciences*, vol. 12, no. 3, pp. 87–91, 2008.

[37] R. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, "Invariant visual representation by single neurons in the human brain," *Nature*, vol. 435, no. 7045, pp. 1102–1107, 2005.

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

[39] M. Lewicki and T. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[40] K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T. Lee, and T. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, 2003.

[41] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[42] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, p. 489, 2006.

[43] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[44] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[45] O. Bryt and M. Elad, "Compression of facial images using the K-SVD algorithm," *Journal of Visual Communication and Image Representation*, vol. 19, no. 4, pp. 270–282, 2008.

[46] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[47] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, pp. 129–159, 2001.

[48] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via L1 minimization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, p. 2197, 2003.

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

[50] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[51] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 655–687, 2008.

[52] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[53] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[54] N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion," *Wavelets in Geophysics*, pp. 299–324, 1994.

[55] B. Ardekani, J. Kershaw, K. Kashikura, and I. Kanno, "Activation detection in functional MRI using subspace modeling and maximum likelihood estimation," *IEEE Transactions on Medical Imaging*, vol. 18, no. 2, p. 101, 1999.

[56] B. Biswal, F. Yetkin, V. Haughton, and J. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI," *Magnetic Resonance in Medicine*, vol. 34, no. 4, pp. 537–541, 1995.

[57] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," *Technical Report TR-CS-2008-08, Dept. of Computer Science, Technion*, 2008.

[58] V. Calhoun, T. Adali, G. Pearlson, and J. Pekar, "Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms," *Human Brain Mapping*, vol. 13, no. 1, pp. 43–53, 2001.

[59] M. McKeown, T. Jung, S. Makeig, G. Brown, S. Kindermann, T. Lee, and T. Sejnowski, "Spatially independent activity patterns in functional MRI data during the Stroop color-naming task," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 3, p. 803, 1998.