# Clustered Components Analysis for Functional MRI

Sea Chen, *Member, IEEE*, Charles A. Bouman*, *Fellow, IEEE*, and Mark J. Lowe

***Abstract*—A common method of increasing hemodynamic response (SNR) in functional magnetic resonance imaging (fMRI) is to average signal timecourses across voxels. This technique is potentially problematic because the hemodynamic response may vary across the brain. Such averaging may destroy significant features in the temporal evolution of the fMRI response that stem from either differences in vascular coupling to neural tissue or actual differences in the neural response between two averaged voxels. Two novel techniques are presented in this paper in order to aid in an improved SNR estimate of the hemodynamic response while preserving statistically significant voxel-wise differences. The first technique is signal subspace estimation for periodic stimulus paradigms that involves a simple thresholding method. This increases SNR via dimensionality reduction. The second technique that we call clustered components analysis is a novel amplitude-independent clustering method based upon an explicit statistical data model. It includes an unsupervised method for estimating the number of clusters. Our methods are applied to simulated data for verification and comparison to other techniques. A human experiment was also designed to stimulate different functional cortices. Our methods separated hemodynamic response signals into clusters that tended to be classified according to tissue characteristics.**

***Index Terms*—BOLD signal estimation, brain activation, clustering methods, EM algorithm, functional MRI, functional neuroimaging, hemodynamic response, independent components analysis, vascular coupling.**

## I. Introduction

FUNCTIONAL magnetic resonance imaging (fMRI) has emerged as a useful tool in the study of brain function. This imaging modality utilizes the fact that the MRI signal is sensitive to many of the hemodynamic parameters that change during neuronal activation (e.g., blood flow, blood volume, and oxygenation). The changes in these parameters cause small intensity differences between properly weighted MR images acquired before and during neuronal activation. Although the contrast can be produced by a number of different mechanisms, blood oxygenation level dependent (BOLD) contrast is the method most commonly employed. BOLD contrast is dependent on an decrease in local deoxy-hemoglobin concentration in an area of neuronal activity [1], [2]. This local decrease in paramagnetic material increases the apparent transverse relaxation constant $T_2^*$, resulting in an increase of MR signal intensity in the area affected. Other methods of functional MR imaging contrast include measurement of cerebral blood flow and volume effects [3]. Although fMRI is widely used, the mechanism of the coupling between brain hemodynamics and neuronal activation is poorly understood.

Although much of the work in fMRI data analysis has revolved around the creation of statistical maps and the detection of activation at different voxel locations [4]–[6], there also has been much interest in understanding the BOLD temporal response. Several groups have proposed models relating the various hemodynamic parameters (blood flow, blood volume, hemoglobin concentration, etc.) to the BOLD signal [7], [8]. These models all predict a BOLD temporal response to changing neuronal activity. Verification of the accuracy of these models requires that the predictions be compared to data. However, the low signal-to-noise ratio (SNR) of fMRI measurements typically requires averaging of many voxels in order to achieve a statistically significant result. Thus, the resulting measurement could possibly be a mixture of many different responses. This presents a possible confound in attempts to develop and validate detailed models of the BOLD response.

Some researchers have attempted to address the issue by using parametric methods [9], [10]. The parametric methods usually assume specific signal shapes (Poisson, Gaussian, Gamma, etc.) and attempt to extract the associated parameters for which the data best fit. Others have taken a linear systems approach in which the response is modeled as an impulse response convolved with the stimulus reference function [11], [12]. Exploratory data analysis methods such as principal components analysis (PCA) [13], [14] and independent components analysis (ICA) [15], [16] are also commonly used by many groups. Recently, clustering methods [17]–[20] have become popular as well.

In this paper, we address the issue of signal averaging by presenting a novel nonparametric clustering method based upon a statistical data model. Our technique is designed to robustly estimate the fMRI signal waveforms rather than detect their presence. Specifically, we first select voxels that contain the fMRI signal using a previously published technique [21]. Then our procedure identifies groups of voxels in fMRI data with the same temporal shape *independent of signal amplitude*. Variations in amplitudes may be due to differences in the concentration of hemodynamic events from partial volume effects or coil geometries. The amplitude variation is explicitly accounted for in our data model. Each distinct response corresponds to a unique *direction* in a multidimensional feature space (see Fig. 1).

Manuscript received April 9, 2003; revised July 10, 2003. This work was supported in part by the National Science Foundation (NSF) IGERT PTDD under Training Grant DGE-99-72770. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was Z. P. Liang. *Asterisk indicates corresponding author*.

S. Chen is with the Division of Imaging Sciences, Department of Radiology, Indiana University, School of Medicine, Indianapolis, IN, and the Department of Biomedical Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: sechen@iupui.edu).

*C. A. Bouman is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: bouman@ecn.purdue.edu).

M. J. Lowe is with the Division of Imaging Sciences, Department of Radiology, Indiana University, School of Medicine, Indianapolis, IN, and the Department of Biomedical Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: mjlowe@iupui.edu).

Digital Object Identifier 10.1109/TMI.2003.819922

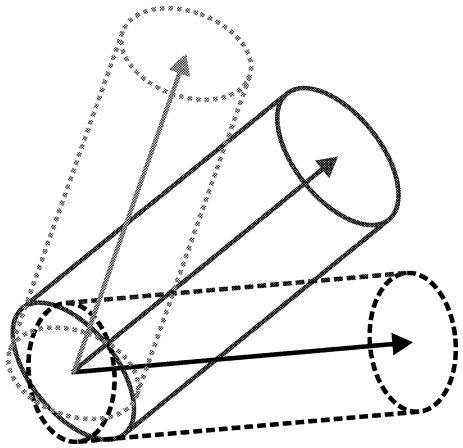

Fig. 1. Visualization of cylindrical clusters extracted by CCA: Because CCA finds cluster directions independent of amplitude, the shape of the vector clouds will be cylindrical instead of the more common spherical clouds around class means extracted by other clustering methods.

Our analysis framework is based upon two distinct steps. In the first step, the dimensionality of the voxel timecourses is reduced and feature vectors are obtained. The noise in the feature vectors is then whitened. The second step consists of our novel clustering technique that we call clustered components analysis (CCA).[1]

The dimensionality reduction used in this paper is similar to the method described by Bullmore, *et al.* [22]. Their method decomposes the temporal response at each voxel into harmonic components corresponding to a sine and cosine series expansion at the appropriate period. We have developed a method to further decrease dimensionality by estimating an $M$-dimensional signal subspace [23]. Although signal subspace estimation (SSE) is not new [24], [25], our method uses a simple thresholding technique and is quite effective. It is implemented by estimating the signal covariance as the positive definite part of the difference between the total signal-plus-noise covariance and the noise covariance. At this point in our analysis technique, each voxel's response is represented by an $M$ dimensional feature vector.

In the second step of our analysis framework, we present a new method for analyzing the multivariate fMRI feature vectors that we call CCA. This method depends on a explicit data model of the feature vectors and is implemented through maximum-likelihood (ML) estimation via the expectation-maximization (EM) algorithm [26], [27]. An agglomerative cluster merging technique based on a minimum description length (MDL) criterion [28] is used to estimate the true number of clusters [29].

Because the truth is not known in a real experiment, synthetic data was generated to test the performance of our method. Other common methods of multivariate data-driven analysis techniques (PCA, ICA, and fuzzy clustering) were applied to the same data set and the results were compared. Finally, a human experiment was performed that stimulated the motor, visual, and auditory cortices. Our methodology was applied to this data. The goal of the human experiment was to produce a set of activation data spanning a broad range of cerebral cortex and a diverse set of neuronal systems. This data set will allow our clustering method to determine the distribution of distinct temporal responses according either to neuronal system or tissue characteristics.

[1]CCA software is available from www.ece.purdue.edu/~bouman.

## II. Theory

### A. Dimensionality Reduction

The first step of our analysis framework is the reduction of dimensionality via decomposition into harmonic components using least squares fitting. This step is similar to that described in [22]. The next step is a novel method of determining the signal subspace by estimating signal and noise covariances and performing an eigendecomposition. The final step is a prewhitening of the noise before application of the CCA.

*1) Decomposition Into Harmonic Components:* In a standard block paradigm, control and active states are cycled in a periodic manner during the fMRI experiment. Therefore, the response signal should also be periodic. By assuming periodicity, harmonic components can be used as a basis for decomposition to reduce dimensionality. However, application of the periodicity constraint is not necessary for the technique described in the next section.

The data set of an fMRI experiment, $D$, can be defined as an $P \times N$ matrix, where $N$ is the number of voxels and $P$ is the number of time points. We first remove the baseline and linear drift components of fMRI data as a preprocessing step [21]. The columns of $D$ are then zero mean, zero drift versions of the voxel timecourses.

The harmonic components, $A_l$, are a sampling of sines and cosines at the fundamental frequency of the experimental paradigm, $\gamma$ (in radians/s), and its higher harmonics. The number of harmonic components, $L$, is limited by the requirement that there be no temporal aliasing. In other words, $L < ((2\pi)/(\Delta t \gamma))$ where $\Delta t$ is the temporal sampling period

$$A_l(t) = \begin{cases} \cos\left(\frac{l+1}{2}\gamma t\right), & \text{if } l \text{ odd} \\ \sin\left(\frac{l}{2}\gamma t\right), & \text{if } l \text{ even} \end{cases} \quad \text{for } l = [1, 2, \ldots, L]. \tag{1}$$

We then form a $P \times L$ design matrix

$$A = [a_1, \ldots, a_L] \tag{2}$$

where $a_l$ is a column vector formed by sampling the $l$th harmonic component at the times corresponding to the voxel samples. Using this notation, the data can then be expressed as a general linear model [6] where

$$D = A\Theta + \nu. \tag{3}$$

$\Theta$ is an $L \times N$ harmonic image matrix containing the linear coefficients, and $\nu$ is the $P \times N$ dimensional noise matrix.

Assuming all information in the signal is contained within the range $A$, an estimate $\hat{\Theta}$ can be computed using a least squares fit, resulting in

$$\hat{\Theta} = (A^t A)^{-1} A^t D \tag{4}$$

where the residual error $\epsilon$ is given by

$$\epsilon = D - A\hat{\Theta} \tag{5}$$

$$= (I - A(A^t A)^{-1} A^t) D \tag{6}$$

$$= (I - A(A^t A)^{-1} A^t) \nu. \tag{7}$$

The data set can be expressed in terms of the estimate of the coefficient matrix and the residuals matrix

$$D = A\hat{\Theta} + \epsilon. \tag{8}$$

We denote the estimation error as $\tilde{\Theta}$, where

$$\tilde{\Theta} = \hat{\Theta} - \Theta \tag{9}$$

$$= (A^t A)^{-1} A^t \nu. \tag{10}$$

*2) Signal Subspace Estimation:* Our next objective is to identify the subspace of the harmonic components that spans the space of all response signals. This signal subspace method improves the SNR by reducing the dimensionality of the data.

The covariance matrices for the signal, signal-plus-noise, and the noise are defined by the following relations:

$$R_s = \frac{1}{N} E[\Theta\Theta^t]$$

$$R_{sn} = \frac{1}{N} E[\hat{\Theta}\hat{\Theta}^t]$$

$$R_n = \frac{1}{N} E[\tilde{\Theta}\tilde{\Theta}^t].$$

Since we cannot observe $\Theta$ directly, we must first estimate $R_{sn}$ and $R_n$, and then use these matrices to estimate $R_s$. With this in mind, we use the following two estimates for $R_{sn}$ and $R_n$

$$\hat{R}_{sn} = \frac{1}{N} \hat{\Theta}\hat{\Theta}^t \tag{11}$$

$$\hat{R}_n = \text{trace}\{\epsilon\epsilon^t\}(A^t A)^{-1}/(N(P - L - 2)) \tag{12}$$

where $\epsilon$ is computed using (6). The expression for $\hat{R}_n$ is derived in Appendix I using the assumption that the noise $\nu$ is white and is shown to be an unbiased estimate for $R_n$. Note that the denominator of the expression reflects the additional reduction in degrees of freedom when the two drift components are removed. Since $\hat{R}_{sn}$ and $\hat{R}_n$ are both unbiased estimates of the true covariances, we may form an unbiased estimate of the signal covariance $R_s$ as

$$\hat{R}_s = \hat{R}_{sn} - \hat{R}_n. \tag{13}$$

The corresponding eigendecomposition is then

$$\hat{R}_s = U_s \Lambda_s U_s^t. \tag{14}$$

Generally, the matrix $\hat{R}_s$ will have both positive and negative eigenvalues because it is formed by the subtraction of (13). However, negative eigenvalue in $\hat{R}_s$ are nonphysical since we know that $R_s$ is a covariance matrix with strictly positive eigenvalues. Therefore, we know that the subspace corresponding to the negative eigenvalues is dominated by noise. We may exploit this fact by removing the energy in this noise subspace. To do this, we form a new $M \times M$ diagonal matrix $\hat{\Lambda}_s$, which contains only the $M$ positive diagonal elements in $\Lambda_s$, and we form a new $L \times M$ modified eigenvector matrix $\hat{U}_s$ consisting of the columns of $U_s$ corresponding to the positive eigenvalues in $\Lambda_s$. The reduced dimension signal component, or eigenimage, can then be written as

$$\hat{Y} = \hat{U}_s^t \hat{\Theta}. \tag{15}$$

The eigenimage $\hat{Y}$ contains the linear coefficients for the eigensequences $\Sigma = A\hat{U}_s$.

The CCA presented in the following section assumes that the noise is white. Therefore, we apply a whitening filter $W$ to form

$$Y = W\hat{Y} \tag{16}$$

as described in Appendix II. The column vectors of $Y$ correspond to $M$ dimensional feature vectors that describe the timecourse of each voxel. The timecourse realizations of the individual voxels may be reconstructed via the following relation:

$$\hat{D} = \Sigma W^{-1} Y \tag{17}$$

### *B. Clustered Component Analysis*

The method of CCA is developed in this section. The goal of the method is to cluster voxels into groups that represent similar shapes and to estimate the representative timecourses. Specifically, we apply the analysis to the feature vectors $Y$ found in (16). The analysis not only allows for the estimation of cluster timecourses, but also estimates the total number of clusters automatically. The algorithm consists of two steps. The first step is the estimation of the timecourses using the EM algorithm. Estimation of the number of clusters occurs in the second step using the MDL criterion. The diagram of Fig. 2 illustrates the basic flow of the CCA algorithm, and Fig. 3 give a detailed pseudocode specification of the CCA algorithm.

*1) Data Model:* Let $Y_n$ be the $n$th column of the matrix $Y$ found in (16). $Y_n$ is a vector of parameters specifying the timecourse or the $M$ dimensional feature vector for voxel $n$. Furthermore, let $E_K = [e_1, \ldots, e_K]$ be the $K$ zero mean, zero drift component directions (representing $K$ clusters) in the feature space, each with unit norm ($e_k^t e_k = 1, \forall k \in [1, 2, \ldots, K]$). The basic data model can be written as

$$Y_n = \alpha_n e_{x_n} + \omega_n \tag{18}$$

where $\alpha_n$ is the unknown amplitude for voxel $n, 1 \leq x_n \leq K$ is the class of the voxel, and $\omega_n$ is zero mean, Gaussian noise.

Because the data have been whitened, we assume that $E[\omega_n \omega_n^t] = I$. The probability density function of the voxel $n$ can be stated as

$$p_{y_n \mid x_n}(y_n \mid k, E_K, \alpha_n) = \frac{1}{(2\pi)^{M/2}} \times \exp\left\{-\frac{1}{2}(y_n - \alpha_n e_k)^t (y_n - \alpha_n e_k)\right\} \tag{19}$$

with log-likelihood function being

$$\log p_{y_n \mid x_n}(y_n \mid k, E_K, \alpha_n) = -\frac{M}{2}\log 2\pi - \frac{1}{2}\left(y_n^t y_n - 2\alpha_n y_n^t e_k + \alpha_n^2\right). \tag{20}$$

In order to resolve the dependence on the unknown amplitude, the ML estimate $\hat{\alpha}_n$ of the amplitude is found

$$\hat{\alpha}_n = \underset{\alpha_n}{\text{argmax}}\{\log p_{y_n \mid x_n}(y_n \mid k, E_K, \alpha_n)\} = y_n^t e_k. \tag{21}$$

The amplitude estimate in (21) is then substituted into the log-likelihood of (20)

$$\log p_{y_n \mid x_n}(y_n \mid k, E_K, \hat{\alpha}_n) = -\frac{M}{2}\log 2\pi - \frac{1}{2}\left(y_n^t y_n - e_k^t y_n y_n^t e_k\right). \tag{22}$$

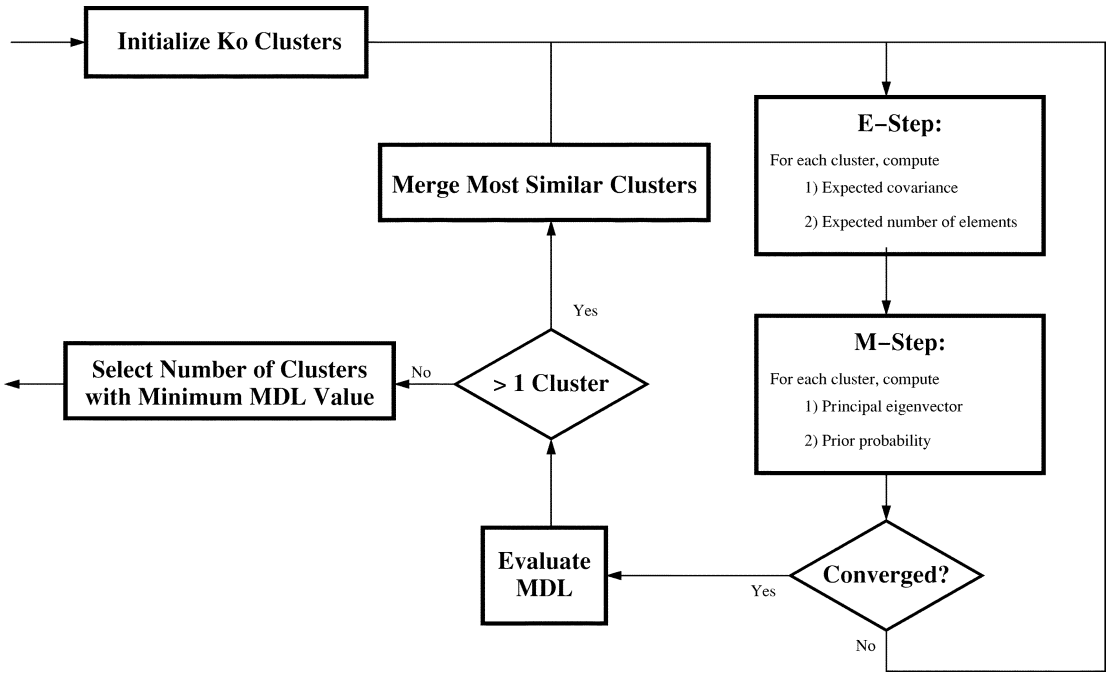


Fig. 2. This schematic diagram shows the basic operations of the CCA algorithm. The EM algorithm is performed by iteratively applying the E-step and M-step until convergence is achieved. After convergence, the MDL criterion is evaluated, and if there is more then one cluster remaining, then the two most similar clusters are merged, and the process is repeated. The final result is chosen to contain the number of clusters corresponding to the minimum value of the MDL criterion.

initialize $K$ to $K_0$
initialize ${E_{K_0}}^{(1)}$ and ${\Pi_{K_0}}^{(1)}$
while $K > 1$
  $i \leftarrow 1$
  do
    compute the posterior probabilities $p_{x_n|y_n}(k|y_n, K, E_K, \Pi_K, \hat{\alpha_n})$ for all $n$ using (28)
    E-step: compute $\bar{N}_{k|K}^{(i)}$ and $\bar{R}_{k|K}^{(i)}$ using (29) and (30)
    M-step: compute $E_K^{(i+1)}$ and $\Pi_K^{(i+1)}$ using (31) and (32)
    $\delta \leftarrow \log p_y(y|K, {E_K}^{(i+1)}, {\Pi_K}^{(i+1)}, \hat{\alpha}) - \log p_y(y|K, {E_K}^{(i)}, {\Pi_K}^{(i)}, \hat{\alpha})$
    $i \leftarrow i + 1$
  while $\delta > \upsilon$ (where $\upsilon$ is the stopping tolerance)
  set $\hat{E}_K = {E_K}^{(i)}$ and $\hat{\Pi}_K = {\Pi_K}^{(i)}$
  compute $MDL_K = MDL(K, \hat{E}_K, \hat{\Pi}_K)$ using (33)
  save $\hat{E}_K$ and $\hat{\Pi}_K$ and $MDL_K$
  find the two clusters $\hat{l}$ and $\hat{m}$ which minimize the distance function $d(l, m)$
  using (34) and (35)
  merge clusters $l$ and $m$ to form ${E_{K-1}}^{(1)}$ and ${\Pi_{K-1}}^{(1)}$ using (36) and (37)
  $K \leftarrow K - 1$
end
choose $\hat{K}$ and the corresponding $\hat{E}_{\hat{K}}$ and $\hat{\Pi}_{\hat{K}}$ which minimize the MDL

Fig. 3. Summary of CCA a algorithm.

From (22), the density function may be written as

$$p_{y_n \mid x_n}(y_n \mid k, E_K, \hat{\alpha}_n) = \frac{1}{(2\pi)^{M/2}} \times \exp\left\{-\frac{1}{2}\left(y_n^t y_n - e_k^t y_n y_n^t e_k\right)\right\}. \quad (23)$$

The class of voxel $n$ is specified by the class label $x_n$, which is an independent, identically distributed discrete random variable taking on integer values from 1 to $K$. We define $\pi_k = P\{X_n = k\}$ as the prior probabilities that a voxel is of class $k$. The set of prior probabilities for $K$ classes are then defined to be $\Pi_K = [\pi_1, \ldots, \pi_K]$, where $\sum_{k=1}^{K} \pi_k = 1$.

Using Bayes rule, the voxel probability density can be written without conditioning on class label

$$p_{y_n}(y_n \mid K, E_K, \Pi_K, \hat{\alpha}_n) = \sum_{k=1}^{K} p_{y_n \mid x_n}(y_n \mid k, E_K, \hat{\alpha}_n)\pi_k \quad (24)$$

$$= \sum_{k=1}^{K} \left(\frac{1}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2}\left(y_n^t y_n - e_k^t y_n y_n^t e_k\right)\right\}\right)\pi_k. \quad (25)$$

We note that this is a Gaussian mixture distribution [30].

The log-likelihood is then calculated for the whole set of voxels

$$\begin{aligned}\log p_y(y \,|\, K, E_K, \Pi_K, \hat{\alpha}) &= \sum_{n=1}^{N} \log \left( \sum_{k=1}^{K} p_{y_n \,|\, x_n}(y_n \,|\, k, E_K, \hat{\alpha}_n)\pi_k \right) \\ &= \sum_{n=1}^{N} \log \left[ \sum_{k=1}^{K} \left( \frac{1}{(2\pi)^{(M-1)/2}} \right. \right. \\ &\quad \left. \left. \times \exp\left\{ -\frac{1}{2}\left( y_n^t y_n - e_k^t y_n y_n^t e_k \right) \right\} \pi_k \right) \right]. \quad (26)\end{aligned}$$

*2) Parameter Estimation Using the Expectation-Maximization Algorithm:* The aim of this section is to estimate the parameters $E_K$ and $\Pi_K$ in the data model. This is done by finding the ML estimates for the log-likelihood given in (26) for a given cluster number $K$

$$(\hat{E}_K, \hat{\Pi}_K) = \arg\max_{E_K, \Pi_K} \log p_y(y \,|\, K, E_K, \Pi_K, \hat{\alpha}). \quad (27)$$

The ML estimates $\hat{E}_K$ and $\hat{\Pi}_K$ in (27) are found by using the EM algorithm [26], [30].

In order to compute the expectation step of the EM algorithm, we must first compute the posterior probability that each voxel label $x_n$ is of class $k$

$$p_{x_n \,|\, y_n}(k \,|\, y_n, K, E_K, \Pi_K, \hat{\alpha}_n) = \frac{p_{y_n \,|\, x_n}(y_n \,|\, k, E_K, \hat{\alpha}_n)\pi_k}{\sum_{l=1}^{K} p_{y_n \,|\, x_n}(y_n \,|\, k, E_K, \hat{\alpha}_n)\pi_l}. \quad (28)$$

In the expectation step of the EM algorithm, the estimated number of voxels per class $\bar{N}_{k\,|\,K}^{(i)}$ and the estimated unnormalized covariance matrix of the class $\bar{R}_{k\,|\,K}^{(i)}$ given the current estimation of the parameters $E_K^{(i)}$ and $\Pi_K^{(i)}$ must be computed. See Appendix III-A for more details. Because the EM algorithm is iterative, the superscripts $(i)$ denote iteration number. The subscript $k\,|\,K$ denotes the parameter corresponding to the $k$th cluster out of a total of $K$ clusters

$$\bar{N}_{k\,|\,K}^{(i)} = \sum_{n=1}^{N} p_{x_n \,|\, y_n}\left( k \,|\, y_n, E_K^{(i)}, \Pi_K^{(i)}, \hat{\alpha}_n \right) \quad (29)$$

$$\bar{R}_{k\,|\,K}^{(i)} = \sum_{n=1}^{N} y_n y_n^t p_{x_n \,|\, y_n}\left( k \,|\, y_n, E_K^{(i)}, \Pi_K^{(i)}, \hat{\alpha}_n \right). \quad (30)$$

In the maximization step of the EM algorithm, the parameters are reestimated from the values found in the expectation step [see (29) and (30)], yielding $E_K^{(i+1)}$ and $\Pi_K^{(i+1)}$. Let $e_{\max}\{R\}$ be the principal eigenvector of $R$

$$e_k^{(i+1)} = e_{\max}\left\{ \bar{R}_{k\,|\,K}^{(i)} \right\} \quad (31)$$

$$\pi_k^{(i+1)} = \bar{N}_{k\,|\,K}^{(i)} \Big/ N \quad (32)$$

for $k \in [1, 2, \ldots, K]$ (see Appendix III-B for more details).

Equations (31) and (32) are alternately iterated with (29) and (30) [using (28)]. The iterations are stopped when the difference in the log-likelihood [see (26)] for subsequent iterations is less than an arbitrary stopping criterion, $v$. We then denote the final estimates of the parameters for a given number of clusters $K$ as $\hat{E}_K$ and $\hat{\Pi}_K$.

*3) Model Order Identification:* Our objective is not only to estimate the component vectors $\hat{E}_K$ and the prior probabilities $\hat{\Pi}_K$ from observations, but also to estimate the number of classes $\hat{K}$. We use the MDL criterion developed by Rissanen [28], which incorporates a penalty term $KM\log(NM)/2$. The term $NM$ represents the number of scalar values required to represent the data, and the term $KM$ represents the number of scalar parameters encoded by $\hat{E}_K$ and $\hat{\Pi}_K$

$$\mathrm{MDL}(K, E_K, \Pi_K) = -\log p_y(y \,|\, K, E_K, \Pi_K, \hat{\alpha}) + \frac{1}{2} KM \log(NM). \quad (33)$$

The MDL criterion is then minimized with respect to $K$. This is done by starting with $K$ large, and then sequentially merging clusters until $K = 1$. More specifically, for each value of $K$, the values of $\hat{E}_K, \hat{\Pi}_K$, and $\mathrm{MDL}(K, \hat{E}_K, \hat{\Pi}_K)$ are calculated using the EM algorithm from Section II-B2. Next, the two most similar clusters are merged, $K$ is decremented to $K-1$, and the process is repeated until $K = 1$. Finally, we select the value of $\hat{K}$ (and corresponding parameters $\hat{E}_{\hat{K}}$ and $\hat{\Pi}_{\hat{K}}$) that resulted in the smallest value of the MDL criterion.

This merging approach requires that we define a method for selecting similar clusters. For this purpose, we define the following distance function between the clusters $l$ and $m$:

$$d(l, m) = \sigma_{\max}(\bar{R}_{l\,|\,K}) + \sigma_{\max}(\bar{R}_{m\,|\,K}) - \sigma_{\max}(\bar{R}_{l\,|\,K} + \bar{R}_{m\,|\,K}) \quad (34)$$

where $\sigma_{\max}(R)$ denotes the principal eigenvalue of $R$. In Appendix IV, we show that this distance function is an upper bound on the change in the MDL value [see (66)]. Therefore, by choosing the two clusters $\hat{l}$ and $\hat{m}$ that minimize the cluster distance

$$(\hat{l}, \hat{m}) = \arg\min_{l,m} d(l, m) \quad (35)$$

we minimize an upper bound on the resulting MDL criterion. The parameters of the new cluster formed by merging $\hat{l}$ and $\hat{m}$ are given by

$$\pi_{(\hat{l},\hat{m})} = \pi_{\hat{l}} + \pi_{\hat{m}} \quad (36)$$

$$e_{(\hat{l},\hat{m})} = e_{\max}\{\bar{R}_{\hat{l}\,|\,K} + \bar{R}_{\hat{m}\,|\,K}\}. \quad (37)$$

The remaining cluster parameters stay constant, and the merged parameters then become the initial parameters for the iterations of the EM algorithm used to find estimates for $K-1$ clusters.

To start the algorithm, a large number of clusters $K_0$ is chosen. For each class

$$\{k : k = 1, 2, \cdots, K_0\} \quad (38)$$

the parameters $\Pi_{K_0}^{(1)}$ and $E_{K_0}^{(1)}$ must be initialized. The parameters $\Pi_{K_0}^{(1)}$ are initialized uniformly as

$$\pi_k^{(1)} = \frac{1}{K_0}. \quad (39)$$

The first $M$ component directions, $e_k^{(1)}$, are initialized to the $M$ principal eigenvectors of the estimated data covariance matrix given by

$$\bar{R}_{k|K}^{(1)} = \frac{1}{N} \sum_{n=1}^{N} y_n y_n^t. \quad (40)$$

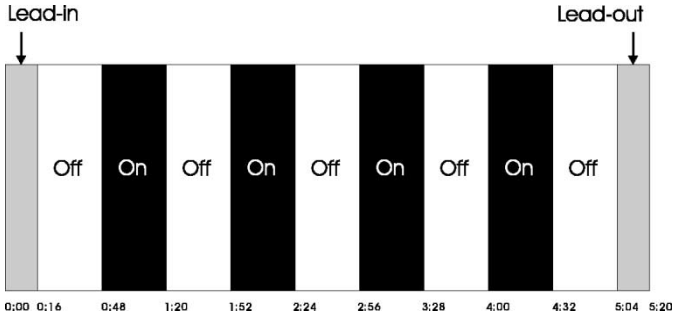


Fig. 4. Experimental paradigm timing.

The remaining $K - M$ components are initialized by randomly sampling the normalized data vectors $y_n/||y_n||$.

## III. Methods

### A. Experimental Paradigm

An experimental paradigm was designed to activate the auditory, visual, and motor cortices. The paradigm was arranged so all activation occurred in sync at a cycle length of 64 s (32 s control, 32 s active). The timing of the paradigm was as follows: 16 s lead in (control), four cycles of the paradigm ($4 \times 64$ s), 32 s control, and 16 s lead out (control). See Fig. 4 for a diagram of paradigm timing. The visual cortex was activated using a flashing 8-Hz checkerboard pattern ($6 \times 8$ squares) with a blank screen control state viewed through fiber-optic goggles (Avotec, Inc., Stuart, FL). The flashing checkerboard has been shown to provide robust activation throughout the visual system [31]. The auditory cortex was activated using backward speech through pneumatic headphones (Avotec). The backward speech has been shown to provide robust activation in the primary auditory cortex [32]. Auditory control was silence through the headphones (note that the ambient scanner noise is heard by the subject throughout the scan). The visual and auditory stimuli were constructed using commercial software (Adobe AfterEffects, Adobe Systems, Inc., San Jose, CA). The motor cortex was activated through a complex finger-tapping task. Left and right fingers were placed opposed in a mirror-like fashion in the rest position and tapped together in a self paced way in the following pattern for activation: thumb, middle, little, index, ring, repeat. This complex finger-tapping task has been shown to provide robust motor cortex activation [33]. Tapping was cued by the onset of visual and auditory stimuli. Rest was the control state for the motor paradigm.

### B. Human Data Acquisition

Whole-brain images of a healthy subject were obtained using a 1.5 T GE Echospeed MRI Scanner (GE Medical Systems, Waukeshau, WI). Axial 2D spin echo $T_1$-weighted anatomic images were acquired for reference with the following parameters: $\mathrm{TE} = 10$ ms $\mathrm{TR} = 500$ ms, $\mathrm{matrix\ dimensions} = 256 \times 128$, 15 locations with thickness of 7.0 mm and gap of 2.0 mm covering the whole brain, field-of-view $= 24 \times 24$ cm.

BOLD-weighted 2D gradient echo echoplanar imaging (EPI) functional images were acquired during a run of the experimental paradigm with the following parameters: $\mathrm{TE} = 50$ ms, $\mathrm{TR} = 2000$ ms, $\mathrm{flip\ angle} = 90°, \mathrm{matrix\ dimensions} = 64 \times 64$, 160 repetitions, and the same locations and field-of-view as the anatomic images.

TABLE I
Parameters Used in Synthetic Data Generation

| Signal | $f_a$ | $f_b$ | $f_c$ | $d_a$ | $d_b$ | $d_0$ |
|---|---|---|---|---|---|---|
| 1 | 0.6 | 0.02 | 0.2 | 1 | 10 | 2 |
| 2 | 0.35 | 0.2 | 0.5 | 3 | 5 | 8 |
| 3 | 0.35 | 0.1 | 1 | 5 | 5 | 15 |

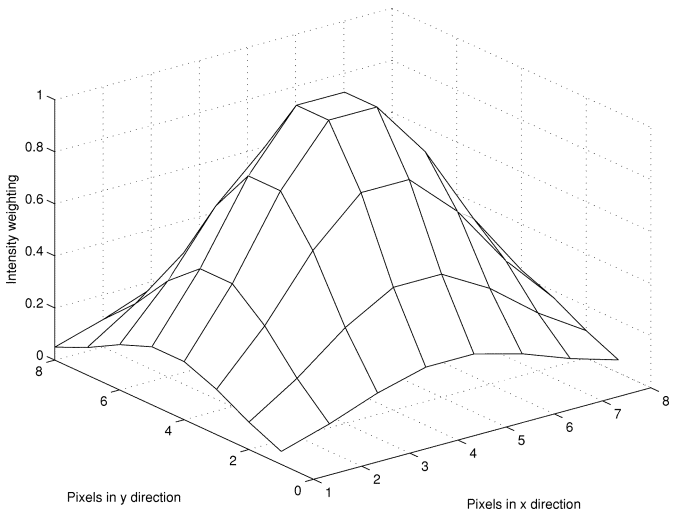


Fig. 5. Gaussian window for variation in amplitude in the synthetic data. Each vertex in the mesh corresponds to a voxel in an $8 \times 8$ square ROI. The mesh values at the vertices modulate the amplitudes of activation at the corresponding voxels.

### C. Synthetic Data Generation

To test the validity of the our methods, synthetic fMRI images were generated using the averaged functional images gathered from the real data set acquired as per Section III-B as baseline images. The BOLD response signals were modeled using the methods given by Purdon *et al.* [34], in the three subsequent equations.

The physiologic model is based upon two gamma functions given by

$$g_a(t) = (1 - e^{-1/d_a})^2 (t+1) e^{-t/d_a} \tag{41}$$

$$g_b(t) = (1 - e^{-1/d_b}) e^{-t/d_b} \tag{42}$$

where $d_a$ and $d_b$ are time constants. The activation signal $s(t)$ is then a combination of these gamma functions convolved (denoted by $*$) with the stimulus reference signal $c(t)$ which equals 0 during the control states and 1 during the active states. $d_0$ denotes a time delay and $f_a$, $f_b$, and $f_c$ are amplitudes which characterize the activation

$$s(t) = f_a(g_a * c)(t - d_0) + f_b(g_b * c)(t - d_0) + f_c(g_a * c)(t - d_0)(g_b * c)(t - d_0). \tag{43}$$

The mixture weights, as well as time constant and time delay parameters, were varied between three locations of $8 \times 8$ in one slice in order to simulate responses from different functional cortices and/or tissue characteristics. The parameters for (41), (42), and (43) are given in Table I for each of the signals/locations.

The amplitudes of these signals were modulated by the baseline voxel intensities $\mu_n$ using 7% peak activation and then multiplied by a normalized Gaussian window $(G)$ (see Fig. 5) to

TABLE II
MEAN SQUARED ERROR FOR ANALYSES ON SYNTHETIC DATA BEFORE AND AFTER SSE

| | PCA | FCM | ICA (c) | ICA (u) | CCA |
|---|---|---|---|---|---|
| Before SSE | $5.73\times10^{-4}$ | $7.37\times10^{-4}$ | $1.02\times10^{-3}$ | $2.49\times10^{-4}$ | $3.49\times10^{-5}$ |
| After SSE | $5.80\times10^{-4}$ | $1.95\times10^{-4}$ | $2.21\times10^{-4}$ | $1.63\times10^{-4}$ | $3.09\times10^{-5}$ |

simulate the variation in amplitudes across the functional region. Additive white Gaussian noise ($\nu$) was then added to all the voxels at a standard deviation of 2% of the mean baseline voxel intensity in the entire brain

$$y_n(t) = \mu_n + 0.07\mu_n G_n s_n(t) + \nu_n(t). \tag{44}$$

### D. Data Processing

*1) Synthetic Data:* In the synthetic data, only the voxels that had been injected with signal were considered for analysis. Mean and drift were removed from the voxel timecourses. Only the timepoints corresponding to the first four cycles were considered (128 out of 160 time points). Because the sequence used was a two-dimensional acquisition, each slice is acquired at staggered timing. Therefore, a different design matrix $A$ was specified for each slice by shifting the harmonic components by the corresponding time delay for the slice. The dimensionality reduction scheme including harmonic decomposition and signal subspace estimation outlined in Section II-A was performed.

PCA was applied to the synthetic data after harmonic decomposition using the harmonic images $\hat{\Theta}$ given in (4). An eigendecomposition was performed on $\hat{R}_{sn}$ given in (11) to obtain the principal components. PCA was also applied to the synthetic data after signal subspace estimation. The principal components are the columns of $\hat{U}_s$ derived from (14). In both cases, the three principal components were chosen corresponding to the three largest variances.

Fuzzy C-means clustering (FCM), using the Matlab fuzzy toolbox (Mathworks, Natick, MA), was also applied to the synthetic data on $\hat{\Theta}$ and unwhitened feature vectors $\tilde{Y}$. The routine was constrained to yield three clusters in both cases.

Spatial ICA, using software available from [35], was applied to $\hat{\Theta}$ and to the $\tilde{Y}$. The analysis was applied unconstrained to yield as many components as channels and also constrained to yield three independent components. For the unconstrained case, the three independent components that best matched (in a least squares sense) the injected signals were chosen.

CCA was applied only after signal subspace estimation on the whitened feature vectors $Y$. The CCA was initialized with $K_0 = 20$ clusters. The resultant estimates were transformed back into the time domain

$$\tilde{e}_k = \Sigma W^{-1}\hat{e}_k, \forall k \in [1, \ldots, K] \tag{45}$$

in a manner similar to (17).

*2) Human Data:* The functional image data were analyzed voxel-by-voxel for evidence of activation using a conventional Student's T-test analysis [21] using in-house software. Regions of interest (ROIs) were drawn on the resulting statistical maps in the cortical regions corresponding to primary activated regions for each of the three stimuli (i.e., precentral gyrus for the motor stimuli, superior temporal gyrus for the auditory stimuli, and the calcarine fissure for the visual stimuli). Only the voxels in the ROIs were considered in the analysis. The data were processed for the dimensionality reduction in the same manner as the synthetic data. CCA was applied to the resultant feature vectors $Y$, and the results were transformed back into the time domain using (45). CCA for the human data was also initialized with $K_0 = 20$ clusters.

## IV. RESULTS

### A. Synthetic Data

The harmonic decomposition and signal subspace estimation yielded $M = 16$ dimensions for the synthetic data. The resultant signals from PCA, FCM, ICA, and CCA applied before and after signal subspace estimation were then least squares fitted to each injected synthetic signal (peak-to-trough amplitude normalized) and were matched by finding the combination with the smallest mean square error. The mean square error results of the analysis methods are shown in Table II. The timesequence realizations are plotted for the analysis methods applied after signal subspace estimation in Fig. 6.

Hard classification results were then calculated by finding the largest membership value (e.g., largest component for PCA and ICA, largest membership value for FCM, and largest posterior probability for CCA) for each voxel. The number of correct voxel classifications for the analyses is shown in Table III. The hard classification results image for CCA is shown in Fig. 7.

### B. Human Data

The signal subspace was found to have $M = 8$ dimensions. CCA returned $\hat{K} = 9$ classes and the components $\hat{E}_9$ from the feature space. The number of voxels in each cluster is displayed in Fig. 8. The majority of the voxels (94%) lie in the first five clusters. The last four clusters contain 10 or less voxels. Therefore, the timesequence realizations of the component directions from the first five clusters are given separately from the last four clusters in Fig. 9. Each voxel was then assigned to the class with the highest *a posteriori* probability. The hard classifications for the first five clusters are shown in Fig. 10 with the accompanying anatomic data. The colors used for the voxels correspond to the class colors in Fig. 9(a).

## V. DISCUSSION

### A. Experimental Results

Our simulation results show that, as a general rule, mean square error decreased by using signal subspace estimation (except for PCA). The number of voxels classified correctly increased for each analysis method as a result of signal subspace estimation. It can also be seen that CCA outperforms each of

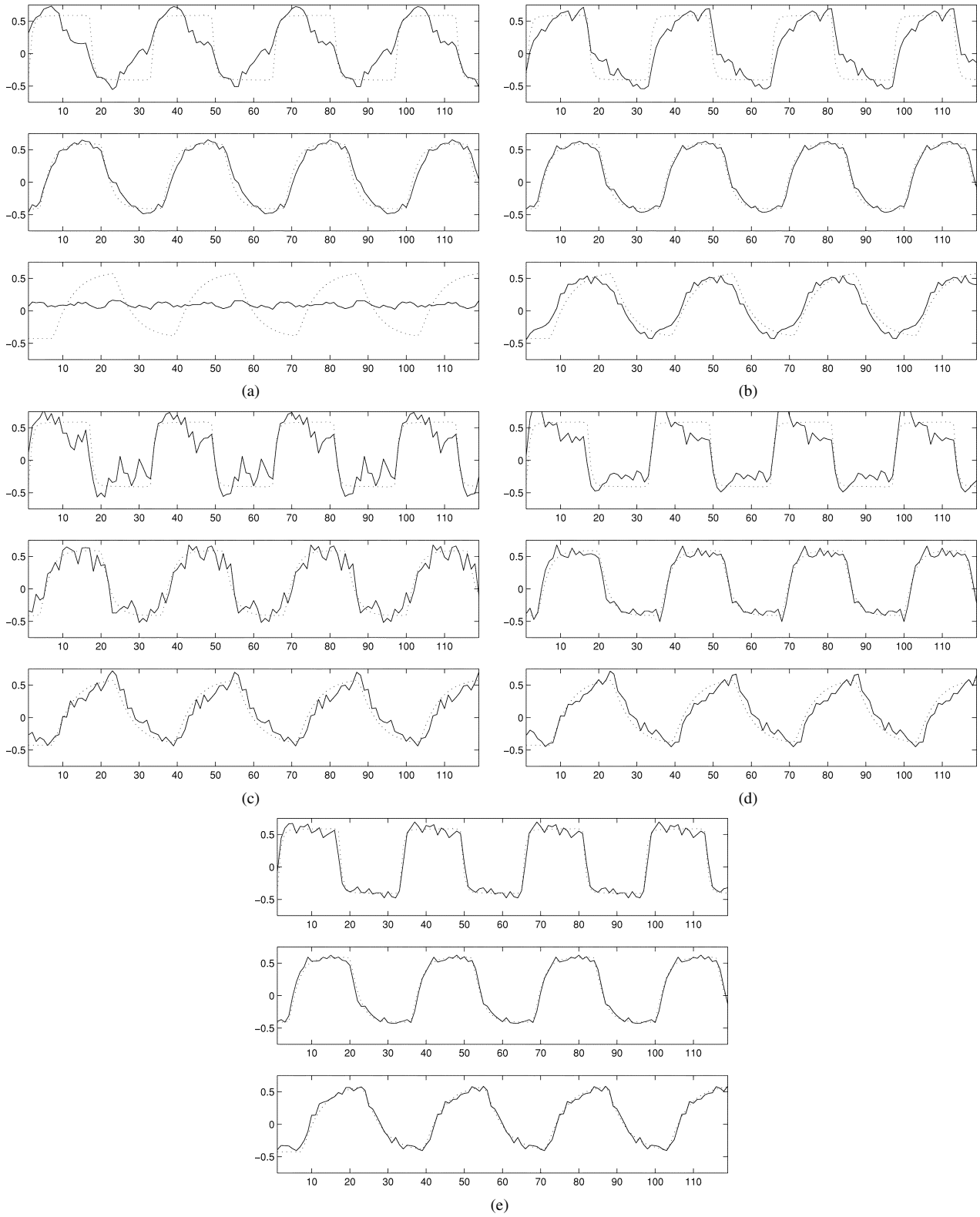


Fig. 6. Estimation methods after signal subspace estimation plotted against injected synthetic signal. (a) PCA, (b) FCM, (c) constrained ICA, (d) unconstrained ICA, and (e) CCA. The solid estimated signals are least squares scaled to the dotted injected signals.

the other analysis methods in both mean square error and correct voxel classification. Although in CCA the changes are small before and after signal subspace estimation in both mean square error and correct classification, the signal subspace estimation speeds the algorithm greatly.

The experimental results from the human data reveal that a distinct functional behavior does not correlate directly with each of the functional ROIs, at least for the motor and auditory cortices. Inspection of Fig. 10 shows that the classes are distributed along patterns of location with respect to sulcal-gyral bound-

TABLE III
NUMBER OF VOXELS CLASSIFIED CORRECTLY ON SYNTHETIC DATA BEFORE AND AFTER SSE OUT OF 192 TOTAL VOXELS

| | PCA | ICA (c) | ICA (u) | FCM | CCA |
|---|---|---|---|---|---|
| Before SSE | 61 | 113 | 38 | 95 | 167 |
| After SSE | 111 | 162 | 77 | 151 | 169 |

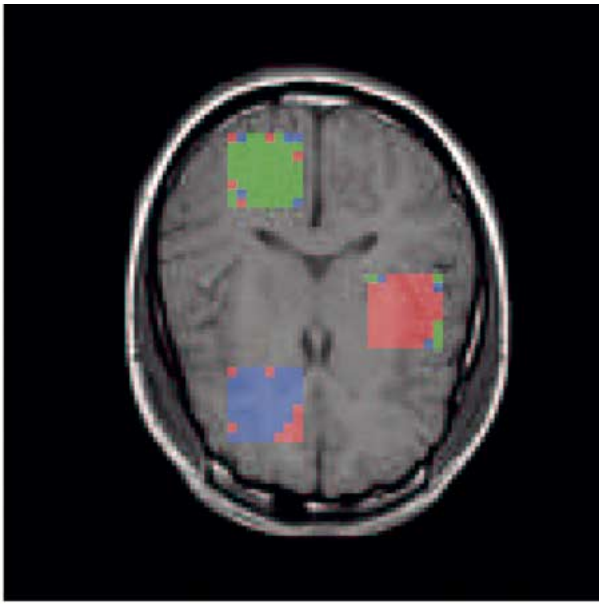

Fig. 7. Hard classification results from CCA on synthetic data.

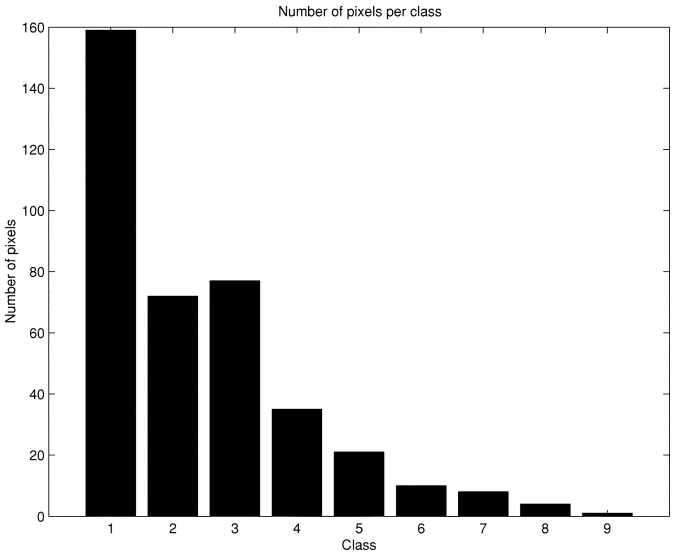


Fig. 8. Histogram of the number of voxels in each class for the human data set.

aries. This may reflect that the temporal evolution of the BOLD signal is dependent much more on the vascularization of the tissue than the functional specifics of neuronal activation.

### *B. Relation to Other Analysis Methods*

In this paper, we have attempted to develop an analysis framework that incorporates the advantages of previous methods. This section will detail how our methods relate to other multivariate methods used by other researchers. We note that over the past few years some commonly used algorithms for analyzing fMRI data have been widely distributed as software packages [36], [37].

Linear time invariant systems methods try to model the hemodynamic impulse response function by using deconvolution. Although it has been shown that for some cases linearity is reasonable [11], [12], for other situations nonlinearities arise [38]. Parametric methods are heavily model driven and not as suitable if data are not described well by the model. Therefore, data-driven methods should be used in fMRI where the mechanisms of the system are unknown. Our approach is data-driven and may be more appropriate in this case.

PCA is a method which uses the eigenanalysis of signal correlations to produce orthogonal components in the directions of maximal variance [39], [40]. However, it is unlikely that the distinct behaviors in fMRI data correspond to orthogonal signals. Backfrieder *et al.* [14], attempted to solve this problem by using an oblique rotation of the components. Most researchers, however, use PCA as a preprocessing step for dimensionality reduction. A threshold is usually arbitrarily set to the number of components kept [13]. We use a method similar to PCA for dimensionality reduction. Our threshold, however, is determined from the data itself after noise covariance estimation from harmonic decomposition and should effectively remove this subjective aspect the analysis.

ICA is used in signal processing contexts to separate out mixtures of independent sources or invert the effects of an unknown transformation [41]. It was adapted to produce spatial independent components for fMRI datasets by McKeown *et al.* [15], [16]. The fMRI data is modeled as a linear combination of maximally independent (minimally overlapping) spatial maps with component timecourses. It has been pointed out, however, that neuronal dynamics may overlap spatially [42]. Our method does not constrain distinct behaviors to be spatially independent. Another shortcoming of ICA is that it does not lend itself to statistical analysis. McKeown and Sejnowski have attempted to solve this problem by developing a method to calculate the posterior probability of observing a voxel timecourse given the ICA unmixing matrix [43]. Because CCA is based upon an explicit statistical model, it does not suffer from this disadvantage.

Clustering algorithms have been applied to both fMRI raw timecourse data [18], [20], [44] and to timecourse features such as univariate statistics and correlations [45]. Because we are trying to estimate the response signal, we use a hybrid method to characterize the timecourses into lower dimensionality representations. The main problem with most of these clustering methods is that the variation in amplitude is not taken into consideration. Our method produces component *directions* due to the amplitude variances (see Fig. 1) rather than traditional cluster means. More recently, Brankov *et al.* [46], proposed an alternative clustering approach which is also designed to account for variations in signal amplitude. Another shortcoming of most clustering methods is that the number of clusters is arbitrarily determined. Baune *et al.* [20], attempt to solve this problem by setting a threshold on Euclidean distances for the merging of clusters. Liang *et al.* [30], also used an information criterion for order identification (Akaike information criterion and MDL) to analyze PET and SPECT data to find image

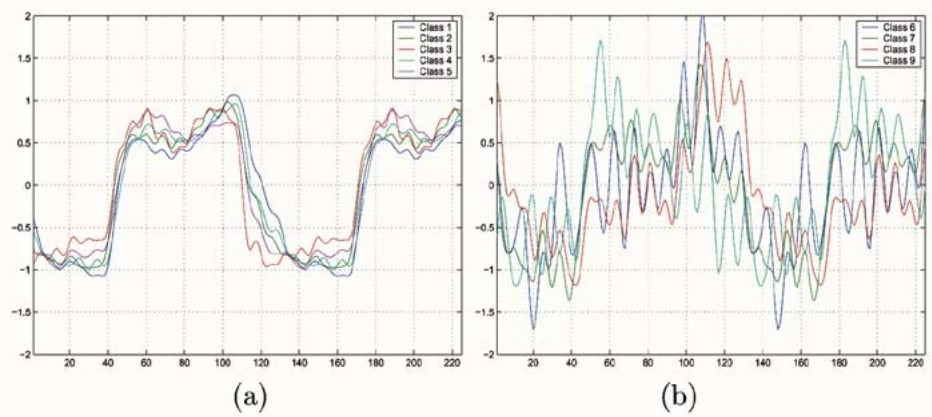


Fig. 9. Timesequence realizations of the feature space for the human data set. (a) Timesequences for the first five clusters, (b) timesequences for clusters 6–9.

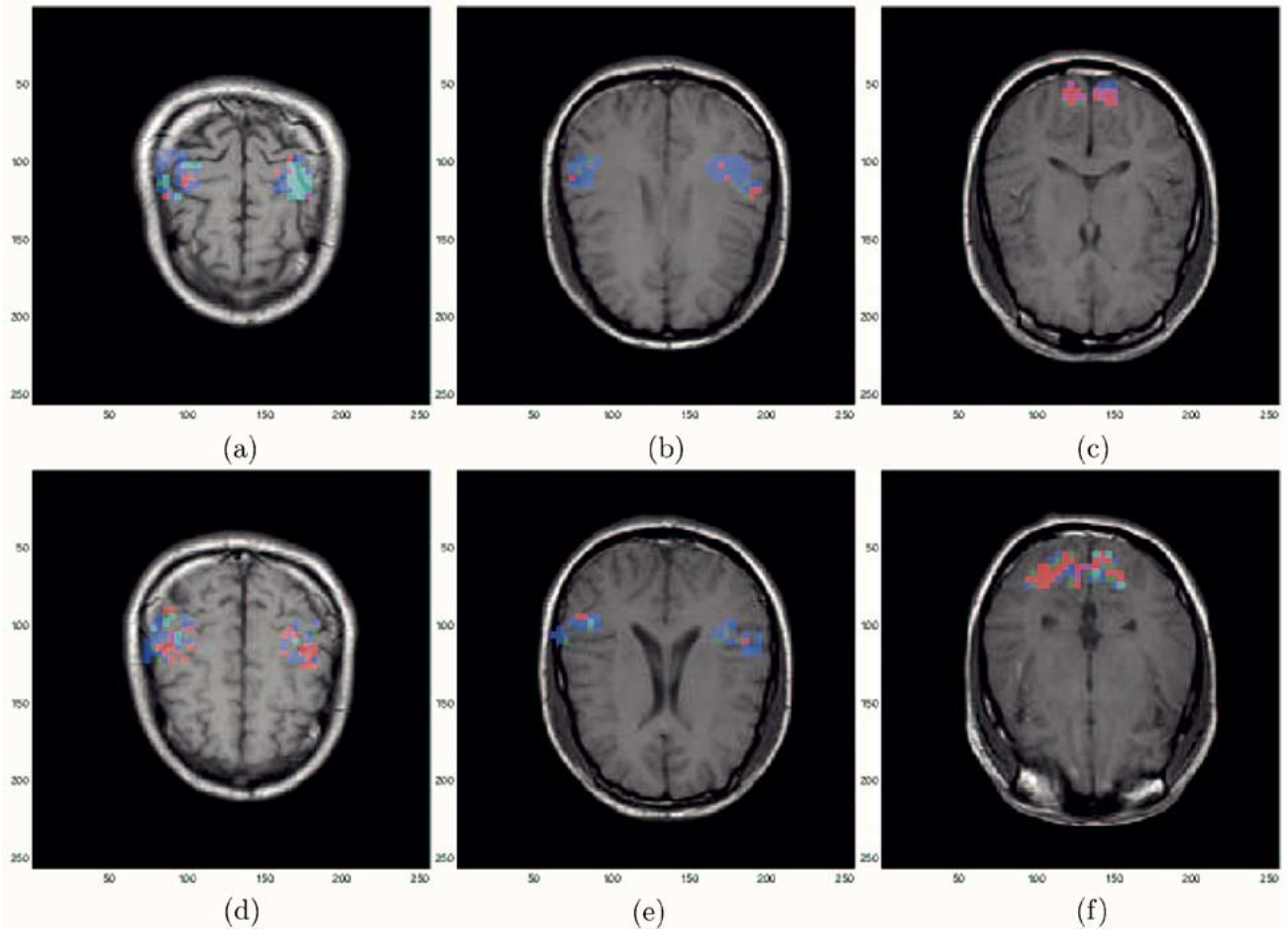


Fig. 10. CCA hard classification on the real data set (first five clusters). The colors correspond to the class colors shown in Fig. 9(a). (a) upper motor cortex slice, (b) upper auditory cortex slice, (c) upper visual cortex slice, (d) lower motor cortex slice, (e) lower auditory cortex slice, and (f) lower visual cortex slice.

parameters. Our method uses the MDL criterion plus a cluster merging strategy to determine the number of clusters in an unsupervised manner.

### C. Algorithm Details

We use a design matrix $A$ which consists of harmonic components used in the dimensionality reduction applied in this paper. Periodicity of the response signal is assumed due to the block design of the stimulus paradigm. This assumption allows for the reduction in dimensionality and the estimation of a noise covariance for signal subspace estimation. However, periodicity is not a trait of all stimulus paradigms. In nonperiodic paradigm designs, the CCA method can still be applied without the SSE step. It can be seen in Tables II and III that CCA performs well even without SSE. Other orthogonal bases, including wavelets and splines, may also be used as the design matrix. These may also allow for estimation of a noise covariance, but we have not explored the details of this approach. CCA can even be applied to event-related approaches where the inter-stimulus interval is random.

In this paper, the assumption is made that the noise in the images is additive white Gaussian. It is known, in fact, that the

noise in magnitude MRI data is Rician [47]. In addition, in dynamic *in vivo* MR imaging, physiologic processes introduce correlated "noise." However, as a first-order approximation, the additive white Gaussian noise model works fairly well. The framework we have presented can be generalized to more complex noise models.

Because the two steps of our method are both entirely self-sufficient, they can be used independently of each other. The SSE methods of the first step can be used in conjunction with multivariate clustering methods other than our CCA. Conversely, the CCA can be used on any dataset which contain feature vectors which have the property of amplitude variation which we discussed in Section I. As a generalization, these methods can also be used in applications outside the realm of fMRI.

## VI. Conclusion

In this paper, we have introduced two novel ideas for the analysis of fMRI timeseries data. The first was a method to reduce the dimensionality and increase SNR by using a signal subspace estimation and simple thresholding strategy. The second was the method of CCA in which the data were iteratively classified independent of signal amplitude. The second method also included a technique to find the number of clusters in an unsupervised manner.

The methodology presented here will allow investigators to improve the estimation of the BOLD signal response by dramatically improving SNR through signal averaging without destroying potentially important statistically distinct temporal elements in the process.

## Appendix I
## Derivation of Noise Covariance

In this Appendix, we derive an estimate for the noise covariance matrix of the noise subspace. Using (7) and defining the matrix $P_A = A(A^tA)^{-1}A^t$, results in the relation $\epsilon = (I - P_A)\nu$. Using the assumption that the noise is white, we have $I\sigma^2 = (1/N)E[\nu\nu^t]$, where $I$ is an identity matrix and $\sigma$ is the variance of the noise. From this we know that $\sigma^2 = (1/(NP))E[\mathrm{tr}\{\nu\nu^t\}]$. Using these results, we derive the following:

$$\begin{aligned} E[\mathrm{tr}\{\epsilon\epsilon^t\}] &= E[\mathrm{tr}\{(I-P_A)\nu\nu^t(I-P_A)^t\}] \\ &= \mathrm{tr}\{(I-P_A)E[\nu\nu^t](I-P_A)^t\} \\ &= \sigma^2 N\,\mathrm{tr}\{(I-P_A)(I-P_A)^t\} \\ &= \sigma^2 N(P-L). \end{aligned}$$

The noise covariance matrix may then be computed

$$\begin{aligned} R_n &= \frac{1}{N}E[\tilde{\Theta}\tilde{\Theta}^t] \\ &= \frac{1}{N}(A^tA)^{-1}A^tE[\nu\nu^t]A(A^tA)^{-1} \\ &= \sigma^2(A^tA)^{-1}A^tA(A^tA)^{-1} \\ &= E[\mathrm{tr}\{\epsilon\epsilon^t\}](A^tA)^{-1}/(N(P-L)) \\ &= E[\hat{R}_n] \end{aligned}$$

where $\hat{R}_n$ is given in (12).

## Appendix II
## Derivation of Whitening Matrix

The corrected noise covariance matrix after subspace processing is denoted as $\bar{R}_n$ where

$$\bar{R}_n = \hat{U}_s^t\hat{R}_n\hat{U}_s. \tag{46}$$

The whitening filter matrix $W$ has the property that $\bar{R}_n^{-1} = WW^t$. Let $\bar{R}_n = U_nS_nV_n^t$ be the singular value decomposition where $S_n$ is a diagonal matrix of singular values. Then, the whitening matrix is is given by

$$W = V_n[\sqrt{S_n}]^{-1} \tag{47}$$

and the whitened feature vectors are given by $Y = W\hat{Y}$.

## Appendix III
## Expectation-Maximization Algorithm

### A. Expectation Step

The $Q$ function used in the EM algorithm is defined as the expectation of the joint log-likelihood function given the current estimates of the parameters. Knowing this, we can write

$$\begin{aligned} &Q\left(K, E_K, \Pi_K; E_K^{(i)}, \Pi_K^{(i)}\right) \\ &\quad = E_{x\,|\,y}\left[\log p_{y,x}(y, X\,|\,E_K, \Pi_K, \hat{\alpha})\,|\,y, K, E_K^{(i)}, \Pi_K^{(i)}, \hat{\alpha}\right] \end{aligned} \tag{48}$$

$$\begin{aligned} &= \sum_{k=1}^{K}[\log p_{y\,|\,x}(y\,|\,k, E_K, \Pi_K, \hat{\alpha})\log p_x(k\,|\,\Pi_K)] \\ &\quad \times p_{x\,|\,y}\left(k\,|\,y, E_K^{(i)}, \Pi_K^{(i)}, \hat{\alpha}\right). \end{aligned} \tag{49}$$

Because of conditional independence, the one-to-one mapping of $x_n \rightarrow y_n$, and $p_x(k\,|\,\Pi_K) = \pi_k$

$$\begin{aligned} &Q\left(K, E_K, \Pi_K; E_K^{(i)}, \Pi_K^{(i)}\right) \\ &\quad = \sum_{k=1}^{K}\left\{\sum_{n=1}^{N}\left[-\frac{1}{2}\left(y_n^ty_n - e_k^ty_ny_n^te_k\right)\right.\right. \\ &\qquad \left.\times\, p_{x_n\,|\,y_n}\left(k\,|\,y_n, E_K^{(i)}, \Pi_K^{(i)}, \hat{\alpha}\right)\right] \\ &\qquad - \frac{M}{2}\log(2\pi)\sum_{n=1}^{N}p_{x_n\,|\,y_n}\left(k\,|\,y_n, E_K^{(i)}, \Pi_K^{(i)}, \hat{\alpha}\right) \\ &\qquad \left.+\,\log\pi_k\sum_{n=1}^{N}p_{x_n\,|\,y_n}\left(k\,|\,y_n, E_K^{(i)}, \Pi_K^{(i)}, \hat{\alpha}\right)\right\}. \end{aligned} \tag{50}$$

Now define

$$\bar{N}_{k\,|\,K}^{(i)} = \sum_{n=1}^{N}p_{x_n\,|\,y_n}\left(k\,|\,y_n, E_K^{(i)}, \Pi_K^{(i)}, \hat{\alpha}\right) \tag{51}$$

and

$$\bar{R}_{k\,|\,K}^{(i)} = \sum_{n=1}^{N}y_ny_n^tp_{x_n\,|\,y_n}\left(k\,|\,y_n, E_K^{(i)}, \Pi_K^{(i)}, \hat{\alpha}\right). \tag{52}$$

We can then write the $Q$ function in its final form

$$Q\left(K, E_K, \Pi_K; E_K^{(i)}, \Pi_K^{(i)}\right) = \sum_{k=1}^{K}\left\{-\frac{1}{2}\operatorname{tr}\left(\bar{R}_{k\,|\,K}^{(i)}\right) + \frac{1}{2}e_k^t \bar{R}_{k\,|\,K}^{(i)} e_k + \bar{N}_{k\,|\,K}^{(i)} + \log \pi_k\right\} - \frac{NM}{2}\log(2\pi). \tag{53}$$

### B. Maximization Step

In order to find $E^{(i+1)}$, we maximize $Q$ with respect to each $e_k, k \in \{1, 2, \ldots, K\}$. We can see that all the terms are constant with respect to $e_k$ except $(1/2)e_k^t \bar{R}_{k\,|\,K}^{(i)} e_k$. So maximizing this factor with respect to $e_k$ is equivalent to maximizing $Q$. The update equation becomes

$$e_k^{(i+1)} = \underset{e_k}{\operatorname{argmax}}\left(e_k^t \bar{R}_{k\,|\,K}^{(i)} e_k\right). \tag{54}$$

It is known from linear algebra theory that the solution to this maximization is the principal eigenvector of $\bar{R}_{k\,|\,K}^{(i)}$. If we let $e_{\max}(R)$ be the principal eigenvector of $R$, we can write the update equation as

$$e_k^{(i+1)} = e_{\max}\left(\bar{R}_{k\,|\,K}^{(i)}\right). \tag{55}$$

Now we need to find $\Pi_K^{(i+1)}$. We can see that all the terms of $Q$ are constant with respect to each $\pi_k$ except for $\bar{N}_{k\,|\,K}^{(i)} \log \pi_k$. Therefore, maximizing this term with respect to $\pi_k$ is equivalent to maximizing $Q$ with respect to $\pi_k$. The problem is a constrained optimization because $\sum_{k=1}^{K} \pi_k = 1$ due to the fact that these are probabilities. If the method of Lagrange multipliers is applied, we find that the update equations for $\Pi_K^{(i+1)}$

$$\pi_k^{(i+1)} = \frac{\bar{N}_{k\,|\,K}^{(i)}}{N}. \tag{56}$$

## Appendix IV
## Derivation of Cluster Merging

In this section, we derive the distance function used for cluster merging in minimization of the MDL criterion. Let $l$ and $m$ denote the indices of the two clusters to be merged. Let $E_K$ and $\Pi_K$ to be the result of running the EM algorithm to convergence with clusters of order $K$, and let $E_{(l,m)\,|\,K}$ and $\Pi_{(l,m)\,|\,K}$ denote new parameter sets in which the parameters for clusters $l$ and $m$ are equated. This means that $\Pi_{(l,m)\,|\,K} = \Pi_K$ and $E_{(l,m)\,|\,K}$ remains the same except for the column vectors corresponding to the clusters $l$ and $m$ which are modified to be

$$e_l = e_m = e_{(l,m)} \tag{57}$$

where $e_{(l,m)}$ denotes the common value of the parameter vectors.

Also define the subscript $E_{(l,m)\,|\,K-1}$ and $\Pi_{(l,m)\,|\,K-1}$ to be parameter sets with $K-1$ clusters in which the $l$ and $m$ clusters have been merged into a single cluster with parameters $e_{(l,m)}$ and $\pi_{(l,m)} = \pi_l + \pi_m$. The change in the MDL criterion produced by merging the clusters $l$ and $m$ is then given by

$$\begin{aligned}&\operatorname{MDL}\left(K-1, E_{(l,m)\,|\,K-1}, \Pi_{(l,m)\,|\,K-1}\right) - \operatorname{MDL}(K, E_K, \Pi_K)\\ &\quad= \operatorname{MDL}\left(K-1, E_{(l,m)\,|\,K-1}, \Pi_{(l,m)\,|\,K-1}\right) - \operatorname{MDL}\left(K, E_{(l,m)\,|\,K}, \Pi_{(l,m)\,|\,K}\right)\\ &\qquad+ \operatorname{MDL}\left(K, E_{(l,m)\,|\,K}, \Pi_{(l,m)\,|\,K}\right) - \operatorname{MDL}(K, E_K, \Pi_K).\end{aligned} \tag{58}$$

From (33), we can see that

$$\operatorname{MDL}\left(K-1, E_{(l,m)\,|\,K-1}, \Pi_{(l,m)\,|\,K-1}\right) - \operatorname{MDL}\left(K, E_{(l,m)\,|\,K}, \Pi_{(l,m)\,|\,K}\right) = -\frac{M}{2}\log(NM). \tag{59}$$

and from the upper bounding properties of the $Q$-function, we know that

$$\begin{aligned}&\operatorname{MDL}\left(K, E_{(l,m)\,|\,K}, \Pi_{(l,m)\,|\,K}\right) - \operatorname{MDL}(K, E_K, \Pi_K)\\ &\quad\le Q(K, E_K, \Pi_K; E_K, \Pi_K) - Q\left(K, E_{(l,m)\,|\,K}, \Pi_{(l,m)\,|\,K}; E_K, \Pi_K\right).\end{aligned} \tag{60}$$

Substituting into (58) results the following inequality:

$$\begin{aligned}&\operatorname{MDL}\left(K-1, E_{(l,m)\,|\,K-1}, \Pi_{(l,m)\,|\,K-1}\right) - \operatorname{MDL}(K, E_K, \Pi_K)\\ &\quad\le Q(K, E_K, \Pi_K; E_K, \Pi_K) - Q\left(K, E_{(l,m)\,|\,K}, \Pi_{(l,m)\,|\,K}; E_K, \Pi_K\right) - \frac{M}{2}\log(NM).\end{aligned} \tag{61}$$

Since we assume that $E_K$ and $\Pi_K$ are the result of running the EM algorithm to convergence, we know that

$$(E_K, \Pi_K) = \underset{E_K', \Pi_K'}{\operatorname{argmax}}\, Q(K, E_K', \Pi_K'; E_K, \Pi_K). \tag{62}$$

Furthermore, the inequality of (58) is most tight when $E_{(l,m)\,|\,K}$ and $\Pi_{(l,m)\,|\,K}$ are chosen to be

$$\left(\hat{E}_{(l,m)\,|\,K}, \hat{\Pi}_{(l,m)\,|\,K}\right) = \underset{E'_{(l,m)\,|\,K}, \Pi'_{(l,m)\,|\,K}}{\arg\max}\, Q\left(K, E'_{(l,m)\,|\,K}, \Pi'_{(l,m)\,|\,K}; E_K, \Pi_K\right). \tag{63}$$

The optimization of (63) is a constrained version of (62). It is easily shown that values of $(\hat{E}_{(l,m)\,|\,K}, \hat{\Pi}_{(l,m)\,|\,K})$ and $(E_K, \Pi_K)$ are equal except for the parameter vector $e_{(l,m)}$ which is given by

$$e_{(l,m)} = e_{\max}\{\bar{R}_{l\,|\,K} + \bar{R}_{m\,|\,K}\} \tag{64}$$

where $e_{\max}\{R\}$ is the principal eigenvector of $R$, and $\bar{R}_{l\,|\,K}$ and $\bar{R}_{m\,|\,K}$ are computed using (30).

Substituting into (61) and simplifying the expression for the $Q$ function results in

$$\begin{aligned}
&\mathrm{MDL}\left(K-1, E_{(l,m)\,|\,K-1}, \Pi_{(l,m)\,|\,K-1}\right) \\
&\quad - \mathrm{MDL}(K, E_K, \Pi_K) \\
&\quad \le Q(K, E_K, \Pi_K; E_K, \Pi_K) \\
&\quad\quad - Q\left(K, \hat{E}_{(l,m)\,|\,K}, \hat{\Pi}_{(l,m)\,|\,K}; E_K, \Pi_K\right) \\
&\quad\quad - \frac{M}{2}\log(NM) \\
&\quad = \sigma_{\max}(\bar{R}_{l\,|\,K}) + \sigma_{\max}(\bar{R}_{m\,|\,K}) \\
&\quad\quad - \sigma_{\max}(\bar{R}_{l\,|\,K} + \bar{R}_{m\,|\,K}) - \frac{M}{2}\log(NM)
\end{aligned} \tag{65}$$

which produces the final result

$$\begin{aligned}
&\mathrm{MDL}\left(K-1, E_{(l,m)\,|\,K}, \Pi_{(l,m)\,|\,K}\right) - \mathrm{MDL}(K, E_K, \Pi_K) \\
&\qquad\qquad \le d(l,m) - \frac{M}{2}\log(NM)
\end{aligned} \tag{66}$$

where $d(l,m)$ is the positive distance function in (34).

## ACKNOWLEDGMENT

The authors would like to thank Brian Cook for help in the development of the stimulus paradigm and Julie Lowe for help in scanning.

## REFERENCES

[1] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proc. Nat. Acad. Sci.*, vol. 87, pp. 9868–9872, 1990.

[2] K. K. Kwong, J. W. Belliveau, D. A. Chesler, I. E. Goldberg, R. M. Weisskoff, B. P. Poncelet, D. N. Kennedy, B. E. Hoppel, M. S. Cohen, R. Turner, H. Cheng, T. J. Brady, and B. R. Rosen, "Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation," *Proc. Nat. Acad. Sci.*, vol. 89, pp. 5675–5679, 1992.

[3] J. W. Belliveau, D. N. Kennedy, R. C. McKinstry, B. R. Buchbinder, R. M. Weisskoff, M. S. Cohen, J. M. Vevea, T. J. Brady, and B. R. Rosen, "Functional mapping of the human visual cortex by magnetic resonance imaging," *Science*, vol. 254, pp. 716–719, 1991.

[4] P. A. Bandettini, E. C. Wong, R. S. Hinks, R. S. Tikofsky, and J. S. Hyde, "Time course EPI of human brain function during task activation," *Magn. Reson. Med.*, vol. 25, pp. 390–397, 1992.

[5] P. A. Bandettini, A. Jesmanowicz, E. C. Wong, and J. S. Hyde, "Processing strategies for time-course data sets in functional MRI of the human brain," *Magn. Reson. Med.*, vol. 30, pp. 161–173, 1993.

[6] K. J. Friston, P. Jezzard, and R. Turner, "Analysis of functional MRI time-series," *Human Brain Mapping*, vol. 2, pp. 69–78, 1994.

[7] G. M. Hathout, S. S. Gambhir, R. K. Gopi, K. A. Kirlew, Y. Choi, G. So, D. Gozal, R. Harper, R. B. Lufkin, and R. Hawkins, "A quantitative physiologic model of blood oxygenation for functional magnetic resonance imaging," *Investigat. Radiol.*, vol. 30, no. 11, pp. 669–682, Nov. 1995.

[8] R. B. Buxton, E. C. Wong, and L. R. Frank, "Dynamics of blood flow and oxygenation changes during brain activation: The balloon model," *Magn. Reson. Med.*, vol. 39, pp. 855–864, 1998.

[9] F. Kruggel and D. Y. von Cramon, "Temporal properties of the hemodynamic response in functional MRI," *Human Brain Mapping*, vol. 8, pp. 259–271, 1999.

[10] V. Solo, P. Purdon, R. Weisskoff, and E. Brown, "A signal estimation approach to functional MRI," *IEEE Trans. Med. Imag.*, vol. 20, pp. 26–35, Jan. 2001.

[11] G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger, "Linear systems analysis of functional magnetic resonance imaging in human V1," *J. Neurosci.*, vol. 16, no. 13, pp. 4207–4221, 1996.

[12] M. S. Cohen, "Parameteric analysis of fMRI data using linear systems methods," *NeuroImage*, vol. 6, pp. 93–103, 1997.

[13] J. J. Sychra, P. A. Bandettini, N. Bhattacharya, and Q. Lin, "Synthetic images by subspace transforms I. Principal components images and related filters," *Med. Phys.*, vol. 21, no. 2, pp. 193–201, Feb. 1994.

[14] W. Backfrieder, R. Baumgartner, M. Samal, E. Moser, and H. Bergmann, "Quatification of intensity variations in functional MR images using rotated principal components," *Phys. Med. Biol.*, vol. 41, pp. 1425–1438, 1996.

[15] M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI data by blind separation into independent spatial components," *Human Brain Mapping*, vol. 6, pp. 160–188, 1998.

[16] M. J. McKeown, T.-P. Jung, S. Makeig, G. Brown, S. S. Kindermann, T.-W. Lee, and T. J. Sejnowski, "Spatially independent activity patterns in functional MRI data during the Stroop color-naming task," *Proc. Nat. Acad. Sci.*, vol. 95, pp. 803–810, 1998.

[17] C. Goutte, P. Toft, E. Rostrup, F. A. Nielsen, and L. K. Hansen, "On clustering fMRI time series," *NeuroImage*, vol. 9, pp. 298–310, 1999.

[18] K.-H. Chuang, M.-J. Chiu, C.-C. Lin, and J.-H. Chen, "Model-free functional MRI analysis using Kohonen clustering neural network and fuzzy $c$-means," *IEEE Trans. Med. Imag.*, vol. 18, pp. 1117–1128, Dec. 1999.

[19] X. Golay, S. Kollias, G. Stoll, D. Meier, A. Valvanis, and P. Boesiger, "A new correlation-based fuzzy logic clustering algorithm for fMRI," *Magn. Reson. Med.*, vol. 40, pp. 249–260, 1998.

[20] A. Baune, F. T. Sommer, M. Erb, D. Wildgruber, B. Kardatzki, G. Palm, and W. Grodd, "Dynamical cluster analysis of cortical fMRI activation," *NeuroImage*, vol. 9, pp. 477–489, 1999.

[21] M. J. Lowe and D. P. Russell, "Treatment of baseline drifts in fMRI time series analysis," *J. Comput. Assist. Tomogr.*, vol. 23, no. 3, pp. 463–473, 1999.

[22] E. T. Bullmore, S. Rabe-Hesketh, R. G. Morris, L. Gregory, S. C. R. Williams, J. A. Gray, and M. J. Brammer, "Function magnetic resonance image analysis of a large-scale neurocognitive network," *NeuroImage*, vol. 4, no. 1, pp. 16–33, Aug. 1996.

[23] S. Chen, C. A. Bouman, and M. J. Lowe, "Harmonic decomposition and eigenanalysis of BOLD fMRI timeseries data in different functional cortices," in *Proc. ISMRM 8th Scientific Meeting*, CO, 2000, p. 817.

[24] B. A. Ardekani, J. Kershaw, K. Kashikura, and I. Kanno, "Activation detection in functional MRI using subspace modeling and maximum likelihood estimation," *IEEE Trans. Med. Imag.*, vol. 18, pp. 101–114, Feb. 1999.

[25] A. F. Sole, S.-C. Ngan, G. Sapiro, X. Hu, and A. Lopez, "Anisotropic 2-D and 3-D averaging of fMRI signals," *IEEE Trans. Med. Imag.*, vol. 20, pp. 86–93, Feb. 2001.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[27] E. Redner and H. Walker, "Mixture densities, maximum likelihood, and the EM algorithm," *SIAM Rev.*, vol. 26, no. 2, Apr. 1984.

[28] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, no. 2, pp. 417–431, Sept. 1983.

[29] C. A. Bouman. (1997, Apr.) Cluster: An unsupervised algorithm for modeling gaussian mixtures. [Online]. Available: http://www.ece.purdue.edu/~bouman

[30] Z. Liang, R. J. Jaszczak, and R. E. Coleman, "Parameter estimation of finite mixtures using the EM algorithm and information criteria with applications to medical image processing," *IEEE Trans. Nucl. Sci.*, vol. 39, pp. 1126–1133, Aug. 1992.

[31] M. J. Lowe, M. Dzemidzic, J. T. Lurito, V. P. Mathews, and M. D. Phillips, "Functional discrimination of thalamic nuclei using BOLD contrast at 1.5 T," in *Proc. ISMRM 8th Scientific Meeting*, CO, 2000, p. 888.

[32] M. D. Phillips, M. J. Lowe, J. T. Lurito, M. Dzemidzic, and V. P. Matthews, "Temporal lobe activation demonstrates sex-based differences during passive listening," *Radiology*, vol. 220, no. 1, pp. 202–207, July 2001.

[33] M. J. Lowe, J. T. Lurito, V. P. Matthews, M. D. Phillips, and G. D. Hutchins, "Quantitative comparison of functional contrast from BOLD-weighted spin-echo and gradient-echo echoplanar imaging at 1.5 T and $H_2\,{}^{15}O$ PET in the whole brain," *J. Cerebral Blood Flow Metabol.*, vol. 20, no. 9, pp. 1331–40, Sept. 2000.

[34] P. Purdon, V. Solo, E. M. Brown, and R. Weisskoff, "Functional MRI signal modeling with spatial and temporal correlations," *NeuroImage*, vol. 14, no. 4, pp. 912–923, Oct. 2001.

[35] S. Makeig. (2001, Sept.) EEG/ICA Toolbox for Matlab. [Online]. Available: http://www.sccn.ucsd.edu/~scott/ica.html

[36] K. J. Friston, A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Human Brain Mapping*, vol. 2, pp. 189–210, 1995.

[37] R. W. Cox, "AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical Research*, vol. 29, pp. 162–173, 1996.

[38] A. L. Vazquez and D. C. Noll, "Nonlinear aspects of the BOLD response in functional MRI," *NeuroImage*, vol. 7, pp. 108–118, 1998.

[39] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 4th ed. Upper Saddle, NJ: Prentice-Hall, 1998.

[40] K. J. Friston, C. D. Frith, P. F. Liddle, and R. S. Frackowiak, "Functional connectivity: The principal-component analysis of large (PET) data sets," *J. Cerebral Blood Flow Metabol.*, vol. 13, no. 1, pp. 5–14, 1993.

[41] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computat.*, vol. 7, pp. 1129–1159, 1995.

[42] K. J. Friston, "Modes or models: A critique on independent component analysis for fMRI," *Trends Cogn. Sci.*, vol. 2, no. 10, pp. 373–375, Oct. 1998.

[43] M. J. McKeown and T. J. Sejnowski, "Independent component analysis of fMRI data: Examining the assumptions," *Human Brain Mapping*, vol. 6, pp. 368–372, 1998.

[44] R. Baumgartner, C. Windischberger, and E. Moser, "Quantification in functional magnetic resonance imaging: Fuzzy clustering vs. correlation analysis," *Magn. Reson. Imag.*, vol. 16, no. 2, pp. 115–125, 1998.

[45] C. Gouttex, L. K. Hansen, M. G. Liptrot, and E. Rostrup, "Feature-space clustering for fMRI meta-analysis," *Human Brain Mapping*, vol. 13, pp. 165–183, 2001.

[46] J. G. Brankov, N. P. Galatsanos, Y. Yang, and M. N. Wernick, "Similarity based clustering using the expectation maximization algorithm," *Proc. IEEE Int Conf. Image Processing*, vol. 1, pp. I-97–I-100, 2002.

[47] J. Sijbers, A. J. den Dekker, J. Van Audekerke, M. Verhoye, and D. Van Dyck, "Estimation of the noise in magnitude MR images," *Magn. Reson. Imag.*, vol. 16, no. 1, pp. 87–90, 1998.