Decoding Behavioral Accuracy in an Attention Task Using Brain fMRI Data

Zhe Wang[*], Yu Zheng[*], Michael Jigo[†], Taosheng Liu[†], Jian Ren[*], Zhi Tian[‡], Tongtong Li[*]
[*]Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA
[†]Department of Psychology, Michigan State University, MI 48824, USA
[‡]Department of Electrical and Computer Engineering, George Mason Unversity, VA 22030, USA
Emails: {wangzh34, zhengy30}@msu.edu, nnana_michael@yahoo.com, {tsliu, renjian}@msu.edu,
ztian1@gmu.edu, tongli@msu.edu

***Abstract*—In this paper, we investigate whether we can distinguish that a subject is making a correct or incorrect behavioral response by analyzing the fMRI data of localized brain regions, obtained from a feature-based attention experiment. For each subject, we first construct the feature vectors for each region of interest (including V1, MT or IPS1) from the fMRI signals. Second, we project the feature vectors onto a lower dimensional subspace using Linear Discriminant Analysis (LDA), where the difference between two classes (correct vs. incorrect response) is maximized. Finally, we apply the Bayesian classifier to the projected data, and find that the classification accuracies corresponding to V1, MT and IPS1 are 87.2%, 90.8% and 81.7%, respectively, when all the trials are considered. Our analysis indicates that: when people make correct or incorrect responses, significant difference exists in the fMRI signals, especially in V1 and MT regions, and the difference can be effectively captured by the LDA-Bayesian classifier. We also prove that: when the original data are normally distributed, LDA, which aims to maximize the difference between different classes, is equivalent to the optimal Maximum Likelihood (ML) based classification method.**

***Index Terms*—Linear Discriminant Analysis, Maximum Likelihood, Bayesian, fMRI**

I. Introduction

Brain is a large network of functionally interconnected brain regions. The "Big Data" problem arises naturally when investigating the huge amount of data generated by the brain fMRI scan, which consists of thousands or even millions of voxels for each subject. Understanding the neural basis of how people make a correct or incorrect response or decision is of great importance to basic and applied research. In this paper, we investigate whether we can tell that a subject is making a correct or incorrect response by analyzing the high dimensional fMRI signals from localized brain regions observed in each individual trial; more specifically, to reduce the dimension of the acquired fMRI data first and then identify an effective classification method.

Our research is based on a feature-based attention experiment [1], carried out by T. Liu and his group at Michigan State University, where they examined the distribution and organization of neural signals related to deployment of feature-based attention following the concept of spatial priority maps [2]–[4]. More specifically, in the experiment, the participants were instructed to attend to one of two overlapping dot fields, one of which rotated in a clockwise (CW) direction and the other rotated in a counter-clockwise (CCW) direction. They were instructed to report the presence or absence of a brief speedup in the cued direction. For each particular trial, if the participant answered "Yes" to a speedup event, it is referred as a correct trial; if the participant answered "No" to a speedup event, it is referred as an incorrect trial.

In this research, we explore the fMRI signals in three brain regions (V1, MT and IPS1) observed during both correct and incorrect trials. The idea is to apply the linear discriminant analysis (LDA) based classifier to the fMRI data to extract the difference between correct and incorrect trials, and use it to determine whether the participant is making a correct response on a particular trial.

LDA is a statistical method in machine learning to separate two or more classes of objects, by projecting them onto a subspace or direction where different classes show most significant differences [5]. Combined with the Bayesian analysis, LDA has been shown to be an efficient technique for dimension reduction and classifications [6]. In [7], Wang et al. applied a Pseudo-Fisher Linear Discriminative Analysis (pFLDA) to classify Alzheimer's Disease patients and normal subjects, and obtained an accuracy of 83%. In [8], Davatzikos et al. applied the LDA based method to classify spatial patterns of brain activity for lie detection. They reported the accuracy of 65.6%.

In this paper, we apply LDA to detect the difference in the fMRI signals of three brain regions (V1, MT and IPS1) obtained during the visual attention experiment mentioned above. The major steps in the procedure include: (i) For each subject, we construct feature vectors for each region from the corresponding fMRI data; (ii) Relying on LDA, we project the feature vectors onto a lower dimensional subspace, where the difference between two classes (correct vs. incorrect response) is maximized; (iii) We then apply the Bayesian classifier to the projected data for classification between the two classes. We show that the final classification accuracies corresponding to V1, MT and IPS1 are 66.7%, 71.5%, 58.5% respectively for the balanced data set (i.e., the number of correct trials and incorrect trials are equal); and 87.2%, 90.8%, 81.7% respectively for the unbalanced data set where all the trials are

This work was partially supported by the National Science Foundation under Grant ECCS-1744604.

978-1-5386-4727-1/18/$31.00 ©2018 IEEE

taken into consideration. The results imply that: *when people make correct or incorrect responses, significant difference exists in the fMRI signals, especially in V1 and MT regions, and the difference can be effectively captured by the LDA-Bayesian classifier.*

In the paper, we also investigate the relationship between LDA-based and Maximum Likelihood (ML) based classification methods. Recall that LDA aims to separate two or more classes by projecting them onto a subspace or direction where the difference between different classes is maximized. Here, we prove that when the original data are normally distributed, LDA is equivalent to maximizing the log-likelihood function of the projected data.

II. Linear Discriminant Analysis and Bayesian Classification

In this section, we revisit the Linear Discriminant Analysis method for dimension reduction and the Bayesian classifier for classification of the projected data. We will also investigate the relationship between LDA and the maximum likelihood based classification method in this section.

A. Linear Discriminant Analysis

Linear Discriminant Analysis aims to separate two or more classes by projecting the data onto a subspace or direction where different classes show most significant differences [5]. Here, we illustrate the basic idea of LDA using a two-class case. Suppose we have a set of $d-$dimensional vector samples $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, where N_1 of them are from the first class, denoted as C_1, N_2 of them are from the second class, denoted as C_2, and $C = \{C_1, C_2\}$. For $i = 1, 2$, the mean and scatter matrix (i.e., the scaled covariance matrix) of each of the two classes are defined as:

$$\begin{aligned} \boldsymbol{\mu}_i &= \frac{1}{N_i}\sum_{\mathbf{x}\in C_i}\mathbf{x}, \qquad (1)\\ S_i &= \sum_{\mathbf{x}\in C_i}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^t. \qquad (2)\end{aligned}$$

Consider the projection of vectors in X to a d_0-dimensional space, where $d_0 \le d$, such that the difference between or among different classes is maximized. More specifically, let $\mathbf{y} = W\mathbf{x}$, $\mathbf{x} \in X$, where W is a $d_0 \times d$ matrix to be determined by the LDA algorithm. In this paper, we choose $d_0 = 1$, and let $W = \mathbf{w}^t$. As a result, the transform can be rewritten as:

$$y = \mathbf{w}^t\mathbf{x}, \qquad (3)$$

where $\mathbf{w}^t$ is a $1 \times d$ vector. For $i = 1, 2$, let

$$\tilde{C}_i = \{y | y = \mathbf{w}^t\mathbf{x},\ \mathbf{x} \in C_i\}. \qquad (4)$$

Define $\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$ as the overall mean, $S_W = \sum_{i=1}^{2} S_i$ as the within-class scatter matrix, and $S_B = \sum_{i=1}^{2} N_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^t$ as the between-class scatter matrix. LDA seeks a transform vector $\mathbf{w}$ that maximizes the following objective function:

$$J(\mathbf{w}) \;=\; \frac{\mathbf{w}^t S_B \mathbf{w}}{\mathbf{w}^t S_W \mathbf{w}}. \qquad (5)$$

It can be proved [5], [6] that to maximize $J(\mathbf{w})$, $\mathbf{w}$ should satisfy

$$S_W^{-1} S_B \mathbf{w} = \lambda\mathbf{w}, \qquad (6)$$

for some constant λ. That is, $\mathbf{w}$ is an eigenvector of $S_W^{-1}S_B$. Following the eigenvalue decomposition of matrix $S_W^{-1}S_B$, LDA chooses the eigenvector corresponding to the largest eigenvalue of the matrix $S_W^{-1}S_B$ as $\mathbf{w}$. After $\mathbf{w}$ is found, we project the original vectors $\mathbf{x} \in X$ to a one dimensional subspace using $y = \mathbf{w}^t\mathbf{x}$. We can then apply a classifier on the projected data y.

B. Bayesian Classification

Here we apply the Bayesian classifier [7] due to its simplicity and effectiveness in dealing with unbalanced dataset. Denoting the projected data as y, the discriminant function for the Bayesian classifier is

$$g_i(y) = \ln p(y|\tilde{C}_i) + \ln P(\tilde{C}_i), \qquad (7)$$

where $i = 1, 2$. We will see in the next subsection that the projected data y can be approximately characterized by the Gaussian distribution, thus $p(y|\tilde{C}_i)$ can be evaluated using the Gaussian probability density function (PDF). Assuming a general case where the random variable y is Gaussian with PDF $N(m_i, \sigma_i^2)$. In this case, from (7), we have

$$g_i(y) = -\frac{(y-m_i)^2}{2\sigma_i^2} - \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln\sigma_i + \ln P(\tilde{C}_i), \qquad (8)$$

where $i = 1, 2$. In a general case, the mean m_i and variance σ_i for each category $\tilde{C}_i$ are different. The common term, $-\frac{p}{2}\ln 2\pi$, in (8) can then be dropped, and the resulting discriminant function has the quadratic form:

$$g_i(y) = a_i y^2 + b_i y + c_i\ , \qquad (9)$$

where

$$\begin{aligned} a_i &= -\frac{1}{2\sigma_i^2}, \qquad (10)\\ b_i &= \sigma_i^{-2} m_i\ , \qquad (11)\\ c_i &= -\frac{m_i^2}{2\sigma_i^2} - \frac{1}{2}\ln\sigma_i + \ln P(\tilde{C}_i). \qquad (12)\end{aligned}$$

The decision can be made as follows: if $g_1(y) > g_2(y)$, the sample y can be assigned to class $\tilde{C}_1$; otherwise, if $g_1(y) < g_2(y)$, y will be assigned to class $\tilde{C}_2$.

C. LDA and the Maximum Likelihood Method

In this subsection, we demonstrate that when the original data from all classes are normally distributed, then LDA is equivalent to the optimal maximum likelihood method. For $i = 1, 2$, assuming each vector $\mathbf{x} \in C_i$ has the same probability density function (pdf):

$$f_X(\mathbf{x}; \boldsymbol{\mu}_i, \Sigma_i) = \frac{1}{\sqrt{2\pi^d |\Sigma_i|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \Sigma_i{}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}. \tag{13}$$

Consider a very general linear transform defined by: $\mathbf{y} = W\mathbf{x}$, where W is a $d \times d$ matrix. For the transformed data, the probability density function becomes:

$$f_Y(\mathbf{y}; \tilde{\boldsymbol{\mu}}_i, \tilde{\Sigma}_i) = \frac{1}{\sqrt{2\pi^d |\tilde{\Sigma}_i|}} e^{-\frac{1}{2}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_i)^t \tilde{\Sigma}_i^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_i)}, \tag{14}$$

where $i = 1, 2$, $\tilde{\boldsymbol{\mu}}_i = W\boldsymbol{\mu}_i$, and $\tilde{\Sigma}_i = W\Sigma_i W^t$.

Recall that in LDA, we try to find W such that the difference among different classes is maximized in the transformed space. Without loss of generality, we assume that the major difference lies in the first dimension of the transformed vector $\mathbf{y}$ only, and the remaining $d-1$ dimensions make little contributions. Under this assumption, $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\Sigma}_i$ can be decomposed into two parts as:

$$\tilde{\boldsymbol{\mu}}_i = \begin{bmatrix} \tilde{\mu}_i^1 \\ \tilde{\boldsymbol{\mu}}^{d-1} \end{bmatrix}, \tag{15}$$

$$\tilde{\Sigma}_i = \begin{bmatrix} \tilde{\Sigma}_i^1 & \mathbf{0} \\ \mathbf{0} & \tilde{\Sigma}^{d-1} \end{bmatrix}, \tag{16}$$

since for each i, $\tilde{\boldsymbol{\mu}}_i^{d-1} \approx \tilde{\boldsymbol{\mu}}^{d-1}$, $\tilde{\Sigma}_i^{d-1} \approx \tilde{\Sigma}^{d-1}$. Accordingly, the matrix W can also be decomposed into two parts as

$$W = \begin{bmatrix} W^1 \\ W^{d-1} \end{bmatrix}. \tag{17}$$

In this case, we have $\tilde{\mu}_i^1 = W^1\boldsymbol{\mu}_i$, $\tilde{\boldsymbol{\mu}}_i^{d-1} = W^{d-1}\boldsymbol{\mu}_i$, $\tilde{\Sigma}_i^1 = W^1\Sigma_i {W^1}^t$ and $\tilde{\Sigma}_i^{d-1} = W^{d-1}\Sigma_i {W^{d-1}}^t$.

For fairness, in LDA based classification, the sample size of the two classes is assumed to be the same, i.e., $N_1 = N_2 = N/2$. With the probability density function given in (14), the log-likelihood function of the original data $\mathbf{x}$ can be written as [9]:

$$\begin{aligned} & L(W) \\ & = \sum_{i=1}^{2} \sum_{\mathbf{y}\in\tilde{C}_i} log\ |W| f_Y(W\mathbf{x}; \tilde{\boldsymbol{\mu}}_i, \tilde{\Sigma}_i) \\ & = nlog|W| - \frac{n}{2}log(2\pi)^d - \sum_{i=1}^{2} \frac{n_i}{2} log|\tilde{\Sigma}_i^1| \\ & - \frac{1}{2}\sum_{i=1}^{2} \sum_{\mathbf{x}\in C_i} (W^1\mathbf{x} - \tilde{\mu}_i^1)^t (\tilde{\Sigma}_i^1)^{-1} (W^1\mathbf{x} - \tilde{\mu}_i^1) \\ & - \frac{n}{2} log|\tilde{\Sigma}^{d-1}| \\ & - \frac{1}{2}\sum_{\mathbf{x}\in C} (W^{d-1}\mathbf{x} - \tilde{\boldsymbol{\mu}}^{d-1})^t (\tilde{\Sigma}^{d-1})^{-1} (W^{d-1}\mathbf{x} - \tilde{\boldsymbol{\mu}}^{d-1}) \end{aligned} \tag{18}$$

To find the optimal W that maximizes $L(W)$, we set

$$\frac{\partial L(W)}{\partial \tilde{\Sigma}_i^1} = 0, \tag{19}$$

$$\frac{\partial L(W)}{\partial \tilde{\Sigma}^{d-1}} = 0. \tag{20}$$

Therefore, we obtain

$$\tilde{\Sigma}_i^1 = W^1 S_W {W^1}^t, \tag{21}$$

$$\tilde{\Sigma}^{d-1} = W^{d-1} S_B {W^{d-1}}^t. \tag{22}$$

Substitute (21) and (22) into (18) and remove the constant items, the optimization of $L(W)$ is equivalent to optimizing the following function:

$$\begin{aligned} L_{eq}(W) &= Nlog|W| - \frac{N}{2} log|W^1 S_W {W^1}^t| \\ &- \frac{N}{2} log|W^{d-1} S_B {W^{d-1}}^t|. \end{aligned} \tag{23}$$

The optimal choice of W will satisfy the differential equation:

$$\frac{dL_{eq}(W)}{dW} = 0. \tag{24}$$

After some manipulations as shown in the Appendix, the solution to the optimization problem (24) is composed of eigenvectors of the matrix $S_W^{-1} S_B$.

If we only keep the eigenvector corresponding to the largest eigenvalue of $S_W^{-1} S_B$, then we obtain the LDA algorithm presented in Section II-A [10]–[12]. As can be seen, if the data follow Gaussian distribution, LDA is equivalent to the optimal maximum likelihood decision making method.

III. LDA-Bayesian Based Classification of Correct/Incorrect Trials

In this section, we will construct feature vectors from the fMRI data of the visual attention experiment, and demonstrate how to conduct the LDA-Bayesian based classification of correct and incorrect trials from the feature vectors.

A. fMRI Data Acquisition

A total of 12 observers participated in the feature-based attention experiment; all had normal or corrected-to-normal vision. Two of the subjects were authors, and the rest were graduate and undergraduate students at Michigan State University, all of whom gave written informed consent and were compensated for their participation. The experimental procedures were approved by the Institutional Review Board at Michigan State University and adhered to safety guidelines for MRI research.

In the attention task, the subjects were cued to attend to one of two overlapping and rotating dot fields, and reported the presence or absence of a brief speedup in the cued direction. The stimuli were composed of two dot fields that rotated clockwise (CW) and counter-clockwise (CCW) respectively, with 60% motion coherence (dot color: gray, luminance: $147.4cd/m^2$; dot size: $0.1°$; density: $1.1dots/deg^2$). Each dot within the pattern had a lifetime of 6 frames to reduce the possibility of tracking individual dots. The dot field was contained within an annulus (eccentricity from $2.5°$ to $8°$) that was centered on a central cross (size: $0.5°$) and displayed on a black background (luminance: $0.05cd/m^2$).

Each trial began with a leftward or rightward-pointing arrow cue that appeared $0.77°$ above the central cross and instructed subjects to attend to CCW or CW motion, respectively. After $0.3s$, the cue was replaced with spatially overlapping CW and CCW dot fields that were displayed for $4.1s$. Each pattern rotated around the center of the annulus at a speed of $45°/s$ before a brief ($0.2s$) speed increment occurred in either direction; on 80% of trials, the speedup occurred in the cued direction (valid trials). On each valid trial, the magnitude of the speedup was adjusted via best PEST to maintain a hit rate of 65%. On the other 20% of trials, a speedup occurred in the uncued direction (invalid trials), using the magnitude of the preceding valid trial (i.e., speedup on invalid trials was not controlled via staircase). An inter-trial interval (ITI) followed the speedup: the ITI was $4.2s$ on 40% of trials, $6.4s$ on 30% of trials, $8.6s$ on 20% of trials, and $10.8s$ on 10% of trials. During this interval, subjects reported the presence or absence of the speedup in the cued direction with corresponding Present or Absent button presses. For each valid trial, if the participant answered "Yes" to a speedup event, it is referred as a correct trial; if the participant answered "No" to a speedup event, it is referred as an incorrect trial. We did not analyze invalid trials due to the small number of trials.

All subjects completed 10 fMRI runs, which yielded 180 valid trials. Each run began with an $8.8s$ fixation period and lasted $338.8s$; the images collected during the fixation period were discarded to avoid magnetic saturation effects. Cue direction (CW vs. CCW) and validity (valid vs. invalid) were randomized within each run of the attention task, and motion direction (CW vs. CCW) and trial type (speedup vs. catch) were both randomized within each run of the baseline task.

Functional data were preprocessed according to standard methods. In addition, for each participant we independently defined several regions-of-interest (ROI) in the occipital and parietal cortex using standard retinotopic mapping procedure (for details of preprocessing and mapping procedure, see [1]). In this paper, we focus on three ROIs: V1, MT, and IPS1 because the previous research [1] has suggested these regions as the important brain areas in the visual perception and attention tasks.

B. Feature Vector Construction

For every participant, we construct the feature vectors for each ROI (V1, MT or IPS1) separately. After preprocessing, for each subject, the fMRI data of each region are organized as a 3D matrix:

$$\{v_{S,R}(m,t,k)|m=1,...,M;t=0,...,4;k=1,...,K\},$$

where $S=1,2,...,12$, represents the subject index; R denotes the ROI region under investigation, which is either V1, MT or IPS1; m is the voxel index within the region; M denotes the total number of voxels in the region, which may vary from region to region and from subject to subject; t denotes the time instant or sampling time index, k the trial index and K the total number of trials. We then add all the 5 signal samples within each trial together to formulate an $M \times K$ matrix $\tilde{v}_{S,R}$, where

$$\tilde{v}_{S,R}(m,k)=\sum_{t=0}^{4} v(m,t,k).$$

For region R of subjects S, the feature vector for each trial k is obtained by stacking the fMRI signal of all the voxels within the region into an $M-$dimensional vector, and can be represented as:

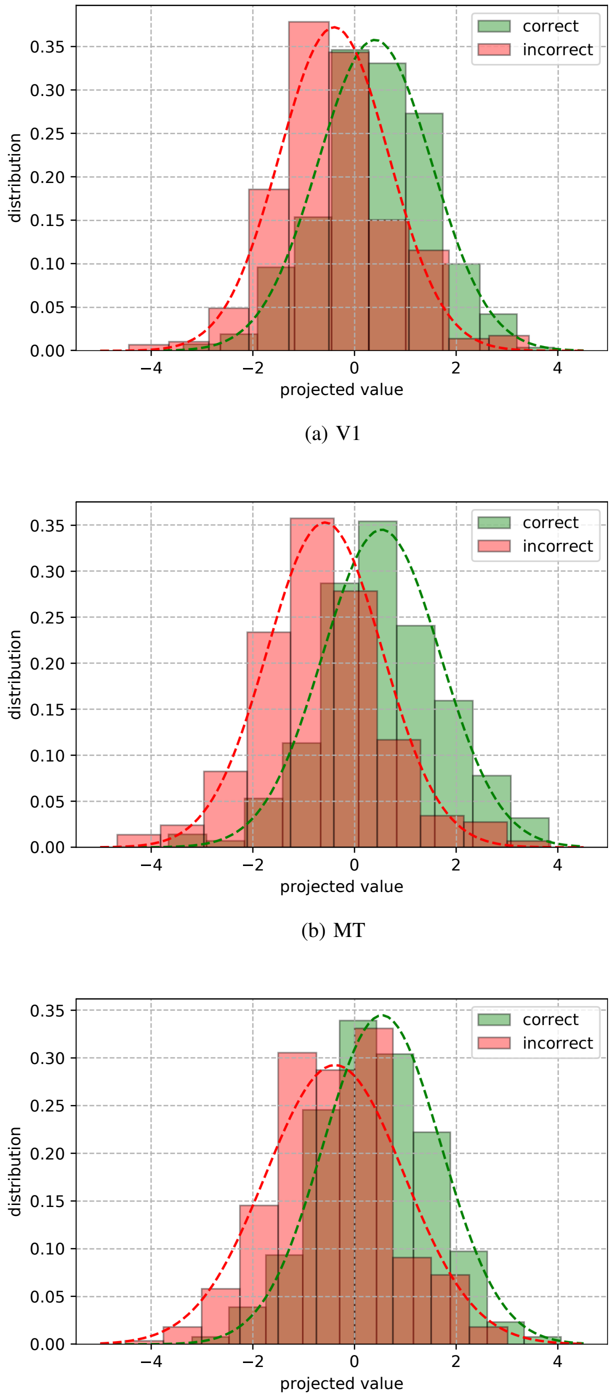
$$\mathbf{x}_k=[\tilde{v}_{S,R}(1,k),...,\tilde{v}_{S,R}(M,k)]^t, k=1,...,K.$$

C. LDA-Bayesian Based Classification

First, we reduce the dimensionality of the feature vectors using LDA by projecting them onto a $1-$dimensional subspace, where the difference between the correct trial class ($\tilde{C}_1$) and the incorrect trial class ($\tilde{C}_2$) is maximized. More specifically, $y_k=\mathbf{w}^t\mathbf{x}_k, k=1,2,...,K$, where $\mathbf{w}$ is the transform vector determined by LDA.

Second, we classify each y_k to class $\tilde{C}_1$ or $\tilde{C}_2$ using the Bayesian classifier, which maximizes the posterior probability $P(\tilde{C}_i|y_k)$, $i=1,2$. The Bayesian classifier is chosen here for its simplicity and effectiveness [13].

More specifically, given a projected data sample $y_k=\mathbf{w}^t\mathbf{x}_k$, in which $\mathbf{w}$ is the transform vector determined by LDA. For $k=1,2,...,K$, if $y_k<b$ (the boundary parameter determined by the classifier), trial k will be assigned to class $\tilde{C}_2$; otherwise, it will classify it to $\tilde{C}_1$. This implies the existence of an attention axis in the space of the feature vectors. For each trial, if the angle between the feature vector and the axis is smaller than a certain threshold, then the participant is more likely to make a correct response in this



(a) V1

(b) MT

(c) IPS1

Fig. 1: Histogram of the projected data from V1, MT and IPS1 regions of a single subject.

trial. Otherwise, the participant is more likely to make an incorrect response.

The histograms of the projected values $\{y_k\}, k = 1, 2, ..., K$, corresponding to region V1, MT and IPS1 of a particular subject are provided in Figure 1. As can be seen, the distribution of $\{y_k\}$ is reasonably close to Gaussian, although there are some deviations due to limited data size. Based on our analysis in Section II.C, if the data follow Gaussian distribution, LDA is equivalent to the optimal maximum likelihood decision making method.

To show that significant difference exists between correct and incorrect trials in the projected data in the three brain regions, i.e., V1, MT and IPS1 regions, we carried out a statistical significant test on the project values $\{y_k\}$, $k = 1, 2, ..., K$. In the two sample t-test [14], all the three brain regions show $p < 0.05$, which rejects the the null hypothesis that the expectation of the projected value are the same for correct and incorrect trials. Thus, it supports that the LDA algorithm successfully separated the data into correct and incorrect trial groups with different means. It is also interesting to see that the projected data are most separable for MT region and least for IPS1 which will affects the classification accuracy in subsequent analysis.

D. *Numerical Results*

The performance of the classification accuracy is tested under two scenarios:

- Case 1: In this case we use a balanced dataset, where we randomly select 40 correct trials and 40 incorrect trials. Then the classification accuracies corresponding to V1, MT and IPS1 are 66.7%, 71.5% and 58.5%, respectively.
- Case 2: In this case we use the unbalanced data, i.e., all the 180 trials are utilized, where 78% of the trials are correct and 22% are incorrect. The classification accuracies corresponding to V1, MT and IPS1 regions are 87.2%, 90.8% and 81.7%, respectively

The resulting range of accuracy is consistent with the result in [1]. The high accuracy for V1 and MT in both cases suggests that significant difference in the fMRI signals exists in those two regions during this experiment.

The main reasons for the lower accuracy in case 1 are: (i) Case 1 has a relatively small data size, which may reduce the accuracy of the classifier due to insufficient data. (ii) It does not utilize the full information, and lacks the prior information on the correctness of the trials in the balanced data case, as a result, it does not fully exploit the benefit brought by the Bayesian classifier.

IV. Conclusions

In this research, we explored the fMRI signals in three brain regions (V1, MT and IPS1) observed during a cued attention experiment. We applied the LDA-Bayesian classifier to the fMRI data to extract the difference between correct and incorrect trials, and used it to determine whether the participant was making a correct response in a test trial. We showed that the classification accuracies corresponding to V1, MT and IPS1 are 66.7%, 71.5%, 58.5% respectively for the balanced data set, where the number of correct trials and incorrect trials are equal; and 87.2%, 90.8%, 81.7% respectively for the unbalanced data set where all the trials are taken into consideration. This implies that: when people make correct or incorrect responses, significant difference exists in the fMRI signals, especially in the V1 and MT regions, and the difference can be effectively captured by the LDA-Bayesian classifier.

Appendix

In this appendix, we give a brief derivation on the solution to the optimization problem in equation (24).

From (24), we can immediately obtain

$$\frac{\partial L_{eq}(W)}{\partial W^1} = 0, \tag{25a}$$

$$\frac{\partial L_{eq}(W)}{\partial W^{d-1}} = 0. \tag{25b}$$

Based on the following identities for computing the derivative with respect to determinant [15], [16]:

$$\frac{\partial \log |\mathbf{X}^t\mathbf{C}\mathbf{X}|}{\partial \mathbf{X}} \;=\; 2(\mathbf{C}\mathbf{X}(\mathbf{X}^t\mathbf{C}\mathbf{X})^{-1}), \tag{26}$$

it turns out that (25) is equivalent to

$$\frac{\partial log|W|}{\partial W^1} \quad = (W^1S_W(W^1)^t)^{-1}W^1S_W, \tag{27a}$$

$$\frac{\partial log|W|}{\partial W^{d-1}} \quad = (W^{d-1}S_B(W^{d-1})^t)^{-1}W^{d-1}S_B. \tag{27b}$$

By stacking equations (27a) and (27b), we get

$$\left[\begin{array}{c} \frac{\partial log|W|}{\partial W^1} \\ \frac{\partial log|W|}{\partial W^{d-1}} \end{array}\right] = \left[\begin{array}{c} (W^1S_W(W^1)^t)^{-1}W^1S_W \\ (W^{d-1}S_B(W^{d-1})^t)^{-1}W^{d-1}S_B \end{array}\right], \tag{28}$$

Following the identity [15]:

$$\frac{\partial \log |\mathbf{X}|}{\partial \mathbf{X}} \;=\; \mathbf{X}^{-t}, \tag{29}$$

the left-hand side of (28) is equivalent to

$$\left[\begin{array}{c} \frac{\partial log|W|}{\partial W^1} \\ \frac{\partial log|W|}{\partial W^{d-1}} \end{array}\right] = \frac{\partial log|W|}{\partial W} = W^{-t}\ . \tag{30}$$

Therefore, by combining (30) and (28), we get

$$\mathbf{I}_d = \left[\begin{array}{c} (W^1S_W(W^1)^t)^{-1}W^1S_W \\ (W^{d-1}S_B(W^{d-1})^t)^{-1}W^{d-1}S_B \end{array}\right]\cdot W^t, \tag{31}$$

which is equivalent to

$$(W^1S_W(W^1)^t)^{-1}W^1S_W(W^{d-1})^t = \mathbf{0}, \tag{32a}$$

$$(W^{d-1}S_B(W^{d-1})^t)^{-1}W^{d-1}S_BW^1 = \mathbf{0}. \tag{32b}$$

It can be further simplified as

$$W^1S_W(W^{d-1})^t = \mathbf{0}, \tag{33a}$$

$$W^{d-1}S_B(W^1)^t = \mathbf{0}. \tag{33b}$$

Suppose that W is nonsingular, we can set W as

$$W = \Phi S_W^{-1/2}. \tag{34}$$

Substituting (34) to (33), we have

$$\Phi^1(\Phi^{d-1})^t \;=\; \mathbf{0}, \tag{35a}$$

$$\Phi^{d-1}S_W^{-1/2}S_BS_W^{-t/2}\Phi^1 \;=\; \mathbf{0}. \tag{35b}$$

Equations (35a) and (35b) are simultaneously satisfied when the rows of Φ corresponds to the orthogonal eigenvecotors of $S_W^{-1/2}S_BS_W^{-t/2}$, that is

$$S_W^{-1/2}S_BS_W^{-t/2}\Phi^t = \Phi^t\Lambda, \tag{36}$$

where $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_d)$. Substituting (34) into (36) we get

$$S_W^{-1}S_BW^t = W^t\Lambda, \tag{37}$$

which means that W consists of the eigenvectors of $S_W^{-1}S_B$. It is sufficient to keep the eigenvector corresponding to the largest eigenvalue if we seek to project the data into a one dimensional space.

References

[1] T. Liu, L. Hospadaruk, D. C. Zhu, and J. L. Gardner, “Feature-specific attentional priority signals in human cortex,” *Journal of Neuroscience*, vol. 31, no. 12, pp. 4484–4495, 2011.

[2] C. Koch and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry,” in *Matters of intelligence*. Springer, 1987, pp. 115–141.

[3] J. M. Wolfe, “Guided search 2.0 a revised model of visual search,” *Psychonomic bulletin & review*, vol. 1, no. 2, pp. 202–238, 1994.

[4] L. Itti and C. Koch, “Computational modelling of visual attention,” *Nature reviews neuroscience*, vol. 2, no. 3, p. 194, 2001.

[5] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[7] K. Wang *et al.*, “Discriminative analysis of early Alzheimers disease based on two intrinsically anti-correlated networks with resting-state fMRI,” *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*, pp. 340–347, 2006.

[8] C. Davatzikos, K. Ruparel, Y. Fan, D. Shen, M. Acharyya, J. Loughead, R. Gur, and D. D. Langleben, “Classifying spatial patterns of brain activity with machine learning methods: application to lie detection,” *Neuroimage*, vol. 28, no. 3, pp. 663–668, 2005.

[9] I. J. Myung, “Tutorial on maximum likelihood estimation,” *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[10] H. Zhou, D. Karakos, S. Khudanpur, A. G. Andreou, and C. E. Priebe, “On projections of gaussian distributions using maximum likelihood criteria,” in *Information Theory and Applications Workshop, 2009*. IEEE, 2009, pp. 431–438.

[11] S. R. Searle, “Matrix algebra useful for statistics,” *New York*, vol. 1982, 1982.

[12] N. Kumar and A. G. Andreou, “Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition,” *Speech communication*, vol. 26, no. 4, pp. 283–297, 1998.

[13] I. Rish, “An empirical study of the naive bayes classifier,” in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. IBM New York, 2001, pp. 41–46.

[14] J. H. McDonald, *Handbook of biological statistics*. sparky house publishing Baltimore, MD, 2009, vol. 2.

[15] K. B. Petersen, M. S. Pedersen *et al.*, “The matrix cookbook,” *Technical University of Denmark*, vol. 7, no. 15, p. 510, 2008.

[16] M. Brookes, “The matrix reference manual,” *Imperial College London*, vol. 3, 2005.