An Information Theoretic Approach to Classify Cognitive States Using fMRI

Itir Onal, Mete Ozay, *Student Member, IEEE*, Orhan Firat, *Student Member, IEEE*, Ilke Öztekin, Fatos T. Yarman Vural, *Member, IEEE*

Abstract— **In this study, an information theoretic approach is proposed to model brain connectivity during a cognitive processing task, measured by functional Magnetic Resonance Imaging (fMRI). For this purpose, a local mesh of varying size is formed around each voxel. The arc weights of each mesh are estimated using a linear regression model by minimizing the squared error. Then, the optimal mesh size for each sample, that represents the information distribution in the brain, is estimated by minimizing various information criteria which employ the mean square error of linear regression model. The estimated mesh size shows the degree of locality or degree of connectivity of the voxels for the underlying cognitive process.**

The samples are generated during an fMRI experiment employing item recognition (IR) and judgment of recency (JOR) tasks. For each sample, estimated arc weights of the local mesh with optimal size are used to classify whether it belongs to IR or JOR tasks. Results indicate that the suggested connectivity model with optimal mesh size for each sample represent the information distribution in the brain better than the state-of –the art methods.

I. INTRODUCTION

There has been a growing body of recent neuroimaging research investigating how information is distributed in the brain using functional Magnetic Resonance Imaging (fMRI) measurements. In this approach, a method called multi-voxel pattern analysis (MVPA), utilizes machine learning algorithms to extract and classify distributed patterns of brain activity [1 - 7]. Generally, the brain state during a cognitive process is measured via fMRI and intensity values of multiple voxels which are concatenated under a feature vector to train a well-known classifier, such as Neural Networks, Naïve Bayes, k-Nearest Neighbor (k-NN) or Support Vector Machine (SVM). The classifier is, then tested with an unknown feature vector representing a cognitive state, or type of information. The classification performance can be used to infer the accuracy of the machine learning model in successfully representing the underlying cognitive state.

This approach represents a cognitive process by a large and fixed size feature vector (as large as the number of active voxels, usually, in the order of several thousands). However, it does not model the degree of connectivity among the voxels.

In this study, the distributive nature of discriminative information in the brain is investigated by an information theoretic approach. Instead of using the feature vector formed by voxel intensity values, a set of spatially local meshes [8] which represents the spatial relationships among voxels is used to represent a sample (cognitive state of a person). Around each voxel, called seed voxel, a local mesh is formed with its spatially nearest neighbors and the relationships among the seed voxel and its neighbors are represented as the arc weights of the mesh. For each local mesh, these arc weights are estimated using a linear regression model. The optimal mesh size for each sample is estimated by maximizing some information theoretic criteria. The error variance is used in calculating various information theoretic criteria for model order selection. Therefore, the problem of estimating the brain connectivity is formulated as a model order selection problem. In this study, three different information criteria, namely Akaike's Information Criterion (AIC) [9], Bayesian Information Criterion (BIC) [10] and Rissanen's Minimum Description Length Method (MDL) [11] are used to find the optimal order of the regression model, in other words to select the optimal mesh size for each sample.

Among them, AIC assumes that, there is an unknown fMRI data generating process in the brain and AIC attempts to approximate this unknown. Hence, it selects the model order, i.e. optimal mesh size around a voxel as the one that "best" approximates this unknown process.

On the other hand, BIC finds the likelihood of the mesh model formed around each voxel that generates the fMRI data by using a prior probability. This likelihood is used for order selection among a finite set of mesh models. Among candidate models, it selects the "true" model as the one that maximizes the posterior probability. Hence, BIC selects the optimal mesh size among a finite set such that a local mesh having the optimal mesh size is most likely to represent the fMRI data among others.

MDL has a totally different approach than the previous two criteria. It aims to find the mesh model that best represents the information. MDL is a formalization of Occam's razor such that, it assumes the best model that represents the information as the one that leads to the best compression of data. Therefore, it finds the optimal mesh size such that local mesh having the optimal mesh size best represents the information in a compressed manner.

The major assumption of this study is that the brain makes a trade-off between the degree of complexity (increasing mesh size) and the degree of fit (decreasing error), as in the above mentioned information criteria.

I. Onal, M. Ozay, O. Firat, and F. T. Yarman Vural are with the Department of Computer Engineering, Middle East Technical University, 06800, Ankara-Turkey (e-mail: itir@ceng.metu.edu.tr, mozay@metu.edu.tr, orhan.firat@ceng.metu.edu.tr, vural@ceng.metu.edu.tr).

İ. Oztekin is with the Department of Psychology, Koç University, 06800, İstanbul-Turkey (e-mail: ioztekin@ku.edu.tr).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

Therefore, the optimal mesh size is the one that wins this trade-off and minimizes an information criterion, such as AIC, BIC or MDL with respect to mesh size *p*. For each sample, which represents a specific cognitive task applied on a person and for each criterion, an optimal mesh size is estimated as the one that minimizes the related information criterion. Hence, in this work, we assume that the degree of connectivity changes in the brain depending on the cognitive processes and depending on an individual person, and this varying distributive nature can be modeled by a local mesh model, where the model order is estimated by the above mentioned information criteria. When compared to MVPA method, in which a fixed-size vector of voxel intensity values are used to train the classifier, the suggested local mesh model with a variable mesh size achieves better classification performance. This result supports the idea of using information theoretic criteria with local relational structure is promising to represent brain connectivity.

II. Methods

In this study, fMRI was used to record neural activation during two working memory tasks, namely item recognition (IR) and judgment of recency (JOR) [12]. For both IR and JOR tasks, each trial began with the presentation of a centered fixation point for 500 ms. Then a study list including five consonants were presented one at a time for 500 ms each. After the presentation of the study list, a task cue was presented to indicate the upcoming memory judgment (IR or JOR) for 750 ms. Following the presentation of task cues, two probe consonants were presented for both tasks for 3000 ms. In IR trials, one consonant is from the study list where the other one was new. Participants were requested to indicate the one belonging to the study list in this task with a button press. In JOR trials on the other hand, both probes were from the study list, and participants were asked to select the probe that was more recent in the study list (Fig. 1). Preprocessing of neuroimaging data steps included slice acquisition timing across slices, realignment of images to the first volume for head movement correction, normalization of anatomical and functional images to a standard template EPI and smoothing of images with a 6-mm full-width half-maximum isotropic Gaussian kernel.

III. Spatially local mesh model

In this study, the fMRI intensity values measured at each voxels $v(t_i, \bar{s}_j)$ at each time instant t_i, $i = 1,2, \dots N$, where *N* is the number of samples at spatial coordinates $\bar{s}_j$, $j = 1,2, \dots M$ and M is the number of voxels, are used to model the cognitive states. The voxels are distributed in the brain in three dimensions. Therefore, $\bar{s}_j$ represents three dimensional voxel coordinates, $\bar{s}_j = (x_j, y_j, z_j)$. fMRI measurements are concatenated under an *NxM* matrix, where each row of this matrix corresponds to a feature vector of intensity values of voxels $v(t_i, \bar{s}_j)$, in a time instant t_i. Therefore, a sample is represented by a vector of voxel intensity values acquired from a single trial. Each sample at time instant t_i is also associated with a task label c_i, where $c_i \in \{IR, JOR\}$.

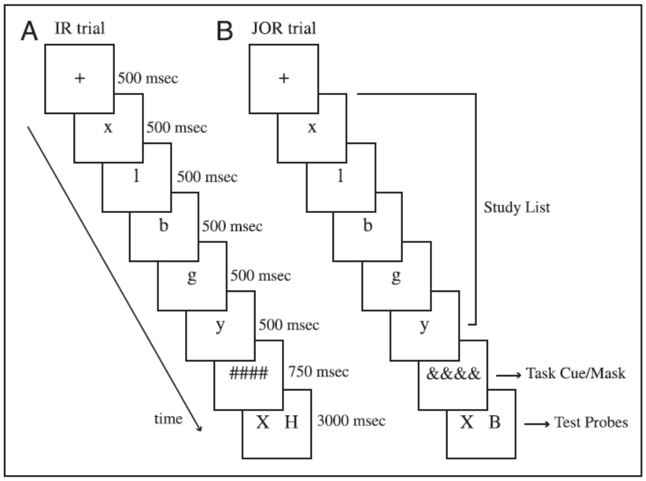

Fig. 1. Oztekin et al., 2009. A sample sequence for an item recognition (IR) trial (shown on panel A) and a judgment of recency (JOR) trial (shown on panel B). After the presentation of fixation point, a study list consisting of five consonants were presented to the participant. Then, different visual masks that cued different tasks were presented. Finally, two test probes were shown in both tasks and either JOR or IR judgment was performed.

In the local mesh model, *p-neighborhood* η_p of each voxel is defined spatially [8]. Each voxel $v(t_i, \bar{s}_j)$ is used as the seed voxel of the local mesh, and the p-nearest neighbors $\{v(t_i, \bar{s}_k)\}_{k=1}^{p}$ of this seed voxel are selected as the ones whose voxel coordinates has the smallest Euclidean distances to that of seed voxel. Hence, a local mesh consists of the seed voxel, and its p-nearest neighbors connected by a set of arcs in a star topology (Fig. 2).

The seed voxel is connected to its p-nearest neighbors with arc weights $a_{i,j,k}$ that represent the relationship between the seed voxel and its neighbors. The arc weights $a_{i,j,k}$ are estimated using the linear regression equation,

$$v(t_i, \bar{s}_j) = \sum_{\bar{s}_k \in \eta_p} a_{i,j,k} \, v(t_i, \bar{s}_k) + \varepsilon_{i,j,p} \, . \qquad (1)$$

In (1), $\varepsilon_{i,j,p}$ is the residual error obtained while estimating the arc weights $a_{i,j,k}$ of the local mesh at time instant t_i , where the seed voxel is $v(t_i, \bar{s}_j)$ and its p-nearest neighbors are $\{v(t_i, \bar{s}_k)\}_{k=1}^{p}$. By minimizing the squared error ${\varepsilon_{i,j,p}}^2$ acquired in (1), the arc weights $a_{i,j,k}$ are estimated for each local mesh of size *p*. Using these arc weights $a_{i,j,k}$, which represent the relationship between the voxel $v(t_i, \bar{s}_j)$ and its neighbors $\{v(t_i, \bar{s}_k)\}_{k=1}^{p}$, a mesh arc vector $\bar{a}_{i,j} = [a_{i,j,1} \; a_{i,j,2} \dots a_{i,j,p}]$ of size 1*xp* is formed. Note that, each voxel is, then, represented in terms of its relationships with its neighbors using this mesh arc vector $\bar{a}_{i,j}$ instead of its own fMRI intensity value $v(t_i, \bar{s}_j)$. For a voxel, all mesh arc vectors for all time instants are combined to form a Nxp mesh arc vector $A_j = [\bar{a}_{1,j} \; \bar{a}_{2,j} \dots \bar{a}_{N,j}]^T$. Finally all Nxp mesh arc vectors for all voxels are combined so that a $Nxp.M$ feature matrix $F = [A_1 \; A_2 \dots A_M]$ is constructed for a participant.

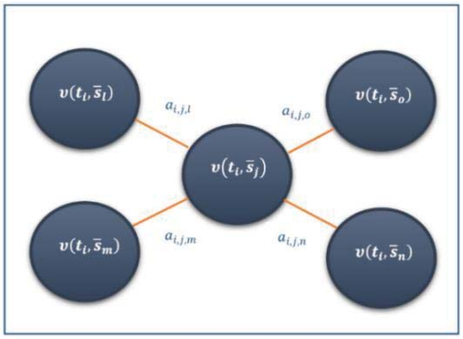

Fig. 2. A local mesh representing a seed voxel $v(t_i, \bar{s}_j)$ at the center and its p-nearest neighbors $\{v(t_i, \bar{s}_k)\}_{k=1}^{p}$ at time instant t_i. The relationship between the seed voxel $v(t_i, \bar{s}_j)$ and its neighbors are represented with the arc weights $a_{i,j,k}$ between them.

IV. Model order selection for estimating the degree of brain connectivity

The size of a local mesh, p represents the degree of connectivity for a voxel and its neighbors. Therefore, depending on the mesh size, relationships among voxels, in other words how the information is distributed in the brain, changes. Consequently, finding the optimal mesh size for each sample is a crucial step in the suggested local mesh model.

At this point, we assume that there is a trade-off between the mesh size which represents the model complexity and error variance which represents the degree of fit among the voxels in a neighborhood. Therefore, the degree of connectivity can be represented by optimizing this trade-off. In this study, three different information theoretic criteria, namely Akaike Information Criterion (AIC), Rissanen's Minimum Description Length Method (MDL) and Bayesian Information Criterion (BIC), are adopted to select the optimal mesh size. In our case, as the mesh size p increases, the complexity of local mesh model increases since each voxel is then represented as a linear combination of more number of its neighbors. On the other hand, as the mesh size increases expected value of squared error $E\left(\overline{\varepsilon_{\iota,p}}^2\right)$ decreases showing that the model fits better. From equations (4, 5 and 6), it can be seen that in all of the information criteria, expected value of squared error is found either as itself $E\left(\overline{\varepsilon_{\iota,p}}^2\right)$ (6) or in logarithmic function $\ln\left(E\left(\overline{\varepsilon_{i,p}}^2\right)\right)$ (4,5). Therefore, an increase in the mesh size decreases the $E\left(\overline{\varepsilon_{\iota,p}}^2\right)$ and $E\left(\overline{\varepsilon_{\iota,p}}^2\right)$ tends to decrease the information criteria. On the other hand, an increase in the mesh size tends to increase the information criteria. Therefore, for each equation (4, 5 and 6), there is a mesh size p that makes the information criterion minimum and that p is selected as the optimal mesh size.

First, the squared error of the linear regression equation at each time instant t_i for each voxel $v(t_i, \bar{s}_j)$ is computed from the following equation:

$$\varepsilon_{i,j,p}{}^2 = \left(v(t_i, \bar{s}_j) - \textstyle\sum_{\bar{s}_k \epsilon \eta_p(\bar{s}_j)} a_{i,j,k}\, v(t_i, \bar{s}_k)\right)^2 . \tag{2}$$

Then, to find the expected value of squared error for each sample, the average of all squared errors with respect to all voxels $v(t_i, \bar{s}_j)$ at time instant t_i is approximated with:

$$E\left(\overline{\varepsilon_{\iota,p}}^2\right) \cong \frac{1}{M}\textstyle\sum_{j=1}^{M} \varepsilon_{i,j,p}{}^2 , \tag{3}$$

where $E(.)$ is the expectation operator.

Finally, the expected value of squared error of Eq.(3), is used to optimize three different information criteria, which are AIC, BIC and MDL, with respect to the model order p, as described in the following section.

A. Akaike Information Criterion (AIC)

If the data generating process, in our case the generation of fMRI data during a cognitive process were known, information loss of the local mesh model of size p would be found using the Kullback–Leibler (KL) divergence between the model and the information distribution with certainty. Hence, the optimal mesh size would be selected as the one having the smallest KL divergence with the underlying cognitive process. However, the information distribution in the brain is unknown and we approximate this unknown by using AIC for a local mesh formed around each voxel. Therefore, we assume that the mesh size p which makes the AIC minimum is the one that best approximates the unknown information distribution and this p is selected as the optimal mesh size. Optimal mesh size around a voxel $v(t_i, \bar{s}_j)$ at each time instant t_i, is estimated using Akaike's Information Criterion [9] for each sample using the following equation

By taking the average of squared errors for all time instants t_i and for all seed voxels $v(t_i, \bar{s}_j)$, the expected error for the mesh size p is found using :

$$AIC_i(p) = \ln\left(E\left(\overline{\varepsilon_{i,p}}^2\right)\right) + \frac{2 \cdot p}{M} \tag{4}$$

where $E\left(\overline{\varepsilon_{\iota,p}}^2\right)$ is the expected value of error, p is the mesh size and M is the total number of active voxels.

B. Bayesian Information Criterion (BIC)

BIC attempts to estimate a true model among the candidates. In our case BIC is used to find the local mesh model of optimal mesh size among all the candidate local mesh models of size p. BIC answers how likely the data is generated by the local mesh model of size p by estimating the posterior probability and selecting the p as the optimal mesh size which gives the highest posterior probability. Unlike AIC, BIC uses prior probability, hence the prior selection affects the accuracy. To find the optimal mesh size with BIC, the following formula is adopted from [10],

$$BIC_i(p) = \ln\left(E\left(\overline{\varepsilon_{i,p}}^2\right)\right) + \frac{\ln(M) \cdot p}{M} \tag{5}$$

where $E\left(\overline{\varepsilon_{\iota,p}}^2\right)$ is the expected value of error, p is the mesh size and M is the number of voxels.

C. Rissanen's Minimum Description Length (MDL)

In this study MDL is used to find the local mesh model of size p that best represents the relationship among voxels. It assumes that, the best model, i.e. the local mesh model having the optimal mesh size, requires smallest description length. A local mesh model of size M, representing the relationship between a voxel and all other voxels would include redundant information. Moreover, it would cause high dimensionality problem. MDL is used to find how the information is represented with the minimum number of relationships among voxels without a high information loss. MDL, in [11] is adopted to represents the information among the voxels in a compressed manner as follows:

$$MDL_i(p) = E\left(\overline{\varepsilon_{i,p}}^2\right) \cdot \left(1 + \left(\frac{p+1}{M}\right) ln(M)\right) \qquad (6)$$

where $E\left(\overline{\varepsilon_{i,p}}^2\right)$ is the expected value of error, p is the mesh size and M is the number of voxels.

After an information criterion is evaluated for different mesh sizes for each sample, the mesh size p which minimizes one of the information criteria is selected as the optimal mesh size for that sample at t_i.

$$\hat{p}_{i,x}^{IC} = \underset{p}{\operatorname{argmin}}(IC_i(p), \forall p \in [2,100],\ t_i \in P_x), \qquad (7)$$

where $IC_i(p)$ is either $AIC_i(p)$ or $BIC_i(p)$ or $MDL_i(p)$ for sample at t_i belonging to participant P_x. Notice that, $\hat{p}_{i,x}^{IC}$ is optimal depending on the choice of a particular criterion.

Note that, optimal mesh size estimated for each sample, depends on three parameters; i) the minimized information criterion, ii) the participant to which the sample belongs and iii) cognitive task studied during the fMRI measurements. Suppose that the optimal mesh size for an information criterion IC and participant P_x is represented by a random variable $\hat{p}_{i,x}^{IC}$. Then, for each P_x, mean and standard deviation (std) of the optimal mesh size $\hat{p}_{i,x}^{IC}$ are approximated by,

$$\mu_x{}^{IC} \cong \frac{1}{N^{te}} \sum_{i=1}^{N^{te}} \hat{p}_{i,x}^{IC}, \qquad (8)$$

$$\sigma_x{}^{IC} \cong \sqrt{\frac{1}{N^{te}} \sum_{i=1}^{N^{te}} \left(\hat{p}_{i,x}^{IC} - \mu_x{}^{IC}\right)^2}, \qquad (9)$$

respectively. In the above approximation, N^{te} represents the number of test samples measured from participant P_x (N^{te} is the same for all participants. Hence, it is independent of x). Moreover, N_T^{te}, where T is either *IR* or *JOR,* represents the number of test samples belonging to task T measured from a participant. In the dataset, $N_{IR}^{te} = N_{JOR}^{te} = \frac{1}{2} N^{te}$ meaning that number of test samples belonging to IR and JOR are equal. Similarly, for each task T the mean and std of the optimum mesh size over all participants can be approximated by,

$$\mu_T{}^{IC} \cong \frac{1}{8N_T^{te}} \sum_{x=1}^{8} \sum_{\forall c_i = T} \hat{p}_{i,x}^{IC}, \qquad (10)$$

$$\sigma_T{}^{IC} \cong \sqrt{\frac{1}{8N_T^{te}} \sum_{x=1}^{8} \sum_{\forall c_i = T} \left(\hat{p}_{i,x}^{IC} - \mu_T{}^{IC}\right)^2}, \qquad (11)$$

where c_i is the label of sample at time instant t_i and $\forall c_i = T$ represents all samples belonging to task T. Recall that there are total of 8 participants.

Optimal mesh size selection for each information criterion is handled independently. Therefore, from each criterion, an optimal mesh size is acquired for each participant and for each cognitive task. This size is used in further classification tasks only related with that information criterion. Hence, there is no unique and common optimal mesh size for a sample estimated by minimizing different information criteria.

V. Results

Since optimal mesh size varies for each sample and for each person under each cognitive task, information criteria can be used to optimize it for each sample to improve classification accuracy. Moreover, a comparison of the classification accuracy across three information criteria can indicate the criterion that yields the best performance. We modeled neural activation recorded from 8 participants while they performed the IR and JOR tasks in a scanner.

Our region of interest (ROI) consisted of 2030 voxels that were identified from a whole-brain voxel-wise contrast assessing the active voxels during both tasks, with a threshold of $p < .001$, uncorrected. For the mesh size p, the dimension of feature vector can be represented as px2030. Therefore, as the mesh size increases, the dimension of feature vector increases linearly. Each participant has 240 training samples (120 for IR and 120 for JOR tasks) and 80 test samples (40 for IR and 40 for JOR tasks).

For each test sample, the optimal mesh size $\hat{p}_{i,x}^{IC}$ is estimated for each information criterion in the interval [2, 100]. This interval is taken large to assure that the minimum of the criterion is not local, instead global. Therefore, for all participants it is assured that the optimal mesh size lies in the interval [2,100]. Then, a spatially local mesh of size $\hat{p}_{i,x}^{IC}$ is formed around each training sample, so that each training sample will turn into a feature vector of arc weights with size $\hat{p}_{i,x}^{IC}$x2030. On the other hand, a spatially local mesh of size $\hat{p}_{i,x}^{IC}$ is also formed around the test sample at t_i. A classifier is

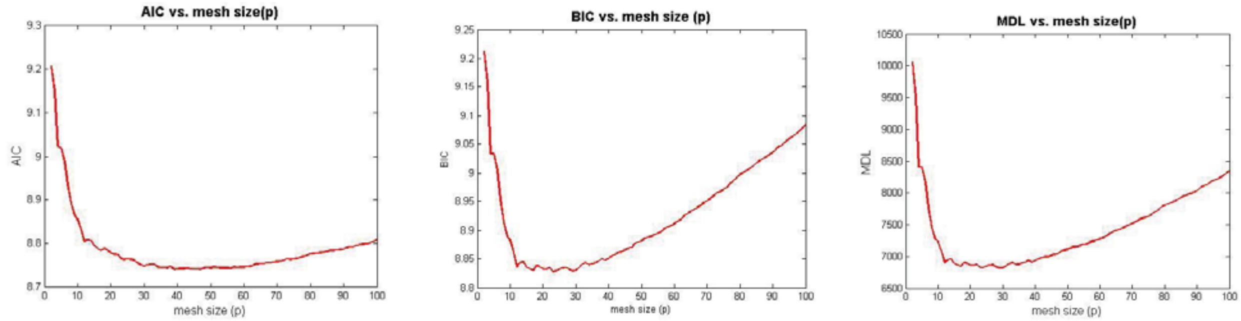

Fig. 3. Information criteria vs. mesh size plots for the same sample of the same participant indicating that the minimum of information criteria may be different for the same sample.

trained using the training samples of size $\hat{p}_{i,x}^{IC}$x2030 and whether the test sample belongs to IR or JOR is found using the classifier. Therefore, if another test sample at t_l has the optimal mesh size $\hat{p}_{l,x}^{IC}$, then local mesh of size $\hat{p}_{l,x}^{IC}$ is formed for each voxel in the training sample and the resulting feature vector will be used to train another classifier. Hence, if $\hat{p}_{i,x}^{IC} \neq \hat{p}_{l,x}^{IC}$ for test samples at t_i and t_l, then these test samples are classified using different classifiers which receive different sizes of feature vectors. For each participant, classification accuracies are found using k-NN method where the k value is found using cross validation in the training data.

Our data is composed of four runs, hence, we employed 4-fold cross validation. At each step, 3 runs are used as training data and the remaining run is used as test data. In Table I, average classification performances of 4-fold cross validation using k-NN are displayed across the 8 participants P_x, $x = 1,2 \ldots 8$. Performance results denoted in bold indicate the best performance for each participant. The last column of Table I presents average k-NN accuracy of MVPA method in which voxel intensity values are directly fed to the classifier. Note that, the information criterion giving the best performance changes from person to person. Among these information criteria, using MDL to select optimal mesh size for each sample is always better than or equal to MVPA method. Hence, results indicate that MDL can be used to select the optimal mesh size such that local mesh model having optimal mesh size can successfully represent the relationships among voxels. While AIC increases the performance using MVPA method or gives the same performance with MVPA for 7 participants, it fails to find the optimum order for participant 7 (P_7). Similarly, selecting BIC is better than or equal to MVPA for 7 participants but, it fails to optimize the result for participant 3 (P_3). However, the average performances acquired using three criteria over all participants are the same (58%) and on average, using each of these criteria gives better accuracy than using MVPA method.

Although the average accuracies are equal for each of these three criteria, the one that has the maximum performance changes from participant to participant. For example, BIC has the highest performance (57%) for P_7 while AIC has the lowest performance (54%). On the contrary, for P_1, AIC has the best performance among others (61%) where BIC has the worst performance (58%). Hence, we can state that brain connectivity, represented with a local mesh model having variable size, is different for each participant. As a result, the best information criterion to select optimal mesh size differs for participants. The performances might seem low for a two – class classification task. Yet, this situation is mainly caused by the design of the experiment, where the encoding phases of both IR and JOR are the same. Table II represents the standard deviations of classification performances among 4-fold. As it can be seen, for some participants standard deviation among 4-fold is high (e.g. using BIC for P_2). On the other hand using BIC for P_3 results in a low standard deviation meaning that the classification performance changes slightly based on the training and test data used.

TABLE I. AVERAGE CLASSIFICATION PERFORMANCES (K-NN) AMONG 4-FOLD USING THREE CRITERIA AIC, MDL AND BIC WITH LOCAL MESH AND USING MVPA METHOD AMONG 8 PARTICIPANTS

	Average k-NN Accuracies			
	AIC	**BIC**	**MDL**	**MVPA**
P1	**0.61**	0.58	0.58	0.58
P2	**0.59**	**0.59**	**0.59**	0.58
P3	**0.59**	0.55	**0.59**	0.59
P4	**0.60**	0.59	0.57	0.55
P5	0.56	**0.58**	**0.58**	0.55
P6	**0.58**	**0.58**	0.57	0.53
P7	0.54	**0.57**	**0.57**	0.55
P8	0.57	**0.58**	**0.58**	0.57
Average	0.58	0.58	0.58	0.56

TABLE II. STANDARD DEVIATION OF CLASSIFICATION PERFORMANCES AMONG 4-FOLD USING THREE CRITERIA AIC, MDL AND BIC WITH LOCAL MESH AND USING MVPA METHOD AMONG 8 PARTICIPANTS

	Standard Deviation of Performances among 4-fold			
	AIC	**BIC**	**MDL**	**MVPA**
P1	5.30	6.07	6.07	2.07
P2	2.60	7.77	6.25	2.58
P3	6.65	1.08	2.84	5.63
P4	3.84	7.23	5.35	3.43
P5	4.63	5.28	5.28	4.25
P6	5.61	5.67	4.06	4.13
P7	4.39	4.38	2.98	1.77
P8	6.50	2.39	2.39	4.84

Optimal mesh size for a sample $\hat{p}_{i,x}^{IC}$ is found using (7) and for three different criteria, three optimal mesh sizes are found. Figure 3 represents information criteria vs. mesh size plots for the same sample belonging to participant P_3 at t_i. For this sample, $\hat{p}_{i,3}^{AIC} = 47$, meaning that optimal mesh size for the sample estimated using AIC is 47. However, $\hat{p}_{i,3}^{MDL} = 29\,, and\ \hat{p}_{i,3}^{BIC} = 23$ since at these values related criterion becomes minimum (Fig. 3). On the other hand, for the same participant, optimal mesh size intervals may change with the criterion (Table III). For example, optimal mesh sizes for the samples belonging P_5 changes in the interval [23, 29] if AIC is used and if BIC is used this interval is changed to [20, 23]. Furthermore, if MDL is used, all samples belonging to P_5 have the optimal mesh size as 23. As a result, it can be stated that behavior of optimal mesh size for each sample changes with the information criterion used.

From Table III it can also be seen that mean ${\mu_x}^{IC}$ and standard deviations ${\sigma_x}^{IC}$ of optimal mesh size distributions vary for each participant. Moreover, for some criterion, std ${\sigma_T}^{IC}$ is 0, indicating that $\hat{p}_{i,x}^{IC}$ is same for all samples. On the other hand high standard deviation indicates a spread of optimal mesh sizes. Furthermore, it can be seen that mean ${\mu_T}^{IC}$ and standard deviations ${\sigma_T}^{IC}$ of optimal mesh size distributions do not differ from task to task (Table IV).

TABLE III. OPTIMAL MESH SIZE INTERVALS, MEAN (μ_x^{IC}) AND STD (σ_x^{IC}) OF THESE INTERVALS FOR 8 PARTICIPANTS ESTIMATED USING THREE CRITERIA AIC, MDL AND BIC

	AIC			MDL			BIC		
	Interval	$\boldsymbol{\mu_x^{AIC}}$	$\boldsymbol{\sigma_x{}^{AIC}}$	**Interval**	$\boldsymbol{\mu_x^{MDL}}$	$\boldsymbol{\sigma_x{}^{MDL}}$	**Interval**	$\boldsymbol{\mu_x^{BIC}}$	$\boldsymbol{\sigma_x{}^{BIC}}$
P_1	17-39	17.59	3.66	17-17	16	0	17-17	16	0
P_2	23-35	32.41	1.53	16-23	21.80	0.95	16-23	21.76	1.00
P_3	39-47	39.11	2.35	23-30	27.44	2.09	23-30	23.64	2.54
P_4	69-70	68.88	0.31	40-42	40.49	0.86	40-42	39.83	0.98
P_5	23-29	25.03	2.99	23-23	22	0	20-23	21.96	0.33
P_6	16-17	15.43	0.49	12-17	14.16	1.24	12-17	13.69	1.37
P_7	25-37	26.42	3.20	25-25	24	0	25-25	24	0
P_8	16-30	19.05	5.33	12-17	15.93	0.55	12-17	15.93	0.55

In this study, we do not propose "the best" criterion to be used in the classification. Rather the results indicate that, following an information theoretic approach to estimate the optimal mesh size of local mesh model and using the corresponding feature vectors in the classification, performs better than classical MVPA methods.

TABLE IV. MEAN (μ_T^{IC}) AND STD (σ_T^{IC}) OF OPTIMAL MESH SIZE INTERVALS FOR 2 DIFFERENT TASKS FOUND USING THREE CRITERIA AIC, MDL AND BIC

	AIC		MDL		BIC	
	$\boldsymbol{\mu_T^{AIC}}$	$\boldsymbol{\sigma_T{}^{AIC}}$	$\boldsymbol{\mu_T^{MDL}}$	$\boldsymbol{\sigma_T{}^{MDL}}$	$\boldsymbol{\mu_T^{BIC}}$	$\boldsymbol{\sigma_T{}^{BIC}}$
IR	30.56	16.48	22.71	8.00	22.12	7.74
JOR	30.39	16.57	22.73	8.00	22.07	7.68

VI. CONCLUSION

In this paper, we propose a model for distinguishing cognitive states based on distributed patterns of neural activation in the brain. In this model, each voxel is represented with its relationships among its neighbor voxels in a local mesh. Then for each class, the optimal mesh sizes are found using three different information criteria. Since the optimal mesh size greatly varies from sample to sample even belonging to same participant, one aim was to develop a method that determines the optimal mesh size for each sample. In addition, we provided a performance comparison of three information theoretic methods to determine the optimal mesh size. Results indicated that all the Information Criteria can successfully estimate the optimal mesh size and improve the overall classification performances on the average. However, MDL always beets the performance of the classical MVPA method for all participants. We also showed that information criteria can be used to select optimal mesh size for each sample instead of using it for each participant as in [13].

In the future studies, local mesh of varying size may be formed around each voxel during the same cognitive process using the information criteria. Moreover, to achieve a more generic success, presented method will be implemented for different cognitive tasks. In the future, this method will be employed to select the optimal mesh size in local meshes formed with functionally nearest neighbors, using functional connectivity matrices [14].

REFERENCES

[1] J.-D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nature reviews. Neuroscience*, vol. 7, no. 7, pp. 523-34, Jul. 2006

[2] J.-D. Haynes, "Decoding and predicting intentions," *Annals of the New York Academy of Sciences*, 1224: 9–2, 2011

[3] N. Kriegeskorte, "Pattern-information analysis: from stimulus decoding to computational-model testing," *Neuroimage*, 56:411–421, 2011

[4] N. Kriegeskorte, R. Goebel and P. Bandettini, "Information-based functional brain mapping," *Proceedings of the National Academy of Sciences of the United States of America*, 103: 3863–3868, 2006

[5] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, 293: 2425–2430, 2001.

[6] D. D. Cox, and R. L. Savoy. "Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex," *Neuroimage*, 19: 261–270, 2003

[7] A. J. O'Toole, F. Jiang, H. Abdi, N. Penard, J. P. Dunlop and M. A. Parent, "Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data," *Journal of Cognitive Neuroscience,* 19: 1735–1752, 2007.

[8] M. Ozay, I. Oztekin, U. Oztekin and F. T. Yarman Vural, and "A Mesh learning model for pattern analysis of brain activity" , *Human Brain Mapping,* under review.

[9] H. Akaike, "A new look at statistical model identification," *IEEE Trans. Automat. Contr.*, vol. 19, pp. 716–723, Dec. 1974

[10] G. Schwarz, Estimating the Dimension of a Model, *The Annals of Statistics,* vol.6, no.2, pp.461-464, 1978.

[11] J. Rissanen, "Universal coding, information prediction and estimation, " *IEEE Trans. Inf. Theory*, vol.30, 629 - 636, 1984.

[12] I. Oztekin, B. McElree, B. P. Staresina and L. Davachi, "Working Memory Retrieval: Contributions of the Left Prefrontal Cortex, the Left Posterior Parietal Cortex, and the Hippocampus," *Journal of Cognitive Neuroscience* vol. 21 no. 3, pp. 581-593

[13] I. Onal, M. Ozay, O. Firat, I. Oztekin and F. T. Yarman Vural, "Analyzing the Information Distribution in the fMRI measurements by estimating the degree of locality," *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2013.

[14] O. Firat, M. Ozay, I. Onal, I. Oztekin and F. T. Yarman Vural, "Functional Mesh Learning for Pattern Analysis of Cognitive Processes," *12th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC),* 2013