(T4-fMRI)Using fMRI to Diagnose Schizophrenia

Department of Computing Science
University of Alberta

Department of Computing Science
University of Alberta

Shashindra Silva
Department of Electrical and Computer Engineering
University of Alberta
jayamuni@ualberta.ca

Abstract

Diagnosing schizophrenia is a challenging task for which diagnostic tests have yet to be developed, but where data is becoming increasingly more available due to the use of Functional Magnetic Resonance Imaging (fMRI). Studies have shown that fMRI can be used in conjunction with Sparse Gaussian Markov Random Field (SGMRF) to achieve high accuracy when diagnosing schizophrenia, however, these studies are limited by fairly small dataset that origninate from single locations. In this work we focus on obtaining high diagnostic accuracies for data that has been combined from multiple locations. We evaluated the performance of two main approaches. First we consider the use of SGMRF on voxel degrees extracted from fMRI data, and second we use Regions of Interest (ROI) as defined by Power *et al.* extracted from the data, and used a the feature set for the classifiers. After averaging the accuracies obtained from 5 different 20% balanced hold-out sets, where each set was sampled from each site with the same proportion. We found that an ROI approach used in conjunction with a SGMRF classifier gave us the best accuracy of 74.16% on our multi-site data. While these accuracies are low in comparison to those obtain from single site analysis, we show that reasonable accuracies can be obtain when combining multiple schizophrenia datasets.

1 Introduction

Schizophrenia is a mental/psychiatric disorder [16, 9] known to affect blood flow in the brain [9] where those who are affected can experience hallucinations, delusions and diminished mental capacities to varying extents [8]. While several features of schizophrenia have proven useful for its diagnosis there are currently no set of features that have sufficient sensitivity or specificity to be used in diagnostic tests [8, 12]. This effectively means that subjectivity plays a role when a physician is diagnosing a patient.

Functional Magnetic Resonance Imaging (fMRI) is a tool for mapping of the functioning human brain caused by underlying brain electrical activity[11]. When a person is doing a task, neuron activity fluctuates and in order to provide the energy needed for this activity, the blood flow increases to supply the neurons with the needed glucose and oxygen[11]. More blood flow also brings more oxygen through blood vessels. This change in the level of oxygenated blood known as oxyhemoglobin and deoxyhemoglobin (oxygenated or deoxygenated blood) changes the magnetic susceptibility of blood (BOLD signal) which is detectable through fMRI [11].

fMRI plays an important role in modern psychiatry research, as it provides the needed tool to determine dissimilarities in brain system that is a root for psychiatric illness[22]. An advantage of fMRI

in medical diagnosis is that it is non-invasive. This means that unlike some other imaging methods, no instruments or dyes are placed in the patients body.[11][6].

One of the approaches that has been used for studying schizophrenia is Sparse Gaussian Markov Random Field(SGMRF) [16][19]. The primary advantage of using this method is that the functional network of the brain can be captured using the precision matrix [16]. By using the resulting network, healthy subjects can be differentiated from schizophrenic ones, by observing differences in the functional connectivity of the brain. Currently automated approaches to schizophrenia diagnosis have been able to yield accuracies of 93% for data that originates from a single location [16] and up to approximately 80% for data that originates from multiple locations [3].

In this work we consider probablistic graphical models (PGMs) as a means of capturing differences in the interconnections between the brains of healthy patients and the brains of schizophrenics. We compare several different methods including those that use voxel degrees, regions of interest, markov network structure and PCA as approaches to feature extraction.

2 Background and Prior Work

2.1 Regions of Interest and Single-Voxel Analysis

There are two main approaches for extracting information from fMRI images. The first is a single-voxel approach and the second is to study regions of interest (ROI) [7]. The trade-off between these two approaches is that a single-voxel approach requires the analysis of every voxel and is subject to the low signal-to-noise ratios of individual voxels, whereas a region based approach is only effective if the selected regions capture all relevant information in an fMRI task [7]. In 2011, Power *et al.* identified 264 putative function regions of interest derived from resting state fMRI, where no specific task being performed during data collection [14]. JD Power argues in his video abstract that these regions are currently the best representation of functional networks in the brain that are available [14]. Several of the approaches used in this study directly build on top of these regions.

2.1.1 Calculating Voxel Degrees

Introduced as a "degree map" in Rish *et al.* [18], voxel degrees are the constituent components of degree maps and is a feature of fMRI data that has previously been successfully used as part of schizophrenia diagnosis [16]. Voxel degrees represent the connectedness of voxels in the brain with the other voxels and are described as "the number of voxel neighbours in a network" [16]. Degrees are calculated in the following steps: First, multiple pair-wise Pearson correlation comparisons between all voxels to create a correlation matrix. Second, a threshold is applied to the correlation matrix. If a correlation value is greater than the threshold there said to be an edge between those two voxels. Finally, for each voxels all of the edges it participates in are summed and the total number of edges becomes the degree of that voxel. Voxel degrees are relevant to our study because they are used in several of our approaches.

2.2 Rish *et al.*

Rish *et al.* [16] was able to achieve very high accuracies when diagnosing patients as schizophrenic using fMRI data from a single location and employing several approaches. The most successful of these approaches include the use of support vector machines (SVM)l and sparse gaussian markov random field (SGMRF) classifiers, described later in the background. The error rate for SVM was 7% using pairwise correlation features, and 14% for the SGMRF using degree 100 most significant voxels in the degree map feature set. As part of our work we use the SGMRF approach detailed in their paper. In the MRF classifier, first they split training data in two different group. For each group they compute the precision matrix (assuming the data is centered unless μ needs to be computed as well). See the appendix for the *Inference* section for how they classify each new subject based parameters obtained from training stage.

Figure 1.a shows interaction between voxels BOLD signal and a stimulus, and Figure 1.b shows a Bayesian network of brain voxel being affected by both stimulus and unobserved brain processes that can change BOLD signal too. Given this figure one can see MRF approach invests on the

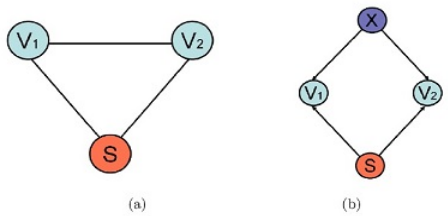

Figure 1: Graphical models of voxel interactions Rish *et al.*

different network on of interaction for healthy versus schizophrenic subject in response to a stimulus to differentiate them.

2.3 Rosa *et al.*

Rosa *et al.* [19] work is on detecting patients with major depressive disorder(MDD). Their study was done using fMRI datasets and applying SGMRF method for obtaining the precision matrix. Then they go and use SVM on the obtained precision matrix for doing the classification task. They obtained 85% accuracy on their first dataset and 78.95% on their second dataset using this method. Both dataset was obtained using a single machine in one place. Their first dataset was consisting of nineteen patients and nineteen controls, and their second dataset was consisting of thirty patients and thirty controls. They also used SVM on the correlation feature obtained using Pearson correlation to show effectiveness of their approach. Which in the case of correlation results were not better than chance. Their approach can be compared to our approach in the *Individual MRF structure classification section*.

2.4 Multi-site Comparisons

Although, there were several prior work on the classification of schizophrenia patients using fMRI, most of them are based on data from a single site. Classification from multi-site data is inherently difficult due to the batch effects resulted from the use of different machines and environments. At the same time, multi-site analysis can easily be generalized for a new data set from a totally different source. Cheng et al. [3] analyzes a multi site data set which is obtained from five different sites with different machines. They used SVM for classify schizophrenia patients and healthy controls and obtained accuracies in the range of $73.53 - 80.92\%$. In this research we will try to obtain results with similar accuracies, but using probabilistic graphical methods.

2.5 Principal Component Analysis

Principal Component Analysis or as it called PCA is a technique which uses complicated mathematic techniques to transform the data from original space to a new dimensionality called principal components, which usually has smaller dimension than the original space. In general PCA uses vector space transformation to increase the dimensionality of large datasets. For doing so PCA uses mathematical projection. [15]

The way PCA does this projection is by maximizing the variance of projected data. As it been shown in the [1] if we project the data on the eigenvector of the data variance will be maximum. So if we want to project our N-dimensional data to the space M-dimension where $M < N$, *"the optimal linear projection for which the variance of the projected data is maximized is now defined by the M eigenvectors $u_1, ..., u_M$ of the data covariance matrix S corresponding to the M largest eigenvalues $\lambda_1, ..., \lambda_M$."* [1]: S and μ (the sample set mean) are defined as:

$$S = \frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)(X_n - \mu)^T \qquad (1) \qquad\qquad \mu = \frac{1}{N}\sum_{n=1}^{N} X_n \qquad (2)$$

2.6 Support Vector Machines

In a 1995 paper by Vladimir Vapnik *et al.* the concept of support Vector Machines (SVMs) was introduced as a statistical tool for classification problems [20]. In our work we use the most simple SVM, a linear SVM, because it does not use non-linear kernels and therefore has no hyperparameters that need to be tuned with cross validation. Instead, a linear SVM learns the "maximum-margin hyperplane" classifier which is a linear combination of the input features and partitions the data space into separate classes. The term "maximum-margin" refers to the SVM's ability to find the maximal separation between classes and the hyperplane, therefore creating the largest "margin" [20]. We can be guaranteed of the optimality of the result as the SVM problem is known to be convex [2]. The simplest form of a linear SVM minimizes $\frac{1}{2}||w||^2$ subject to the constraint $y_i(x_i w + b) - 1 \geq 0, \forall i$ where w is the weight vector, x_i is the instance's features, y_i is the instance label and b is the bias term of the model. Unfortunately, this version of the SVM only works in the case where the data is linearly separable. To extend the SVM to linearly inseparable data, the addition of positive slack variables is required, such that the constraints become $\forall i \; y_i(x_i w + b) - 1 + \xi_i \geq 0$ and $\xi_i \geq 0$, where xi_i is the slack variable for the instance i [2].

2.7 SGMRF

One variation of Markov Random Field is Gaussian Random which is generally used for continuous space of variables and has well-defined mathematic properties that can be computed. Multivariate Gaussian density function over a set of random variables X is defined as below:

$$p(X) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu)\right\}. \tag{3}$$

In the equation μ is the mean and Σ is the covariance matrix. We can set μ to zero and replace Σ^{-1} with C the equation (3) which can be written as the following form. It also should be noted that $C = \Sigma^{-1}$ and is known as the precision matrix in the literature.

$$p(X) = (2\pi)^{-n/2}|C|^{1/2} \exp\left\{-\frac{1}{2}X^t C X\right\}. \tag{4}$$

It has been shown by [10] that X_i and X_j are independent if and only if their corresponding entries in the precision matrix (C) are zero. It can be concluded that missing edge in the MRF will lead to zero entries in the precision matrix[17]. The Problem of learning probabilistic graphical model for a given dataset is reduced to learning the precision matrix given this proof.

The Log-likelihood of the dataset, assuming each row of the data is a p-dimensional vector and consisting n samples $D = \{X_1, X_2, ..., X_n\}$, can be written as follow. It should also be noted that we assume each sample is identically independent distributed(iid).

$$L(D) = \frac{n}{2}|C| - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^T C (X_i - \mu) + const, \tag{5}$$

Here, const is a constant which is not dependant on μ or C. We also can center the data in such a way that $\mu = 0$, so the second term reduces to $\frac{1}{2}\sum_{i=1}^{n} X_i^T C X_i$ which is equal to $\frac{n}{2}$ tr(AC), where tr is trace of the matrix. The above formula will can be written as shown in equation (6) where the log-likelihood maximization problem is shown in equation (7).

$$L(D) = \frac{n}{2}[|C| - |tr(AC)|] + const \tag{6}$$

$$\max_{C \succ 0} |C| - tr(AC) \tag{7}$$

Here $|...|$ represents determinant and A is the empirical covariance matrix calculated by $A = \frac{n}{2}\sum_{i=1}^{n} X_i^T X_i$, or maximum likelihood estimation fo Σ. It should also be noted the $C \succ 0$

constraint makes C positive definite.

To achieve sparse formulation for Gaussian Markov Random Field we can use the Laplace prior. If we impose the Laplace prior on the precision matrix $p(C_{ij}) = \frac{\lambda_{ij}}{2} e^{-\lambda_{ij}|C_{ij}|}$ we can achieve sparsity. This will change the equations (6) and (7) to the following forms below where the l_1-norm of C is $||C||_1 = \sum_{ij} C_{ij}$, and $\rho = \frac{n}{2}\lambda$.

$$L(D) = \frac{n}{2}[ln|C| - |tr(AC)|] - \lambda||C||_1, \quad (8) \qquad \max_{C \succ 0} ln|C| - tr(AC) - \rho||C||_1, \quad (9)$$

2.7.1 Method for obtaining precision matrix

One obvious way for learning the parameters for joint distribution probability is using regularized likelihood maximization such as AIC and BIC. However, finding the simplest model which fits the data is a NP-hard problem using this approach. There are also few limitations that make using this approach less desirable. First Empirical covariance matrix may not even exist, specially when the number of features in the dataset is more than the number of samples. Which is the case in fMRI studies. fMRI datasets usually have fraction of samples to features. Second problem using this approach is not having zero elements in the inverse of empirical covariance matrix. Hence, to construct the MRF using this matrix one should include explicit sparsity constraint.[17]

These two major problems cause us to search for other solution to build the precision matrix given the dataset. Fortunately one can use the alternative approximation approach for the above problem. Recently there have been some successes for getting the approximate precision matrix given the dataset. Glasso[5], block-coordinate descend(BCD) known as COVSEL and projected gradient approach are among them. For further reading on these methods readers can refer to [17]. In this study we are using Varsel and Glasso as the preferred method for obtaining the precision matrix and subsequently MRF.

3 Methodology

3.1 Data Set

Data used for this study were downloaded from the Function BIRN Data Repository (http://fbirnbdr.birncommunity.org:8080/BDR/). The original data contained nine sites and 235 subjects. However, during the preprocessing steps some of the subjects were removed because they could not be preprocessed properly, resulting in a final set of 95 subjects and five sites. Data were preprocessed by Dr. Mina Gheiratmand by using FSL software (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). Our data contained 46 schizophrenic subjects and 49 healthy subjects. Each subject had four runs (repeated scans) so effectively our data set had 380 subjects. Each run had 137 time slices and each time slice had the signal amplitude of over $100,000$ voxels. The voxels were referred using the 3D coordinates and thus the dimensionality of a single run was $\mathcal{R}^{N_1 \times N_2 \times N_3 7 \times 137}$, wh We removed 80% of subjects from each site as our holdout set and used the remaining set as our training set. Furthermore, we made sure that the holdout set was also balanced, so that the ratio between the patients and healthy subjects was same in the holdout and training sets for each site. We made sure that all the runs from the same subject either belonged to the holdout set or to the training set. The training set was used for finding the hyperparameters of the system using five-hold cross validations. After finding the hyper parameters we used the full training set to train the system with the selected hyper parameter and obtained the accuracy of the hold out set. We repeated this procedure five times for different hold out sets to get an average accuracy.

3.2 SGMRF with log degrees of ranked voxels

First, we executed the approach proposed by Rish *et al.* [16] on our data set. We had degrees of each voxels as well as the smoothed log values of the voxels for each subject. Furthermore, as

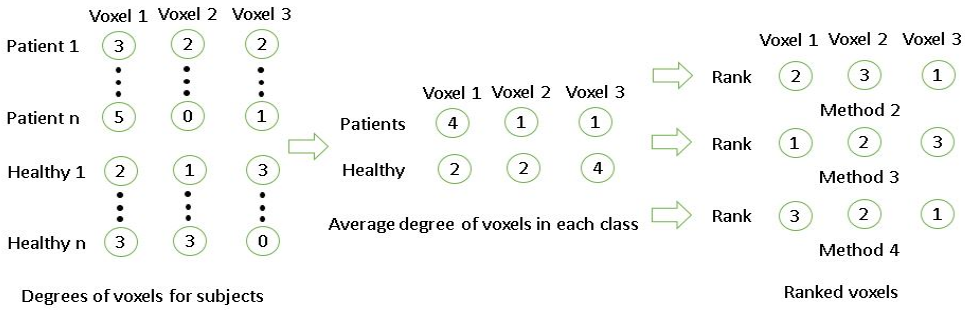

Figure 2: Voxel ranking methods to select best voxels. In step 2, the differences of the average degrees between patients and healthy subjects are 2, -1, and -3 respectively for the three voxels.

some voxels of some subjects had zero values for the BOLD signal, an universal mask was used to select the voxels which had a non-zero BOLD signal value through the whole data set. This resulted in a 28719 voxels per each subject. However, using all the available voxels resulted in significantly increased computation time and thus selecting the best voxels (based on the training data) to separate the healthy and schizophrenic subjects was very important.

To select the best voxels out of the possible 28719 voxels, we experimented with multiple approaches. These included selecting voxels based on t-tests, based on absolute differences of mean voxel degrees, based on differences of mean voxel degrees ($schizophrenic - healthy$) and based on differences of mean voxel degrees ($healthy - schizophrenic$).

The last three voxel ranking methods are explained in Figure 2. Out of the above four approaches we obtained the highest accuracy on our cross validation set by the third approach. Also the number of voxels k to be used in the SGMRF model was decided by running the experiment for different k values and finding that $k = 20$ performed well during cross validation. By using the selected voxels, we obtained the precision matrices for schizophrenic and healthy subjects. Similar to the value of k, the sparsity coefficient λ was also obtained through a hyper parameter search on our cross validation set.

3.3 SGMRF with degrees from Dorsolateral Prefrontal Cortex area

Connections in certain areas of the brain specifically dorsolateral prefrontal cortex (DLPFC) [13] and thalamocortical [3] circuitry. Thus, we extracted the degrees of voxels in these areas and used them to build the SGMRF model. Since we had only 4504 voxels for each subject, we did not perform any voxel rankings in this approach. Similar to the previous approach using the smoothed log values of the degrees resulted in higher accuracy.

3.4 Methods using Power *et al.*'s ROIs

The following three methods used in this subsection all involve the 264 regions of interest that are described in work by Power *et al.*. Due to missing data resulting from incomplete scans of subjects, 8 regions were removed from analysis leaving 256 ROIs. See appendix for list of omitted region coordinates. We use the approach described by Vega *et al.* [21] to summarize regions of interest by taking region averages for each of the $5mm$ radius spheres that define regions. Finally this resulted in a 137×256 matrix for each subject where each row is a time point across all regions and each column is the average time series for a single region. The permitted SGMRF hyperparameter λ values were 0.7, 0.1, 0.07, 0.01 and 0.007 for this section.

3.4.1 ROI with Subject Concatenation

For this approach we follow the method described by Vega *et al.* [21]. Training set subjects in each class are first concatenated so that two large matrices are created, one with dimension $ns * 137 \times 256$ and the other with dimension $nh * 137 \times 256$ where ns is the number of schizophenic subjects in the

training set and nh is the number of health subjects. To make learning the SGMRF structure faster we also used the same feature normalization as was described by Vega *et al.* [21], where for each class and each region in that class the region mean was subtracted from each time point and each time point was divided by the region standard deviation. Next, with the use of Glasso, a SGMRF structure was learned for each class which resulted in two sparse precision matrices that encode the independences learned. Finally, when presented with a new subject from the hold-out set, equations (8) and (9) were used to determine the likelihood of the subject belonging to each class followed by which class the new subject should be assigned.

A modified version of this approach was also implemented and tested but has been omitted for brevity and due to it obtaining poor results.See appendix for background information. In this variant, a Fourier transform was used on each of the subject's time series data to obtain Fourier coefficients. These Fourier coefficients were then used in place of the original time series for learning a classifier.

3.4.2 Region Degrees

Like the work described in Rish *et al.* [18], region degrees are also used in this approach but in this case, we only considered the degrees that result from the 256 ROIs. This process is illustrated in Figure 3. The only difference in the process described previously is that we explicitly remove the self correlations in the correlation matrix (the diagonal), and we conceptually represent the process as binarizing a matrix and summing over its rows. We also chose to use a threshold value of 0.7 as was done in Rish *et al.* [18].

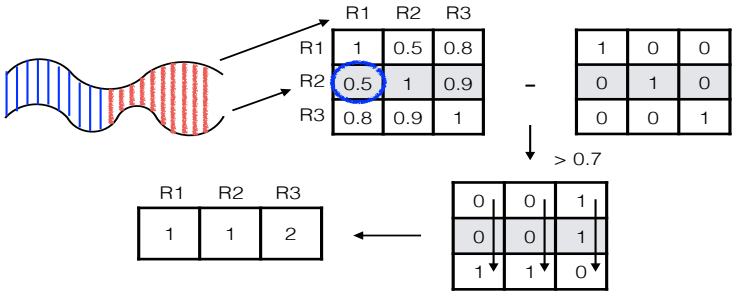

Figure 3: Pair-wise comparisons are made between region time series data to produce a correlation matrix. The diagonal is removed and a threshold is applied to binarize the matrix. Sums are collected for each region to produce the "degree" of that region.

Now that we had a vector of region degrees (1×256) for each subject, a SGMRF was used to again to build a classifier. This followed similarly from the previous approach except that now the input to Glasso was a $ns \times 256$ matrix for schizophrenic patients and a $nh \times 256$ matrix for healthy subjects. Again this resulted in two 256×256 precision matrices which we could use in likelihood calculations for subjects from the hold-out set. Additionally, we also used a linear SVM classifier trained directly on region degree data to compare its performance to the SGMRF classifier.

3.4.3 Individual MRF Structure Classification

This approach is similar to the region concatenation approach described previously except that here we did not concatenate subjects and instead learned a precision matrix (SGMRF structure) for each subject individually. For example, if we had ns_{total} schizophrenics in our dataset and nh_{total} healthy subjects ($ns_{total} + nh_{total}$ 137×256 matrices) then we would use Glasso to generate $ns + nh$ precision matrices.

To build a classifier, a linear SVM is trained on the precision matrices generated from the training set and then tested on the precision matrices generated from the hold-out set.

3.5 PCA and some other features

We have tried different features using the approach proposed by Rish *et al.*. Features that are used for this approach are as following: [Correlation , Log-Disconnection, Log-Disconnection Regions,

Eignevalues, MBI stat, Degree]. This features were provided to us with the dataset. So we wanted to see how good results would be implementing the exact same approach as suggested by Rish *et al.*. In this experiment we first selected the hyper parameters; number of features and λ; using 5-fold cross validation and then test those parameters on the hold-out set. The hold-out set has been selected randomly from different site with criteria of being balanced with regards to healthy and schizophrenic, and according to the ratio of each group's subjects in each site. Result that we achieved had accuracy of less than 50% for all features except than degrees and disconnection which achieved 65.55% and 63.33% accuracy respectively.

Another approach that we have used for feature selection was using PCA for transforming the dataset to lower dimension. We hoped this could magnify distinguishing features in the dataset while reducing it size. Unfortunately results were not as we hoped and were mostly near the chance level. It should be mentioned that we chose three different sizes for the number of principal components that we used :3, 35 and 70.

4 Results and Discussion

	Rish (Full)	Rish (DLPFC)	ROI + Concatenation	Region Degrees	MRF structure	Other Features
Classifier	SGMRF	SGMRF	SGMRF	SVM	SVM	SGMRF
Accuracy	72.32%	63.52%	74.17%	60.65%	69.44%	65.55%

Table 1: Accuracy on 5 Striated Hold-out Sets

As Seen in Table 1 methods using a SGMRF classifier were able to obtain the best results, of which 74.17% was the highest average accuracy recorded. This seems to suggest that a SGMRF seems to be a more effective tool for this particular problem. Because the SGMRF is a generative tool whereas the SVM is discriminative, it may be the case that the SVM is simply unable to capture the differences in network structure between patients and controls and it is these network differences that define schizophrenia as a disease. When comparing the ROI approach with patient concatenation and the Rish approach using all voxels against the Rish approach that only considers the DLPFC regions we see the former two get much better results. This seems to indicate that the DLPFC region is not sufficient to capture the difference between patients and controls. Another noteworthy difference is that of the differences between the approach creating degree features from the region averages and the other better performing methods. This discrepancy may be due to (Not sure

5 Conclusions and Future Work

In this work we studied the use of fMRI and machine learning to diagnose schizophrenia, given a set of healthy and schizophrenic subjects. We found that a region based approach using SGMRFs yielded the highest accuracy (74.16%) when averaged over 5 hold-out sets where the hold-out set was composed of 20% of each site used in the study. We conclude from this that ROIs are more effective in the diagnosis of schizophrenia then whole brain analysis because extranious noise is being removed from the data.

In this study we did not select the number of principal components based on the variance captured by the components, future work could be done to explore this method and determine if it could improve classification accuracy. Other future work involves the use of ensemble methods. We have found that several methods perform well individually, so an attempt at combining them together seems like a natural next step.

6 Acknowledgements

We would like to thank Dr. Mina Gheiratmand for preprocessing the fMRI data and co-coaching our project, Sugai Liang for her advice and Roberto Vega for all his advice and help. Finally, we would also like to thank Dr. Irina Rish for the feature extraction code she provided and Dr. Greiner for his supervision of this project.

References

[1] Christopher M Bishop. Pattern recognition. *Machine Learning*, 2006.

[2] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[3] Wei Cheng, Lena Palaniyappan, Mingli Li, Keith M Kendrick, Jie Zhang, Qiang Luo, Zening Liu, Rongjun Yu, Wei Deng, Qiang Wang, Xiaohong Ma, Wanjun Guo, Susan Francis, Peter Liddle, Andrew R Mayer, Gunter Schumann, Tao Li, and Jianfeng Feng. Voxel-based, brain-wide association study of aberrant functional connectivity in schizophrenia implicates thalamocortical circuitry. *Npj Schizophrenia*, 1, May 2015.

[4] Pierre Duhamel and Martin Vetterli. Fast fourier transforms: a tutorial review and a state of the art. *Signal processing*, 19(4):259–299, 1990.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Graphical lasso in r and matlab.

[6] Amiram Grinvald, Hamutal Slovin, and Ivo Vanzetta. Non-invasive visualization of cortical columns by fMRI. *Nature Neuroscience*, 3(2):105–107, feb 2000.

[7] Ruth Heller, Damian Stanley, Daniel Yekutieli, Nava Rubin, and Yoav Benjamini. Cluster-based analysis of fmri data. *NeuroImage*, 33(2):599–608, 2006.

[8] Assen Jablensky. The diagnostic concept of schizophrenia: its history, evolution, and future prospects. *Dialogues Clin Neurosci*, 12(3):271–287, 2010.

[9] Nomura Kenji. Pressure-induced performance decrement in verbal fluency task through prefrontal overactivation: A near-infrared spectroscopy study. *Front. Neurosci.*, 4, 2010.

[10] S.L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996.

[11] PM Matthews and P Jezzard. Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(1):6–12, 2004.

[12] Philip McGuire, Oliver D Howes, James Stone, and Paolo Fusar-Poli. Functional neuroimaging in schizophrenia: diagnosis and drug discovery. *Trends in Pharmacological Sciences*, 29(2):91 – 98, 2008.

[13] S G Potkin, J A Turner, G G Brown, G McCarthy, D N Greve, G H Glover, D S Manoach, A Belger, M Diaz, C G Wible, J M Ford, D H Mathalon, R Gollub, J Lauriello, D O'Leary, T G M van Erp, A W Toga, A Preda, K O Lim, and FBIRN. Working memory and DLPFC inefficiency in schizophrenia: The FBIRN study. *Schizophrenia Bulletin*, 35(1):19–31, November 2008.

[14] Jonathan D. Power, Alexander L. Cohen, Steven M. Nelson, Gagan S. Wig, Kelly Anne Barnes, Jessica A. Church, Alecia C. Vogel, Timothy O. Laumann, Fran M. Miezin, Bradley L. Schlaggar, and Steven E. Petersen. Functional network organization of the human brain. *Neuron*, 72(4):665–678, nov 2011.

[15] Mark Richardson. Principal component analysis. *URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si*, 2009.

[16] Irina Rish, Guillermo Cecchi, Benjamin Thyreau, Bertrand Thirion, Marion Plaze, Marie Laure Paillere-Martinot, Catherine Martelli, Jean-Luc Martinot, and Jean-Baptiste Poline. Schizophrenia as a network disease: Disruption of emergent brain function in patients with auditory hallucinations. *PLoS ONE*, 8(1):e50625, jan 2013.

[17] Irina Rish and Genady Grabarnik. *Sparse Modeling: Theory, Algorithms, and Applications*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 2014.

[18] Irina Rish, Benjamin Thyreau, Bertrand Thirion, Marion Plaze, Marie-laure Paillere-martinot, Catherine Martelli, Jean-luc Martinot, Jean-Baptiste Poline, and Guillermo A Cecchi. Discriminative network models of schizophrenia. In *Advances in Neural Information Processing Systems*, pages 252–260, 2009.

[19] Maria J. Rosa, Liana Portugal, John Shawe-Taylor, and Janaina Mourao-Miranda. Sparse network-based models for patient classification using fMRI. In *2013 International Workshop on Pattern Recognition in Neuroimaging*. Institute of Electrical & Electronics Engineers (IEEE), jun 2013.

[20] Armin Shmilovici. Support vector machines. In *Data Mining and Knowledge Discovery Handbook*, pages 257–276. Springer, 2005.

[21] Roberto Vega, Khare Kriti, and Sayem Mohammad Siam. Gender and age group classification using functional magnetic resonance imaging and gaussian markov random fields. *Unpublished*, 2015.

[22] Xiaoyan Zhan and Rongjun Yu. A window into the brain: Advances in psychiatric fMRI. *BioMed Research International*, 2015:1–12, 2015.

7 Appendix

7.1 Fourier Transform

A Fourier transform allows for the translation of any signal from the time domain to the frequency domain. More specifically, it takes a signal and decomposes it into a series constituent $\sin$ and $\cos$ componets. When taking the Fast Fourier Transform or FFT of real data the resulting peaks in the frequency domain are conjugate symmetric [4], meaning that methods that only require the real component of the data need only work with half of the resulting component coefficients in the transform. These Fourier coefficients can be used in place of the original signal as features provided to a machine learning classifier.

7.2 Omitted ROI coordinates

The regions removed are characterized by region number with the associated MNI coordinates described in Power *et al.* provided in parenthesis: 81 (-44, 12, -34), 82 (46, 16, -30), 128 (52, 7, -30), 184 (17, -80, -34), 247 (33, 12, -34), 248 (-31, -10, -36), 249 (49, -3, -38) and 250 (-50, -7, -39).

7.3 Inference

Once we have constructed probabilistic graphical model through precision matrix we can use that to do probabilistic inference for variables of interest. To make it more clear let assume X is our dataset and $Z \subset X$ is the subset of our dataset which is observed with assigned values of z, and let $Y \subseteq X - Z$ be the set of unobserved variables. Now we can use inference to compute posterior probability of $P(Y|Z = z)$. In this task Y is a binary variable, and classification task is to find the assignment for y which makes the $P(Y|Z = z)$ maximum. $y^* = argmax_y P(Y = y|Z = z)$. Bayes rule give us:

$$P(Y = y|Z = z) = \frac{P(Z = z|Y = y)P(Y = y)}{P(Z = z)} \tag{10}$$

And since denominator is fixed for each assignment Y=y. We compute the $argmax$ only using the numerator.

$$y^* = \underset{y}{\operatorname{argmax}}\, P(Z = z|Y = y)P(Y = y) \tag{11}$$

So for a given dataset we can learn the model $P(Z = z|Y = y)P(Y = y)$ for each class label, and then given a test dataset assign the most likely class label using the equation(11) [17]

7.4 Description of the features

1. Correlation: All pair-wise correlations between super-voxels in the (spatially) down-sampled data. The original data is 53x64x37 voxels. The number of all pair-wise correlations is extremely large for the original data. Data have been down sampled to 13x16x12. Then all of the pair-wise correlations computed, which is $(731x731 - 731)/2$. 731 is the number of nonzero elements in the subjects universal brain mask (intersection of all subjects brains) in the down sampled data.
2. Eigenvalues: (Nonzero) eigenvalues of the correlation matrix

3. Log-disconnection: instead of thresholding for correlation coefficient $r > 0.7$ and finding the degrees, It has been thresholded for $r < 0.4$ to have a measure of dis-connectivity of a voxel. In contrast, with the degree features.
4. Log-disconnection regions: as the log-disconnection feature, but with only difference of being limited to three regions: superior frontal gyrus, middle frontal gyrus and Thalamus.
5. MBI stat: statistics of the average brain intensity signal (for each subject) as the feature. These were mean, standard deviation, skewness and Kurtosis.
6. Degree: as described in the report.