Final Report: Improved Discrimination of Asperger Patients using fMRI and Machine Learning

Amanda Funai, Hari Bharadwaj, Will Grissom

12/14/07

1 Introduction

Asperger Syndrome is an autism spectrum disorder that reduces a patient's ability to interact socially, and restricts their interests and abilities. Using functional magnetic resonance imaging (fMRI), it has been shown that Asperger patients exhibit reduced activity between the nodes of a resting-state neuronal network comprised of the posterior cingulate cortex (PCC), the medial prefrontal cortex (MPFC), and the lateral parietal cortex, compared to healthy controls [1].

The study of Ref. [1] proposed a technique to discriminate Asperger patients and healthy controls using the self-organizing map (SOM) algorithm [2] to automatically generate a cluster representing this resting-state network from resting-state fMRI data. If robust, such a method could have important applications in screening for this disorder, both in clinical and research settings. However, this method has a few potential weaknesses. First, it performs discrimination using a single statistic. Second, the SOM algorithm clusters voxel timecourses based on Euclidean distances, and therefore may not robustly cluster voxels that are functionally connected but possess inter-voxel delays. Third, the original method required user interaction to choose the cluster that best represents the resting-state network, which may bias the results. Therefore, we are proposing three innovations to this algorithm. To address the first and last issues, we will apply filter and wrapper methods directly to the fMRI images, bypassing SOM and the need for user interaction. We have also implemented several different discrimination methods to compare to the simple z-score threshold test implemented in

Ref. [1]. These include Fisher discriminant analysis and Support Vector Machines (SVM's). To address the second issue, we are implementing a kernelized version of SOM, which allows for using a different metric rather than the direct Euclidean norm between timecourses and cluster exemplars. Each of the these innovations will be ultimately compared and contrasted to find the best technique to differentiate between normal and Asperger's patients. In the following sections we will describe the original discrimination method and our innovations, and provide a comparison between their performance.

2 Background

2.1 The Self-Organizing Map Algorithm

The Self-Organizing Map (SOM) algorithm [2] generates a map of length-n exemplar vectors $\boldsymbol{m}_i$ that optimally describe a set of observed length-n vectors $\boldsymbol{x}_k$. The algorithm is initialized by defining a grid of nodes, each of which is associated with a random exemplar vector $\boldsymbol{m}_i$. At each iteration t, updates to the exemplars are made as follows. An observation $\boldsymbol{x}\left(t\right)$ (in this scenario an observation is a voxel's magnitude fMRI timecourse) is compared to each exemplar vector, to find the prototype index $\hat{c}$ that satisfies:

$$\hat{c} = \underset{c}{\operatorname{argmin}} \left\|\boldsymbol{x}\left(t\right) - \boldsymbol{m}_c\left(t\right)\right\|^2 . \tag{1}$$

The exemplars are then updated according to:

$$\boldsymbol{m}_i\left(t+1\right) = \boldsymbol{m}_i\left(t\right) + h\left(i;\hat{c},t\right)\left(\boldsymbol{x}\left(t\right) - \boldsymbol{m}_i\left(t\right)\right), \tag{2}$$

where $h\left(i;\hat{c},t\right)$ is a *neighborhood* function, that updates exemplars according to their distance from $\boldsymbol{m}_{\hat{c}}$. $h\left(i;\hat{c},t\right)$ is initially a wide function that narrows as iterations progress, until only $\boldsymbol{m}_{\hat{c}}$ is updated. We take $h\left(i;\hat{c},t\right)$ to be the Gaussian function

$$h\left(i;\hat{c},t\right) = \alpha\left(t\right) exp\left(-\frac{\left\|\boldsymbol{r}_i - \boldsymbol{r}_{\hat{c}}\right\|^2}{2\sigma^2\left(t\right)}\right), \tag{3}$$

where $0 < \alpha\left(t\right) < 1$ is a learning rate that decreases with t, $\boldsymbol{r}_i$ and $\boldsymbol{r}_{\hat{c}}$ are the positions *on the SOM grid* of the updated and 'best' exemplars, respectively, and $\sigma^2\left(t\right)$ controls the width of the neighborhood function, and also decreases with iteration. Note that we treat a different observation vector at each iteration, but when t reaches the number of observation vectors, we may loop back to the first observation vector and continue updating.

2.2 SOM-based discrimination of healthy controls and adults with Asperger's Disorder

The original method of Ref [1] used resting-state fMRI datasets. These were obtained by acquiring 270 volumes of images, taken once every 0.75 seconds, with the subjects lying down in the scanner, awake but resting. The total dataset is comprised of 8 adult Asperger's subjects, 10 healthy adult subjects that were matched in IQ and age range to the Asperger's subjects, and 8 separate unmatched healthy adult subjects.

Each subject's data was independently processed as follows. The SOM algorithm was applied to each subject's data to obtain a 10 $\times$ 10 grid of exemplars. SOM was again applied to the 100 exemplars, to obtain 16 superclusters. The correlation (z-score) between each voxel timecourse and each supercluster exemplar was then calculated, yielding 16 z-score maps. These were then examined and the 'best' map corresponding to the network of interest was determined by hand.

To perform discrimination, the 'best' z-score maps were transformed to a normalized brain atlas. Then a mask for the network of interested was calculated by thresholding the summed 'best' z-maps of the control subjects. This mask was applied to the 'best' map of each subject to calculate the average z-score within the masked regions. This statistic was used to discriminate between groups, via a threshold that was chosen to achieve zero false negatives, and is therefore the maximum average z-score of the patient group.

Figure 1 shows an example mask and z-maps generated by this technique. Figure 2 shows a box plot of the z-scores for each group. While the average z-score is a good discriminator between the matched controls and the patients, it is a poor discriminator between unmatched controls and patients. We performed leave-one-out cross validation to estimate the error of this method as 0.23 (6 errors out of 26 subjects).

3 Approach I: Wrappers and Filters

Our first approach explored various feature extraction methods to discriminate between patients with Aspergers and healthy controls. We worked directly on the brain images themselves, without using any physiological assumptions. These approaches will require no user interaction.

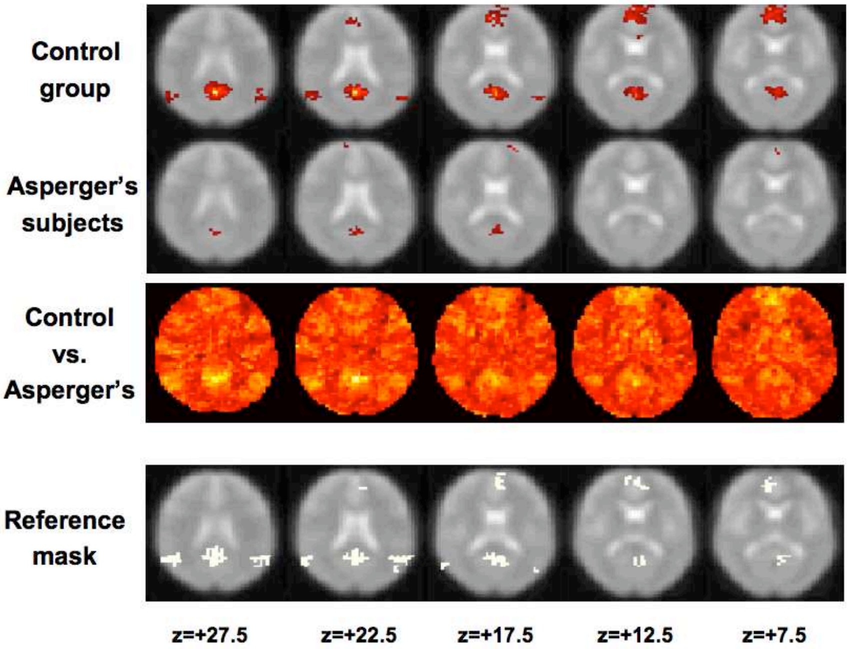

Figure 1: Top three rows: Thresholded z-scores for the cluster encompassing the resting-state network of interest, superimposed on anatomical images. Asperger's subjects exhibit reduced functional connectivity between regions. Bottom row: Reference mask used to calculate average z-scores for classification. Each column is a different slice in the volume (moving superior to inferior through the brain). Image adapted from Ref. [1].

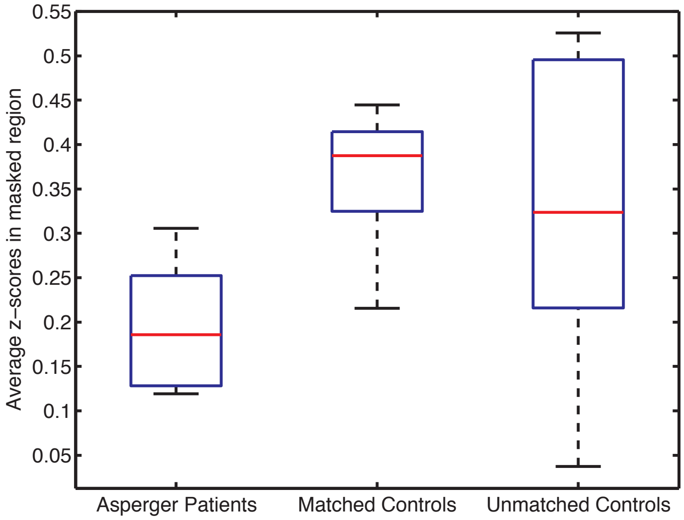

Figure 2: Average z-scores for Asperger's subjects and healthy controls in the reference mask. Thresholding the statistic is a good discriminator between patients and matched controls, but a poor one between patients and unmatched controls.

3.1 Technical description

Feature extraction, or feature subset selection, is a difficult problem [3]. Even defining relevant features can be difficult (Ref. [4] cites 6 different definitions) especially when optimality and relevance are not directly related. For example, a Bayes classifier would use every feature; in practice, however, a smaller subset of features performs better - and often an ‘optimal’ feature set may not even contain some strongly relevant features. This difficulty precludes methods based on minimizing error or other criteria, so heuristic algorithms are most common.

Filtering methods attempt to find the most relevant feature without assuming a final classification algorithm. In a preprocessing step, an algorithm (i.e., a ‘filter’) is applied to each feature which gives a numerical value used for ranking. The top m features are then used in classification. T-test statistics, the filters used in this paper, look at the variance among the training data of each feature and are defined as

$$t^{(j)} = \left| \frac{\bar{x}_{+}^{(j)} - \bar{x}_{-}^{(j)}}{s/\sqrt{n}} \right|, \tag{4}$$

where j is the feature index and s is the pooled sample standard deviation. In our application, $\bar{x}_{+}^{(j)}$ and $\bar{x}_{-}^{(j)}$ are the average value of voxel j among Asperger’s patients and healthy controls, respectively. The larger $t^{(j)}$, the more variance (and hopefully information) is captured by the feature. Unfortunately, these methods often pick features that, while strongly correlated with the class label, hurt performance under the final algorithm used for classification. In addition, while filtering may find the best m features, these are not necessarily the m *best* features.

Wrapper methods use the accuracy of the classification method to determine the best subset of features. One chooses how to search the feature space, how to predict the accuracy of the given subset, and the classification algorithm to be used. Usually the feature space is too large to search through every possible subset of features so different strategies are used, such as forward selection. Cross-validation can be used for accuracy prediction. Classification methods include decision trees, Naive Bayes, and LDA. In this report, a forward search method is used - this method starts with an empty set of features and applies the classification algorithm to each feature and returns an accuracy prediction. The best feature is chosen and the wrapper loops through all the features again, applying the classification algorithm to the best feature plus each individual feature. This continues until a stopping rule is satisfied. Although wrapper methods appear to be a brute force method, they perform well, [3].

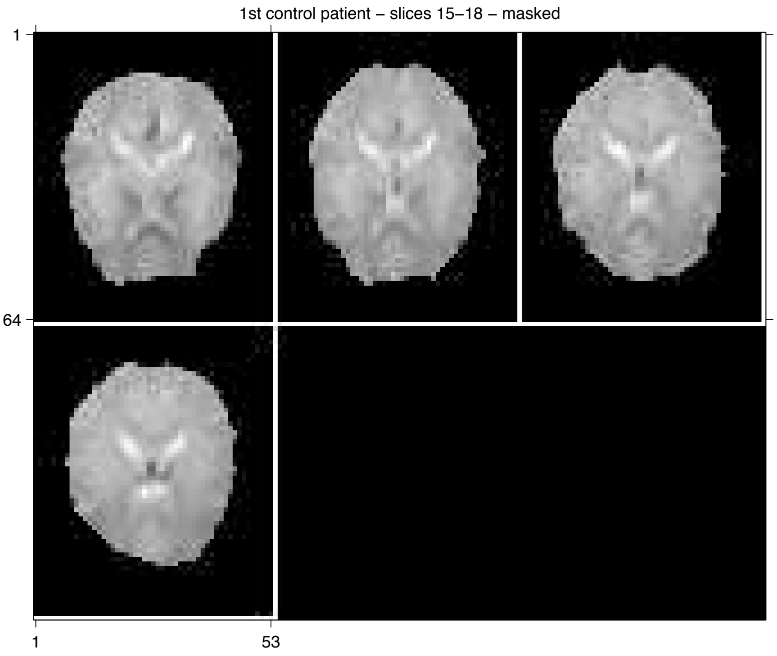

Figure 3: Control patient - four slices of the brain, shown masked

One particularly difficult problem scenario in classification is high dimensionality with a small number of samples - the problem we face here. Methods using a Gaussian assumption and calculating a covariance matrix (for example, naive Bayes or LDA) perform especially poorly as the small number of samples causes the covariance matrix to be unstable [5]. In fact, [6] proves that the classification rate will tend to .5 under these circumstances as noise accumulates. One solution they suggest is to use a filtering method first to reduce the dimensionality of the problem and then applying an appropriate classifier.

3.2 Methods, Results and Discussion

In this approach we operate directly on the subjects' fMRI timecourses, after transforming them to a normalized brain atlas. We used four brain slices for each subject. An example of one of the control patients is shown in Figure 3. For this approach, only the first time point was used. The brain in each image was also masked to remove the background. Overall, of the original 33886080 dimensions per subject, we used 5609.

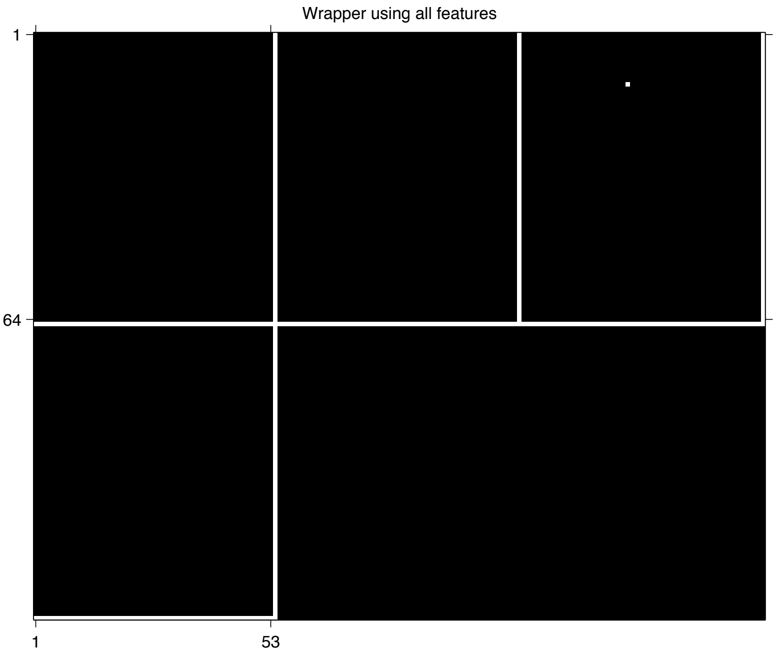

Figure 4: SVM selected features (error = 0.12) using all features

The first algorithm we applied was a forward search wrapper using LDA and Naive Bayes classifiers, and leave-out-out cross-validation (L1OCV) as an error metric. The search stopped when the error was reduced by less than 0.1%. Both wrapper methods always achieved an error rate of 0.31. The linear estimator always classified each data point as belonging to a healthy control. According to the literature, this rate is to be expected with such a small sample size, due to inaccuracy in covariance matrix estimation.

The next algorithm we tried was a forward search wrapper using L1OCV and the same stopping rule as before, but using the SVM algorithm with a linear kernel. Because this algorithm maximizes the margins between classes instead of fitting the class data to Gaussian curves, SVM should perform better than LDA or naive Bayes. The error rate for an SVM wrapper is 0.12 selecting two features, which constitutes an improvement over the original SOM method's error of 0.23. The two chosen features are shown in Fig. 4.

To address the curse of dimensionality, we incorporated the t-test filter method first, and then applied the best features to the SVM wrapper [6]. The features with the highest 100 t-test statistics were used in the previously-described SVM

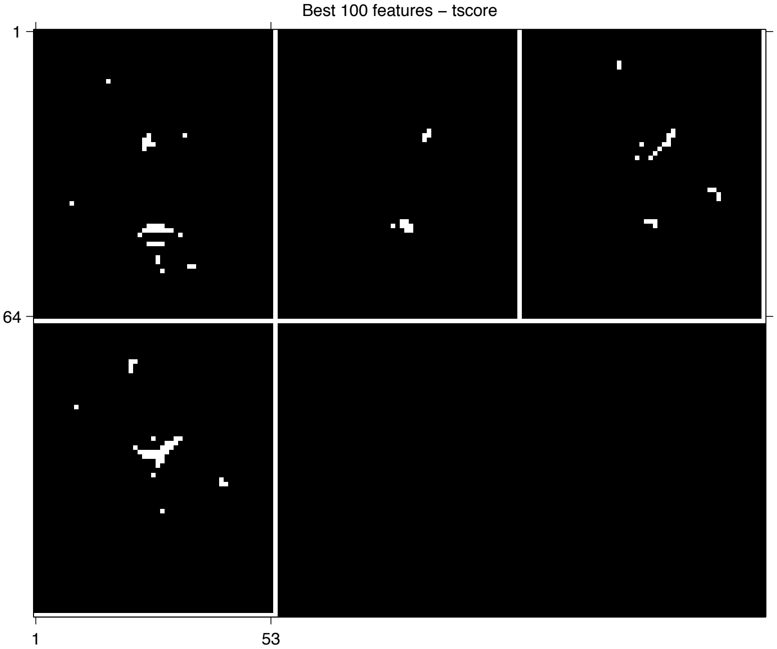

Figure 5: Best 100 features using t-test statistic filtering (shown in white)

wrapper. The top 100 and 500 features from filtering are shown in Figs. 5 and 6. The regions covered by the top features encompass the areas suggested by physiology, but also cover other areas of the brain. Using these features, four features were selected (Fig. 7), achieving the same error rate of 0.12.

The wrapper and filtering combined with wrapper method achieved an improved classification error of 0.12, despite the small size of the data set. However, a classifier based on only two to four pixels is not likely to perform well when applied to new MR images. Ideally, we would select more pixels, for example 25 or 50 to create a mask (like the SOM) to use on future test samples. Unfortunately, the curse of dimensionality combined with the small sample size prevents this approach. Because the wrapper method could not be run to find the best m features due to time constraints, we used the top m features (as found by filtering) and used SVM to find the L1OCV error as shown in Table 1.

We also explored shrinking the images to a smaller image (1/4 to 1/2 the original size) and applying filtering and then the wrapper method, but the subtle details of the original were lost and the SVM wrapper gave an error or .3077 for each.

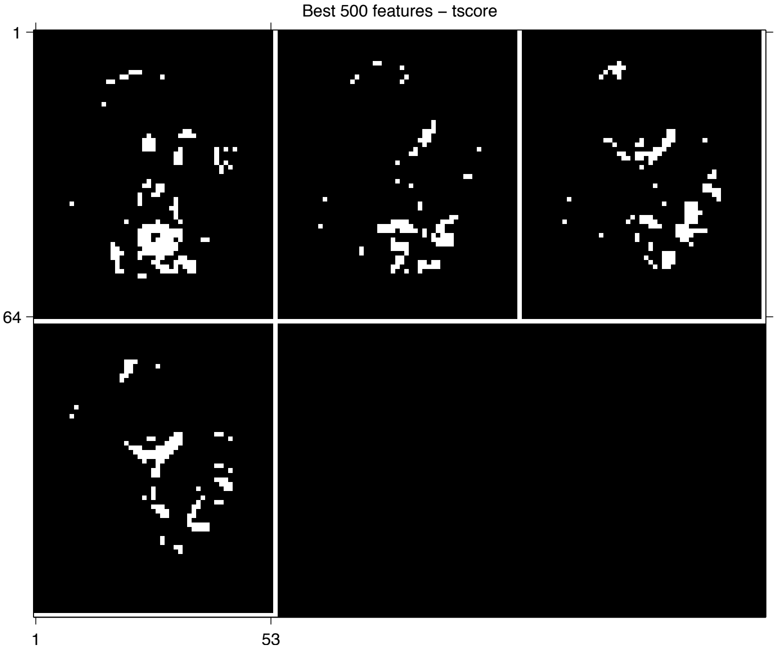

Figure 6: Best 500 features using t-test statistic filtering (shown in white)

Table 1: Error rates using best m features found by filtering using SVM with a linear kernel

Number of Features	Error
1	0.38
5	0.31
25	0.54
50	0.54
100	0.62
200	0.69
500	0.54
1000	0.65

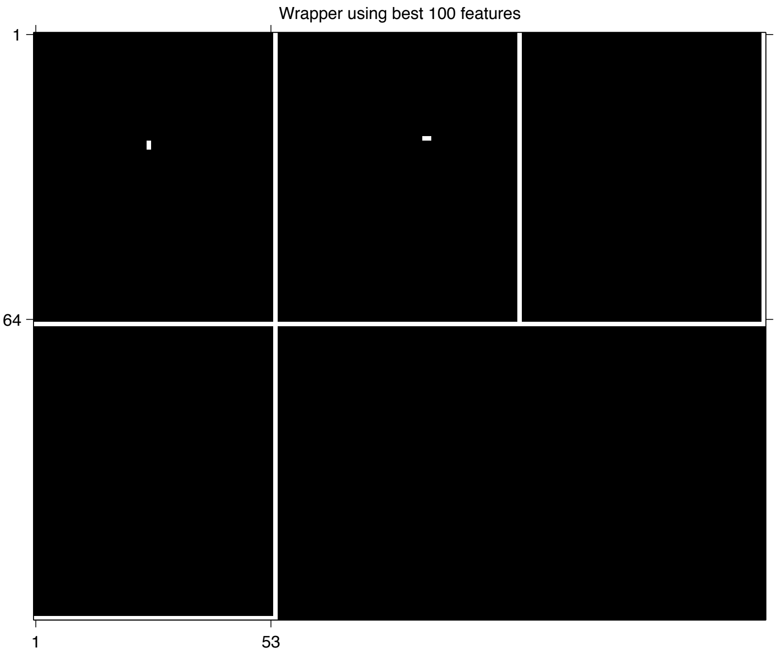

Figure 7: SVM-selected features (error = 0.12) using top 100 features

4 Approach II: Kernelized SOM

4.1 Drawbacks to the standard SOM; our approach

While this Euclidean distance metric (Eq. 1) used in the standard SOM may be effective within one spatial region, correlations between distinct spatial regions may not be fully reflected by it due to, e.g., delays between the time-courses of voxels in the same functional network. In fact, it has been shown that temporal delays between the time-courses of connected brain regions are significant enough to indicate causality between the regions [7]. As a consequence, it is reasonable to expect that a clustering method that clusters two like time-courses with some delay between them with a higher degree of confidence may outperform standard SOM in this classification task. Therefore, we explored the application of SOM's to alternative representations of the data, using the kernelized SOM method [8]. The kernelization of SOM is simple; assuming the input timecourses $\boldsymbol{x}\left(t\right)$ have been normalized so as to possess unit norm, the Euclidean distance metric used by SOM effectively reduces to an inner product:

$$\|\boldsymbol{x}(t) - \boldsymbol{m}_c(t)\|^2 = 2 - 2\langle \boldsymbol{x}(t), \boldsymbol{m}_c(t)\rangle . \tag{5}$$

We can kernelize the SOM by replacing $\boldsymbol{x}(t)$ with a non-linear function of the data, $\phi(\boldsymbol{x}(t))$. Note that the data should be normalized to have unit norm in the function space. In particular, we used functions capable of highlighting correlations between voxels with time-courses that have some delay between them.

The success of this approach hinges on two assumptions. First, we assume that delays between the time-courses of functionally connected areas actually exist in our data, and second, we assume that the alternative representations of our data will allow clustering of greater confidence and subsequently greater discrimination power between patients and controls. That is, we assume that voxels which are functionally connected but which have a delay between them will possess low values of $\|\boldsymbol{x}(t) - \boldsymbol{m}_c(t)\|^2$ that obscure differences due to reduced levels of connectivity.

4.2 Methods

To make the SOM robust to delays between timecourses, we chose nonlinear functions that threw away varying degrees of timing data. Prior to running the SOM algorithm, we transformed the timecourses to the Fourier domain using the DFT. To obscure a delay of t_{del} seconds, we quantized the phase of each Fourier harmonic according to:

$$\hat{\theta}_k = \left\lfloor \theta_k \frac{N\Delta t}{2\pi k t_{del}} \right\rfloor \frac{2\pi k t_{del}}{N\Delta t}, \qquad k = 1, \ldots, \frac{N}{2}, \tag{6}$$

where k is the index of the Fourier harmonic, $\Delta t = 0.75$ seconds is the sampling period of the fMRI experiment, and $N = 270$ is the number of volumes sampled. The width of each quantization bin is equal to the phase shift corresponding to a delay of t_{del} seconds. The quantized-phase data was then transformed back to the time domain, and SOM proceeded as usual. We tested this kernelized SOM for several values of t_{del}, from 1 to 10 seconds. Leave-one-out cross-validation error was recorded for each value of t_{del}, for comparison to the standard SOM of section 2.1.

Table 2: Error rates for various levels of phase quantization. The entry $t_{del} = 0$ corresponds to the original SOM.

t_{del}	Error
0	0.23
1	0.38
2	0.38
3	0.35
5	0.35
10	0.31

4.3 Results

Table 2 lists the leave-one-out cross validation errors for 5 quantization levels, compared to the original SOM ($t_{del} = 0$). From the table, it is evident that phase quantization increased the estimated error, indicating that phase quantization does not improve discrimination performance. Figure 8 shows a boxplot of the average z-scores in the masked region for the phase-quantized SOM with $t_{del} = 10$ seconds. Compared to Fig. 2, the phase-quantized SOM has resulted in a wider range of average z-scores for the matched controls and more overlap of z-scores with the Asperger's patients. This will result in decreased discrimination performance. The phase-quantized SOM did, however, consistently (i.e., for all t_{del} values tested) reduce the variance of z-scores for the unmatched controls, as well as this group's overlap with the Asperger's patients. This will result in improved discrimination performance between patients and this group. Overall, the phase-quantized SOM does not improve our ability to discriminate between Asperger's and controls. The reason for this could be simply that significant delays between the connected brain regions do not exist in this data.

5 Approach III: Spectral Coherence and Linear Discriminant methods

This section reports attempts made to improve discrimination using linear classifiers (LDA, FDA and SVM's) and another set of features derived from the coherence metric. One of the drawbacks of the similarity or the connectivity measure used in [1] is that it uses Euclidean distances between the timecourses, and therefore does not specifically take advantage of the nature of the physical quantity

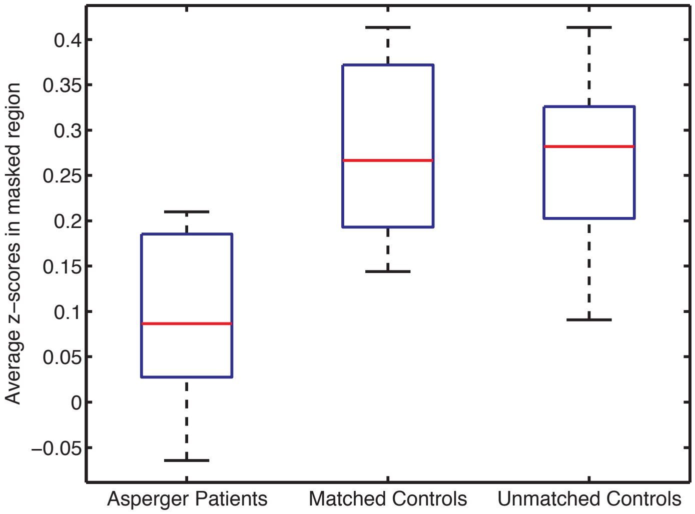

Figure 8: Average z-scores for Asperger's subjects and healthy controls in the reference mask, for the phase-quantized SOM method with $t_{del} = 10$. Compared to Fig. 2 for the original SOM, the phase-quantized SOM produces average z-scores with a wider range for the matched controls, leading to increased discrimination error for that group. However, it produces a narrower range for the unmatched controls, so that the net effect of quantization is a small increase in estimated error.

which it attempts to capture. By definition, functionally 'connected' brain regions share a causal relationship with each other [9]. Spectral coherence has been effectively used as a metric that captures this causal synchrony in electroencephalography (EEG), another functional neuroimaging modality [10].

5.1 Spectral coherence as a feature capturing functional connectivity

Coherence is a statistical quantitative measure of the phase consistency between two signals (section 5.2.2). Consider, for example, coherence between voltages at two nodes in a linear, noise-free electric circuit. Whereas each node voltage will oscillate at the AC generator frequency with generally different phase, phase differences remain fixed over time. Coherences between all paired voltages in the circuit are equal to one in such linear circuits. Thus coherences measured between separate electric circuits or between distinct cortical voxel locations provide measures of mutual influences or long-range synchrony, but the magnitudes of such influences in complex, non-linear dynamic systems can be quite different at different frequencies or different spatial scales [11]. Hence it may be hypothesised that the coherence measure would better represent 'similarity' between two time-courses for functional connectivity studies than Euclidean distance. The actual coherencies are analogous to correlation coefficients and are dependent on unobservable probability density functions of associated stochastic processes and hence can only be estimated [11]. The procedure to estimate coherence are discussed in section 5.2.

5.2 Methods

5.2.1 Data preprocessing and ROI selection

The data obtained from the patients, matched controls and unmatched controls were transformed to a normalized brain atlas. A reference binary mask was derived by thresholding the z-scores obtained using the procedure outlined in [1] to select a region of interest (ROI). Pairwise coherences between all the voxels that fall in the ROI were then calculated as described in the 5.2.2. However since the number of coherence metrics to be computed was combinatorially large, the image was blurred by averaging 32 neighboring voxels (4×4 in-plane and 2 slices) into one voxel, resulting in about 600 metrics to be estimated for each subject.

5.2.2 Feature extraction and Classification

The coherence between any two time courses was estimated as follows:

1. The time-courses were divided into $N = 5$ non-overlapping epochs of the same length $(40$ seconds)

2. The cross spectral density of the 2 time courses for a given epoch i and the estimated cross spectral density was computed as:

$$\begin{aligned} G_{12i}(f) &= G_{1i}(f)G^*_{2i}(f) \qquad (7) \\ \widehat{G}_{12}(f) &= \frac{1}{N}\sum_{i=1}^{N} G_{12i}(f), \qquad (8) \end{aligned}$$

   where G_{1i} and G_{2i} are individual Fourier transforms. 9-point FFTs were obtained from each 54-point epoch in time. Each epoch was 40 seconds long, implying that the frequency resolution in the coherence estimates was about 0.025 Hz.

3. The estimate of spectral coherence based on N epochs was computed as (with the individual power spectra being calculated similarly)

$$\widehat{\gamma}^2_{12}(f) = \frac{|\widehat{G}_{12}(f)|^2}{\widehat{G}_1(f)\widehat{G}_2(f)}. \qquad (9)$$

The average length-9 coherence vector over each subject was taken as a feature characterising functional connectivity over the ROI. Typical coherence vectors obtained from a patient and from a matched control are shown in Fig. 9. This feature vector was passed through the FDA, LDA and SVM discriminators. For comparison, the mean z-scores over each of 6 selected slices from the ROI were strung into a z-score feature vector of length 6. Linear discrimination was also used in [1], but with z-scores averaged over the entire ROI, resulting in a threshold comparator. Using the slice averages may avoid the loss of accuracy that may possibly be caused by offsetting of the overall mean by one or two ‘noisy’ slices and allow the discriminators to ‘learn’ which slices are more important.

Linear discriminators were explored in this section. LDA, FDA and SVM’s were implemented to classify Asperger patients from healthy controls. SVM’s were kernelised and used a Gaussian kernel.

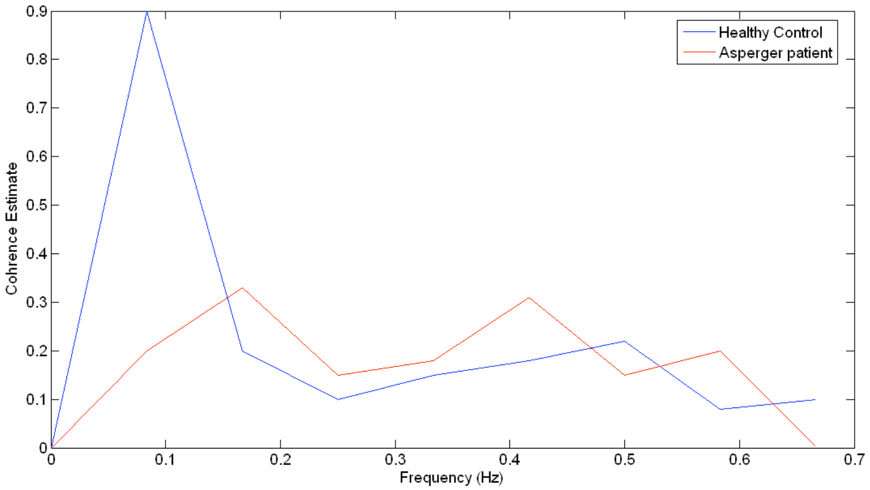

Figure 9: This shows two typical 9 point coherence feature vectors derived from a control (blue) and a patient (red). Observe that in this example plot, the coherence of a healthy control in the < 0.1 Hz range is higher than that of the patient.

5.3 Results and Discussion

Leave-one-out cross validation errors were computed for each of the classifiers and feature type. The matched controls group and the separate controls group were lumped together as one class ($n_1 = 18$) and the patients were another class ($n_2 = 8$). Classifier performance is summarized in Table 3. LDA and FDA performed identically. In fact, they were identical classifiers, i.e., they generated the same hyperplane as would be expected from theory.

Thus it can be seen that stringing the slice z-score averages instead of using the overall averages improved the classification from [1] by about 4% when used

Table 3: Leave-one out misclassification errors for LDA, FDA and SVM's using the coherence and the z-score vectors discussed in section 5.2.2.

Classifier	Patients	Controls	Overall
LDA-FDA:z-scores	0.12	0.27	0.23
SVM:z-scores	0.12	0.22	0.19
LDA-FDA:Coherence	0	0.27	0.19
SVMs:Coherence	0	0.27	0.19

with SVM's. Using the coherence estimates improves the classification by 4% for all the discriminators. Whether the advantage of this marginally superior performance would scale to larger data-sets is unknown. However, one major advantage of using the coherence estimates of blurred images is that it is computationally inexpensive compared to the calculations required for SOM. Calculating FFT's for small timecourses can be done rapidly. In particular, the SOM algorithm requires a sequential updating of the exemplars whereas calculation of the coherence estimates can be done in parallel and hence advantageous when used with the single instruction multiple data procesors that are widely in use today. Though all the methods are inherently offline, the coherence method would accelerate diagnosis and group analyses.

6 Conclusion

We have demonstrated three approaches to improving discrimination between patients with Asperger's and healthy controls, based on resting-state fMRI data. The methods we implemented aim to address the shortcomings of the SOM-based method of Ref. [1].

In our first approach, we implemented filter and wrapper methods on the data itself, in order to build classifiers based on more data than the single average z-score statistic used by the original method. These methods also obviate the need for user interaction in the classification process, thereby avoiding the possibility of user-bias. The filter and wrapper methods were able to achieve improved discrimination accuracy, but the small number of subjects and high dimensionality of each subject's data generally posed a problem to these methods. Among these methods, the forward search method combined with the SVM algorithm, and the t-test filter combined with the SVM algorithm yielded the best results, cutting the estimated error in half.

In our second approach, we addressed a different shortcoming of the original approach, namely that the Euclidean distance used in the SOM clustering algorithm does not fully coincide with our notion of functional connectedness. To allow the possibility that there may exist functionally connected regions with some time delay between their otherwise like timecourses, we permitted phase lags between regions by quantizing the phase of the data's Fourier harmonics, prior to running the SOM algorithm. This non-linear transformation of the data, combined with SOM, comprises a kernelized SOM algorithm. This approach, however, did not yield improved discrimination power. Therefore, we may conclude that time delays do not significantly reduce the apparent functional connectedness of distinct brain regions in this scenario. This may be because delays simply do not exist in the data, or that they exist but do not cause the connected regions of interest to appear dissimilar, compared to other regions.

Our final approach to improving discrimination combined the goals of the first two approaches. In this approach, an alternative to average cluster z-scores was used as a metric, namely coherence between voxels in the pre-identified network of interest. Coherence is a metric of connectedness that, in contrast to Euclidean distance, remains high for timecourses that possess a delay between them. This new metric, combined with classification methods that function on more than one statistic, were shown to generally provide improved classification error compared to the original SOM-based method. Compared to the other two approaches, this method is probably the most deserving of further investigation, since the coherence metric fits our notion of connectedness, and the more sophisticated classification methods that function on this metric are likely to provide the best error among all the methods we tried, after further investigation.

7 Group member contributions

1. Amanda Funai: Filter and wrapper methods (Approach I, Section 3). Wrote introduction to progress report. Poster compilation.

2. Hari Bharadwaj: Spectral coherence and linear discriminant methods (Approach III, Section 5). Poster compilation.

3. William Grissom: Compilation of reports and proposal, transformation of the data to a normalized atlas, implementation of original SOM-based method, Sections 1, 2, and 4 (kernelized SOM).

References

[1] S J Peltier, O Ousley, R Welsh, C Kilts, K Harenski, and X Hu. Model-free discrimination of resting-state activity between controls and adults with Asperger's Disorder. *Proceedings, Human Brain Mapping*, page 369, 2007.

[2] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[3] I Guyon and A Eliseff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–82, 2003.

[4] R Kohavi and G H John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–74, 1997.

[5] P. Guo and M. Lyu. Classification for high-dimension small-sample data sets based on kullback-leibler information measure, 2000.

[6] J. Fan and Y. Fan. High Dimensional Classification Using Features Annealed Independence Rules. *ArXiv Mathematics e-prints*, January 2007.

[7] A Roebroeck, E Formisano, and R Goebel. Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, 25(1):230–242, 2005.

[8] K. W. Lau, H. Yin, and S. Hubbard. Kernel self-organising maps for classification. *Neurocomputing*, 69(16-18):2033–2040, 2006.

[9] K J Friston, C D Frith, P Fletcher, P F Liddle, and R S Frackowiak. Functional topography: multidimensional scaling and functional connectivity in the brain. *Cereb Cortex*, 6(2):156–164, 1996.

[10] R.D. Katznelson. *Deterministic and stochastic field theoretic models in the neurophysics of EEG*. PhD thesis, University of California at San Diego, 1982.

[11] P L Nunez, R Srinivasan, A F Westdorp, R S Wijesinghe, D M Tucker, R B Silberstein, and P J Cadusch. EEG coherency I: Statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalogr Clin Neurophysiol*, 103(5):499–515, 1997.