Decoding Cognitive States from fMRI Timeseries

Catie Chang
catie@stanford.edu
CS229 Final Project Report

I. Introduction

Conventional analysis of functional magnetic resonance imaging (fMRI) data follows a regression-based approach, in which one identifies the neural correlates of a particular cognitive function by correlating individual voxel timecourses with a known pattern of stimulus presentations. However, one can reverse the direction of analysis; rather than using knowledge of the stimulus pattern to infer brain activity, one can ask whether brain signals can be used to predict perceptual or cognitive states. This problem falls naturally into the domain of machine learning, and can be viewed as an instance of learning classifiers from extremely high-dimensional, sparse, and noisy data [1].

Recently, Mitchell et al. [1] demonstrated the feasibility of training Gaussian Naive Bayes (GNB), k nearest neighbors (kNN), and Support Vector Machine (SVM) classifiers to discriminate between a finite set of cognitive states. For instance, given fMRI data from a single time interval $[t1, t2]$, the classifier could determine whether the subject was viewing a picture or reading a sentence during that interval. Cox and Savoy [2] used Linear Discriminant Analysis (LDA) and SVM to investigate the neural representation of various object categories, and Davatzikos et al [3] applied SVMs and LDA to the problem of lie detection. The ability to automatically classify fMRI data has found applications in clinical diagnosis as well; Zhang et al [4] use KNN, GNB, SVM, and Adaboost classifiers to discriminate between the brains of healthy versus drug-addicted subjects.

Besides demonstrating the power of machine learning algorithms, classifying fMRI data can provide valuable insight into mechanisms of brain functioning; that is, activation patterns which play a critical role in discriminating between perceptual or cognitive conditions may likely have strong neurobiological relevance. In this way, classification may be regarded as a hypothesis-generating method. Furthermore, the approach of characterizing entire spatiotemporal activation patterns represents one important departure from massively univariate analyses (e.g. SPM [5]) commonly found in functional neuroimaging literature.

In this project, we first apply well-known machine learning algorithms (e.g. LDA, SVM) to the problem of discriminating between cognitive states both a mental arithmetic task (block design) and a working memory task (event-related design). In each task, classifiers are trained using short time intervals of fMRI data known to correspond to either of two conditions: (1) an experimental condition, or (2) a control condition. After training, the classifier must determine whether an unseen segment of fMRI data corresponds to an experimental or control condition. For both single-subject and multiple-subject classification, we are able to achieve accuracies comparable to those reported in [1], [6], [3].

Next, we characterize the most discriminating regions of the brain for both cognitive tasks by examining the weight vectors (w) returned by the SVM classifiers. We demonstrate that these *discriminating volumes* overlap to some extent with the t-maps resulting from the univariate general linear model (GLM) analysis. This work is similar to that described in [6].

Finally, we explore the use of Dynamic Bayesian Networks (DBNs) in modeling functional interactions between brain regions in the working memory task. Unlike the LDA and SVM classifiers, which use only spatial patterns of voxels to characterize differences between cognitive states, DBNs can capture the statistical relationships between brain regions over time. Hence, modeling fMRI data using DBNs can be used for both classifying new data and generating hypotheses about functional interactions between brain regions throughout different stages of cognitive processing.

II. Classification

Training Classifiers

In both the arithmetic and working memory tasks, the classification function assumes the following form:

$$f:\ \text{fMRI-sequence}(t_1,t_2) \rightarrow \{\text{State0, State1}\}$$

fMRI-sequence(t_1, t_2) refers to the set of fMRI images collected during the time interval $[t_1, t_2]$. Our method for encoding fMRI-sequence(t_1, t_2) is based on [1]. Each training example $x^{(i)}$ is formed by concatenating the signal values from p selected voxels over (t_1, t_2). Thus, $x^{(i)}$ is an $(mp \times 1)$ vector, where m is the number of fMRI images acquired during the interval (t_1, t_2). Each $x^{(i)}$ has an associated binary-valued label $y^{(i)}$, where $y^{(i)} = 0$ if (t_1, t_2) corresponds to a trial of class*State0* and $y^{(i)} = 1$ if (t_1, t_2) corresponds to trial of class *State1*.

In the mental arithmetic task, *State1* corresponds to a trial in which the subject views a valid (though possibly incorrect) equation (e.g. "4+5=9", or "2+4=7"), and must determine whether or not the equation is correct. *State0* corresponds to a trial in which the subject views a non-equation string of numbers and symbols (e.g. "4@5!9"), and must determine whether or not the equation contains the number "5". In the working memory task, the subject is prompted to remember a string of 5 numbers that are either all the same ("low load"; *State0*) or all different ("high load"; *State1*).

Dimensionality Reduction Because the number of voxels in the brain is large (over 550,000 for our data), feature

TABLE I

LOOCV errors obtained across multiple subjects, for the Arithmetic task.

Method	Kernel	Errors	Mean Error
LDA	n/a	0.1875, 0.1562, 0.1938, 0.2563, 0.2875, 0.2250, 0.1813	0.2125
SVM	RBF	0.1688, 0.0938, 0.1688, 0.1500, 0.1938, 0.1938, 0.1688	0.1625
SVM	Linear	0.1625, 0.1250, 0.1875, 0.1813, 0.2563, 0.1625, 0.1562	0.1759

TABLE II

LOOCV errors obtained across multiple subjects, for the Working Memory task.

Method	Kernel	Errors	Mean Error
LDA	n/a	0.23, 0.44, 0.4, 0.25, 0.38, 0.25, 0.33, 0.33, 0.38, 0.31, 0.29, 0.29, 0.17, 0.33	0.3125
SVM	RBF	0.21, 0.40, 0.38, 0.25, 0.38, 0.25, 0.33, 0.21, 0.25, 0.417, 0.25, 0.25, 0.15, 0.30	0.286
SVM	Linear	0.21, 0.33, 0.38, 0.25, 0.42, 0.19, 0.31, 0.23, 0.25, 0.38, 0.19, 0.27, 0.21, 0.27	0.2768

TABLE III

LOOCV errors obtained across multiple subjects using the SVM classifier on a PCA basis

subject	WM	MA
1	0.2917	0.1688
2	0.1458	0.0750
3	0.3333	0.1500
4	0.2500	0.2062
5	0.2917	0.0563
6	0.2083	0.0938
7	0.2708	0.1000
8	0.1667	n/a
9	0.2917	n/a
10	0.2083	n/a
11	0.3125	n/a
12	0.1667	n/a
13	0.1250	n/a
14	0.2292	n/a
Mean Error	0.2351	0.1214

selection/dimensionality reduction is desirable. We implemented the following two feature selection methods:

(i) Intersection of functional and anatomical ROIs Here, we begin with a set of 116 anatomical ROIs (e.g. hippocampus, cerebellum. These ROIs comprise all available templates supplied by the Marsbar Matlab toolbox [7]). We then determine which voxels in each region are significantly ($p < .01$) active in the experimental condition versus the control condition by separately correlating each voxel's timecourse with the stimulus waveform for each condition (a boxcar signal which is high when the stimulus is on and low when the stimulus is off), and applying a t-test. The significant voxels in a particular ROI are grouped into a single "supervoxel", and the timecourses of each constituent voxel are averaged to obtain the supervoxel's timecourse. Note that some ROIs may contain no significant voxels. When classifying across multiple subjects, we only consider ROIs that contain at least one significant voxel from each subject.

(ii) PCA using voxels from the entire brain Here, no feature (voxel) selection is performed; instead, PCA is applied to reduce the dimensionality of the full ($n \times p$) training matrix, where n is the total number of training examples and p is equal to the number of voxels in the brain. This method allows us to construct "discriminating volumes", i.e. spatial maps that display voxels playing the most critical role in the classification [6]. In addition, unlike *(i)*, this method makes no a priori assumptions about voxel activation, thereby opening the possibility of discovering previously-unsuspected brain regions that might prove critical to the cognitive process of interest. Here, we did not choose to discard any of the principal components (PCs) corresponding to nonzero eigenvectors; hence, there is no information loss in the PCA transformation.

We choose these feature selection methods because they easily extend to the problem of training classifiers across multiple subjects. For instance, the ROI abstraction directly allows for the mapping of corresponding features across subjects [8]. Other feature selection methods proposed in the current literature, e.g. selecting the p most active voxels in the brain [1], do not generalize well to the multiple-subjects case.

In the mental arithmetic task, each block consisted of 15 timepoints. Because there tends to be higher variability in brain activation at the beginning and end of each experimental block, our training and test sets drew data from the 9 timepoints centered in the middle of each block. For the working memory task, we drew 1 data point from each trial (8 seconds after the stimulus onset) to use in the training and test sets.

Other preprocessing steps Prior to forming the feature vectors, we low-pass filter the voxel timeseries, subtract the mean, and remove linear drifts. fMRI data contains high-frequency, task-unrelated noise as well as linear "scanner drift" which can introduce high degrees of variation between training examples of the same class; after low-pass filtering and detrending, training examples belonging to a single class appear more uniform. For the second feature selection method, we downsampled all brain volumes by a factor of 2 before applying PCA (due to memory limitations).

Classifiers We applied the following classifiers: (1) LDA (from the Discrim toolbox for Matlab by Michael Kiefte, Dalhousie Univ.), (2) SVM (SVM toolbox for Matlab by Anton Schwaighofer: http://ida.first.fraunhofer.de / anton/software.html)

Testing Classifiers

We quantify the LOOCV error obtained when training both on (a) single subjects and (b) across multiple subjects. For single subjects, we quantify the performance of the classifiers and feature selection methods using leave-one-out cross-validation. Since fMRI data is strongly correlated in time, the point that is left out (say, at time t), will be related to the training points at $[t-k,\ t+k]$, thereby producing biased error estimates. Thus, when testing on point t, we exclude points $t \pm 3$ from the training set [1].

For multiple subjects, the LOOCV error is the prediction error obtained from testing on one subject (after having trained on the remainder of subjects).

Tables 1 and 2 display prediction errors when feature selection method *(i)* was applied. Table 1 shows the prediction errors on each of 7 subjects for the mental arithmetic task, and Table 2 shows the prediction errors on each of 14 subjects for the working memory task. Results indicate that SVM, compared with LDA, is more capable of generalizing across subjects. The across-subject errors in both tasks are comparable to those reported by [8]. When using SVMs to detect whether subjects were reading a sentence or viewing a picture, they achieved across-subject errors of 14-25%. On the syntactic ambiguity study (determining whether subjects were reading an ambiguous or unambiguous sentence), SVM errors were around 35-45%. In a face-matching task, Miranda et al. [6] achieved error rates of 20-25% (faces versus locations), and 10-15% (faces versus control).

Table 3 displays the across-subject prediction errors when feature selection method *ii)* was applied. The mean error on the mental arithmentic class was 12.14%, and the mean error on the working memory task was 23.51%.

In general, the error for the arithmetic task is much lower than that of the working memory task. This result may reflect the fact that the arithmetic task has a *block-design* structure, while the working memory task has an *event-related* structure. In a block design, several trials of the same condition (experimental or control) are presented in close succession; thus, the characteristic activity pattern for each condition is sustained over several seconds. On the other hand, the experimental and control trials of an event-related design are randomly interleaved, which often means that (a) two trials of the same type are separated in time, and (b) two trials of different conditions may sometimes occur in quick temporal succession, thereby allowing the the slowly-varying brain reponse (hemodynamic response function, or HRF) to the first trial to obscure that of the second trial type. The signal-to-noise ratio is also higher in a block design compared to an event-related design, since HRFs are known to sum in a rougly linear fashion. Thus, event-related designs pose a more difficult challenge to classifiers.

It is also worth noting that the error for the SVM across-subjects classification using feature selection method *(ii)* in the working memory task was highly sensitive to the set of data points used for training/testing. For instance, while using the first and/or second timepoints following the onset of each trial, the across-subject classification error rarely dropped below 45%. However, using the fourth timepoint yielded errors in the 20-30% range.

III. Mapping Discriminating Volumes

The SVM classifier finds a hyperplane that maximizes the the margin of separation between the two classes. The normal vector w defining the optimal hyperplane (and, consequently, the optimal linear transformation of the data points) can be interpreted as the direction along which the two classes differ the most. Thus, mapping the w-vector into a brain volume yields a spatial map indicating the relative discriminating importance of each voxel. While the LDA classifier also yields a separating hyperplane, we chose to generate the discriminating volumes based on the SVM because of SVM's superiority in generalizing to unseen data.

Because SVM was performed in principal components space, we first project the resulting normal vector (w^p) back into voxel-space with the transformation $w = Vw^p$, where V is the matrix of eigenvectors (PCs).

The discriminating volume and corresponding GLM map for the working memory task is shown in Figures 1a and 1b, respectively. Note the presence of clusters in the parietal, motor, and left basal ganglia regions in both maps; these areas are known to play a role in working memory-related processing. The majority of the differences occur in the top and bottom rows of the montages, which do not contain brain structures relevant to working memory processing. The discriminating volume and GLM map for the mental arithmetic task is shown in Figures 1c and 1d. Note that a few of the regions with high weights in the discriminating volume are overlap with regions with high t-values resulting from the GLM analysis (e.g. L/R putamen, L/R insula). However, many salient clusters in the GLM map are missing from the discriminating volume (e.g. cerebellar regions, SMA, cingulum).

In Figure 1, the GLM maps are thresholded to show voxels with significance $p < 0.05$, while the discriminating volumes are thresholded to show voxels whose w-values are above 30% of the maximum absolute value in w. Though it is not exactly correct to compare w-values with p-values, [6] observed that after converting w-values to p-values using nonparametric permuations tests, voxels whose values exceeded 30% of the max matched very closely the voxels which turned up as significant at $p < .05$.

In the working memory task, 56% of the thresholded (>30% of the max) discriminating volume intersected with the GLM map (however, only 14% of the GLM map intersected with the thresholded discriminating volume). For the mental arithmetic, 47% of the thresholded discriminat-

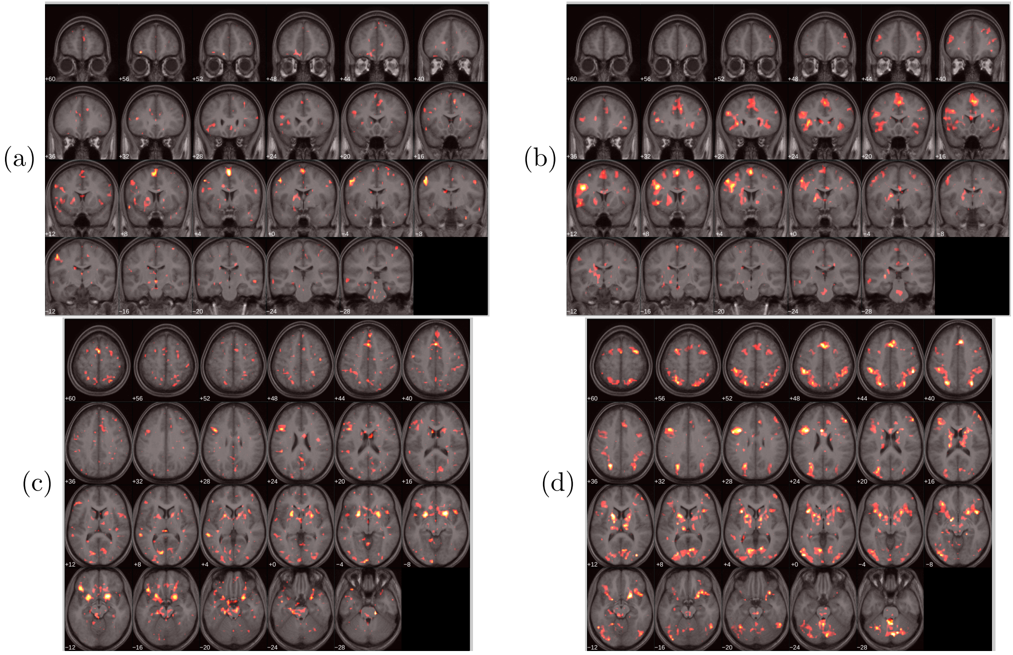


Fig. 1. (a) Discriminating volume for the working memory task - coronal slices. (b) GLM t-map for the working memory taks - coronal slices. (c) Discriminating volume for the mental arithmetic task - axial slices. (d) GLM t-map for the mental arithmetic task - axial slices. For display purposes, elements in the discriminating volume maps (i.e. w, which is a unit vector) are scaled by a factor of 10^4.

ing volume intersected with the GLM map (and again, 14% of the GLM map intersected with the thresholded discriminating volume).

How might we interpret differences between the GLM maps and the discriminating volumes? The GLM is a univariate approach which tries to fit a linear model to the timeseries, whereas the SVM considers only the spatial pattern at each moment in time. While the two methods of analysis provide different perspectives on the data, it is unclear which is more "correct". The SVM and other methods of multi-voxel analysis do, however, provide valuable alternatives to linear modeling; and indeed, there are many cases in which simple linear timeseries models are inappropriate (e.g. experimental designs in which timeseries regressors are highly collinear, or when the actual shape of the hemodynamic response differs significantly from the canonical, modeled response). It is notable that both here and in [6], SVM classification using no prior model of the hemodynamic response, as well as completely unbiased feature selection, identifies regions similar to those identified by a GLM analysis.

IV. Dynamic Bayesian Nets

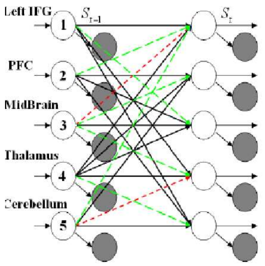
A Bayesian network is a graphical representation of joint probability distributions over sets of random variables. Nodes in the graph represent variables in a system, and arcs represent conditional dependences. Dynamic Bayesian Networks (DBNs) explicitly model temporal processes. In a DBN, nodes are arranged into columns, where each column represents a particular time frame in the process. Arcs connecting nodes between columns represent causal relationships.

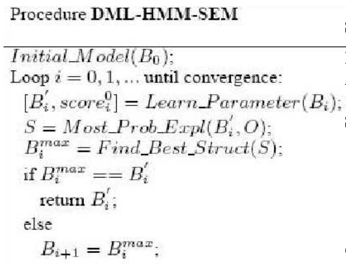
DBNs have recently been applied to the problem of modeling relationships between brain regions in fMRI data. Unlike the GLM, DBNs do not assume linearity of the BOLD response or voxel-wise independence. Burge et al. [9] used DBNs to model differences in brain networks between a group of dementia patients and a group of healthy subjects. They constructed a DBN with 150 nodes in each column, where node i in column t represented the mean signal of the ith ROI at timeframe t. Timecourses were quantized to 2 or 4 levels, so that each node had a discrete conditional probability table (CPT). A structure-learning algorithm was applied to search for the network topology that best explains the data (the Markov assumption is invoked to reduce the search space; only two columns, t and $t+1$, are needed). After the networks were learned, classification accuracy on test data from each group was obtained as a means of validating the resulting structures; their classification accuracy comparable to that of SVMs and Gaussian Naive Bayes (GNB) classifiers.

Very recently, Zhang et al. [10] used DBNs to model the interactions between 5 ROIs in groups of drug-addicted and healthy subjects. A Dynamically Multi-Linked Hidden Markov Model (DML-HMM) [11] was used, and a Structural Expectation-Maximization (SEM) algorithm was performed to learn both the structure and parameters that maximize $P(\text{data}|\text{model})$ (Figure 2).

In this project, I constructed DBNs to explore possible differences in functional interaction between the high-load and low-load conditions in the working memory task described above (Section II). I chose to model the relationships between 6 ROIs that are hypothesized to be important in working memory. As in [9], observations were taken as the mean signal values in each ROI, and were quantized to 4 levels. I experimented with several dif-

(a)





```
Procedure DML-HMM-SEM
$Initial\_Model(B_0)$;
Loop $i = 0, 1, \ldots$ until convergence:
  $[B_i', score_i^0] = Learn\_Parameter(B_i)$;
  $S = Most\_Prob\_Expl(B_i', O)$;
  $B_i^{max} = Find\_Best\_Struct(S)$;
  if $B_i^{max} == B_i'$
    return $B_i'$;
  else
    $B_{i+1} = B_i^{max}$;
```

Fig. 2. (a) Example of a DML-HMM (Zhang 2005, NIPS). (b) SEM algorithm for parameter and structure learning.

ferent models/topologies. The first model was similar to [9] in that each node represented the fully-observed mean ROI signal value. Structure-learning was performed using the REVEAL algorithm [12] with both the BIC and ML scoring criteria, as implemented in the Bayes Net Toolbox (http://bnt.sourceforge.net). The ML criterion constructs arcs that maximize the mutual information between the parents and child of each "familiy", which tended to always result in fully connected graphs. The BIC criterion has a penalty term which lessens the tendency to overfit:

$$BIC = -2 \log L(\theta) + K \log N$$

Here, $L(\theta)$ is the maximum likelihood of the data under the current structure, K is the number of parameters of the model, and N is the size of the training set. However, under the BIC criterion, my structural searches only produced diagonal (identity) adjacency matrices for both the high-load and low-load data sets. I then tried using the ML criterion, while restricting the maximum number of parents of each child to 3. The structures learned for the high-load and low-load condition were different. However, classifying new test data via $\underset{class}{argmax}\ P(\text{data}|\text{Structure}_{class})$ could not be performed significantly better than chance. This is likely a consequence of having too few datapoints in the test set; however, it is also possible that the chosen ROIs do not exhibit strong causal relationships, or relationships that differ between the high- and low-load conditions.

I also constructed DML-HMM models and implemented the SEM algorithm for structure and parameter learning. As in [10], hidden nodes were assumed to be discrete and binary-valued. Arcs could exist between hidden states only. As shown in Figure 3, the resulting structures for both the low-load and high-load conditions differ only in one column, and again, classification accuracy did not exceed chance.

(a)

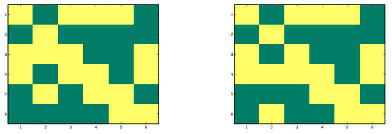


Fig. 3. (a) DML-HMM-SEM adjacency matrix for the high-load condition. (b) DML-HMM-SEM adjacency matrix for the low-load condition. Low classification accuracy implies that these structures are not reflective of differences in neural connectivity between conditions.

Future work could involve selecting different sets of ROIs, setting up different priors in the CPDs based on existing neural connectivity models, and letting each observation take continuous (rather than quantized) values whose densities are modeled, say, as mixtures of Gaussians.

Acknowledgements

Many thanks to Professor Vinod Menon (Dept. of Psychiatry & Behavioral Sciences, Dept. of Neuroscience) for the datasets, fast computers, and helpful discussions about fMRI!

References

[1] Tom M. Mitchell, Rebecca Hutchinson, Radu Stefan Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman, "Learning to decode cognitive states from brain images.," *Machine Learning*, vol. 57, no. 1-2, pp. 145–175, 2004.

[2] David D. Cox and Robert L. Savoy, "Functional magnetic resonance imaging (fmri) "brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex.," *NeuroImage*, vol. 19, no. 2, pp. 261–270, 2003.

[3] C. Davatzikos, K. Ruparel, Y. Fan, and D.G. Shen, "Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection," *NeuroImage*, vol. 28, no. 3, pp. 663–668, 2005.

[4] *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*. IEEE Computer Society, 2005.

[5] K. Friston, A.P. Holmes, J-B Poline, P.J. Grasby, S.C.R Williams, R.S.J. Frackowiak, and R. Turner, "Analysis of fmri time-series revisited," *NeuroImage*, vol. 2, no. 1, pp. 45–53, 1995.

[6] J. Mourao-Miranda, A.L.W. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional mri data," *NeuroImage*, 2005.

[7] M. Brett, J-L Anton, R. Valabregue, and J-B Poline, "Region of interest analysis using an spm toolbox," *International Conferance on Functional Mapping of the Human Brain*, 2002.

[8] X. Wang, R. Hutchinson, and T.M. Mitchell, "Training fmri classifiers to detect cognitive states across multiple human subjects," *NIPS*, 2003.

[9] J. Burge, V.P. Clark, T. Lane, H. Link, and S. Qiu, "Evidence for altered neural networks in dementia," *Tech report, University of New Mexico*, vol. TR-CS-2004-28, 2004.

[10] Lei Zhang, Dimitris Samaras, Nelly Alia-Klein, Nora Volkow, and Rita Goldstein, "Modeling neuronal interactivity using dynamic bayesian networks," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. MIT Press, Cambridge, MA, 2006.

[11] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," *ICCV'03 Ninth IEEE International Conference on Computer Vision*, vol. 2, pp. 742, 2003.

[12] S. Liang S. Furhrman and R. Somogyi, "Reveal, a general reverse engineering algorithm for inference of genetic network architectures," *Pac Symp Biocomput.*, pp. 18–29, 1998.