Real-time functional MRI Classification of Brain States using Markov-SVM Hybrid Models: Peering inside the rt-fMRI black box.

Ariana Anderson ariana82@ucla.edu **Dianna Han** **Pamela K. Douglas**

Jennifer Bramen **Mark S. Cohen**

Abstract

Real-time functional MRI (rt-fMRI) methods provide the ability to predict and detect online changes in cognitive states. Applications require appropriate selection of features, preprocessing routines, and efficient computational models in order to be both practical to implement and deliver interpretable results. We predict video activity in nicotine-addicted subjects using both regional spatial averages and preconstructed independent component spatial maps we refer to as an "IC dictionary." We found that this dictionary predicted better than the anatomical summaries and was less sensitive to preprocessing steps. When prior state information was incorporated using hybrid SVM-Markov models, the online models were able to predict even more accurately in real-time whether an individual was viewing a video while either resisting or indulging in nicotine cravings. Collectively, this work proposes and evaluates models that could be used for biofeedback. The IC dictionary offered an interpretable feature set proposing functional networks responsible for cognitive activity. We explore what is inside the black box of real-time fMRI, and examine both the advantages and shortcomings when machine learning methods are applied to predict and interpret cognitive states in the real-time context.

1 Introduction

Functional MRI (fMRI) is a proven imaging technique to detect and characterize changes in cognitive states. Current technology and algorithms are fast enough to create reliable maps of the topography of brain activity in a fraction of a second, producing a research field knows as real-time fMRI (rt-fMRI) ([1],[2],[3],[4],[5]). In rt-fMRI, incoming fMRI signal is analyzed immediately, providing representations of underlying conditions or states and quality control ([6],[7]). Thus far, rt-fMRI has been applied to functional localization and biofeedback with some success ([8],[9],[10]). Functional localizers based on rt-fMRI have been used to collect high resolution maps of motor, language, and somatosensory areas and to allow detection and correction of motion and other artifacts during the scan ([11],[12],[13],[14],[15],[16]). As another important application field, biofeedback attempts to teach subjects how to modulate their neural activity through a brain-computer interface (BCI). Intuitively, detecting cognitive states with high accuracy and providing biofeedback rapidly are essential for such applications, and models that can perform training and predicting online are of specific interest and importance.

When biofeedback models are applied effectively to rt-fMRI data streams, neural feedback based BCI can enable closed loop self-modulation of neuronal activity ([17],[18],[19]). Through operant training and visual feedback cues, subjects have learned to modulate insular cortex activity ([20]), navigate through mazes ([21]), communicate desired motor movements ([22],[23]), and manage chronic pain ([24],[25]). Coupled with machine learning (ML), rt-fMRI may open up the possibility for new experimental design and theapeutics ([18]), particularly for medication refactory conditions.

All of these, and many other findings ([19],[26]), have led to much excitement in the neuroimaging community. The challenges of applying ML algorithms to rt-fMRI are almost as great as the enthusiasm behind the research. The choice of features is ill-defined, because of both anatomical variations among subjects and inconsistency of functional activity across time known as nonstationarity. Many models that predict in realtime are actually trained offline because of the computational expense involved in selecting features, model training, model evaluation, and online prediction all within a typical TR of 2 seconds. Algorithms may be difficult to interpret because they are not tailored to the structure (temporal by spatial) of fMRI data or the unique challenges it poses, such as nonstationarity, temporal autocorrelations, or the spatial correlations among voxels that are known to exist. Thus, a successful online rt-fMRI model would be resilient to nonstationarity and signal drift, use a predetermined interpretable feature set, and harness the known autocorrelations to increase predictive accuracy.

1.1 fMRI data challenges

fMRI data is serially autocorrelated due to the hemodynamic response function (HRF) and to noise, which can result from electrical interference, movement, or even from cardiac or respiratory functions ([27],[28],[29]). In addition, it is reasonable to assume that the cognitive states are themselves correlated, with the immediately previous state affecting the present. Although there is strong temporal covariance within fMRI data, few attempts have been made to utilize the information encoded in their temporal patterns for prediction. SVM, linear discriminant analysis, and naïve Bayes are frequent choices of classification machines ([30],[31]), yet permuting the observations' order does not change the final models, as they make no assumptions on the covariance structure of the observations orderings. This may be a suboptimal approach to both offline and real-time classification; by treating all states as mutually independent, current fMRI models may be omitting valuable information that could aid in classification.

The family of Markovian models are known for their power in modeling spatial and sequential data. They generally assume that the current state depends only on the most recent history (the previous state) or neighborhood characteristics. They have been applied in the past to offline (but not realtime) fMRI data analysis, both in image segmentation and in state modeling. Markov Random Field (MRF) theory has been typically used to process and analyze fMRI data ([32],[33],[34]). Hidden Markov Model (HMM) analyses are also employed for fMRI activation detection, including voxel-based modeling ([35]). Woolrich, et al., used Bayesian inference including a Markov Chain Monte Carlo (MCMC) sampling technique to extend modeling to group analysis [36]. However, all these methods have not yet been applied to rt-fMRI analysis. We wish to harness the memory property of these models to inform the regular machine learning models, in the anticipation that capturing the structure contained in the serial autocorrelations among states will provide additional information for classification.

The observed fMRI signal is known to drift both temporally and spatially due to physiological changes, subject movements, neuronal plasticity and instrument stability problems ([36],[26],[19]). Models trained offline yet tested online are particularly sensitive to signal drift and nonstationarity, since the incoming data may bear little resemblance to the data they were trained to recognize. In practice, such drifts are even more difficult to accommodate when biofeedback is used to alter the subjects' response during the scan, because the model changes the subjects' response while simultaneously predicting it. Moreover, nonstationarity makes interpretation of features and their weights exceedingly difficult over time, as it is nontrivial to decouple drift induced by neural feedback from drift inherent in measurement of time-varying cognitive processes.

This issue is typically dealt with both by detrending the data and by selecting model-training windows. LaConte found that in order to achieve acceptable accuracy across scans it was necessary to both detrend the fMRI signal *and* linearly detrend the output of the SVM classifier. ([22]). Although linear detrending is simple in binary classification, it becomes computationally infeasible with multivariate outcomes having high-dimensional partitions between states. Online models that use sliding windows or weighted time averages to train on a small portion of the past history and predict incoming observations are less sensitive to nonstationarity inherent to fMRI data, but the statistical power of these models is reduced. Training time typically requires multiple minutes, and the time required to generate feedback typically takes many seconds ([20]), partly due to the large number of possible features available in a three-dimensional image.

Previously, features used for rt-fMRI models have been limited largely to GLM-based approaches and hypothesis-driven ROI (Region of Interest) based analyses where the mean signal in one or more ROIs is used to predict a state or condition. However, the interpretation of these features and their arbitrary scaling can be problematic, as well as the computational time involved in evaluating each of their potential for classification. Operator choices include: which regions to select, their summary statistics, and how these regions should be scaled ([19]). As multiple areas can operate in cognitive processing, the question arises as to how these individual regions should be combined and weighted into networks. In addition, the interpretation of ROI-based methods offer little insight into hidden cognitive representations, as the signal fluctuations within a single ROI may be the result of the influence of multiple underlying brain networks or cognitive states. If several asynchronous and independent networks involve the same single region, it is possible for this region to appear inactive regardless of its underlying activity.

Unsupervised learning methods, especially Independent Component Analysis (ICA) have been used extensively to address the problem of network activity in fMRI ([37],[38]) in the attempt to extract spatial features that co-vary in time. The spatial features are constrained to have statistically independent time-courses, allowing feature maps to operate on the same region with independent activities. These methods have been adapted to rt-fMRI data in a sliding-window to identify components and time-courses associated with specific activities ([39]). DeMartino, et al., have developed an approach that brings the temporal structure of the ICs into the alignment process ([40]). Specifically, they form an IC fingerprint whose dimensions helps to characterize the IC by temporal features. Anderson demonstrated the creation of an IC Dictionary on a larger scale by performing bootstrapped clustering of 21,256 ICs pooled over 279 scans taken from 51 subjects performing a video-craving task ([41]). These feature maps were taken to be representative of the latent cognitive processes operating while nicotine-addicted subjects were watching videos designed to induce cravings, forming a sparse feature space by which to represent the different states.

1.2 Model proposal

Although machine learning methods have demonstrated impressive power to classify fMRI data, there is still a strong need to balance the power of mathematical models with their neuroscientific interpretability. Currently, rt-fMRI models are hindered by data drift and an overwhelming set of possible features, limiting their ability to predict in real-time for biofeedback. Models may not be capitalizing on all available information by essentially ignoring the existing autocorrelations in the data. In fact, models that offer exceptional classification accuracy as a black box may provide little benefit and little insight for understanding underlying cognitive state changes by selecting features to optimize classification accuracy, without regard to interpretability. The adoption of multivariate pattern analysis (MVPA) for both feature selection and data modeling has to be bounded by model interpretability while harnessing all available information in the data.

Building on these findings, we create and evaluate a set of rt-fMRI models that collectively evaluate: 1.) The extent of nonstationarity within fMRI data, and how preprocessing steps such as demeaning affect the ability to classify cognitive states within and across scans from the same subject in online and offline models. We also assess feature choices (blindly-nominated ROIs vs. *a priori* defined IC maps) on both classification accuracies and model interpretations, and 2.) How incorporating prior fMRI states in the form of a Markov transition matrix can inform and update the SVM models' class likelihoods. Our objective is both to evaluate the effectiveness of various classification models and to identify which systems are most responsible for discriminating during real-time classification. Collectively, these methods ask and answer questions important to real-time classification, namely the impact of nonstationarity on both the model learning and interpretation, and whether using *a priori* information in the form of IC templates or Markov state transitions can increase our understanding and identification of latent cognitive processes during real-time analysis. We explore the tradeoff between sophisticated MVPA methods and practical interpretability, and whether the black box algorithms that perform blindly feature selection and classification are in fact superior to models that use cognitive state-based features *a priori* defined.

2 Methods

2.1 Data

The dataset consisted of 51 subjects scanned pre- and post-treatment in a smoking-cessation study. Data were collected as the subjects viewed videos under three video conditions interspersed with resting periods and brief auditory stimuli that was unintentionally muffled. The video cues were passive viewing of cue-neutral videos, passive viewing of smoking provocation videos, and viewing after being instructed verbally to resist craving. The full experimental design along with the data collection procedure is presented in ([42]). The fMRI analysis followed a standard pipeline established in our lab using FSL ([43]). Preprocessing included motion correction using MCFLIRT; nonbrain removal using BET; slight spatial smoothing using a Gaussian kernel of FWHM 5mm; high pass temporal filtering with σ=50.0 s. Registration to high resolution and/or standard images was carried out using FLIRT.

2.2 Dictionary creation

Following the methods presented in ([41]), we created a dictionary of common ICs expressed as intensity topologies in the probabilistic atlas provided in FSL. These ICs are dimensions in our classification process. Single session ICA results from 279 scans were first aligned to a common atlas space, projected into a lower-dimension anatomical-based atlas space by averaging within ROIs specified by the Harvard-Oxford cortical and subcortical structural atlases ([44]), and then pooled together. These 20,000 ICs were clustered using bootstrapped k-means clustering to obtain a set of 20 template ICs, which were then back-projected into the full voxel space. These exemplars we refer to as the IC dictionary, a set of 20 components possibly corresponding to the underlying functional networks present during and across the treatments and tasks. Examples appear in Figure 1.

2.3 Feature extraction

We used two sets of features and compared their effectiveness: the IC dictionary and the ROI summaries. For the ROI summaries, each time point in a scan a volume was reduced into a 110-dimensional feature by averaging the signal within each region specified by the Harvard-Oxford atlas ([44]). For the feature of IC-functional correlations, the correlation r^2 of a functional volume V_t at time t with one of 20 reference ICs is used to create a feature vector $\overrightarrow{x_t} = \Big(r^2(IC_1, V_t), r^2(IC_2, V_t), \ldots, r^2(IC_{20}, V_t)\Big)$ The entire volume V_t is then compressed into a 20-dimensional vector, where each element of the vector expressed the relative contribution of each IC-Dictionary element to the subjects activity at that time point. Support for the IC-functional correlation model comes from evidence that the found ICs themselves align well with functionally identifiable brain networks such as motor control, memory and executive function ([45]).

Our objective is to learn the model g that optimally maps the observed feature vector $\overrightarrow{x_t}$ at time t to the set of N possible cognitive states $\mathcal{C}$, or $g : \overrightarrow{x_t} \to \mathcal{C}$. We evaluated classifier and model drift by determining whether demeaning the data (within each feature) aided the classification accuracy for using ROIs and the ICs as features.

2.4 Models

We evaluated nonstationarity by investigating: 1.) the effect of demeaning within each feature set, and 2.) the differences between the models trained within a scan (online) and the models trained across scans (offline). The classifiers were variations on Support Vector Machines (SVM) ([46]), which seeks to find a hyperplane that separates training data into positive and negative classes (it is straightforward to extend this criteria to multiple classes). Markov transition matrices were added and omitted to the online and offline models to evaluate the effect of adding the state transitions on the overall classification accuracy ([47]).

These models were used to predict four different encodings of the stimulus, and the average accuracy across encodings was used to evaluate the model strength. The original task consisted of a sequence of visual and auditory stimuli interspersed with rest periods (the audio stimulus was muffled unintentionally); the video stimuli were three different movies intended to create states of

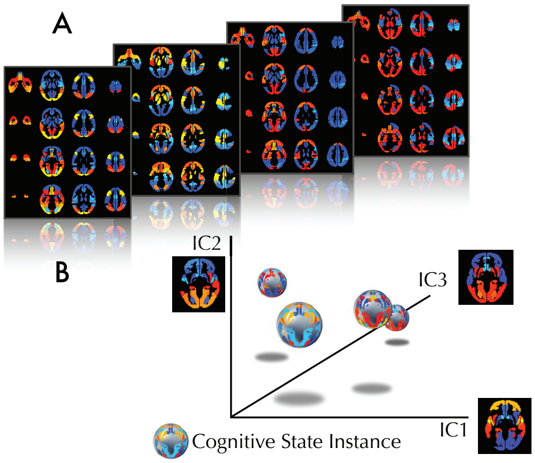

Figure 1: Representative spatial topologies of four of the 20 discovered dictionary ICs (independent components). B) Conceptual framework: Any given cognitive state is modeled as a point in a 20 dimensional feature space defined by the dictionary ICs.

crave, crave-resist, and crave-neutral. Using the known timing ([42]), we coded the response variable for the models in four different ways, giving the classifier successively more complicated states to distinguish among. These tasks were: 1.) Video/Audio, where just the portions of video and audio were classified. 2.) Task/Rest, where Video and Audio were coded identically as a generic task. 3.) Video/Audio/Rest, where each condition was coded separately. 4.) Audio On/Rest/Video Crave/Video Resist/Video Neutral, where the models had to predict membership of five states.

If we use $C_{t,i}$ to denote the cognitive state i of a system $C \in \mathcal{C}$ at the time point t, the system states will form a discrete-time Markov chain with transition matrix $\mathbf{A}$ if for any states $\{j, i, i_{t-1}, \dots, i_0\}$, $P(C_{t+1} = j \big| C_t = i, C_{t-1} = i_{t-1}, \dots, C_0 = i_0) = P(C_{t+1} = j \big| C_t = i)$. The rows $\overrightarrow{a_i}$ of the transition matrix $\mathbf{A}$ contains the transition probabilities to all possible states $j \in 1, \dots, N$ given the previous state i. Each element $a_{i,n} \in \overrightarrow{a_i}$ gives the probability of transitioning to state $n \in N$ given the previous state i.

Combining the SVM (radial basis kernal with $\gamma = \frac{1}{20}$) and the Markovian dependency, we applied four models to our dataset:
Model A: SVM Online trains an online SVM model g on the data from time $(1, t-1)$ and tests it on the data at time t. For N possible states at time t, the SVM model g outputs the current likelihood of each state c_t given the previously observed data, such that $\overrightarrow{c_t} = g(\overrightarrow{x_t} | \overrightarrow{x_{t-1}}, \overrightarrow{x_{t-2}}, \dots, \overrightarrow{x_0})$, where $\overrightarrow{c_t} = \big(p(C_{t,1}), p(C_{t,2}), \dots, p(C_{t,N})\big)$. This requires the model to be updated at every time point, but the computational cost is negligible because of the low number of explanatory variables (either 20 or 110, depending on the features selected).

Model B: SVM Markov Online updates the SVM class probabilities $\overrightarrow{c_t}$ using a Markov transition matrix. It estimates a transition matrix $\mathbf{A}$ at every time point given the history of the process. The predicted class label C_j at a time point t given the current state i is decided by $C_j = \max_{n \in N} \left\{ a_{i,n} \overrightarrow{c_t} \right\}$. This is a variation of a model presented by ([47]).

Model C: SVM Test trains an SVM model offline, g, and tests it online during a new scan from the same subject, pre-treatment.

Model D: SVM Markov Test creates a model offline, g, using a training scan, and tests it online using the testing scan. The offline model updates the SVM probabilities with the Markov transition matrix, $\mathbf{A}$, also learned from the training data.

3 Results

There were 64 different classifiers depending on which model was selected, how the response variables were encoded, the choice of features (ROI vs IC Dictionary), and whether or not demeaning was used. By averaging across options, we obtain with high certainty an understanding of how changing each part affects the classification accuracy as a whole. We discovered the ROI-based models were impacted by demeaning and training (online or offline), but IC dictionary models did not have substantial changes in accuracy based on these changes. Including the temporal information using a Markov transition matrix increased the predictive accuracy by roughly 23%. This varied little regardless of the feature choice (ROI or IC) and the training choice (online or offline). The average accuracy over all possible response encodings are shown in Figure 2, where the average chance accuracy is 52.1%.

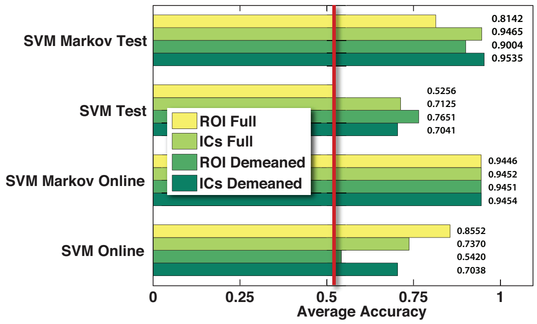

Figure 2: Accuracy by Parameters, compared to Chance

For the ROI feature, demeaning increased classification accuracy for the SVM-Test Model (offline model) by nearly 24%, but decreased the classification accuracy for SVM-Online by 31.3% averaged over all conditions. However, demeaning had little impact on the predicted accuracy when the IC feature was used. Although classification accuracy was slightly better (4%) for all online models, this difference was contingent again on whether ROIs were selected as features, and whether demeaning was performed. This difference again varied more for the ROI feature than for the IC feature, and was also influenced by initial feature demeaning. The online model (SVM-Online) was superior by 33% for predicting on ROI Full features, yet the offline model (SVM Test) was superior for predicting on ROI Demeaned features by 22%. Collectively, these results show that the decision of training offline or online and the choice of demeaning features or not both have much more impact when using ROIs as features than using ICs, showing that IC-features produce more stable models. Incorporating Markov transition matrices increased the accuracy for all models.

We argue that the ICs are plausible cognitive states because known visual and spatial networks are present. ICs have higher intensity in regions like the calcarine sulcus and cuneus (known to be involved in vision) and ventral striatum and medial orbito-frontal gyrus (known to be involved in craving). [41] compared results from the model free IC dictionary creation to those obtained using the GLM and found overlap between the two results. Further, as expected, regional activity was spread out over multiple ICs rather than clustering into one IC for vision and another for craving, suggesting the ICs are able to separate sub-networks in a way the simple GLM analysis could not.

4 Discussion

Collectively, these results show that using IC features are less sensitive to model training options (offline *versus* online) and preprocessing steps (feature demeaning *versus* no demeaning). These IC dictionary models were only strengthened when incorporating the covariance structure among states, using a Markov transition matrix.

We emphasize again that a mathematical model that allows little insight into actual neural processes provides little value for neuroscience. The SVMs can produce importance values that represent the relative contributions of various features to the overall classification accuracy, by simply permuting elements within each feature ([48]). Although the weighting of each feature is a relatively simple calculation, the features themselves are more difficult to interpret if ROIs are used instead of the ICs. With 110 ROIs as possible features, the calculations and rankings become cluttered when determining importance. Even more severe though is the inability to interpret ROIs individually. As most cognitive states are recognized to be the contribution of multiple regions acting collectively, and regions often are involved in multiple tasks/functional networks, discovering that a given region is significant in classification gives little interpretability into what the underlying processes are.

The substantial improvement in classification accuracy when incorporating a Markov transition matrix can be attributed to the fact that the transition matrix eliminates certain kinds of errors: in a blocked design the transition probability between many states becomes zero. In other words, the Markov transition matrix effectively removed unlikely state transitions. For example, when the previous state is a video cue it is almost impossible for the current state to be auditory. Intuitively, the transition matrix acts as a high-pass filter for the SVM probability predictions, sharpening the probability of some output classes and diminishing it for others. Further analyses may focus on the problem of separability: can induced drift be distinguished from inherent physiologic drift? Modeling drift that occurs as a result of neural feedback may be useful in assessing the efficacy of neural feedback based therapy, and plasticity that may occur as a result.

Using IC features results in much more stable, robust predictions over time. This leads to an even deeper realization about some intrinsic properties of fMRI data: they are fluid and nonstationary; and transforming them with respect to a reference set, in this case, the IC dictionaries, helps to ground them. This is similar to using triangulation to measure an object; by having a grounded reference point, the certainty in the final measurement increases. ROI boundaries typically are selected based on anatomical or cytoarchitectonic features ([49]). As such, considerable functional inhomogeneity can be present within a given ROI ([50]), and slight spatial perturbations in seed based analysis can generate significant changes in connectivity ([51]). By contrast, IC spatial maps are nominated functionally by the data themselves over larger brain areas than a single ROI, making ICs resistant to local measurement changes that might occur due to spatial drift.

Considerable evidence supports the notion that the ICs themselves represent meaningful functional systems in the brain. For example, there is considerable stability of discovered ICs across individuals ([52]) and, most significantly, the spatial signatures of these ICs align well with previously reported patterns seen in conventional activation studies. A single ROI may provide “mingled” but non-deterministic information about the cognitive states. For example, the hippocampus has important roles in explicit memory encoding ([53],[54]), but it does not perform such tasks in isolation: memory encoding also relies on neighboring medial temporal lobe structures and the prefrontal cortex ([55]). Combining ROIs to create plausible networks is itself a multivariate problem, and performing MVPA methods to determine importance within a model is ill-defined. Because of this, we believe that defining the features beforehand using the IC dictionary offers a substantial advantage not just for constructing models, but for interpreting them. The face credibility arises from the fact that such networks do, indeed, co-occur in more complex activities (e.g., spatial working memory). This concept is in any event much more plausible than assigning unique functions to individual brain regions.

We are aware that the models presented here have several limitations. Our IC dictionary may be sensitive to the choice of atlas. Although the Harvard-Oxford atlas we used is well-accepted, it is not the only one available, and it is based upon structural instead of functional architecture. The number of features used here (20) is also a parameter that was not investigated fully: as there were thousands of ICs initially available, we could have constructed up to thousands of features to represent the data. The ICs we used as features were dependent upon the data from which they were constructed. Although we used ICs from a larger set of craving-related data, there is the possibility that the resulting ICs may be sensitive to task performance and may not provide as much utility if the task were instead changed to, for example, a memory task. Our particular model for cognitive states is explicitly linear. This is, of course, naïve, not only because the weighting of a given IC on a state may itself be nonlinear, but also because we would expect significantly more complicated interactions may exist among functional networks in the brain. For example, in a multi-sensory environment, it is well established that auditory stimuli can effect visual perception ([56]), and *vice versa*. The linear SVM is poorly equipped to model such effects. Although we used a Markov transition matrix to incorporate temporal dependencies in the response, this does not explicitly capture the known hemodynamic response and temporal dependencies in the features but instead capitalizes on the information in the response, or the state being observed. It is possible that modeling these intrinsic temporal properties more accurately may remove much of the noise that exists in the fMRI signal, thus improving classification accuracy. It is also likely that this approach is sensitive to the experimental design used; for blocked-design, the transitions between states are clustered. With event related designs though, the probability between states are likely to be more evenly distributed, causing the Markov matrix to be a low-pass filter that would even out the probabilities estimated by the SVM model. This effect may be mitigated though both because of the HRF, which leaves a lingering effect of prior activity in current observations, and also because of the nature of cognitive states and their inability to change instantaneously. Because of these and other limitations, we present this analysis as a launching point for future work.

5 Conclusion

These results not only indicate that IC-functional correlation features have better statistical performance than an ROI-based analysis, but also point to a path of knowledge-based feature selection. Our IC dictionary is a lower dimensional feature space that respects the concept that the functional architecture of the brain prominently includes interactions among isolated regions. Though not a test of this hypothesis, they are supportive of it. Indeed, this is part of a larger community trend that moves us somewhat further from the purely localizationist neophrenological perspective, championed originally by Franz Gall in the late nineteenth century ([57]), towards the holonomic views of Karl Pribram who posited a largely dispersed representation of function in the neocortex ([58]).

Introducing a Markov matrix greatly increased the classification accuracy by taking advantage of the information contained in the experimental design itself. Additional support for incorporating Markov transitions is that cognitive state changes are unlikely to be random events. There is an underlying structure in them with autocorrelations among subsequent observations, and harnessing the information contained in these autocorrelations can improve classification accuracy. Because of nonstationarity within the actual data, the features themselves along with their correlations may change over time. The offline cross-validation error used to estimate the testing error was in fact biased, with as much as 30% difference between the accuracy predicted and the accuracy obtained. We believe that this bias is not caused by methodological errors, but rather by errors in the assumptions. SVM, along with most other machine learning models, assume that the relationships within and among the covariates do not depend on time. Rather, it views observations as a set of high-dimensional points embedded within a metric space. No assumptions are made on the covariance structure among the points, which can be a particular weakness when in fact the points exhibit a strong temporal dependency. We feel that incorporating state transitions may be an important advance in developing machine methods for real-time classification of brain states and more generally for the treatment of nonstationary data. By incorporating the temporal dependencies explicitly into the model, we are harnessing known structure to classify an unknown fluid outcome in real-time imaging.

Our results show that it is overall very difficult to perform accurate classification on fMRI data, both because of nonstationarity in the data and because of the difficulties in defining and interpreting features. Although we present ICs as informed features here, MVPA methods in general are capable of proposing features within the modeling process by selecting and weighting ROIs ([40],[59]). These methods, including ours, are sensitive to such choices as the atlas to be used and how these ROIs are to be combined. We advocate ICs as features though because their computational efficiency makes them more applicable for real-time feedback.

The choice of features, whether defined *a priori* or *post hoc*, needs to be made with the ultimate goal in mind: to construct models that not only classify with the maximum accuracy, but also allow researchers to glean input into the mechanics underlying the data. Although an interesting question to ask is always *how* the model performs, an even more exciting question is *why* it works. This question is best answered by examining all the materials that went into constructing it: the choice of data preprocessing steps, the very nature of the data, and the final selection of how to map the features to the responses. When viewed this way, peering into the black box of models can be more illuminating than observing what it produces.

Acknowledgments

This work is supported by funding under R33DA026109 to M.S.C.

References

[1] M. S. Cohen and R. M. Weisskoff. Ultra-fast imaging. *Magn Reson Imaging*, 9:1–37, 1991.

[2] R. W. Cox, A. Jesmanowicz, and J. S. Hyde. Real-time functional magnetic resonance imaging. *Magn Reson Med*, 33:230–236, Feb 1995.

[3] D. Gembris, J. G. Taylor, S. Schor, W. Frings, D. Suter, and S. Posse. Functional magnetic resonance imaging in real time (FIRE): sliding-window correlation analysis and reference-vector optimization. *Magn Reson Med*, 43:259–268, Feb 2000.

[4] N. Weiskopf, R. Sitaram, O. Josephs, R. Veit, F. Scharnowski, R. Goebel, N. Birbaumer, R. Deichmann, and K. Mathiak. Real-time functional magnetic resonance imaging: methods and applications. *Magn Reson Imaging*, 25:989–1003, Jul 2007.

[5] A. R. Bleier, F. A. Jolesz, M. S. Cohen, R. M. Weisskoff, J. J. Dalcanton, N. Higuchi, D. A. Feinberg, B. R. Rosen, R. C. McKinstry, and S. G. Hushek. Real-time magnetic resonance imaging of laser heat deposition in tissue. *Magn Reson Med*, 21:132–137, Sep 1991.

[6] J. T. Voyvodic. Real-time fMRI paradigm control, physiology, and behavior combined with near real-time statistical analysis. *Neuroimage*, 10:91–106, Aug 1999.

[7] M.S. Cohen. Real-time functional magnetic resonance imaging. *Methods*, 25(2):201–20, 2001.

[8] N. H. Goddard, J. D. Cohen, W. F. Eddy, C. R. Genovese, D. C. Noll, and L. E. Nystrom. Online analysis of functional mri datasets on parallel platforms. *The Journal of Supercomputing*, 11:295–318, 1997.

[9] K. Grill-Spector, R. Sayres, and D. Ress. High-resolution imaging reveals highly selective nonface clusters in the fusiform face area. *Nat. Neurosci.*, 9:1177–1185, Sep 2006.

[10] W. Schneider, D. C. Noll, and J. D. Cohen. Functional topographic mapping of the cortical ribbon in human vision with conventional MRI scanners. *Nature*, 365:150–153, Sep 1993.

[11] T. Gasser, O. Ganslandt, E. Sandalcioglu, D. Stolke, R. Fahlbusch, and C. Nimsky. Intraoperative functional MRI: implementation and preliminary experience. *Neuroimage*, 26:685–693, Jul 2005.

[12] T. Gasser, E. Sandalcioglu, B. Schoch, E. Gizewski, M. Forsting, D. Stolke, and H. Wiedemayer. Functional magnetic resonance imaging in anesthetized patients: a relevant step toward real-time intraoperative functional neuroimaging. *Neurosurgery*, 57:94–99, Jul 2005.

[13] T. Gasser, A. Szelenyi, C. Senft, Y. Muragaki, I. E. Sandalcioglu, U. Sure, C. Nimsky, and V. Seifert. Intraoperative MRI and functional mapping. *Acta Neurochir. Suppl.*, 109:61–65, 2011.

[14] C. Schwindack, E. Siminotto, M. Meyer, A. McNamara, I. Marshall, J. M. Wardlaw, and I. R. Whittle. Real-time functional magnetic resonance imaging (rt-fMRI) in patients with brain tumours: preliminary findings using motor and language paradigms. *Br J Neurosurg*, 19:25–32, Feb 2005.

[15] D. T. Gering and D. M. Weber. Intraoperative, real-time, functional MRI. *J Magn Reson Imaging*, 8:254–257, 1998.

[16] M. Moller, M. Freund, C. Greiner, W. Schwindt, C. Gaus, and W. Heindel. Real time fMRI: a tool for the routine presurgical localisation of the motor cortex. *Eur Radiol*, 15:292–295, Feb 2005.

[17] X Hong M Rohan MS Cohen, R Terwilliger and P Roemer. Real-time observation of mental activity: the autocerebroscope., 1997.

[18] Christopher R. DeCharms. Applications of real-time fmri. *Nat Rev Neurosci*, 9(9):720–9, 2008.

[19] S. M. LaConte. Decoding fMRI brain states in real-time. *Neuroimage*, 56:440–454, May 2011.

[20] A. Caria, R. Veit, R. Sitaram, M. Lotze, N. Weiskopf, W. Grodd, and N. Birbaumer. Regulation of anterior insular cortex activity using real-time fMRI. *Neuroimage*, 35:1238–1246, Apr 2007.

[21] S-S. Yoo, T. Fairneny, N-K. Chen, S-E. Choo, L.P. Panych, H. Park, S-Y. Lee, and F.A. Jolesz. Brain-computer interface using fmri: Spatial navigation by thoughts. 15(10):1591–1595, 07 2004.

[22] S. M. LaConte, S. J. Peltier, and X. P. Hu. Real-time fMRI using brain-state classification. *Hum Brain Mapp*, 28:1033–1044, Oct 2007.

[23] Henrik Ohlsson, Joakim Rydell, Anders Brun, Jacob Roll, Mats Andersson, Anders Ynnerman, and Hans Knutsson. Enabling bio-feedback using real-time fmri.

[24] R.C. Decharms. Reading and controlling human brain activation using real-time functional magnetic resonance imaging. *Trends Cogn Sci*, 2007.

[25] R. C. deCharms, F. Maeda, G. H. Glover, D. Ludlow, J. M. Pauly, D. Soneji, J. D. Gabrieli, and S. C. Mackey. Control over brain activation and pain learned by using real-time functional MRI. *Proc. Natl. Acad. Sci. U.S.A.*, 102:18626–18631, Dec 2005.

[26] N. Weiskopf, R. Sitaram, O. Josephs, R. Veit, F. Scharnowski, R. Goebel, N. Birbaumer, R. Deichmann, and K. Mathiak. Real-time functional magnetic resonance imaging: methods and applications. *Magn Reson Imaging*, 2007.

[27] E. Zarahn, G. K. Aguirre, and M. DÉsposito. Empirical analyses of BOLD fMRI statistics. *NeuroImage*, 5:179–197, 1997.

[28] MS Cohen and RM DuBois. Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. *J Magn Reson Imaging*, 10:33–40, 1999.

[29] Martin M. Monti. Statistical analysis of fMRI time-series: A critical evaluation of the GLM approach. *Preprint submitted to Frontiers Special Topics*, 2006.

[30] F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fmri: a tutorial overview. *NeuroImage*, 45(1 Suppl), March 2009.

[31] P. K. Douglas, S. Harris, A. Yuille, and M. S. Cohen. Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. *Neuroimage*, 56:544–553, May 2011.

[32] W. Liu, P. Zhu, J. S. Anderson, D. Yurgelun-Todd, and P. T. Fletcher. Spatial regularization of functional connectivity using high-dimensional Markov random fields. *Med Image Comput Comput Assist Interv*, 13:363–370, 2010.

[33] Markus Svensén, Frithjof Kruggel, and D. Yves von Cramon. Markov random field modelling of fmri data using a mean field em-algorithm. In *Proceedings of the Second International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, EMMCVPR '99, pages 317–330, London, UK, 1999. Springer-Verlag.

[34] M.B. Nagori T.N. Mane S.A. Agrawal, M.S. Joshi. Evaluation of markov blanket algorithms for fmri data analysis. In *International Conference on Information and Network Technology*, 2011.

[35] Fangyuan Nan, Yaonan Wang, and Xiaoping Ma. fMRI Activation Detection by MultiScale Hidden Markov Model. In Sanguthevar Rajasekaran, editor, *Bioinformatics and Computational Biology*, volume 5462, chapter 28, pages 295–306. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[36] M. W. Woolrich, T. E. Behrens, C. F. Beckmann, M. Jenkinson, and S. M. Smith. Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage*, 21:1732–1747, Apr 2004.

[37] Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

[38] M. Mckeown, S. Makeig, G. Brown, T. Jung, S. Kindermann, A. Bell, and T. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components, 1998.

[39] F. Esposito, E. Seifritz, E. Formisano, R. Morrone, T. Scarabino, G. Tedeschi, S. Cirillo, R. Goebel, and F. Di Salle. Real-time independent component analysis of fMRI time-series. *Neuroimage*, 20:2209–2224, Dec 2003.

[40] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage*, 43:44–58, Oct 2008.

[41] Ariana Anderson, Jennifer Bramen, Pamela K. Douglas, Agatha Lenartowicz, Andrew Cho, Chris Culbertson, Arthur L. Brody, Alan L. Yuille, and Mark S. Cohen. Large sample group independent component analysis of functional magnetic resonance imaging using anatomical atlas-based reduction and bootstrapped clustering. *International Journal of Imaging Systems and Technology*, 21(2):223–231, 2011.

[42] A. L. Brody, M. A. Mandelkern, R. E. Olmstead, J. Jou, E. Tiongson, V. Allen, D. Scheibal, E. D. London, J. R. Monterosso, S. T. Tiffany, A. Korb, J. J. Gan, and M. S. Cohen. Neural substrates of resisting craving during cigarette cue exposure. *Biol. Psychiatry*, 62:642–651, Sep 2007.

[43] Stephen M. Smith, Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E. J. Behrens, Heidi Johansen-berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnjak, David E. Flitney, Rami K. Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J. Michael Brady, and Paul M. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23:208–219, 2004.

[44] R. S. Desikan, F. Segonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31:968–980, Jul 2006.

[45] Stephen M. Smith, Peter T. Fox, Karla L. Miller, David C. Glahn, P. Mickle Fox, Clare E. Mackay, Nicola Filippini, Kate E. Watkins, Roberto Toro, Angela R. Laird, and Christian F. Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*, 106(31):13040–13045, August 2009.

[46] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[47] Garczarek UM. Classification rules in standardized partition spaces. *Doctoral Dissertation: University of Dortmund*, 2002.

[48] Yin-Wen Chang and Chih-Jen Lin. Feature ranking using linear svm. *Journal of Machine Learning Research - Proceedings Track*, 3:53–64, 2008.

[49] J. A. Maldjian, P. J. Laurienti, R. A. Kraft, and J. H. Burdette. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage*, 19:1233–1239, Jul 2003.

[50] G. Marrelec and P. Fransson. Assessing the influence of different ROI selection strategies on functional connectivity analyses of fMRI data acquired during steady-state conditions. *PLoS ONE*, 6:e14788, 2011.

[51] D. S. Margulies, J. L. Vincent, C. Kelly, G. Lohmann, L. Q. Uddin, B. B. Biswal, A. Villringer, F. X. Castellanos, M. P. Milham, and M. Petrides. Precuneus shares intrinsic functional architecture in humans and monkeys. *Proc. Natl. Acad. Sci. U.S.A.*, 106:20069–20074, Nov 2009.

[52] F Barkhof P Scheltens CJ Stam SM Smith CF Beckmann JS Damoiseaux, SA Rombouts. Consistent resting-state networks across healthy subjects. *Proc National Academy of Science*, 2006.

[53] W. B. Scoville, B. Milner, W. B. Scoville, and B. Miller. Loss of recent memory after bilateral hippocampal lesions. 1957. *J Neuropsychiatry Clin Neurosci*, 12:103–113, 2000.

[54] A. D. Ekstrom, M. J. Kahana, J. B. Caplan, T. A. Fields, E. A. Isham, E. L. Newman, and I. Fried. Cellular networks underlying human spatial navigation. *Nature*, 425:184–188, Sep 2003.

[55] L. R. Squire and B. J. Knowlton. The medial temporal lobe, the hippocampus, and the memory systems of the brain. In *Memory*, pages 765–779.

[56] L. Shams, Y. Kamitani, and S. Shimojo. Illusions. What you see is what you hear. *Nature*, 408:788, Dec 2000.

[57] F.J. Gall. *Anatomie et physiologie du système nerveux en général et du cerveau en particulier: avec des observations sur la possibilité de reconnoitre plusieurs dispositions intellectuelles et morales de l'homme et des animaux*. chez N. Maze, libraire, 1819.

[58] K.H. Pribram. *Languages of the brain: experimental paradoxes and principles in neuropsychology*. Brandon House, 1981.

[59] N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U.S.A.*, 103:3863–3868, Mar 2006.