

Interpretable classifiers for FMRI improve prediction of purchases

Logan Grosenick, Stephanie Greer, Brian Knutson

Abstract—Despite growing interest in applying machine learning to neuroimaging analyses, few studies have gone beyond classifying sensory input to directly predicting behavioral output. With spatial resolution on the order of millimeters and temporal resolution on the order of seconds, functional magnetic resonance imaging (FMRI) is a promising technology for such applications. However, FMRI data’s low signal-to-noise ratio, high dimensionality, and extensive spatiotemporal correlations present formidable analytic challenges. Here, we apply different machine-learning algorithms to previously acquired data to examine the ability of FMRI activation in three regions – the nucleus accumbens (NAcc), medial prefrontal cortex (MPFC), and insula – to predict purchasing. Our goal was to improve spatiotemporal interpretability as well as classification accuracy. To this end, sparse penalized discriminant analysis (SPDA) enabled automatic selection of correlated variables, yielding interpretable models that generalized well to new data. Relative to logistic regression, linear discriminant analysis, and linear support vector machines, SPDA not only increased interpretability but also improved classification accuracy. SPDA promises to allow more precise inferences about when specific brain regions contribute to purchasing decisions. More broadly, this approach provides a general framework for using neuroimaging data to build interpretable models, including those that predict choice.

Index Terms—FMRI, prediction, classification, purchasing, accumbens, frontal, insula, human, single-trial, spatiotemporal, discriminant, sparse, lasso, elastic net, PDA, SVM

I. INTRODUCTION

THE development of event-related functional magnetic resonance imaging (FMRI) has revolutionized cognitive neuroscience. Currently, among neuroimaging techniques, only FMRI allows investigators to visualize changes in subcortical activity at a temporal resolution of seconds and at a spatial resolution of millimeters [1]. Using FMRI, investigators visualize changes in vascular oxygenation (hereafter, activation) that occur 4–6 seconds after changes in neural activity. These changes in activation correlate more closely with postsynaptic changes in dendritic potentials than with presynaptic changes in axonal firing rates [2]. Although the FMRI signal lags behind these postsynaptic changes, this lag can be modeled and deconvolved, affording second-to-second temporal inference. Nonetheless, many FMRI methods are derived from those previously developed for positron emission

tomography (PET) scanning, which has a minimum temporal resolution of 120 sec, and so have only recently begun to adapt in order to take advantage of the increased temporal resolution of FMRI.

Traditionally, subcortical regions have been of great interest to affective neuroscientists, since appetitive and aversive behavior can be unconditionally elicited from subcortical circuits via electrical stimulation [3]. A little more than a decade of FMRI research has begun to validate some of these findings in humans, suggesting that one subcortical circuit including the nucleus accumbens (NAcc) plays a role in anticipation of gains, while another circuit including the deep cortical region in the anterior insula plays a role in anticipation of loss [4]. Additionally, a region in the mesial prefrontal cortex (MPFC) appears to play a role in correcting inaccurate gain predictions [5]. Together, these findings implicate these evolutionarily conserved brain regions in the representation of expected value and subsequent choice [6].

The ability to visualize anticipatory activation reverses the traditional logic of neuroimaging design and analysis. Instead of simply examining how sensory input influences brain activation, investigators can potentially examine how brain activation influences subsequent motor output. Thus, beyond localization, researchers can now begin to answer novel questions about where and when brain activation predicts behavior. By temporally staggering information presentation prior to the point of choice, scientists can further attempt to determine whether different brain regions respond to different types of information (e.g., anticipation of gain, anticipation of loss), and whether this activation then contributes to subsequent choice.

Despite the possibility of using FMRI activation to directly predict choice, few studies have done so. The majority of classification studies have instead used FMRI activation to classify concurrent sensory input, such as the category of perceived visual stimuli [7]–[14], or more recently the identity of natural images [15]. Additional studies have used FMRI activation to classify lying vs. telling the truth [16], or recall of different object categories [17]. At present, only two FMRI studies have used classification models to predict choice. The first study used a simple logistic regression (LR) model to predict purchasing behavior with averaged data from bilateral NAcc, bilateral MPFC, and right insula [18] (data reanalyzed below). The second study used a linear discriminant analysis (LDA) to predict choice on the next trial of a reversal-learning task based on activation from nine regions in the previous trial, and found that a combination of NAcc, MPFC, and anterior cingulate activation best predicted the next choice [19]. While

Manuscript received November 18, 2008; revised April 25, 2008.

L. Grosenick is with the Neuroscience Institute at Stanford, Stanford University, Stanford California USA

S. Greer and B. Knutson are with the Psychology Department, Stanford University, Stanford California USA

Copyright (c) 2007 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

the classification models used in these studies (i.e., LR and LDA) can accurately predict upcoming choice, they do not yield interpretable coefficients when applied to data having a large number of correlated variables. These methods therefore yield little insight into the neural mechanisms underlying classifications made using spatially and temporally unaveraged fMRI data.

While classification models have been used to identify relevant brain regions in space, only one previous study has extended these models to identify relevant brain regions in time [12]. Specifically, researchers used a linear Support Vector Machine (SVM) to identify when brain regions discriminated between viewing positive and negative pictures. This SVM, however, did not eliminate unimportant coefficients, complicating interpretation. Here, we sought not only to identify but also to automatically select in space and time only those coefficients relevant to prediction of purchasing – both across as well as within subjects.

In principle, automatic variable selection should improve both model parsimony and interpretability. In practice, however, investigators must take care to ensure the stability and uniqueness of coefficients across samples when fitting correlated data. Traditionally, penalization (or regularization) has provided an effective means of stabilizing correlated coefficients [20]. More recently, related models such as the Least Absolute Shrinkage and Selection Operator (LASSO) have extended penalized linear regression to include automatic variable selection, which sets irrelevant coefficients to exactly zero (removing them from the model). Indeed, given certain conditions, LASSO models have been shown to possess the “oracle property” [21], in which they are guaranteed to asymptotically identify important coefficients while eliminating unimportant ones. Unfortunately, correlated inputs (which are prevalent in fMRI data) may violate the conditions under which this desirable property holds [22]. However, a generalization of the LASSO model called the Elastic Net (ENET) offers a promising alternative in such situations [23]. Interestingly, one parameterization of the ENET model is equivalent to Univariate Soft Thresholding (UST), which yields coefficients identical to a voxelwise thresholded univariate general linear model [24] – providing continuity with currently popular neuroimaging analyses. Importantly, all of these regression models (i.e., LASSO, ENET, and UST) can be converted to discriminant classifiers through the application of a method known as Optimal Scoring [25]. These models comprise instances of a larger class of Penalized Discriminant Analysis (PDA) classifiers [25]. Because these models include a penalty that drives small coefficients to zero, we refer to them as sparse PDA (or SPDA) models.

The goal of this paper was to introduce a new, interpretable method for single-trial classification of fMRI, and to apply it to prediction of purchases. Additionally, we explored whether preprocessing methods (i.e., spatial smoothing and temporal filtering) influenced results. We compared SPDA analyses (i.e., PDA-LASSO, PDA-ENET, and PDA-UST) with more standard analyses (i.e., logistic regression, linear discriminant analysis, and SVM), using previously collected data [18]. Overall, we found that SPDA models increased classification

rates, while markedly improving spatiotemporal interpretability.

II. DATA COLLECTION AND PREPROCESSING

Data from 25 healthy right-handed subjects were included in these analyses (one subject’s original fMRI data could not be recovered and so was omitted). Along with the typical magnetic resonance exclusions (e.g., metal in the body), subjects were screened for psychotropic drugs and ibuprofen, substance abuse in the past month, and history of psychiatric disorders (DSM IV Axis I) prior to collecting informed consent. Subjects were paid \$20.00 per hour for participating and also received \$40.00 in cash to spend on products. In addition to the 25 subjects who were included in the analysis, 6 subjects who purchased fewer than four items per session (i.e., < 10%) were excluded due to insufficient data to model, and 8 subjects who moved excessive amounts (i.e., > 2 mm between whole brain acquisitions) were excluded.

While being scanned, subjects participated in a “Save Holdings Or Purchase” (SHOP) Task. During each task trial, subjects saw a labeled product (product period; 4 sec), saw the product’s price (price period; 4 sec), and then chose either to purchase the product or not (by selecting either “yes” or “no” presented randomly on the right or left side of the screen; choice period; 4 sec), before fixating on a crosshair (2 sec) prior to the onset of the next trial (see Supplement 1 for illustration of the task layout).

Each of 80 trials featured a different product. Products were pre-selected to have above-median attractiveness, as rated by a similar sample in a pilot study. While products ranged in retail price from \$8.00-\$80.00, the associated prices that subjects saw in the scanner were discounted down to 25% of retail value to encourage purchasing. Therefore the cost of the products during the experiment ranged from \$2.00 to \$20.00. Consistent with pilot findings, this led subjects to purchase 30% of the products on average, generating sufficient instances of purchasing to model.

To ensure subjects’ engagement in the task, two trials were randomly selected after scanning to count “for real”. If subjects had chosen to purchase the product presented during the randomly selected trial, they paid the price that they had seen in the scanner from their \$40.00 endowment and were shipped the product within two weeks. If not, subjects kept their \$40.00 endowment. Based on these randomly drawn trials, seven of twenty-five subjects (28%) were actually shipped products.

Subjects were instructed in the task and tested for comprehension prior to entering the scanner. During scanning, subjects chose from 40 items twice and then chose from a second set of 40 items twice (80 items total), with each set in the same pseudorandom order. Subjects having to make a choice on the same item twice allowed us to explore the effects of item repetition on the relation of neural activity to choice (item sets were counterbalanced across subjects). After scanning, subjects rated each product in terms of how much they would like to own it and what percentage of the retail price they would be willing to pay for it. Then, two trials were randomly drawn to count “for real”, and subjects received the outcome of each of the drawn trials.

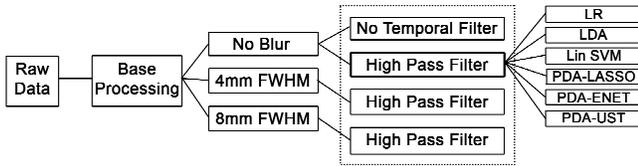


Fig. 1. Data preprocessing and analysis flowchart.

Functional images were acquired with a 1.5 T General Electric MRI scanner using a standard birdcage quadrature head coil. Twenty-four 4-mm-thick slices (in-plane resolution 3.75 X 3.75 mm, no gap) extended axially from the midpons to the top of the skull, providing whole brain coverage and adequate spatial resolution of subcortical regions of interest (e.g., midbrain, NAcc, OFC). Whole brain functional scans were acquired with a T2*-sensitive spiral in-/out- pulse sequence (TR=2 s, TE=40 ms, flip=90), which minimizes signal dropout at the base of the brain [26]. High-resolution structural scans were also acquired to facilitate localization and coregistration of functional data, using a T1-weighted spoiled grass sequence (TR=100 ms, TE=7 ms, flip=90).

After reconstruction, preprocessing was conducted using Analysis of Functional Neural Images (AFNI) software [27]. For all functional images, voxel time-series were sine interpolated to correct for nonsimultaneous slice acquisition within each volume, concatenated across runs, corrected for motion, and normalized to percent signal change with respect to the voxel mean for the entire task. Data not normalized to percent signal change were also run, but consistently resulted in very similar or slightly worse model performance and so are not considered further here. To compare different preprocessing algorithms applied to volume averages, data were submitted to varying levels of spatial smoothing (i.e., either 0 mm, 4 mm, or 8 mm full width at half-maximum Gaussian blur) and temporal filtering (i.e., either none or high pass filtering admitting frequencies < 90 sec). Four scan runs were averaged over each volume of interest (VOI) and submitted to the same logistic regression format used to predict purchasing in a previous report [18] (Figure 1). Subsequently unaveraged (voxel-wise), high-pass filtered data with no blur were submitted to spatiotemporal classification (Figure 1).

Spatiotemporal data were arranged as in previous spatiotemporal analyses [12], but extracted from predefined regions of interest rather than whole-brain volumes. Specifically, data was arranged as an $N \times p$ data matrix \mathbf{X} with N corresponding to the number of trial observations on the p input variables, each of which was a particular voxel at a particular time point. This yielded 16 voxels each for bilateral NAcc and MPFC, and 14 voxels for bilateral insula, all taken at 9 time points each taken every 2 seconds, for a total of 414 input variables per trial. Fixed effects then added 24 additional dummy-coded variables, resulting in a total of 438 input variables (Talairach coordinates listed in Supplement 2). Data were sub-sampled once by alternatively choosing purchase and not purchase trials at random without replacement until no trials remained in one of one classes and the number of sampled trials in

each class was equal. Altogether, these data included 1118 trials for the first presentation dataset, and 1094 trials for the second presentation dataset, yielding a total of 2212 trials for the combined first and second presentation datasets across subjects.

III. PENALIZED DISCRIMINANT ANALYSIS FRAMEWORK

FMRI data is high-dimensional and can have strong correlations in both space and time. As a result, selection of appropriate models for classifying FMRI data requires careful consideration of how different models handle correlated data in many dimensions. Application of standard logistic regression (LR) or linear discriminant analysis (LDA) to FMRI data may suffer from degenerate sample covariance matrices, which can potentially limit both generalizability to new test data (increasing classification error) and degrade coefficient interpretability [25]. Appropriate penalization of the covariance matrix, however, can improve generalizability and yield more interpretable models [25], [28]. Further, some form of automatic variable selection – in which the process of fitting the model selects an optimal subset of the variables – is desirable given the large number of input variables. Such variable selection has been shown to improve both model generalization to new data and model interpretability.

Modern regression tools exist for simultaneous penalization and automatic variable selection, but must be modified to be appropriate for binary classification. With the ‘Optimal Scoring’ procedure [25], [29], a function can be estimated that converts the continuous output of any regression method into binary (or n-ary) classes. This procedure can be used to modify penalized regression models to classify categorical output variables (e.g., the decision whether or not to purchase a product). When applied generally to penalized regressions, this approach is called Penalized Discriminant Analysis (PDA) [25]. Here, we focus on “sparse” PDAs (SPDAs) – where the penalized regressions include an L_1 (“Laplacian”) penalty on the coefficients that sets unimportant coefficients to exactly zero. These should generate parsimonious, interpretable sets of model coefficients that yield insight into which data in time and space contribute to choice on a trial-by-trial basis.

In general, the penalized linear regression models we consider have coefficient estimates given by (in “Lagrange” form):

$$\hat{\beta} = \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}^T \beta\|_2^2 + \lambda J(\beta) \quad (1)$$

where \mathbf{y} is a real-valued vector of outputs (dependent variables), \mathbf{X}^T is the transpose of our input variable matrix (independent variables), β is the vector of coefficients and $\hat{\beta}$ the coefficient estimates, the function $J(\beta)$ is some penalty function in terms of the model coefficients β_j ; $j \in \{1, \dots, p\}$, λ is a penalty parameter, and $\|\mathbf{a}\|_2 = \sqrt{\sum_i a_i^2}$ denotes the L_2 (“Euclidean”) norm of vector \mathbf{a} .

The Optimal Scoring procedure modifies a regression model with continuous-valued outputs so that it can classify a vector of categorical outputs \mathbf{g} by simultaneously optimizing over a function $\theta(\mathbf{g}) : \mathbf{g} \rightarrow \mathbb{R}^{1 \times N}$. This function converts a vector of categorical values (e.g. 0’s and 1’s) to a vector of real values.

Given such a function, equation (1) can be altered to:

$$\hat{\beta} = \arg \min_{\theta, \beta} \|\theta(\mathbf{g}) - \mathbf{X}^T \beta\|_2^2 + \lambda J(\beta) \quad (2)$$

minimized under the constraint $N^{-1}\|\theta(\mathbf{g})\|_2^2 = 1$. As $\theta(\mathbf{g})$ and β can be found separately, standard methods for fitting a particular regression model can then be applied to appropriately transformed data (see [25]). This can be implemented for a binary classification as follows: (1) construct the dummy-coded $N \times 2$ indicator matrix \mathbf{Y} with 1's in the columns corresponding to 1's in the original output and 0's otherwise; (2) choose a 2×2 initial scoring matrix Θ_0 satisfying the constraints $\Theta_0^T \mathbf{D}_p \Theta_0 = \mathbf{I}$, where $\mathbf{D}_p = \mathbf{Y}^T \mathbf{Y} / N$, and let $\Theta_0^* = \mathbf{Y} \Theta_0$; (3) fit a multi-response regression model of your choosing on Θ_0^* , yielding fitted $N \times 2$ matrix $\hat{\Theta}_0^*$ and the vector of fitted regression values $\eta(x)$; (4) obtain the eigenvector matrix Φ of $\Theta_0^{*T} \hat{\Theta}_0^*$ and thus the optimal scores $\Theta = \Theta_0 \Phi$. Then the final model is $\eta_f(x) = \Phi^T \eta(x)$. This procedure is only slightly different for larger numbers of classes and is motivated and explained in [25], [29].

Since the Optimal Scoring procedure transforms regression models into discriminant classifiers, and since discriminant classifiers can be written in terms of a combined regression and Optimal Scoring procedure, this framework allows extension of concepts from regression (e.g., degrees of freedom) to discriminant classification, which in turn allows model comparison using various goodness-of-fit criteria (e.g., Akaike Information Criterion (AIC) [30], Bayesian Information Criterion (BIC) [31], Mallor's Cp [32]) while preserving the ability to visualize data as discriminant coordinates.

To allow both penalization and automatic variable selection, a reasonable choice for creating a sparse PDA is the LASSO [33], which uses the penalty function $J(\beta) = \|\beta\|_1$ in equations (1) and (2) above, where $\|\beta\|_1 = \sum_i |\beta_i|$ is the L_1 ("Laplacian") norm of coefficient vector β . When the number of non-zero coefficients in the model is expected to be sparse (e.g., N for $p \gg N$), the LASSO provides an attractive option, since it performs simultaneous variable subset selection and prediction, and is easily computed using the LARS algorithm [34]. The LASSO has also been shown to perform well at prediction in many problems, competing favorably with ridge regression (where $J(\beta) = \|\beta\|_2^2$) and the more general "bridge regression" (where $J(\beta) = \sum_i |\beta_i|^\gamma$; $\gamma \geq 0$) [35].

Although the LASSO performs well in variable selection and prediction, it also has limitations, particularly given sufficiently correlated input variables [22], or if the number of observations N is small relative to the number of input variables p (spatiotemporal fMRI data suffers from both problems) [23]. Specifically, the LASSO can select at most N variables when $N < p$, and is not well-defined unless the bound on the L_1 -norm of the coefficients is below a certain value [23]. Given a group of highly correlated input variables, the LASSO is likely to arbitrarily select just one variable from the group, generating unstable coefficients over different samplings of the same variables and failing to capture correlated groups of relevant variables [23]. Model performance also suffers given correlated inputs. For instance, ridge regression empirically dominates the LASSO in typical $N > p$ regression

settings when the input variables are correlated [33]. Further, LASSO loses its desirable 'oracle property' [21] (the ability to asymptotically choose only relevant input variables) when the input variables are sufficiently correlated [22].

Because of these problems related to correlated input variables, the LASSO may not always be best suited for the analysis of spatiotemporal fMRI data. A generalization of the LASSO called the elastic net (ENET) addresses correlations between inputs by implementing a hybrid penalty with both ridge (L_2) and LASSO (L_1) properties [23]. ENET coefficient estimates are given by:

$$\hat{\beta}^{ENET} = \sqrt{(1 + \lambda_2)} \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}^T \beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \quad (3)$$

and thus implement a hybrid penalty involving two penalty parameters λ_1 and λ_2 , with the former essentially modulating automatic variable selection while the latter allows relevant but correlated variables to remain together in the model. As described in [23] Theorem 2, the estimates $\hat{\beta}^{ENET}$ can be rewritten as:

$$\hat{\beta}^{ENET} = \arg \min_{\beta} \beta^T \left(\frac{\mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{I}}{\sqrt{1 + \lambda_2}} \right) \beta - 2\mathbf{y}^T \mathbf{X} \beta + \lambda_1 \|\beta\|_1 \quad (4)$$

where standard LASSO estimates obtain when $\lambda_2 = 0$:

$$\hat{\beta}^{LASSO} = \arg \min_{\beta} \beta^T (\mathbf{X}^T \mathbf{X}) \beta - 2\mathbf{y}^T \mathbf{X} \beta + \lambda_1 \|\beta\|_1 \quad (5)$$

Thus, ENET represents a stabilized version of the LASSO, in which the estimated covariance matrix $\hat{\Sigma} = \mathbf{X}^T \mathbf{X}$ is shrunk towards the $p \times p$ identity matrix \mathbf{I} as λ_2 increases, since the stabilized sample covariance matrix in equation (4) can be written:

$$\frac{\mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{I}}{\sqrt{1 + \lambda_2}} = (1 - \gamma) \hat{\Sigma} + \gamma \mathbf{I} \quad (6)$$

(with $\gamma = \lambda_2 / (1 + \lambda_2)$) [23].

Conversely, letting $\lambda_2 \rightarrow +\infty$ (and thus $\gamma \rightarrow 1$), creates a special case of ENET which yields "Univariate Soft Thresholding" (UST) coefficients [36]:

$$\hat{\beta}^{UST} = \arg \min_{\beta} \beta^T \beta - 2\mathbf{y}^T \mathbf{X} \beta + \lambda_1 \|\beta\|_1 \quad (7)$$

which can be equivalently written as:

$$\hat{\beta}_j^{UST} = \left(|\mathbf{x}_j^T \mathbf{y}| - \frac{\lambda_1}{2} \right)_+ \text{sign}(\mathbf{x}_j^T \mathbf{y}) \quad (8)$$

for $j \in \{1, \dots, p\}$. where $(\cdot)_+$ denotes taking only the positive part of the quantity in parentheses, and where $\text{sign}(\cdot)$ yields +1 for positive values, -1 for negative values, and 0 otherwise. These estimates have particular relevance to fMRI analysis, since the values $\mathbf{x}_j^T \mathbf{y}$ are simply the univariate linear coefficient estimates, which are then thresholded at the value $\lambda_1/2$ chosen from the data via cross-validation, below which they are set to zero. In other words, UST coefficients are equivalent to a mass-univariate general linear model map with a data-driven threshold. This equivalence directly links the coefficients for the family of ENET methods parameterized by (λ_1, λ_2) to mass-univariate statistical maps currently popular in most fMRI analyses.

The current investigation compared PDA-LASSO estimates $\hat{\beta}^{PDA-ENET}(\lambda_1, 0)$ and approximate PDA-UST estimates $\hat{\beta}^{PDA-ENET}(\lambda_1, 10000)$ with optimized PDA-ENET fit $\hat{\beta}^{PDA-ENET}(\lambda_1, \lambda_2)$. Specifically, the PDA-ENET model was optimized freely over (λ_1, λ_2) pairings, while a PDA-LASSO model was fit with the same λ_1 , but with $\lambda_2 = 0$, and a PDA-UST model was also fit with the same λ_1 but with $\lambda_2 = 10000$. Comparisons facilitated evaluation of (i) which model the freely optimized PDA-ENET most closely resembled, and (ii) the effects of using no λ_2 regularization and thus using the standard sample covariance matrix estimate (PDA-LASSO), versus shrinking the estimated covariance matrix to identity (PDA-UST) – with the unconstrained PDA-ENET estimates lying somewhere at or between these two extremes.

Finally, for purposes of comparison with other popular models for FMRI classification, we also applied a linear support vector machine (linear SVM) classifier to the FMRI data. Linear SVMs essentially seek to maximize the margin (area) surrounding a linear separating hyperplane (the decision boundary) between different classes [37], [38]. Denoting the width of the margin as $2C$, and defining a hyperplane as $\{x : f(x) = x^T \beta + \beta_0 = 0\}$, the SVM problem is typically written:

$$\min \frac{1}{2} \|\beta\|_2^2 + \gamma \sum_{i=1}^N \xi_i \quad \text{subject to} \quad \begin{cases} \mathbf{y}(\mathbf{X}^T \beta + \beta_0) \geq \mathbf{1} - \xi \\ \xi_i \geq 0 \quad \forall i \end{cases} \quad (9)$$

where the margin surrounding the separating hyperplane is related to the coefficients by $C = 1/\|\beta\|$, and where ξ is a vector of “slack variables”, which are zero for correctly classified trials and give the distance from the separating hyperplane for incorrectly classified trials [37], [38]. Interestingly, it is possible to rewrite this “standard” SVM formulation as a penalized regression:

$$\hat{\beta}^{SVM} = \arg \min_{\beta, \beta_0} \|(\mathbf{1} - \mathbf{y}(\mathbf{X}^T \beta + \beta_0))_+\|_2^2 + \lambda_{SVM} \|\beta\|_2^2 \quad (10)$$

where $\lambda_{SVM} = 1/2\gamma$ [37].

In this formulation, linear SVM takes the form of a penalized regression in which the penalization term λ_{SVM} is related to the width of the margin and the model is penalized for allowing some observations to fall on the wrong side of the separating hyperplane. However, this penalization does not result in automatic variable selection as in the SPDA models above. Instead, it returns non-zero coefficients for essentially all input variables.

Despite evidence indicating that optimization over a regularization parameter (as is standard for penalized regressions) can significantly improve SVM classifier results [39], no published studies applying SVMs to FMRI data have done so. Instead, most studies adopt software default penalization parameters (e.g. $\lambda_{SVM} = 0.005$ in SVMlight). Here, we optimize linear SVM fits over the parameter λ_{SVM} , using five-fold cross-validation and the SVMSPATH algorithm to fit the entire regularization parameter path and to estimate an optimal value for λ_{SVM} [39].

As is standard for penalized models, inputs were centered and standardized prior to entry into the model [37]. Multivari-

ate SPDA, SVM, LDA, and LR models were fit with the freely available Elastic Net, SVMSPATH, and MASS packages for the R Statistical Computing Environment [40]. For the spatially averaged LR models we used the MATLAB statistical toolbox (Mathworks, Natick, MA). The Elastic Net package uses the EN-LARS algorithm which fits the entire λ_1 -regularization path in approximately the time required for a single ordinary least squares fit [23]. Full models were fit for each value of $\lambda_2 \in \{0, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}$.

The speed of the EN-LARS algorithm allowed fitting of all models over a five-fold internal cross-validation to estimate optimal values for (λ_1, λ_2) . Each of these five internal cross-validations was in turn nested within the training set of a larger two-fold cross validation used to estimate out-of-sample (“test”) error with the estimates of (λ_1, λ_2) chosen during the internal cross-validation. Each two-fold cross-validation was carried out 10 times by repartitioning the same data into different training and test sets five times, and using each partition once as the training set and once as the test set. All models were applied to identical training and test partitions to ensure that they would be comparable (this eliminated the potential confound of rate variability across models resulting from having fit them to different partitions of the data). This arrangement allowed us to compare the models using the 5×2 cross-validation test, a modified paired-t-test that has been shown to have reasonable power and an acceptable type I error rate (unlike, e.g., standard paired t-tests, whose type I error rate is inflated for such comparisons) [41].

Coefficient estimates were obtained for the cross-validations by taking the median value of each coefficient over the cross-validations. Since a regression using an L_1 penalty on the coefficients is equivalent to a maximum a posteriori (MAP) estimator of the mode with a Laplacian prior [37], the median (rather than the mean) provides an appropriate estimator of central tendency.

In summary, we compared six linear classifiers to explore the effects of penalization and automatic variable selection on classification accuracy and model interpretability. Two classical models (LR and LDA) provided a baseline, as they include no penalization or automatic variable selection. The more modern linear SVM included penalization, but no variable selection. Finally, the three sparse PDA models include both penalization and automatic variable selection. The PDA-LASSO and PDA-UST are special cases of the PDA-ENET with $\lambda_2 = 0$ and $\lambda_2 = +\infty$ (here approximated by setting $\lambda_2 = 10000$), respectively. These SPDA models include both penalization and automatic variable selection. Each of the three SPDA models treated the sample covariance matrix differently. Therefore, comparing these three models allowed us to examine the effects of differentially penalizing the estimated covariance matrix on both model coefficients and classification accuracy.

IV. RESULTS AND DISCUSSION

A. Preprocessing Results for Averaged Data

Spatially smoothing the data (at 8 mm FWHM) appeared to decrease the contributions of NAcc and insula to the

TABLE I

Z-SCORES AND COEFFICIENTS FOR LOGISTIC REGRESSION MODELS FIT TO AVERAGED DATA (N=25)

Spatial Blur	0 mm	0 mm	4 mm	8 mm
Temporal Filter	No Filter	HPF	HPF	HPF
Constant	-1.92 -.329(.17)	-2.06 -.353(.17)	-2.06 .353(.17)	-1.99 -.341(.17)
NACC (Bilateral)	5.70* .574(0.44)	6.98* .991(.14)	6.90* 1.07(.16)	6.35* 1.11(.18)
MPFC (Bilateral)	5.44* .445(0.08)	6.63* .696(.11)	6.79* .753(.11)	7.10* .856(.12)
Insula (Right)	-7.72* -.636(0.08)	-5.91* -.780(.13)	-5.89* -.889(.15)	-5.35* -1.00(.19)
Observations	3,761	3,761	3,761	3,761
Pseudo- R^2	0.102	0.109	0.109	0.107
AIC	4147.6	4118.2	4116.8	4126.0
Rate	63.6(1.3%)	64.7(1.3%)	65.0(1.3%)	65.0(1.3%)

Regression includes subject fixed effects. Z-scores above coefficients with standard errors in parentheses (Significance: * $p < .001$, two-tailed). AIC: Akaike Information Criterion (lower score indicates better fit to the data).

Classification Rates: Taken from the mean of 200 iterations of five-fold cross validation.

logistic regression model fit to data averaged within ROIs. However, spatial smoothing did not significantly change the fit or classification rate of the logistic regression model overall (Table 1). These findings are consistent with the claim that in small gray matter regions adjacent to white matter (e.g., NACC and insula) partial voluming may reduce data quality. Temporally filtering the data (by admitting frequencies < 90 sec) appeared to increase the contribution of the NACC and MPFC while decreasing the contribution of the insula to the predictive model. However, the effects of temporal filtering did not significantly change the fit or classification rate of model overall (Table 1). Thus, the ability to predict purchases from brain activation remained fairly robust across different spatial smoothing and temporal filtering regimens. Given that spatial blur would artificially increase correlations between variables for the voxel-wise analysis, and on the basis of the above comparisons, we used data with no spatial blur and a temporal high pass filter for the remaining analyses.

B. Classification Across Subjects Using Voxel-wise Data

Six predictive models were applied to spatially unsmoothed high-pass filtered data, yielding the held-out (“test”) sample rates and associated p-values (binomial) shown in Table 2. The associated p-values correspond to the classification accuracy (chance level = 50%). All six models classified significantly above chance level for the combined datasets as well as for first presentation dataset ($p < .01$). The three SPDA models (but not the others) also classified significantly above chance level for the second presentation dataset.

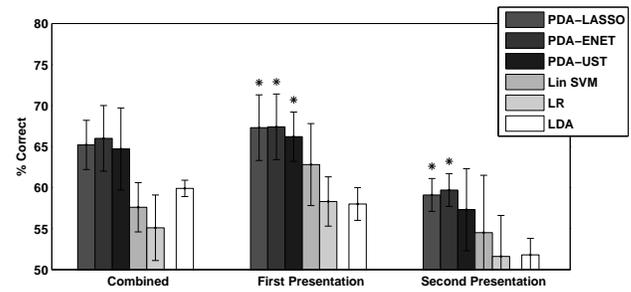


Fig. 2. Mean classification accuracy (% correct on held-out test set) and standard deviations for each model on first, second, and combined presentation datasets (compared against LDA model performance within dataset via 5×2 CV tests; * $p < .05$)

TABLE II

ACROSS-SUBJECT CLASSIFICATION RATES FOR ALL MODELS FIT TO VOXEL-WISE DATA

	Combined	Presentation 1	Presentation 2
PDA-LASSO	65.2 ± 3% (3.24 × 10 ⁻²⁴)	67.3 ± 4% (2.35 × 10 ⁻¹³)	59.1 ± 2% (2.66 × 10 ⁻⁵)
PDA-ENET	66 ± 4% (1.83 × 10 ⁻²⁶)	67.4 ± 4% (2.35 × 10 ⁻¹³)	59.7 ± 2% (8.21 × 10 ⁻⁶)
PDA-UST	64.7 ± 5% (1.33 × 10 ⁻²²)	66.2 ± 3% (1.59 × 10 ⁻¹⁴)	57.3 ± 5% (8.37 × 10 ⁻⁴)
Lin SVM	57.6 ± 3% (4.88 × 10 ⁻⁷)	62.8 ± 5% (1.56 × 10 ⁻⁹)	54.5 ± 7% (0.040)
LR	55.1 ± 4% (8.37 × 10 ⁻⁴)	58.3 ± 3% (1.37 × 10 ⁻⁴)	51.6 ± 5% (0.494)
LDA	59.9 ± 1% (5.92 × 10 ⁻¹¹)	58 ± 2% (1.92 × 10 ⁻⁴)	51.8 ± 2% (0.442)
Total Trials	2212	1118	1094
Test Trials	1106	559	547

Classification rates with standard deviations; p-values in parentheses below indicate significance of classification accuracy above chance (binomial) on held-out test set.

Bold indicates that the classification accuracy is significantly greater than the corresponding classification accuracy for the LDA ($p < .05$).

We used 5×2 CV tests on model classification rates to compare performance of each model against LDA as a baseline. All three SPDA classifiers performed significantly better than the LDA model on the first presentation dataset. Additionally, the PDA-ENET and PDA-LASSO performed significantly better than the LDA model on the second presentation dataset. There were no significant differences between classification rates for either the SVM model or the logistic regression model and the LDA model on any dataset ($p > .05$). Together, these findings indicate that SPDA models outperform logistic regression, LDA, and linear SVM models. Classification accuracy can be taken as a goodness-of-fit measure for these models (e.g., it is equivalent to a variant of the pseudo- R^2 for logistic

regression). Therefore, improved classification accuracy not only demonstrates that SPDA classifiers better predict purchasing decisions, but also supports greater confidence in the underlying model and its coefficients.

In comparing the SPDA models, the PDA-ENET was freely optimized over the λ_2 parameter, which could take values ranging from $\lambda_2 = 0$ (the PDA-LASSO solution) to $\lambda_2 = 10000$ (the approximate PDA-UST solution) Since the estimated values for λ_2 for all three datasets were close to $\lambda_2 = 1$, the PDA-ENET appeared to balance characteristics of both the PDA-LASSO and PDA-UST models. Thus, while coefficient estimates improved from the PDA-LASSO solution after shrinking the sample covariance matrix towards the identity matrix, stopping shrinkage prior to the PDA-UST solution (and thus using a stabilized sample covariance matrix in the model) provided the highest classification accuracy.

C. Interpreting Model Coefficients

In addition to improving classification rates, SPDA models also markedly enhanced coefficient interpretability in both space and time. Coefficient values indicate the degree to which a particular voxel at a particular time point contributed to discriminating the eventual choice to purchase an item or not. The SPDA models automatically selected a relevant set of variables for classification, setting the remaining coefficients to zero. Comparison models (i.e., LR, LDA, linear SVM) did not perform automatic variable selection, yielding more complicated coefficient maps that would require heuristic thresholding for subsequent interpretation. Even after such thresholding, coefficients may resist interpretation, since correlation between variables can yield unstable coefficients that vary greatly in magnitude and sign (for examples see Supplement 3). Both lack of variable selection and unstable coefficients potentially limit the extent to which a model can generalize to new data.

For purposes of interpretation, we plotted the estimated SPDA coefficients as heat maps organized spatially by voxel and temporally by time point (TR=2 sec) (Figure 3). Below each coefficient map, average values within each region are plotted over time. LR and LDA models produced uninterpretable coefficient maps (Supplement 3), and the linear SVM model coefficient map is depicted for comparison (Figure 4). Supplement 4 includes a video of the coefficients changing over time (overlaid on brain volumes).

Across SPDA models, while the PDA-LASSO provides a sparser solution, the coefficients are also less interpretable, due to the model's tendency to choose only one of a group of correlated inputs. PDA-ENET and PDA-UST, on the other hand, can include grouped coefficients corresponding to correlated input variables. Consistent with the classification rates reported in Table 2, coefficients selected by all SPDA models differed markedly for the first and second presentation datasets, suggesting that the neural correlates of making an initial purchasing decision versus a repeated purchasing decision differ. We therefore discuss these results separately below.

For the first presentation dataset, all SPDA models showed a contribution of the NAcc starting during product presentation and continuing through price presentation. The MPFC's contribution was strongest during price presentation and continued

during the choice period. While all SPDA models showed similar NAcc and MPFC contributions, the insula contribution varied across models. Specifically, the insula's contribution was clearest in the PDA-LASSO coefficients but no longer evident in the PDA-ENET or PDA-UST coefficients. These models differ in their treatment of the estimated covariance matrix. While the PDA-LASSO does not penalize the estimated covariance matrix and thus includes relationships between inputs, the PDA-UST shrinks the estimated covariance matrix to the identity matrix and thus treats the input variables as independent. Since insula contributions were most apparent in the PDA-LASSO coefficients and absent in the PDA-UST coefficients, they likely resulted from spurious correlations with other variables in the model (see also [19]).

For the second presentation dataset, all SPDA models indicated that the regions of interest contributed differently than in models fit to the first presentation dataset. The insula contributed more robustly in the price and choice periods across all three models, seemingly independent of the contributions of other inputs. In contrast, NAcc and MPFC contributions were weaker and less coherent in space and time than for the first presentation dataset. These findings suggest that initial purchasing decisions may utilize different neural circuits than repeated purchasing decisions (see also [42]).

Together, these findings confirm and extend the original analyses of these data [18], but without imposing prior assumptions about the relevance of certain time points. Instead, the SPDA models include all time points and automatically select those relevant to classification. The original correlational analysis suggested that different regions processed different types of information prior to the purchase decision. Specifically, NAcc activation during the product and price periods correlated with preference for the displayed product. The SPDA analyses indicate that NAcc activation began to contribute information about the upcoming purchasing decision precisely when the products were presented, and continued to contribute as the product remained on screen the price period. Similarly, in the original analyses, MPFC activation during the price period correlated with price differential (or perceived bargain). In the SPDA analyses, MPFC activation began to contribute information about the upcoming purchasing decision when the price was presented. Finally, in the original analyses, right insula activation during the choice period correlated with choosing not to purchase the product. In the SPDA analyses, right insula activation began to contribute information about the upcoming purchasing decision when the price was presented, but only during trials involving repeated decisions. Based on this finding, one might speculate that price has a greater impact than product preference in repeated versus initial purchasing decisions, but behavioral verification of such a hypothesis awaits future study. Thus, we find that sparse classifiers lacking temporal constraints can facilitate identification of when during a trial each voxel contributes to subsequent choice, and can validate functional inferences in a paradigm in which information is presented in a staggered and incremental fashion. Together, these methods support a spatially and temporally specific model of these neural circuits as they respond to different stimuli to promote upcoming

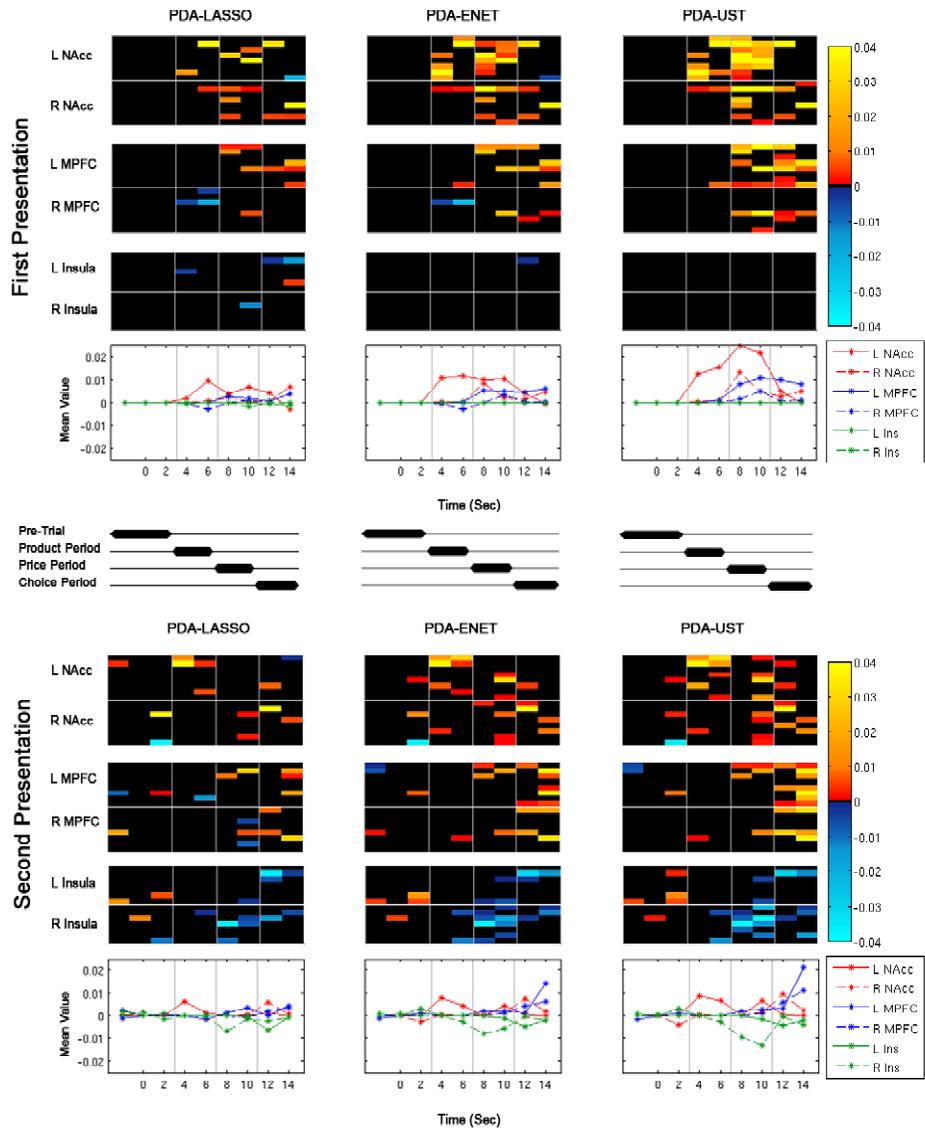


Fig. 3. Spatiotemporal coefficient maps and spatial averages for the three SPDA classifiers. Each panel represents the model constructed by each SPDA method (PDA-LASSO left, PDA-ENET center, PDA-UST right) when trained on a particular dataset (first presentation at the top, second presentation at the bottom). Lines in the center are included to illustrate time periods associated with the task. These periods are lagged by 4 seconds to account for hemodynamic response. Actual trial timing is marked on the x-axis of line graphs. The heat maps display the coefficient estimates organized by brain region (along the y-axis) and time (along the x-axis) each rectangle represents one voxel at one time point. Black rectangles indicate that the classifier has automatically excluded the input variable corresponding to that rectangle. All non-zero coefficient estimates are represented by a color corresponding to the value on the color bar. Line graphs below the heat maps illustrate average contribution of each region of interest for each point in time. Each point here represents the mean value across 7-8 voxels for one distinct time point.

purchasing decisions.

For comparison, we also examined heat maps of coefficients from a linear SVM model, which has been recently applied to fMRI data with the aim of producing more interpretable spatiotemporal results [12]. Since SVM classification for the second presentation data was not significant, only results for the first presentation data are considered. The SVM heat map suggests that it may rely on similar inputs to those selected by the SPDA models, with peaks in NAcc coefficient values occurring during the product and price periods. However, the coefficient map is much noisier than those produced by the SPDA classifiers. The primary difference between the

SPDA and SVM coefficient maps has to do with the models' treatment of "noise" variables. This is particularly apparent during the first presentation pre-trial period. Since subjects have no information about the product or price during this pre-trial period, they cannot contemplate an upcoming choice. While SPDA models set these pre-trial coefficients to zero, the SVM leaves them all in the model. This limitation makes distinguishing signal from noise a challenge in interpreting SVM models (particularly in the temporal domain), and in the present context appears to harm linear SVM classification rates.

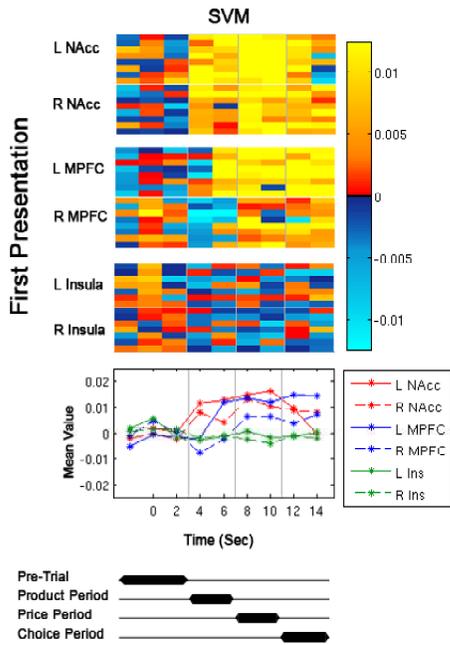


Fig. 4. Spatiotemporal coefficient heat map and spatial averages for the SVM model on first presentation dataset. Representation is the same as in figure 3.

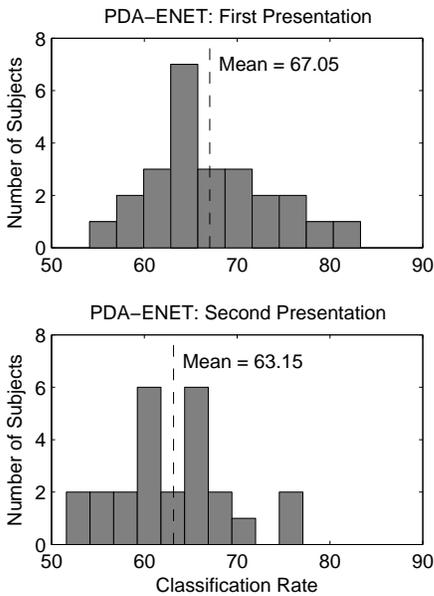


Fig. 5. Within-subject ENET classification rates for first and second presentation datasets.

D. Classification Within Subjects

While the preceding analyses focused on classification across all subjects, much of the current fMRI classification literature instead focuses on models fit within individual subjects. To examine the performance of SPDA within individuals, we fit PDA-ENET models to each individual’s data using three-fold cross validation to estimate out-of-sample (“test”) error. Model parameters were fit using an additional internal three-fold cross-validation within each training sample. A histogram of the resulting classification rates for each of the 25 subjects (Figure 5) shows a mean classification rate of 67% for the first presentation dataset (binomial $p = 1.5e-7$ across test trials). Nine subjects had rates $> 70\%$, with a maximum individual test rate of 83.3%. For the second presentation dataset, the mean classification rate was 63% (binomial $p = 1.3e-4$ across test trials). Three subjects had rates $> 70\%$, with a maximum individual rate of 77.1%. These findings show that while the mean PDA-ENET classification accuracy within subjects is similar to that reported above across subjects, the variance of these rates for fits within subjects is substantial. Consideration of individual differences may thus prove crucial to developing brain computer interfaces and to real-time classification applications in this domain.

Since the average number of trials available for fitting individual models was $> 10\%$ of the number of input variables, the success of the PDA-ENET model within subjects suggests that this model classifies well even given fewer observations than input variables (i.e., in the $N \ll p$ setting). It is worth noting that fitting such data is not possible with LDA or LR classifiers. Together, these findings suggest that the PDA-ENET model presents a viable option for single-trial prediction of purchasing decisions, both across subjects and within subjects. Finally, the algorithm for fitting the PDA-ENET model is highly efficient (i.e., can fit a model for all possible values of λ_1 in approximately the same time as a single ordinary least squares fit), making it well-suited to the temporal demands of real-time classification.

V. CONCLUSION

Relative to other linear models, sparse penalized discriminant analysis (SPDA) improved fMRI prediction of purchases by both increasing classification accuracy and enhancing interpretability. Specifically, the PDA-ENET model achieved about a 67% correct classification rate, better than the rates obtained with other models (e.g., LR, LDA, linear SVM), as well as previously-reported results [25]. Interestingly, the PDA-ENET model yielded better classification rates for initial versus repeated decisions, and these improvements held both across and within subjects. Beyond improving classification accuracy, SPDA models also allowed coefficient interpretation in time as well as space. Improving upon previous analyses, SPDA models automatically selected coefficients in time and space that significantly predicted future purchasing decisions. These coefficients indicated that regions of interest began to predict purchasing only when relevant information was presented (i.e., product information in the case of the NAcc and price information in the case of the MPFC and insula). Thus,

combined with temporally staggered information presentation, SPDA models enable investigators to infer which brain regions respond to different types of information to influence impending decisions. While the present analyses focused on specific regions of interest, fast algorithms for fitting SPDA models will allow extension to whole brain analyses. Further, since SPDA models produced robust results both across and within subjects, they offer the potential for predicting purchases in real time, and may eventually extend to applications involving neurofeedback or brain/computer interfaces.

ACKNOWLEDGMENT

During preparation of this manuscript, BK was supported by a FINRA Investor Education Grant and National Institute of Aging Grant R21 030778.

REFERENCES

- [1] S. A. Huettel, A. W. Song, and G. McCarthy, *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer Associates, Inc, 2004.
- [2] N. K. Logothetis, "The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal," *Philosophical Transactions of the Royal Society of London*, vol. 357, pp. 1003–1037, 2002.
- [3] J. Panksepp, *Affective neuroscience: The foundations of human and animal emotions*. New York: Oxford University Press, 1998.
- [4] C. M. Kuhnen and B. Knutson, "The neural basis of financial risk-taking," *Neuron*, vol. 47, pp. 763–770, 2005.
- [5] B. Knutson, G. W. Fong, S. M. Bennett, C. M. Adams, and D. Hommer, "A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: Characterization with rapid event-related fmri," *NeuroImage*, vol. 18, pp. 263–272, 2003.
- [6] B. Knutson, J. Taylor, M. Kaufman, R. Peterson, and G. Glover, "Distributed neural representation of expected value," *Journal of Neuroscience*, vol. 25, pp. 4806–4812, 2005.
- [7] T. A. Carlson, P. Schrater, and S. He, "Patterns of activity in the categorical representations of objects," *Journal of Cognitive Neuroscience*, vol. 15, no. 5, pp. 701–717, 2003.
- [8] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fmri) "brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex," *NeuroImage*, vol. 19, pp. 261–270, 2003.
- [9] J.-D. Haynes and G. Rees, "Predicting the stream of consciousness from activity in human visual cortex," *Current Biology*, vol. 15, pp. 1301–1307, 2005.
- [10] Y. Kamitani and F. Tong, "Decoding the visual and subjective contents of the human brain," *Nature Neuroscience*, vol. 8, pp. 679–685, 2005.
- [11] T. Mitchell, T. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, "Learning to decode cognitive states from brain images," *Machine Learning*, vol. 57, pp. 145–175, 2004.
- [12] J. Mourao-Miranda, K. J. Friston, and M. Brammer, "Dynamic discrimination analysis: A spatial-temporal svm," *NeuroImage*, vol. 36, pp. 88–99, 2007.
- [13] A. O'Toole, F. Jiang, H. Adbi, and J. V. Haxby, "Partially distributed representations of objects and faces in ventral temporal cortex," *Journal of Cognitive Neuroscience*, vol. 17, no. 4, pp. 580–590, 2005.
- [14] X. Wang, R. Hutchinson, and T. Mitchell, "Training fmri classifiers to discriminate cognitive states across multiple subjects," *Advance in Neural Information Processing Systems*, 2003.
- [15] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, pp. 352–356, 2008.
- [16] C. Davatzikos, K. Ruparel, Y. Fan, D. G. Shen, M. Acharyya, J. W. Loughhead, R. C. Gur, and D. D. Langleben, "Classifying spatial patterns of brain activity with machine learning methods: application to lie detection," *NeuroImage*, vol. 28, no. 3, pp. 663–668, 2005.
- [17] S. M. Polyn, V. Natu, J. Cohen, and K. A. Norman, "Category-specific cortical activity precedes retrieval during memory search," *Science*, vol. 310, no. 5756, pp. 1963–1966, 2005.
- [18] B. Knutson, S. Rick, G. E. Wimmer, D. Prelec, and G. Loewenstein, "Neural predictors of purchases," *Neuron*, vol. 53, pp. 147–156, 2007.
- [19] A. N. Hampton and J. P. O'Doherty, "Decoding the neural substrates of reward-related decision making with functional mri," *Proceedings of the National Academy of Science*, vol. 104, pp. 1377–1382, 2007.
- [20] A. Hoerl and R. Kennard, "Ridge regression: Applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, 1970.
- [21] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [22] P. Zhao and B. Yu, "On model selection consistency of lasso," Statistics Department UC Berkeley, Tech. Rep., 2006.
- [23] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B*, vol. 67, no. 2, pp. 301–320, 2005.
- [24] K. Friston, A. Holmes, K. Worsley, J.-P. Poline, C. Fritn, and R. Frackowiak, "Statistical parametric maps in functional imaging a general linear approach," *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1995.
- [25] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *The Annals of Statistics*, vol. 23, no. 1, pp. 73–102, 1995.
- [26] G. H. Glover and C. S. Law, "Spiral in/out bold fmri for increased snr and reduced susceptibility artifacts," *Magnetic Resonance in Medicine*, vol. 46, pp. 512–522, 2001.
- [27] R. W. Cox, "Afn: Software for analysis and visualization of functional magnetic resonance images," *Computers in Biomedical Research*, vol. 29, pp. 162–173, 1996.
- [28] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.
- [29] T. Hastie, A. Buja, and R. Tibshirani, "Flexible discriminant analysis by optimal scoring," *Journal of the American Statistical Association*, vol. 89, 1994.
- [30] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.
- [31] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [32] C. L. Mallows, "Some comments on cp," *Technometrics*, vol. 42, no. 1, pp. 87–94, 1973.
- [33] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.
- [34] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.
- [35] I. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Chemometrics*, vol. 35, no. 2, pp. 109–135, 1993.
- [36] D. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.
- [37] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001.
- [38] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer, 2000.
- [39] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for support vector machine," *The Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, 2004.
- [40] R. D. C. Team, "R: A language and environment for statistical computing." *R Foundation for Statistical Computing*, vol. Vienna, Austria, 2008. [Online]. Available: <http://www.R-project.org>
- [41] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, pp. 1895–1923, 1998.
- [42] H. Kim, R. Adolphs, and J. P. O'Doherty, "Temporal isolation of neural process underlying face preference," *Proceedings of the National Academy of Science*, vol. 104, no. 46, pp. 18 253–18 258, 2007.