Improving FMRI prediction of purchases with
Penalized Discriminant Analyses

Logan Grosenick¹, Stephanie Greer², Brian Knutson²

November 21, 2007

¹ Neuroscience Institute at Stanford, Stanford University, Stanford California USA
² Psychology Department, Stanford University, Stanford California USA

Abstract

Despite growing interest in applying machine learning algorithms to brain data, few studies have gone beyond classifying sensory input to using brain activity 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) data is a promising candidate for such applications. However, FMRI data’s low signal-to-noise ratio, high dimensionality, and extensive spatial and temporal correlations present formidable analytic challenges. Here, we apply different preprocessing and machine learning algorithms to a previously-acquired dataset [23] 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 increase both classification rates and spatiotemporal interpretability. To this end, penalized discriminant analyses based on the LASSO, Elastic Net, and Univariate Soft Thresholding enabled automatic selection of correlated subsets of variables. Relative to logistic regression and linear discriminant analyses these methods provided both better classification rates and interpretability. These findings validate temporal inferences about which brain regions utilize which types of information in subsequent purchasing decisions. Penalized discriminant analysis methods may improve future attempts to use FMRI data to make single-trial behavioral predictions.

1 Introduction

The development of event-related functional magnetic resonance imaging (FMRI) has revolutionized cognitive neuroscience. Currently, among neuroimaging tech-
Techniques, only fMRI allows investigators to visualize changes in subcortical activity at a temporal resolution of seconds and at a spatial resolution of millimeters [17]. With fMRI, investigators visualize changes in vascular oxygenation (hereafter, “activation”) that occur 4–6 seconds after changes in neural activity. Such activation correlates more closely with postsynaptic changes in dendritic potentials than with presynaptic axonal firing rates [25]. Although the fMRI signal lags behind these postsynaptic changes, the lag can be modeled and deconvolved, allowing second-to-second temporal inference. Nonetheless, many fMRI methods derive from methods previously adapted 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 greater temporal specificity afforded by fMRI.

Traditionally, subcortical circuits have been of great interest to affective neuroscientists, since appetitive and aversive behavior can be unconditionally elicited from subcortical regions via electrical stimulation [30]. A little more than a decade of fMRI research has begun to validate some of these findings in humans, suggesting that a subcortical circuit including the nucleus accumbens (NAcc) plays a role in anticipation of gains, while a circuit including the deep cortical region the insula plays a role in anticipation of loss [24]. Additionally, a region in the medial prefrontal cortex (MPFC) appears to play a role in correcting erroneous gain predictions [21]. Together, these findings implicate these ancient parts of the brain in the representation of expected value and subsequent choice [22].

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 have the potential to 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 activation in these substrates then contributes to subsequent choice.

Despite the possibility of using brain 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 [2, 3, 16, 19, 27, 28, 29, 36]. Additional studies have used fMRI activation to classify lying vs. telling the truth [5], and recall of different object categories [31].

At present, only two fMRI studies have used classification analysis to predict choice. The first used a simple logistic regression model to predict purchasing behavior with averaged data from bilateral NAcc, MPFC, and right insula [23] (data reanalyzed below). The second used a linear discriminant analysis 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 [12].
Here, we sought to extend the results of the first study by improving the classification rates while preserving interpretability by implementing multivariate models that automatically select voxels important to subsequent choice, resolved in both time and space.

While classifiers have been used to spatially identify voxels relevant to a model, they have only been used to identify relevant voxels simultaneously in both space and time only once before [28]. In this example, the researchers applied a linear Support Vector Machine (SVM) to whole brain data over multiple time points to identify when relevant voxels discriminated between viewing positive and negative pictures. This approach yielded interesting coefficients in space and time, however, the SVM kept all coefficients in the model, making interpretation difficult. In this paper, in contrast, we sought to automatically identify subsets of voxels that best predicted purchasing in both space and time across and within subjects. In principle, automatic variable selection should select out a reduced number of relevant coefficients, which should facilitate both model parsimony and coefficient interpretability. In practice, however, in the case of highly correlated fMRI data, researchers must be careful to ensure the stability and uniqueness of those model coefficients.

Traditionally, coefficient penalization (or regularization) has proven an effective means of stabilizing correlated coefficients [18]. More recently, related models such as the Least Absolute Shrinkage and Selection Operator (LASSO) have extended penalized linear regression to include automatic variable selection, allowing researchers to zero irrelevant coefficients. Indeed, given certain conditions, LASSO models have been shown to possess the “oracle property” [7], in which they are guaranteed to asymptotically identify relevant coefficients while eliminating irrelevant ones. Unfortunately, correlated inputs violate the conditions under which this desirable property holds [37]. However, a generalization of the LASSO model called the Elastic Net model provides a promising alternative in such situations (ENET) [38]. 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 [10] – providing some continuity with currently popular neuroimaging analyses. Importantly, all of these models can be converted to discriminant classifiers through the application of a method related to LDA and canonical correlation analysis known as Optimal Scoring [13]. In this way these models can be reconceptualized as part of a larger class of Penalized Discriminant Analysis (PDA) classifiers [13].

The goal of this paper was to improve single-trial FMRI classification of purchasing by increasing both classification rates and spatiotemporal interpretability. Thus, we reanalyzed previously collected data using set of PDA models (i.e., PDA-LASSO, PDA-ENET, and PDA-UST) and compared classification rates against those obtained with more standard logistic regression and linear discriminant models. Additionally, we examined whether spatial and temporal smoothing would influence results. Since data came from small subcortical regions bounded by white matter, we reasoned that spatial smoothing might degrade classification due to partial voluming of signal, while temporal smooth-
ing might improve classification due to removal of low frequency signal changes unrelated to neural function.

2 Data Collection and Preprocessing

Data from 25 healthy right handed subjects were included in these analyses (one original subject’s FMRI data could not be recovered and so was not included). 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.

During scanning, subjects participated in a "Save Holdings Or Purchase" (SHOP) Task. In 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 by 75% of retail value to encourage purchasing. Consistent with pilot findings, this led subjects to purchase 30% of the products on average, generating sufficient instances of purchasing to adequately power statistical modeling.

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. Subjects were actually shipped products following 7 (28%) of the total of 25 scans.

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, to allow examination of the effects of repetition on the capacity for neural activation to predict 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.

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 [11]. 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, preproessing was conducted using Analysis of Functional Images (AFNI) software [4]. For all functional images, voxel time series were sinc 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. To compare different preprocessing algorithms, data were submitted to varying levels of spatial filtering (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 datasets were averaged over each volume of interest (VOI) and submitted to the same logistic regression format used previously to predict purchasing [23] (Figure 1). Subsequently un-averaged, high-pass filtered data with no blur were used for the spatiotemporal classification (Figure 1).

[Figure 1 about here]

Spatiotemporal data were arranged as in previous spatiotemporal analyses [28], but extracted from predefined regions of interest rather than whole brain volumes. Specifically, data was arranged as an $N \times p$ data matrix $X$ 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 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, yielding 438 input variables (Talairach coordinates listed in Supplement 2). These data were then subsampled to ensure equal numbers of buy and not buy trials, enabling assessment of binary classification rates using the binomial distribution with a predicted 50% success rate. Altogether, these data included 1118 trials across subjects for the first presentation dataset, and 1094 for the second presentation dataset, yielding 2212 total trials for the combined presentation dataset across subjects.
3 Penalized Discriminant Analysis Framework

Voxel-wise FMRI data has high dimensionality and a high degree of correlation between contiguous measurements in space and time. Application of standard Logistic Regression (LR) or Linear Discriminant Analysis (LDA) to FMRI data may thus suffer from degenerate sample covariance matrices, which can potentially limit both generalizability to new test data (increasing classification error) and degrade coefficient interpretability [13]. Appropriate penalization of the covariance matrix, however, can improve generalizability and yield more interpretable models [9, 13]. Further, some form of automatic variable selection – in which the model chooses an optimal subset of the variables – seems desirable given the large number of correlated input variables. Such variable selection should aid both interpretation and the model’s generalization to new data. Modern regression tools exist for both penalizing and performing automatic variables selection, but must be modified for application to binary classification.

With the ‘Optimal Scoring’ (OS) procedure [13], an optimal function can be estimated that converts the continuous output of any regression method into binary (or n-ary) classes. Thus, this OS procedure modifies penalized regression models to optimally classify categorical output variables (e.g. the decision whether or not to purchase a product). Here, we focus on penalized linear models, in the hope of generating interpretable model coefficients that yield insight into which data in time and space contribute to subsequent decisions. When applied to penalized regressions, this method has been called Penalized Discriminant Analysis (PDA) [13]. The penalized linear regression models we consider have coefficient estimates given by (in “Lagrangian” form):

\[
\hat{\beta} = \arg \min_{\beta} ||y - X^T \beta||^2 + \lambda J(\beta)
\]  

where \(y\) is a real-valued vector of outputs (dependent variables), \(X^T\) is the transpose of our input variable matrix (independent variables), \(\beta\) is the vector of coefficients, the function \(J(\beta)\) is some penalty function in terms of the model coefficients \(\beta_j; j = 1, \ldots, p\), and \(\lambda\) is a penalty parameter controlling how much the penalty term contributes to the solution for the estimated model coefficients \(\hat{\beta}\). Note that \(||x|| := \sum_i \sqrt{x_i^2}\), denotes the Euclidean norm of a vector.

The OS procedure modifies a regression model with continuous-valued outputs so that it can classify a vector of categorical outputs \(g\) by simultaneously optimizing over a function \(\theta(g) : g \rightarrow \mathbb{R}^{1 \times N}\). This function converts a vector of categorical inputs (e.g. 0’s and 1’s) to an output vector of real numbers. Given such a function, equation (1) can be altered to:

\[
\hat{\beta} = \arg \min_{\theta, \beta} ||\theta(g) - X^T \beta||^2 + \lambda J(\beta)
\]

minimized under the constraint \(N^{-1}||\theta(g)||^2 = 1\). As \(\theta(g)\) and \(\beta\) can be found separately, standard methods for fitting a particular regression model can then be applied to appropriately transformed data (see [13]). Since the OS procedure transforms regression models into discriminant classifiers, and since discriminant
classifiers can be written in terms of a combined regression and OS procedure, this framework allows extension of concepts from regression (e.g., degrees of freedom) to discriminant classification, which in turn facilitates model comparison using goodness-of-fit criteria (e.g., Akaike Information Criterion (AIC) [1], Bayesian Information Criterion (BIC) [32], Mallow’s Cp [26]) while preserving the ability to visualize data as discriminant coordinates.

As we are interested in both penalization to stabilize our coefficients and in automatic variable selection, a natural choice for our base regression method in a PDA is the LASSO [34], which uses the penalty function \( J(\beta_j) = |\beta_j| \) in equations (1) and (2) above. When the number of non-zero coefficients in the model is expected to be sparse (\( \leq N \) for \( p >> N \)), the LASSO provides an attractive alternative, since it performs simultaneous variable subset selection and prediction, and is easily computed using the LARS algorithm [6]. The LASSO has also been shown to perform well at prediction in many problems, competing favorably with ridge regression (where \( J(\beta_j) = \beta_j^2 \)) and the more general "bridge regression" (where \( J(\beta_j) = |\beta_j|^\gamma; 0 < \gamma < 2 \) [8].

Although the LASSO performs well in variable selection and prediction, it also has limitations, particularly in the case of correlated input variables or when the number of observations \( N \) is small relative to the number of input variables \( p \) (spatiotemporal FMRI data suffers from both problems). Specifically, the LASSO can select at most \( N \) variables when \( N < p \), and is not well-defined unless the \( L_1 \)-norm of the coefficients is below a certain value [38]. Given a group of highly correlated input variables, the LASSO is likely to arbitrarily select just one variable from the group, generating potentially unstable results over multiple fits and failing to capture correlated groups of relevant variables [38]. Model fit 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 [34]. Further, LASSO loses its desirable 'oracle property [7] - or asymptotically choosing only relevant input variables - when the input variables are correlated [37].

Because of problems with correlated input variables, the LASSO may not be ideally suited for the analysis of spatiotemporal FMRI data. A generalization of the elastic net (ENET) addresses correlations between inputs by implementing a hybrid penalty with both ridge and LASSO properties [38]. ENET coefficient estimates are given by:

\[
\hat{\beta}_{ENET} = \sqrt{(1+\lambda_2)} \arg \min_{\beta} ||y - X^T\beta||^2 + \lambda_1|\beta| + \lambda_2\beta^2
\]

and thus implement a hybrid penalty involving two penalty parameters \( \lambda_1 \) and \( \lambda_2 \), with the former essentially modulating automatic variable selection while the latter allows correlated variables and ensures unique coefficient solutions. As described in [38] Theorem 2, the estimates \( \hat{\beta} \) can be rewritten as:

\[
\hat{\beta}_{ENET} = \arg \min_{\beta} \hat{\beta}^T \left( \frac{X^TX + \lambda_2 I}{\sqrt{1+\lambda_2}} \right) \beta - 2y^TX\beta + \lambda_1|\beta|_1
\]

where standard LASSO estimates obtain when \( \lambda_2 = 0 \).
\[ \hat{\beta}^{LASSO} = \arg \min_{\beta} \hat{\beta}^T (X^T X) \beta - 2y^T X \beta + \lambda_1 |\beta|_1 \]  

Thus, ENET can be thought of as a stabilized version of the LASSO, in which the sample covariance matrix \( \hat{\Sigma} = X^T X \) is shrunken towards the \( p \times p \) identity matrix \( I \) as \( \lambda_2 \) increases, since the stabilized sample covariance matrix in equation (4) can be written:

\[ \frac{X^T X + \lambda_2 I}{\sqrt{1 + \lambda_2}} = (1 - \gamma) \hat{\Sigma} + \gamma I \]  

(with \( \gamma = \lambda_2 / (1 + \lambda_2) \) [38].)

Conversely, increasing \( \lambda_2 \), in the extreme to \( \lambda_2 = +\infty \) (and so \( \gamma \to 1 \)), creates a special case of ENET which yields "Univariate Soft Thresholding" (UST) coefficients (Donoho et al., 1994):

\[ \hat{\beta}^{UST} = \arg \min_{\beta} \hat{\beta}^T \beta - 2y^T X \beta + \lambda_1 |\beta|_1 \]  

which can be equivalently written as:

\[ \hat{\beta}_j^{UST} = \left( |x_j^T y| - \frac{\lambda_1}{2} \right)_+ \text{sign} (x_j^T y) \]  

for \( j \in \{1, \ldots, p\} \), 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 are of particular interest in the case of FMRI analysis, since the values \( x_j^T y \) are simply the univariate linear coefficient estimates, which are then thresholded at the data-driven threshold \( \lambda_1 / 2 \), below which they are set to zero. In other words, UST coefficients are equivalent to an appropriately thresholded mass-univariate general linear model map [10]. This equivalence provides a direct bridge between the coefficients for the family of ENET methods parameterized by \( (\lambda_1, \lambda_2) \) and mass-univariate statistical maps currently popular in FMRI analyses.

The current investigation compared PDA-LASSO estimates \( \hat{\beta}^{ENET} (\lambda_1, 0) \) and approximate PDA-UST estimates \( \hat{\beta}^{ENET} (\lambda_1, 10000) \) with optimized PDA-ENET fit \( \hat{\beta}^{ENET} (\lambda_1, \lambda_2) \). Specifically, the PDA-ENET model was optimized freely over a grid of \( (\lambda_1, \lambda_2) \) pairings, then a PDA-LASSO model was fit with the same \( \lambda_1 \), but with \( \lambda_2 = 0 \), and a PDA-UST model was also fit with the same \( \lambda_1 \) but with \( \lambda_2 = 10000 \). These comparisons enabled evaluation of (i) which model the freely optimized PDA-ENET most closely resembled, and (ii) the effects of using no \( \lambda_2 \) regularization and estimating the full covariance matrix (PDA-LASSO) versus shrinking the covariance matrix to identity (PDA-UST) – with the unconstrained PDA-ENET estimates lying somewhere between these two extremes.

Finally, for purposes of comparison with the PDA models, we applied a linear support vector machine (linear SVM) classifier to the spatiotemporal data for both presentations and their combination. Linear SVM seeks to maximize the
margin (area) surrounding a linear separating hyperplane (the decision boundary) between different classes [14, 35]. 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

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

(9)

where the margin surrounding the separating hyperplane is related to the coefficients by $C = 1/\|\beta\|$, and where $\xi$ 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 [14, 35]. One can also rewrite this "standard" SVM formulation as a penalized regression:

$$\beta^{SM} = \arg \min_{\beta, \beta_0} \left\| \left(1 - y(x^T \beta + \beta_0)\right)_+ \right\| + \lambda_{svm} \|\beta\|^2$$

(10)

where $\lambda_{svm} = 1/2\gamma$ [14].

In this formulation, linear SVM takes the form of a penalized regression in which the penalization term $\lambda_{svm}$ is related to the width of the margin. However, the penalization here does not result in automatic variable selection as in the PDA models above - rather it returns coefficients for all input variables.

We are unaware of any studies applying SVMs to fMRI data that have optimized over this penalization parameter, despite evidence demonstrating that optimization can have a significant impact on SVM classifier results [15]. 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 $\lambda_{svm}$, using 5-fold cross-validation and the SVMPATH algorithm to fit the entire regularization parameter path and estimate an optimal value for $\lambda_{svm}$ [15].

As is standard for penalized models [14], we centered and standardized inputs so that they had equal variance and the intercept term was zero. To fit the PDA and SVM models, we used the freely available Elastic Net and SVMPATH packages for the R Statistical Computing Environment [33]. For LDA and LR models we used the MATLAB statistical toolbox (Mathworks, Natick, MA). The Elastic Net package uses the EN-LARS algorithm which fits the entire $\lambda_1$-regularization path in approximately the time required for a single ordinary least squares fit [38]. We fit full models 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 5-fold internal cross-validation to estimate optimal values for $(\lambda_1, \lambda_2)$. Each of these 5 internal cross-validations was in turn nested within the training set of a larger 5-fold cross validation used to estimate out-of-sample error rate with the estimates of $(\lambda_1, \lambda_2)$ chosen during the internal cross-validation.

In summary, we employed 6 linear classifiers in order to explore the effects of penalization and automatic variable selection on class discrimination and model interpretability. Two models, LR and LDA, provide a baseline for comparison and utilize no penalization or variable selection, while the linear SVM includes a
penalization term, but no variable selection. We implemented three PDA models. 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 models include both penalization and automatic variable selection. Each of the three PDA models demonstrate a different treatment of the sample covariance matrix. Therefore, comparing these three models allows us to examine the effects of penalization on the estimated covariance matrix.

4 Results and Discussion

A. Preprocessing

Spatially filtering the data (at 8 mm FWHM) appeared to decrease the contributions of NAcc and insula to the predictive model fit to the data averaged within ROIs. However, spatial filtering did not significantly change the fit or classification rate of the model overall (Table 1). These findings are consistent with the prediction 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 of brain activation to predict purchases remained fairly robust across different spatial and temporal smoothing regimens. On the basis of these comparisons, we used a dataset with no spatial blur and a temporal high pass filtering for the remaining analyses.

[Table 1 about here]

B. Classification Across Subjects

Six predictive models were applied to the spatially unsmoothed high-pass filtered data, yielding the held-out (“test”) sample rates and associated p-values (binomial) shown in Table 2. All six models classified above chance for the combined dataset as well as for first presentation dataset ($p < .01$). The three PDA models (but not the others) also classified above chance for the second presentation dataset. Of the PDA models, the PDA-ENET model classified at higher rates than the comparison standard (i.e., the LDA model) on all three datasets. Additionally, the PDA-LASSO and PDA-UST models classified at higher rates than the LDA model on the second presentation dataset, and the PDA-LASSO model classified at higher rates than the LDA model on the first presentation dataset. There were no significant differences between classification rates for the LDA model and either the SVM model or the logistic regression model on any dataset ($p > .05$; uncorrected). Classification rates were significantly higher for the first versus the second presentation dataset across all six models (confirmed by paired t-tests across models, $p < .01$; uncorrected; Figure 2). Together,
these findings indicate that PDA models show superior classification to logistic regression, LDA, and SVM models.

[Table 2 about here]

[Figure 2 about here]

In comparing the PDA models, the PDA-ENET model was freely optimized over the $\lambda_2$ parameter, which could take optimal solution values ranging from $\lambda_2 = 0$ (the PDA-LASSO model solution) to $\lambda_2 = 10000$ (the approximate PDA-UST model solution). Since the optimal solutions for $\lambda_2$ for all three datasets were close to $\lambda_2 = 1$, the PDA-ENET model appeared to balance characteristics of both 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 including sample covariance-related information in the model) provided the best classification rates.

C. Interpreting Model Coefficients

In addition to improving classification rates, the PDA models also increased coefficient interpretability. Coefficient values indicate the degree to which a particular voxel at a particular point in time contributed to discriminating the eventual choice to purchase or not. The PDA models used here automatically selected a “relevant” set of spatiotemporal inputs for classification, setting the remaining coefficients to zero. Comparison models (i.e., LR, LDA, SVM) did not perform automatic variable selection and so assigned non-zero coefficients to all spatiotemporal inputs. This results in more complicated coefficient maps that require heuristic thresholding for subsequent selection of relevant inputs. Further, even after thresholding, coefficients may resist interpretation, as correlation between input variables to non-penalized models (LR and LDA) yields noisy coefficients that vary greatly in magnitude and sign (for examples see Supplement 3). Both lack of variable selection and unstable coefficients can limit model generalization to new data.

For purposes of interpretation, we plotted the PDA model coefficients as heat maps organized spatially by region and temporally by time point in each trial (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), however, an SVM model coefficient map is depicted for purposes of comparison (Figure 4). Supplement 4 includes a video of the coefficients changing over time overlaid on brain volumes.

[Figure 3 about here]

Comparing PDA models, while the PDA-LASSO model 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 input variables. PDA-ENET and PDA-UST model coefficients, on the other hand, can include grouped coefficients
corresponding to correlated input variables. Further, in the case of correlated coefficients, PDA-LASSO solutions are not unique, but PDA-ENET and PDA-UST do yield unique solutions [14]. Consistent with the classification rates reported above, coefficients selected by all PDA models for the first and second presentation datasets differed, suggesting that the neural correlates of making an initial purchasing decision versus a repeated purchasing decision vary. We therefore discuss these results separately below.

For the first presentation dataset, all PDA models showed a strong contribution of the left NAcc starting during product presentation and continuing through price presentation. The right NAcc contributed more during price presentation. The MPFC’s bilateral contribution was strongest during price presentation and the left MPFC continued to contribute during the choice period. While all PDA models showed similar NAcc and MPFC contributions, the insula contribution varied across models. Specifically, the insula’s contribution was clearest during the price period in the PDA-LASSO model but no longer evident in the PDA-UST model, with the contribution to the PDA-ENET model falling in between. These models differ in their treatment of the estimated covariance matrix. While the PDA-LASSO model imposes no penalty on the estimated covariance matrix and thus includes interactions between the input variables, the PDA-UST model shrinks the estimated covariance matrix to identity and thus treats the input variables as independent. Since insula contributions were most apparent in PDA-LASSO and PDA-ENET models, they may have resulted from interactions with other input variables (see also [12]).

All PDA models fit to the second presentation dataset indicate that the regions of interest contributed differently in this model than they did in the model 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 [20]). Based on the pattern of contributions revealed by the coefficient maps, one might speculate that perceived bargain has a greater impact on purchasing than product preference in repeated versus initial purchasing decisions, but verification of such a hypothesis awaits future research.

Together, these findings confirm and extend the original analyses of these datasets [23], but without imposing prior assumptions about the relevance of certain time points. Instead, the PDA models include all time points and automatically select those relevant to classification. The original correlational analysis suggested mechanisms underlying information represented in each region of interest. Specifically, NAcc activation during the product and price periods correlated with preference for the displayed product. The PDA analyses indicate that NAcc activation contributed the most information about the upcoming purchasing decision during these same time periods in each trial. Similarly, in the original analyses, MPFC activation during the price period correlated with price differential (or perceived bargain). In PDA analyses, MPFC activation
contributed the most information about the upcoming purchasing decision during the price period. Thus, models lacking temporal constraints can facilitate identification of when during each trial each region contributes to subsequent choice, and can validate functional inference in a paradigm in which information is presented in a staggered fashion. Together, these methods support a spatially and temporally specific model of how people use these neural circuits to make initial purchasing decisions.

[Figure 4 about here]

For comparison, we also examined heat maps of coefficients from a linear SVM model, a model recently applied to FMRI data classification with the purpose of producing interpretable spatiotemporal results [28]. Since SVM model classification for the second presentation dataset was not significant, only results for the first presentation dataset are considered. The SVM model heat map suggests that its classification may rely on similar inputs to those selected by the PDA models — with peaks in NAcc coefficient values occurring during the price and product periods. However, the coefficient map is much noisier than those produced by the PDA classifiers. The primary difference between the PDA and SVM coefficient maps has to do with the models’ treatment of noise variables. This is particularly apparent during the 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 PDA models mostly zero out these pre-trial coefficients, the SVM leaves them 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 seems to reduce linear SVM classification rates.

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 these models within individuals, we fit PDA-ENET models to each individual’s data using 3-fold cross-validation to estimate out-of-sample (“test”) error. Model parameters were fit using an additional internal 3-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 rate on held-out test trials within subjects is similar to that from the across-subjects classification above, the variance of these rates for fits within subjects is substantial. This suggests that consideration of individual differences may prove important in developing brain computer interfaces and in real-time classification
applications. Since the average number of trials available for fitting individual models was >10% the number of input variables, the robust performance of the PDA-ENET model within subjects suggests that this model classifies well even given few observations (i.e., in the \( N << p \) setting). 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 \( \lambda_1 \) in approximately the same time as a single ordinary least squares fit), making it well-suited to the temporal demands of real-time classification.

[Figure 5 about here]

5 Conclusions

Relative to other linear models, penalized discriminant analysis (PDA) models improved fMRI prediction of purchases by increasing classification and enhancing interpretability. Specifically, the PDA-ENET model achieved a ~67% classification rate, better than the rates obtained with other linear models (e.g., LR, LDA, SVM), as well as previously-reported results Knutson et al. [23]. The PDA-ENET model yielded better classification rates for initial versus repeated decisions, and these rates held both across and within subjects.

Beyond improving classification rates, PDA-ENET models also facilitated interpretation. Validating and refining earlier time-constrained analyses, PDA-ENET models automatically selected spatiotemporal input variables that maximally contributed towards classifying future purchasing decisions. This selection 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, PDA-ENET models potentially enable investigators to infer which regions respond to different types of information to influence impending decisions.

Spatially, only voxels in regions of interest were included in these classifications, but as efficient algorithms exist for their computation, these methods could be extended to include whole brain analyses. Because PDA-ENET models produced robust classifications both across and within subjects, they present a promising method for predicting purchases in real time, and may eventually extend to applications involving neurofeedback or brain/computer interfaces.

Acknowledgements:

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


**Figure Captions:**

Figure 1. Data preprocessing and analysis flowchart.

Table 1: Logistic regression models predicting purchasing or not purchasing using different spatial and temporal filters (n = 25).

Table 2. Across-subject classification rates and p-values (uncorrected) for first, second, and combined presentation datasets.

Figure 2. Mean classification rates (% accuracy 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 t-tests; * p < .05; ** p < .01, uncorrected)

Figure 3. Spatiotemporal coefficient maps and averages for three versions of PDA classifiers. Charts at the top, represent models for only the first presentation of the products; charts at the bottom represent models for only the second presentation of the products. 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 coefficients organized by brain region (along the y-axis) and time (along the x-axis) each rectangle represents a coefficient corresponding to one voxel at one time point. Black rectangles indicate that the classifier has automatically excluded the data point corresponding to that rectangle. All non-zero coefficients are represented by a color corresponding to the value on the color bar. Line graphs below the heat maps illustrate average coefficients for each time point across all voxels in each of six regions of interest. Each point here represents the mean value across 7-8 voxels for one distinct time point.

Figure 4. Spatiotemporal coefficient heat maps and averages for the SVM model on first presentation dataset. Representation is the same as that explained in figure 3.

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