FORECASTING PURCHASING BEHAVIOR USING FMRI DATA

LOGAN GROSENICK

Abstract

Despite growing interest in applying machine learning to neuroimaging data, few studies have gone beyond classifying sensory input to using brain data to forecast 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 candidate 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 Penalized Discriminant Analysis to a previously-acquired data [12] to investigate using fMRI activation in three regions – the nucleus accumbens (NAcc), medial prefrontal cortex (MPFC), and insula – to forecast purchasing behavior and generate an interpretable spatiotemporal model.

1. Introduction

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 a spatial resolution of millimeters [10]. With fMRI, investigators visualize changes in vascular oxygenation (hereafter, "activation") that occur 4-6s after changes in neural activity. This activation correlates closely with postsynaptic changes in dendritic potentials [15]. 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 have only recently adapted to take advantage of this greater temporal specificity.

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 [16]. 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 of the insula plays a role in anticipation of loss [14]. Additionally, a region in the mesial prefrontal cortex (MPFC) appears to play a role in correcting erroneous gain predictions [11]. Together, these findings implicate these ancient parts of the brain in the representation of expected value and related choice [13].

The ability to visualize anticipatory activation allows a reversal of the traditional logic of neuroimaging design and analysis. Instead of examining how sensory input influences brain activation, investigators have the potential to examine how brain activation influences subsequent motor output. The goal of this work was to advance single-trial fMRI forecasting of purchasing behavior by simultaneously yielding good classification rates and interpretable coefficients in space and time.

We reanalyzed previously collected data using set of penalized discriminant analysis (PDA) [8] models that both constrain correlated coefficients and perform automatic variable selection (PDA-ENET). We compared classification rates and model coefficients from these models against those obtained via logistic regression (LR), linear discriminant anlysis (LDA), and linear support-vector machine (SVM).

2. Data

During scanning, 25 subjects participated in a "Save Holdings Or Purchase" (SHOP) Task. In each of 80 task trials, 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. Subjects chose from 40 items twice and then chose from a second set of 40 items twice (two presentations of 80 unique items total), with each set in the same pseudorandom order, to examine the effects of item repetition (item sets were counterbalanced across subjects). For more details on data collection and preprocessing see [12]. For spatiotemporal analysis, we defined a $N \times p$ data matrix $\mathbf{X}$ with N corresponding to the number of trial observations of the p input variables, each of which was a particular voxel from an ROI at a particular time point. This resulted in a total of 414 input variables per trial – augmented to 438 input variables with fixed effects.

3. Penalized Discriminant Analysis

Voxel-wise fMRI data has high dimensionality and strong correlations between contiguous measurements in space and time. Application of standard Logistic Regression (LR) or Linear Discriminant Analysis (LDA) to fMRI data thus suffer from poor or degenerate covariance matrix estimates, which can limit model generalizability to new data and limit coefficient interpretability [8]. Appropriate penalization of the covariance matrix, however, can improve generalizability and yield interpretable models [4, 7, 8]. Further, automatic variable selection is desirable given the large number of correlated input variables. Such variable selection should aid in both interpretation and in the model's generalization to new data. Modern regression tools exist for both penalizing and performing automatic variables selection, but we must modify them to perform binary classification.

The 'Optimal Scoring' (OS) procedure [7, 8] 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 $\theta(\mathbf{g}) : \mathbf{g} \mapsto \mathbb{R}$. We may then write our penalized regression coefficient estimates in 'Lagrangian' form as:

$$\hat{\beta} = \arg\min_{\theta,\beta} ||\theta(\mathbf{g}) - \mathbf{X}^T\beta||_2^2 + \lambda J(\beta) \tag{3.1}$$

subject to $N^{-1}||\theta(g)||_2^2 = 1$, where $\theta(\mathbf{g})$ is a real-valued vector, $\mathbf{X}^T$ is the transpose of our input matrix, β is the vector of coefficients, the function $J(\beta)$ is a penalty function in terms of β, λ is a penalty parameter, and $||\cdot||_2^2$ is the L_2 norm.

One natural choice for our regression method in our PDA is the LASSO [17], which uses $J(\beta) = ||\beta||_1$ in equation (3.1), where $||\cdot||_1$ is the L_1 norm. When the number of non-zero coefficients in the model is expected to be sparse ($\leq N$ for $p >> N$), the LASSO [17] is attractive because it performs variable subset selection and is easily computed using the LARS algorithm [2].

Although the LASSO performs well in variable selection and prediction, it also has limitations, particularly in the case of correlated input variables or when $N < p$. 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 [19]. Given a group of highly correlated input variables, the LASSO is likely to randomly select just one variable from the group, generating unstable results over multiple fits and failing to capture correlated groups of relevant variables [19]. It's performance also suffers given correlated inputs, for instance, ridge regression empirically dominates the LASSO even in typical $N > p$ regression settings with correlated inputs [17]. Further, LASSO loses it's desirable 'oracle properties' [3] given grouped inputs [18].

A generalization of the LASSO called the elastic net (ENET) addresses the grouped variable problem by implementing a hybrid penalty with both ridge and LASSO properties [19]. ENET coefficient estimates are given by:

$$\hat{\beta}^{ENET} = \sqrt{(1+\lambda_2)} \arg\min_{\beta} ||\mathbf{y} - \mathbf{X}^T\beta||_2^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2 \tag{3.2}$$

which – as detailed in [19] Theorem 2 – can be rewritten as:

$$\hat{\beta}^{ENET} = \arg\min_{\beta} \hat{\beta}^T \left(\frac{\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}}{\sqrt{1+\lambda_2}} \right) \beta - 2\mathbf{y}^T\mathbf{X}\beta + \lambda_1||\beta||_1 \tag{3.3}$$

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

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \hat{\beta}^T \left(\mathbf{X}^T\mathbf{X} \right) \beta - 2\mathbf{y}^T\mathbf{X}\beta + \lambda_1||\beta||_1 \tag{3.4}$$

Thus ENET is like a stabilized version of the LASSO, with the estimate covariance matrix $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}$ shrunk towards the $p \times p$ identity matrix $\mathbf{I}$ as λ_2 increases [19].

Conversely, letting $\lambda_2 \to +\infty$, results in a special case of ENET called "Univariate Soft Thresholding" (UST) [1]:

$$\hat{\beta}^{UST} = \arg\min_{\beta} \hat{\beta}^T\beta - 2\mathbf{y}^T\mathbf{X}\beta + \lambda_1||\beta||_1 \tag{3.5}$$

which can be equivalently written as:

$$\hat{\beta}_j^{UST} = \left(|\mathbf{x}_j^T\mathbf{y}| - \frac{\lambda_1}{2} \right)_+ \text{sign}\left(\mathbf{x}_j^T\mathbf{y}\right) \tag{3.6}$$

for $j \in \{1, ..., p\}$. These estimates are of particular interest in the case of fMRI analysis, as they are equivalent to a thresholded mass-univariate GLM map [5]. This provides a direct bridge between the coefficients for the family of ENET methods and "statistical parametric maps" popular in fMRI analyses.

Inputs were centered and standardized to have equal variance and an intercept of zero. To fit the PDA and SVM models, we used the freely available Elastic Net and SVMPATH packages in R [9, 19]. The Elastic Net package uses the EN-LARS algorithm which fits the entire λ_1-regularization path in about the time required for an OLS fit. We fit models for each value of $\lambda_2 \in \{0, 0.0001, ..., 1000, 10000\}$. The EN-LARS algorithm allowed easy fitting of all models over a 5-fold internal cross-validation to estimate values for (λ_1, λ_2). Each of these 5 internal cross-validations was nested within the training set of a larger 5-fold cross validation used to estimate out-of-sample error rate with the estimates of (λ_1, λ_2) chosen via internal cross-validation. Logistic Regression (LR) and LDA were run in MATLAB.

TABLE 1

Data Set:	Presentation 1	Presentation 2	Combined
LR	59.6% (0.005)	53.9% (0.34)	62.7% (2.2e-5)
LDA	61.2% (0.001)	52.8% (0.50)	63.3% (3.8e-8)
Lin SVM	61.8% (0.0005)	52.1% (0.58)	63.8% (1.2e-08)
PDA-LASSO	65.5% (5e-6)	62.1% (6.5e-4)	64.0% (7e-9)
PDA-ENET	65.2% (9.4e-6)	60.4% (2.7e-3)	66.9% (1.7e-12)
PDA-UST	63.5% (1e-4)	60.1% (4.3e-3)	62.7% (1.1e-7)
Total Observations	1118	1094	2212
Test Trials	225	217	442

4. Results and Discussion

Six classifiers (LR, LDA, linSVM, and 3 PDA models) were applied to the data, yielding the held-out test rates and associated binomial p-values shown in Table 1. All models classified above chance for the combined data and the 1st presentation data ($p < .01$). The PDA models (but not the others) also classified above chance for the 2nd presentation data. Of the PDA models, t-tests showed that the PDA-ENET had significantly higher rates than LDA on all three datas ($p < 0.05$). Additionally, PDA-LASSO and PDA-UST had significantly higher rates than LDA on the 2nd presentation data ($p < 0.05$), and PDA-LASSO had higher rates than LDA on the 1st presentation data. There were no significant differences between rates for LDA and either SVM or LR on any data. Rates were significantly higher for the 1st versus the 2nd presentation data across all six models.

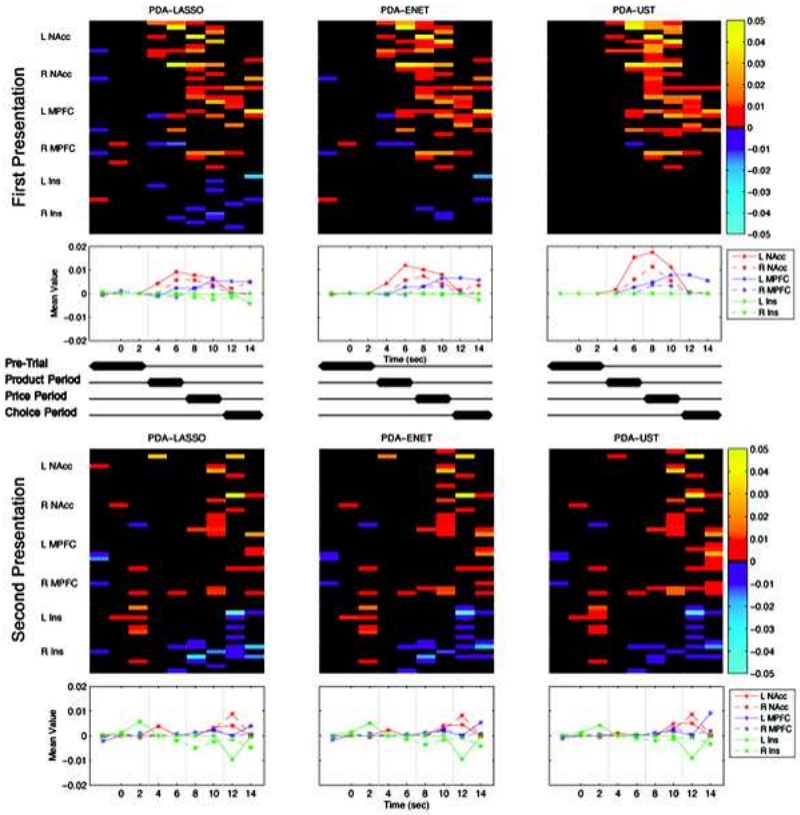
PDA-ENET was freely optimized over the λ_2 parameter, which could take optimal solution values ranging from $\lambda_2 = 0$ (PDA-LASSO solution) to $\lambda_2 = 10000$ (the approximate PDA-UST solution). Since the values for λ_2 chosen via 5-fold CV for all three datas were close to $\lambda_2 = 1$, the optimized PDA-ENET solution appeared to balance characteristics of both PDA-LASSO and PDA-UST models.

Critically, in addition to yielding the best classification rates to date for such data, the PDA models also increased coefficient interpretability. The PDA models used here automatically selected a set of spatiotemporal inputs for classification, zeroing the coefficients of remaining inputs. Comparison models (i.e., LR, LDA, SVM) did not perform automatic variable selection and so assigned non-zero coefficients to all spatiotemporal inputs. This, and the lack of regularization in the LR and LDA case, made the LR, LDA, and linSVM coefficients quite uninterpretable.

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 1). Below each heat map, average values within each region are plotted over time. Product, Price, and Purchase periods are also diagrammed.

For the first presentation data, all PDA models showed a strong contribution of the left NAcc starting during product presentation and continuing through price

FIGURE 4.1



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. 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 [6]).

All PDA models fit to the second presentation data indicate that the regions of interest contributed differently in this model than they did in the model fit to the first presentation data. The insula contributed more robustly – both 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 data. These findings suggest that initial purchasing decisions may utilize different neural circuits than repeated purchasing decisions. Further, they show that PDA-ENET is a viable option for single-trial based forecasting of purchasing behavior, and yields interpretable coefficients in space and time.

References

[1] D.L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[2] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshitani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[3] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[4] Jerome H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[5] K.J. Friston, A.P. Holmes, K.J. Worsley, J.-P. Poline, C.D. Fritn, and R.S.J. Frackowiak. Statistical parametric maps in functional imaging a general linear approach. *Human Brain Mapping*, 2(4):189–210, 1995.

[6] 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*, 104:1377–1382, 2007.

[7] Trevor Hastie, Andreas Buja, and Robert Tibshirani. Flexible discriminant analysius by optimal scoring. *Journal of the American Statistical Association*, 89, 1994.

[8] Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

[9] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for support vector machine. *The Journal of Machine Learning Research*, 5:1391–1415, 2004.

[10] S. A. Heuttel, A. W. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging*. Sinauer Associates, Inc, Sunderland, MA, 2004.

[11] 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*, 18:263–272, 2003.

[12] B. Knutson, S. Rick, G. E. Wimmer, D. Prelec, and G. Loewenstein. Neural predictors of purchases. *Neuron*, 53:147–156, 2007.

[13] B. Knutson, J. Taylor, M. Kaufman, R. Peterson, and G. Glover. Distributed neural representation of expected value. *Journal of Neuroscience*, 25:4806–4812, 2005.

[14] C. M. Kuhnen and B. Knutson. The neural basis of financial risk-taking. *Neuron*, 47:763–770, 2005.

[15] 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*, 357:1003–1037, 2002.

[16] J. Pankesepp. *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press, New York, 1998.

[17] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[18] Peng Zhao and Bin Yu. On model selection consistency of lasso. Technical report, Statistics Deparment UC Berkeley, 2006.

[19] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67(2):301–320, 2005.