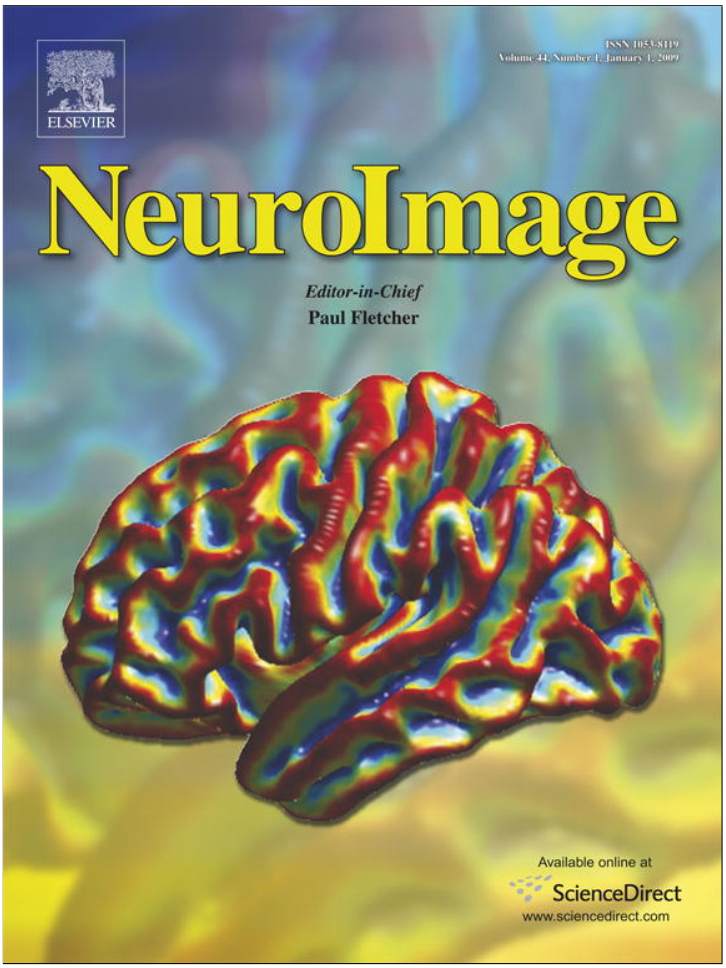

NeuroImage 44 (2009) 112–122

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Prediction and interpretation of distributed neural activity with sparse models

Melissa K. Carroll [a], Guillermo A. Cecchi [b,*], Irina Rish [b], Rahul Garg [b], A. Ravishankar Rao [b]

[a] Department of Computer Science, Princeton University, 35 Olden Street, NJ 08540, USA
[b] Computational Biology Center, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

ARTICLE INFO

Article history:
Received 26 February 2008
Revised 23 July 2008
Accepted 17 August 2008
Available online 27 August 2008

ABSTRACT

We explore to what extent the combination of *predictive* and *interpretable* modeling can provide new insights for functional brain imaging. For this, we apply a recently introduced regularized regression technique, the Elastic Net, to the analysis of the PBAIC 2007 competition data. Elastic Net regression controls via one parameter the number of voxels in the resulting model, and via another the degree to which correlated voxels are included. We find that this method produces highly predictive models of fMRI data that provide evidence for the distributed nature of neural function. We also use the flexibility of Elastic Net to demonstrate that model robustness can be improved without compromising predictability, in turn revealing the importance of localized clusters of activity. Our findings highlight the functional significance of patterns of distributed clusters of localized activity, and underscore the importance of models that are both predictive and interpretable.

Introduction

In the absence of formal neuronal theories of global brain function, the analysis of Functional Magnetic Resonance Imaging (fMRI) has been frequently reduced to modeling the relationship between specific image voxels and the associated mental tasks. The highest priority of early fMRI analysis was the *interpretation* of analyses to infer relevant voxels, usually involving testing hypotheses that related localized Regions of Interests (ROIs) to function. Over time, the putative regions became more localized and hypothesized functions more specific, such that, for instance, different brain regions were associated with viewing a face versus viewing a house. However, Haxby et al. (2001) published a seminal paper in which models built solely from sub-maximally responding voxels were able to discriminate between mental states, underscoring both the extent of distribution of brain function and the need for models that accurately *predict* mental states from fMRI data. Since then, a diverse collection of sophisticated predictive modeling methods have been introduced to the fMRI literature (Cox and Savoy, 2003; Norman et al., 2006), achieving impressive prediction performance that has surprised many neuroscientists. Interest in predictive modeling has been so strong that a competition, the PBAIC (Pittsburgh-EBC-Group, 2007), has been introduced to reward the most accurately predicting models. However, although predictive accuracy is vital to model assessment, it is important to keep in mind the ultimate objective of fMRI data analysis that underscores neuroscientific discovery, and thus include model interpretability as a necessary evaluation criterion.

It is well known in statistical data analysis that proper variable selection is as critical for prediction as for interpretation (Tibshirani, 1996); therefore one development that has contributed to strong predictive performance is *sparse* modeling, in which resulting models use information from only a relatively small subset of predictive variables. Standard predictive models from fMRI data are built using individual image voxels as predictors and single time point (time-to-response, or TR) volumes as examples (Mitchell et al., 2004), leading to datasets typically consisting of a large number (e.g. 10^4) of predictors but many fewer examples (e.g. 10^2 or 10^3). Learning statistical models from such data is particularly challenging since it is easy to *overfit* the training data and produce models that generalize poorly. However, the overfitting problem can be tempered by reducing the dimensionality of the data, and thus prediction performance can be significantly improved by employing methods that identify the relevant predictive variables or combinations of them. Many predictor selection techniques (e.g., as in Mitchell et al., 2004) use a straightforward *filtering* approach, treating the selection stage as separate from the modeling stage, but sparse modeling approaches combine the selection and modeling states into one process, sometimes called *embedded* selection. Other methods, such as ICA (Calhoun et al., 2003), use *dimensionality reduction* to extract new predictors by linearly combining voxels. Sparse modeling methods, such as LASSO (Tibshirani, 1996) (for regression) and Sparse PCA (Ulfarsson and Solo, 2007) (for dimensionality reduction), compare favorably to non-sparse methods on prediction performance, since they incorporate multivariate information into the selection process and, in some cases, through their incremental nature facilitate the optimization of predictors.

In principle, a sound fMRI model should exclude all irrelevant voxels while retaining all relevant ones, and the set of selected voxels

* Corresponding author.
E-mail address: gcecchi@us.ibm.com (G.A. Cecchi).

1053-8119/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.08.020

will be reliable or *robust*, i.e., consistent across multiple well-designed experiments exploring a task. Therefore, a voxel selection method should both facilitate the optimization of the number of included voxels, and produce robust models. Although the prediction performance of a model is a good measure of model validity, Zou and Hastie (2005) suggest that prediction performance alone does not guarantee validity. They prove that in datasets in which many relevant predictors are correlated with each other, such as genetics data, existing sparse techniques will tend to include only one representative predictor from each cluster of correlated predictors. The authors introduce a new algorithm, the Elastic Net, which they demonstrate matches, and in some cases surpasses, the prediction performance of sparse methods such as LASSO, yet can achieve the *grouping effect* of assigning similar weights to correlated predictors. Elastic Net even offers the implementer the ability to adjust the degree to which the grouping effect is enforced through one of its input parameters, while the other parameter controls the model sparseness, i.e. the number of voxels selected. Since activity levels of individual voxels are known to correlate highly with each other, some sparse modeling methods, such as LASSO, might fail to include all of the relevant voxels, but Elastic Net in such cases appears to fulfill interpretability goals of voxel selection while still surpassing the prediction performance of traditional non-sparse methods.

To reconcile traditional fMRI analysis approaches, which were guided by the intuition that functional units are localized regions rather than voxels, with the success of models of spatially distributed activity, we hypothesized that true neural response will be marked by distributed patterns of localized clusters of activity. The Elastic Net parameter controlling the "grouping" effect, usually labeled λ_2, is a particularly interesting variable for testing this hypothesis. Since nearby voxels tend to be highly correlated with each other, we hypothesized that as this grouping parameter is adjusted to increase the degree to which groups of correlated voxels are included, the model would feature more localized clusters, and hence the "scattering", or "spatial distribution," of the model voxels (formally defined later) would decrease.

In this paper, we describe the Elastic Net method and evaluate its behavior on an fMRI predictive modeling task, from the PBAIC 2007, with regard to model prediction performance and robustness, as well as to a novel spatial distribution metric we introduce herein. Our main results are the following: (1) optimizing the number of voxels included in the model via cross-validation on the data instead of choosing a fixed number of voxels is vital to prediction and, by facilitating this optimization, Elastic Net outperforms the traditional regression method of Ordinary Least Squares (OLS) in prediction performance; (2) for a fixed grouping parameter, better predictive performance of cross-validated models is positively correlated with increased spatial distribution; (3) increasing the grouping parameter tends to increase the robustness of a model; (4) models do become less spatially distributed as the grouping parameter is increased.

In summary, this paper demonstrates the promise of Elastic Net for fMRI data analysis, presents experimental results supporting our hypothesis about distributed patterns of localized clusters of neural activity, and illustrates the importance of producing models that are both predictive and interpretable.

Methods

Elastic net

In this section, we provide a formal description of the Elastic Net method. When both the fMRI data and predicted mental states are quantified as real-valued time series, as in the PBAIC data, a common approach is to formulate the prediction task as a regression problem, in which individual TRs are viewed as independent and identically distributed (i.i.d.) samples (a simplifying assumption), the voxel activity levels are the predictive variables (*predictors*), and the mental state is the predicted, or *response*, variable. Formally, let $X_1,\cdots, X_N$ be a set of N predictors, let Y be the response variable, and let M be the number of samples; $\mathbf{X}=(\mathbf{x}_1|\cdots|\mathbf{x}_N)$ denotes the $M\times N$ data matrix, where each $\mathbf{x}_i$ is an M-dimensional vector consisting of the values for predictor X_i for all M instances, while the M-dimensional vector $\boldsymbol{y}$ denotes the corresponding values for the response variable Y. Many existing regression techniques, including Elastic Net, assume a preprocessing step that performs location and scale transformations, so that the response variable is centered to have zero mean and all predictors have been standardized to have zero mean and unit length:

$$\sum_{i=1}^{M} y_i = 0, \sum_{i=1}^{M} x_{ij} = 0 \text{ and } \sum_{i=1}^{M} x_{ij}^2 = 1,\ 1 \le j \le N.$$

Then the linear regression problem is to learn the coefficients β_i in the following model:

$$\hat{\boldsymbol{y}} = \mathbf{x}_1\beta_1 + \cdots \mathbf{x}_N\beta_N = \mathbf{X}\beta \tag{1}$$

where $\boldsymbol{y}$ is an approximation of $\boldsymbol{y}$. A standard approach is to use the Ordinary Least Squares (OLS) regression which finds a set of β_i that minimize the sum-squared approximation error $||\boldsymbol{y}-\mathbf{X}\beta||_2^2$ of the above linear model. More advanced techniques include *regularized* regression methods that add a *regularization constraint* of some form to the basic least-squares minimization problem in order to avoid overfitting and improve the prediction accuracy. This constraint usually takes the form of a bound on norms of the coefficients of the model, i.e., the β values in Eq. (1). The two most common types of regularization imposed are bounds on the L_1- and L_2-norms, i.e., the sum of the absolute values or squares of the coefficients respectively. Note that from a Bayesian point of view, regularized regression can be viewed as finding regression coefficients that maximize the model's posterior probability under an assumption about the prior on the coefficient values; for example, L_2-regularization corresponds to a Gaussian prior and L_1-regularization corresponds to a Laplace prior. These norms are specific cases of L_q-norm, denoted $||\mathbf{z}||_q$, where $q \ge 1$:

$$||\mathbf{z}||_1 = \sum_{i=1}^{N} |z_i|,\ ||\mathbf{z}||_2^2 = \sqrt{\sum_{i=1}^{N} z_i^2},\ ||\mathbf{z}||_q = \left(\sum_{i=1}^{N} |z_i|^q\right)^{1/q} \tag{2}$$

Several state-of-the-art regularized regression methods exist, differentiated mainly by the type of regularization they employ. Examples include: *Ridge regression* (Hoerl and Kennard, 1988), which uses L_2-norm regularization, *LASSO* (Tibshirani, 1996), which uses L_1-norm regularization, and *Bridge regression* (Frank and Friedman, 1993; Fu, 1998), which uses L_q-norm regularization, with Ridge and LASSO corresponding to $q=2$ and $q=1$, respectively. Interestingly, in the L_q-norm regularization family where $q \ge 1$, only the L_1-norm regularization can produce a *sparse* model (Fan and Li, 2001), i.e. a model in which only a small subset of the predictors have nonzero coefficients (Tibshirani, 1996). Therefore, most of the modern sparse modeling methods include L_1-norm regularization.

The *Elastic Net* (*EN*) regression (Zou and Hastie, 2005) was designed to produce models that achieve both sparsity and the grouping effect mentioned in the Introduction by using a weighted combination of L_1- and L_2-norm penalties on top of the least-squares problem, and can be written formally as minimizing the following objective function:

$$L_{\lambda_1,\lambda_2}(\beta) = ||\boldsymbol{y}-\mathbf{X}\beta||_2^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2. \tag{3}$$

It is easy to see from Eq. (3) that Elastic Net becomes equivalent to LASSO when $\lambda_2=0$ and $\lambda_1>0$, while for $\lambda_1=0$ and $\lambda_2>0$ it is equivalent to Ridge regression. When both λ_1 and λ_2 are zero, the Elastic Net problem simply reduces to OLS regression. Also, as shown in Zou and

Hastie (2005, Eq. (16)), when λ_2 approaches infinity, the Elastic Net becomes equivalent to *univariate soft thresholding*, i.e. to correlation-based voxel selection with a particular threshold value.

The Elastic Net optimization problem (Eq. (3)) can be solved using the LARS-EN procedure proposed by Zou and Hastie (2005), which is a relatively simple modification of the most popular algorithm for solving LASSO, Least Angle Regression using a Stagewise procedure (LARS) (Efron et al., 2004). LARS-EN has two input parameters: the *grouping* parameter λ_2 and the *sparsity* parameter k that specifies the maximum number of *active* predictors, i.e. the predictors having nonzero coefficients in β (also called the *active set*). It can be shown (Efron et al., 2004) that each value of k corresponds to a unique value of λ_1 in Eq. (3), with larger λ_1 (i.e., larger weight on L_1-norm penalty) enforcing sparser solutions and thus corresponding to a *smaller* number of nonzero coefficients k. Herein, we will slightly abuse the notation, and *denote the sparsity parameter as* λ_1, *while always interpreting it as the active set size*. LARS-EN produces the collection of solutions, called the *regularization path*, for all values of λ_1 varying from 1 to its specified maximum value. As such, the sparsity parameter is also referred to as the *early stop* parameter since it serves as a stopping criterion for the LARS-EN incremental procedure, which adds predictors to the active set at each iteration (though removal is also possible). Note that LARS-EN, like the original LARS, is highly efficient, as it finds the entire regularization path at the cost of a single OLS fit. In addition, knowing the regularization path facilitates choosing the best solution β and its corresponding parameter λ_1 using cross-validation (described in Cross-Validated sparsity parameter λ_1). Further details about LARS-EN are provided in Appendix A.

Data

The data used in these experiments were supplied by the 2007 Pittsburgh Brain Activity Interpretation Competition (PBAIC) (Pittsburgh-EBC-Group, 2007) (see reference for more detail). Subjects were engaged in a Virtual Reality task, during which they had to perform a number of tasks, designed around the theme of "anthropology field work" in a hypothetical neighborhood. The field work included, among others, the acquisition of pictures of neighbors with particular characteristics (e.g., a piercing), the gathering of specific objects (e.g., fruits, weapons), and the avoidance of a growling dog. These tasks were rated as continuous variables over the time course of the experiment. The PBAIC organizers refer to the rating vectors as "features," but to avoid confusion with the statistics and machine learning literature, in which predictor variables are referred to as features, we will refer to these vectors as "response variables" or "response vectors." Several objective response variables (e.g., picking up the objects) were measured simultaneously with the functional data, while a few subjective response variables (e.g., valence) were estimated off-line. There were three independent runs (i.e., sessions) for each of 3 subjects; fMRI data for all of the runs are available, but the response data is available only for the first two runs. Each run includes fMRI data for the 33,000–35,000 voxels (depending on the subject) and 24 response vectors over 704 TRs each. All experiments were performed using fMRI data that had been passed through a high-pass filter (described in more detail in PBAIC Competition), and response vectors that had been convolved with a standard hemodynamic response function (HRF).

Metric definitions

Spatial distribution metric: We hypothesized that models would be marked by patterns of distributed clusters of localized neural activity. To evaluate model distribution and clustering, we computed a *spatial distribution metric* that estimates the spread of voxels throughout the brain. This metric is an adapted version of Thiel's redundancy measure, usually utilized to characterize spatial point patterns (Okabe et al., 2000). We first translate the β values obtained by the corresponding regression model back to their original (x, y, z) coordinates in brain space, i.e., the spatial maps of the models. We then calculate the degree of spatial distribution as follows: (1) the maps are binned using a fixed grid of $3\times3\times3$, the minimum bin size yielding a meaningful number, resulting in B bins (5808 in this case); (2) a normalized distribution is computed, such that each bin b is represented by

$$p_b = Q^{-1}\sum_{i\in b}|\beta_i|$$

where $Q=\sum_{j=1}^{A}|\beta_j|$ and A is the number of nonzero β weights, or number of active voxels, in the model; (3) the entropy of the distribution was computed as

$$H = -\sum_{b=1}^{B} p_b \log p_b$$

(4) the final distribution measure was computed as

$$d = H/H_0$$

where $H_0=\log A$ corresponds to the maximum entropy. Thus the spatial distribution will vary between 0 and 1, tending towards 0 for maximally clustered models, in which all or a majority of the β weight mass is located in one 27-voxel bin, and tending toward 1 for maximally distributed models, in which the weight mass is evenly spread among as many 27-voxel bins as possible. Spatial distribution was computed using this metric for each Elastic Net run using the resulting model's β weights. Note, therefore, that the same voxels trained using Elastic Net with different λ_2 values will result in slightly different spatial distribution scores, as the voxels will be weighted differently. The performance scores will naturally differ as well. Further details about the derivation of this algorithm and empirical tests of its validity are provided in the supplemental material.

Robustness metric: As mentioned in Introduction, we hypothesized that greater inclusion of voxels from within correlated clusters would result in greater overlap in included voxels between two models generated on different datasets. For computing robustness across experimental runs, we thus counted the total number of unique voxels selected over both experimental runs, and the number of voxels co-occurring in the two models. In our results, we discuss robustness for cases in which λ_1 was variable and optimized using cross-validation; therefore, we report robustness as the percent of the total number of unique voxels included in either model that appeared in both models.

Experiments

For each experiment, 144 models were trained: one for each of the 3 subjects, 24 response vectors, and 2 fMRI runs of 704 TRs (for cross-validation purposes).

Prediction performance metric: To evaluate the prediction performance of a model, the coefficients β were saved and applied to the voxel time series of 704 TRs not used for training. A Pearson correlation coefficient was obtained for the predicted response vector versus the actual convolved response vector for the 704 test TRs. This correlation coefficient is reported as the prediction performance.

Voxel selection

Using a filtering approach to selection is similar to setting the λ_2 parameter to a value of infinity, as only univariate associations with the response vector are considered. Therefore, we might expect the extent of inclusion of voxels from these clusters to increase on a continuum from low λ_2 values through filtering selection methods. In addition, we

sought to compare filtered and embedded voxel selection as well as naive, or random, voxel selection. For all model runs, the following methods were first used to obtain a ranking of voxels:

1. Correlation-based (filtering): A Pearson correlation coefficient was obtained for each voxel with the response vector of interest. Voxels were ranked by the absolute value of this coefficient.
2. Elastic/LASSO-based: Elastic Net was trained on the full dataset of 30,000+ voxels with $\lambda_1 = 1000$. The λ_2 parameter was selected from among 0.0, 0.1 and 2.0 as part of the experiments, in which 0.0 corresponds to pure LASSO. The λ_2 value is indicated in all results. Recalling that, on some iterations, LARS-EN removes voxels, the number of iterations taken to achieve an active set size of 1000 routinely slightly exceeds 1000 iterations. However, for simplicity, the voxels that were active during the first iteration in which the active set size was 1000 were selected as the top 1000 voxels and their rank was approximated as the order in which they were last added to the active set.
3. Random: For comparison purposes, random rankings were generated by randomly ordering the voxels. A different (unique) random ordering was used for each of the 144 models.

The top 1000 voxels from each of these 3 methods for each of the 144 runs were retained. Two sets of experiments were then performed, for each of the runs, described in the following two sections.

Fixed sparsity parameter λ_1

The top v voxels were selected from this set of 1000, where v was 10, 300, or 1000, approximating the active set if λ_1 was set to v. A regression model was then trained using only these voxels. The models were trained in two ways:

1. Linear regression (OLS): OLS was performed on the set of v voxels against the convolved response vector. When 1000 voxels were used, OLS results should not be considered valid since the data matrix is not invertible because the number of voxels (1000) exceeds the number of TRs (704). Therefore, OLS results are only reported when selecting 10 or 300 voxels.
2. Elastic Net/LASSO/Ridge regression: the subset of v voxels was used for training in Elastic Net without an early stop, i.e. with λ_1 set to its maximum value v, thus leaving only L_2 regularization. The λ_2 parameter was, for the Elastic/LASSO-selected voxels, set to the same value as was used in the selection phase and, for the other selection methods, varied as part of the experiments and indicated in the results. Due to regularization, the results of this method are still valid when the number of voxels exceeds the number of TRs. Note that this procedure is not equivalent to running Elastic Net on the full, original set of voxels with $\lambda_1 = v$, even when Elastic Net was used to generate the rankings, since, in the selection run, some voxels may have been added before the active set was size v but dropped before the active set size was 1000, and thus did not appear in any of the top v sets where $v < 1000$. Still, these results are likely close approximations of using the given λ_1 values.
   Also note that the Elastic Net results for a given λ_2 value are equivalent to applying Ridge regression with the λ_2 value to the pre-selected group of voxels. Since computing the Ridge regression model for datasets this size is typically infeasible, in practice a small subset of voxels are pre-selected, typically by examining univariate correlation (Chigirev et al., 2006), one of the voxel selection methods explored in these experiments.

Cross-Validated sparsity parameter λ_1

Cross-validation refers to the optimization of model-building parameters based on prediction performance on held-out data, e.g, data from a separate experimental run, for maximizing generalization of the selected model to new datasets. Testing on different data than that used in training is important to ensure that the model is not overfit to the training data. The PBAIC data was supplied in a form meant to facilitate cross-validation. In this set of experiments, we trained models on the data for both runs 1 and 2 for each subject and tested them on the data from whichever of the two runs was not used for training. The fact that Elastic Net produces the full regularization path allows us to easily select an optimal λ_1 value using cross-validation. We simply select, for each response variable, the β values from the iteration producing the model with highest prediction performance.

As in the fixed λ_1 experiments, in the cross-validated λ_1 experiments, Elastic Net regression was performed on the full, original set of voxels with a λ_1 of 1000. In this case, however, the β values were retained for the model generated on each Elastic Net iteration and were used to compute the prediction performance at each iteration. The model with the highest prediction performance was retained and the number of active voxels in the model was chosen as the final λ_1. Thus, λ_1 is determined separately for every final model produced.

Cross-validation of LASSO was similarly tested by running Elastic Net with a λ_2 value of 0.0. LASSO is applicable when $N > M$, as in this case; however, in such cases, Efron et al. (2004) states, "a LASSO fit can have no more than M-1 (mean centered) variables with nonzero coefficients.". Due to this limitation and computational inefficiencies in running LASSO with LARS-EN, these experiments only considered values of N up to 300, instead of the 1000 for Elastic Net.

We do not perform experiments in which the number of voxels used in an OLS model is selected using cross-validation, since the nature of OLS computation requires performing a separate OLS run for each value considered. In contrast, LARS-EN computes the entire regularization path, facilitating cross-validation of the λ_1 parameter in Elastic Net and LASSO, one advantage of these methods over OLS and Ridge. We feel that 10 and 300 voxels are representative numbers of voxels that might be tested when applying OLS and Ridge.

Note that the results reported for this second set of experiments should be taken with a grain of salt, as the λ_1 value was chosen by considering performance on the same dataset for which performance was evaluated. Thus, the performance results in particular will not reflect true generalization error, for which a third dataset would be needed. The most straightforward way of measuring generalization error would be to apply the models to data from the third run of the experiments. However, as of this writing, the third run response vectors have not yet been made public. As an approximation, the current test datasets can be randomly subdivided into an optimization dataset, used for determining λ_1, and a test dataset, used to evaluate prediction performance. Making such a division fairly is difficult for data from structured experiments, like the video game runs, yet we did so and found that prediction performance on the test datasets was nearly identical to that on the optimization datasets, even when applying a form of cross-validation to the OLS models(such that the λ_1 chosen resulted in the best mean optimization performance). These results are included in supplemental material. In addition, in a similar situation, Mitchell et al. (2004) found little difference in prediction performance estimates or the number of selected voxels whether or not an optimization dataset was used. The purposes of these experiments, rather than being to precisely compare prediction performances, are primarily to illustrate the importance of choosing an appropriate λ_1 value and explore the effects of other model parameters, neither of which is dependent on generalization performance.

All experiments were also performed for intermediate values of λ_2. The results were consistent with an interpolation between 0.1 and 2.0, but were not included for brevity.

Results

The goal of our study was to exploit the flexibility of Elastic Net to explore the four axes of prediction, interpretation, distribution and localization. For this, we experimented with both fixed and cross-validated λ_1 values, as well as two λ_2 values, 0.1 and 2.0. Since cross-validating the λ_1 value is most likely to affect prediction performance, and predictive models are associated with spatially distributed patterns, it makes sense to consider λ_1 along with these properties. Likewise, since we hypothesized that manipulating λ_2 would affect robustness, which relates to interpretation, and analyses focused more on interpretation tend to consider localized structures, we will consider λ_2 along with interpretation and localization.

Prediction, cross-validation, and spatial distribution

A goal of our prediction experiments was to compare Elastic Net, as a predictive modeling tool, to the traditional regression method of Ordinary Least Squares (OLS). In addition, we sought to compare Elastic Net, run with a nonzero λ_2 value, to LASSO, which can be considered Elastic Net run with a λ_2 value of 0.0. Recall that OLS requires choosing a fixed number of voxels for all models, such as the values considered in our experiments (10 and 300), while Elastic Net facilitates the selection of a model-dependent optimal number of voxels. As shown in Fig. 1, Elastic Net and LASSO perform comparably to each other, yet the ability to easily optimize the parameter λ_1 (i.e. the number of included voxels) can significantly improve prediction performance on the test (or optimization) dataset. This improvement, regardless of any voxel pre-selection method applied, causes both Elastic Net and LASSO to significantly outperform OLS in prediction. The prediction improvements for Elastic Net resulting from cross-validation are shown in Fig. 2 to occur consistently across all response variables. Also recall that the Elastic Net results in Fig. 1 can be considered equivalent to running Ridge regression on the same pre-selected groups of voxels.

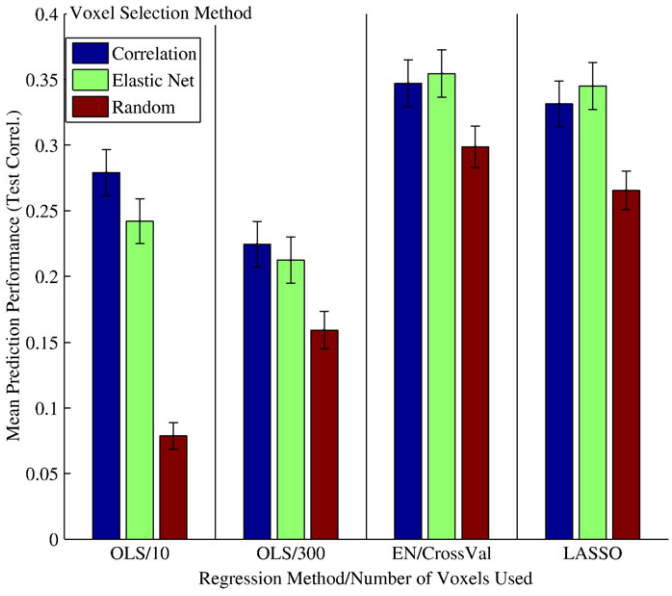

Fig. 1. Elastic Net and LASSO, when optimized, produce more predictive models than OLS with fixed numbers of selected voxels, regardless of the initial voxel selection approach. Prediction performance across the 3 voxel selection methods is shown for OLS, using 2 sample values for the number of voxels selected, and Elastic Net and LASSO, when cross-validation is used to select an optimal number of voxels. Mean correlation of model predictions with test data are shown, with standard error bars, averaged over the 24 response vectors, 3 subjects, and 2 cross-validation runs. For the Elastic Net voxel selection used by OLS and Elastic Net, and for the Elastic Net runs, a λ_2 value of 2.0 was used. The average number of voxels used over all 144 Elastic Net C-Val models was 296. The average used by each feature is shown in Fig. 5(c).

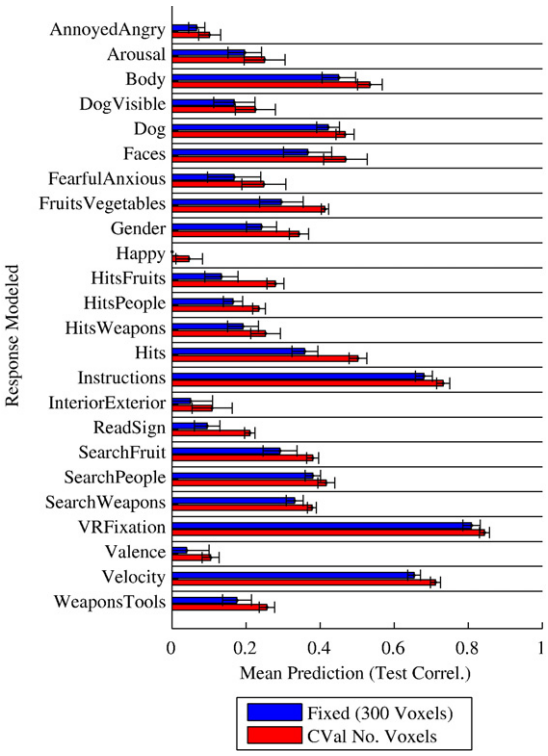

Fig. 2. Selecting an optimal number of voxels through cross-validation improves prediction performance compared to using a fixed number of voxels. Mean correlation of model predictions with test data for each predicted response vector are shown, with standard error bars, averaged over the 3 subjects and 2 cross-validation runs. An Elastic Net λ_2 value of 2.0 was used. For Fixed, 300 voxels were used. For C-Val, the number of voxels varied; averages by feature are shown in Fig. 5(c).

Our spatial distribution metric reveals a likely key reason for the improvement in prediction performance associated with selecting an optimal number of voxels. As Fig. 3 shows, in most cases, cross-validating, compared to using a fixed number of voxels, produces models that are significantly less spatially distributed; this finding suggests that, in many cases, very poor performance is associated with models overfit to training data due to the inclusion of many irrelevant voxels, which are most likely distributed randomly throughout the brain. However, note that in Fig. 1, random voxel selection, in which small random subsets of the full set of voxels were selected, resulted in prediction performance that was comparable to that of the more principled approaches, especially as a larger, fixed number of voxels was used. These seemingly disparate results can be reconciled by examining the effects of cross-validation. After cross-validating to remove the randomly distributed irrelevant voxels, we observe in Fig. 4 the importance of representations that incorporate information from distributed yet relevant voxels. When using a fixed number of voxels (Fig. 4a), greater model spatial distribution is correlated with poorer prediction performance, but when the number of voxels is chosen using cross-validation (Fig. 4b), the reverse is true: better predicted response vectors tend to be those for whose optimal learned models are more spatially distributed. These results, and the strong performance of random voxel selection provide strong evidence for the distributed nature of pattern representation in the brain.

Robustness and grouping

Having examined prediction performance using the standard criteria, we now consider robustness, our novel metric defined earlier. Recall that the main effect of increasing λ_2 is the inclusion of more voxels that are correlated with other relevant voxels and hence omitted by traditional sparse methods that only consider prediction performance. Indeed, as shown in Fig. 5a, increasing the λ_2 value from

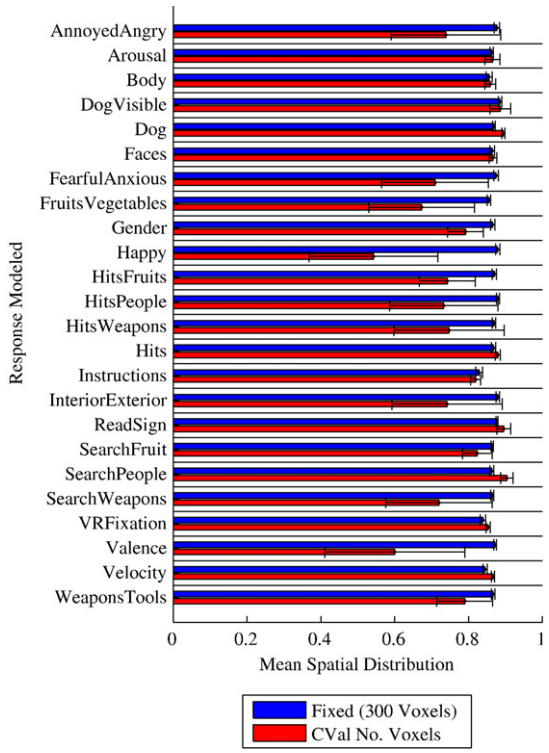

Fig. 3. Selecting an optimal number of voxels through cross-validation, rather than selecting a fixed number of voxels, frequently results in models that are significantly less spatially distributed. Mean spatial distribution values of learned models for each predicted response vector are shown, with standard error bars, averaged over the 3 subjects and 2 cross-validation runs (C-Val.). An Elastic Net λ_2 value of 2.0 was used. For Fixed, 300 voxels were used. For C-Val, the number of voxels varied; averages by feature are shown in Fig. 5(c).

0.1 to 2.0 may slightly improve prediction performance, but ultimately has little effect on prediction. However, since selection of voxels from within correlated clusters is likely arbitrary, including more voxels from such clusters should increase the frequency with which voxels are selected for separate experimental runs. As is clearly evident in Fig. 5b, increases in λ_2 are usually associated with increases in the *robustness* (as described in Methods) across the 2 cross-validation runs for each response vector+subject combination; moreover, as Fig. 5c shows, this change is frequently associated with the inclusion of a greater number of voxels; these additional voxels are likely those relevant voxels highly correlated with other relevant voxels. Hence, by including more relevant yet correlated voxels, increasing λ_2 improves model robustness without compromising prediction performance. We also found that model robustness directly correlates with model prediction performance ($r=0.677$, $p<0.001$ when $\lambda_2=0.1$; significant also for $\lambda_2=2.0$), suggesting that these two measures point to certain response variables as being generally "easier" to model, producing models that are both more predictive and more robust.

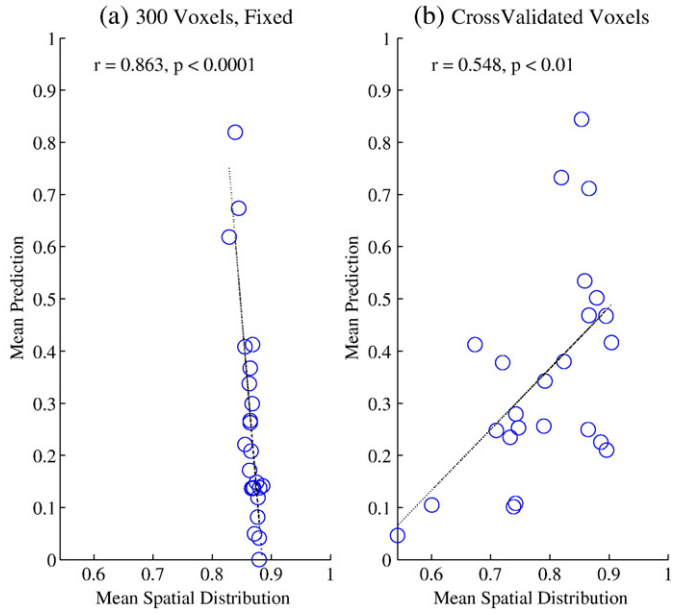

Fig. 4. When selecting a fixed number of voxels (a), higher spatial distribution of learned models is associated with poorer prediction performance; when the optimal number of voxels is selected through cross-validation (b), the reverse is true. Mean model spatial distribution values are shown plotted against their matching mean prediction performance (correlation with test data). Values are averaged over the 3 subjects and 2 cross-validation runs. A fit linear regression line is overlaid and correlation statistics are indicated. An Elastic Net λ_2 value of 2.0 was used.

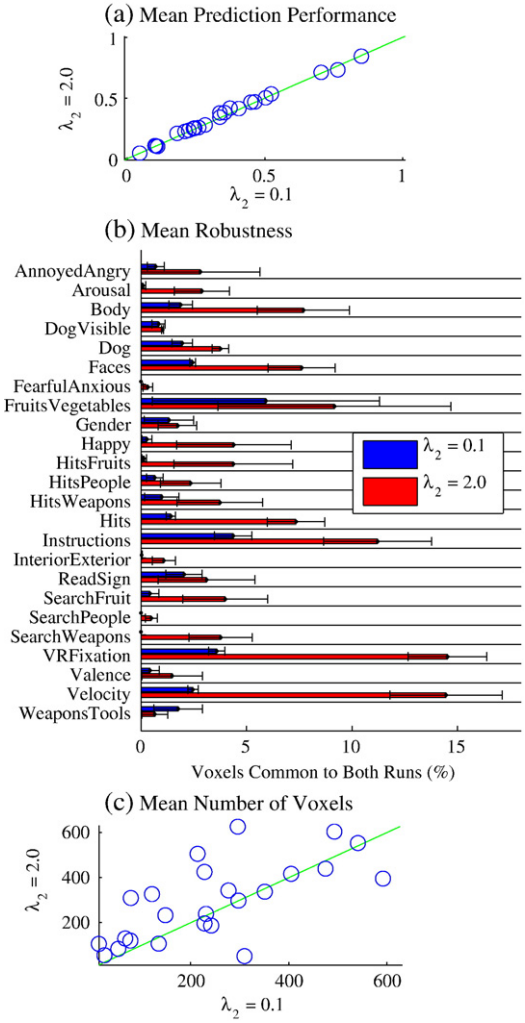

Fig. 5. Even among equally predictive, cross-validated, models (a), increasing the λ_2 parameter increases model robustness (b) while slightly increasing the number of included voxels (c). (a) Mean model prediction performance (correlation with test data) scores are shown, with scores obtained with a higher λ_2 value plotted against their matching scores, obtained with a lower λ_2 value. The unit line indicates no change; points above and to the left of that line reflect improved performance. (b) The mean percentage of specific voxels selected in the models associated with both matched cross-validation runs, out of all specific voxels used in either of the two runs, are shown, with standard error bars, for each predicted response vector. (c) The mean number of voxels selected when a lower λ_2 value (0.1) is used are plotted against the matching mean number of voxels selected when a higher λ_2 value (2.0) is used. The unit line indicates no change; points above and to the left of that line reflect a greater number of voxels. Means in (a) and (c) are over the 3 subjects and 2 cross-validation runs, in (b) over the 3 subjects.

Localization

We have shown that controlling the learning algorithm to select an optimal number of voxels for prediction helps reveal information about the spatial structure of neural response, namely the relationship between prediction, robustness, and spatial distribution. As Fig. 5 depicts, the additional control provided by Elastic Net's λ_2 parameter enables adjustment over model properties not directly related to prediction performance; this control too turns out to facilitate testing of neuroscientific hypotheses. In particular, recall the hypothesis that clusters of correlated voxels would exist and would frequently be localized within the brain. As Fig. 6 demonstrates, increases in λ_2, shown to be associated with greater inclusion from among correlated clusters of voxels, are also associated with decreases in model spatial distribution. Since our spatial distribution metric is essentially independent of the number of voxels included in the model, this decrease in spatial distribution implies the inclusion of more spatially proximal groups of voxels; therefore, the correlated clusters from which more voxels are included are likely to be spatially proximal, consistent with neuroscientific intuition.

This is further exemplified in Figs. 7 and 8: for the Instruction feature (auditory playback of instructions to begin a new block), the figures show the density maps of the absolute value of the regressors, normalized to the subject's anatomy, for λ_2=0.1 and λ_2=2.0. The higher value of the grouping parameter implies a relatively small number of locally extended clusters, whereas the smaller value produces a more globally distributed ensemble of locally restricted clusters. Observe that the maps are not threshold versions of each other: the λ_2=2.0 map tends to overlap with the λ_2=0.1 one, but it takes more contiguous territory; the latter is more spotty and distributed (note that, due to cross-validation, the maps do not necessarily have the same number of active voxels).

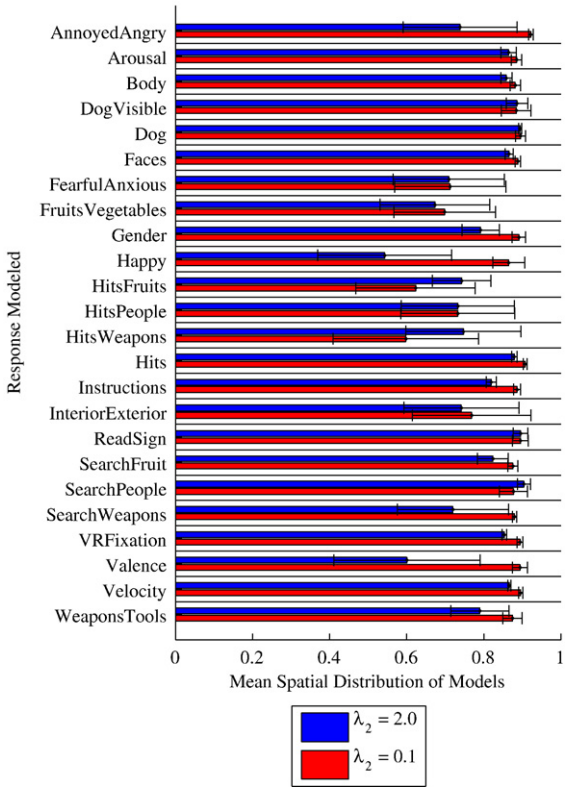

Fig. 6. Increasing the λ_2 parameter is associated with less spatially distributed (more locally clustered) models. Mean spatial distribution of learned models for each predicted response vector are shown, with standard error bars. Means in both plots are over the 3 subjects and 2 cross-validation runs.

More specifically, the difference between the two maps has a functional underpinning. For λ_2=2.0, the most prominent clusters (a total of 25 with more than 50 voxels in the normalized Talairach space — see supplementary material) correspond, as expected, to language areas including BA 21, 22, 37, 41, 42, 44 and 46, along with cerebellar and occipital activations in BA 18 and 19. For λ_2=0.1, the top clusters are shown in Table 1; the first four clusters overlap with clusters identified by the λ_2=2.0 map, but not the last three, which were not even selected as smaller clusters. In order to understand the origin of these differences, we approximated the temporal evolution of each cluster with the first principal component, and compared them against each other. What emerges from this analysis (see supplementary material) is that the λ_2=2.0 clusters tend to be highly correlated with each other, presenting a temporal profile that closely follows the Instructions paradigm. On the other hand, the λ_2=0.1 clusters tend to differ more from each other. This is, in particular, the case for the clusters that do not overlap with λ_2=2.0; i.e. they are less correlated with the top clusters identified by λ_2=0.1, and with the λ_2=2.0 clusters, even those with comparable volume. One of these clusters is highlighted in Fig. 7; it is listed as the last one in Table 1, and corresponds to Brodmann Area 7 (Precuneus), usually associated with working memory processes (Callicott et al., 1999; de Fockert et al., 2001). It is worth noting that the other λ_2=0.1 clusters, even those that do not overlap with any λ_2=2.0 cluster, at least seem to span areas with similar functionality. This highlights the notion that indeed different values of λ_2 reveal functional properties that to some extent can be considered qualitatively different.

PBAIC competition

Even though the focus of this paper is the interplay between prediction and interpretation, it is instructive to describe the technical aspects of our submission to the PBAIC 2007 competition, with the Elastic Net as the core modeling step. Typical of this competition, the full model building was an iterative process. Since Run 3 prediction performance scores were available for each feature, this process included feedback on generalization performance, which was used to revise the modeling decisions. We describe below the full model design used in our final competition submission.

Pre-processing: We observed that both OLS and Elastic Net were hampered by the presence of very low frequencies, not removed by simple detrending techniques. After experimentation with cross-validation, we empirically determined the number of Fourier modes that needed to be removed from the voxel's time traces in order to maximize performance. *Training:* Data from Run 1 was used to train the model, Run 2 was used for parameter optimization, and Run 3 was used for model evaluation. The predictors used to train the algorithm were the vectorized time traces of the voxels, expanded with 4 time-shifted copies, corresponding to shifts +1,+2,−1,−2, padded with zeros. Voxel pre-selection was based on the correlation between the time traces (including the shifted ones) and the corresponding convolved response vector. The pre-selection included all voxels with a correlation above 0.3 times the maximal correlation, selected separately for each response vector and time-shift. The selected predictors (voxels and shifted voxels) were used to train Elastic Net with parameters described in the next section. *Validation:* The final λ_1 values were chosen based on prediction performance on the Run 2 data, separately for every subject and response variable. *Prediction:* The prediction for Run 3 was based on the model selected by the validation step. *Post-processing:* Some of the binary features (e.g. Instructions) were "re-binarized", i.e. the initial prediction was de-convolved (with the hemodynamic response function), binarized and convolved again; this step helped dramatically for the final prediction score.

Even with a stringent correlation cutoff for pre-selection, 10,000 predictors (voxels and time-shifted voxels) may remain, making

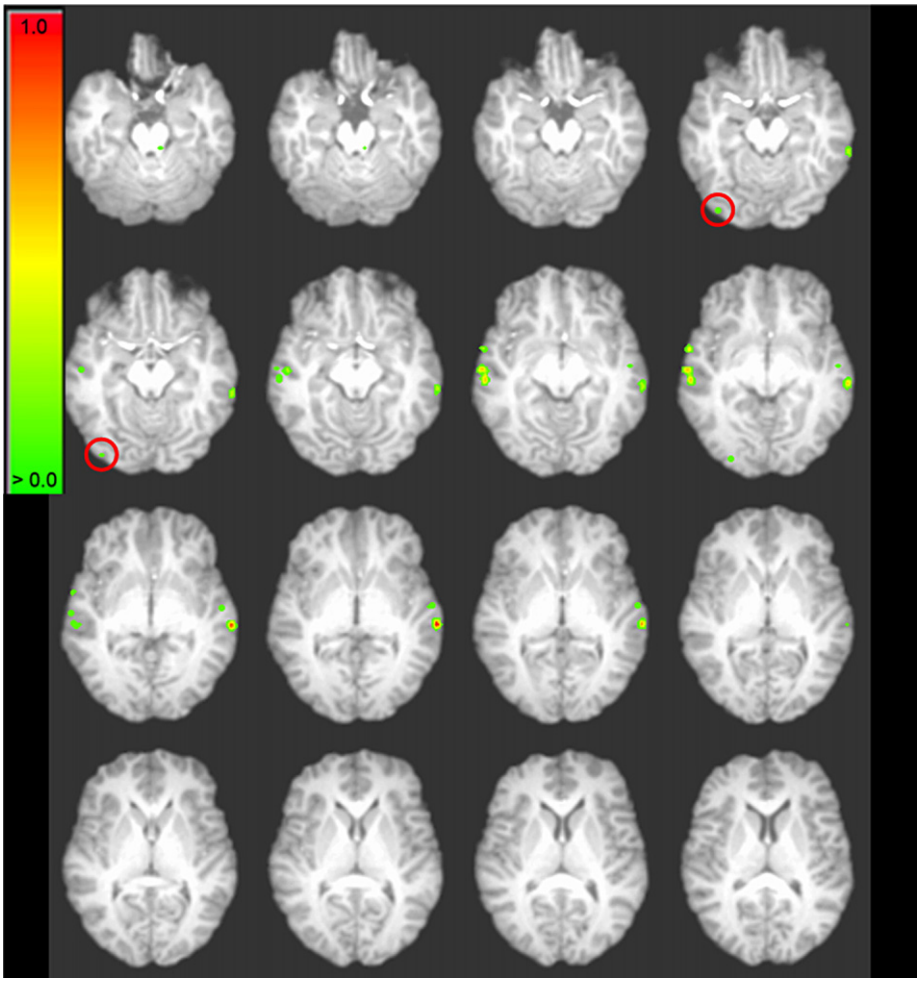

Fig. 7. Distribution and clustering of the models. Absolute values of β weights for the Instruction feature, subject 1, for λ_2=0.1 (radiological view), with associated colorbar. The highlighted cluster is identified in this model but not in the λ_2=2.0 model.

training with no early stopping computationally infeasible given current constraints. Observing that, with the exception of the fixation response variable, optimal λ_1 values tended to be <1000, λ_1 was set to a low value. Through trial-and-error, the optimal trade-off between computation and prediction was found to be 300; we observed a marked difference with smaller values, and only marginal improvements for larger ones, up to 500. We also experimented with training using only non-time-shifted voxels, and were able to train using all voxels (with no pre-selection applied) when λ_1 was set to 300. However, this change does not seem to add significant improvements to the predictions, although we expect that once memory management issues are resolved, the addition of the entire set of time-shifted voxels will prove significant. Finally, we explored a range of λ_2 values. Consistent with previous work and theory, we observed that the presence of a nonzero λ_2 term has a positive effect in the prediction, but the effect saturates very quickly with the parameter value.

The final submission was ranked 12th based on the average score of the features; a fair comparison with other regression methods cannot be made, as post-processing had a decisive effect on performance, and we did not do binarization in all the features amenable to it (for example, the Instructions feature Pearson correlation can be improved from 0.7 to 0.99).

Discussion

Indisputably, prediction is an essential component of scientific modeling, and great effort should be put into maximizing it; however, as shown in this paper, equally predictive models can still be markedly different. In fMRI analysis, the core goal underlying predictive modeling is production of a model that can be interpreted to pinpoint all relevant voxel activity and exclude all irrelevant activity. Therefore, it is crucial to not lose sight of the interpretation of the resulting models in the quest to optimize prediction performance.

Innumerable techniques have been developed for choosing an appropriate set of voxels from which to build models. Not surprisingly, these techniques are most often evaluated based on the prediction performance of the resulting models. As our results confirm, selecting a set of voxels that optimizes prediction performance is indeed critical for model interpretation. When fixed numbers of voxels are chosen, the resulting models are frequently overfit to training data, and therefore implicate many voxels that are actually irrelevant. Since these spurious voxels will tend to be randomly distributed throughout the brain, these models might misrepresent more general neuroscientific characteristics, for instance by overestimating the spatial distribution of response. Elastic Net is able to select an optimal number of voxels by automatically selecting voxels as part of the modeling process and computing the full regularization path, which facilitates cross-validation. The standard error bars for spatial distribution in Fig. 3 indicate considerable variability in spatial distribution even when cross-validating when λ_2=2.0, yet our results indicating increased model robustness as λ_2 is increased suggest that with higher, better optimized, values of λ_2, the spatial distribution of the models for a given response vector might converge.

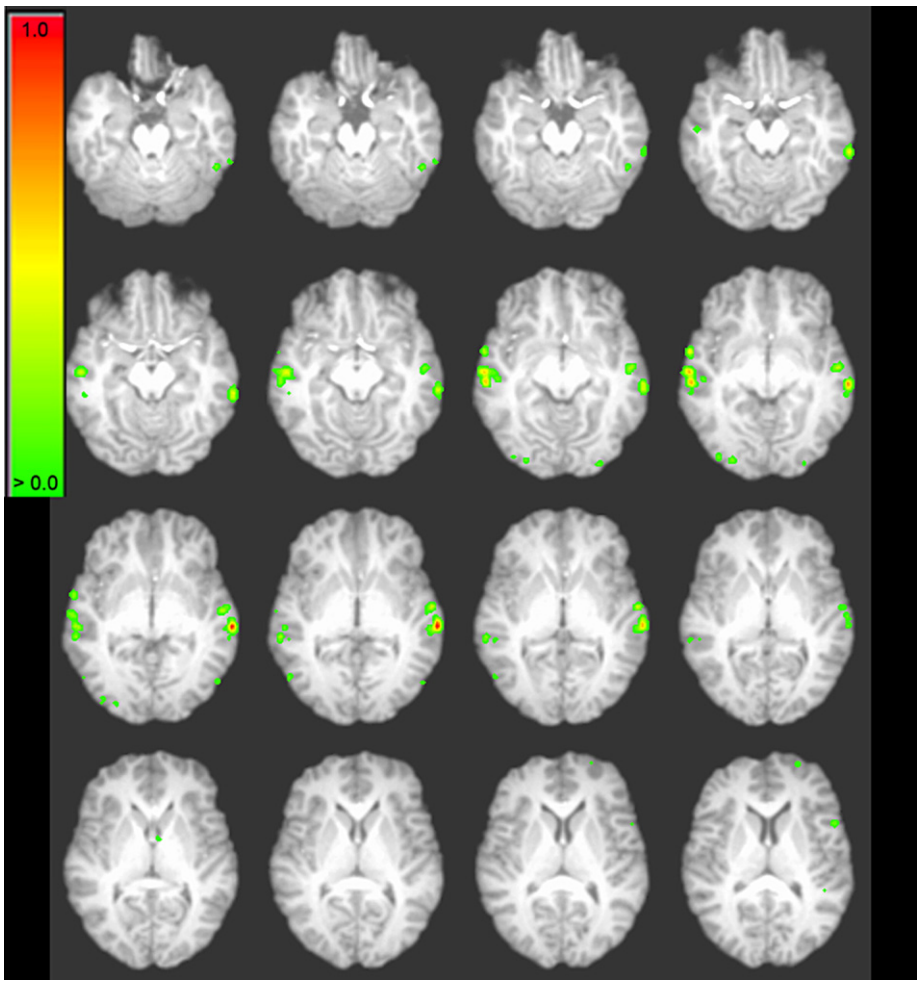

Fig. 8. Distribution and clustering of the models. Absolute values of β weights for the Instruction feature, subject 1, for λ_2=2.0 (radiological view), with associated colorbar. The clusters are bigger and include many, but not all of the λ_2=0.1 clusters.

When choosing a validated set of voxels, the true importance of spatial distribution becomes clear; the most predictive models draw on information from the most distributed regions of the brain. However, we should be conservative when drawing conclusions about the relationship between spatial distribution and prediction performance from these results. We found no significant correlation between the degree to which a particular response variable's mean model spatial distribution increased when cross-validation was used, and the corresponding improvement in prediction of that response variable. Perhaps these results imply that modeling highly distributed neural responses is easier than modeling highly localized responses; such a finding would be consistent with the observed exclusion of voxels from localized clusters when low λ_2 values are used. Therefore experimenting with even higher λ_2 values may address this question. Still, in conjunction with the inferior yet impressive predictive performance of random voxels, these findings underscore that the highly distributed nature of neural response necessitates the use of multivariate methods, not tied to localized regions, which can be validated by prediction performance.

We have also shown that being preoccupied with prediction performance can be equally destructive. Models that function as highly predictive "black boxes" might be useful for neuro-engineering "mind reading" efforts, but for informing neuroscience, these models should also be reliable and valid. At the least, we would expect useful models of the PBAIC response vectors to incorporate the same, or very similar, sets of voxels when trained on data from the two experimental runs. Elastic Net goes beyond other sparse modeling approaches by facilitating, with only one additional parameter, an increase in model robustness without compromising prediction performance. Examining the findings more closely by manipulating the λ_2 parameter, we observe that neural response is marked by clusters of correlated voxel activity. Existing sparse methods, which effectively use a very low λ_2 value, will tend to include only one voxel from each cluster; the voxel chosen is in effect arbitrary due to minor fluctuations in the dataset. By selecting more voxels from within these clusters that are redundant for prediction yet relevant to the task, Elastic Net with a higher λ_2 value achieves more robust, and hence valid, models.

We can be even more emphatic, but speculative, in our interpretation of the results, by extrapolating the observation that,

Table 1
Principal clusters of activation for λ_2=0.1

Voxels	x_p	y_p	z_p	∩	Description
844	62	−24	−1	Y	Left BA 21/22
807	−62	−15	−6	Y	Right BA 21/22
167	52	−12	−7	Y	Left BA 21/22
153	−62	2	−5	Y	Right BA 21/22
100	−26	−84	28	N	Right BA 18/19
58	−40	−80	−13	N	Right BA 18/19
58	20	−66	57	N	Left BA 7*

The coordinates correspond to the peak activation (i.e. absolute value of the regressor) within the cluster; the ∩ column indicates whether the cluster overlaps with a cluster identified by λ_2=2.0 (*highlighted in Fig. 7).

at least for a good feature such as Instruction, two alternative models (as shown in Figs. 7 and 8) can achieve an almost identical prediction performance: the only possible way to avoid this model degeneracy is to introduce further functional constraints, including local and global neural dynamics, and perhaps challenge the idea that brain function arises from sufficiently repeatable spatio-temporal patterns. By the same token, the fact that the maps are not thresholded versions of each other also emphasizes that brain processes occurring at several temporal and spatial scales can be detected by fMRI. To make progress along these challenging lines, however, we need stronger and more encompassing theoretical tools than we have at our disposal.

These results suggest the promise of Elastic Net for fMRI modeling, although the field is still quite far from producing models that are as predictive, and certainly as robust, as needed to answer the pressing modern neuroscientific questions. For instance, regardless of λ_2 value, the models built on the PBAIC 2007 data all showed poor robustness, with no response vector averaging better than 17% overlap. This finding highlights the important point, however, that models produced are only as good as the data used to train them. While the PBAIC experiments were run carefully, the experimenters will agree that the response variables analyzed are highly challenging to model. The response variables are sometimes overly broad and sometimes overly specific, and little is known about how they might map to underlying function. Our results reflect this variability among the response variables, with some responses clearly more "easily" modeled than others. The diversity and complexity of these response variables are in fact the very reasons they appear in such a generalized competition; the experimenters have been highly impressed by even the prediction capabilities observed using all methods thus far. We intend to continue exploring the issue of robustness empirically as well as theoretically on better understood fMRI datasets.

Still, the effect of λ_2 on robustness is clear from these results, and the control provided by λ_2 is shown in this paper to have value even beyond improving robustness. This parameter has a well-understood effect on the resulting models; as such, it is an attractive candidate independent variable for testing neuroscientific hypotheses. In our experiments, this variable served as a proxy for the degree to which members of correlated clusters are included in models. Since greater inclusion is associated with decreased spatial distribution, we can conclude that these correlated clusters are frequently localized in space. Combined with our findings about the importance of spatial distribution, a picture emerges of neural response characterized by patterns of localized clusters distributed highly throughout the brain. Many existing techniques exist solely to attempt to extract these localized clusters directly, for instance by setting thresholds on the extent of cross-correlation (Forman et al., 1995). Elastic Net, by selecting voxels automatically during modeling, offers the advantage of automatically selecting these clusters without the need for separate methods or thresholds, and does so in a multivariate context, on a per response variable basis, so that all clusters relevant to a task, but none that are irrelevant, are selected.

As demonstrated preliminarily in this paper, once a model has proven itself trustworthy, the next step is naturally to extract information from it. Models exhibiting satisfactory prediction performance and robustness can be easily explored visually by plotting the β weights from the models according to corresponding voxel location. The presence of voxels within specific ROIs can be determined and techniques can be developed to explicitly "cluster" the voxel sets so as to visually isolate the localized clusters. It might also be interesting to compare the "clusters" generated by Elastic Net with the "dimensions" extracted through dimensionality reduction techniques such as ICA and Sparse PCA. In addition, we have shown that the spatial distribution metric introduced in this paper, while simple, can reveal insights about neural functioning. Subsequent efforts might use this measure to explore differences in neural representation across tasks and between subjects. Predictive modeling techniques clearly offer great promise for knowledge discovery, but this ultimate goal of their production must be considered.

Appendix A. LARS-EN Algorithm

The Elastic Net functional in Eq. (3) can be easily transformed into the LASSO functional (Zou and Hastie, 2005)

$$L(\gamma, \beta^*) = ||\boldsymbol{y}^* - \mathbf{X}^*\beta^*||_2^2 + \gamma||\beta^*||_1 \quad \text{(A1)}$$

where $\gamma = \lambda_1/\sqrt{(1+\lambda_2)}$, and $\mathbf{X}^*$, $\boldsymbol{y}^*$ are properly "augmented" versions of $\mathbf{X}$ and $\boldsymbol{y}$ of size $(n+p)\times p$ and $(n+p)\times 1$, respectively.

The LARS algorithm is very similar to Forward Stagewise regression, which is a "cautious" version of "forward stepwise regression" (Weisberg, 1980), a simple iterative approach to variable selection and regression. However, LARS is more efficient than Forward Stagewise as it makes larger steps, which are still cautious compared to the straightforward greedy method. LARS starts with an empty set of predictors and selects the one having the largest absolute correlation with the response; however, it proceeds along the selected direction only up to the point that another predictor becomes equally correlated (in the absolute sense) with the *current residual*. Then, LARS chooses a new direction equiangular between the two predictors and continues moving along this direction until some third predictor enters the "most correlated" set (also called the *active set*). LARS chooses the new direction equiangular between the three active predictors, and so on, until it includes the desired number of predictors, specified as an input to the algorithm. It was shown by Efron et al. (2004) that, under a very minor modification, LARS finds an *optimal* solution to the LASSO problem in Eq. (4). This result and the computational efficiency of LARS made it the algorithm of choice for solving various sparse regression problems, including the Elastic Net. The LARS-EN algorithm for solving Elastic Net (Zou and Hastie, 2005) essentially uses LARS (Efron et al., 2004) to minimize the above functional, while exploiting the specific structure of the augmented data matrix to improve the efficiency of LARS. In our analysis, we used a publicly available package (Sjöstrand, 2005).

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2008.08.020.

References

Calhoun, V.D., Adali, T., Hansen, L.K., Larsen, J., Pekar, J.J., 2003. ICA of Functional MRI Data: An Overview. In: 4th international symposium on Independent Component Analysis and Blind Signal Separation (ICA2003).

Callicott, J.H., Mattay, V.S., Bertolino, A., Finn, K., Coppola, R., Frank, J.A., Goldberg, T.E., Weinberger, D.R., 1999. Physiological characteristics of capacity constraints in working memory as revealed by functional MRI. Cereb. Cortex 9 (1), 20–26.

Chigirev, D., Stephens, G., The-Princeton-EBC-Team, 2006. Predicting base features with supervoxels. Abstract presented, 12th HBM meeting, Florence, Italy.

Cox, D., Savoy, R., 2003. Functional magnetic resonance imaging (fMRI) brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19, 261–270.

de Fockert, J., Rees, G., Frith, C., Lavie, N., 2001. The role of working memory in visual selective attention. Science 291, 1803–1806.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32 (1), 407–499.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96 (2), 1348–1360.

Forman, S., Cohen, J., Fitzgerald, M., Eddy, W., Mintun, M., Noll, D., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn. Reson. Med. 33, 636–647.

Frank, I., Friedman, J., 1993. A statistical view of some chemometrics regression tools. Technometrics 35 (2), 109–148.

Fu, W., 1998. Penalized regression: the bridge versus the lasso. J. Comput. Graph. Stat. 7 (2), 397–416.
Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293 (5539), 2425–2430.
Hoerl, A., Kennard, R., 1988. Ridge regression. Encycl. Stat. Sci. 8 (2), 129–136.
Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M., Newman, S., 2004. Learning to decode cognitive states from brain images. Mach. Learn. 57, 145–175.
Norman, K., Polyn, S., Detre, G., Haxby, J., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10 (3), 424–430.
Okabe, A., Boots, B., Sugihara, K., Chiu, S., 2000. Spatial Tesselations: Concepts and Applications of Voronoi Diagrams2nd Ed. Wiley.
Pittsburgh-EBC-Group, 2007. PBAIC Homepage: http://www.ebc.pitt.edu/2007/competition.html.
Sjöstrand, K., 2005. Matlab implementation of LASSO, LARS, the elastic net and SPCA. Version 2.0. URL http://www2.imm.dtu.dk/pubdb/p.php?3897
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc., Ser. B. 58 (1), 267–288.
Ulfarsson, M., Solo, V., 2007. Sparse variable principal component analysis with application to fMRI. In: Proc. IEEE International Symposium on Biomedical Imaging (ISBI07).
Weisberg, S., 1980. Applied Linear Regression. Wiley, New York.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the Elastic Net. J. R. Stat. Soc., Ser. B. 67 (2), 301–320.