Multi-Voxel Pattern Analysis of fMRI Data

Jarrod A. Lewis-Peacock[1,2] and Kenneth A. Norman[3,4]

[1] Department of Psychology, University of Texas at Austin
[2] Imaging Research Center, University of Texas at Austin
[3] Department of Psychology, Princeton University
[4] Princeton Neuroscience Institute, Princeton University

Total words in manuscript: 6350
Figures: 2 (1 color)
Keywords: fMRI, multi-voxel pattern analysis, MVPA

Please address correspondence to:

Jarrod A. Lewis-Peacock
Department of Psychology
University of Texas at Austin
108 E. Dean Keeton
Austin, TX 78712
Phone: (512) 475-6836
Fax: (512) 471-6175
Email: jalewpea@utexas.edu

Abstract

The central goal of cognitive neuroscience is to understand how information is processed in the brain. To accomplish this goal, researchers studying human cognition are increasingly relying on multi-voxel pattern analysis (MVPA); this method involves analyzing spatially distributed (multi-voxel) patterns of functional MRI activity, with the goal of decoding the information that is represented across the ensemble of voxels. In this chapter, we describe the major subtypes of MVPA, we provide examples of how MVPA has been used to study neural information processing, and we highlight recent technical advances in MVPA.

1. Introduction

Cognitive neuroscience theories deal with information processing: What information is represented in different brain structures, how is this information transformed over time, and how is it used to guide behavior? Functional MRI (fMRI) constitutes a powerful tool for addressing these questions: While a subject performs a cognitive task, we can obtain estimates of local blood flow (a proxy for local neural processing) from tens of thousands of distinct neuroanatomical locations (*voxels*, or volumetric pixels), within a matter of seconds.

Traditional univariate fMRI analysis methods have focused on characterizing how cognitive variables modulate the activity of individual brain voxels (or small clusters of voxels; e.g., Gonsalves and Cohen, 2010). The goal of this chapter is to describe a different approach to fMRI analysis that focuses on extracting information about a person's cognitive state (i.e., a snapshot of the person's current internal thought space) from spatially distributed, multi-voxel patterns of fMRI activity. This approach is referred to as multi-voxel pattern analysis (MVPA) (Haxby et al., 2001; Kamitani and Tong, 2005; Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006; Pereira, Mitchell, & Botvinick, 2009). Over the past decade, this approach has become ubiquitous in fMRI research, and its adoption has led to novel discoveries about the brain bases of perception, attention, imagery and working memory, episodic memory, semantic knowledge, language processing, and decision making (see Rissman & Wagner, 2012; Tong and Pratte, 2012).

Given the goal of detecting the presence of a particular cognitive state in the brain, the primary advantage of MVPA methods over individual-voxel-based methods is increased sensitivity for detecting this information. Conventional fMRI analysis methods try to find voxels that show a statistically significant response to the experimental conditions. To increase sensitivity to a particular condition, these methods spatially average across voxels that respond significantly to that condition. Although this approach reduces noise, it also reduces signal in two important ways: First, voxels with weaker (i.e., non-significant) responses to a particular condition might carry some information about the presence/absence of that condition. Second, spatial averaging blurs out fine-grained spatial patterns that might discriminate between experimental conditions.

Like conventional methods, the MVPA approach also seeks to boost sensitivity by looking at the contributions of multiple voxels. However, to avoid the signal-loss issues mentioned above, MVPA does not routinely involve uniform spatial averaging of voxel responses. Instead, the MVPA approach uses a weighted average of responses, treating each voxel as a distinct source of information about the participant's cognitive state. The technique finds ways to optimize these weights, and then aggregates this (possibly weak) information across voxels to derive a more precise sense of what the participant is thinking. The multi-voxel response can be thought of as a combinatorial code for representing distinctions between cognitive states (see Figure 1). Because MVPA analyses focus on high-spatial-frequency (and often idiosyncratic) patterns of response, they are typically conducted within individual subjects, although recent advances in data alignment procedures have paved the way for the expansion of classification analyses beyond the individual subject (Haxby et al., 2011; See Section 4. *New Developments*).

-- [Figure 1 here] --

Broadly speaking, the term MVPA has come to encompass two distinct methods. The first method involves using *pattern classification* methods imported from machine learning to learn a mapping between multi-voxel brain states and cognitive state information. This approach flips standard univariate fMRI analysis on its head: Standard voxel-based analysis uses multiple regression to predict the activity of individual voxels, based on the participant's cognitive state. By contrast, classification-based MVPA uses multiple regression to predict the participant's cognitive state, based on the activity of multiple voxels. The second major subtype of MVPA does not use pattern classifiers; rather, it examines the *similarity structure* of multi-voxel patterns (i.e., which patterns are similar to one another) and uses this similarity structure information to draw conclusions about what information is reflected in these patterns.

In Section 2, we provide an overview of these two subtypes of MVPA. In Section 3, we illustrate how MVPA has been used to study information representation and processing in the brain. In Section 4, we discuss recent advances in MVPA that allow for finer-grained mappings between brain activity and cognitive states within individuals, and also new methods for aligning and combining brain data across individuals.

Importantly, while this chapter is focused on fMRI, we should emphasize that most of the MVPA methods described here can be applied to other imaging modalities as well (for applications to EEG and MEG data, see, e.g., Jafarpour, Horner, Fuentemilla, Penny, & Duzel, 2013; for applications to direct neural recording data, see, e.g., Hung, Kreiman, Poggio, & DiCarlo, 2005).

2. Mechanics of MVPA

Here, we will review the basic procedures of MVPA. All pattern analyses start with *preprocessing* of the raw fMRI BOLD data, including temporal and spatial realignment, noise filtering, and z-scoring of the data (over time, within each voxel) within each run. Next, *feature selection* chooses which voxels will be included in the analysis. All voxels in the brain can be used, but it is often advantageous to limit the analysis to certain voxels. One way to select features is to limit the analysis to specific anatomical regions (e.g. Haxby et al., 2001 focused on ventral temporal cortex in their study of visual object processing). Univariate statistics used in conventional fMRI analysis (e.g., Mitchell et al., 2004) and newer multivariate "wrapper methods" (Guyon and Elisseeff, 2003) can also be used for feature selection (e.g., one can discard the voxels that — taken on their own — do the worst job of discriminating between conditions). Finally, *pattern assembly* involves sorting the data into discrete "brain patterns" corresponding to the pattern of activity across the selected voxels at a particular time in the experiment. Patterns can be assembled using the preprocessed fMRI signal for each trial or, alternatively, by using multiple regression to estimate the unique neural response in each voxel for each trial (Mumford, Turner, Ashby, & Poldrack, 2012). Brain patterns are labeled according to which cognitive state (or experiment condition, stimulus, response, etc.) generated the pattern; this labeling procedure needs to account for the fact that the hemodynamic response measured by the scanner is delayed and smeared out in time, relative to the instigating neural event. Once the patterns have been assembled, MVPA can proceed along two main branches of analysis:

classifier-based MVPA and *pattern-similarity MVPA*. We will now discuss both methods in turn.

2.1. Classifier-based MVPA

There are two steps to classifier-based MVPA. The first step, *classifier training*, involves feeding a subset of labeled patterns into a multivariate pattern classification algorithm. Based on these patterns, the classification algorithm learns a function that maps between voxel activity patterns and cognitive states. As illustrated in Figure 2, brain patterns can be viewed as points in a multidimensional voxel space; the goal of the classifier is to find a decision boundary in this space that best separates the patterns associated with the to-be-discriminated cognitive states.

-- [Figure 2 here] --

The second step is generalization testing: Given a new pattern of brain activity (not previously presented to the classifier), can the trained classifier correctly determine the cognitive state associated with that pattern?

The most commonly used classifiers are linear classifiers, which derive a linear decision boundary between classes. At training, the classifier learns a weight for each voxel, plus an intercept term; collectively, these weights determine the equation of the hyperplane (in two dimensions, a line) that forms the decision boundary. At test, the classifier uses these weights (and intercept) to compute a weighted sum of voxel activity values, and it uses this weighted sum to determine whether the test pattern falls on one side or the other of the decision boundary. There are a wide range of linear classification algorithms; the main difference between these algorithms relates to *which features of the data* they use when modeling the data. The multidimensional clouds of data for each class can be characterized in terms of their mean value and also their covariance matrix. This matrix specifies the spread of the cloud along each voxel dimension (i.e., how tall/wide is the cloud) and also the covariance between each pair of dimensions (i.e., the tilt of the cloud). The simplest classifier is the minimum distance classifier (e.g., Haxby et al., 2001), which estimates the mean value for each class based on the training data and then classifies new points based on their proximity to these means. However, Figure 2 demonstrates how ignoring the covariance matrix can produce non-optimal decision boundaries. More complex linear classifiers (e.g., Fisher's linear discriminant; Duda, Hart, & Stork, 2012) can converge on the optimal decision boundary by creating a more sophisticated model of the data: In addition to estimating the mean, they also model the class-conditional probability densities (i.e., they estimate the full covariance matrix within and between voxel dimensions for each class). Figure 2 shows how the boundary learned by a linear-discriminant classifier factors in the "tilt of the ellipse" for each data cloud. Nonlinear classifiers (e.g., k nearest neighbor, multilayer neural networks) can form even more complex decision boundaries.

Classification of fMRI data is a challenging problem, for several reasons: First, the number of data points (brain patterns) that are available for training tends to be small relative to the number of parameters in the model. For example, Polyn, Natu, Cohen, & Norman (2005) trained a classifier on 450 total brain patterns (150 for each of three stimulus classes) per participant, where each brain pattern consisted of approximately 7,000 voxels. In this situation, the covariance matrix has millions of unique entries (corresponding to all of the voxels, plus all of the unique *pairings* of voxels); each of these entries is a parameter that needs to be estimated based on only a few hundred training patterns. Further adding to the complexity of this problem, the brain patterns are very noisy (i.e., the clouds are highly dispersed). In this kind of situation, where the data are noisy and the number of parameters being estimated by the classifier dwarfs the number of training patterns, classifiers are prone to *overfitting* the noise in the training data: That is, the classifier may learn idiosyncratic features of the training examples rather than the actual distinction between the classes, thereby leading to poor generalization.

Overfitting is the main obstacle to achieving good fMRI classification. One way to combat overfitting is to collect more data, but there are practical limits on collecting more data per participant. In Section 4 of this chapter, we will discuss new developments in MVPA that allow us to obtain more data by combining across subjects. The other way to combat overfitting is to try to limit the complexity of the classifier. For example, Gaussian Naive Bayes classifiers (GNB; Pereira et al., 2009) simplify the modeling of the covariance matrix by treating the *n* dimensions of the data as independent (such that the off-diagonal elements of the covariance matrix are zero), thus reducing the number of parameters to estimate from n^2 to n. Support vector machines (SVMs; Cox and Savoy, 2003) achieve complexity control by defining the category boundary in terms of a small number of support vectors (i.e., training exemplars close to the decision boundary). Another way to limit the number of free parameters is to limit the number of voxels used for classification (e.g., restricting classification of oriented gratings to low-level visual cortex, Kamitani and Tong, 2005; or restricting classification of faces and scenes to ventral temporal cortex, Kuhl, Rissman, Chun, & Wagner, 2011). This is a useful approach when there is a priori knowledge of strong selectivity for the classes in particular brain regions.

An effective way to reduce the complexity of linear classifiers is to add a *regularization* parameter to the model that punishes undesirable properties of the solution (e.g. large weights on individual voxels). Common forms of regularization are L2 regularization, which penalizes the sum of squares of the voxel weights, and L1 regularization, which penalizes the absolute value of the weights. As the regularization parameter is increased, L2 regularization pulls in extreme voxel weights (resulting in a smoother distribution of weights), whereas L1 regularization causes some weights to be driven to zero (a sparser solution). In both cases, the regularization parameter limits the space of possible solutions, thereby reducing the flexibility of the classifier and reducing overfitting.

Practically speaking, all of the above forms of complexity control (GNBs, SVMs, voxel reduction, and regularization) have been shown to improve generalization performance, relative to linear classifiers that do not incorporate complexity control. The only exception is when the number of voxels is very small or the number of training patterns is very large, at which point it becomes feasible to estimate the full covariance matrix. Importantly, with fMRI data, nonlinear classifiers virtually never outperform linear classifiers on generalization tests — the added flexibility of these classifiers leads to overfitting.

2.2. Pattern-similarity MVPA

The second major form of MVPA is pattern similarity analysis (e.g., Kriegeskorte, Mur, Bandettini, 2008a). Here, brain patterns are viewed as points in high-dimensional voxel space, where the distance between points indicates the similarity of the patterns. Rather than specifying which features of the data to separate with a classifier, pattern similarity analysis summarizes the space using a matrix that records the distance between each pair of points. This matrix can be viewed as a neural "fingerprint" of the representational space. Although information about the exact positions of the points is lost, the information about similarity structure contained in the pairwise similarity matrix is highly diagnostic of what information is coded in that region (e.g., if items with similar shapes elicit similar neural patterns, but items with similar sizes do not, this indicates that the region is more sensitive to shape than size information).

The final step in pattern-similarity MVPA is to compare the neurally derived similarity matrix to some other similarity matrix (e.g., to a matrix holding a cognitive model's predictions about the conceptual similarity between stimuli). The comparison between these matrices is used to evaluate the quality of the model's predictions. A key benefit of the pattern-similarity approach is that — in contrast to the pattern-classification approach outlined above — it is not necessary to explicitly specify (ahead of time) the dimensions of cognitive variance that are of interest. Rather, all of the requisite analyses can be carried out post-hoc (e.g., to see if an area represents the size of an object, look at whether objects that are similar in size gave rise to similar neural patterns).

3. Applications of MVPA

In this section, we will describe three common uses of multi-voxel pattern analysis: (1) classifier-based thought tracking, (2) classifier-based information mapping, and (3) information mapping based on pattern similarity. We will discuss the goals of each analysis and we will review some recent applications of each method.

3.1. Classifier-based thought tracking

The goal of the classifier-based thought tracking approach is to measure participants' thoughts on a trial-by-trial basis, to characterize the dynamics of these thoughts, and to assess how they relate to behavior. This approach is used when the main concern is tracking a particular latent cognitive state, and there is relatively less concern about how that cognitive state is represented in the brain (although this approach can be applied to specific regions of interest to localize cognitive representations).

Compared to univariate methods, MVPA squeezes more information about the participant's cognitive state out of each snapshot of fMRI data, thereby increasing the effective temporal resolution of fMRI analysis and making it possible to record trajectories of cognitive states over time. However, even with the added sensitivity of MVPA, not all cognitive states are equally "visible" to fMRI. In this situation, researchers often find it useful to take the cognitive state of interest and link it to something that we know is highly visible with fMRI: stimulus *category* information (e.g., faces and scenes). Consider an analogy: When injecting contrast dyes in neuroanatomy, we don't care whether the dye stains cells green or red, so long as the colors are visible under the microscope and so long as the different stains we are using (to measure different cellular properties) have distinct colors. Likewise, when we attach cognitive states to faces or scenes, for example, we don't do this because we care about faces or scenes per se; rather we do this because thoughts about faces and scenes are highly visible and differentiable with fMRI.

This type of MVPA analysis has been used to study various aspects of memory and cognition. For example, Polyn et al. (2005) used classifiers in a free recall experiment and showed that category-specific patterns of activity emerged about 6 seconds prior to verbal recalls from a given category. In a more recent study, Zeithamova, Dominick, & Preston (2012) used classifier-based thought tracking to explore the process of memory integration. Classifiers tracked the reinstatement of object and scene category information during repeated exposures to AB and BC stimulus pairs (e.g., frog-bucket and bucket-scene). Across subjects, the degree of reactivation of the C item (in this example, the scene) during AB exposures was positively correlated with later performance on a transitive inference memory test for the A-C association; the authors explain this result in terms of participants binding the (reactivated) C item to the A item at encoding. For other recent examples of classifier-based thought tracking, see Lewis-Peacock, Drysdale, Oberauer, & Postle (2012) and Detre, Natarajan, Gershman, & Norman (2013).

3.1.1. Cautionary notes for thought-tracking studies

The ideal situation for thought-tracking is to get independent readouts of the relevant cognitive states, but achieving this goal can be difficult. Classifiers are opportunistic: If two categories are anticorrelated in the training set (e.g., all training patterns are either faces or scenes, never both) the classifier will learn this negative correlation, and it will come to treat the lack of scene activity as strong evidence for the presence of faces (see Kuhl et al., 2011, for discussion of this issue). Training on additional categories alleviates this problem by reducing the size of the negative correlation between categories at study (e.g., if there are faces, scenes, and objects, then the absence of faces does not perfectly predict the presence of scenes).

3.2. Classifier-based information mapping

A second application of MVPA is less concerned with getting a useful readout of information processing during individual trials, and more concerned with assessing whether a particular fine-grained distinction is represented in a particular brain region (e.g., Pereira and Botvinick, 2011). This analysis is similar in concept to the mass-univariate approach, in that the goal is to determine which brain regions are responsive to a particular cognitive process. However, rather than considering how the activity in each individual voxel is predicted by a person's (presumed) cognitive state, classifier-based information mapping uses information from multiple voxels simultaneously to predict the person's cognitive state.

This analysis can be done using many different a priori regions of interest, or it can done using the searchlight method (e.g., Kriegeskorte, Goebel, Bandettini, 2006). This method consists of constructing a "searchlight" of voxels and sliding this searchlight all around the three-dimensional brain volume. For each placement of the searchlight, you consider the multi-voxel pattern of activity within that searchlight. A classifier is trained on these patterns, and then used to assess how informative these patterns are about the cognitive state of interest.

This approach has been used to discover new insights into cognition and the localization of function in the brain. For example, Soon, Brass, Heinze, & Haynes (2008) used the searchlight technique to discover brain regions whose activity patterns were predictive of future decisions. They found that the outcome of a simple decision (to press a left or right button) could be decoded from prefrontal and parietal cortices up to 10 seconds prior to this decision entering awareness.

3.2.1. Cautionary notes for classification-based information mapping

An important caveat for the information-mapping approach is that above-chance decoding - which signals that a brain region contains information about a particular cognitive distinction - does not necessarily imply that this region is involved in guiding behavior based on that distinction (e.g., Williams, Dang, & Kanwisher, 2007). Furthermore, information-mapping is opportunistic and may produce false positive results; see Todd, Nystrom, & Cohen (2013) for discussion of how MVPA can be more susceptible than univariate analysis to experimental confounds (e.g., task difficulty). Finally, multivariate decoding is not necessarily more sensitive than univariate decoding (Jimura and Poldrack, 2012). If the underlying signal has a coarse spatial scale, then univariate approaches using spatial smoothing at this scale will outperform MVPA. In this case, the extra parameters used to model the data in MVPA can lead to overfitting (Kriegeskorte et al., 2006).

3.3. Pattern similarity analysis

The goal of pattern similarity analysis of fMRI data (e.g., Kriegeskorte et al., 2008a) is to make inferences about the similarity of mental concepts based on the similarity of patterns of brain activity elicited by those concepts. The strength and versatility of this approach comes from the many different ways that similarity matrices can be computed and thus permits many types of comparisons. For example, Kriegeskorte et al. (2008b) showed that the similarity structure of neural patterns in human IT cortex (measured using fMRI) resembles the similarity structure of neural patterns in monkey IT cortex (measured using electrophysiology). Furthermore, they showed that the similarity structure of neural patterns in both human and monkey IT (specifically, clustering into animate vs. inanimate objects) could not be explained purely in terms of the low-level visual features of the stimuli.

Pattern similarity analysis is frequently used to study the representation of item-specific information — the logic here is that regions that differentiate items within a category should show greater pattern similarity between two instances of the same item, compared to two distinct items from the same category. For example, Ritchey, Wing, Labar, & Cabeza (2012) found evidence that encoding-retrieval similarity at the individual item level predicted memory success. Similarity between an item's neural representation and the neural representations of *other* studied items has also been used to predict memory performance. For example, LaRocque et al. (2013) found that greater levels of across-item pattern similarity in perirhinal and parahippocampal cortices were associated with better recognition memory performance.

3.3.1. Cautionary notes for pattern-similarity analysis

As described above, classifiers compute weighted combinations of features that discriminate between classes; uninformative or noisy features may be effectively "filtered out" by being assigned small weight values. In contrast, pattern similarity analyses do not compute weights for each voxel − these analyses treat all voxels as equally important. For this reason, pattern similarity analyses are more susceptible to contamination from uninformative or noisy features than classifiers. Another concern is that pattern similarity results can be influenced by univariate effects. For example, imagine that a 33-voxel searchlight contains a 10-voxel subregion that tracks memory strength, such that all 10 of these voxels activate together for remembered (but not forgotten) items. This will increase the average pattern similarity between remembered items. Naively, one might interpret this effect in terms of neural representations "converging" in representational space, when (in fact) it is merely due to a univariate effect being superimposed on the searchlight region.

4. New developments

In this section, we will discuss recent advances in MVPA that complement and extend existing approaches.

4.1. Decoding and encoding representational spaces

A major limitation of the classifier studies discussed above is that the classifiers are *specialists*: They can only discriminate between cognitive states that they were trained to discriminate, and the training process is highly laborious. The classifier needs to be trained on a large number of "snapshots" of these cognitive states (on the order of hundreds or more, depending on how subtle the differences are between the cognitive states) before it can discriminate between them reliably. You can train a classifier to discriminate between brain patterns elicited by lions and camels, but this classifier won't tell you anything about the difference between oranges and grapes. This fact places a strong limitation on the kinds of questions that can be addressed in any particular study.

Recently, several studies have sought to surmount this limitation by reconceptualizing the decoding problem: Instead of treating stimulus classes as distinct entities, these studies draw on the psychological literature on representation and conceptualize psychological states as points in a high-dimensional representational space. For example, the meaning of a particular concrete noun can be conceptualized as a point in a high-dimensional "meaning space", where each dimension corresponds to a particular aspect of the noun's meaning (e.g., can it be eaten? can it be manipulated? can it be used as shelter?) (Just, Cherkassky, Aryal, & Mitchell, 2010).

Once stimuli have been placed in an n-dimensional feature space, classifiers can be trained to decode *each feature dimension* (e.g., what does the brain look like for nouns that describe edible vs. inedible items). These classifiers can then be applied to a novel brain pattern and used to decode the coordinate of that stimulus in the n-dimensional feature space. The decoded set of coordinates can then be compared to a "dictionary" of the meaning vectors associated with particular words, and – based on this – the classifier can make a guess about which word the person is thinking about at that moment. Alternatively, some studies have used a complementary *encoding* approach where, instead of predicting feature vectors based on brain patterns, these studies learned to predict brain patterns based on a combination of feature vectors: i.e., if a word has a particular meaning vector, what should its fMRI pattern look like? See Naselaris, Kay, Nishimoto, & Gallant, 2011 for further discussion of encoding and decoding models.

The power of this “feature space” idea is that it is usually possible to learn the neural correlates of particular feature dimensions based on a limited subset of stimuli; once the brain-to-feature mapping has been learned by a decoding model, the model can be used to decode the feature vector for any stimulus that resides within the representational space, regardless of whether that stimulus appeared at training; likewise, once the feature-to-brain mapping has been learned by an encoding model, it can be used to predict the brain response to any stimulus that resides within the representational space. For example, Mitchell et al. (2008) used this approach to decode which of two novel words (i.e., words not presented during classifier training) the participant was thinking about, with 77% accuracy. Several other studies have used this feature-based decomposition approach to decode the contents of visual stimuli based on brain activity (e.g., Kay, Naselaris, Prenger, & Gallant, 2008).

Importantly, if a particular feature-space model yields above-chance decoding (or above-chance prediction of brain patterns in a particular region), this tells us that the model has *some* relationship to how those stimuli are coded in the brain, but there could be other models that do a better job. Given two competing models of neural coding, one way to discriminate between them is to build encoding models based on the two different “feature spaces”, and to see which one of them does a better job of predicting the observed fMRI activity (Serences and Saproo, 2012).

4.2. Improving across-subject classification

Earlier, we discussed the “data starvation” problem in MVPA analysis: The number of brain snapshots is typically low relative to the number of parameters being estimated by the classifier, resulting in a high danger of overfitting. The easiest way to combat this problem would be to combine data across participants. However, this will only work to our benefit if the brain patterns corresponding to particular cognitive states are reasonably consistent across participants – otherwise, the added within-class variability (resulting from across-participant differences) will offset the beneficial effects of having more data. The key question thus becomes: How can we align data across participants in a manner that minimizes across-participant variability in cognitive representations?

The standard approach to across-subject alignment is to transform each participant's data into a common template space based on anatomical landmarks, and then to combine the transformed data. This procedure has proved to be very useful for standard univariate fMRI analyses, but there have been relatively few reports of anatomical alignment alone leading to good across-subject classification. There are two likely reasons for this: First, the transformations result in spatial blurring which might erase high-spatial-frequency information in the data that would otherwise be useful for the classifier. Also, people have different experiential histories that shape how concepts are represented in their brains – no amount of anatomical alignment will correct for such differences.

To address these issues, Haxby and colleagues have developed a new across-subject alignment procedure called *hyperalignment* (Haxby et al., 2011) that aligns brains not based on anatomical landmarks but rather based on the functioning of those brains (i.e., aligning parts of the brain that behave similarly, regardless of "where" exactly in the brain these parts came from). The basic hyperalignment algorithm performs a Procrustes transformation that rotates, scales, and shifts temporal trajectories of voxels to best align datasets from different brains. Haxby and colleagues hyperaligned a dataset of 21 participants viewing the movie Raiders of the Lost Ark, and performed across-subject classification for movie segments, faces & objects, and animal species. They found that hyperalignment produced far superior classification performance compared to alignment based purely on anatomy; hyperalignment even matched the accuracy of within-subject classification. Good classification indicates good alignment of patterns across participants, which suggests that adding more participants to the training set should help generalization even further. Therefore, hyperalignment is an extremely promising approach to minimizing the "data starvation" problem of MVPA.

5. Conclusions

Multi-voxel pattern analysis allows us to detect information in the brain that was not visible using previously developed methods of fMRI analysis. Getting a better handle on the informational contents of a person's brain puts researchers in a better position to test theories of how information-processing works in the brain, and how cognitive states shape behavior.

MATLAB software for performing a classifier analysis (the Princeton Multi-Voxel Pattern Analysis toolbox) can be found at http://www.pni.princeton.edu/mvpa. Alternatively, there is a separately developed Python version of the toolbox (PyMVPA) available at http://neuro.debian.net. MATLAB software for performing pattern-similarity MVPA can be found at http://www.mrc-cbu.cam.ac.uk/methods-and-resources/toolboxes.

Although MVPA offers advantages over other forms of analysis, there are limitations to what it can accomplish (Davis and Poldrack, 2013). Going forward, it will be beneficial to compare MVPA to other measurement and analysis techniques to get a better sense of which aspects of neural processing we can and cannot detect with MVPA. Comparisons of MVPA with univariate analysis (Jimura and Poldrack, 2012) and fMRI adaptation (Epstein and Morgan, 2012) suggest that these methods are interrogating different aspects of the neural code. It will also be useful to compare MVPA to neurophysiology data from human and nonhuman primates (e.g., Kriegeskorte et al., 2008b) to better understand the strengths and limitations of this powerful, but relatively recent advance in the cognitive neuroscience toolkit.

6. References

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*(2), 261-270.

Davis, T., & Poldrack, R. A. (2013). Measuring neural representations with fMRI: Practices and pitfalls. *Annals of the New York Academy of Sciences*.

Detre, G. J., Natarajan, A., Gershman, S. J., & Norman, K. A. (2013). Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia*.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Epstein, R. A., & Morgan, L. K. (2012). Neural responses to visual scenes reveals inconsistencies between fMRI adaptation and multivoxel pattern analysis. *Neuropsychologia*, *50*(4), 530-43.

Gonsalves, B. D., & Cohen, N. J. (2010). Brain imaging, cognitive processes, and brain networks. *Perspectives on Psychological Science*, *5*(6), 744-752.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157-1182.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425-2430.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404-16.

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523-34.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science, 310*(5749), 863-6.

Jafarpour, A., Horner, A. J., Fuentemilla, L., Penny, W. D., & Duzel, E. (2013). Decoding oscillatory representations and mechanisms in memory. *Neuropsychologia*, *51*(4), 772-80.

Jimura, K., & Poldrack, R. A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, *50*(4), 544-52.

Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS One*, *5*(1), e8622.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679-85.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352-5.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863-3868.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in System Neuroscience, 2*.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Bandettini, P. A. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126-41.

Kuhl, B. A., Rissman, J., Chun, M. M., & Wagner, A. D. (2011). Fidelity of neural reactivation reveals competition between memories. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(14), 5903-8.

LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spector, K., & Wagner, A. D. (2013). Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. *The Journal of Neuroscience*, *33*(13), 5466-74.

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience*, *24*(1), 61-79.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M. A., & Newman, S. D. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, *57*, 145-175.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. -M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191-5.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636-43.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400-10.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424-30.

Pereira, F., & Botvinick, M. (2011). Information mapping with pattern classifiers: A comparative study. *NeuroImage*, *56*(2), 476-96.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1), S199-209.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*, 1963-1966.

Rissman, J., & Wagner, A. D. (2012). Distributed representations in memory: Insights from functional brain imaging. *Annual Review of Psychology*, *63*, 101-28.

Ritchey, M., Wing, E. A., Labar, K. S., & Cabeza, R. (2012). Neural similarity between encoding and retrieval is related to memory via hippocampal interactions. *Cerebral Cortex*.

Serences, J. T., & Saproo, S. (2012). Computational advances towards linking BOLD and behavior. *Neuropsychologia*, *50*(4), 435-46.

Soon, C. S., Brass, M., Heinze, H. -J., & Haynes, J. -D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*(5), 543-5.

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*, *77*, 157-65.

Tong, F., & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual Review of Psychology*, *63*, 483-509.

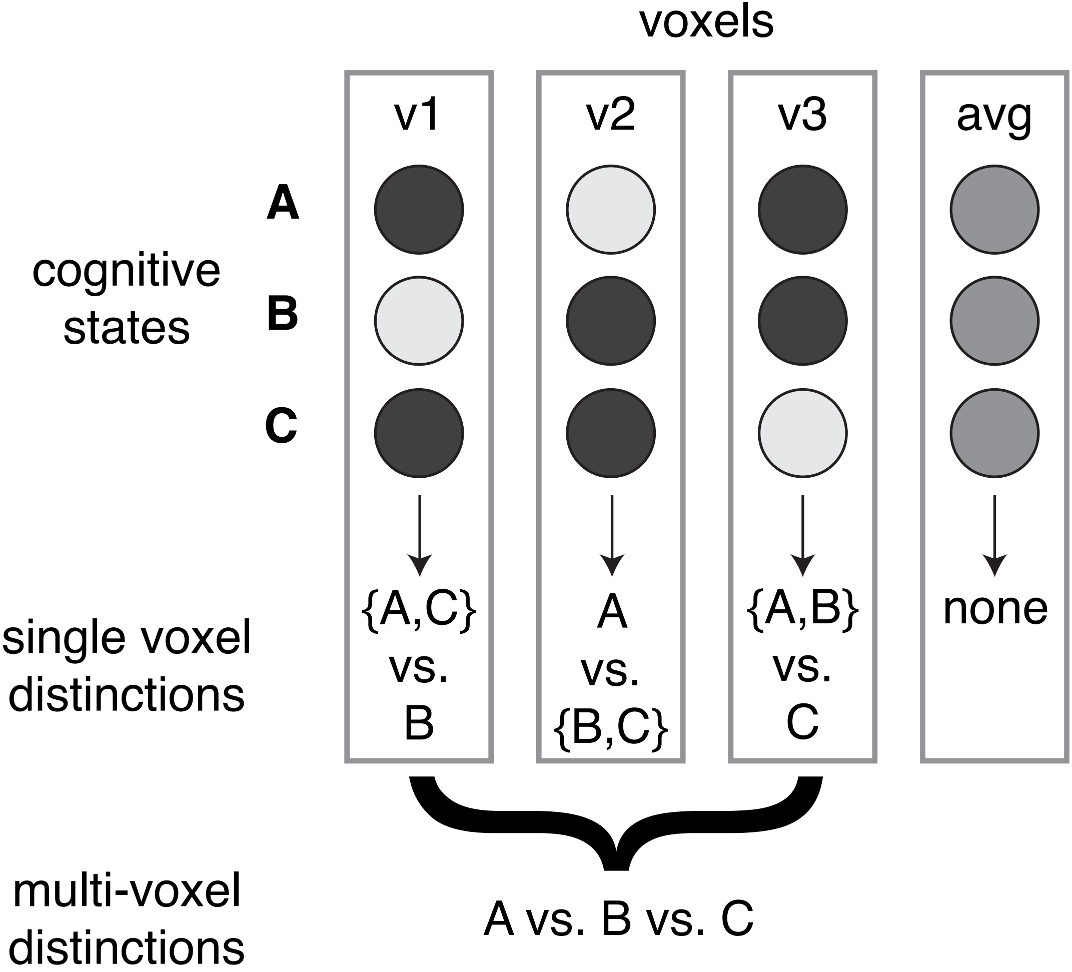
Williams, M. A., Dang, S., & Kanwisher, N. G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nature Neuroscience*, *10*(6), 685-686.

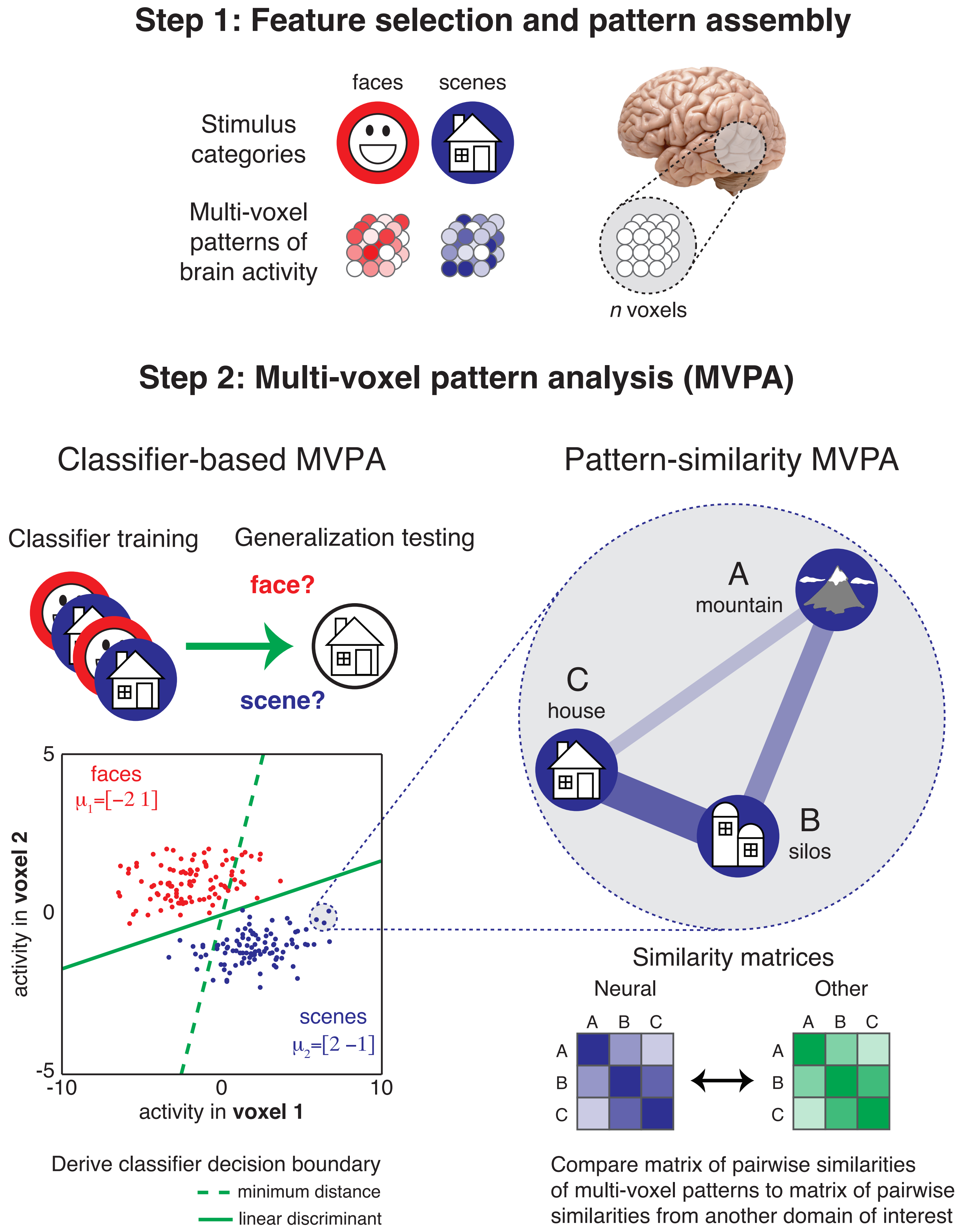
Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*(1), 168-79.

Figure Captions

Figure 1. Observing the multi-voxel response pattern allows us to distinguish all three cognitive states A, B and C. Considering each voxel in isolation provides only partial discrimination.

Figure 2. The two main types of multi-voxel pattern analysis of fMRI data.


voxels
v1
v2
v3
avg
A
B
C
cognitive states
{A,C} vs. B
A vs. {B,C}
{A,B} vs. C
none
single voxel distinctions
multi-voxel distinctions
A vs. B vs. C


Step 1: Feature selection and pattern assembly
faces
scenes
Stimulus categories
Multi-voxel patterns of brain activity
n voxels
Step 2: Multi-voxel pattern analysis (MVPA)
Classifier-based MVPA
Pattern-similarity MVPA
Classifier training
Generalization testing
face?
scene?
5
faces
μ1=[−2 1]
0
-5
-10
0
10
activity in voxel 2
activity in voxel 1
scenes
μ2=[2 −1]
Derive classifier decision boundary
minimum distance
linear discriminant
A
mountain
C
house
B
silos
Similarity matrices
Neural
Other
A B C
A
B
C
A B C
A
B
C
Compare matrix of pairwise similarities of multi-voxel patterns to matrix of pairwise similarities from another domain of interest