Distinguishing conjoint and independent neural tuning for stimulus features with fMRI adaptation

Daniel M. Drucker, Wesley Thomas Kerr, Geoffrey Karl Aguirre

Department of Neurology, University of Pennsylvania

RUNNING HEAD: Conjoint and Independent Neural Tuning

Correspondence:

Geoffrey Karl Aguirre

Department of Neurology
University of Pennsylvania
3 West Gates
3400 Spruce Street
Philadelphia PA 19104
Tel: 215-662-3390
Fax: 215-349-5579
aguirreg@mail.med.upenn.edu
Abstract

A central focus of cognitive neuroscience is identification of the neural codes that represent stimulus dimensions. One common theme is the study of whether dimensions, like color and shape, are encoded independently by separate pools of neurons or are represented by neurons conjointly tuned for both properties. We describe an application of fMRI-adaptation to distinguish between independent and conjoint neural representations of dimensions by examining the neural signal evoked by changes in one versus two stimulus dimensions and considering the metric of two-dimension additivity. We describe how a continuous carry-over paradigm may be used to efficiently estimate this metric. The assumptions of the method are examined as are optimizations. Finally, we demonstrate that the method produces the expected result for fMRI data collected from ventral occipito-temporal cortex while subjects viewed sets of shapes predicted to be represented by conjoint or independent neural tuning.
Introduction

A major goal of cognitive neuroscience is to determine how neural populations represent stimulus variation. Measurement of the tuning of neurons for variations in stimuli is one approach. For any given neuron, a tuning curve exists which describes the response modulation of that neuron as a function of different levels of a stimulus dimension. Neurons in primary visual cortex, for example, demonstrate tuning for stimulus orientation, with a smooth decrement of response as the angle of the stimulus differs from the optimum. A population of these neurons, each responding maximally to a different orientation, can, as a whole, accurately encode the value of the stimulus dimension.

The brain accurately represents numerous stimulus dimensions. Do different neurons represent individual stimulus dimensions, or could one neuron be tuned to represent multiple dimensions? For any two given dimensions of a stimulus (for example, the orientation and spatial frequency of a grating) two extremes of representation can be imagined. In a conjoint representation, a given neuron would have both a preferred spatial frequency and a preferred orientation. The two dimensions in this case are jointly encoded by a single population of neurons, each neuron responding optimally to a particular value of each dimension, and its response dropping with change in either one. Alternatively, the two dimensions of the stimulus could be encoded by two independent populations, with each neuron tuned for one of the dimensions but with no tuning preference for the other. In the particular case of orientation and spatial frequency, it has been shown by single-unit recording that V1 neurons generally encode these and several other visual dimensions conjointly (DeValois et al. 1982; Mazer et al. 2002). This paper describes an application of functional MRI (fMRI) to distinguish conjoint from independent representation of two stimulus dimensions within a spatially restricted population of neurons.

Functional MRI measures cortical responses with a spatial resolution on the order of millimeters. Adaptation (Grill-Spector and Malach 2001) has been used to measure the behavior of neural populations at sub-voxel scales by measuring the graded reduction in population response that accompanies repetition (or near repetition) of a stimulus property. The population of neurons within a voxel can be argued to represent (be tuned to) a stimulus dimension if the
presentation of pairs of stimuli with ever greater differences along that dimension results in a progressive recovery from adaptation and thus an ever greater fMRI response. Proportional recovery from adaptation of this kind has been described for the angular displacement of gratings (Fang et al. 2005), and the similarity of faces (Jiang et al. 2006), shapes (Drucker and Aguirre 2009), and colors (Aguirre 2007). While potentially powerful, the inferences provided by adaptation studies are nuanced, particularly with regard to precise cortical localization (for a comprehensive discussion, see Bartels 2008).

We have recently extended the proportional adaptation approach to the presentation of continuous stimulus sequences (Aguirre 2007). Using that method, we have shown that fMRI can be used to efficiently measure the neural representation of multiple stimulus dimensions simultaneously (Aguirre 2007; Drucker and Aguirre 2009). This allows us to relate the similarity of a set of stimuli to the similarity of the responses that they evoke within a neural population.

In this paper, we show that measurement of the recovery from adaptation for changes within a stimulus space can be used to distinguish between conjoint and independent neural representations. This is accomplished by measuring the recovery from adaptation for stimulus changes in both perceptual dimensions and stimulus changes in each dimension alone. If the recovery for a combined change is simply the additive combination of the recovery for each dimension in isolation, we take this as evidence for independent neural populations. When the neural recovery for a combined change is sub-additive, this may reflect populations consisting of neurons which represent the two stimulus dimensions conjointly.

The central insight that motivates our approach has been described previously. Engel (2005) tested for neural populations jointly tuned to two stimulus axes by measuring subadditivity of adaptation responses. He employed a paired-adaptation paradigm in which an adapting grating was followed by a test grating that differed in color, orientation, or both. Two levels of each factor (color and orientation) were tested. The combined stimulus change was shown to have an interacting component for some populations, in agreement with the previous findings suggesting radially-symmetric receptive fields within the stimulus space (e.g., Livingstone and Hubel 1984; further references in Engel 2005). We extend Engel’s approach in several ways.
We first demonstrate that this sub-additivity inference follows from the shape of neural receptive fields within a stimulus space. Next, we describe how the covariates of a general linear model applied to a continuous carry-over fMRI experiment may be used to test for sub-additivity. We consider how the sensitivity and specificity of the method are affected by non-linearities and rotations of the perceptual space, neural representation, or hemodynamic transform. The properties of the method given populations consisting of neurons that are neither perfectly independent nor conjoint in their representation are examined, and we provide a generalization in terms of the exponent of a Minkowski characterization of the neural space. Optimizations of the approach are then described, including the sampling of stimuli from the perceptual space and the selection of efficient stimulus sequences for BOLD fMRI. Finally, we provide the results of a pair of experiments conducted on two parameterized shape stimulus sets, the dimensions of which are predicted to be represented in one case by conjoint neural population tuning and in the other by independent tuning. We show that fMRI responses in ventral occipito-temporal cortex have the expected properties.

Theory

Stimulus Spaces and Conjoint and Independent Neural Populations

A set of stimuli may be constructed from parametric variations along two dimensions. For example, a set of outlines may differ in their shape and color (Figure 1A). The response of a neuron to the stimuli is characterized by its receptive field within the stimulus space. We may contrast two idealized models of neuronal receptive fields arranged to represent variations in the color and shape of a set of visual stimuli. A neuron may have a conjoint receptive field within the stimulus space, such that maximal firing is only elicited by a stimulus of a particular color and a particular shape (Figure 1B). Alternatively, a neuron may have an independent receptive field, such that firing is maximal for a particular stimulus in one dimension (e.g., shape), but is not altered by the other stimulus dimension (in this case, color) (Figure 1C). Of course, intermediate tuning functions are possible; these cases are considered later. Our first theoretical goal is to distinguish between the extreme conjoint and independent models of receptive field organization within a particular voxel.
These two possibilities could be distinguished directly by measuring the tuning of individual neurons. But the signal obtained with BOLD fMRI averages the population neural response from a voxel, making this measurement unavailable. If independently tuned populations of neurons were separated from one another by distances of many millimeters across the cortical surface, then BOLD fMRI would be able to resolve the presence of separate cortical areas tuned to the different stimulus dimensions. Alternatively, if there were sufficient heterogeneity in the spatial distribution of independently tuned neurons, then one might be able to examine the pattern of responses across voxels (Norman et al. 2006; Kamitani and Tong 2005) to distinguish stimuli. The inferences that may be drawn from such a distributed analysis are discussed in Appendix A. The neural representation of some stimulus dimensions, however, may be intermixed at a spatial scale below the resolving power of BOLD fMRI. In this situation, the presentation of any one stimulus in isolation might evoke the same average response across the population, rendering the conjoint and independent tuning possibilities (indeed, the individual stimuli) indistinguishable by fMRI. To distinguish conjoint and independent tuning in this case, we must measure the properties of the neural population using adaptation methods.

The sensitivities of intermixed neurons may be revealed by neural adaptation and fMRI. If the experiment presents a pair of stimuli that are (e.g.) the same shape but differ in color, both a conjointly tuned and independently tuned population of neurons would show some recovery from adaptation for this stimulus transition. A pair of stimuli that are the same color but differ in shape would produce the same result. Such an experiment would reveal the presence of a population of neurons that are tuned to represent color and shape, but it would not distinguish between the conjoint and independent possibilities.

To do so, one measures the recovery from adaptation associated with a combined stimulus change. Consider a pair of stimuli which differ in both color and shape. Within a population of independently tuned neurons, the transition in shape will be within the unidimensional receptive field (Figure 1C) of some neurons, producing a recovery from adaptation; for other neurons the transition in color will have a comparable effect. The total recovery from adaptation for the population will be simply the recovery seen in the population of shape-tuned neurons plus the recovery of the color-tuned neurons.
The population of conjointly tuned neurons demonstrates a different behavior. In this model system, neurons have radially symmetric receptive fields within the stimulus space (Figure 1B). As a consequence, for an individual neuron, the effect of a stimulus change along both dimensions is not an additive effect of each change in isolation. Instead, it is the Euclidean distance of the change in the stimulus space. This can be intuited by considering that, within a radially symmetric field, rotational invariance must hold. Consequently, a change of one unit along one dimension is equivalent to a diagonal change of one unit along both dimensions. Within the symmetric field, that combined change can be decomposed into changes of 0.7 units along each axis, demonstrating that the neuron considers a combined stimulus change to be less than the sum of individual stimulus changes (specifically, $1.0 \text{ unit of stimulus change} \neq 0.7 + 0.7 \text{ units of individual stimulus changes}$). Appendix B provides a formal proof that this property of individual neurons predicts sub-additive recovery from adaptation for a model population of neurons.

In summary, we may distinguish between conjoint and independent tuning of neurons in a population by comparing the recovery from adaptation for combined transitions to that seen for isolated transitions along each stimulus dimension. We now consider the design of an fMRI experiment to do so.

*Construction of a BOLD fMRI experiment*

In theory, one could conduct the test described above by measuring the BOLD fMRI response to three stimulus pairs: a pair that differs only in color, a pair that differs only in shape, and a pair that differs in both color and shape. As will be developed below, such a limited test is not robust to non-linearities in the measurements. A more robust test, with the ability to check for deviations from the assumed model, is provided by measuring multiple transitions within the stimulus space over a range of distances. At least three samples along each stimulus dimension are needed; here we consider a stimulus space with four samples along each dimension. To fully characterize the tuning properties of the population of neurons under study, we measure all possible transitions between the stimuli in the perceptual space. This may be accomplished efficiently using a continuous carry-over approach (Aguirre 2007), in which stimuli are presented continuously and sequentially using a serially counterbalanced stimulus order. The
participant views a stream of stimuli, perhaps while performing an attention task that is irrelevant to the stimulus similarity (e.g., detecting an infrequent target not from the stimulus space). We measure the fMRI response to each stimulus and model it as a function of its relationship to the prior stimulus: how much change is there in shape, color, or both (Figure 2)? This design allows us to examine the recovery from adaptation between all possible stimulus pairs (and characterize the neural response to each stimulus free of first-order context; see Appendix A).

The analysis of data collected from such an experiment is based on covariates that model recovery from adaptation for the stimulus changes; we will consider two covariates first. One covariate models the degree of change in color for each stimulus compared to the prior stimulus, while a second covariate models the amount of change in shape. These covariates are then convolved with a standard hemodynamic response function and used to model the BOLD fMRI data. How would these covariates model data from a voxel that contained a population of neurons with independent tuning for the two stimulus dimensions? Presuming equal and linear transforms of stimulus changes to neural recovery from adaptation and to BOLD fMRI signal—assumptions which are examined below—equal loading upon the two covariates would be sufficient to model the continuous neural recovery from adaptation present in the data: the color covariate would model the recovery from adaptation produced by the population of color tuned neurons within a voxel, while the shape covariate would model the behavior of the independently tuned shape responsive neurons. Because the linear addition of these two covariates is a sufficient model for the data, we term these the “city-block” covariates, as the fMRI response to the transition between any two stimuli is well described by the rectilinear (purely additive) distance between the stimuli.

How would the model behave given a voxel that contained a population of neurons with conjoint tuning to the stimulus dimensions? The covariates would be unable to simultaneously model the recovery from adaptation associated with single stimulus changes and that from combined stimulus changes. This is because the signal from combined stimulus changes would be less than that predicted by the sum of the isolated stimulus changes. The variance attributable to this sub-additivity can be modeled with a “Euclidean contraction” covariate, which takes the value of the difference between the City-block and Euclidean distance between the stimuli (a.k.a. a “farm-gate contraction”; Wyszecki and Stiles 2000). This covariate captures the degree to
which the neural response deviates from pure additivity for the two stimulus dimensions. If this
covariate models a significant amount of the variance in the observed responses, then we may be
able to reject the independent (purely additive) neural representation model.

For ease of interpretation (although not of statistical necessity) the Euclidean contraction
covariate is orthogonalized with respect to the City-block covariates so that it will have zero
loading in presence of an independently tuned neural population. The additional components of
the model (e.g., main effects versus null-trials, stimulus repetitions, stimuli that follow null-
trials) are considered at length elsewhere (Aguirre 2007). An example set of data, covariates, and
results is available for download (http://cfn.upenn.edu/aguirre/wiki/example_carry-over_data_and_analysis).

Distortions of the measured response

To test for the presence of a population of conjointly tuned neurons we manipulate a set
of perceptual stimuli and measure an evoked BOLD fMRI response. Several transformations of
the independent and dependent variables intervene between the data and our desired inference.
Here we examine the specificity and sensitivity of the method in the face of these distorting
transformations. In an idealized model, the stimulus space presented to the subject evokes equal,
regular, and linear differences in neural representation, so that proportional steps in the space
produce proportional changes in the similarity of neural response, and in turn a linearly
proportional recovery from adaptation for the population. These changes in neural activity are
then transformed into BOLD fMRI signal change by linear convolution with the hemodynamic
response function. Of course, non-linearities and asymmetries may exist at each of these steps.

Particularly troublesome are compressive non-linearities that act symmetrically upon
both dimensions of the stimulus space representation. Such a distortion could cause
independently tuned neural populations to appear conjointly tuned, as larger stimulus transitions
will be less than predicted from smaller stimulus transitions, mimicking the behavior of a
Euclidean distance metric. A plausible cause of such a symmetric, compressive, non-linearity is
saturation of the transformation of neural activity to hemodynamic response. As we will argue,
however, the relatively small signal modulation produced by neural adaptation justifies a small-
scale linear approximation in this case.
We examined the effect of hypothetical distortions in the setting of simulated data from an experiment. The MATLAB code used for the simulation is available for download (http://cfn.upenn.edu/aguirre/code/simulation/sequence/distort/calcdistort.m). The model begins with a set of regularly spaced stimuli\(^1\) (Figure 3A), and thus a similarity matrix defined by the distance between the stimuli. We considered next the similarity matrix of neural responses for a population of neurons that represent the stimuli. The neural representation could perfectly reflect the original stimulus space, or may contain distortions (Figure 3B). For example, the changes in the stimuli along one axis may be more salient than the changes along the other axis, and this change may be a linear or non-linear transform of the original stimulus space. Further, a non-linear transform of both axes may be present, such that (for example) the neural representations of the stimuli from one corner of the space are much more similar than the neural representations of stimuli from the other corner.

Given a sequence of stimuli in a continuous carry-over design, we can then model the time-course of neural response that would be expected given recovery from adaptation proportional to the dissimilarity of neural responses to each stimulus (see Appendix B and Verhoef 2008). The dissimilarity between stimuli (and therefore the recovery from adaptation) can be modeled assuming either conjoint neural tuning and a Euclidean distance metric, or independent neural tuning and a city-block distance metric. This results in two different models of neural activity over time, corresponding to the independent and conjoint possibilities. This neural signal is then transformed into a hypothetical BOLD fMRI signal by convolution with a standard hemodynamic response function (Aguirre 1998).

Prior to convolution, however, the possibility of another non-linearity is introduced. While generally conforming to a linear system (Boynont et al. 1996), non-linearities in the transform of neural activity to BOLD fMRI signal may exist. Of particular importance are compressive non-linearities, in which a doubling of the neural response results in less than a doubling of the BOLD signal (Vazquez and Noll 1998). We considered several non-linearities (Figure 3C) imposed over the full range of neural response (i.e., from no neural activity to a

\(^1\) The di-octagonal space, discussed below, was used. For ease of explanation, however, Figures 3 and 4 illustrate the simulation using a grid-spacing of stimuli.
hypothetical maximal neural activity in an area). Perceptual adaptation, however, has not been observed to modulate neural activity over its entire range. When measured in visually responsive cortical areas, the modulation of BOLD fMRI signal by adaptation has been at most 20% of the maximal response of the region (Aguirre 2007; Fang et al. 2005; Kourtzi et al. 2003). Therefore, we selected the most non-linear 20% range of each distorting function, and examined the effect of the distortion applied to the neural signal prior to convolution.

A set of 100 different orders of stimulus presentation were considered. Each was a counterbalanced sequence selected for having high overall Efficiency for detection of the City-block and Euclidean contraction effects (described below). A simulated BOLD fMRI data sequence was then created for each possible crossing of the distortion of neural representation with distortion of the hemodynamic transform, for both the conjoint and independent distance metrics. The final, simulated BOLD fMRI sequence was then analyzed with the covariates described in the previous section; i.e., we obtained the loading of the simulated data on the model covariates. The average loading on the model covariates was obtained across the 100 different sequences of stimulus presentation.

Ideally, the loading upon the Euclidean contraction covariate should remain zero for an independently tuned neural population, and have a positive loading for a conjointly tuned population, regardless of the distortions introduced. Table 1 presents the average loading upon the Euclidean contraction covariate for assumed conjoint and independent neural populations in the face of distortions of the neural similarity space, and non-linear transforms of neural activity into BOLD fMRI signal. Appropriately, the covariate has a positive loading (1.4) for the conjoint population and zero loading for the independent populations when no distortion is applied. Further, the model is robust to distortions. A positive loading for the conjointely tuned neural population is found in all cases. This indicates that the model remains sensitive for the test of conjoint neural tuning across a variety of distortions and non-linearities. When independently tuned neural populations are assumed, the loading upon the Euclidean contraction covariate generally remains at zero or is negative, preserving the specificity of the approach (i.e., one would not mistakenly reject the null-hypothesis of independent populations).
In a few cases, the non-linearities produce a positive loading upon the Euclidean contraction covariate even when independent neural populations were assumed. These cases are potentially problematic as they constitute improper bias. Fortunately the degree of bias was generally small. In the face of non-linearities in the BOLD response, the improper loading upon the Euclidean contraction covariate was about one tenth of that measured for the city-block effect. Therefore, even in the face of quite severe, compressive non-linearities in the BOLD hemodynamic response, the relatively small scale of neural adaptation changes maintains the specificity of the method. It is worth noting that non-linearities in the BOLD response may be further discounted as the cause of a finding of conjoint representation if, for a separate stimulus space, independent representation is demonstrated.

Logarithmic transformations of both axes of the neural representation were, however, a problematic case, as they tend to produce larger loadings (0.4) upon the Euclidean contraction covariate. Several steps may be taken to guard against this possible non-linearity. First, and as detailed below, behavioral testing conducted with the stimulus space should be used to confirm that roughly equal perceptual salience accompanies equal changes within the stimulus space. Next, post-hoc testing may be conducted upon the discretized responses to stimulus pairs to examine the relationship between stimulus change and recovery from adaptation. This relationship may be examined for compressive non-linearities. An example of such an analysis is provided in the demonstration experiment below and described in Appendix C. Finally, the Euclidean contraction covariate effect can be statistically judged not against a loading of zero, but against 40% of the city-block covariate effect. If the Euclidean contraction effect is significantly larger than this proportion, then our simulations suggest that a symmetric, compressive non-linearity at the neural level cannot account for the result.

The average loading upon the covariates are presented here for a set of possible sequences of stimulus ordering. While the model maintains its expected sensitivity and specificity in the face of distortions and non-linearities on average, the model may be less robust in the instance of a particular pairing of a distortion with a particular sequence. Therefore, it is advisable to use a variety of sequences within and across scanning sessions so that the robust aggregate performance of the model is retained.
To summarize, the proposed method retains sensitivity for the test for conjointly tuned neural populations in the presence of a variety of distortions and non-linearities in the transformation of neural population codes to BOLD fMRI signal. Specificity, i.e., the absence of a positive test outcome in the setting of independently tuned neural populations, is retained for the majority of considered distortions. Where improper bias was found as a consequence of a hemodynamic non-linearity, the bias was small. Improper bias could be induced by symmetric, compressive non-linearities in the neural representation. Experimental design optimizations and post-hoc techniques to guard against this situation were offered.

Rotation of the assumed stimulus dimension axes

A different violation of the model assumptions occurs when the underlying neural representation is independent for the stimulus dimensions, but its neural instantiation is not aligned with the assumed dimensional axes of the study. For example, consider an experiment designed to examine the neural representation of rectangles. The stimulus space used in the experiment consists of rectangles that vary in height and width, and the experimenter models these two parameters. It may be the case, however, that a population of neurons actually has independent tuning for the sum and difference of height and width (roughly corresponding to area and aspect ratio); a 45° rotation of the axes as modeled by the experimenter. Figure 4A illustrates a population of neurons with receptive fields rotated 22.5° with respect to the assumed axes of the study. As shown in the last two rows of Table 1, even assuming equal and linear transformations, a positive loading upon the Euclidean contraction covariate would be obtained in the case of rotation, leading to the erroneous conclusion of a conjoint representation of height and width.

When presented with significant loading upon the Euclidean contraction covariate, and thus tentative evidence for conjoint representation, an additional test is required to reject the possibility of rotated, independent axes. This special case may be detected by examining the behavior of the covariates under rotations of the assumed model (Figure 4B). The analysis is repeated assuming that the stimuli presented actually were organized by (e.g.) area and aspect ratio (as well as partial rotations). In practice, this is accomplished by calculating distances for the City-block covariates assuming different axes than those actually used. For example, if a 45°
rotation of the model is being tested, then a stimulus change in both height and width of a rectangle is represented in the covariates as a single change in the one covariate (now behaving as a “size” covariate) and no change in the other covariate (now behaving as an “aspect ratio” covariate). Figure 4C presents the loading upon the City-block and Euclidean contraction covariates that will be obtained in the presence either of conjointly tuned neurons or independent populations of neurons with receptive fields oriented 22.5° away from that initially assumed in the experiment. If the neural representation is truly conjoint, then the loading upon the Euclidean contraction covariate will be unchanged with rotation of the model (ignoring the effects of gamut; discussed below). In contrast, a minima for the Euclidean contraction covariate is found when the assumed axes for the stimuli match the actual axes in neural representation (Figure 4C). Because of discrete (16 point) sampling of a continuous stimulus space, the Euclidean contraction covariate receives an artifactual, negative loading at rotations adjacent to the veridical 22.5° value. If we perform the simulation with an increase in the sampling density (a 7x7 stimulus array and 49 samples), the true, monotonic function is revealed. This imperfection should have little effect on the inference.

In summary, when significant loading upon the Euclidean contraction covariate is obtained in an experiment, an additional test is necessary to reject the possibility of independent, but misaligned, neural populations. Post-hoc testing of the performance of the model under assumed rotations of the stimulus axes can distinguish between the independent but rotated, and the conjointly tuned cases.

Extension to a generalized neural space metric

So far we have considered two competing, extreme models of neural representation: two independent neural populations that represent stimulus dimensions separately, and a conjoint population that represents the two dimensions together. Earlier, we considered how these concepts are related to receptive fields that are either linear or radially symmetric within a stimulus space. Intermediate receptive fields are possible, however, with oval shapes of varying elongation. In such cases the population would not be wholly independent, but instead represent one dimension to a greater extent than the other. These intermediate cases are considered readily within the framework of the Minkowski exponent that defines the representational space.
We have seen how independent neural populations may be expected to produce recovery from adaptation proportional to the rectilinear, City-block distance of the stimulus transition, while a conjoint population will produce a sub-additive recovery proportional to the Euclidean distance. City-block and Euclidean metrics correspond to metrics of $r=1$ and $r=2$ within a generalized Minkowski measure of distance:

$$d(a,b) = \left( \sum_{k=1}^{n} |a_k - b_k|^r \right)^{1/r}$$  \[ Eq1 \]

where the total distance $d$ in an $n$-dimensional space between stimuli $a$ ($a_1, a_2, \ldots, a_n$) and $b$ ($b_1, b_2, \ldots, b_n$) is related to the distance along dimension $k$ of $n$ dimensions.

Oval receptive fields correspond to $1 < r < 2$. The earlier simulation that produced the expected beta values for the independent and conjoint neural representation cases are repeated in Figure 5A for these intermediate (and other) Minkowski values. As can be seen, there is a smooth and decelerating increase in loading upon the Euclidean contraction covariate as the Minkowski exponent of the underlying neural representation increases. Thus, significant positive loading upon the Euclidean contraction covariate may be taken as a rejection of independent representation of the stimulus dimensions, although not necessarily as an endorsement of a fully equal and conjoint organization.

One might further consider $r < 1$ or $r > 2$. A Minkowski value of less than one corresponds to a neural similarity measure in which the response to a change in both dimensions is greater than the sum of the changes in each. This non-geometric representation (as it violates the triangle inequality assumption) can be related to featural models of similarity (Tversky 1977; Goldstone and Son 2005). Testing for the predicted negative loading upon the Euclidean contraction covariate is a straightforward extension of the current technique. Conversely, a Minkowski of greater than 2 reflects greater weight being placed upon the dimension with the greater perceptual change. In the limit, $r = \infty$ corresponds to a similarity measure which reflects only the dimension with the larger change (the “dominance” metric).

In practice, the actual loading upon the elements of the model will be in terms of fMRI signal change, and will vary from region to region and for different stimulus spaces. Therefore,
the absolute magnitude of the effect cannot be precisely interpreted, making the ratio of the Euclidean contraction to the city-block effect a more useful measure (Figure 5B). Generally, positive loading supports non-independent coding and negative loading supports “featural” coding. Moreover, the relative loading upon the covariate in a single cortical area is interpretable. The Euclidean contraction measured for two different stimulus spaces, and thus the relative degree of combined stimulus representation, could be of neuroscientific interest. For example, one may wish to compare different stimulus changes that define a multi-dimensional perceptual space to determine which dimensions are “privileged”, in the sense that they have explicit and independent neural coding.

**Optimizations**

The test for a conjointly tuned neural population amounts to the measurement of variance attributable to the Euclidean contraction covariate. We consider here optimizations of the approach to maximize power for this test.

**Normalization of perceptual space:** In our earlier consideration of asymmetries and distortions of the perceptual, neural, and hemodynamic transforms we found that these distortions can alter loading upon the Euclidean contraction covariate. Therefore, a primary optimization is to select stimuli that are likely to map to linear and equal changes in the neural representation. Measurements of neural receptive fields from single unit recording may be used to establish the spacing and gamut of the stimuli on each dimension to best linearize and normalize the neural representations. In most cases, of course, such a direct measure of neural response will be unavailable, as this is typically the goal of the experiment itself. Behavioral testing (e.g., Kruskal and Wish 1978) may be used to ensure that perceived changes along each dimension are equivalent and the scaling of the similarity is uniform throughout the space, with the hope that this normalized perceptual space will correspond to a normalized neural representation.

**Stimulus space sampling:** Two continuously varying stimulus dimensions will together define a space with an infinite number of unique points. In experimental settings, only a few representative points are chosen to provide an estimated characterization of the entire space. Some stimulus sets can more efficiently recover the complete space than others. Previous work
with well-defined perceptual stimulus spaces has mostly used stimuli sampled from a “square” grid, as in Figure 6A. The grid spaces stimuli equally spaced along the two dimensions, with all stimuli within a row or column varying along only one dimension.

We propose an alternate sample space (inspired by Shepard 1964) consisting of two nested octagons (a “di-octagon”), that contains the same number of stimuli as the grid but more desirable properties (Figure 6B). First, the gamut (or range in stimuli) is comparable whether one or both dimensions are under consideration. That is, the maximal distance between two points along one of the chosen dimensions (e.g., between 0,29 and 100,29) will always be the same as the maximal distance between two points defined by both dimensions (e.g., 0,29 and 71,100). This renders combined stimulus changes as salient as stimulus changes along a single dimension. In a square grid, by contrast, all distances along a single dimension are shorter than the distances between points that vary on both dimensions. A related benefit is that the distribution of distances across pairs is more uniform in the di-octagon, as opposed to the grid which is skewed towards pairings at short distances. Post-hoc model evaluation is also strengthened. As described above, a rotation of the assumed stimulus space is used to check for independent, but misaligned neural dimensions. The di-octagonal space allows this model rotation to be conducted without unduly affecting the gamut, simplifying the interpretation of the loading upon the Euclidean contraction covariate under these hypothetical rotations.

Finally, the di-octagonal space increases the range of the Euclidean contraction covariate, thus improving power. The variance of the Euclidean contraction covariate is maximized in the comparison of stimulus transitions along a single dimension to equal stimulus transitions along both dimensional axes. The di-octagonal space is configured to maximize the number of stimulus pairs that represent a pure change (0 degrees difference from a single dimensional axis) and pairs that represent equal change on both axes (45 degrees from a single dimensional axis), as compared to a square grid, which has many stimulus pairs at other angles, and thus less potentially informative variation.

**Sequence selection**: The sequence that dictates the order of stimulus transitions may also be optimized. In prior work (Aguirre 2007) we have shown that the order of counterbalanced stimulus presentation used in a BOLD fMRI experiment can impact sensitivity. A
counterbalanced presentation order can be provided by an m-sequence (Buracas and Boynton 2002) or a type 1 index 1 sequence (Nonyane and Theobald 2008). Permutations of the assignment of stimuli to labels in these sequences can be examined for their relative Efficiency (Friston et al. 1999; Aguirre 2007). For the di-octagonal stimulus space of 16 stimuli, a 17 element sequence is needed if null-trials are to be included as well. No n=17 m-sequence exists. We therefore searched permutations of n=17, type 1, index 1 counterbalanced sequences and measured the Efficiency of the sequence for detection of loading upon the Euclidean contraction covariate (assuming a conjoint neural representation) and loading upon the City-block covariate (Figure 7). The relative Efficiency of a sequence for the Euclidean contraction and City-block covariates is uncorrelated. Therefore, a sequence may be selected that optimizes Efficiency for detection of one or the other or a balance between the two. Notably, searching across label permutations nearly doubles the expected Efficiency of a sequence selected at random. Code to search for Efficient sequences, as well as pre-selected sequences, are available for download (http://cfn.upenn.edu/aguirre/).

Example Experiment

We turn now to an application of the metric estimation method to an fMRI experiment. We wished to apply the model to examine the neural representation of two stimulus spaces: one predicted to be a conjointly tuned population of neurons, and one predicted to be a population of independently tuned neurons. Our selection of stimuli was motivated by the psychological study of integral and separable perceptual spaces. Some visual properties of objects are apprehended separately (e.g., color and shape), whereas other dimensions are perceived as a composite (e.g., saturation and brightness); these have been termed separable and integral dimensions (Shepard 1964). We hypothesized that integral perceptual dimensions are represented by populations of neurons that represent the dimensions conjointly, while separable dimensions are represented by independent neural populations; similar ideas have been proposed recently (Arguin and Saumier 2000; Cant 2008; Kayaert et al. 2005; Stankiewicz 2002).

We constructed two sets of simple 2D closed contours that varied along two parameterized dimensions. The first set (Figure 8A) consisted of ineffable “popcorn” shapes, defined by radial frequency components. Abstract radial frequency components are not thought
to be a central organizing component of visual cortex (Albright and Gross 1990), so we have no reason to predict that the two dimensions would be independently represented. Further, we behaviorally characterized these shapes as integral (Supplementary figure 1) using the Garner sorting task (Garner and Felfoldy 1970), replicating Op de Beeck (2003), from which this shape space is derived. The second set of “moon” shapes (Figure 8B), which vary in curvature and thickness, were behaviorally characterized as separable (Supplementary figure 1; again replicating Op de Beeck 2003), and thus predicted to be represented by a neural population with independent tuning for curvature and thickness. In addition, there is evidence strongly suggesting that shape curvature and aspect ratio are independently coded (Kayaert et al. 2005; Arguin and Saumier 2000; Stankiewicz 2008; Op de Beeck et al. 2003). The stimulus spaces were iteratively adjusted to equate perceptual step sizes, as judged by reaction time for same / different judgments in 12 subjects (supplementary methods).

During separate fMRI scanning sessions, we studied six participants while they viewed a continuous visual stream of each of the two sets of shapes. Subjects were required to monitor and report the position of a bisecting line, which was randomly tilted and shifted within preset limits (Supplementary Figure 2), in order to maintain attention. (See supplementary methods for full fMRI scanning details).

For the hypothesis regarding conjoint representation to be posed, it is first necessary to identify a neural population that exhibits a proportional recovery from adaptation for both of the stimulus dimensions. For each subject, we identified within ventral occipito-temporal cortex voxels that showed recovery from adaptation to both stimulus axes for both stimulus spaces (see supplementary methods). Figure 9A shows the position of these voxels across subjects. Most voxels were concentrated around the right posterior fusiform sulcus, corresponding to ventral LOC (Drucker and Aguirre 2009). Because of their method of selection, the identified voxels are guaranteed to have some recovery from adaptation. Figure 9B shows the magnitude of recovery along each stimulus dimension demonstrating the roughly equivalent effect of each axis, and between the two stimulus spaces.

We then examined the loading upon the Euclidean contraction covariate (Figure 9D). The effect of the Euclidean contraction may be tested within the voxels selected as demonstrating
recovery from adaptation, as these two effects are orthogonal. The “popcorn” space, defined by perceptually integral, ineffable dimensions, had significant loading upon the Euclidean contraction covariate across subjects \([t(5df)=4.7, p=0.0055]\), allowing us to reject the hypothesis that the “popcorn” dimensions are represented by independent neural populations. The magnitude of the effect was too large to be explained by undetected distortions of the neural space, based upon our simulations: the loading upon the Euclidean contraction covariate was greater than 0.4 of the primary adaptation (city-block) effect for each subject \([t(5df)=2.6, p=0.048]\). In contrast, for the “moon” space, defined by the separable dimensions of curvature and thickness, loading upon the Euclidean contraction did not differ from zero \([t(5df)=0.1, p=0.92]\), allowing for the possibility that independent neural populations code for the stimulus space. The direct comparison of the popcorn and moon results, predicted to yield a larger Euclidean contraction value for the popcorn space, trended towards significance \([t(4 df)=2.5, p=0.067, \text{one tailed}]\).

A positive loading on the Euclidean contraction covariate, as in the “popcorn” case, should be further evaluated with systematic rotations of the assumed stimulus dimensional axes. The purpose of this post-hoc test is to ensure that the result cannot be explained by independent neural populations with receptive fields that are misaligned with the axes that were assumed to define the stimulus space. Model rotation will leave the loading on the Euclidean contraction covariate unaltered in the case of a conjointly tuned neural population. In the case of independently tuned but misaligned populations, however, the loading should drop to zero at the appropriate model orientation. Figure 9E shows that the Euclidean contraction covariate maintained positive loading across all model orientations, confirming a non-independent neural population. The slight dip and rise in the observed function may be taken as evidence that the actual distance metric of the neural representation is between 1 and 2, with the orientation of the oval tuning functions aligned with a 45° rotation of the stimulus axes. A similar result would obtain if the studied voxels contained a mixture of both independent and conjointly tuned neural populations.

Finally, we conducted an additional post-hoc test to determine if compressive non-linearities in the neural representation of the stimulus space could be responsible for the finding of conjoint neural tuning for the popcorn space (see Appendix C). The recovery from adaptation
for stimulus pairs that changed along a single dimension was obtained, and used to compare the position of a stimulus to the recovery of BOLD response. For both dimensions of the popcorn space, the relationship was close to linear. As the method of voxel selection strongly biased us towards a pool of voxels with such a linear relationship in this example, the application of this post-hoc test will be more relevant when applied to data selected from (e.g.) an anatomical region of interest.

Discussion

We have described an application of fMRI that can test for non-independent neural tuning for stimulus dimensions. The method examines the recovery from adaptation associated with changes in stimuli along one or two stimulus dimensions. By examining the additivity in response to combined stimulus changes, the metric of the neural representation can be tested. This in turn informs as to the underlying neural implementation of that representation. We have shown that the approach is generally robust to non-linearities and distortions of the measurements, and have proposed post-hoc tests that can guard against improper bias. An application of the method to a stimulus space expected to have a conjoint neural representation yielded the predicted result.

A ready criticism of the approach is that it uses a measure of neural adaptation recovery to index the similarity of neural responses to stimuli. Single unit recordings have identified circumstances in which the stimulus selectivity of the adaptation recovery effect does not completely reflect the tuning of the un-adapted neuron (e.g. Sawamura 2006). Importantly, however, the method we describe depends on a slightly different assumption: that the magnitude of neural adaptation recovery is proportional to the magnitude of the stimulus change. Recent work (Verhoef 2008) has found just this response in neurons in macaque infero-temporal cortex. Moreover, our simulations show that the test is robust to even substantial deviations from a linear relationship between adaptation recovery and stimulus similarity.

A deeper issue concerns the localization inference provided by fMRI adaptation methods (Bartels 2008). As BOLD fMRI is thought to be more sensitive to synaptic input activity and local processing then to cell-body spike rate, it is possible that local adaptation effects will reflect the tuning properties of inputs to the region, as opposed to the tuning of the neurons within the
region itself (Tolias 2005). This possibility does not render the results of adaptation studies uninteresting, but does nuance their interpretation. It also provides further justification for the carry-over approach (Aguirre 2007), in which the direct (non-adaptation) effects of stimuli may also be observed using pattern analysis methods (albeit with inferential challenges as well; see Bartels 2008 and Appendix A).

Our method of metric estimation of neural tuning joins long-standing efforts in psychology and neuroscience to determine the separability or modularity of mental operations. In the cognitive neurosciences, fMRI has frequently been used to demonstrate that two tasks evoke activity in separate brain regions, leading to the inference that the tasks are subserved by different modules (Sternberg 2001, p. 186). This work itself derives from much earlier attempts to use the additivity of behavioral measures, such as reaction time, to deduce the structure of mental processes. Beginning with Donders’ notion of pure insertion in the 19th century, this was most notably discussed and expanded by Sternberg (1969; 2001). A related domain of study has been the determination of the metric of stimulus spaces (Shepard 1980), with implications for the separate analyzability of perceptual dimensions. A theoretical challenge considered in this extant literature, relevant to our current study as well, is the reification of studied stimulus dimensions found to be independent. Although a particular study may find independent tuning for a pair of stimulus dimensions, it does not automatically follow that neurons are therefore tuned “for” those axes. It remains possible that the dimensions selected for study are manifestations of some further, as yet unstudied, organizational scheme.

Previous efforts to identify independently represented stimulus dimensions have proceeded across several sensory domains. Support for independent coding for dimensions such as chromaticity and spatial frequency was found in behavior using a Shepard additivity approach (Monnier 2006). The general behavioral separability of spatial and temporal frequency (Reisbeck and Gegenfurtner 1999) is reflected in electrophysiological measures (Priebe et al. 2003). Similarly, shape selectivity has been broadly found to be independent of cues such as position and size in macaque IT (Sary et al. 1993; Janssen et al. 2000), and surface texture (Köteles et al. 2008). In the human auditory system, different dimensions of timbre were found to be processed separately, based on additivity of their mismatch negativities (Caclin et al. 2006).
Our method amounts to using a linear model to test the metric of a space; an approach which has been considered problematic (Hubert 1992). Iterative goodness-of-fit measures have been generally used instead, although these can also be confounded by isometries between metric spaces as measured by finite numbers of stimuli (Arabie 1991). For example, a set of stimuli might be best fit in two dimensions with a Euclidean metric, but in three with a City-block metric. Consequently, the linear method we have described is not automatically generalizable to the study of stimulus spaces beyond that examined here. Specifically, we have argued by simulation for the validity of our model for two dimensions with 16 regularly spaced samples. While there is a vast array of neuroscientific questions that might be asked within this domain, our method would need to be verified anew for application to a different form of stimulus space.

An avenue for future investigation is the use of iterative models to fit the fMRI data generated by our method. In contrast to the three covariate approach we have examined, it is possible to model the degree of neural adaptation recovery comprehensively. This could be accomplished using a basis set of binary covariates that symmetrically model each stimulus transition (see Appendix C). The beta weights upon these covariates could then be used to create a diagonally-symmetric, 16x16 matrix of neural similarity for the stimuli, indexed by recovery from adaptation. The resulting similarity matrix could then be submitted to an iterative multi-dimensional (MDS) or probabilistic (PROSCAL; http://www.proscal.com/) scaling analysis for the measurement of space metrics. Our expectation is that the greater flexibility and reduced assumptions of such an approach will necessarily result in a loss of statistical power for the specific inferences we have focused upon in this paper.

We have considered in this paper several types of non-linearities and distortions that can exist in neural representation or recovery from adaptation. While we find that the method is generally robust to these deviations, there naturally exists the possibility of further violations of the assumptions of the model that we have not evaluated. A general approach that may be taken to reduce the risk of mistaken inference is to study two different stimulus spaces for a given cortical region of interest. If a different result is obtained for each, that is, additivity for one but not the other, this can serve as evidence that the obtained result is not the result of non-linearities in recovery from adaptation or hemodynamic response at that site.
We envision the use of the metric estimation test to study the representation of stimulus properties across sensory cortical areas. By revealing the presence of independently tuned neural populations, the fundamental axes of perceptual representation might be identified. Interestingly, a given stimulus space may be represented conjointly in one region of cortex, but independently in another. For instance, while several stimulus properties are known to be conjointly represented by neurons in V1 (Mazer 2002), separate tuning for these properties appears at higher cortical levels. Our method can show, within a single experiment, the progression from conjoint to independent representation across cortex for a set of dimensions.

Finally, our linear metric estimation method may be applied to other imaging modalities besides fMRI. The key feature of the measure to be tested is that it derives from a population of neurons for which signal adaptation recovery is monotonically related to stimulus similarity. For example, Furl and colleagues (Furl 2007) have studied recovery from adaptation to facial expression with magnetoencephalography (MEG). They demonstrated that the M170 component from the superior temporal sulcus (STS) has a monotonic recovery from adaptation to the degree of change in facial expression. Using the method we have described, one might now test the further hypothesis that independent neural populations in the STS are tuned to the facial expressions of fear and anger. Using a similar logic to that employed in our work, Caclin et al. (2006) used the additivity of a component of auditory event-related potentials to infer partially independent neural populations for processing dimensions of timbre.

In summary, we have presented a linear test and its optimizations for measuring the metric properties of neural tuning using fMRI. This approach builds upon our earlier development of the continuous-carry over design (Aguirre 2007) which introduced a method for efficiently characterizing neural similarity spaces. With our current work, we have now refined the approach to allow measurement of the metric relationship between the dimensions of a similarity space.

Acknowledgments

WTK produced the proof contained in Appendix B.
Grants

This work was supported by K08 MH 7 2926-01 and a Burroughs-Wellcome Career Development Award.

References


Drucker DM, Aguirre GK. Different spatial scales of object similarity representation in lateral and ventral LOC. *Cerebral Cortex*, 2009.


**Appendix A: Relationship to multi-voxel distributed pattern analysis**

As a carry-over experiment presents stimuli in a continuous, counter-balanced fashion, it allows the recovery of the average neural response to each stimulus from each voxel, independent of first-order context (Aguirre 2007). From these measures a distributed neural
similarity matrix can be constructed containing the correlations between the patterns of across-voxel responses to each possible pairing of stimuli (Drucker and Aguirre 2009). Using iterative MDS and related methods (see Discussion), these data may be examined for the metric properties of the distributed neural space, perhaps to support claims of conjoint or independent representation of the stimulus axes. Recently, Schwarzlose et al. (2008) used pattern-correlation to argue that spatial position and object identity are independently represented in several object-selective areas. For this application of multi-voxel techniques, however, it is important to note that any such inferences drawn from the distributed pattern regard the mean responses within a voxel, and not necessarily the tuning of individual neurons within the voxel. This is because the property being sought (independent or conjoint tuning) may be manifest independently at the scale either of individual neuronal tuning or at the scale of spatial arrangement of tuned neurons within voxels.

Figure A1 illustrates readily how populations of neurons with either conjoint or independent tuning can be spatially pooled within voxels such that the across-voxel pattern of activity will be either additive or subadditive in its metric. For example, the lower-left panel illustrates how neurons that are conjointly tuned for a pair of stimulus dimensions may nonetheless be systematically distributed across cortex. Consider neurons that are conjointly tuned for the orientation and spatial frequency of a grating, yet are organized into macroscopic orientation columns without regard to spatial frequency tuning, and in spatial frequency columns without regard to orientation. Such a cortical organization would yield a multi-voxel pattern result that finds voxels that are independently tuned for the two stimulus dimensions, yet the tuning of the neurons themselves is conjoint. Figure 1A further illustrates how neurons with different metric properties could be heterogeneously organized across voxels to produce all crossings of neural and voxel metrics for conjoint and independent tuning.

Formally, therefore, the metric properties of the neural similarity space at the level of individual voxel tuning is orthogonal to the metric properties of distributed neural similarity across voxels. Fortunately, using the carry-over approach described here, both within and across-voxel metric forms may be recovered simultaneously and efficiently in the same experiment.
Appendix B: Proof of additivity and sub-additivity of fMRI adaptation in neural populations with independent or conjoint tuning

We will demonstrate that the within voxel neural adaptation to two perceptual dimensions reflects the Minkowski r relationship between these dimensions for a model neuronal population. Specifically, a change in two perceptually independent dimensions is additive while a change in two perceptually conjoint dimensions is sub-additive.

We assume that the adaptation to one stimulus, $V_{Stim1|Stim2}$, is proportional to the difference in neural response of the current stimulus, A, from the previous stimulus, B, according to the following formula which is motivated by the finding that neural adaptation is proportional to stimulus similarity (Verhoef 2008):

$$V_{Stim1|Stim2} = w \cdot \sum_{neurons} |A - B|.$$

We use $w$ to signify an arbitrary proportionality constant. Further, we assume that the receptive fields (RFs) of neurons that are tuned to perceptually conjoint dimensions are radially symmetric Gaussians with variance $\sigma^2$. The RFs of neurons tuned to perceptually independent dimensions are assumed to be Gaussians with variance $\sigma^2$ in one dimension and uniform in the other dimension. These Gaussians are tiled orthogonally across the stimulus space. The probability density function of these Gaussians will be notated as $G(x)$. Although we assume Gaussians for notation purposes, any even decay function may be chosen.

We will prove our claim in a small example system with four stimuli arranged in a 2 x 2 square, labeled A through D, and four neurons with RFs tiled so that the maxima of the Gaussians align with the location of the stimuli. Assume that the stimuli are placed a distance, d, apart in stimulus space. $V_{AB}$ and $V_{AC}$ are constructed to represent adaptation along one perceptual dimension. $V_{AD}$ is constructed to represent adaptation along both dimensions. This example is easily extended to a larger neuron population without changing the theoretical nature of the system.

In the case of independent dimensions, the neural response to each stimulus will be:
$A = [G(0), G(d), G(0), G(d)]$
$B = [G(0), G(d), G(d), G(0)]$
$C = [G(d), G(0), G(0), G(d)]$
$D = [G(d), G(0), G(d), G(0)]$

From this, it is simple to calculate that $V_{AB} = V_{AC} = 2w \cdot (G(0) - G(d))$ and $V_{AD} = 4w \cdot (G(0) - G(d))$. Simple algebra confirms additivity ($V_{AB} + V_{AC} = V_{AD}$).

In the case of conjoint dimensions, the neural response to each stimulus will be:

$A = [G(0), G(d), G(d), G(d\sqrt{2})]$
$B = [G(d), G(0), G(d\sqrt{2}), G(d)]$
$C = [G(d), G(d\sqrt{2}), G(0), G(d)]$
$D = [G(d\sqrt{2}), G(d), G(d), G(0)]$

The $\sqrt{2}$ term appears due to the radially symmetric RFs. It is then simple to calculate that $V_{AB} = V_{AC} = 2w \cdot (G(0) - G(a)) + 2w \cdot (G(a) - G(a\sqrt{2}))$. Cancelling the two $G(a)$ terms, we see that $V_{AB} = V_{AC} = 2w \cdot (G(0) - G(d\sqrt{2}))$, which we notice is equal to $V_{AD}$. Therefore the system is clearly sub-additive ($V_{AB} + V_{AC} = 2V_{AD}$), and indeed more so than would be suggested by $V_{AB}^2 + V_{AC}^2 = V_{AD}^2$.

We have demonstrated above that the within voxel neural adaptation to a change along two perceptually independent dimensions is additive, whereas the within voxel neural adaptation to a change along two perceptually conjoint dimensions is sub-additive.

Appendix C: Measurement of discretized recovery from adaptation

The metric estimation approach described in the body of this paper uses continuous covariates to model recovery from adaptation for changes along each stimulus dimension. A
different analysis scheme may be employed in which a basis set of covariates models the BOLD fMRI signal associated with each transition between one stimulus and the next (120 in total if a symmetric directional effect is assumed). The loading upon these covariates may be used to construct a neural-adaptation similarity matrix, which may then be submitted to MDS or PROSCAL algorithms. See http://cfn.upenn.edu/aguirre/wiki/example_carry-over_data_and_analysis for an example set of data and basis covariates.

Here, we consider a post-hoc analysis that may be performed on this basis set of covariates to detect the presence of compressive non-linearities in the stimulus representations and BOLD response. This can be accomplished by directly comparing the stimulus distance and the distances implied by the BOLD fMRI signal in adaptation recovery. To do so, we consider the pure (one-dimensional) changes on each dimension, and measure them with respect to a reference stimulus. The BOLD signal change associated with ever larger transitions is obtained, and used to construct a representation of the degree of difference in BOLD signal between stimuli, related to the degree of difference in the stimuli themselves.

There are 16 points in the di-octagon space (Figure C1, Panel A); each position may be assigned a label (1-16). We wish to measure pure changes along a single dimension \( k \). Distances are measured relative to an origin; we adopt the position along the \( k \) dimension occupied by stimuli 7 and 8. We then obtain the BOLD response to transitions between \( 7 \rightarrow 15 \) and \( 8 \rightarrow 9 \); the mean of these measurements is segment \( A \), which now acts as an anchor point for measurements proceeding rightward. Relative to the origin, \( A \) represents the dimension \( k \) position of points 6, 15, 9, and 1. The mean BOLD response to transitions \( 6 \rightarrow 5, 15 \rightarrow 13, 9 \rightarrow 11, \) and \( 1 \rightarrow 2 \) distances is then obtained and termed \( B \). Segment \( C \) is provided by the transitions \( 13 \rightarrow 4 \) and \( 11 \rightarrow 3 \). The segment \( AB \) is provided by \( 7 \rightarrow 13 \) and \( 8 \rightarrow 11 \), the segment \( BC \) by \( 15 \rightarrow 4 \) and \( 9 \rightarrow 3 \), and the segment \( ABC \) by \( 7 \rightarrow 4 \) and \( 8 \rightarrow 3 \). The pure dimension \( k \) transition \( 16 \rightarrow 12 \) is not usable as it cannot be related to the origin.

Using these values we may then plot four points: (i) the fixed origin at zero, representing the dimension \( k \) position of points 7 and 8; (ii) the dimension \( k \) position of points 9 and 15 at \( A \); (iii) the dimension \( k \) position of points 5, 13, 11, and 2 at \( A+B \) and at \( AB \); and (iv) the dimension \( k \) position of points 4 and 3 at \( A+BC, AB+C, \) and \( ABC \).
The values obtained for these points may then be compared to the positions of the stimuli along the $k$ axis in the original stimulus space. The analogous process may be followed for interrogation of the orthogonal dimension $j$.

Figure C1, Panel B shows the resulting relationship between stimulus position and constructed BOLD response for a model di-octagonal space without any distortion, and for one in which a logarithmic transform has been applied along the $k$ dimension. Goodness-of-fit measures of linear and non-linear models to actual data may be used to test for the presence of a non-linearity in the integrated response (across neural representation and BOLD transform) across a dimension.
Figure Legends

Figure 1 – (A) An example stimulus space, defined by changes in color and shape of two-dimensional contours. (B) A population of idealized neurons with receptive fields conjointly tuned to stimulus color and shape. Each neuron responds optimally to a particular position in the stimulus space, and its response drops off as a function of distance from that point, regardless of which dimension is changed. (C) A population of idealized neurons independently tuned to respond to a particular color or a particular shape. Each neuron responds optimally to a particular position in one dimension of the stimulus space, and its response drops off as a function of distance from that point along one dimension only, but is not responsive to changes in the other dimension.

Figure 2 – In an example carry-over experiment a continuous sequence of stimuli are presented. The fMRI data that result are modeled with three covariates. The first two “City-block” covariates model the change from one stimulus to the next in color and in shape. The third covariate models the difference between the rectilinear, City-block distance between stimuli and the sub-additive, Euclidean distance.

Figure 3 – (A) A simulation of the test for conjoint and independent populations. The model begins with a set of stimuli with regular spacing. A population of neurons represent these stimuli through the pattern of their firing, and we model the similarity of the patterns of firing for stimulus pairs. The distance between neural patterns for the different stimuli can be linearly related to the physical distance between the stimuli, or distorted along either dimension in a linear or non-linear manner. A particular sequence of stimuli indicates a series of transitions between stimuli and a magnitude of neural activity modulated by adaptation based upon the overlap of neural population response to the two stimuli. The aggregate neural response may be linearly related to the overlap between the population responses, or modulated by a non-linear function. The vector of neural responses is transformed to a BOLD fMRI signal by convolution with a standard hemodynamic response function. The letters superimposed on the arrows reference the non-linearities presented in the other panels of this figure. (B) Depictions of the neural distances between stimuli resulting from the linear and non-linear distortions of the stimulus space examined in the model. (C) Non-linearities in the transformation of neural activity to BOLD fMRI response. Three different non-linearities are considered, each covering
the theoretical maximum range of neural firing for a region (blue curves). Neural adaptation generally produces a small (20%) modulation of signal strength (superimposed red segment). We modeled the worst-case hemodynamic non-linearity in which the neural adaptation modulation occurs within the maximally non-linear component of each function.

Figure 4 – (A) A population of neurons may represent a stimulus space with two sets of linear receptive fields, but these may have an orientation that is rotated with respect to the dimensional axes that define the stimuli; in this case by 22.5°. (B) This circumstance may be detected by repeating the analysis and assuming a rotated version of the actual stimulus space. (C) For a population of neurons with conjoint tuning, the loading on the Euclidean contraction covariate is unaltered with model rotation. Given independent tuning, in contrast, model rotation alters the Euclidean contraction beta value, with a value of zero reached when the assumed rotation of the model matches the actual alignment of the receptive fields. Two independently-tuned curves are shown, corresponding to experimental designs that have either 16 or 49 samples of the stimulus space. The negative “dip” of the independently tuned function with 16 samples under rotation is therefore shown to be an artifact of discreet sampling of a continuous stimulus space.

Figure 5 – (A) Performance of the model for a range of Minkowski exponent values defining the neural population. A red plus marks where the Euclidean contraction covariate has a value of zero for a Minkowski exponent of one, corresponding to independent neural populations. (B) The curves from panel A, expressed as the ratio of the Euclidean contraction to the City-block effect.

Figure 6 – (A) An example color stimulus space defined by hue and brightness with grid-spaced sampling. (B) A di-octagonal sampling of the same stimulus space.

Figure 7 – The Efficiency (Friston 1999) of n=17, type 1 index 1 counterbalanced sequences was calculated for the Euclidean contraction and City-block model covariates. 50000 label permutations (Aguirre 2007, Appendix A) were tested. Blue circles indicate the sequences with the highest Efficiency for one or both covariates.

Figure 8 – (A) A set of two-dimensional, closed contours defined by variations in radial frequency contours. The dimensions that define the stimulus space are perceptually integral. (B)
A set of closed contours that vary along two identifiable dimensions (curvature and thickness). These dimensions are perceptually separable.

Figure 9 – (A) The location of voxels showing adaptation for all four stimulus dimensions (the two “popcorn” dimensions and the two “moon” dimensions) across 6 subjects. The data are displayed atop an inflated, ventral cortical surface. (B) The signal change associated with linear recovery from adaptation for the four stimulus dimensions within the selected voxels, averaged across subjects. Because of the method of voxel selection, some signal change is guaranteed. The figure illustrates the comparable degree of response for all four dimensions. (C) Significant loading upon the Euclidean contraction covariate was observed across 6 subjects within the selected voxels for the popcorn stimulus space, suggesting that the two stimulus dimensions are not represented by independent neural populations. There was no loading upon the same measure for the moon stimulus space, thus failing to reject independent representation. (D) The possibility of independently tuned, but rotated, stimulus dimensions was tested for the popcorn space by examining the Euclidean contraction effect under assumed rotations of the stimulus space. The function never reached zero, confirming non-independence of the neural representation. (E) The recovery from adaptation for discretized stimulus changes along a single stimulus axis was obtained across subjects for each dimension of the stimulus space. A notably linear relationship was observed, arguing against a compressive non-linearity as the cause of the finding of conjoint representation. The value at the first stimulus position was fixed at zero. See Appendix C for details.

Figure A1 – Four hypothetical arrangements of neurons within voxels. For each of the cases, we consider four voxels, each voxel containing several neurons. Shown is the stimulus space for each voxel, with the receptive fields of the neurons contained within the voxel plotted as circles (conjointly tuned) or bars (independently tuned). As can be readily appreciated, voxels with conjoint or independent tuning can be constructed with neurons with either conjoint or independent tuning. Consequently, the metric property of neural similarity calculated across voxels is not necessarily related to the metric tuning properties of the neurons within voxels.

Figure C1 – (A) The di-octagonal “popcorn” stimulus space is shown. Below is the standard numbering and positions of the stimuli along the two dimensions, and with a
logarithmic transformation applied to one axis. (B) The relationship between aggregate BOLD response and stimuli may be examined in an attempt to detect non-linearities in the neural representation and/or hemodynamic transform. Shown are the stimulus transitions that contribute measurements of the BOLD fMRI response, which are then used to construct the graph shown. The solid, linear relationship is for simulated responses in which a linear mapping of stimulus space to neural adaptation and to BOLD response is produced. The dashed line indicates the simulated result for the same measurement when a logarithmic transform has been applied to the neural representation along dimension \( k \).
### Tables

<table>
<thead>
<tr>
<th>distortion of neural space</th>
<th>distortion of hemodynamic transform</th>
<th>Euclidean contraction (scaled)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>conjoint</td>
<td>independent</td>
</tr>
<tr>
<td>original</td>
<td>none (linear)</td>
<td>1.4</td>
</tr>
<tr>
<td>linear</td>
<td>none (linear)</td>
<td>1.2</td>
</tr>
<tr>
<td>sigmoid</td>
<td>none (linear)</td>
<td>0.9</td>
</tr>
<tr>
<td>quadratic</td>
<td>none (linear)</td>
<td>1.3</td>
</tr>
<tr>
<td>log</td>
<td>none (linear)</td>
<td>1.3</td>
</tr>
<tr>
<td>sigmoid x2</td>
<td>none (linear)</td>
<td>0.5</td>
</tr>
<tr>
<td>quadratic x2</td>
<td>none (linear)</td>
<td>1.3</td>
</tr>
<tr>
<td>log x2</td>
<td>none (linear)</td>
<td>1.4</td>
</tr>
<tr>
<td>original</td>
<td>sigmoid</td>
<td>1.4</td>
</tr>
<tr>
<td>linear</td>
<td>sigmoid</td>
<td>1.0</td>
</tr>
<tr>
<td>sigmoid</td>
<td>sigmoid</td>
<td>1.0</td>
</tr>
<tr>
<td>quadratic</td>
<td>sigmoid</td>
<td>1.3</td>
</tr>
<tr>
<td>log</td>
<td>sigmoid</td>
<td>1.2</td>
</tr>
<tr>
<td>sigmoid x2</td>
<td>sigmoid</td>
<td>0.7</td>
</tr>
<tr>
<td>quadratic x2</td>
<td>sigmoid</td>
<td>1.3</td>
</tr>
<tr>
<td>log x2</td>
<td>sigmoid</td>
<td>1.4</td>
</tr>
<tr>
<td>original</td>
<td>quadratic</td>
<td>1.4</td>
</tr>
<tr>
<td>linear</td>
<td>quadratic</td>
<td>1.2</td>
</tr>
<tr>
<td>sigmoid</td>
<td>quadratic</td>
<td>0.9</td>
</tr>
<tr>
<td>quadratic</td>
<td>quadratic</td>
<td>1.3</td>
</tr>
<tr>
<td>log</td>
<td>quadratic</td>
<td>1.4</td>
</tr>
<tr>
<td>sigmoid x2</td>
<td>quadratic</td>
<td>0.5</td>
</tr>
<tr>
<td>quadratic x2</td>
<td>quadratic</td>
<td>1.3</td>
</tr>
<tr>
<td>log x2</td>
<td>quadratic</td>
<td>1.4</td>
</tr>
<tr>
<td>original</td>
<td>log</td>
<td>1.5</td>
</tr>
<tr>
<td>linear</td>
<td>log</td>
<td>1.0</td>
</tr>
<tr>
<td>sigmoid</td>
<td>log</td>
<td>0.9</td>
</tr>
<tr>
<td>quadratic</td>
<td>log</td>
<td>1.2</td>
</tr>
<tr>
<td>log</td>
<td>log</td>
<td>1.0</td>
</tr>
<tr>
<td>sigmoid x2</td>
<td>log</td>
<td>0.5</td>
</tr>
<tr>
<td>quadratic x2</td>
<td>log</td>
<td>1.1</td>
</tr>
<tr>
<td>log x2</td>
<td>log</td>
<td>1.3</td>
</tr>
<tr>
<td>rotate 22.5°</td>
<td>none (linear)</td>
<td>1.2</td>
</tr>
<tr>
<td>rotate 45°</td>
<td>none (linear)</td>
<td>1.2</td>
</tr>
</tbody>
</table>
Table 1 - Average loading upon the Euclidean contraction covariate, scaled by the loading upon the City-block model covariates, for a simulation of independent and conjointly tuned neural populations. The average was obtained over 100 different, counterbalanced sequences that were selected for having high combined Efficiency for both the City-block and Euclidean contraction effects. The distortion of neural space was applied to a single stimulus dimension unless noted as “x2”, in which case it was applied to both stimulus dimensions (Figure 3B). Indicated in bold are those cases where a positive loading upon the Euclidean contraction covariate was obtained despite assuming independent neural populations tuned to the two stimulus axes, raising the possibility of improper bias.
A Stimulus Space

B Conjoint Tuning

C Independent Tuning

Neural Population Code
City-block model
\[
\begin{align*}
&\text{hue change} & 0 & 2 & 3 & 2 & 1 & \cdots \\
&\text{shape change} & 1 & 2 & 0 & 3 & 2 & \cdots
\end{align*}
\]

Euclidean contraction
\[
\begin{align*}
&\frac{(\text{hue} + \text{shape})}{\text{shape}} & 0 & -1.2 & 0 & -1.4 & -0.8 & \cdots
\end{align*}
\]
Stimulus space
Simulation of non-linearities
Neural space distortions
Stimulus sequence
Adapted neural response
Independent neural activity
adaptation effect
min max
BOLD signal
Conjoint BOLD fMRI signal

A Simulation of non-linearities

B Neural space distortions

C Hemodynamic transform distortions

adaptation effect
BOLD signal
neural activity
min max

linear
quadratic
sigmoidal
logarithmic
quadratic x 2
sigmoidal x 2
logarithmic x 2

line
quadratic
sigmoidal
logarithmic
Euclidean Contraction $\beta$ value

Effect of model rotation

A

Independent, rotated neural tuning

B

Model rotation

C

Effect of model rotation

Euclidean Contraction $\beta$ value

Conjoint

Independent (n=49)

Independent (n=16)

0 22.5 45

model rotation [deg]
Minkowski index

\[ (\text{covariate } \beta, \text{value}) \]

independent

\[ \text{Euclidean contraction : city-block} \]

\[ (\text{Minkowski index}) \]

\[ 0.5, 1, 2 \]

contraction
Di-oct stimulus arrangement [n=17]
A  ROI overlap across subjects

B  Adaptation effect

C  Euclidean contraction

D  Effect of model rotation

E  Relation of stimulus change to BOLD response (popcorn)
Conjoint Neural Tuning

Independent Neural Tuning

Conjoint Voxel Tuning

Independent Voxel Tuning
A

B

aggregate BOLD response

stimulus change along k