**Using multiple high-dimensional feature spaces to model brain activity recorded during naturalistic experiments**

by

Anwar O. Nunez-Elizalde

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Neuroscience

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jack L. Gallant, Chair
Professor Richard B. Ivry
Professor Joni D. Wallis
Professor Bin Yu

Summer 2018

**Using multiple high-dimensional feature spaces to model brain activity recorded during naturalistic experiments**

## Abstract

Using multiple high-dimensional feature spaces to model brain activity recorded during naturalistic experiments

by

Anwar O. Nunez-Elizalde

Doctor of Philosophy in Neuroscience

University of California, Berkeley

Professor Jack L. Gallant, Chair

The human cerebral cortex comprises many functionally distinct areas that represent different information about the world. It has been challenging to map these areas efficiently. In this dissertation, I present a new approach that addresses this problem. In chapter one, I present a novel voxelwise encoding model based on Tikhonov regression. I discuss the theoretical basis for Tikhonov regression, demonstrate a computationally efficient method for its application, and show several examples of how Tikhonov regression can improve predictive models for fMRI data. I also show that many earlier studies have implicitly used Tikhonov regression by linearly transforming the regressors before performing ridge regression. In chapter two, I present a critique of an alternative method used to study brain representations called representational similarity analysis. I show that this method makes strong assumptions about the relationship between representational models and brain responses. I also show that representational similarity analysis can lead to incorrect conclusions when used to compare representational models. In chapter three, I present a rich paradigm for efficient non-invasive functional brain mapping. In this paradigm, subjects watch interesting short films while their brain activity is measured. Multiple feature spaces are used to model the brain responses to the short films. Each feature space constitutes a hypothesis about the type of representations that might be important for brain regions involved in watching, listening, and understanding the short films. The novel voxelwise encoding model developed in chapter one is then used to find the most predictive feature spaces across the cortical surface and also to recover maps that capture how the individual feature spaces are represented within cortical regions. The results suggest a high degree of homogeneous selectivity for feature spaces across large regions of the cortical surface within individual subjects. These patterns are

highly consistent across all subjects. Finally, I explore the functional organization of the middle temporal cortex and show that the visual feature spaces can capture novel functional subdivisions in this region.

# Contents

# List of Tables

# Acknowledgments

I would like to thank all the people who have helped me during my time in the lab. Tolga Çukur for guidance during my rotation, for help with sequence development, and for experimental design and piloting for the decision-making project. Shinji Nishimoto for help with decoding semantics from silent movies during my rotation, for experimental design and input for the decision-making project, and in general for just knowing everything. Dustin Stansbury for help with building HRF-separable encoding models using generalized least squares. Brittany Griffith for her dedication and hard work making beautiful brain segmentations and cortical surfaces. An Vu for help with lots of fun, yet inconclusive, hours of 7T piloting. Michael Oliver for many an interesting discussion on statistical learning and philosophy. Mark Lescroart for discussions on estimation, statistics and inference, and for making karaoke happen. Natalia Bilenko for help coming up with the short films visual imagery task, and for introducing me to the wonders of LPCA. Alex Huth for our collaboration on Tikhonov regression and RSA, for many helpful suggestions in several projects, and for valuable conversations on neuroscience and computation. James Gao for help in the early days of "mritools" and "brain bullets", for allowing me to create torturous audio-visual stimuli and memorize those stimuli during way too many hours of multi-band sequence testing, for maintaining a steady cluster with magical powers, and for knowing all things hardware and software. Leila Wehbe for our collaboration on the deep multi-view autoencoder and visual imagery, for advice in various projects, for being an excellent debate partner on all matter of topics, and for providing support and encouragement. Fatma Deniz for our collaboration on the bilingual project which proved I did not waste hours and hours watching videos online while in undergrad, for help with assessing the generalization of the short films to stories, for providing valuable feedback on the short films manuscript, and for tremendous support and perspective during my time in graduate school. The rest of the Gallant Lab for providing an excellent environment for research, and in particular Tianjiao Zhang and Christine Tseng for moving forward with projects I could not. And finally, Jack Gallant for maintaining a high standard of scientific research and for pushing in that direction even when difficult, for valuable insights into the philosophy of scientific research and academia, and for being extremely supportive and understanding during times of personal need.

# Chapter 1

# Spatiotemporal encoding models with multivariate normal priors

## 1.1 Overview

Predictive models for neural or fMRI data are often fit using regression methods that employ priors on the model parameters. One widely used method is ridge regression, which employs a spherical Gaussian prior that assumes equal and independent variance for all parameters. However, a spherical prior is not always optimal or appropriate. There are many cases where expert knowledge or hypotheses about the structure of the model parameters could be used to construct a better prior. In these cases, non-spherical Gaussian priors can be employed using a generalized form of ridge known as Tikhonov regression. Yet Tikhonov regression is only rarely used in neuroscience. In this chapter we discuss the theoretical basis for Tikhonov regression, demonstrate a computationally efficient method for its application, and show several examples of how Tikhonov regression can improve predictive models for fMRI data. We also show that many earlier studies have implicitly used Tikhonov regression by linearly transforming the regressors before performing ridge regression.

## 1.2 Introduction

Cognitive and systems neuroscience has in recent years become increasingly reliant on predictive encoding models. In the fMRI literature, encoding models have produced insights into the cortical representations of visual (Thirion et al., 2006, Kay et al., 2008b, Nishimoto et al., 2011, Huth et al., 2012, Lescroart et al., 2015), auditory (de Heer et al., 2017, De Angelis et al., 2017), and linguistic (Mitchell et al.,

2008, Wehbe et al., 2014, Huth et al., 2016) information. To efficiently estimate the parameters of encoding models, many studies use L2-regularized (ridge) regression (Hoerl and Kennard, 1970). L2 regularization improves regression models by imposing a multivariate normal prior on the model parameters, where the mean of the prior is zero and the covariance is spherical. Compared to unregularized regression, ridge makes models better at generalizing to new data and decreases overfitting by shrinking model parameter estimates towards zero and improving estimation for features that are nearly collinear. However, assuming a spherical covariance is rarely optimal, and in many cases prior information or expert knowledge can be used to construct informative non-spherical priors. In this chapter we explore how non-spherical priors can be applied to several encoding model problems and show that this can greatly improve model performance. We also show that some previously published encoding models can be reinterpreted in terms of non-spherical priors, providing new insights into why those models were successful. Finally, we offer practical advice and efficient methods for estimating encoding models with non-spherical priors.

Although encoding models have proven highly successful for modeling fMRI data, there are several complications that make them difficult to use. First, in many feature spaces it is difficult to assign a specific interpretation to the features. This problem is especially acute for feature spaces learned using unsupervised methods, such as the word embedding space word2vec (Mikolov et al., 2013). When using these feature spaces to predict neural or BOLD responses, it is difficult to interpret what exactly a given voxel represents.

Second, it is often unclear how the regularization method used for regression interacts with the choice of feature space. For example, feature spaces that are identical up to a linear transformation (i.e. $\mathbb{L}_1(s) = \mathbb{L}_2(s)P$) can yield drastically different results even though both span the same space.

Third, although the basic shape and variability of the HRF are reasonably well understood (Glover, 1999), many studies do not use this prior information when estimating the HRF (Kay et al., 2008a, Nishimoto et al., 2011, Huth et al., 2012), and most studies simply assume a single canonical HRF for all voxels (Penny et al., 2011).

Fourth, it is becoming increasingly important to characterize how and where different feature spaces overlap in terms of variance explained (Lescroart et al., 2015, de Heer et al., 2017). This is usually done by combining different feature spaces into one encoding model. However, ordinary regularization techniques wrongly assume that all feature spaces require the same level of regularization.

Here we address all of these issues by constructing encoding models using carefully designed multivariate normal priors. In the standard encoding model formulation, complex features are extracted from the stimuli and then regularized regression is

used to learn model parameters subject to simple priors. In the new framework presented here, we extract simple, interpretable features from the stimuli, and then use Tikhonov regression (Tikhonov et al., 1977) to learn parameters subject to complex multivariate priors. This is made possible by a duality between imposing a prior and extracting features from the stimuli, so the exact same model can be represented in both ways. This simple change in perspective has significant consequences for model interpretation, because it shows that a complex feature space can be decomposed into a combination of a simple feature space and a multivariate prior. This framework is also highly modular, making it easy to combine different spatial and temporal priors and test many different kinds of priors.

We evaluate each proposed application of our framework on empirical data from naturalistic experiments on vision and language. We show that non-spherical multivariate normal priors can improve prediction accuracy in a variety of settings. In order to encourage the adoption of the framework presented here, we have released an open-source Python software package that efficiently implements all the models described in this chapter (`http://github.com/gallantlab/tikypy`).

### 1.2.1 Linearized predictive encoding models

In a typical fMRI experiment, brain images $y(t) \in \mathbb{R}^m$ are recorded at times $t = 1 \dots T$ while a subject is exposed to stimuli $s(t)$. Each brain image consists of $m$ voxels, $y_\ell(t)$ for $\ell = 1 \dots m$. The goal of the encoding model framework is to find a function $f_\ell$ that maps stimuli to BOLD responses in each voxel: $f_\ell(s(1), \dots, s(t)) \approx y_\ell(t)$. Because the space of possible functions is extremely large, it is common to work under a hypothesis that limits the complexity of $f$. Although there often are many reasonable hypotheses that one can make about $f$ (Wu et al., 2006), the only type that we shall consider here is where $f$ is a linear combination of features that are extracted from the stimulus, usually by a nonlinear function. In this case, $f$ is called a "linearized" model, and the function that extracts features from the stimulus, $\mathbb{L}_s(t) \in \mathbb{R}^{1\times p}$, is called the "linearizing transformation" (Wu et al., 2006). Formally, $\mathbb{L}_s(t)$ maps a $u$-dimensional stimulus at time $t$ into a $p$-dimensional vector of stimulus features $x_i(t)$.

$$\mathbb{L}_s(t) : s(t) \in \mathbb{R}^{1\times u} \mapsto x(t) \in \mathbb{R}^{1\times p}$$

Under the linearized model formulation, the brain response is modeled as a linear combination of the stimulus features, usually over a fixed time window $d$,

$$y_\ell(t) = \left[ \begin{array}{cccc} x(t) & x(t-1) & \dots & x(t-d) \end{array} \right] \beta_\ell + \epsilon_\ell(t),$$

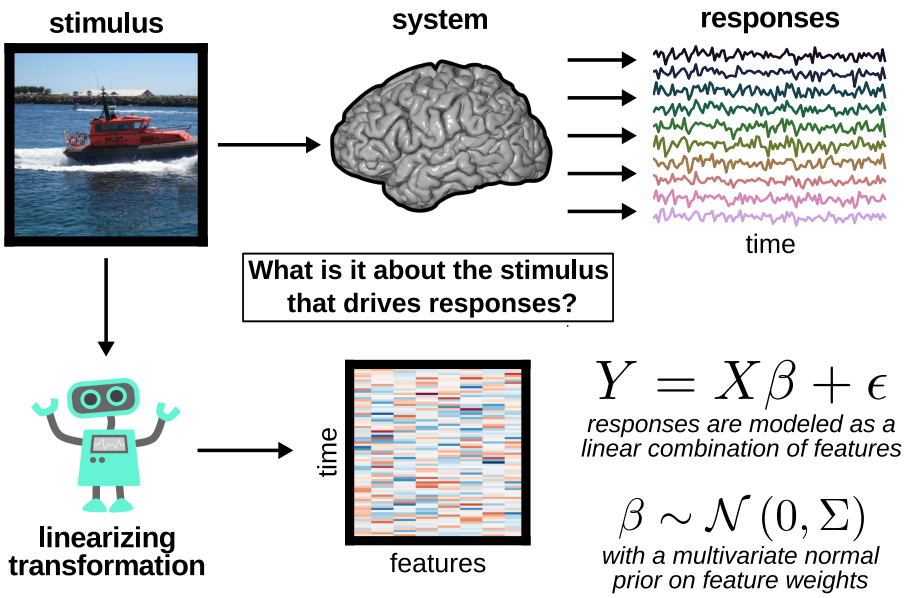


**Figure 1.1: Modeling the stimulus-response relationship with linearized predictive encoding models and multivariate normal priors.** In a typical experiment, a series of stimuli are shown to the subject and the brain responses recorded. Features are extracted from the stimuli using a computational model, human labels, or any other method. The brain responses are then modeled as a linear combination of the features. When using very large models, some form of regularization is often used. A common approach in computational neuroscience is to impose a multivariate normal distribution on the feature weights. When the MVN is spherical, this is called ridge regression (Hoerl and Kennard, 1970). In general, the MVN prior can also have non-spherical structure. This is referred to as Tikhonov regression (Tikhonov et al., 1977). A goal of modeling the data using this approach is to have an accurate model that can predict brain responses to novel stimuli and is also interpretable.

where $\epsilon_\ell(t) \sim \mathcal{N}(0, \sigma_\ell^2)$ is stationary, zero-mean normal noise, $x(t-d) \in \mathbb{R}^{1\times p}$ is the feature vector delayed $d$ time points, and $\beta_\ell \in \mathbb{R}^{pd\times 1}$ is a set of linear weights over the $p$ features at each of the $d$ delays.

To write the simultaneous equation for all voxels we replace $y_\ell(t)$ with a matrix $Y \in \mathbb{R}^{T\times m}$ that contains the response of each voxel at each timepoint, and we replace $\beta_\ell$ with a matrix $\beta \in \mathbb{R}^{pd\times m}$ that contains the weight vector for every voxel. We write the matrix of linearized stimulus features as $X \in \mathbb{R}^{T\times pd}$,

$$X = \begin{bmatrix} x(0) & \mathbf{0}_p & \cdots & \mathbf{0}_p \\ x(1) & x(0) & \cdots & \mathbf{0}_p \\ \vdots & \vdots & \vdots & \vdots \\ x(t) & x(t-1) & \cdots & x(t-d) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_p & \mathbf{0}_p & \cdots & x(T) \end{bmatrix} = \begin{bmatrix} X_{\delta(0)} & X_{\delta(1)} & \cdots & X_{\delta(d)} \end{bmatrix}, \quad (1.1)$$

where each row of $X$ contains the feature vectors for the past $d$ timepoints. Each block of $p$ columns $X_{\delta(j)}$ contains the linearized stimulus feature matrix delayed by $j$ time points. This is referred to as a finite impulse response model (Oppenheim et al., 1983). This allows us to rewrite the basic model as:

$$Y = X\beta + \epsilon$$

where $\epsilon_t \sim \mathcal{N}_m(0, \text{diag}\{\sigma_1^2, \cdots, \sigma_m^2\})$ is zero-mean, independent noise for each voxel at each time point. The only free parameter in this formula is the weight vector $\beta \in \mathbb{R}^{pd \times m}$.

We can find an estimate of $\beta$ by maximizing the probability of the data $Y$ given the stimulus features $X$

$$\hat{\beta} = \underset{\beta}{\text{argmax}}\ P\left(Y|X,\beta\right).$$

This estimate of $\beta$ is called the maximum a posteriori (MAP) estimate. We can derive various analytic solutions depending on the form of the distribution we assume for $P(Y|X,\beta)$. In this chapter, we assume that the responses can be modeled as multivariate normal random variables.

The likelihoood of the data can be expressed as

$$P\left(Y|X,\beta\right) \propto \frac{1}{\det\left(\Sigma_\epsilon\right)} \exp\left(-\frac{1}{2}\text{trace}\left(\left(Y - X\beta\right)^\top \Sigma_\epsilon^{-1}\left(Y - X\beta\ \right)\right)\right),$$

where $\Sigma_\epsilon = \text{diag}\{\sigma_1^2, \cdots, \sigma_m^2\}$ contains the variance of the noise for each voxel. If we assume that the noise variance is the same in each voxel, then we can set $\Sigma_\epsilon = \sigma^2 I$. We can also switch to using the log of the likelihood instead of the likelihood. The log-likelihood of the data can then be expressed as

$$\log P\left(Y|X,\beta\right) \propto -\frac{1}{2}\sum_{\ell}^{m}\left(\frac{1}{\sigma^2}\left\|y_\ell - X\beta_\ell\right\|_2^2\right).$$

Finally, note that maximizing the log-likelihood is equivalent to minimizing the negative log-likelihood of the data. The $\beta$ estimate for all voxels can be found simultaneously by solving

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[ \frac{1}{2} \left\| Y - X\beta \right\|_2^2 \right],$$

which is equivalent to finding the $\beta$ that minimizes the squared difference between predicted and actual responses (i.e. ordinary least squares regression).

However, finding the value of $\beta$ that exactly minimizes this squared error function often produces results that do not generalize to new stimuli. This is due to overfitting. Overfitting occurs when model parameters capture the noise $\epsilon$ in addition to the underlying signal. This is a common problem when the data available to estimate the model parameters is small. When building a predictive encoding model our goal is not simply to explain the data that is given, but to predict new data.

### 1.2.2 Ridge regression

To avoid overfitting it is common to employ regularized regression techniques (Friedman et al., 2001). Regularization imposes a prior distribution on $\beta_\ell$. This prior limits how well a model can explain the given data. The goal becomes to maximize the probability of the observed data, by finding the $\beta$ that maximizes the product of the likelihood and the prior

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \, P\left(Y|X,\beta\right) P\left(\beta\right).$$

One commonly used regularization technique is ridge regression, which imposes a zero-mean multivariate normal prior on the individual voxel weights $\beta_\ell$ (Hoerl and Kennard, 1970).

$$\beta_\ell \sim \mathcal{N}_p(0, \lambda^{-2} I_p),$$

$$P\left(\beta_\ell\right) \propto \exp\left( -\frac{1}{2} \beta_\ell^\top \lambda^2 I_p \beta_\ell \right) = \exp\left( -\frac{1}{2} \left\| \lambda \beta_\ell \right\|_2^2 \right).$$

Ridge regression can be implemented by adding a penalty term to the error function, where the penalty is proportional to the sum of the squared weights,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[ \left\| Y - X\beta \right\|_2^2 + \left\| \lambda \beta \right\|_2^2 \right], \tag{1.2}$$

and the strength of the regularization is controlled by $\lambda$, the regularization coefficient. The closed-form solution for the ridge regression problem is given by

$$\hat{\beta} = (X^\top X + \lambda^2 I)^{-1} X^\top Y$$

### 1.2.3 Tikhonov regression

Ridge regression imposes a zero-mean, spherical multivariate normal prior on the feature weights. However, expert knowledge can be used to create a more sophisticated, non-spherical multivariate normal prior on the weights,

$$\beta \sim \mathcal{N}_p(0, \lambda^{-2}\Sigma),$$

$$P(\beta) \propto \exp\left(-\frac{\lambda^2}{2}\beta^\top \Sigma^{-1}\beta\right),$$

where $\Sigma \in \mathbb{R}^{p\times p}$ is the positive semidefinite prior covariance matrix. Note that $\lambda$ is present as a scaling factor on $\Sigma$. This determines how much influence the prior has on the estimated weights.

If we factorize the inverse of the prior covariance matrix by taking its matrix square root $\Sigma^{-1} = C^\top C$, where $C \in \mathbb{R}^{p\times p}$, then

$$P(\beta) \propto \exp\left(-\frac{\lambda^2}{2}\beta^\top C^\top C\beta\right) = \exp\left(-\frac{1}{2}\left\|\lambda C\beta\right\|_2^2\right)$$

The problem can then be solved by maximizing the product of the likelihood and this new prior, or, as above, by minimizing the negative log likelihood,

$$\hat{\beta}_T = \underset{\beta}{\operatorname{argmin}} \left[\|Y - X\beta\|_2^2 + \|\lambda C\beta\|_2^2\right].$$

This is known as Tikhonov regression (Tikhonov et al., 1977). Here $C$ can be thought of as a penalty matrix that punishes $\beta$ when it does not conform to the prior. However, since there are many matrix square roots, $C$ is not uniquely determined by the prior covariance, and in fact any $C$ that satisfies the given relation will produce the same $\hat{\beta}_T$. Also note that when $C = I_p$ Tikhonov regression reduces to ridge regression (Hoerl and Kennard, 1970).

The Tikhonov minimization problem has a closed form solution,

$$\hat{\beta}_T = (X^\top X + \lambda^2 C^\top C)^{-1} X^\top Y.$$

However, this formulation does not immediately admit efficient computational solutions, making it less useful for solving large-scale problems. Fortunately there is a computationally efficient method for solving Tikhonov regression problems. This method, which is often referred to as the "standard form" (Hansen, 1998), transforms a Tikhonov problem into a ridge regression problem. This transformation is accomplished in three steps.

First, a linear transformation is applied to $X$, giving

$$A = XC^{-1}.$$

Second, ridge regression is carried out with $A$, giving

$$\hat{\beta}_A = (A^\top A + \lambda^2 I_p)^{-1} A^\top Y.$$

Third, the estimated weights are projected back into the original space to give the Tikhonov estimate,

$$\hat{\beta}_T = C^{-1}\hat{\beta}_A.$$

(For a proof of this see Appendix 4.1). Because the standard form uses ridge regression internally, it is clear that Tikhonov regression in the standard form will admit the same efficient computational solutions as ridge regression.

The standard form transformation can be used to convert any Tikhonov regression problem into a ridge regression problem by way of a linear transformation of $X$. By the same logic, any linear transformation of $X$ followed by ridge regression is equivalent to some Tikhonov regression problem, and thus some non-spherical multivariate prior on the model weights. This relationship has interesting implications for a number of neuroimaging studies that have applied ridge regression to linearly transformed stimuli, because the models employed by those studies can be re-interpreted as Tikhonov regression with non-spherical priors. We use this technique to explore and re-interpret the models used in some previous studies.

## 1.3 Using multivariate normal priors

### 1.3.1 Feature priors

#### 1.3.1.1 Word embeddings

Several earlier studies have used word embedding spaces to model how the brain represents the meaning, or semantic content, of words (Mitchell et al., 2008, Wehbe et al., 2014, Huth et al., 2016). In this approach, each word is converted into a vector with anywhere from 20 (Mitchell et al., 2008) to 1000 (Huth et al., 2016) embedding dimensions. These vectors are constructed using word co-occurrence statistics from large corpora of text (Turney and Pantel, 2010), and are designed such that words with similar or related meanings (such as 'month' and 'week') are assigned similar vectors, but words with dissimilar meanings (such as 'month' and 'tall') are not. After converting words to vectors, regression models are used to predict BOLD responses as a function of the embedding dimensions.

Formally, this approach starts by defining a matrix of word indicator variables $X \in \mathbb{R}^{n \times p}$, where $X_{t,i} = 1$ if word $i$ was presented at time $t$ and 0 otherwise. Here $n$ is the total number of time points and $p$ is the total number of words in the experiment. Then, in order to replace each word with its $q$-dimensional embedding vector, the indicator matrix is multiplied with an embedding matrix $E \in \mathbb{R}^{p \times q}$ whose rows contain the word embedding vectors. Finally, regression is performed in the embedding space, yielding the linear model

$$Y = (XE)\beta + \epsilon.$$

Interestingly, this formulation appears identical to the standard form transformation of Tikhonov regression (see Appendix 4.1). If the model weights, $\beta$, are estimated using ridge regression (Wehbe et al., 2014, Huth et al., 2016), then this approach is equivalent to Tikhonov regression where the features are word indicators (i.e. the feature matrix is simply $X$), and the prior covariance is given by dot products between embedding vectors, $\Sigma = EE^\top$ (and $C^{-1} = E$).

Thus, the word embedding approach is equivalent to putting a multivariate normal prior on the model weights across words, such that the prior covariance between weights for different words is equal to the dot product between their embedding vectors. If words that have similar meanings have similar embedding vectors, then the dot product between those vectors will be high, and the weights for those words will covary strongly. This re-interpretation of the word embedding approach seems in many ways to be more natural and intuitive than thinking of it as regression in the word embedding space, which is highly abstract and difficult to explain.

#### 1.3.1.2 Evaluating feature MVN priors

To illustrate the Tikhonov approach to word embeddings we estimated two different linear models using the data from (Huth et al., 2016). Both models use words as features, but one model applies an identity prior to the weights while the other applies a semantic similarity prior based on a word embedding space. The data come from an experiment where subjects listened to approximately two hours of naturally spoken narrative stories while undergoing continuous BOLD fMRI. The stories were transcribed and then the transcripts were aligned to the audio to determine exactly when each word was spoken. These aligned transcripts were then used to generate the word indicator matrix, $X$, which contains the number of times each word was spoken during each time slice (here of length 2.0045 seconds, the $T_R$ of the fMRI scan).

The first linear model was estimated separately for each voxel in the fMRI scan using an identity prior on the model weights:

$$Y = X\beta_X + \epsilon$$

$$\hat{\beta}_X = (X^\top X + \lambda^2 I)^{-1} X^\top Y$$

The second linear model was estimated using a semantic prior based on a word embedding space. This embedding space was constructed by computing the statistical co-occurrence of each word in the stories with 985 common English words (see Huth et al., 2016, for details). To apply the semantic prior, the word indicator matrix, $X$, was projected onto the embedding matrix, $E$, and then ridge regression was used to estimate the weights:

$$Y = XE\beta_E + \epsilon$$

$$\hat{\beta}_E = ((XE)^\top (XE) + \lambda^2 I)^{-1} (XE)^\top Y$$

Finally, we used both sets of weights to predict BOLD responses on a separate 10-minute story that had not been used for model estimation, and then computed the correlation between predicted and actual BOLD responses. This model evaluation procedure resulted in two correlation coefficients for each voxel: one for the identity prior and one for the semantic prior. To compare these values we aggregated the data from all seven subjects, and then computed a 2D histogram of the correlation values (Figure 1.2).

Figure 1.2 shows that model prediction performance is nearly always higher with the semantic prior than with the identity prior, often substantially so. Of approximately 150,000 voxels included in the analysis, about 300 were significantly predicted by the identity prior model, and about 15,000 were significantly predicted by the semantic prior model ($n = 290, q(FDR) < 0.05$). The difference in model prediction performance is large and significant (Wilcoxon $W = 10^9, p < 10^{-12}$). At worst, we see that some voxels are predicted about as well by both models.

These results suggest that the semantic prior is a much better reflection of the true underlying voxel weights than the identity prior, and thus supports the earlier conclusion that those voxels represent information about the semantic content of language (Mitchell et al., 2008, Wehbe et al., 2014, Huth et al., 2016).

### 1.3.2 Temporal priors

The temporal activation pattern of the BOLD response is referred to as the hemodynamic response function (HRF). While the neurovascular mechanisms underlying

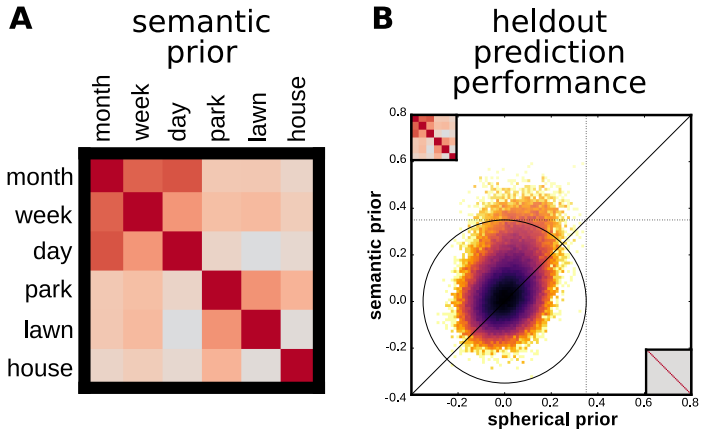


**Figure 1.2: An encoding model estimated with a semantic prior yields more accurate predictions than ridge regression**. Two subjects listened to spoken stories while their brain responses were measured with fMRI (Huth et al., 2016). Brain responses were modeled as a linear combination of the spoken words in the stories. We tested whether model accuracy improves when using information about the semantic similarity of words. **(B)** A semantic prior was constructed from word co-occurrence statistics estimated in a separate corpus data. The semantic prior captures the idea that words that occur close together are semantically related. The semantic prior was used to estimate a voxel-wise encoding model using Tikhonov regression. The same model was also estimated with a spherical prior (ridge regression). **(A)** We assessed model accuracy by computing the correlation between predicted and actual voxel responses to a novel, held-out stimulus. The prediction accuracy was significantly higher (Wilcoxon $W = 10^9, p < 10^{-12}$) when using the semantic prior (mean pearson $r = 0.037$) relative to the spherical prior (i.e. ridge regression; mean pearson $r = 0.005$). This suggests that brain responses are better modeled by including information about the meaning of words (semantics).

the HRF are not well understood, the shape of the HRF has been extensively studied in humans (Boynton et al., 1996, Glover, 1999). At a first approximation, the neural activation evoked by a stimulus leads to changes in blood-oxygenation that peak 4 to 6 seconds after stimulus onset. Several studies have shown that the shape of the HRF is highly variable across voxels and brain regions both within and across subjects (Aguirre et al., 1998, Handwerker et al., 2004, Kay et al., 2008a). It is important to take this variability into account by estimating the shape of the HRF for each voxel when modeling BOLD responses.

A common approach to estimating the HRF is the use of finite impulse response (FIR) models (Kay et al., 2008a). In an FIR model brain responses are modeled as a linear combination of ($p$) features over a fixed time window ($d$) prior to the stimulus onset (see Equation 1.1). The number of parameters in an FIR model is much larger ($p \times d$) than the original number of features ($p$), and grows linearly with

the length of the time window $d$. This increase in the number of parameters can lead to overfitting. In order to reduce overfitting, it is important to regularize FIR models.

When ridge regression is used to estimate FIR models, the implicit assumption is that feature weights are independent across time. This happens because ridge imposes a spherical prior on the temporal covariance of each feature, $\beta_i \sim \mathcal{N}_d\left(0, \lambda^{-2} I_d\right)$, where $\beta_i \in \mathbb{R}^d$ is the vector of weights for feature $i$ across the time window. In the Tikhonov framework, we can relax this assumption by specifying temporal priors that are not spherical,

$$\beta_i \sim \mathcal{N}_d\left(0, \lambda^{-2}\Sigma^T\right).$$

An insight worth highlighting is that applying a temporal prior is equivalent to convolution followed by ridge regression (Appendix 4.2). This follows from the fact that FIR models can be understood as convolution. In the context of Tikhonov regression, this means that applying a temporal prior of the form $\Sigma^T = (C^\top C)^{-1}$ is equivalent to convolving each feature timecourse with a set of temporal filters given by the columns of $C^{-1}$. When $C^{-1} = I$ the features are convolved with Kronecker delta functions at different delays, which is identical to using delays.

#### 1.3.2.1 Smoothness temporal prior

One simple and widely studied temporal prior holds that feature weights are smooth across time. This type of prior is typically applied by defining the penalty matrix $\mathbb{D} \in \mathbb{R}^{d \times d}$ to be a discrete difference operator that penalizes differences between neighboring weights in time,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[\|Y - X\beta\|_2^2 + \|\lambda \mathbb{D}\beta\|_2^2\right].$$

In the Tikhonov framework, this corresponds to a multivariate normal prior with covariance $\mathbb{D}^{-2}$ (Wu et al., 2006)

$$\beta_i \sim \mathcal{N}_d\left(0, \lambda^{-2}\mathbb{D}^{-2}\right).$$

This and similar approaches have been used in several studies (Goutte et al., 2000, Marrelec et al., 2003, Casanova et al., 2008, Bazargani and Nosratinia, 2014).

#### 1.3.2.2 HRF temporal prior

A more empirically-grounded possibility is to use published mathematical descriptions of the HRF to form a prior (Boynton et al., 1996, Friston et al., 1998, Glover,

1999). Previous work has resulted in the characterization of the commonly used "canonocal" HRF. This canonical HRF ($h_1$), its temporal derivative ($h_2$), and its derivative with respect to time-to-peak (dispersion; $h_3$) together provide an informed basis set that can capture some of the empirical variation observed in HRF shapes (Friston et al., 1998). The basis set is a matrix $\mathbb{H} \in \mathbb{R}^{d \times 3}$

$$\mathbb{H} = \begin{bmatrix} | & | & | \\ h_1 & h_2 & h_3 \\ | & | & | \end{bmatrix},$$

where each $h_j$ is a basis vector of length $d$. However, this basis set is not always flexible enough to capture all voxel- or region-specific variability of the HRF (Woolrich et al., 2004, Kay et al., 2008a, Pedregosa et al., 2015). In such cases, an FIR model with enough statistical power can better estimate the shape of the HRF. In practice, however, the FIR model might be difficult to estimate correctly because the large number of parameters ($p \times d$) can lead to overfitting.

Instead of choosing between FIR and HRF-based models, the Tikhohnov framework offers an intermediate approach by allowing us to trade off between both options. To achieve this, we compute the dot product of the HRF temporal basis set and use it as a non-spherical temporal prior on the feature weights,

$$\beta_i \sim \mathcal{N}_d \left(0, \lambda^{-2} \mathbb{H}\mathbb{H}\mathbb{H}^\top\right).$$

As $\lambda^{-2}$ decreases, the effect of the prior on the FIR weights is minimal. On the other hand, as $\lambda^{-2}$ increases, the prior has more effect on the FIR weights.

#### 1.3.2.3 Evaluating temporal MVN priors

In order to evaluate and compare these temporal priors we estimated three encoding models. The first encoding model was estimated with ridge regression, which imposes a spherical temporal prior $\Sigma^T = I_d$. In the second model, we used Tikhonov regression to impose a smoothness temporal prior on the FIR delays $\Sigma^T = \mathbb{D}^{-2}$. Finally, in a third model we imposed a temporal prior constructed from the covariance of an HRF basis set $\Sigma^T = \mathbb{H}\mathbb{H}\mathbb{H}^\top$ (Friston et al., 1998). All models had the same number of parameters and only differed in the temporal prior used.

We used data from an fMRI experiment in which three subjects watched natural movies while their brain activity was recorded (Huth et al., 2012). A total of 6,555 motion-energy features were extracted from these movies using a three-dimensional Gabor pyramid (Adelson and Bergen, 1985, Watson and Ahumada, 1985, Nishimoto et al., 2011). We used 10 temporal delays in order to account for the HRF (0-20

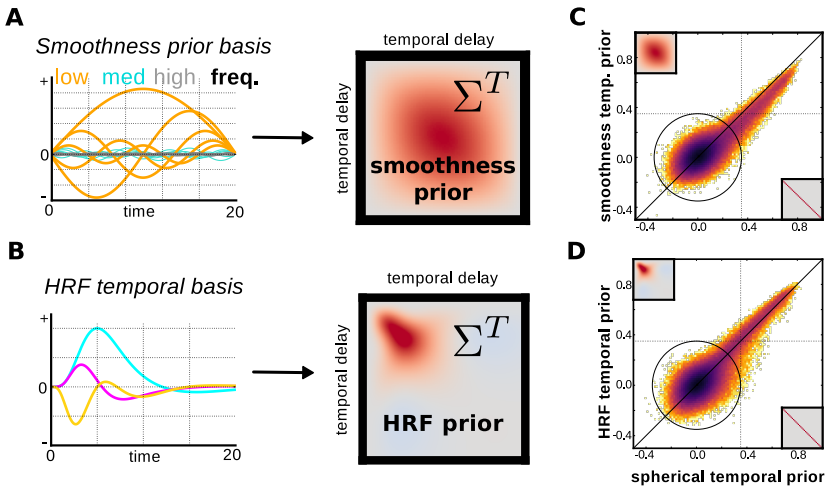


**Figure 1.3: An HRF temporal prior improves prediction performance relative to a spherical temporal prior, a smoothness prior does not.** We modeled BOLD responses collected from three subjects while they watched natural movies as a linear combination of motion-energy features (Nishimoto et al., 2011). In order to account for the HRF, we included 10 temporal delays (0-20 seconds). We estimated three separate encoding models, each one with a different MVN prior on the covariance of the temporal delays. We tested two non-spherical temporal MVN priors and one spherical prior (ridge regression). **(A)** The smoothness priors corresponds to a second order difference operator penalty on the temporal weights. This captures the idea that BOLD responses are smoothly varying in time. **(B)** We also constructed a hemodynamic response function (HRF) temporal prior from previous studies. The HRF temporal prior is computed as the temporal covariance of three basis functions (Friston et al., 1998). **(C)** The smoothness prior acts as a low-pass filter on the BOLD responses. The extracted basis functions are a Fourier basis. The smoothness prior does not improve prediction performance on held-out data. This is because it enforces high covariance in the mid-way in the time-course of the HRF, which is not an appropriate prior. **(D)** The HRF temporal basis improves prediction performance in some well-predicted voxels. The improvement in performance is nevertheless small.

seconds). This resulted in an FIR model with a total of 65,550 channels and 3,600 time points. We selected the regularization parameter, $\lambda$, using a cross-validation procedure (5-fold cross-validation repeated 20 times). This was done separately per voxel for each of the three encoding models estimated. We evaluated model performance for each model by computing the correlation coefficient between predicted and actual BOLD responses on a held-out dataset, which was not used for estimation. The held-out dataset consisted of 270 samples and was constructed by taking the mean temporal BOLD signal across 10 repetitions of a 540 second movie (Schoppe et al., 2016).

A total of approximately 230,000 voxels from four subjects were used in the

analyses (Figure 1.3). We find that the HRF basis set temporal covariance prior provided better predictions than either the spherical prior or the smoothness prior for the best voxels in population (top 10,000 voxels, Wilcoxon $W = 10^{7.24}, p < 10^{-12}$ and $W = 10^{5.23}, p < 10^{-12}$, respectively). The differences in mean prediction performance in the top 10,000 voxels for the models estimated with the HRF ($r = 0.55 \pm 0.001$), spherical ($r = 0.53 \pm 0.001$) and smoothness ($r = 0.45 \pm 0.001$) priors were small but consistent. However, across the total population of voxels the spherical prior yielded better prediction performance (Wilcoxon $W's > 10^{10}, p's < 10^{-12}$). These results suggest that for well-predicted voxels at least the HRF prior has a small but consistent advantage.

## 1.4 Combining spatial and temporal priors

When both a feature prior and a temporal prior are available, they can be used to construct a single spatiotemporal multivariate normal prior. Spatiotemporal priors allow us to incorporate prior information about the feature weights' covariance and the temporal delays' covariance when estimating predictive encoding models. However, as the number of features ($p$) and temporal delays ($d$) increase, the spatiotemporal prior matrix becomes large ($(p \times d)^2$). This makes the estimation of (non-spherical) spatiotemporal encoding models impractical for neuroimaging. In this section, we present a solution to that makes the estimation of these models tractable when $n < p$.

The spatiotemporal prior is constructed by computing the Kronecker product ($\otimes$) between the feature prior $\Sigma^X \in \mathbb{R}^{p \times p}$ and the temporal prior $\Sigma^T \in \mathbb{R}^{d \times d}$,

$$\Sigma = \Sigma^T \otimes \Sigma^X = \begin{bmatrix} \Sigma^T_{1,1}\Sigma^X & \cdots & \Sigma^T_{1,d}\Sigma^X \\ \vdots & \ddots & \vdots \\ \Sigma^T_{d,1}\Sigma^X & \cdots & \Sigma^T_{d,d}\Sigma^X \end{bmatrix}.$$

The resulting spatiotemporal prior is $\Sigma \in \mathbb{R}^{pd \times pd}$. Notice that when both the feature and the temporal priors are spherical, the spatiotemporal prior is also spherical.

The Tikhonov solution to an encoding model with a spatiotemporal multivariate normal prior $\Sigma^T \otimes \Sigma^X$ can be expressed as (see Appendix 4.3):

$$\hat{\beta}_T = \left(\Sigma^T \otimes \Sigma^X\right) X^\top \left(X \left(\Sigma^T \otimes \Sigma^X\right) X^\top + \lambda^2 I\right)^{-1} Y.$$

This is equivalent to the ridge regression solution when both priors are spherical ($I_d \otimes I_p = I_{pd}$). However, computing this solution involves constructing an extremely large $(pd)^2$ spatiotemporal prior matrix which we would like to avoid. Luckily, the

properties of the Kroenecker product allows to derive a computationally efficient solution in cases where $n < p$ (Appendix 4.4). This formulation makes it tractable to fit large encoding models with non-spherical spatiotemporal priors.

### 1.4.1 Evaluating spatiotemporal MVN priors

To illustrate the power of spatiotemporal priors, we estimated four different encoding models using the data from (Huth et al., 2016). We modeled voxel responses to the stimulus as a linear combination of words, and estimated models that differed only in the spatiotemporal prior used:

$$Y = X\beta + \epsilon$$

$$\beta \sim \mathcal{N}_{pd}\left(0, \lambda^{-2}\Sigma^T \otimes \Sigma^X\right).$$

The first and simplest model we evaluated was ridge regression. Ridge regression corresponds to a spatiotemporal prior where both feature and temporal priors are spherical ($I_d \otimes I_p$). The second model used a word embedding prior $\Sigma^X$ constructed from word co-occurrence statistics estimated from a large text corpus (described above), and a spherical temporal prior ($I_d \otimes \Sigma^X$). The third model was constructed using a spherical feature prior and a HRF temporal prior $\Sigma^T$ constructed from a set of HRF basis functions ($\Sigma^T \otimes I_p$). Finally, the fourth model evaluated used a spatiotemporal prior that combines both the word embedding feature prior and the HRF temporal prior ($\Sigma^T \otimes \Sigma^X$).

The models were constructed using 10 $T_R$ temporal delays (20 seconds) in order to account for the hemodynamic lag. A temporal prior $\Sigma^T \in \mathbb{R}^{10\times10}$ was constructed from the temporal covariance of an HRF basis set during the same time period. The FIR matrix $X$ was built using 10 temporal delays for each of the $3,000$ channels. This resulted in an FIR feature matrix with a total of $30,000$ features and $3,737$ time points.

We find that the model estimated with the semantic-temporal prior performs better than the same model estimated with either the semantic or the temporal prior on their own (Figure 1.4). From a total of about 150,000 voxels, approximately 22,500 were significant ($n = 270, q(FDR) < 0.05$) when using the semantic-temporal prior ($r = 0.045 \pm 0.0003$), approximately 5,500 with the temporal prior ($r = 0.019 \pm 0.0002$), and 15,000 with the semantic prior ($r = 0.037 \pm 0.0002$) . The semantic-temporal prior performs much better than the temporal prior model alone (Wilcoxon $W = 10^{9.67}, p < 10^{-12}$). This is not surprising since the semantic-temporal prior includes the semantic prior and that on its own improves prediction performance

(see Figure 1.2). However, we find that the semantic-temporal prior improves performance over and above the semantic prior alone (Wilcoxon $W = 10^{9.71}, p < 10^{-12}$). In sum, we can gain the best from both worlds by combining feature and temporal priors into a single spatiotemporal prior and thereby improve the prediction performance of encoding models.

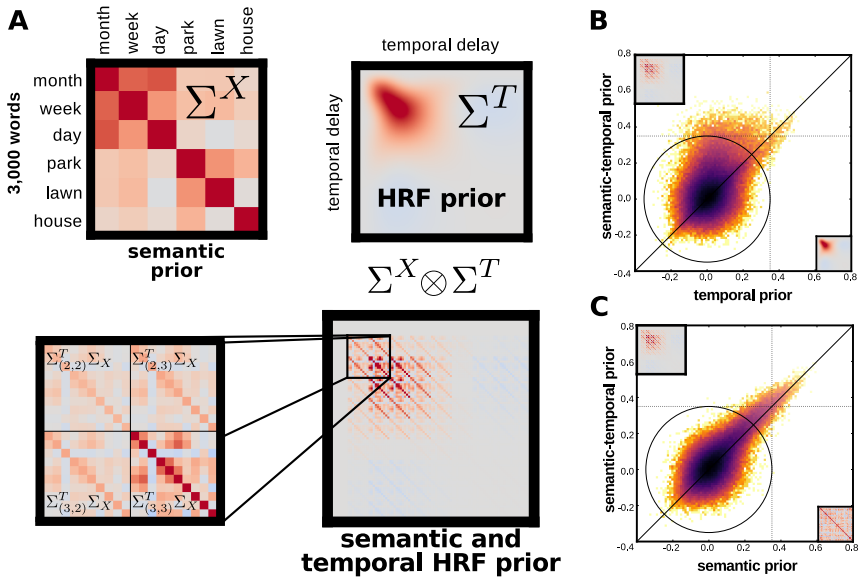


**Figure 1.4: Spatiotemporal multivariate normal priors improve prediction accuracy by combining temporal and feature priors.** We combined semantic and HRF multivariate normal priors. The model is estimated using cross-validation and used to predict BOLD responses to held-out stimuli. The spatiotemporal multivariate prior consistently yields better prediction accuracy than using either prior on its own. **(A)** The spatiotemporal multivariate normal prior is constructed by computing the Kroenecker product of the semantic and temporal priors. This is achieved by scaling each spatial prior by each element in the temporal prior and then concatenating all the resulting matrices. **(B,C)** Prediction accuracy on held-out data improves when using the spatiotemporal multivariate normal prior relative to using either the spatial or the temporal prior alone.

## 1.5 Combining spatiotemporal priors

It is becoming increasingly important to characterize how and where different feature spaces overlap in terms of variance explained (Borcard et al., 1992, Lescroart et al., 2015, de Heer et al., 2017). This is usually done by combining different feature spaces into one single joint model (Lescroart et al., 2015, Çukur et al., 2016, de Heer et al.,

2017). However, estimating joint models with ordinary regularization techniques (e.g. ridge, LASSO, elastic net) assumes that all feature spaces require the same level of regularization. This is often an incorrect assumption. In practice, the level of regularization for a feature space depends on factors such as the feature space covariance, the number of features, and the fraction of variance explained by that feature space. The choice of regularization level for each feature space is critically important to the prediction accuracy of models that combine multiple feature spaces.

Suppose we have two feature spaces $X_1 \in \mathbb{R}^{n \times p}$ and $X_2 \in \mathbb{R}^{n \times q}$ that are combined into a single encoding model:

$$Y = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon$$

Using ridge regression to estimate such a model is equivalent to choosing the same level of regularization on each feature space. The ridge prior on the joint feature weights can be expressed as:

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \sim \mathcal{N}_{p+q} \left( 0, \begin{bmatrix} \lambda^{-2} I & 0 \\ 0 & \lambda^{-2} I \end{bmatrix} \right),$$

where the regularization level $\lambda$ is selected via cross-validation or other methods. It is clear that the prior on each feature space is the same ($\lambda^{-2} I$). However, feature spaces $X_1$ and $X_2$ might need different levels of regularization. Estimating joint models with the same level of regularization on each feature space can lead to poor prediction performance. This is because the globally optimal $\lambda$ will often be suboptimal for the individual feature spaces. This issue applies to ridge (Hoerl and Kennard, 1970), LASSO (Tibshirani, 1996), and elastic-net (Zou and Hastie, 2005) models.

### 1.5.1 Banded ridge regression

Instead, we can impose separate priors on the weights for each feature space:

$$\beta_1 \sim \mathcal{N}_p(0, \lambda_1^{-2} I_p)$$
$$\beta_2 \sim \mathcal{N}_q(0, \lambda_2^{-2} I_q).$$

The Tikhonov framework allows us estimate the joint model with a separate prior on each feature space,

$$Y = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon$$
$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \sim \mathcal{N}_{pq} \left( 0, \Sigma^T \otimes \begin{bmatrix} \lambda_1^{-2} I_p & 0 \\ 0 & \lambda_2^{-2} I_q \end{bmatrix} \right),$$

where $\lambda_1$ and $\lambda_2$ can take different values. For the sake of clarity, assume a spherical temporal prior ($\Sigma^T = I_d$). Estimating this model is equivalent to solving:

$$\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = \underset{\beta_1,\beta_2}{\operatorname{argmin}} \left[\|Y - X_1\beta_1 - X_2\beta_2\|_2^2 + \|\lambda_1\beta_1\|_2^2 + \|\lambda_2\beta_2\|_2^2\right]$$

The solution is:

$$\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = \left( \begin{bmatrix} X_1^\top X_1 & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{bmatrix} + \underbrace{\begin{bmatrix} \lambda_1^2 I_p & 0 \\ 0 & \lambda_2^2 I_q \end{bmatrix}}_{C^\top C} \right)^{-1} \begin{bmatrix} X_1^\top \\ X_2^\top \end{bmatrix} Y$$

Notice that the penalty ($C^\top C$) becomes the ridge penalty when when $\lambda_1 = \lambda_2$. However, when $\lambda_1 \neq \lambda_2$ the structure of the penalty becomes "banded" with the first $p$ values along the diagonal equal to $\lambda_1$ and the next $q$ values equal to $\lambda_2$.

We can also transform the Tikhonov problem into standard form:

$$A = XC^{-1} = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \lambda_1^{-1} I_p & 0 \\ 0 & \lambda_2^{-1} I_q \end{bmatrix} = \begin{bmatrix} \frac{X_1}{\lambda_1} & \frac{X_2}{\lambda_2} \end{bmatrix}.$$

This is a surprisingly simple expression. It says that scaling the features is equivalent to adjusting the strength of the prior. This is due to the inverse relationship between feature scaling and feature weights. All else being equal, dividing the features by a constant is equivalent to multiplying the weights by that constant. Finally, the kernelized standard form solution becomes

$$\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix} \lambda_1^{-2} X_1^\top \\ \lambda_2^{-2} X_2^\top \end{bmatrix} \left( \sum_{i=1}^{m=2} \lambda_i^{-2} X_i X_i^\top + \gamma^2 I \right)^{-1} Y.$$

### 1.5.2 Evaluating banded ridge regression

We evaluated banded ridge regression using data from a natural movie experiment (Huth et al., 2012). We constructed a single encoding model that combined two previously published feature spaces. The first feature space $X_1 \in \mathbb{R}^{3600\times6555}$ captured low-level visual properties from the stimulus (Nishimoto et al., 2011). The second feature space $X_2 \in \mathbb{R}^{3600\times1705}$ captured high-level visual properties consisting of object and action categories (Huth et al., 2012). We modeled voxel responses as a linear combination of these feature spaces:

$$Y = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix} + \epsilon$$

We estimated this joint model using standard ridge regression and banded ridge regression. The only difference between these models is the feature prior used: ridge regression uses a spherical prior ($\Sigma_R^X = \lambda^{-2} I_{p+q}$) whereas banded ridge regression uses a non-spherical prior:

$$\Sigma_T^X = \begin{bmatrix} \lambda_1^{-2} I_p & 0 \\ 0 & \lambda_2^{-2} I_q \end{bmatrix}.$$

Low-level motion-energy features were extracted from the natural movies using a three-dimensional Gabor pyramid (Nishimoto et al., 2011). This yielded a total of $6,555$ features which differed in orientation, spatial and temporal frequency, location, size, and direction of motion. The high-level object and action category features were tagged by hand from each one second segment of the movies and labeled using WordNet synsets (Miller, 1995, Huth et al., 2012). The hyponyms for each synset were inferred from the WordNet graph and also included. This process yielded a total of $1,705$ object and action category features. An FIR model was then constructed by including 10 $T_R$ temporal delays for each feature order to account for the hemodynamic response function. The resulting model consisted of $8,260$ stimulus features times 10 delays ($82,600$ total features) and $3,600$ time points. For simplicity, we used a spherical temporal prior. The feature prior hyperparameters for both ridge ($\lambda$) and banded ridge ($\lambda_1$ and $\lambda_2$) models were selected per voxel via 5-fold cross-validation. The performance of each model was assessed by computing the correlation between model predictions and actual responses using a held-out dataset not used for model estimation.

#### 1.5.2.1 Results

Banded ridge regression provided far better joint model predictions than standard ridge regression (Figure 1.5; Wilcoxon $W = 10^{9.92}, p < 10^{-12}$). Of the approximately 230,000 voxels, about 40,000 were significantly predicted with banded ridge ($q(FDR) < 0.05$, mean pearson $r = 0.06 \pm 0.0003$). In contrast, approximately 20,000 voxels were significantly predicted with ridge regression ($q(FDR) < 0.05$, mean person $r = 0.3 \pm 0.0002$). We used the estimates from the banded ridge joint regression to compute the prediction performance of each feature space on its own. This gives us a separate prediction performance value per voxel for each the motion-energy features and for the object category features. These prediction performance values are plotted on the cortical sheet in Figure 1.5. There is a strong separation in prediction performance between early visual cortex being best predicted by motion-energy features, and higher visual cortex better predicted by object category features.

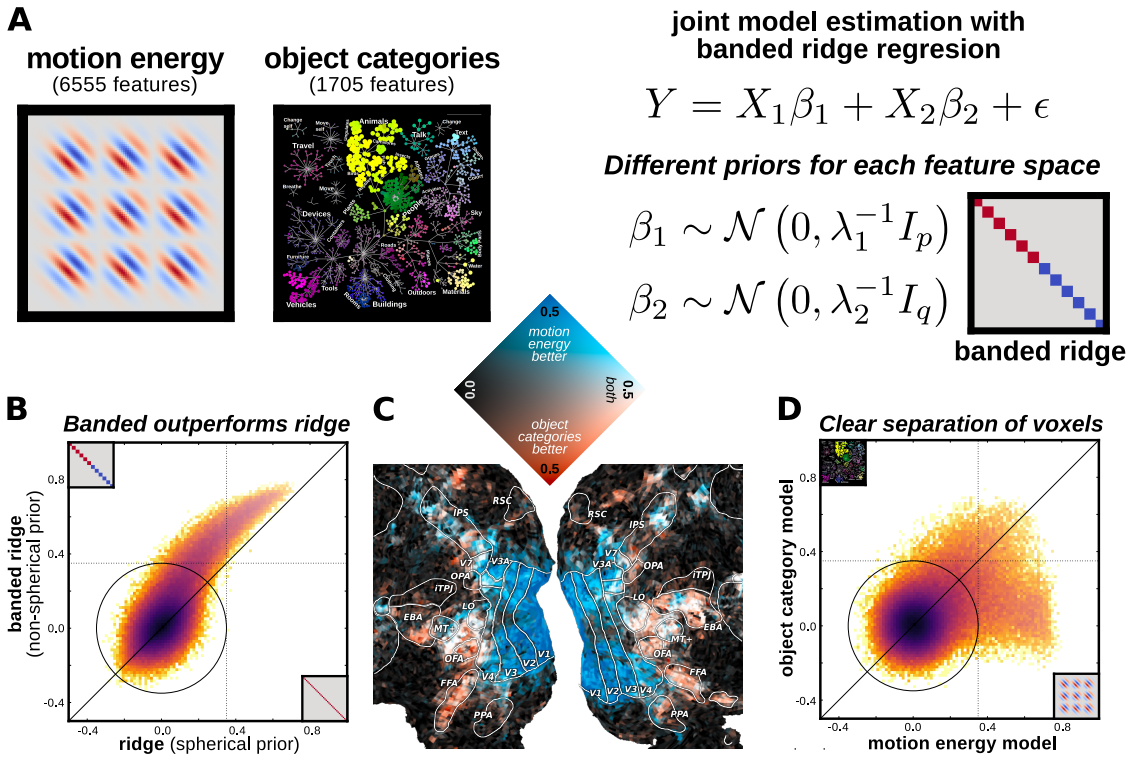


**Figure 1.5: Joint model estimation with banded ridge improves prediction accuracy relative to ridge regression. (A)** In this experiment, we model responses as a function of motion-energy and object category features. We estimate the joint model using independent spherical multivariate normal priors for each feature space, which together constitute a single non-spherical multivariate normal prior. We refer to this method as banded ridge. **(B)** Banded ridge joint model estimation yields much better prediction accuracy on held-out data than ridge regression. **(C)** Once the joint model weights are estimated, we can assess the prediction accuracy of any single model by setting the weights for other models to zero. This gives us the feature space-specific prediction performance after controlling for the other feature space(s). Prediction accuracy is plotted on the cortical sheet. We can see a clear separation between regions in the early visual cortex that are well-predicted by the motion-energy model (blue). Similarly, higher visual cortex is better predicted by the object category features (red). A subset of voxels are predicted similarly well by both models. **(D)** 2D histogram shows a separation in the voxel populations. A large set of voxels are well predicted by motion-energy features, and not object category features. Conversely, many voxels are better predicted by object category features.

A big benefit of banded ridge regression is that it removes spurious correlations between feature spaces. When the motion-energy model is estimated by itself with ridge regression, even voxels in higher visual cortical regions can be well-predicted

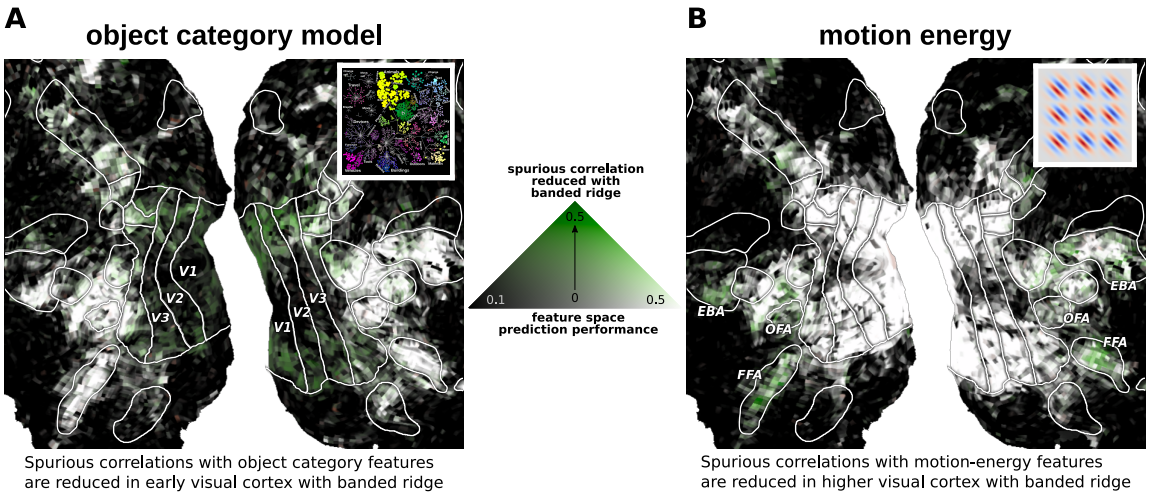


**Figure 1.6: Explaining away feature space-specific variance after banded ridge estimation.** A joint model that included motion-energy and object category features was estimated using banded ridge regression. The two feature spaces effectively compete to predict voxel responses. After estimation, the weights for one model were set to zero and the other model was used to compute the feature space-specific prediction performance. The feature space-specific performance after joint estimation was compared against the performance of the standard single model ridge estimation. The joint banded ridge estimation gives better estimates by allowing feature spaces to "explain away" variance from each other. **(A)** Regions in early visual cortex (green) are predicted with higher accuracy when the model is estimated using ridge relative to banded ridge. The reason these predictions go away is because the variance can be properly allocated to the motion-energy model instead of being captured by the object category model. In other words, the motion-energy features explain away some of the variance from the object category features. **(B)** The object category features explain away some of the variance from the motion-energy features. This can be seen in the regions of anterior visual cortex (green).

(Figure 1.6A). This can occur because of stimulus correlations. For example, suppose there is a consistent correlation between between vehicles and left- and right-direction selective motion-energy filters in the lower visual field. Estimating the motion-energy model on its own will yield high predictions. By estimating the object category and motion-energy models together, the variance can be correctly assigned to the object category model. In cases where a close to perfect correlation exists, the banded ridge estimation will split the variance among the feature spaces. In sum, banded ridge regression yields better estimates of the variance that can be explained by any one feature space.

## 1.6 Discussion

The results highlighted in this chapter show that the Tikhonov framework works very well for estimating predictive models of BOLD responses in the context of naturalistic experiments. The Tikhonov framework can be used to incorporate prior information about how features in the model covary, how the measured signals vary in time, and how multiple feature spaces can be used to build a single preditive model.

The reader should be aware that our results do not necessarily generalize to every experimental condition or dataset. In general, the experimenter should treat the choice of fitting procedure (e.g. FIR, grouped L1, OLS, etc) as a hyperparameter on its own right and use statistical learning theory to make a decision. The Tikhonov framework is presented as another method in the toolkit available to researchers. The banded ridge model proposed is of particular utility when estimating joint models that combine several feature spaces to predictive brain activity.

We have shown a computationally efficient framework for incorporating multivariate normal priors into spatiotemporal encoding models. And that this framework is flexible enough to work well in a variety of cases. The software used to estimate all the models presented in this chapter is publicly available. We hope this facilitates the adoption of this framework.

# Chapter 2

# Representational similarity analysis can lead to incorrect conclusions about representation

## 2.1 Overview

An important goal in functional brain imaging is to determine what type of information is represented within and across brain regions. In recent years, several methods to study brain representations have been developed. A simple and widely used method is representational similarity analysis (RSA). RSA quantifies similarities between brain and model representations and does not require the estimation of a statistical model. However, there exists little work assessing its validity. We show that RSA makes strong assumptions about the relationship between representational models and brain responses. One reason is that RSA does not require a statistical model and can therefore fail to detect significant relationships even when such relationships are present. In addition, RSA can lead researchers to incorrect conclusions when used to compare representational models to brain responses. In contrast, encoding models explicitly estimate the relationship between brain responses and representational models which leads to better performance than RSA.

## 2.2 Introduction

An important goal of functional brain imaging is to identify which types of information different brain regions represent. Computational methods are often used to make inferences about the types of representations that are encoded within and

across brain regions. A commonly used and relatively new method for making inferences about brain representations is representational similarity analysis (RSA; Kriegeskorte et al., 2008b,a). RSA works by estimating a representational similarity matrix (RSM) in a brain region and comparing it to a multitude of candidate RSMs built from computational and/or behavioral representational models. The representational model that is most similar to the RSM of a brain region is chosen as the representational model that best characterizes the information represented in that brain region. However, there is little work to-date that explores the validity of RSA for making inferences about brain representations.

RSA has been widely adopted in part because of its simplicity (Kriegeskorte and Kievit, 2013). RSA does not require estimating a statistical model to relate brain responses to representational models. This is appealing because estimating a statistical model can require significant computational resources and time. With the advent of deep neural networks, the candidate representational models that are available to test are very large and estimating a statistical model that relates them to brain responses is a complex endeavor. By obviating the need of model estimation, RSA presents an advantage over methods that require significant computational resources and time.

In this chapter, we evaluate the validity of RSA for making inferences about brain representations. We show that RSA is problematic for making inferences about representations exactly because it does not require estimating a statistical model to relate brain responses to representational models. We show that RSA can fail to detect a significant relationship between a representational model and brain responses when a relationship exists. We also show that RSA can lead to the wrong answer when used to adjudicate between representational models. This means that RSA can lead researchers to incorrect conclusions about the type of information that is represented in brain regions.

This chapter is organized as follows. We begin by describing RSA, then show simple examples where it can fail. We present simulations to show that RSA can fail to detect a relationship between a representational model and brain responses. We then show in real data that RSA can lead to the wrong conclusion about the type of information encoded in a brain region. Finally, we use simulations to quantify how often and in what cases RSA can give the wrong answer. We compare RSA to encoding models in our analyses.

## 2.3 Description of representational similarity analysis

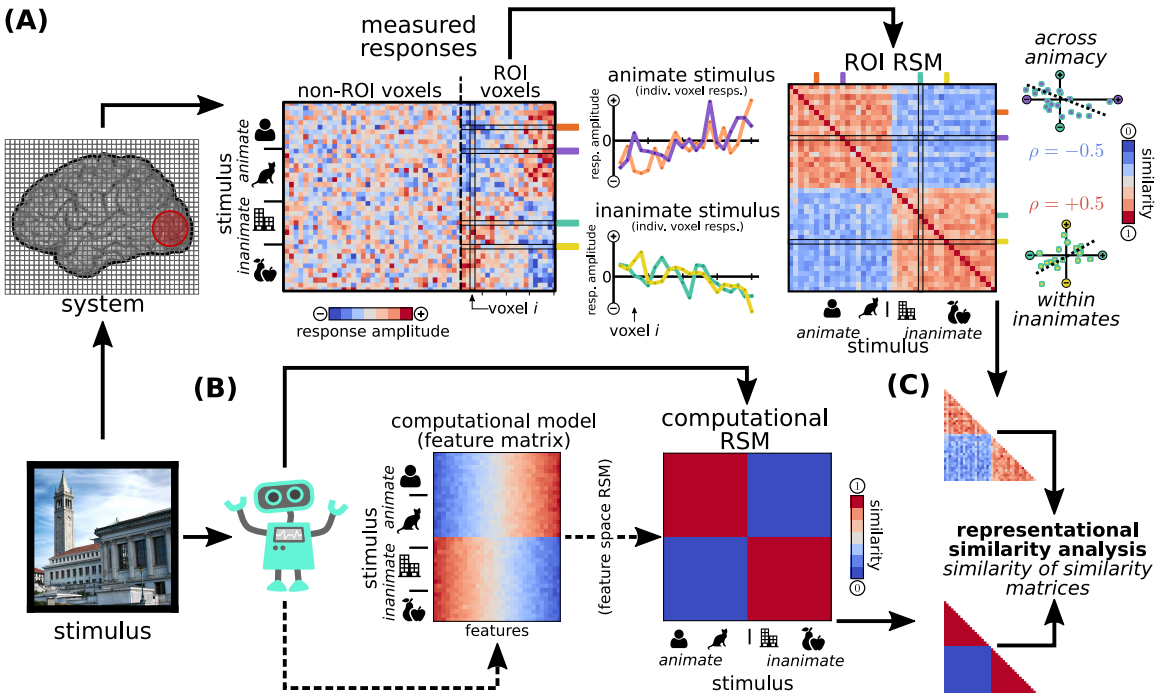


**Figure 2.1: Description of representational similarity analysis.** **(A)** Measured brain responses from a region of interest (ROI) are used to construct a representational similarity matrix (RSM). The activation pattern for two animate stimuli within the ROI are shown (purple and orange). The ROI RSM is constructed by correlating the activation pattern of each stimulus pair. Each entry in the RSM captures the representational similarity between each pair of stimuli in the ROI. In this example, there is a high degree of similarity within animate and within inanimate categories, and a low degree of similarity across animate and inanimate categories. **(B)** A representational model is used to construct an RSM in the same way. The representational model RSM captures the idea that stimuli within a category are represented similarly (red), and that stimuli across categories are represented less similarly (blue). **(C)** The final step of RSA is to compute the similarity between the lower (or upper) triangles of the two RSMs.

### 2.3.1 The representational similarity and dissimilarity matrix

The first step in RSA is to characterize the information represented in a brain region (Figure 2.1). This is achieved by measuring the activation pattern evoked by each stimulus (or task condition) within the region of interest. The activation pattern of every stimulus is then compared with the activation pattern of every other stimulus. This is achieved by computing the correlation between the activation pattern of each pair of stimuli. This results in a symmetric stimulus-by-stimulus representational similarity matrix (RSM).

The next step is to figure out what representations are encoded within that brain region. To achieve this, representational models derived from machine learning (e.g. convolutional neural networks), behavior (e.g. subject ratings), or domain-specific theory (e.g. stimulus animacy) can be used and tested. These representational models are used to extract features (e.g. object category) from which RSMs are built. The resulting RSMs are then compared with the brain RSM.

### 2.3.2 Similarity of representational similarity matrices

RSA computes the similarity between the brain and model representations captured by the RSMs. Because the RSMs are symmetric, only the upper (or lower) triangular entries of the matrices are used to compute their similarity. In general, the similarity is computed as:

$$RSA\left(RSM_X, RSM_Y\right) \equiv similarity\left(triang\left(RSM_X\right), triang\left(RSM_Y\right)\right)$$

where $RSM_Y$ is the brain RSM and $RSM_X$ is the representational model RSM. There are many ways of estimating the similarity between RSMs but the Pearson correlation is a common choice in the literature (Walther et al., 2015). Permutation tests are typically used in order to assess the significance of the similarity between RSMs (Nili et al., 2014).

## 2.4 Simple examples of RSA failures

RSA is in stark contrast to statistical parametric mapping (SPM; Penny et al., 2011) and encoding models (Wu et al., 2006, Naselaris et al., 2011). SPM and encoding models both explicitly estimate a statistical model that relates brain responses and features derived from machine learning, experimental conditions, behavioral, or domain-specific representational models. The set of features (a.k.a. regressors,

predictors) used to build the statistical model defines a feature space. An inference is then made about which of these feature spaces best explains brain responses and how these features are represented in the brain. This process demands a lot of computational resources and time.

RSA is very simple. However, the work to-date that has evaluated the use of RSA to make inferences about representations estimating statistical models has found various issues (Thirion et al., 2015, Cai et al., 2016, Ritchie et al., 2017). In this section, we expand on this literature and present novel simple examples that illustrate how RSA can fail. We highlight that the main issue with RSA is precisely its simplicity. By failing to estimate a statistical model relating brain responses to representational models, RSA makes very strong assumptions about the relationship. The assumptions concern the extent of the representation in the brain, and the importance of the representational model features within the region of interest. These assumptions are not always optimal for inferring brain representations. For this reason, RSA can lead to poor statistical detection power and incorrect conclusions.

## 2.4.1 Assumptions of RSA about extent of the representation in the brain

### 2.4.1.1 RSA can fail when only a sub-region of the ROI is important

RSA can fail to detect a significant relationship between a representational model RSM and a brain RSM when only a sub-region of the ROI encodes the representational model (Figure 2.2). RSA assumes that the extent of the representation in the brain matches the selected ROI. When the ROI is not defined independently per subject (e.g. it is derived from a brain atlas), the mismatch between the actual functional region and the ROI can lead to decreased statistical power. This can occur both when the ROI used is bigger or smaller than the true brain region (Worsley et al., 1996).

### 2.4.1.2 Searchlight RSA can fail when the sphere radius is not optimal

The problem of mismatch between ROI size and brain representation is not ameliorated with searchlight analysis (Kriegeskorte et al., 2006). Searchlight analysis involves moving a sphere of fixed radius across the brain to select voxels and is commonly used in the context of RSA (Nili et al., 2014). Searchlight RSA assumes that the representational model is encoded in the brain as a sphere of fixed radius. It is unclear whether this is optimal for all representations (e.g. cortical regions are not bound within spheres in three dimensions). Moreover, the sphere radius is an im-

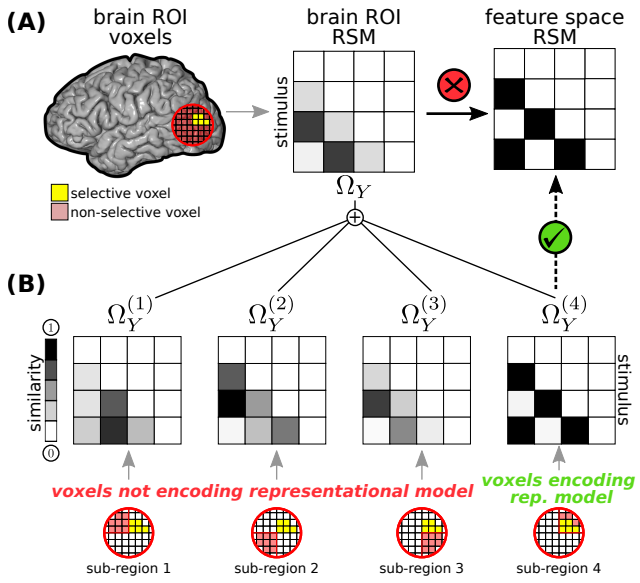


**Figure 2.2: RSA can fail when only a sub-region of the ROI or searchlight is important.** **(A)** The voxels contained within a sphere (red circle) are used to construct a region of interest (ROI) representational similarity matrix (RSM). When there is a mismatch between the ROI used and the functional region, RSA can fail to find a significant relationship with the representational model RSM. (red x-mark). In effect, the unimportant voxels wash out the voxels that encode the representational model. This can occur when the ROI is derived sub-optimally from a common atlas or via a searchlight of fixed radius. **(B)** In this example, RSA can find a significant relationship by dividing the ROI into sub-regions. However, if the ROI were instead too small, RSA would fail to find a relationship because not all the voxels important for the representational model are included. In order to avoid this issue the searchlight radius needs to be estimated per subject per representational model, or by defining ROIs per subject.

portant parameter that is rarely estimated (a radius of 15mm is default; Nili et al., 2014). Assuming an arbitrary sphere radius suffers from issues similar to using a non-optimal filter to spatially smooth brain images (Friston et al., 1993, Worsley et al., 1996).

### 2.4.2 Assumptions of RSA about the importance of model features in the region of interest

The previous issues are inherent to the spatial pooling necessary to construct the brain RSM. They can be ameliorated by defining ROIs per subject (anatomically or functionally), or by independently estimating the radius of the searchlight per subject per representational model. Besides constructing the brain RSM, the other

input to RSA is the representational model RSM. The representational model RSM is constructed without any relation to brain responses. This assumes that the representational model features are all equally important for the brain responses. This leads to a second class of issues.

#### 2.4.2.1 RSA can fail to to detect a significant relationship when model features are not equally important

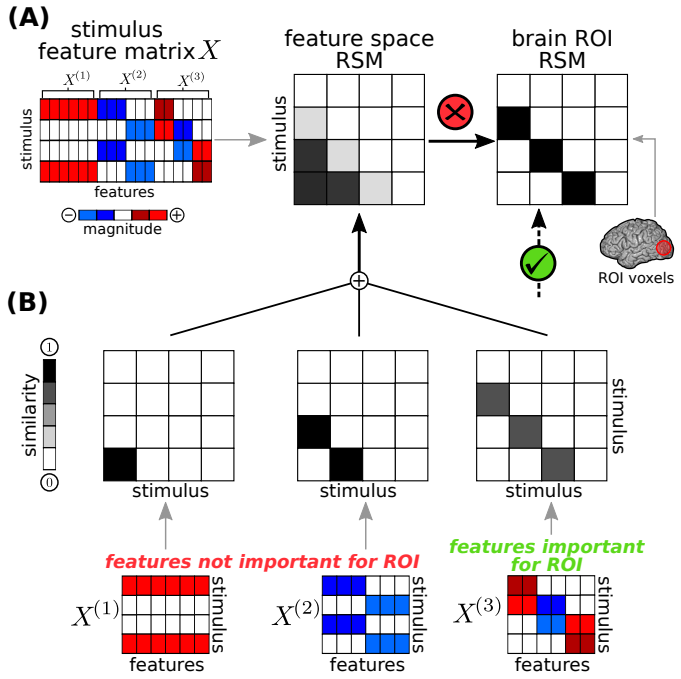


**Figure 2.3: RSA can fail to detect a significant relationship when representational model features are not all equally important**. A simple example where a important representational model features get washed out by unimportant features. **(A)** The representational model is composed of three sets of features and only one subset of these features is important in driving brain activity in the ROI. **(B)** The unimportant features will make the representational model RSM very different from the ROI RSM. This results in a statistical power decrease of RSA. In this case, the relationship between brain responses and the representational model is not significant (red x-mark).

A simple example of this problem is when only a subset of features are important for an ROI. In such cases, the important features can get washed out by the unimportant features when constructing the representational model RSM (Figure 2.3). In our example, RSA will fail to detect a significant relationship between the representational model and brain responses. The lack of feature selectivity with respect to

brain responses is at the heart of the problem. For this particular example, a statistical model that can learn to ignore the unimportant features would be appropriate (e.g. L1-regularized regression, Tibshirani, 1996).

#### 2.4.2.2 Model selection with RSA can fail when true model features are not equally important

RSA is commonly used to compare representational model RSMs with the brain RSM (Kriegeskorte et al., 2008a). When using RSA in this way, the similarity between a representational model RSM and the brain RSM is interpreted as the strength with which the representational model is encoded in the brain region. The representational model that is most similar is inferred to be the one most likely represented within the ROI. In statistical learning, the general procedure for choosing the most likely model from a set of models is called model selection (Friedman et al., 2001).

Because RSA assumes that all the features in a representational model are equally important within a brain region, it can lead to incorrect conclusions about brain representations when comparing representational models. In particular, when the assumptions of RSA are not well-met by the correct representational model an alternative, incorrect representational model might be more similar to the brain RSM. This would lead to the wrong conclusion about what the brain region represents.

An example of how RSA can easily lead to the wrong conclusion about representation is shown in Figure 2.4. The representational model consists of a set of Gabor wavelets and the brain RSM is constructed from left hemisphere V1 responses to natural images. The Gabor model RSM is constructed by correlating all wavelet responses to every pair of images across the full visual field. Only the Gabor wavelets on the right the visual field are important for driving left hemisphere V1 responses. RSA , however, assumes that all the Gabor model features are equally important. This means that the unimportant left visual field Gabor wavelets can potentially wash out the effect of the important right visual field Gabor wavelets. The similarity between the Gabor RSM and the brain RSM will be low. An alternative representational model might by chance be more similar to the brain RSM. Under these conditions, RSA will lead researchers to make the wrong conclusion about V1 representation.

## 2.5 Model assessment with RSA

As shown in Section 2.4.2.1, RSA can fail to detect a statistical relationship when not all features are encoded equally within the ROI. In technical terms, this implies

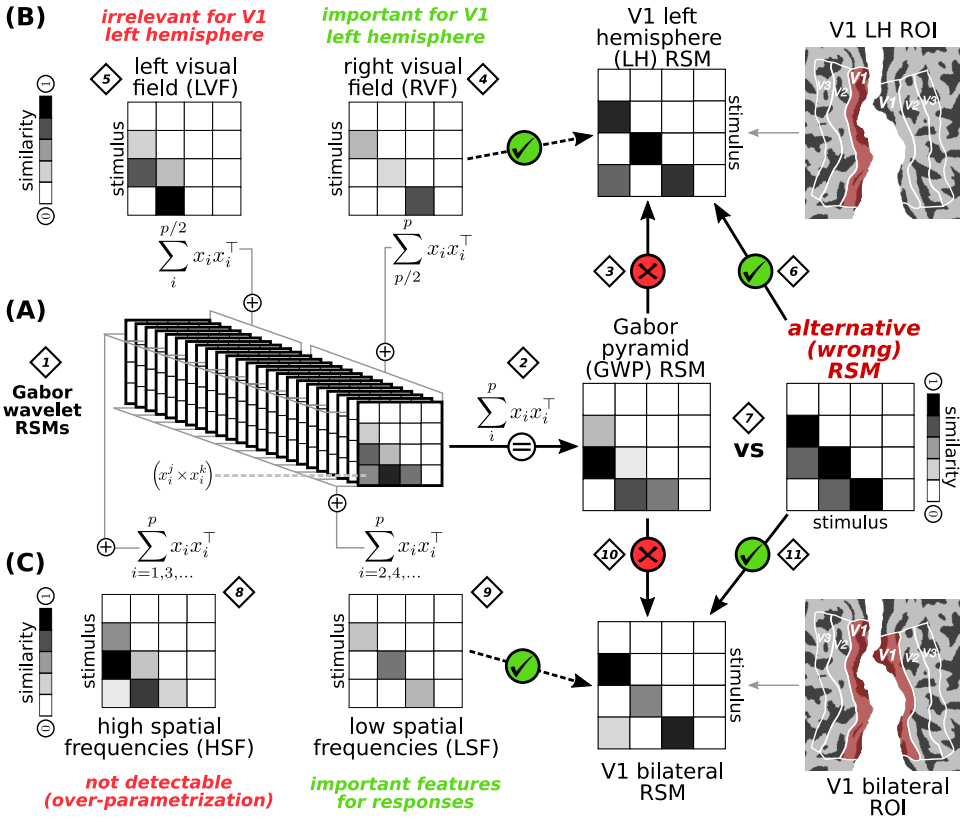


**Figure 2.4: RSA can fail to find the correct representational model when features are not equally important. (A)** A Gabor model is used to compute an RSM (1,2). (3) We compare the Gabor model RSM against the left hemisphere (LH) V1 ROI. The Gabor RSM is not similar to the LH V1 RSM (red x-mark) because the important features for the left hemisphere are washed out by the unimportant features. **(B)** Left hemisphere V1 only processes information from the right visual field (4). If we construct a right visual field Gabor RSM, the resulting RSM is very similar to the LH V1 RSM (green check-mark). The full Gabor RSM is not similar to the LH V1 RSM because the important features are washed out by the unimportant left visual field features (5). (6) This issue becomes especially problematic when using RSA for model comparison. An alternative (incorrect) representational model RSM can be more similar to the LH V1 RSM by chance (green check-mark). (7) When testing whether a Gabor or an alternative representational model RSM better captures the representations of LH V1, RSA chooses the incorrect model. **(C)** (8) The same can occur when the full Gabor RSM is over-parametrized. For example, when very high spatial frequencies that are not detectable at the resolution of fMRI are included. The important low frequency Gabor feautures, (9) are washed out. RSA will fail to detect a significant relationship between the Gabor RSM and the V1 RSM (10). (11) This can lead to incorrect conclusions about representation if the alternative model is by chance similar to the V1 RSM.

that a statistical model estimated to predict brain responses as a linear combination of representational model features will have feature weights that are completely uncorrelated with each other across voxels. In other words, RSA implies that the feature weights are orthogonal for every pair of features across voxels. We perform simulations to evaluate how the orthogonality of feature weights affects the ability of RSA to detect a significant relationship between a representational model and a brain region.

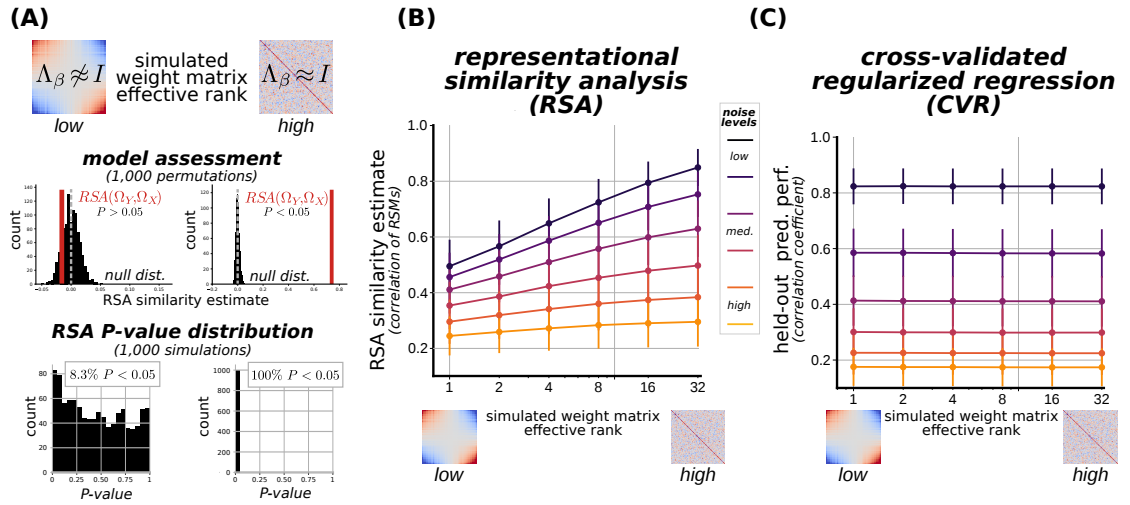


**Figure 2.5: RSA can fail to detect a relationship when some features are more important than others as the similarity between RSMs decreases. (A)** We simulated voxel responses using a linear model where the features were either all equally important or not (low and high effective rank of the feature weights). In both cases, RSA was used to detect whether a significant relationship between the representational model and the brain RSMs exists by shuffling the representational model RSM 1,000 times. We repeated the simulations 1,000 times. RSA fails to find a significant relationship when the weight matrix rank is low (only 8.3% p-values $< 0.05$). **(B)** As some model features become more important than others (i.e. less orthogonal), the RSA similarity decreases. In high SNR conditions (low noise), there is a large difference in the RSA estimates depending on the orthogonality of the feature weights. (Error bars indicate standard deviation). **(C)** Cross-validated ridge regression does not suffer from this issue because the feature importance is captured when the feature weights are estimated. The heldout prediction performance is dominated by noise and not the structure of the feature weights. Note that the y-axes are not comparable across panels since one is held-out prediction performance and the other is the RSA similarity (i.e. correlation of correlations).

### 2.5.1 RSA fails when not all features are equally important

We simulated 2,000 experiments each consisting of 128 voxel responses, 96 stimuli, 100 features, and Gaussian noise ($\sigma = 3$). Brain responses were generated with a linear model ($Y = X\beta + E$). For 1,000 simulations, the generated feature weights were approximately orthogonal ($\beta\beta^\top \approx I_p$). In the other 1,000 simulations, the feature weights were very far from orthogonal and all the units in the population had approximately the same weight vector. After generating the data, we conducted RSA as described in Section 2.3. For each of the 1,000 simulations, we assessed the significance of the relationship with a permutation test by shuffling the RSM matrix 1,000 times. When the feature weights were close to orthogonal, RSA reliably detected the statistical relationship between the representational model and brain region (all 1,000 $p's = 0.001$); Figure 2.5A). This is expected because the RSA assumption is met and all the features are equally important ($\beta\beta^\top \approx I_p$).

However, when this assumption is violated and the feature weights are far from orthogonal RSA fails to find a statistical relationship in 917 of the 1,000 simulations (8.3% $p < 0.05$). RSA fails because not all the representational model features are equally useful in driving activity in the region of interest. We repeated the experiment using a cross-validated ridge regression (CVR) model to estimate the relationship between the representational model and each voxel in the region of interest. Using this approach, we were indeed able to reliably identify a significant relationship between the representational model and the brain responses in all 2,000 simulations (all $p's < 0.05$, not shown).

### 2.5.2 RSA similarity decreases when some features are more important than others

We next evaluated how RSA and CVR models are affected as the feature weights vary from non-orthogonal to orthogonal. This was achieved by generating feature weights per voxel from covariance matrices with varying levels of effective rank (1, 2, 4, 8, 16, 32; Pedregosa et al., 2011). We manipulated the number of stimuli (100, 1,000), features (100, 1,000), voxels (128, 256, 512), noise levels (1, 2, 3, 4, 5, 6; i.i.d Gaussian s.d.), and feature matrix effective rank (1, 5, 10, 20). The voxel responses to the stimuli were generated using a linear model. This resulted in a total of 4,350 simulations for each of the six noise levels.

RSA similarity is affected by both how orthogonal the feature weights are and the noise level (Figure 2.5B). As some features become more and more important than others (i.e. feature weights become less orthogonal), the similarity between brain and representational model RSMs decreases and so does the ability of RSA

to detect a significant relationship. Note that a high SNR dataset (low noise level) will produce very different RSA similarity values depending on how orthogonal the feature weights are (varying from 0.5 to 0.9).

In contrast, cross-validated ridge regression prediction performance depends little on the orthogonality of the feature weights. CVR prediction performance is mainly affected by the amount of noise in the data, not whether the features are all equally important (Figure 2.5C). This is because CVR explicitly estimates a statistical model that relates brain responses to representational model features. The estimated feature weights capture the relative importance of the representational model features explicitly.

The ability of RSA to detect significant relationships depends on the orthogonality of the feature weights. This affects the likelihood of detecting a relationship between a representational model and a brain region. These results are not in and of themselves a reason for much concern since different methods can have varying levels of statistical power under different conditions. There might even be situations where RSA might have higher statistical power relative to regression models. It is the use of RSA for model selection, however, that is a major concern.

## 2.6 Model selection with RSA

RSA is commonly used to compare representational models and decide which one better captures brain representations. However, if the assumptions of RSA are better met for one representational model than for the other, the conclusion can be exactly wrong. We demonstrate this using real and simulated data.

### 2.6.1 RSA fails to choose a Gabor model as the representational model for V1

We used functional MRI data from a vision experiment to evaluate the use of RSA for model selection. We used RSA to test whether V1 representations are better captured by a Gabor model computed on (i) natural images, or (ii) object silhouette segmentations (red and blue, respectively; Figure 2.6). A wealth of evidence has shown that Gabor wavelets computed on natural images are a good model of V1 in neurophysiology (Daugman, 1984) and fMRI (Kay et al., 2008b). While a Gabor model is not the "ground-truth" representational model for V1, it is a good approximation (Carandini et al., 2005). There is strong a priori expectation that a Gabor model computed on natural images should capture V1 representations more accurately than a object silhouette representational space.

A total of 1,260 natural images were shown to two subjects while BOLD responses were recorded with fMRI (see Stansbury et al., 2013, for details). The hemodyamic response function and the response to each stimulus was estimated for each voxel separately using generalized least squares (Stansbury et al., 2013). The silhouette of each object in each image was drawn by hand and the resulting segmented image was binarized. These silhouette images were used to extract object silhouette features. We extracted luminance images by converting the original RGB images to the CIE L*a*b* color space (McLaren, 1976). The luminance images were used to extract image Gabor features. We used two Gabor wavelet pyramids to extract features for each of the (i) image Gabor and (ii) object silhouette representational models. One pyramid was small and the other was large. This yielded a total of four feature matrices: (i) small and large Gabor feature matrices, and (ii) small and large object silhouette feature matrices. Each of these feature matrices was used separately to compute an RSM.

The small Gabor wavelet pyramid contained spatial frequency filters at 0, 2, 4, 8, 16 and 32 cycles per image. This yielded a total of 570 features per stimulus. The large Gabor wavelet pyramid was constructed with same spatial frequency filters as the small pyramid and an additional set of high spatial frequency filters at 64 and 96 cycles per image. The large pyramid yielded 6,302 features per stimulus. At the resolution of fMRI, the high spatial frequencies are not very useful in explaining additional variance in the V1 BOLD responses. The large version of the representational models can be thought of as an over-parametrization. This violates the assumption that all the representational model features matter equally within V1. For the large pyramid, the majority of the Gabor model features are unimportant and may wash out the important features (Figure 2.6A).

We tested whether V1 representations are more similar to image Gabor RSMs or object silhouette RSMs using RSA (Kriegeskorte et al., 2008a,b). We bootstraped the difference in similarity to the V1 RSM 1,000 times and computed p-values from this distribution. We performed RSA with both models using the coefficients of a regression model estimated between the brain RSM and the representational model RSMs (Nili et al., 2014). We also estimated an encoding model for each voxel separately using a training set of 1,260 images with cross-validated ridge regression. We measured prediction performance by computing the correlation coefficient between predicted and actual voxel responses to 126 held-out images. We bootstrapped the difference in prediction performance 1,000 times. We also evaluated the effect that using a different number of stimuli has on both of these analyses.

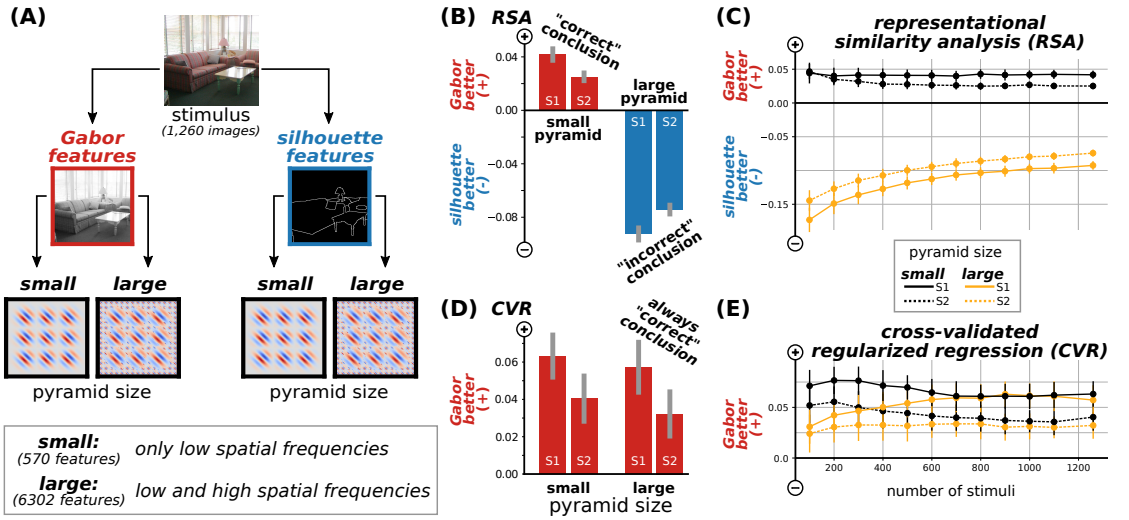


**Figure 2.6: RSA can fail to choose a Gabor model as the representational model for V1**. We used fMRI data collected from two subjects while they viewed 1,260 natural images to test representational models of V1 (Stansbury et al., 2013). **(A)** Two representational models were tested. One is the Gabor model (red) and the other the object silhouette model (blue). We constructed two versions of each representational model one small and one large. **(B)** RSA was used to compare the representational models to the bilateral V1 RSM. When using the large version of the representational models, RSA finds that object silhouette RSM captures V1 representations better than the Gabor RSM. This likely incorrect result has been reported before with RSA (Kriegeskorte et al., 2008a). When computing the RSM for the large Gabor representational model, the unimportant (high spatial frequency) Gabor wavelets wash out the contribution of important (low spatial frequency) Gabor wavelets that dominate the measured V1 responses at the resolution of fMRI. **(C)** These analyses were conducted by subsampling the number of stimuli used. RSA consistently leads to the incorrect conclusion for the large representational models. **(D,E)** Cross-validated ridge regression (CVR) consistently gives the correct answer.

#### 2.6.1.1 Results

RSA yields the expected result when we compare the Gabor and silhouette representational models built from the small pyramid (570 features; Figure 2.6B, red). We see that as the number of stimuli increases, the RSA comparison remains stable. However, when we compare the representational models using the RSMs built with the large pyramid (6,302 features), RSA gives the opposite answer. V1 representations are better captured by the object silhouette representational model (Figure 2.6C, orange; Kriegeskorte et al., 2008a). We cannot say that this is the incorrect conclusion because we do not have access to the true model. However, it certainly goes against expectations and suggests that RSA does not handle noisy features or high dimensional feature spaces well. RSA can give different answers for different

parameterizations of the same representational space because it assumes that all of the representational model features matter equally in the region of interest. From the outset, the representational model must be constructed in such a way that it already closely matches the brain representations without estimating a statistical model.

In contrast, encoding models estimated with cross-validated ridge regression give consistent results for each subject and feature space size (Figure 2.6C). As the number of stimuli increases the difference in prediction performance between the silhouette and Gabor representational models increases in favor of the Gabor model. The prediction performance is lower when using the large version relative to the small version of the representational models. This is expected because increasing the number of features requires more data to estimate the statistical model. This is important especially if many of the features are not useful in driving brain activity. The difference estimate of the regression model remains positive in all comparisons with the Gabor features always better than the object sihouette features.

### 2.6.2 RSA has lower statistical power than regression for model selection

In the previous experiment, we did not have access to ground-truth representational model for brain responses, nor how it relates to measured BOLD responses. Thus, we cannot conclude that RSA lead us to the incorrect conclusion. To determine the conditions under which RSA can give the wrong answer we performed a series of simulations where the ground-truth representational model is known (Figure 2.7A).

We simulated voxel responses to stimuli as a linear combination of ground-truth features plus noise ($Y = X\beta + E$). We then sampled stimulus features similar either to the ground-truth representational model or to the voxel stimulus-by-stimulus response covariance. The stimulus features were then used to construct one RSM for the ground-truth representational model and another RSM for the alternative representational model. RSA was then used to select the model that best captures the representations of the simulated brain responses.

The simulated data varied in the number of stimuli (100, 300), features (100, 1,000), voxels (128, 256, 512), feature weight effective rank (1, 3, 5, 7, ..., 32), similarity between the sampled candidate features $X$ and the ground-truth representational model ($10^{-3}$ to 1; 14 log-spaced samples), and the similarity between the "alternative" features $Z$ and the empirical voxel responses ($10^{-5}$ to 1; 10 log-spaced samples). A total of 25,000 simulations for each of six noise levels were performed (i.i.d. Gaussian with 1, 2, 3, 4, 5, or 6 s.d.). The stimulated representational model features were used to construct a candidate (i.e. correct) RSM and an alternative

(i.e. incorrect) RSM. RSA was then used to select the representational model RSM that best captures the simulated brain RSM. Significance was assessed via bootstrap (1,000 samples with replacement per simulation).

For each simulation we used RSA to test whether the candidate representational model captured the simulated brain responses better than the alternative representational model (Figure 2.7A). The p-value of the difference between the representational models was computed by bootstrapping the difference estimate 1,000 times. We quantified the statistical power of RSA by counting the number of times the candidate (i.e. correct) representational model $X$ was found to be better than the alternative (i.e. incorrect) representational model $Z$ at every significance threshold.

We also used cross-validated ridge regression to estimate a statistical model relating simulated voxel responses and the feature spaces generated from the candidate and alternative representational models. The mean prediction performance across voxels was computed. The significance of the difference between representational models was computed by bootstrapping the heldout prediction performance difference 1,000 times.

#### 2.6.2.1 Results

The RSA similarity between the brain RSM and the incorrect representational space RSM can be higher than with the correct representational space (Figure 2.7B; example with 300 stimuli, 1,000 features, and i.i.d. Gaussian noise with 2 s.d.). This occurs when not all the representational model features are from the correct model are equally important and the incorrect representational model is similar to the brain RSM. As the correct representational model features become less and less important RSA tends to give the wrong answer.

We estimate the statistical power of RSA by counting the number of times it gives the correct answer (Figure 2.7C). At the typical p-value threshold of 0.05, the ability of RSA to select the correct representational model quickly decreases. RSA is worse than cross-validated regularized ridge regression for selecting the correct representational space across all noise levels and significance thresholds explored in our simulations.

The reason for the low statistical power of RSA is illustrated in Figure 2.7D. The RSA similarity between the brain RSM and the correct representational model RSM decreases when the representational model features are not all equally important. This means that feature weights are not close to orthogonal ($\Sigma_\beta \neq I_p$) and so the similarity between the correct representational model RSM and the brain responses will be low. The similarity between an incorrect representational model RSM and the brain RSM might be higher than with the correct representational model by

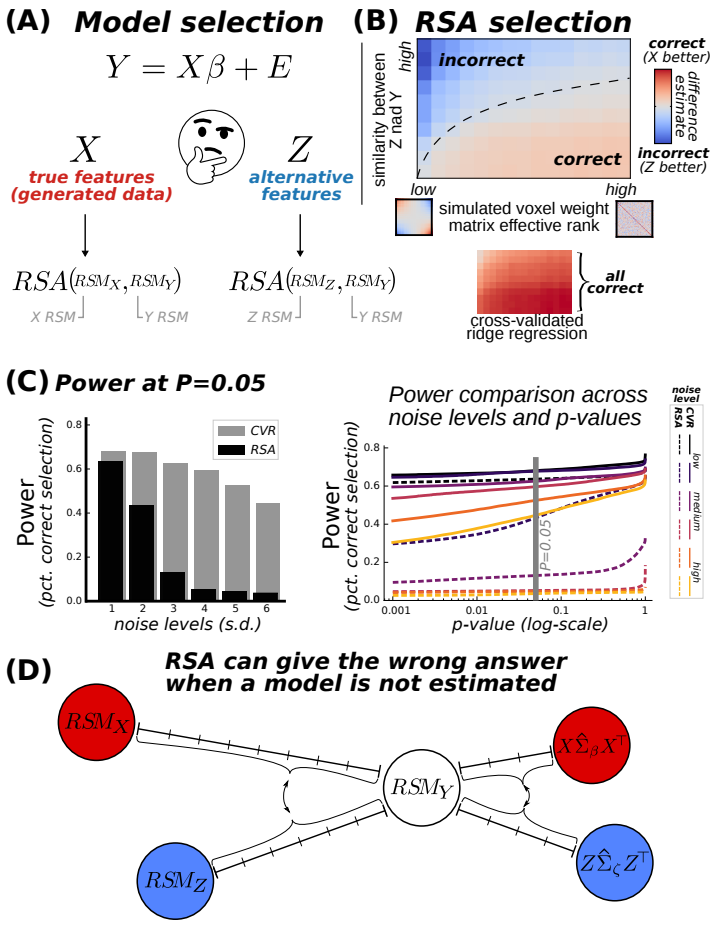


**Figure 2.7: RSA has low statistical power to adjudicate between representational models. (A)** We simulated brain responses to stimuli derived from a known feature space. **(B)** RSA was used to adjudicate between correct and incorrect representational models. RSA gives the incorrect answer as the features of the correct representational model become less equally important (horizontal axis) and the incorrect representational model becomes more similar to the brain RSM (vertical axis). **(C)** This occurs across many simulations varying in the number of stimuli, features, voxels, and noise levels (150,000 simulations). At the typical statistical significance threshold of $p < 0.05$, the ability of RSA to find the correct answer quickly decreases with noise. RSA has consistently less statistical power to find the correct representational model than cross-validated ridge regression across significance thresholds and noise levels. **(D)** Estimating a statistical model to relate brain responses to representational model features before conducting representational similarity analysis can provide more reliable answers if the region of interest is properly chosen. However, there is no evidence that the similarity of similarity matrices is a reliable statistic for inferring representations (Thirion et al., 2015), nor that it has any advantages over classical statistics like the coefficient of determination ($R^2$).

chance. Whenever this conditions occur, RSA will lead researchers to incorrect conclusions about representations. This problem can be ameliorated by estimating a statistical model to relate the correct representational model features to the brain responses directly. This will give an estimate of the importance of the feature weights ($\hat{\beta}$). The estimated feature weights can then be used to compute a representational model RSM that does not assume that all the features are equally important ($\hat{\Sigma}_\beta$). However, it is unclear whether there are any advantages to using this neuroimaging specific approach versus standard statistical learning (Khaligh-Razavi et al., 2017).

## 2.7 Discussion

RSA is a very simple method that requires few computational resources and time. However, the wide availability and low cost of computational resources has made it increasingly easier, faster and cheaper to estimate statistical models that explicitly relate representational models and brain responses. These technological advancements make the main advantages of RSA less beneficial than they were 10 years ago.

The computational simplicity of RSA comes at the cost of making strong assumptions. Assumptions about the extent of the representation within the brain, and about the how the representational model features are represented within a brain region. We have shown that making use of not well-defined ROIs or searchlights of arbitrary radius puts strong constraints on the extent of the representation in the brain and therefore decreases statistical power.

We have also shown that whenever representational model features are not equally important in a brain region, RSA can fail to detect a significant relationship between the representational model and the brain responses. This can lead researchers to make incorrect conclusions about representation when adjudicating between representational models. Empirically, the assumption that representational model features are all equally important within brain regions is not in agreement with the literature (e.g. Huth et al., 2012, 2016).

### 2.7.1 RSA computed on encoding models

Mixed-RSA is a step in the right direction within the RSA literature (Khaligh-Razavi et al., 2017). It relies on estimating a statistical model to relate brain responses to the representational model. This is nothing more than using representational similarity analysis on a standard encoding model (Wu et al., 2006). The main benefit of this approach is spatial pooling over voxels. When a searchlight is used, this

achieves spatial pooling with a sphere instead of Gaussian blurring as is usual in fMRI. However, spatial pooling comes with its own set of issues when performed sub-optimally (Section 2.4.1).

There are cases where standard RSA can find a significant relationship between RSMs, but RSA based on an encoding model cannot. This is interpreted as providing evidence that the encoding models are difficult to train and therefore difficult to rely on (Khaligh-Razavi et al., 2017). This is a misunderstanding and such an outcome does not mean that an encoding model approach is unreliable. It means that the chosen encoding model is suboptimal. In this particular case, it suggests that the stimulus triggered average (STA) is a better encoding model than the chosen encoding model (Appendix 5.1). Assessing and choosing the optimal encoding model from the classes of statistical models that exist (e.g. STA, OLS, ridge, LASSO, GLS, etc.) is a non-trivial task.

Leaving aside all this issues, there is a best case scenario for RSA: when it is conducted on an encoding model appropriately chosen from the class of available statistical models and with an appropriately defined ROI. Even in this case, as far as we know, there is no evidence to suggest that there is any advantage to making inferences about representations with RSA versus standard statistics (e.g. mean $R^2$ over the voxel population). In fact, recent work has shown that when the assumptions of RSA are met, there is little benefit relative to a particular class of encoding models based on ridge regression (Diedrichsen and Kriegeskorte, 2017). Thus, it remains unclear whether RSA as a neuroimaging specific technique adds any value beyond spatial pooling.

### 2.7.2 Encoding models provide a direct answer to the first order question

Encoding models (and SPM) explicitly estimate the relationship between representational model features and brain responses. In the voxel-wise modeling approach, the brain representations are inferred directly from the estimated feature weights (i.e. tunning). This can be performed per voxel, within a region of interest, or across the cortical sheet. Researchers can directly evaluate which features from the representational model are represented in each voxel or construct a representational space from the voxels within a region of interest. The voxel-wise encoding model paradigm is a powerful technique that avoids strong assumptions about the extent of representation, or about the way in which representational model features are encoded within regions. Voxel-wise modeling can also be used to assess the how well the estimated model is able to generalize to novel stimuli not used for model estimation.

Furthermore, voxel-wise encoding models explicitly state the assumptions made. When using regularized regression, for example, many different priors can be used. Tikhonov regression allows researchers to formulate complex priors that might help in constructing predictive voxel-wise models (Tikhonov et al., 1977). These priors can be compared using standard statistical techniques or Bayesian approaches. One limitation is that voxel-wise encoding models require much more data than is common for a typical cognitive neuroscience experiment. However, large high quality datasets are worth the cost. We hope our work shows that making inferences about representational models with RSA should be taken with caution.

# Chapter 3

# Discovering brain representations across multiple feature spaces using brain activity recorded during naturalistic viewing of short films

## 3.1 Overview

We present a rich paradigm and a novel computational model for efficient non-invasive functional brain mapping. In our paradigm, subjects watch interesting short films while their brain activity is measured. Multiple feature spaces are used to model the brain responses to the short films. Each feature space constitutes a hypothesis about the type of representations that might be important for brain regions involved in watching, listening, and understanding the short films. A computational model is then used to find the most predictive feature spaces across the cortical surface and also to recover maps that capture how the individual feature spaces are represented within cortical regions. Our results suggest a high degree of homogeneous selectivity for feature spaces across large regions of the cortical surface within individual subjects. These patterns are highly consistent across all subjects. We are also able to recover known retinotopic, tonotopic and semantic functional maps from this single experiment. Finally, we explore the functional organization of the middle temporal cortex and show that the visual feature spaces can capture novel functional subdivisions in this region.

## 3.2 Introduction

Mapping cognitive function to brain areas is a long standing goal of cognitive neuroscience. Studies in patients and animals were the first to observe that specific cognitive functions are impaired when specific brain areas are damaged (Finger, 2001). The earliest observations lead to the discovery that vision was localized to the occipital cortex (Munk, 1881, Anonymous, 1883) Another early observation was that damage to a region in ventro-lateral prefrontal cortex caused difficulty in producing speech (Broca, 1861a,b). By documenting the cognitive deficits that were caused by damage to specific areas of the brain, researchers began to build a map of the brain wherein cognitive functions are localized to specific brain areas.

Since the advent of non-invasive human brain imaging techniques, work in cognitive neuroscience has continued this functional localization trajectory. One approach to map specific cognitive functions to brain regions is the use of subtraction tasks derived from psychology. In this approach, the difference in brain activity between conditions is used to infer where in the brain a cognitive function is localized. This approach is widely used and has lead to the discovery of several functionally specialized areas (e.g. Kanwisher et al., 1997, Epstein and Kanwisher, 1998). However, this approach is limited. In order to built a functional map of the brain an intractable number of subtraction tasks need to be conducted.

Another approach to map cognitive functions to brain areas is the use of computational models with naturalistic stimuli and tasks (Wu et al., 2006). In this approach, complex stimuli are presented to subjects and computational models are built to learn the relationship between the stimulus and the measured brain signals. This approach has proven successful in characterizing the information that is represented in many brain areas (Kay et al., 2008b, Nishimoto et al., 2011, Stansbury et al., 2013, Huth et al., 2016). Estimating computational models that relate brain activity to naturalistic stimuli requires a large amount of data. Hence, in order to broadly sample the relevant stimulus space (e.g. natural images), naturalistic experiments typically explore only one cognitive domain. This limits the types of representations that can be explored within one single experiment. Furthermore, building and estimating these computational models can be challenging.

We present a naturalistic paradigm that can be used to explore a variety of brain representations simultaneously. In this paradigm, subjects naturally watch interesting short films while their brain activity is measured with functional magnetic resonance imaging (fMRI). The short films contain speech, video, music, environmental sounds, emotions, human interaction, narrative structure and many other components. The short films are labeled with more than a dozen high-dimensional feature spaces. Each feature space constitutes a hypothesis about the type of infor-

mation that might be important for brain regions involved in watching, listening, and understanding the short films (e.g. semantic content present in speech). We then use a new framework to overcome the difficulties related to building and estimating computational models for these complex stimuli. In this framework, brain activity is modeled as a function of all features across all feature spaces simultaneously. The resulting voxelwise encoding model reveals which specific feature spaces are important for every region of the cortical surface and the tuning properties of single voxels in these regions.

In this chapter, we begin by describing the experimental paradigm. We enumerate the feature spaces used to model the measured blood oxygen level dependent (BOLD) responses to the short films. We describe a novel joint voxelwise encoding model developed to analyze these complex stimuli. Finally, our results show that we can recover rich maps of functional selectivity across the cortical surface from one simple task: watching entertaining short films.

## 3.3 Methods

### 3.3.1 Experimental design

#### 3.3.1.1 Short film stimuli

A set of 590 short films were downloaded from various online sources (Vimeo.com, YouTube.com, and shortoftheweek.com). A total of 247 films 3 to 8 minutes in duration were screened by one rater (author AN) on a 1-5 scale reflecting how interesting and engaging each short film was. A total of thirty short films were selected (ratings 4 or 5). Each short film was approximately 5 minutes in duration after editing (4:45±1:08 s.d., min. 2:41, max. 6:59).

Short films were edited to remove credits and title text using video editing software (Kdenlive; open source software maintained by Bushuev et al.). The sound level of each short film was normalized to the approximate average sound level of all the short films. A single MP4 audio-visual file was created for each imaging run (Richardson, 2004). Each MP4 contained two edited short films with a minimum gap of 12 seconds (mean 12.5) between them (24Hz 16:9 1024x576 HD video, 48kHz audio). The resulting fifteen MP4 files were approximately 10 minutes in duration (10:06±55 seconds s.d., min. 8:14, max. 11:46).

#### 3.3.1.2 Stimulus presentation

Short films were presented via customized versions of wxPython (Rappin and Dunn, 2006) and MPlayer (MPlayer Team, www.mplayerhq.hu). The short films were projected onto a tangent screen inside the bore of the magnet (Avotec, Inc., Stuart, FL). The projected display had a size of 34x25 degrees of visual angle (1024x768 pixel resolution). Each short film had a size of 34x20 degrees within the display. The top and bottom 2.5 degrees of the display was filled with grey. For two subjects (SP and JG), the stimulus was presented at 30% of the full field of view corresponding to approximately 10x6 degrees of visual angle [[(see Supplemental Materials [eye movement artifacts])]]. The audio was presented via MRI compatible headphones (Sensimetrics Corporation, Malden, MA). Sound volume was adjusted at the beginning of very scanning session per subject.

#### 3.3.1.3 Scanning procedure

Subjects were instructed to watch the short films as they would normally. Eye movements were not constrained but the eyes were tracked continuously during free viewing (see Section 3.3.1.5 for details). Subjects wore custom-made head restraints to minimize head movement and to improve cross-session alignment (Caseforge, Berkeley, CA). Respiration rate and heart pulse rate were measured continuously (Avotec, Inc., Stuart, FL).

A total of thirty short films were presented to the subjects across three scanning sessions. In each session, there were eight imaging runs and each run contained two short films. Four of the imaging runs in each session contained the short films used to estimate the voxelwise encoding model (24 short films; 8 short films per session x 3 sessions). The additional four imaging runs in each session were used to present four repetitions of the same two short films and these were used to test the estimated voxelwise model (6 short films; 2 short films x 3 sessions x 4 repetitions). Only 3 repetitions of the test stimuli were collected for one subject (SS).

The average imaging run lasted 11.5 minutes (11:28±55 seconds s.d., min. 9:36, max. 13:08). Data were acquired across three separate imaging sessions per subject and each session lasted approximately 2.5hrs.

#### 3.3.1.4 Functional MRI acquisition and preprocessing

MRI data were collected on a 3T Siemens TIM Trio scanner at the UC Berkeley Brain Imaging Center using a 32-channel Siemens volume coil. Functional scans were collected using gradient echo EPI with repetition time (TR) = 2000ms, echo time (TE) = 31ms, flip angle = 70 degrees, voxel size = 2.24 x 2.24 x 4.13 mm (slice

thickness = 3.5 mm with 18% slice gap), matrix size = 100 x 100, and field of view = 224 x 224 mm. 30 axial slices were prescribed to cover the entire cortex and were scanned in interleaved order. A custom-modified bipolar water excitation radiofrequency (RF) pulse was used to avoid signal from fat. The field was shimmed and a fieldmap was acquired between imaging runs. Anatomical data were collected using a T1-weighted multi-echo MP-RAGE sequence on the same 3T scanner. Cortical surfaces meshes were generated from the T1-weighted anatomical scans using Freesurfer software (Dale et al., 1999). Cortical flatmaps were generated using pycortex (Gao et al., 2015)

Each functional run was motion-corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0 (Jenkinson et al., 2012). All volumes in the run were then averaged across time to obtain a high quality template volume. FLIRT was also used to automatically align the template volume of each run to the overall template, which was chosen to be the temporal average of the first functional run for each subject. These automatic alignments were manually checked and adjusted as necessary to improve accuracy. B0 inhomogeneities in the magnetic field were corrected using FSL 5.0 FUGUE. The cross-run transformation matrix, the motion-correction transformation matrices obtained using MCFLIRT, and the B0 inhomogeneity corrections were combined into a single warp field. The original data was resampled directly into the overall template space using trilinear interpolation.

Low-frequency voxel response drift was identified using a $2^{nd}$ order Savitsky-Golay filter with a 120-second window. This drift was subtracted from the signal. Voxel time courses were z-scored separately for each run, i.e., the mean response for each voxel was subtracted and the remaining response was scaled to have unit variance.

RETROICOR (Glover et al., 2000) was used to extract physiological nuisance parameters from the pulse-oximeter and respiration belt signals (Biopac Systems, Inc., Goleta, CA). An ordinary least squares model was estimated and used to predict BOLD responses from a linear combination of physiological nuisance regressors. We then subtracted the signals predicted by the physiological regressors. The resulting residuals were then used to conduct all analyses.

#### 3.3.1.5 Eye tracking

We recorded subject eye movements using an infrared video camera while subjects watched the short films (Avotec, Inc., Stuart, FL). These videos of the eye were used to estimate the subjects point of fixation in each video frame of the short films. Eye videos were recorded at a resolution of 320x240 pixels and a minimum of 30 Hz. A TTL pulse was received by the video recording software and used to temporally align the eye video with the time of the first scan in each run.

Using custom software, we first located the pupil position on every frame of the eye videos. For every eye video (8 runs x 3 sessions x 5 subjects), we manually drew a rectangular window around the eye. We then used the Hough transform to estimate a circle corresponding to the pupil (Duda and Hart, 1972, Itseez, 2015). The Hough transform hyperparameters were manually adjusted per video in order to account for differences in lighting conditions (i.e. infrared light level). The extracted circle center coordinates and radius were then median filtered with a window of 500 ms. The signals were then resampled to match the short films frame rate (24 Hz) using polynomial interpolation.

We then estimated the point of fixation on the screen from the estimated pupil location. At the beginning of each imaging run, 35 calibration points spanning a Cartesian grid were presented one at a time in random order for 2000 ms each. The calibration points were presented in bright green and were overlayed on a series of 5-10 second naturalistic audiovisual clips of varying luminance. The 5 central calibration points were presented twice at the beginning of each video in order to provide a robust estimate around the center of the screen. The total duration of the eye calibration points at the beginning of every imaging run was 82 seconds. An additional 5 central calibration points were presented in the middle (between the two short films presented) and the end of the imaging run. These points were used to evaluate eye tracking quality.

We estimated a warp field for each imaging run to map between pupil position and screen position using the 35 calibration points. We used a leave-one-out cross-validation procedure to choose the optimal warp field for each imaging run. We evaluated linear, smoothing spline and multiquadratic kernel smoothing along with corresponding hyperparameters (Oliphant, 2007). For each combination of hyperparameter and smoothing options, one calibration point was left out and the rest of the points were used to estimate the warp field. The estimated warp field was then used to predict the location of the left-out calibration point on the screen. The squared error between predicted and actual screen position was computed per calibration point. The mean across all errors was used as the metric to evaluate the optimal warp field. The optimal warp field was estimated for each imaging run, session and subject separately.

We assessed the quality of the eye tracking procedure by visual inspection. To do this, we created videos in which the estimated pupil location was overlayed on the eye video. We also created videos in which the estimated fixation point was overlayed on the short films.

#### 3.3.1.6 Spoken word transcripts

All the spoken words in the short films were manually transcribed. The transcriptions were first automatically aligned to the sound wave using the Penn Phonetics Lab Forced Aligner (Yuan and Liberman, 2008). Due to the presence of environmental sounds and music, the automatic alignment did not work well for most of the short films. Further manual adjustments were performed in order to align the spoken words to the sound wave using the open source software Praat (Boersma and van Heuven, 2001). These temporally aligned transcripts were used to construct feature spaces for different aspects of speech such as semantics, syntax and thematic roles.

### 3.3.2 Feature spaces

A total of 15 high dimensional feature spaces were used to model the visual, auditory, and speech content of the short films. An additional two feature spaces were constructed from nuisance regressors. The feature spaces were constructed using a combination of computational tools and hand labeling. For organizational convenience, the 15 feature spaces are grouped into 8 different categories (Figure 3.1).

#### 3.3.2.1 Motion-energy features

A spatiotemporal Gabor pyramid was used to extract low-level visual features (Adelson and Bergen, 1985, Watson and Ahumada, 1985, Nishimoto and Gallant, 2011). The pyramid consisted of a total of 11,845 three-dimensional Gabor filters spanning a square grid that covered the screen. The filters consisted of two spatial dimensions and one temporal dimension. Filters were created using six spatial frequencies (0, 2, 4, 8, 16, and 32 cycles per image), three temporal frequencies, (0, 2 and 4 Hz), and eight directions of motion (0, 45, 90, 135, 180, 225, 270 and 315 degrees). The short film frames were downsampled to 96x170 pixels to minimize computational cost. The RGB frames from the short films were extracted and converted to the CIE L*a*b* color space (McLaren, 1976) and the color information was discarded (see Nishimoto et al., 2011, for more details).

**Stimulus motion-energy**

Each of the 11,845 filters in the spatiotemporal Gabor pyramid was convolved with the luminance video. The resulting filter activation quadrature pairs were squared and summed. The output was downsampled from 24Hz to the functional image acquisition rate (2000ms) using sinc interpolation.

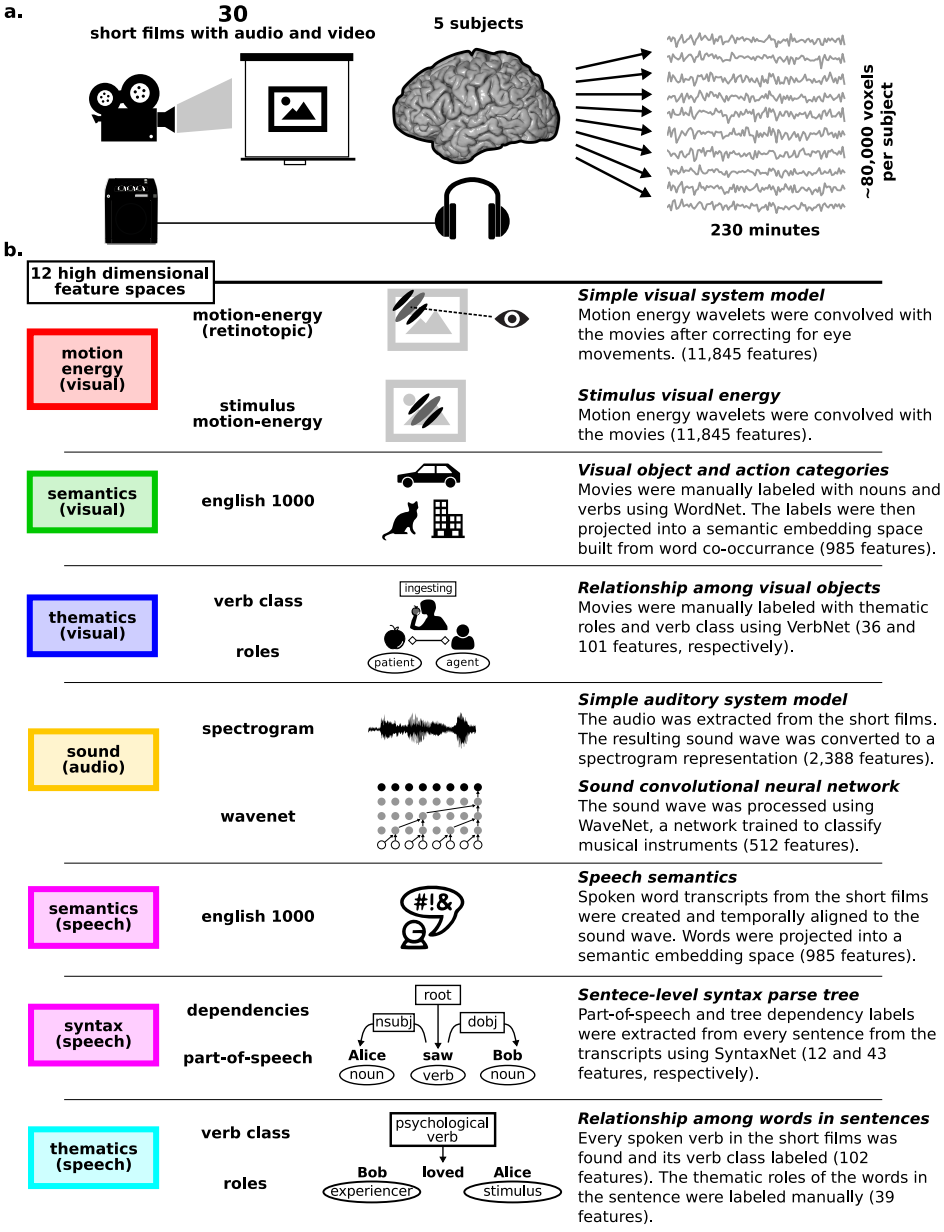


**Figure 3.1: Experimental procedure and feature spaces.** **(a)** Five participants watched thirty short films with audio and visual content while BOLD responses were measured using fMRI. **(b)** More than a dozen high dimensional feature spaces that describe the visual, auditory, and speech content of the short films were extracted using different computational procedures and manual labeling.

**Retinotopic motion-energy**

To correctly represent the visual information processed by early visual cortex, the motion energy features were recomputed to account for subject eye movements. Eye tracking data was used to create a retinotopic video. Each frame in the short films was centered on the point of fixation. The retinotopic RGB video was converted to the CIE L*a*b* color space. Eye movements effectively double the stimulus size. This resulted in a total of 47,186 filters. Virtually no difference in prediction performance was found when the highest spatial frequency filters (32cpi) were removed (data not shown). This resulted in 11,845 filters with seven spatial frequencies (0, 1, 2, 4, 8, and 16 cpi). The resulting retinotopic luminance video was convolved with the 11,845 filters (see above). The features were downsampled from 24Hz to the functional image acquisition rate (2000ms) using sinc interpolation.

#### 3.3.2.2 Visual semantic features

The visual semantic content of the short films was tagged. One observer manually labeled each second of the short films with WordNet labels describing the salient objects and actions in the visual scene (Miller, 1995, Huth et al., 2012). This was spot checked by three additional observers. This resulted in a total of 1,133 unique WordNet labels (i.e. synsets). The number of labels per second varied between 0 and 20, with an average of 6.34±3.05 s.d.

In order make the object category and action label results comparable to speech semantics, the WordNet labels were projected onto the same semantic vector space used for speech (see Section 3.3.2.5; Huth et al., 2016). To achieve this, the words belonging to each WordNet label were obtain from every second of the short films (e.g. the synset "water.n.01" becomes its lemmas "water" and "H20"). For each second every word obtained from the labels was projected onto the semantic feature space. This resulted in several 985-dimensional vectors each representing one words projection onto the semantic feature space. The average across the vectors was then taken resulting in a single 985-dimensional vector for every second of the short films. These features were then downsampled to match the functional image acquisition rate (2000 ms) using a 3-lobe Lanczos filter.

#### 3.3.2.3 Visual thematics

The relationship between objects and the type of events that occur in the short films was labeled using VerbNet (Kipper et al., 2006). VerbNet is a verb lexicon that can capture information about "Who (*actor*) does what (*verb class* and *verb category*) to whom (*undergoer*) where (*location*) and by means of what (*selective restrictions*)?"

VerbNet verb class and thematic roles were labeled for each second of the short films. In a visual scene where "A girl hits the ball with a stick" the verb class is *hit-18.1* (hit), and the thematic roles are agent (girl), patient (ball) and instrument (stick). The verb class also imposes selective restrictions on the thematic roles. For example, an agent must have *intentional control* in some events (e.g. Joe [actor with intentional control] hit Jon [undergoer]) but not in others (e.g. The meteor [actor without intentional control] crushed Jon [undergoer]). These restrictions (e.g. with or without intentional control) are determined on the thematic role by the verb class.

The thematic roles (36 binary features), selective restrictions (38 binary features) and their interaction (71 binary features), and the verb categories (101 binary features) were labeled for each second of the short films and used to construct three separate feature spaces. The feature spaces were downsampled to the temporal acquisition rate of the functional images (2000ms) by computing the mean of every two seconds and binarizing the result.

VerbNet provides a categorization of verb classes into categories based on the type of events described. For example, in a visual scene where "A girl hits the ball with a stick" the verb category is *contact by impact*. The verb classes hit-18.1, swat-18.2, spank-18.3, bump-18.4 all belong to the *contact by impact* verb category (Kipper et al., 2006). VerbNet verb categories were labeled for each VerbNet verb class for each second of the short films. This resulted in 101 binary features. The verb category feature space was downsampled to the temporal acquisition rate of the functional brain images (2000ms) by computing the mean of every two seconds and binarizing the result.

#### 3.3.2.4 Auditory features

The 48kHz sound wave from the short films was used to construct two auditory feature spaces.

**Spectrogram**

Spectrogram features were extracted by computing the spectral power density of the sound wave for every two second intervals of the short films. A frequency resolution of 2Hz was used over the range of 0 to 5kHz. The spectral power desntiy was log-transformed. This resulted in a total 2,388 features sampled at 2000ms.

**WaveNet**

Auditory features related to musical instruments were extracted using a convolutional neural network (Van Den Oord et al., 2016). The network was pre-trained to model

the spectral variation of 1,006 types of musical instruments (Engel et al., 2017). The key property of WaveNet is its ability to capture the temporal auto-correlation of the sound wave in a compact representation. A total of 512 WaveNet features were extracted for each second of the short films. Features were downsampled to match the rate of acquisition of the functional images (2000ms) by taking the mean of every two seconds.

#### 3.3.2.5 Speech semantic features

Each word in the short film transcripts was projected onto a 985-dimensional semantic feature space constructed from word co-occurrence statistics (Huth et al., 2016). The co-occurrence semantic feature space was a matrix of 985 rows and 10,470 columns. The 985 rows describe 985 basic words from Wikipedia's *List of 1000 basic words*, the 10,470 columns are words selected from a very large corpora that included transcripts of Moth Radio Hour stories, popular books from Project Gutenberg, Wikipedia pages and reddit.com user comments. The semantic feature space was constructed in a previous study (see Huth et al., 2016, for details). For each word within the short film transcripts the corresponding column in the semantic feature space was selected, creating a list of 985-dimensional semantic vectors. A 3-lobe Lanczos filter was then used to downsample the feature vectors to the acquisition rate of the functional brain images (2000 ms).

#### 3.3.2.6 Speech syntax features

The syntactic properties of each spoken word were labeled. A pre-trained neural network was used to create a parse tree for each sentence of the short film transcripts (Andor et al., 2016). Two feature spaces were extracted from the parse trees. The first was constructed from the part-of-speech tags (e.g. noun, verb) and consisted of 12 binary features. The second feature space captured the word dependencies in the sentence (i.e. direct object, indirect object, etc.) and consisted of 43 binary features. Each word in a sentence was assigned a feature in each of the two syntactic feature spaces. For each syntactic feature (e.g. noun), a time course was created with a value of 1 whenever a word was labeled with that feature and 0 otherwise. The syntactic features were then downsampled to the rate of acquisition of the functional images (2000ms) using a 3-lobe Lanczos filter.

#### 3.3.2.7 Speech thematic features

VerbNet was used to label thematic roles, verb classes and verb categories for each sentence of the short film transcripts (Kipper et al., 2006). Thematic roles model

how words in a sentence relate to each other and verb categories capture the types of events that occur in the short films. VerbNet allows us to capture a more complex representation of sentences by modeling "Who (*actor*) does what (*verb class* and *verb category*) to whom (*undergoer*) where (*location*) and by means of what (*selective restrictions*)?"

Each verb in in each sentence of the transcripts was labeled with a verb class (e.g. eat becomes *eat-39.1*). For each verb, thematic roles were then labeled manually for each word (e.g. "Alice" is *agent*, "apple" is *patient*). The verb category was also labeled for each verb in the sentence. For example, the verb classes eat-39.1, chew-19.2, devour-39.4 all become *verbs of ingesting*. These VerbNet labels were used to construct a thematic role feature space (39 features), a verb class feature space (274 features), and a verb category feature space (102 features).

To explore the effect of thematic roles on semantic representation, separate semantic feature spaces were constructed for each of the three thematic roles (actor, recipient, place). For example, each word that was assigned to the role of "actor" in a sentence was projected to the semantic vector space. This gave us three thematic role specific 985-dimensional semantic feature spaces (actor semantics, recipient semantics, place semantics). This allowed us to capture whether semantic representations are modulated by the particular thematic role that they undertake. We also constructed a verb-specific semantic space.

All seven thematic feature spaces described above were downsampled to the rate of the functional image acquisition (2000ms) using a 3-lobe Lanczos filter.

#### 3.3.2.8 Nuisance regressors

**Eye movement**

Eye movements can lead to changes in the magnetic field and cause spurious BOLD signals in regions close to the eyes. This is particularly important for regions around orbitofrontal cortex and anterior temporal cortex. Additional analyses were conducted in order to address this issue (data not shown).

The x-y location and size of the pupil was extracted from the eye tracking analysis. These three features were used as nuisance regressors. For each of the eye movement dimension, we created a rectified cubic polynomial expansion. This yielded a total of 36 features. The features were downsampled from 24Hz to 2000ms using cubic polynomial interpolation in order to match the temporal resolution of the sampled BOLD signals. The eye movement feature space can also capture cortical signals associated with the execution of eye movements.

#### Word rate

To capture the main effect of language presence in the short films (independent of semantics, syntax or thematics), a binary vector was created to record when a word was present. The vector was set to one whenever a word was spoken in the short films and zero otherwise. The vector was sampled at the time of word onset and was then downsampled to the rate of acquisition of the functional images (2000ms) using a 3-lobe Lanczos filter. A cubic polynomial expansion was created from this single feature which was then rectified. This yielded a total of 13 features.

### 3.3.3 Analyses to recover visual retinotopy and auditory tonotopy

#### 3.3.3.1 Retinotopy

The retinotopic motion energy feature space was used to model visual responses to the short films. In order to account for the slow hemodynamic response, feature weights were estimated per voxel at delays of 4, 6, and 8 seconds. The model was estimated separately for each voxel using ridge regression (Hoerl and Kennard, 1970). The optimal regularization parameter across all voxels for each subject was estimated via 5-fold cross-validation. The estimated model weights were then used to recover retinotopic maps for each subject.

For each voxel, the maximum weight across the 11,845 spatio-temporal Gabor filters was found. The x-y position of the filter with the largest weight was then used to compute the visual angle and eccentricity for each voxel. This provides a rough estimate of the voxels spatial receptive based on the filters location on the screen. A better but more computationally demanding approach is to simulate voxel responses to dynamic Gaussian noise stimuli and compute the stimulus triggered average. The spatial receptive field can then be estimated by fitting a two-dimensional Gaussian to the stimulus triggered average. This will be done in future work.

#### 3.3.3.2 Tonotopy

A simple spectrogram model was used to model auditory responses to the short films. The spectrogram was constructed by computing the log power spectral density from 0-5kHz in steps of 52Hz over 100ms windows from the sound pressure wave (48kHz). This resulted in 96 features sampled at 100ms for each of the 30 short films. The resulting spectrograms were then downsampled to 2000ms using a sinc filter in order to match the temporal acquisition of the BOLD responses.

Ridge regression was used to model voxel responses as a linear combination spectrogram features. In order to account for the relatively fast onset of the hemodynamic response function in auditory cortex, features were delayed by 2, 4, 6 and 8 seconds (Oppenheim et al., 1983). A single ridge regularization parameter per subject was estimated via 5-fold cross-validation.

For each voxel, the weight amplitude for each spectral feature was computed by taking the average of the estimated weights across the four delays. The largest weight amplitude across the 96 spectrogram features was then found for each voxel. The spectral frequency associated with the largest weight was selected as the best frequency for that voxel.

### 3.3.4 Joint voxelwise encoding model

A single joint model that included all 15 feature spaces and the nuisance regressors was estimated for each voxel separately. This allowed us to partition the explained variance of each voxel into different feature spaces within the joint model (Figure 3.2). This is important when feature spaces are correlated with each other and the shared variance needs to be split among them.

Combining all the feature spaces into a single joint model is difficult from a statistical learning perspective. The joint model has a relatively large number of parameters ($\sim$30,000). Estimating the model with ordinary least squares is not feasible because the number of parameters exceeds the number of data points. In such cases, regularized regression techniques like ridge regression (Hoerl and Kennard, 1970), LASSO (Tibshirani, 1996), and elastic net (Zou and Hastie, 2005) are commonly used to estimate the model parameters (i.e. weights). However, these approaches assume that the joint model weights are all similarly distributed and therefore require the same amount of regularization. This is not a good assumption for the joint model weights because different feature spaces might need different levels of regularization. This can happen when the feature spaces contained in the joint model are of different dimensionality. Estimating a single joint model using ridge regression, LASSO, or elastic net will be suboptimal if different feature spaces require different levels of regularization.

#### 3.3.4.1 Banded ridge regression

We developed a novel voxelwise modeling framework to overcome the complexity of combining multiple feature spaces into a single joint model. In our framework, BOLD responses are modeled as a linear combination of all the feature spaces using linear regression with a non-spherical spatio-temporal multivariate normal prior on

the weights. This approach allows us to impose different levels of regularization on each feature space within the joint model. The regularization parameter for each feature space defines the covariance of the multivariate normal (MVN) prior and is estimated empirically via cross-validation. In essence this is a special case of Tikhonov regression (Tikhonov et al., 1977). We refer to our approach as banded ridge regression (see Chapter 1 Section 1.5 for details).

There are two components in the spatiotemporal MVN prior used in the model. The first is a non-spherical MVN prior on the covariance of the feature weights. This component allows us to apply a different level of regularization to each feature space. The second component is a non-spherical MVN prior on the temporal covariance of the weights and is based on the shape of the hemodynamic response function (Friston et al., 1998). The two MVN priors are then combined into a single MVN prior by computing the Kronecker product between them.

**Feature prior**

In order to account for the fact that different feature spaces require different levels of regularization, we imposed a non-spherical multivariate normal prior on the feature weights. The non-spherical prior imposes a spherical multivariate normal prior on each feature space separately, while taking into account the correlations across feature spaces. The precise shape of the non-spherical prior was estimated via cross-validation.

**Temporal prior**

We modeled the hemodynamic response function using a finite impulse response (FIR) filter per voxel and for each subject separately. This was implemented by modeling the BOLD responses at ten temporal delays corresponding to 0, 2, 4, 6, , 16 and 18 seconds. We imposed a MVN prior on the temporal covariance of the FIR filter. The temporal prior was constructed from a set of HRF basis functions (Penny et al., 2011). This temporal prior allows us to include knowledge about the shape of the HRF into the estimation of the FIR filter for each voxel (Marrelec et al., 2003).

#### 3.3.4.2 Cross-validation

We used cross-validation to estimate the model hyperparameters. This required testing different hyperparameters each one controlling the regularization level for each feature space while taking into account correlations across feature spaces. There were 17 total hyperparameters one per feature space. Evaluating ten hyperparameter

values for each of the 17 feature spaces results in a total of $10^{17}$ hyperparameter sets to test. This is a computationally intractable number.

To overcome this problem, we turned to global search techniques to estimate the optimal hyperparameters per voxel (Bergstra et al., 2013). We performed 300 iterations of a tree-structured Parzen search algorithm to find the optimal hyperparameter for each feature space across all voxels (Bergstra et al., 2011). This process was repeated twenty five times independently. For every set of hyperparameters tested in each iteration, we performed 5-fold cross-validation twice. We used the coefficient of determination ($R^2$) between the predicted and the actual voxel responses as our performance metric for each validation fold.

#### 3.3.4.3 Model estimation

We computed the average prediction performance across cross-validation folds per voxel for each of the 7500 (300 x 25) hyperparameter sets tested. The hyperparameter set that yielded the maximum average cross-validated prediction performance was selected for each voxel. This hyperparameter set was then used to estimate the joint voxelwise model across the full training set.

A separate model was estimated for each of the ∼80,000 voxels for each of the eight subjects (∼400,000 total voxels). Each single joint model consisted of a total of ∼30,000 features and was fit on 3,572 data samples (120 minutes; 24 short films).

#### 3.3.4.4 Model evaluation

BOLD responses to six short films not used to fit the model were used to assess the ability of the joint model to predict new data (27 minutes). Each of the six short films was presented four times to the subjects (three times for subject SS), and then the four voxel time courses were averaged. This was done to increase the SNR of the measured voxel time courses.

The estimated joint model weights were then used to predict the voxel responses to the six test short films (27 minutes). Model prediction performance was computed per voxel as the coefficient of determination ($R^2$) between predicted and actual responses. We also used the individual feature space weights to compute feature space specific prediction performance.

In order to make our results more comparable to previous studies that were based on the correlation coefficient ($r$; Huth et al., 2016), we use the square root of the coefficient of determination ($R^2$) to visualize the results ($r$ and $\sqrt{R^2}$ are more directly comparable).

## 3.4 Results

In a single experimental paradigm and data collection effort, we aim to determine the functional selectivity for visual, auditory, and linguistic features (Figure 3.1) across the cortical surface in individual subjects. Subjects watch 2 hours of audio-visual short films while whole-brain BOLD activity is recorded with functional MRI (Figure 3.2a). The short films are labeled with more a dozen high-dimensional feature spaces that reflect their visual, auditory and conceptual content. A novel voxelwise encoding model is used to model BOLD responses to the short films using all feature spaces simultaneously. The estimated model is then used to predict voxel responses to 27 minutes of novel short films not used for model estimation. We validate our approach by assessing the accuracy and statistical significance of the model predictions. We then identify which feature space captures the most variance for each voxel in the cortical surface, We use the estimated model weights to recover known visual retinotopic, auditory retinotopic maps and semantic maps. Finally, we show novel functional subdivisions within the middle temporal cortex. A summary of results can be found in Table 3.1.

| Result | Section | Figure |
|---|---|---|
| Joint model containing more than a dozen feature spaces significantly predicts cortical responses | 3.4.1 | 3.2, 3.3 |
| Joint model performs better on average than any single space alone | 3.4.1.1 | 3.4 |
| areas: V1, V2, V3, V3ab, V4, V7, and hMT+<br>feature space: motion energy (5/5) | 3.4.1.2 | 3.5 |
| areas: far visual periphery directly posterior to RSC and PPA<br>feature space: motion energy (5/5) | 3.4.1.2 | 3.5 |
| areas: RSC, OPA, PPA (5/5), and IPS (4/5)<br>feature space: visual semantics | 3.4.1.2 | 3.5 |
| areas: STG, Brocas area, sPMv, inferior TPJ, SFG, and MPC<br>feature spaces: speech semantic, syntactic and word rate (5/5) | 3.4.1.2 | 3.5 |
| area: primary auditory cortex<br>feature spaces: spectrogram and wavenet (5/5) | 3.4.1.2 | 3.5 |
| areas: region anterior to hMT+ and posterior STS<br>feature space: visual thematics (5/5) | 3.4.1.2 | 3.5 |
| area: far anterior precuneus (posterior to S1f)<br>feature space: visual thematic features (5/5) | 3.4.1.2 | 3.5 |
| area: anterior precuneus (aPCu)<br>feature space: motion energy (7/10 hemispheres) | 3.4.1.2 | 3.5 |
| Large regions of cortex are best predicted by<br>the same feature space | 3.4.1.2 | 3.5 |
| Maps of feature space selectivity are highly consistent<br>across subjects in most of the cortical surface | 3.4.1.2 | 3.5 |

**Table 3.1 continued from previous page**

| Result | Section | Figure |
|---|---|---|
| area: putative VIP+ (anterior to IPS and posterior to dorsal S1)<br>feature spaces: visual thematics (3/5) and visual semantics (2/5) | 3.4.1.2 | 3.5 |
| area: putative LIP+ (region in the middle of IPS)<br>feature space: motion energy in (5/10 hemispheres) | 3.4.1.2 | 3.5 |
| area: posterior central sulcus<br>feature space: visual semantics (4/5) | 3.4.1.2 | 3.5 |
| area: frontal eye fields (FEF)<br>feature space: visual semantics (2/5) and visual thematics (3/4) | 3.4.1.2 | 3.5 |
| area: dorsal lateral parietal cortex (ventral to middle of IPS)<br>feature space: spectrogram and wavenet features (4/5) | 3.4.1.2 | 3.5 |
| Retinotopic maps can be estimated from motion energy features<br>during free view (1/5) | 3.4.2.1 | 3.6 |
| Tonotopic maps can be recovered from the short films (5/5) | 3.4.2.2 | 3.7 |
| A1 and R can be identified within primary auditory cortex from<br>estimated tonotopic maps (5/5) | 3.4.2.2 | 3.7 |
| area: region posterior to A1 (putative CL) is tuned for low frequencies<br>feature space: spectrogram (8/10 hemispheres) | 3.4.2.2 | 3.7 |
| Semantic models for vision and speech can<br>be estimated from short films (5/5) | 3.4.2.3 | 3.8 |
| Visual and speech semantics predict largely<br>non-overlapping regions (5/5) | 3.4.2.3 | 3.8 |
| Maps of semantic tuning in vision and speech<br>can be recovered from the short films (1/5) | 3.4.2.3 | 3.9 |
| Semantic tuning of visual regions derived from object<br>and action categories projects onto the same<br>speech-derived semantic concepts (1/5) | 3.4.2.3 | 3.9 |
| Motion energy and visual semantic features predict voxel activity<br>in broadly different cortical regions (5/5) | 3.4.3.1 | 3.10 |
| Low-level sound features and high-level speech derived<br>auditory features predict largely non-overlapping regions<br>of auditory cortex (PAC and STG 5/5 subjects) | 3.4.3.2 | 3.11 |
| area: region anterior to hMT+ (dorsal EBA in 4/5 subjects)<br>feature space: visual thematics (5/5) | 3.4.4.1 | 3.12 |
| area: region ventral to hMT+ (ventral EBA in 3/5 subjects)<br>feature space: visual semantics (5/5) | 3.4.4.1 | 3.12 |
| Activity in region anterior to hMT+ (dorsal EBA in 4/5 subjects)<br>is better predicted by visual thematics than motion energy<br>and visual semantics (5/5) | 3.4.4.1 | 3.12 |
| Second best predictive feature spaces for the visual semantics<br>selective region anterior to hMT+ are motion energy<br>and visual semantics (5/5) | 3.4.4.2 | 3.13 |

**Table 3.1 continued from previous page**

| Result | Section | Figure |
|---|---|---|
| An area anterior to hMT+ is functionally distinct from a area ventral to hMT+ that is best predicted by visual semantics (5/5) | 3.4.4.3 | 3.14 |

**Table 3.1:** Summary of results for short films experiment.

We begin by showing that the joint voxelwise model is able to predict BOLD responses to novel short films in many voxels across the cortical surface (Section 3.4.1). We show that the joint model performs as well or better than any single feature space alone for most voxels. We then show that the patterns of feature space selectivity across the cortical surface is highly consistent across all five subjects.

To demonstrate the validity of feature maps recovered using short films, we use the motion energy and spectrogram feature spaces to recover known visual retinotopic and auditory tonotopic maps (Section 3.4.2). We explore the differences in tuning and prediction performance of semantics in both vision and speech. We also explore feature selectivity boundaries in visual and auditory cortices (Section 3.4.3).

Finally, we analyze the functional selectivity for visual features in middle temporal cortex (Section 3.4.4). We compare prediction performance of motion-energy, visual semantics, and visual thematics. Our data suggest that there is a high degree of functional specialization in this region.

### 3.4.1 Joint model significantly predicts voxel activity across the cortical surface

The joint voxelwise encoding model contains a large number of parameters across the feature spaces (∼30,000). This gives the joint model the flexibility needed to accurately model the individual voxel responses to the short films. However, model flexibility comes at the cost of "overfitting". Overfitting can occur when the estimated model parameters are dominated by the specific stimuli used or by the noise of the data used to estimate the model. A model that is overfit to the estimation data does not generalize well to novel stimuli and data. This is a problem because inferences based on overfit models are unlikely to generalize to new observations (Friedman et al., 2001). To address these potential issues, we first evaluate the ability of our model to generalize outside of the stimuli and data used to estimate it.

The estimated joint model was used to compute predictions of individual voxel responses to the six short films collected during the test set (27 minutes total). We use the coefficient of determination ($R^2$) computed between the predicted and

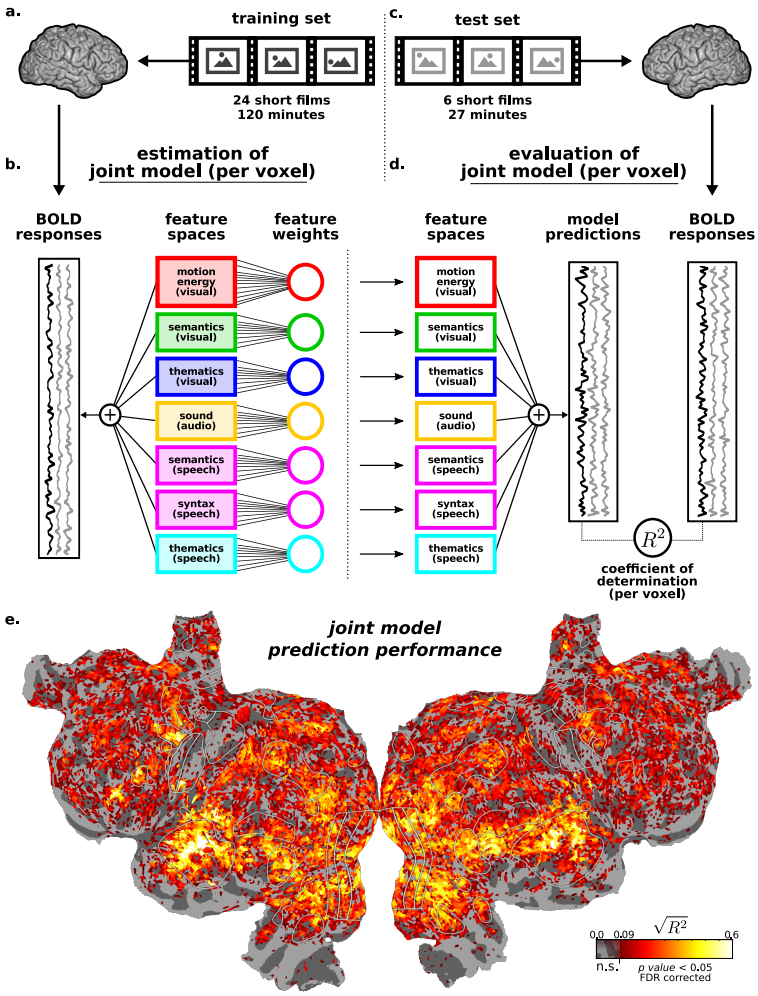


**Figure 3.2: Voxel-wise modeling and joint model prediction performance.** **(a)** A training set consisting of 24 short films (120 minutes) was used to estimate a single joint model for each voxel. **(c)** A separate held-out test set consisting of 6 short films (27 minutes) was used to evaluate the joint model. **(b)** Visual, auditory, and speech related features were extracted using twelve high dimensional feature spaces (28,893 features in total). BOLD responses were modeled as a linear combination of all features across feature spaces. **(d)** The estimated model weights were used to predict BOLD responses to the 27 minute held-out test set. Model prediction performance was quantified as the coefficient of determination ($R^2$) between the predicted and actual BOLD responses to this held-out test set **(e)** Joint model prediction performance of one subject is plotted onto the cortical surface. Yellow and white colors depict well predicted voxels by the joint model. Grey voxels are not significant (FDR-corrected for multiple comparisons, $q(FDR) > 0.05$). Most of the cortical voxels within the visual, temporal, parietal, and prefrontal cortices are well predicted using the joint model.

the actual voxel responses as our measure of prediction performance. The joint voxelwise model significantly predicts the activity of voxels distributed broadly across the cortical surface in all five subjects ($q(FDR) < 0.05$; Figure 3.2c). Regions that are well predicted include sensory regions such as early visual cortex (EVC), primary auditory cortex and speech regions such as Broca, superior premotor ventral (sPMv) and superior temporal gyrus (STG). The model also provides significant predictions in prefrontal (PFC) and lateral parietal (LPC) association regions, though these predictions are not as accurate as those made for sensory and speech regions. This pattern of results is found consistently across all five subjects (Figure 3.3). Thus, the joint model provides good predictions of voxel responses that generalize beyond the stimuli used to estimate the model.

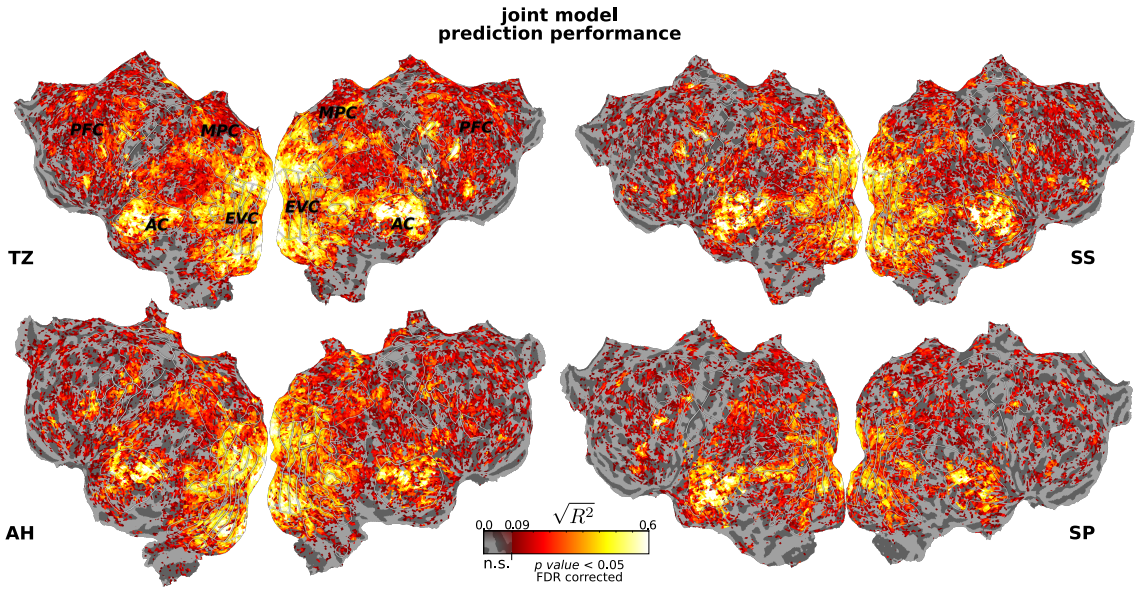


**Figure 3.3: Joint model prediction performance for all the other subjects.** Voxelwise joint model prediction performance for all the other subjects are plotted onto each subjects flattened cortical surface. Yellow and white colors depict well predicted voxels by the joint model. Gray voxels are not significant ($q(FDR) > 0.05$). All subjects show a similar prediction performance where most of the cortical voxels within the visual, temporal, parietal, and prefrontal cortices are well predicted using the joint model.

#### 3.4.1.1 Joint model performs better than any single feature space model alone

The joint model can significantly predict voxel activity in several regions across the cortical surface, however it might be suboptimal for some voxels. This can occur when voxels are located in regions where only a single feature space is important (e.g. motion energy in early visual cortex) and the estimated joint model does not learn to ignore all unimportant feature spaces (Friedman et al., 2001). We therefore tested whether and where models built from single feature spaces alone provided more accurate predictions of voxel activity than the joint model. For each subject, we used ridge regression to model individual voxel responses as a function of each feature space separately (Hoerl and Kennard, 1970). This resulted in 17 ridge regression models one for each feature space per voxel and per subject. For each voxel, the maximum prediction performance across the 17 ridge regression models was identified and compared against the prediction performance of the joint voxelwise model.

We find that voxels in early visual cortex and hMT+ are better predicted by the single motion energy ridge regression model than the joint model across subjects (Figure 3.4a, blue). This suggests that the joint model is suboptimal for voxels in early visual cortex and hMT+. The inclusion of additional feature spaces negatively impacts the prediction performance in those regions because the motion energy features are the single most important feature space.

However, the joint model performs approximately 10% better on average across cortical voxels in all subjects than any other feature space alone (voxel population mean for joint model $R^2$=0.040 versus $R^2$=0.037 for the maximum ridge model across feature spaces; Wilcoxon $W = 10^{10.13}, p < 10^{-12}$, 240,765 voxels) in predictable voxels ($R^2 > 0$). These results suggests that the joint model performs as well or better than any one single feature space alone for the majority of cortical voxels across all subjects.

#### 3.4.1.2 Feature space selectivity is homogeneous across large regions of the cortical surface and is highly consistent across subjects

In the joint model, voxel responses to the short films are modeled using all 17 feature spaces simultaneously. Each feature space constitutes a hypothesis about the information that is represented in the brain (e.g. frequency content of sound). The question of how the brain represents information about the world and how those representations are used to perform abstract cognitive functions has a long history in neuroscience. Cognitive neuroscience and functional neuroimaging in particular

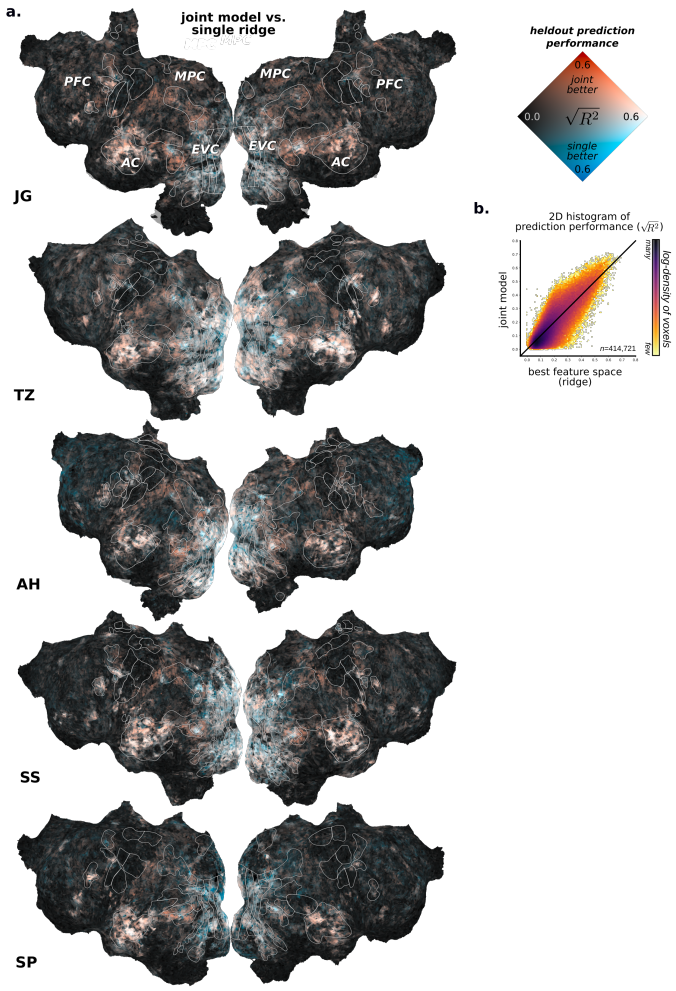


**Figure 3.4: Comparison of joint model against single feature space prediction performance for all subjects.** We estimated single ridge regression models for each feature space independently. We then found the maximum ridge regression prediction performance for each voxel across all feature spaces. **(a)** We compared the maximum ridge model against the joint model prediction performance. The comparison for each individual subject is shown on the cortical surface. Voxels in early visual cortex (EVC) and hMT+ are better predicted by a single feature space than by the joint model (blue). Anterior visual cortex, auditory cortex, regions surrounding hMT+, and intraparietal sulcus (IPS) are better predicted by the joint model (red). White voxels are well predicted by both and black voxels are not well predicted by either. **(b)** Log density histogram of maximum single ridge model (x-axis) versus joint model (y-axis) prediction performance. The joint model performs approximately 10% better on average than the best individual feature space estimated alone (Wilcoxon $W = 10^{10.13}, p < 10^{-12}$) across all predictable voxels from all five subjects ($R^2 > 0$; 240,765 of 414,721 voxels).

have focused on how information is represented in the brain. Broadly, two views have emerged. In one view, brain areas are functionally specialized for the processing of single cognitive functions (e.g. Kanwisher, 2010). In another view, representations are widely distributed and overlapping without clear functional localization (e.g. Haxby et al., 2001). We sought to examine whether the representations captured by the labeled feature spaces are localized or distributed across the cortical surface using the estimated joint model.

We first identified the feature space within the estimated joint model that provided the highest prediction performance on its own for each voxel per subject. To do this, the weights from the estimated joint model are used to compute the prediction performance for each feature space separately. For each voxel, we then identify the feature space with the highest $R^2$ value. Finally, we color-code the feature spaces and plot each voxel with the color corresponding to the best feature space on the cortical surface (Figure 3.5a). Note that voxels with low prediction performance are excluded from this analysis ($\sqrt{R^2} < 0.1$ corresponding to non-significant voxels at FDR-corrected $p < 0.05$ for all subjects).

We find that the patterns of feature space selectivity are homogeneous across large regions of the cortical surface (Figure 3.5). For example, voxels in early visual cortex are best predicted by the motion-energy feature space across all subjects (Figure 3.5 red). Similarly, primary auditory cortex is well predicted by the spectrogram and WaveNet feature spaces (Figure 3.5 yellow). These patterns are highly consistent and they replicate across all five subjects. The most inconsistent area across subjects is the frontal eye fields (FEF) which is best predicted by visual semantic features (JG, TZ) or visual thematics (AH, SP, SS). These results show that large contiguous regions of the cortical surface are involved in the processing the same type of information (e.g. visual semantics). However, more than one contiguous region is typically involved in the processing of the same type of information (e.g. interparietal sulcus and inferiotemporal cortex for visual semantics). This suggests that the functional organization of the cortical surface can be best characterized by a combination of both functional specialization and distributed representations.

### 3.4.2 Feature maps across the cortical surface

The joint model results suggest that large regions of the cortical surface are specialized for the processing of specific information captured by the feature spaces (e.g. early visual cortex and low-level visual information captured by motion energy). We next explore how these regions represent the feature space they are best predicted by (i.e. tuning). To do this, we construct feature maps that capture the tuning proper-

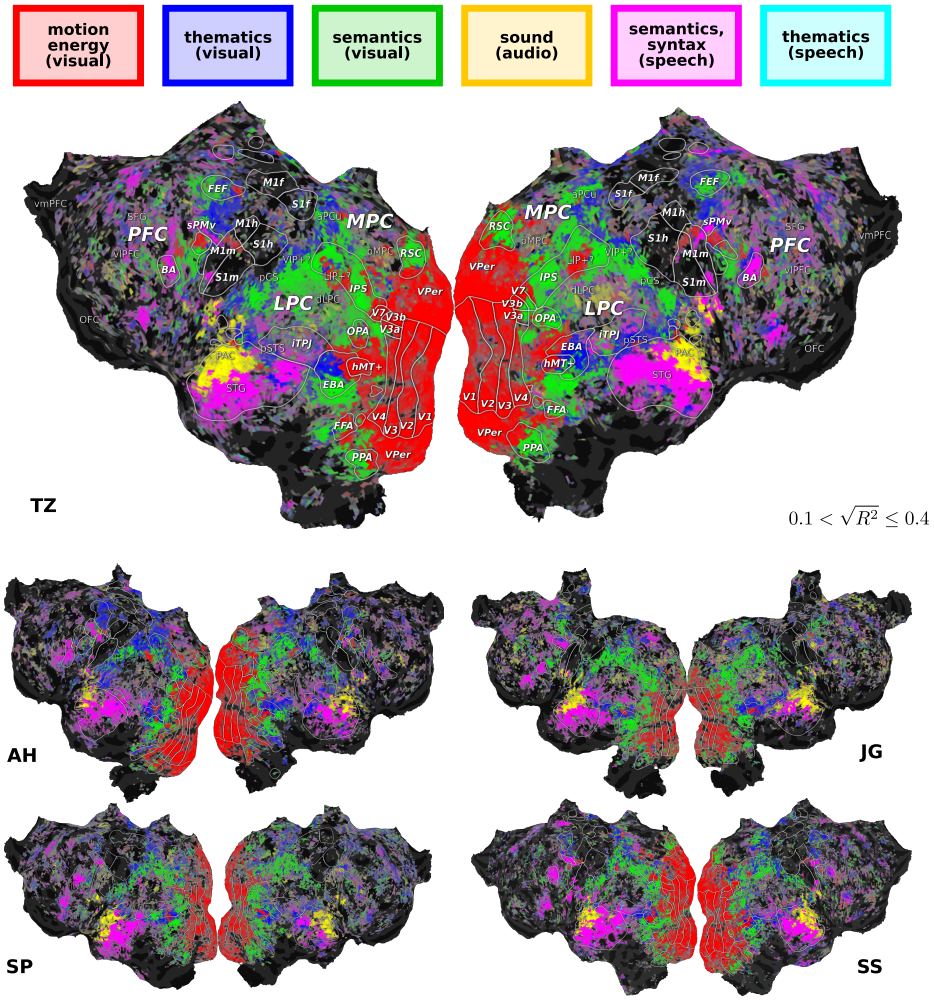


**Figure 3.5: Selectivity for the visual, audio, and speech feature spaces for all subjects**. We found the best predictive feature space per voxel for each subject from the estimated joint model. Prediction performance was quantified as the coefficient of determination ($R^2$) between feature space specific predictions and the actual BOLD responses to the six held-out short films (27 minutes). We divided the feature spaces into groups and assigned each group a color (see Figure 3.1). We colored each voxel according to the most predictive feature space and plotted the result on the cortical surface. We excluded voxels that were not well-predicted ($\sqrt{R^2} < 0.1, q(FDR) < 0.05$ per subject; black). The pattern of selectivity is consistent across subjects. Early visual cortex (EVC) is best predicted by motion-energy features (red), higher visual cortex (anterior to EVC) by semantic features (green), early auditory cortex (dorsal region labeled AC) by sound features (yellow), higher auditory cortex (ventral region labeled AC) by speech features (pink), and prefrontal cortical regions (PFC) also by speech features.

ties of voxels in functionally specialized cortical regions (e.g. visual angle maps from motion-energy features in early visual cortex).

In order to validate our approach, we first explore whether we can recover known brain maps from our data. We focused on recovering the retinotopic map of the visual system from motion-energy features (Sereno et al., 1995, Nishimoto et al., 2011), the tonotopic map of the auditory system from spectrogram features (Fay et al., 1992, Talavage et al., 1997), and semantic maps derived from visual and speech semantic features (Huth et al., 2012, 2016).

#### 3.4.2.1 Retinotopic maps in early visual cortex

Perhaps the most well-known sensory map in the human brain is that of the visual field in early visual cortex. The light that arrives in the eye produces an image on the retina (Palmer, 1999). This retinotopic representation is preserved all the way to cortex and is called retinotopy. This is the reason why neighboring cortical regions represent neighboring parts of the visual field. Retinotopic maps are typically estimated from optimized experiments (Sereno et al., 1995, Hansen et al., 2007, Dumoulin and Wandell, 2008) but can also be estimated from natural movies (Nishimoto et al., 2011). To our knowledge, no study has shown retinotopic maps from free viewing of naturalistic stimuli in humans. In order to validate our approach, we begin by estimating retinotopic maps from the visual responses to the short films.

Figure 3.6 shows retinotopic maps estimated from visual responses to the short films using the retinotopic motion-energy feature space for one subject. For each voxel, the spatial receptive field location was estimated as the location of the motion energy filter with the maximum weight estimate in the model. The filter position was then used to compute the optimal visual angle and eccentricity for that voxel.

The visual angle reversal between V1/V2, V2/V3, and V3/V3ab can be seen in Figure 3.6. The visual field of view used in the short films extends beyond that used to definite the visual area boundaries drawn on the cortical surface (24x24 degrees in retinotopic mapping experiment and 34x20 in short films). Voxels in the far periphery bordering anterior visual cortex show tuning for far eccentricities (Figure 3.6). The estimated retinotopic maps can be improved with better spatial receptive field estimates based on previous studies (Nishimoto et al., 2011). Nevertheless, these preliminary results show that visual angle and eccentricity maps can be estimated from the free viewing of short films.

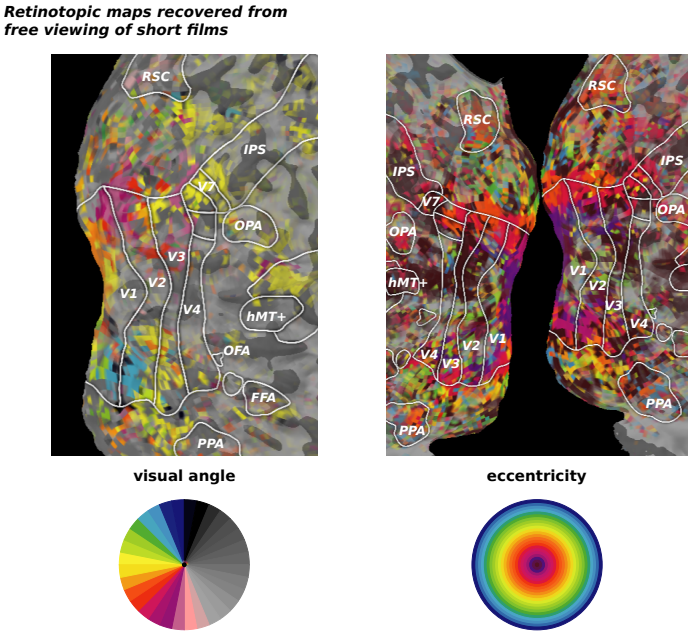


**Figure 3.6:** Estimated visual retinotopy maps from free viewing of short films.

### 3.4.2.2 Tonotopic maps in primary auditory cortex

Another well known sensory map in the human brain is that of sound frequency in primary auditory cortex (PAC). The sound pressure that arrives at the ear is translated into a frequency representation in the cochlea (Fay et al., 1992). This frequency representation is preserved all the way to cortex and is called tonotopy. PAC consists of two distinct areas A1 and R each receiving thalamic input (Merzenich and Schreiner, 1992). These areas contain mirrored representations of sound frequency. Studies showing tonotopy in PAC typically involve the use of high field strength scanners (7T; Formisano et al., 2003) or optimized experiments with simple stimuli (Talavage et al., 1997, 2004. The short films contain a much broader range of sounds than is typically explored in fMRI (but see Lewis et al., 2005, 2011, 2012). We therefore explored whether we could estimate tonotopic maps in PAC from our data in individual subjects.

Figure 3.7 shows the tonotopic maps estimated from auditory responses to the short films modeled using a simple spectrogram model. To find the optimal frequency for each voxel in PAC, we found the frequency with the maximum weight from the estimated model. The optimal frequency for voxels in PAC is displayed for all five

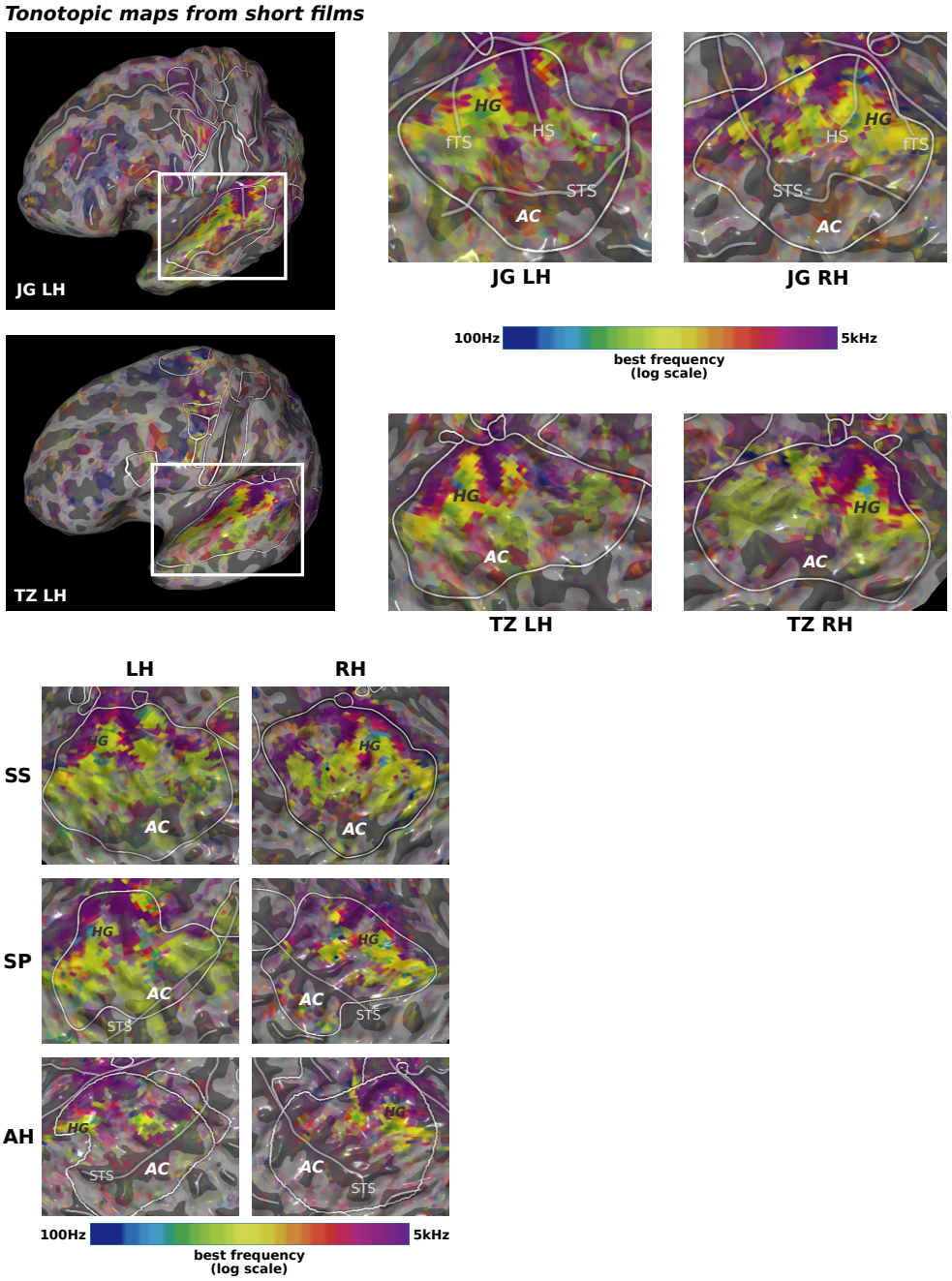


**Figure 3.7:** Auditory tonotopy maps estimated from the short films.

subjects in Figure 3.7. The tonotopic pattern is clearly visible in each individual subject. Voxels located on Heschls gyrus are selective for low frequencies whereas regions anterior and posterior to it are selective for higher frequencies. There is a smooth anterior to posterior gradient in PAC with selectivity for high-to-low-to-high frequencies. Previous studies have identified A1 as the posterior low-to-high region and R as the anterior high-to-low region. Interestingly, a region in the lateral belt located posterior to Heschl sulcus is tuned for low frequencies and was found in 8/10 hemispheres (putative CL; Talavage et al., 2004, Humphries et al., 2010, Moerel et al., 2014). These results demonstrate that tonotopic maps can be estimated in individual subjects from PAC activity measured while subjects listen to the rich auditory content present in the short films.

#### 3.4.2.3 Visual and speech semantics

Previous work has explored visual semantic representations in movies without sound and also semantic representations from spoken stories (Huth et al., 2012, 2016, respectively). We can use the short films to explore both of these semantic representations in one single experiment. The visual object and action categories in the visual scenes and the words spoken in the short films can be used to extract feature spaces related to visual semantics and speech semantics, respectively. This allows us to explore semantics conveyed by the visual modality and contrast them to those conveyed through speech.

We find that regions in lateral and medial parietal (LPC, MPC), superior temporal gyrus (STG), as well as prefrontal cortex regions involved in language processing (Brocas area, sPMv) are significantly predicted by the speech semantic feature space (Figure 3.8a). The regions well predicted by visual semantics are located in anterior visual cortex, interparietal sulcus (IPS), and posterior central sulcus (PCS; Figure 3.8b). There is little overlap between the regions that are well predicted across the visual and linguistic modalities (Figure 3.8c). This suggests a separation between semantics conveyed through vision and speech.

We next explored the semantic tuning for visual and speech semantics separately (Figure 3.9). To achieve this we projected the semantic weights from each modality into the semantic space from a previous study (Huth, et al., 2016). This projection allows us to interpret the 985-dimensional semantic tuning per voxel in visual and speech modalities in a lower-dimensional space consisting of three dimensions. The speech semantic tuning in association regions replicates previous work (Huth, et al., 2016). The visual semantic tuning of anterior visual cortex, IPS and posterior CS maps onto the same part of the speech-derived semantic space. This means that all

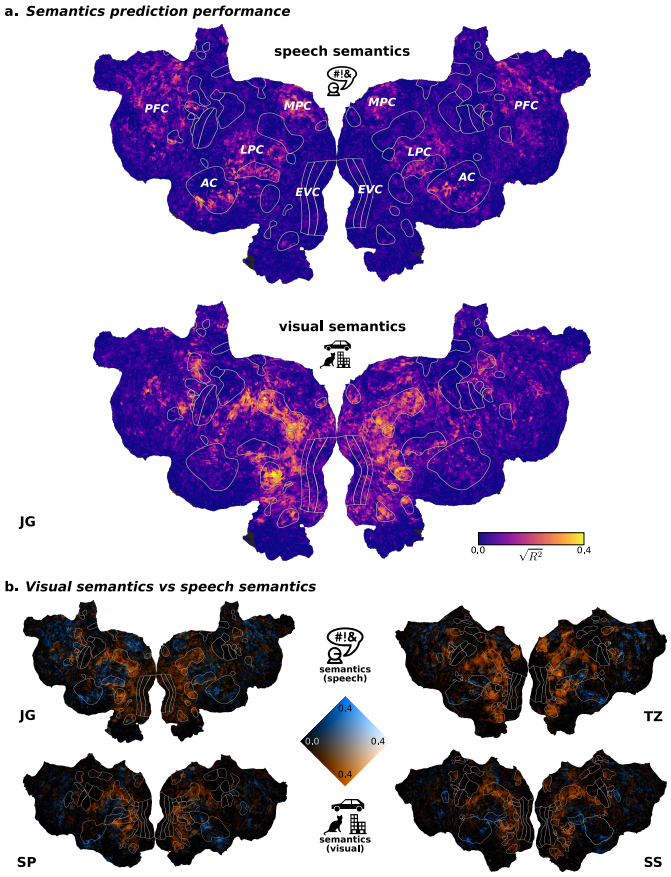


**Figure 3.8: Comparison of prediction performance of speech and visual semantics.** The estimated joint model was used to compute prediction performance for each of the speech and visual semantic feature spaces separately. The speech semantics feature space the meaning of the spoken words that occur in the short films. The visual semantics feature space captures the category of all objects and actions that appear in the short films **(a)** Prediction performance was obtained from the speech semantics model by computing the coefficient of determination between predicted and actual responses for each voxel. Regions of lateral and medial parietal cortex (LPC, MPC), higher auditory cortex (AC), and ventrolateral PFC are well-predicted by speech semantics. **(b)** Same as **(a)** for vision semantics. **(c)** We visualize the prediction performance of speech and visual semantic feature spaces across voxels on the cortical surface using a two-dimensional colormap. There is a high degree of separation in the regions that are well-predicted by each feature space. Voxels for which the visual semantics predict well and the speech semantics predict poorly are colored in orange. Blue corresponds to voxels for which the speech semantics predict well and the visual semantics predict poorly. Black corresponds to low prediction performance in both feature spaces. White corresponds to high prediction performance in both feature spaces.

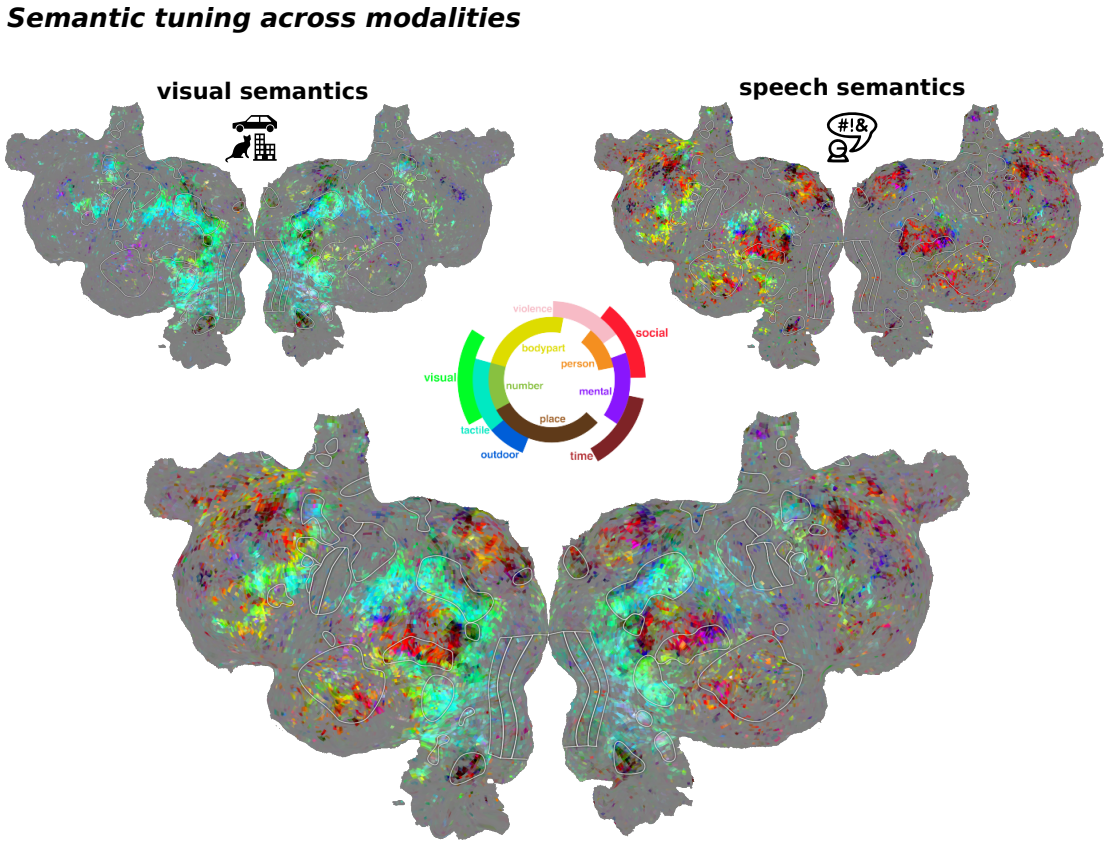


**Figure 3.9: Semantic tuning in vision and speech.** We extracted the weights for each of the visual and speech semantics feature spaces from the joint model. Each feature space contained a 985-dimensional weight vector per voxel. We projected the weight vectors from each feature space into a semantic space derived from a previous study where participants listened to spoken stories (Huth et al., 2016). **(a)** The semantic feature weights learned from the spoken words in the short films match the previous results. A high degree of specificity is found in regions of the semantic system in lateral and medial parietal cortex (LPC, MPC), superior temporal gyrus (STG) and ventrolateral prefrontal cortex. **(b)** The visual semantic feature weights from all voxels project into similar parts of the semantic space corresponding to visual and tactile concepts. The regions are located in higher visual cortex (RSC, PPA, OPA, FFA, EBA) and interparietal sulcus (IPS). **(c)** We computed the average of the semantic weights in vision and speech. We then projected the resulting voxel weight vectors into the same semantic space. Note that is little overlap in the regions encoding each feature space and so the average projection looks like the overlay of both maps.

the visual concepts represented in these regions project onto the same part of the semantic space derived from the language domain.

### 3.4.3 Feature space selectivity boundaries across the cortical surface

#### 3.4.3.1 Boundary between early and anterior visual cortex

Beyond early visual cortex, areas specialized in the visual processing of object categories have been found in occipitotemporal and inferior temporal cortices (see Grill-Spector, 2003, for an overview). These include face selective areas such as the fusiform face area (FFA; Kanwisher et al., 1997) and the occipital face area (OFA; (Halgren et al., 1999), and scene selective areas such as the occipital place area (OPA; Nakamura et al., 2000, Hasson et al., 2003, Dilks et al., 2013), the parahippocampal place area (PPA; Epstein and Kanwisher, 1998) and retrosplineal cortex (RSC; Maguire, 2001). The scene selective regions RSC and PPA are located directly anterior to the visual far periphery. RSC is typically found on the dorsal medial wall of occipital cortex (Maguire, 2001), anterior to the posterior-occipital sulcus that separates parietal and occipital cortices (Ono et al., 1990). PPA is typically located ventro-medial to the collateral sulcus. Previous studies have found subdivisions within RSC and PPA that do not correspond to low-level visual features (Çukur et al., 2016). However, previous studies could not explore cortical activity at the far visual periphery. We therefore sought to examine the degree of overlap between low-level (motion-energy) and high-level (semantics) visual feature representations in the the visual system with particular interest on the boundary between early and anterior visual cortex.

We first computed the prediction performance of motion energy versus visual semantic features from the estimated joint voxelwise model separately (Figure 3.5, red and green). The prediction performance for each of the motion energy and visual semantics feature spaces is shown on the cortical surface for one subject (Figure 3.10a,b). The motion energy feature space accurately predicts voxel activity throughout early visual cortex and hMT+. The visual semantic model accurately predicts voxel responses in anterior visual cortex.

In order to directly examine the degree of overlap in the representation of low- and high-level visual features, we visualize the prediction performance of both the motion energy and the visual semantic feature spaces simultaneously on the cortical surface of each individual subject (Figure 3.10c). There is a clear separation between early and anterior visual cortex. Early visual cortex is better predicted with motion-energy features and anterior visual cortex is better predicted by visual semantic features. There is little overlap in prediction performance except in RSC and PPA in two subjects (TZ, AH). However, most voxels within these two regions are better predicted by the visual semantic features. This suggests that representations of

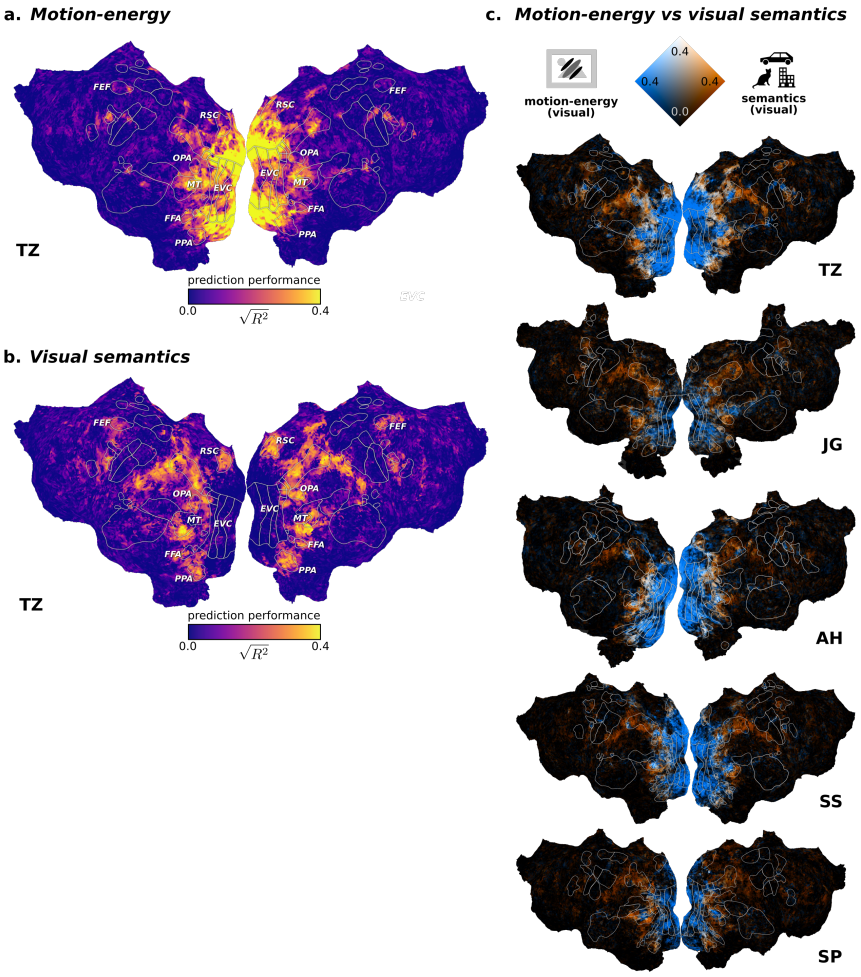


**Figure 3.10: Prediction performance of motion-energy and visual semantics at the boundary of early and higher visual cortex.** The estimated joint model was used to compute prediction performance for each of the motion-energy and visual semantic feature spaces separately. **(a)** As expected, early visual cortex (EVC) and hMT+ are well-predicted by motion-energy features. **(b)** Visual semantic features predict well in higher visual cortex (e.g RSC, OPA, FFA, PPA) and IPS. **(c)** We visualize the prediction performance for each of the motion-energy and visual semantics feature spaces on the cortical surface using a two-dimensional colormap. There is a high degree of separation between regions well-predicted by the motion-energy and the visual semantics feature spaces. White corresponds to voxels where both feature spaces provide high prediction performance on their own. Blue (orange) corresponds to high motion-energy (visual semantics) and low visual semantics (motion-energy) prediction performance. Black corresponds to regions where both models have low prediction accuracy.

low-level motion energy features and high-level visual semantic features are largely separate in the visual system.

#### 3.4.3.2 Boundary between primary and greater auditory cortex

Beyond primary auditory cortex, the superior temporal gyrus (STG) in humans has been shown to process complex auditory content such as natural (Moerel et al., 2013) and environmental (Lewis et al., 2005, 2011, 2012) sounds, music and speech (Norman-Haignere et al., 2015). The boundary and precise functional organization of auditory cortex beyond PAC is not well understood in humans (Saenz and Langers, 2014). Recent work has shown evidence that STG is involved in the processing of various aspects of speech (de Heer et al., 2017). We explored whether we could functionally divide auditory cortex into PAC and speech-selecitive STG. To do this, we compared the prediction performance of low-level auditory feature spaces (spectrogram and WaveNet) against features derived from speech (semantics, syntax, word rate).

We found the maximum prediction performance for the spectrogram and wavenet feature spaces (Figure 3.5 yellow) and compare it against the maximum prediction performance across all feature spaces derived from speech (Figure 3.5 pink). We show both of these values values for each voxel on the cortical surface simultaneously (Figure 3.11). There is a clear separation between PAC and STG where PAC is selective for low-level auditory features (spectrogram and WaveNet) and STG is selective for speech releated features. Regions at the border of PAC and STG are selective for both feature spaces (low-level auditory and speech) though this can be due to voxel bleed-over (Gao et al., 2015). This result suggest that the short films can be used to divide auditory cortex into PAC and speech-selective STG.

### 3.4.4 Functional subdivisions of human middle temporal cortex encoding for visual thematics, semantics and motion-energy

Work in non-human primates has shown that regions within middle temporal cortex are selective for visual motion (Albright, 1984, Britten et al., 1992, Nishimoto and Gallant, 2011). The human homologue is hMT+. It is located in posterior inferior temporal sulcus (Huk et al., 2002), is highly mylenated (Van Essen et al., 2013, Sereno et al., 2013), and damage to this region can lead to deficits in motion perception (Hess et al., 1989).

The functional organization of the region surrounding hMT+ is less understood. Studies in humans have shown that regions around hMT+ show greater responses to

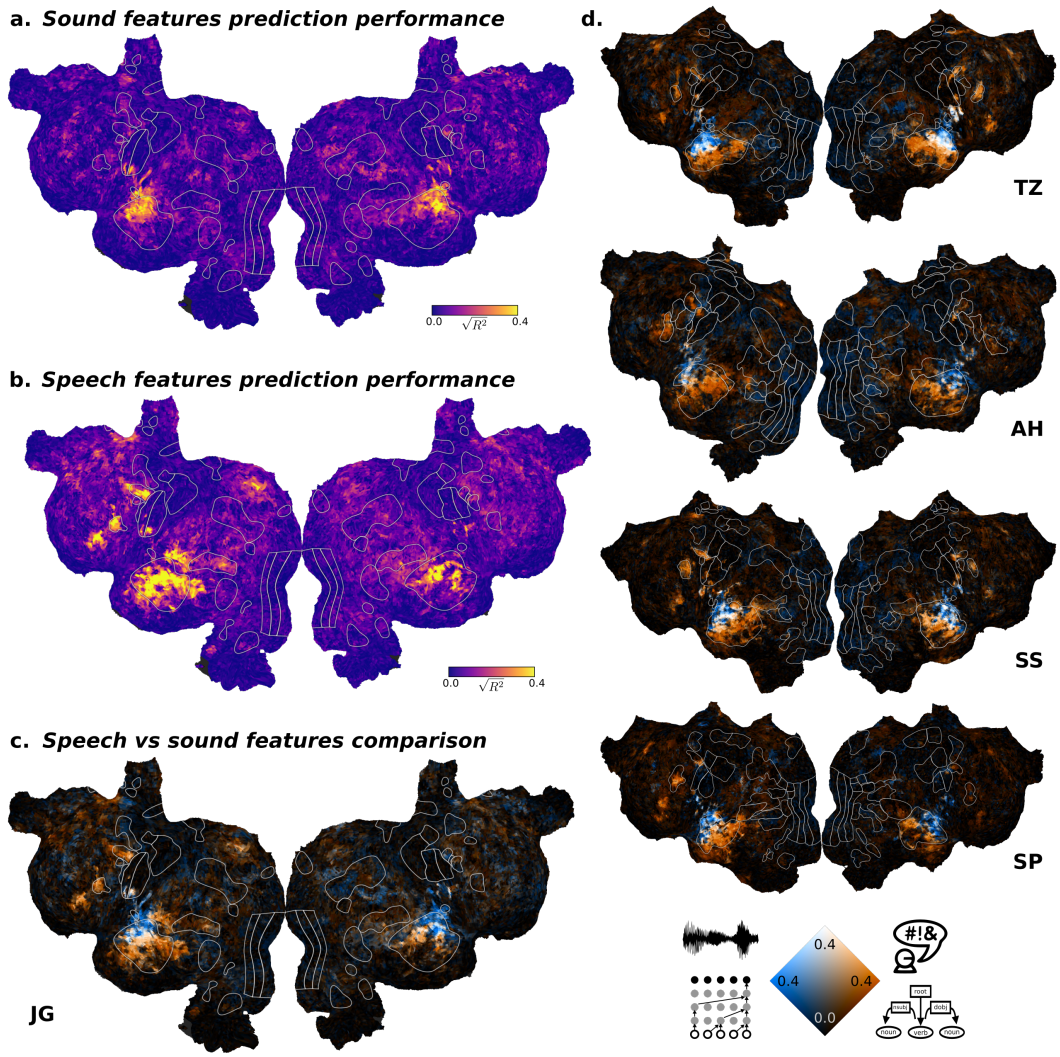


**Figure 3.11: Prediction performance comparison of low-level sound features and high-level speech features in primary auditory cortex and STG. (a)** The maximum prediction performance was identified for low-level sound features (spectrogram and WaveNet). The activity of voxels located in primary auditory cortex (PAC) is well predicted by these features. **(b)** The maximum prediction performance was also identified across all high-level speech feature spaces (semantics, syntax, word rate). Voxel activity in superior temporal gyrus (STG) is well predicted by these high-level speech features. The prediction performance for all cortical voxels performance is shown on the cortical surface for one subject. **(c,d)** A direct comparison of prediction performance shows that voxels located in PAC are best predicted by the low-level sound features across all five subjects. On the other hand, activity of voxels in STG best predicted by high-level speech features.

images of body parts than to images of other objects (Downing et al., 2001). This region is referred to as the extrastriate body area (EBA). It is currently unknown whether EBA is a homogeneous functional area or whether it consists of multiple functional areas (Weiner and Grill-Spector, 2013). Previous work based on anatomical landmarks suggests that EBA can be subdivided into at least three distinct areas (Weiner and Grill-Spector, 2011). One is located posterior to hMT+ in the lateral occipital sulcus; one ventral to hMT+ in the inferior temporal gyrus; and one anterior to hMT+. However, no functional specialization beyond the preference for body part images has been found in these three areas surrounding hMT+.

We sought to find whether the regions surrounding hMT+ are specialized for the processing of particular types of visual information beyond body parts. To do this, we explored the functional selectivity of regions surrounding hMT+ across all visual feature spaces in individual subjects. We find that there is a consistent pattern of functional specialization in regions surrounding hMT+. In particular, a region ventral to hMT+ is best predicted by the visual semantics feature space. This feature space has been shown to capture selectivity for body parts and the actions that those body parts are involved in (Huth et al., 2012). Furthermore, a region anterior to hMT+ is best predicted by visual thematics. The visual thematics feature space captures information about the relationship between objects and the types of events that occur in the visual scene. These results suggest that regions surrounding hMT+ can be divided into at least two functional areas: a ventral area (ventral EBA in 3/5 subjects) involved in representing visual semantics and an anterior area (dorsal EBA in 4/5 subjects) involved in visual thematics. The results are highly consistent and present in all five individual subjects.

#### 3.4.4.1 A region anterior to hMT+ is best predicted by visual thematics

Previous work has shown that visual semantic features can accurately predict voxel responses in anterior visual cortex including EBA and also that voxels in EBA are tuned for the presence of body parts in the visual scene (Huth et al., 2012) consistent with earlier findings (Downing et al., 2001). The visual semantic feature space consists of labels for the object and action categories that are present in the visual scene. These features do not capture the relationship between objects and action categories, nor the types of events that occur in the visual scenes. We hypothesized that information about the relationship between objects and action categories, and the types of events that occur in the visual scene might be a useful representation for the brain. This information can capture complex relationships between visual objects and categories such as biological motion ("Joe dances raegetton"), affordances

("Joe throws the ball"), human movements ("Joe drinks water").

In order to capture information about the relationship between objects and actions and the types of events that occur in the visual scenes, we labeled every second of video in the short films using the VerbNet verb lexicon (Kipper et al., 2006). VerbNet is based on the linguistic concept of thematic roles (Fillmore, 1968). Thematic roles can provide a more complex representation of a visual scene by capturing "Who (*actor*) does what (*verb class*, and *verb category*) to whom (*undergoer*) where (*location*), and by means of what (*selective constraints*)?" The verb category features capture the type of event occurring in the visual scene. For example, a visual scene depicting "Joe eats an apple" or "Joe drinks water" are both a type of "ingesting" event and so are labeled with the verb category "verbs of ingesting". Verb categories allow a higher level of abstraction than the specific verb ("eat.v.01", "drink.v.01") and verb class ("eat-18.1") that are present in the visual scenes. We refer to the feature spaces extracted from VerbNet as visual thematics (see Section 3.3.2.3 for more details).

We computed the prediction performance of the visual thematics feature space alone using the weights estimated from the joint model (Figure 3.12a). We find that a region anterior to hMT+ is well predicted by visual thematics. This region is located in the dorsal part of the extrastriate body area (EBA) as defined in a separate localizer experiment in 4/5 subjects (bodies > objects; Downing et al., 2001). We then compared the prediction performance of the visual thematics feature space against each of motion-energy and the semantic feature spaces (Figure 3.12b and c, respectively). We find that the region anterior to hMT+ is better predicted by visual thematics than motion-energy features. We also find that hMT+ is better predicted by motion-energy features than visual thematics (Figure 3.12b). We then compared visual thematics against visual semantics. We find that the same region anterior to hMT+ is better predicted by visual thematics than visual semantics (Figure 3.12c). Furthermore, we find that a region ventral to hMT+ is better predicted by visual semantics than visual thematics. This ventral region corresponds to the ventral part of EBA in 3/5 subjects.

In order to evaluate whether visual thematics was the best feature space for the region located anterior to hMT+, we compared the prediction performance of the visual thematics feature space against both visual semantics and motion-energy in each individual subject. For each voxel, we select the maximum prediction performance obtained from either the motion-energy or visual semantics feature space. We then compared this maximum against the prediction performance of the visual thematic feature space for each voxel in individual subjects. A clear pattern emerges in every subject (Figure 3.12d). The region anterior to hMT+ corresponding to the dorsal part of EBA in 4/5 subjects is better predicted by the visual thematics fea-

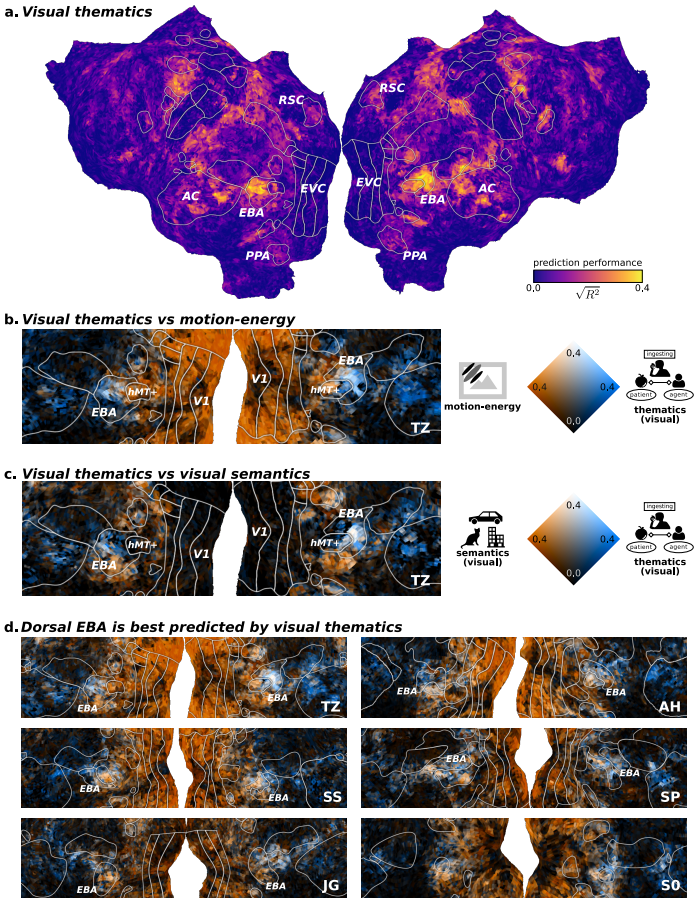


**Figure 3.12: Comparison of visual thematics against motion-energy and visual semantics in middle temporal cortex.** **(a)** The estimated joint model was used to compute prediction performance for the visual thematics feature space. This is shown for one subject on the cortical surface. A region anterior to hMT+ is well predicted by visual thematic features. This region corresponds to the dorsal part of the independently defined extrastriate body area (EBA; body parts > objects) in 4/5 subjects. The prediction performance for of visual thematics against motion-energy **(b)**, and visual thematics against visual semantics **(c)** is shown for one subject using a two-dimensional colormap. **(b,c)** Visual thematic features predict a region anterior to hMT+ better than either of the other two feature spaces (blue). Voxels in white are well predicted by both feature spaces, and voxels in black are not predicted by either. **(d)** The highest prediction performance across motion-energy and visual semantic feature spaces for each voxel in regions surrounding hMT+ **(b,c)** is compared against against visual thematics . There is a high degree of separation between the visual thematics selective region located anterior to hMT+ and the rest of early and higher visual cortex. This is highly consistent across all five subjects.

ture space than either of the other twofeature spaces in every subject. These results suggest that an area located anterior to hMT+ is functionally specialized for visual thematics.

#### 3.4.4.2 Motion energy and visual semantics are the second best feature spaces for voxels in middle temporal cortex selective for visual thematics

In order to test whether important feature spaces for regions surrounding hMT+ are missing from our analyses, we identified the second best predictive feature spaces from the set of all feature spaces (Figure 3.1). We first selected all the voxels where the visual thematic feature space provided the highest prediction performance relative to all other feature spaces. We then selected only the significant voxels from this set ($q(FDR) < 0.05$ for all subjects). For each voxel in the cortical surface, we show the prediction performance of the visual thematics against the second best feature space (Figure 3.13a). Voxels surrounding hMT+ are also well-predicted by the second best feature spaces. Regions posterior to medial S1 in anterior medial parietal cortex are not well predicted by the second best model. This result suggests that non-visual thematic feature spaces provide consistently lower yet comparable prediction performance for voxels surrounding hMT+.

We next plotted the prediction accuracy and identity of the second best feature space on the cortical surface (Figure 3.13b). We focused on the region surrounding hMT+ in all subjects (Figure 3.13c). We find that the second best feature spaces for visual thematics selective voxels in regions surrounding hMT+ are motion-energy and visual semantics for each individual subject. The motion-energy features tend to be the second best features for voxels closer to hMT+. Conversely, visual semantics tend to be the second best feature space for voxels located farther from hMT+. These results suggest that in the region surrounding hMT+, the relevant feature spaces are visual thematics, visual semantics and motion-energy.

#### 3.4.4.3 Prediction performance of thematic roles, semantics and motion-energy feature spaces in middle temporal cortex

In order to discover whether more functional subdivisions are present in middle temporal cortex, we next examine the functional selectivity of each voxel for the visual feature spaces. We visualized the prediction performance for each feature space using an RGB color space (Figure 3.14). The prediction performance of the visual semantic feature space was colored in red. The prediction performance of the visual thematic and motion-energy feature spaces was colored in green and red,

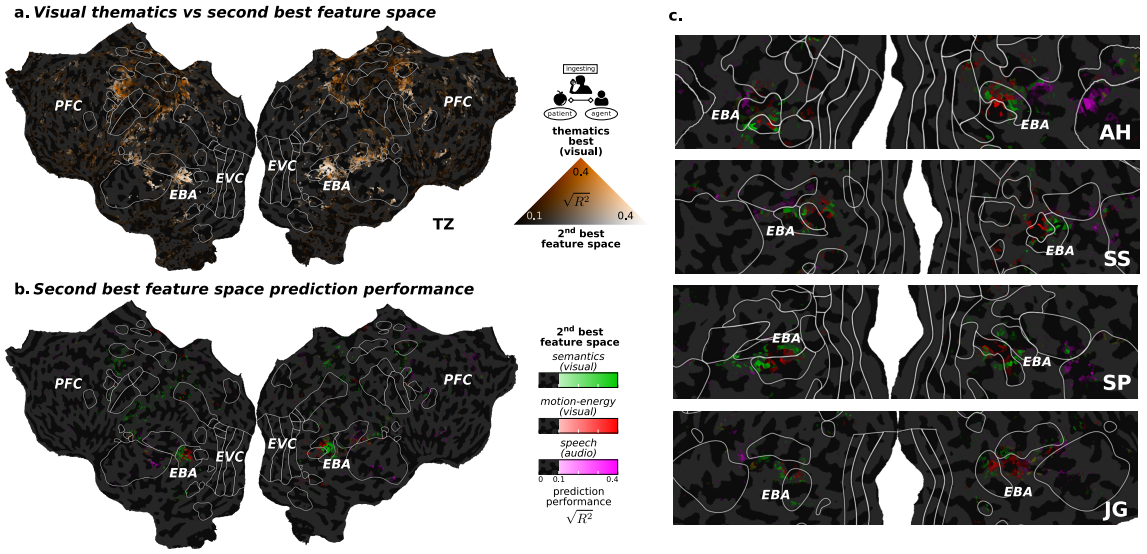


**Figure 3.13: Second best predictive feature space for voxels best predicted by visual thematics. (a)** The compared visual thematics against the second best feature space. We selected all voxels where the visual thematics feature space provided the highest significant prediction performance ($q(FDR) < 0.05$). The results are shown for one subject on their cortical surface. The second best model also predicted well in voxels around EBA. **(b)** We next colored each voxel according to which feature space provided the second best prediction performance. For voxels close to hMT+ the second best feature space is motion-energy (red) and for voxels farther it is visual semantics (green). **(c)** Same as **(b)** for all other subjects.

respectively. Voxels that are well predicted by both motion-energy (blue) and visual thematics (green) are displayed in cyan. Voxels that are well predicted by both visual semantics (red) and thematics (green) are colored in yellow.

We find a clear pattern of feature space selectivity in regions surrounding hMT+ in each individual subject (Figure 3.14a). The region anterior to hMT+ corresponding to dorsal EBA in 4/5 subjects is better predicted by visual thematics (yellow-greenish regions), and a region ventral to hMT+ is better predicted by visual semantics (red-yellowish). Finally, hMT+ itself is best predicted by motion energy features. The pattern of selectivity in individual subjects is very consistent. These results suggests that the putative body part selective regions surrounding hMT+ are functionally specialized. In particular, it suggests that an area anterior to hMT+ is specialized in the representation of visual thematic features and a region ventral to hMT+ is specialized in the representation of visual thematics (Figure 3.14b)

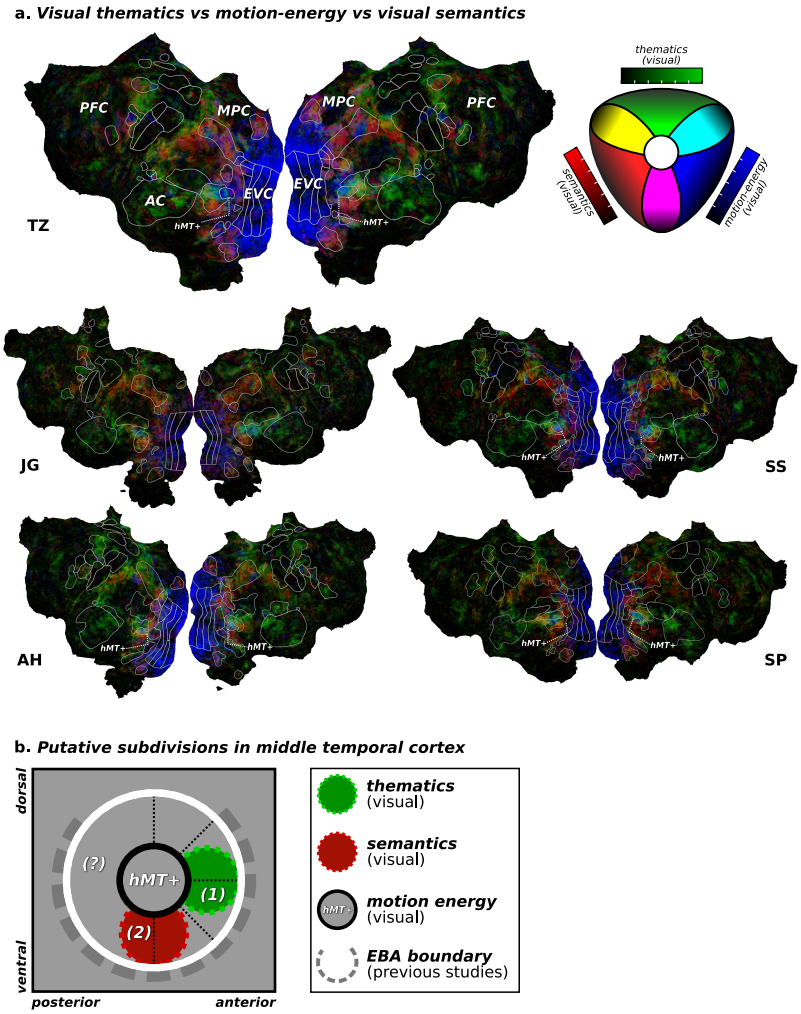


**Figure 3.14: Comparison of all visual feature spaces for all subjects.** **(a)** The model performance of each of the visual feature spaces (motion energy, visual semantics, and visual thematics) are mapped to each subject's flattened cortical surface. Blue voxels are best predicted by the motion-energy features and are located in the early visual cortex. Red voxels are best predicted by the visual semantics features. Green voxels are best predicted by the visual thematics features are located in several semantically selective regions, including a region anterior to hMT+ corresponding to the dorsal part of the extrastriate body area (EBA) in 4/5 subjects. Voxels in high-level visual regions such as FFA, PPA, RSC, and voxels at the boundary of early visual cortex are well predicted by both the motion energy features and the visual semantics features. These voxels are depicted in purple and are consistent across subjects. **(b)** Putative subdivision in middle temporal cortex based on the results.

## 3.5 Discussion

In this chapter, we demonstrated that naturalistic viewing of short films is a powerful paradigm that can be used to discover brain representations across vision, audio and speech from a single experiment. We presented a new voxelwise modeling approach for combining more than a dozen feature spaces into a single joint model. This joint model allowed us to predict voxel responses to 27 minutes of novel short films and to recover rich functional maps from individual subjects. These functional maps characterize both the types of feature spaces that are important across the cortical surface and how these feature spaces are represented in cortical areas.

We found that feature space representations are localized to broad continuous regions of the cortical surface. Early visual cortex is well predicted by motion energy features, anterior visual cortex by visual semantics and thematics, STG, the lateral and medial parietal cortex, ventrolateral PFC, and superior frontal gyrus by speech features, and primary auditory cortex by sound features. These results are highly consistent across subjects. We can also recover known maps that capture how the feature spaces are represented within cortical regions.

We were able to recover retinotopic maps from naturalistic free viewing of short films. To our knowledge, this has not been achieved before. The recovered retinotopic maps are still coarse because the spatial receptive fields (RF) estimated from the motion energy features are poor. We can improve the quality of the recovered retinotopic maps by obtaining better estimates of the spatial RF based on previous approaches used in fixated viewing of naturalistic silent movies (Nishimoto et al., 2011). Nevertheless, even with poor RF estimation, we were able to show high eccentricity tuning for voxels in the far periphery. Previous work has mapped out the far visual periphery with a customized setup that involves positioning the stimulus screen 10-12cm away from the subject (Pitzalis et al., 2006). The naturalistic viewing of short films is a much easier way to map the far periphery and recovers both hemispheres. Furthermore, the short films paradigm allows us to model low-level (motion energy) and high-level (semantics) visual responses at the border of the far periphery and anterior visual cortex at the same time which eliminates spatial alignment issues across imaging sessions.

We were also able to recover tonotopy maps from primary auditory cortex (PAC) in individual subjects. The quality of the tonotopy maps is better than those estimated from a separate 40 minute tonotopy experiment (data not shown). We can clearly differentiate A1 and R from PAC using the tonotopic maps derived from the short films (Merzenich and Schreiner, 1992). Interestingly, a region posterior to Heschl sulcus in the lateral belt was found to be selective for low frequencies in 8/10 hemispheres. This region might correspond to the secondary belt region called CL

(Talavage et al., 2004, Humphries et al., 2010, Moerel et al., 2014) though the functional organization of auditory cortex outside of A1 and R is not well understood in humans (Saenz and Langers, 2014). The speech, environmental sounds, and music contained in the short films produces high functional SNR throughout auditory cortex. This provides a good dataset to test feature spaces that might capture the types of representations that exist in auditory cortex.

### 3.5.1 Two regions surrounding hMT+ are functionally distinct

Previous research has shown that regions surrounding hMT+ are involved in the representation of body parts (Downing et al., 2001, Weiner and Grill-Spector, 2010). However, the precise functional organization of this extrastriate body area (EBA) surrounding hMT+ is not well understood (Weiner and Grill-Spector, 2013). Work based on anatomical landmarks has argued that the region surrounding hMT+ can be subdivided into three distinct body selective areas and is not a single homogeneous EBA (Weiner and Grill-Spector, 2011). The three distinct body part selective areas lie posterior, anterior, and ventral to hMT+. However, even if anatomically distinct, there is little evidence to suggest that these three areas are functionally distinct.

We found that an area anterior to hMT+ is best predicted by the visual thematics feature space in all our subjects. This area corresponds to the dorsal part of EBA in 4/5 subjects. The visual thematics feature space captures information about the relationship between objects and the types of events in the visual scene. This is different from visual semantics which only capture the object and action categories present in the scene, not their relationship. This suggests that the dorsal part of EBA represents information beyond the mere presence of visual body parts. We also found that an area ventral to hMT+ is best predicted by visual semantics and it corresponds to ventral EBA in 3/5 subjects. Finally, we do not find consistent results across subjects regarding the type of information that the area posterior to hMT+ represents.

Taken together, our results suggests that two areas one anterior and one ventral to hMT+ might be functionally distinct. However, our analyses do not reveal what specific features are represented within these areas. The functional tuning of the area anterior to hMT+ could be related to biological motion, particular body movements, body part interactions, or more complex feature combinations. In future work, we will explore the representation of the visual thematic features within this area in order to aid our interpretation of its functional role. We also plan to find an optimal set of stimuli that maximally differentiates activity in these two areas. These stimuli

will then be used to create an experiment to functionally localize the two areas.

### 3.5.2 Variance partitioning and multiple feature space representations

The results in this chapter are presented in terms of which feature space best predicts individual voxel responses. However, the model used explicitly combines multiple feature spaces to predict voxel responses and it performs significantly better than any individual feature space alone (except for motion energy in early visual cortex and hMT+). We did not explore whether there exist cortical areas that are significantly involved in representing multiple feature spaces. In order to answer this question, a variance partitioning analysis needs to be conducted (Lescroart et al., 2015, de Heer et al., 2017). However, current methods make it intractable to conduct variance partitioning analysis with 17 feature spaces. It would require estimating all $2^{17}$ unique models one for each combination of feature spaces (de Heer et al., 2017).

We can estimate unique and shared variance for each feature space by combining different approaches. First, we can measure how much unique variance one individual feature space explains by estimating the full joint model and a separate joint model that is missing only one feature space. The difference between the full joint model and this "knock out" model is an estimate of how much unique explained variance can be attributed to one feature space. The unique variance component can be estimated as the difference between the joint model and the knock out model. The shared variance component can be estimated by subtracting the unique variance component from a model that only includes one feature space. A model that only includes a single feature space is able to capture both the unique variance of that feature space and the variance shared with all other feature spaces. By subtracting the unique variance component from the full variance component, we can estimate the shared variance component. We will develop this new variance partitioning analysis in future work.

### 3.5.3 Related work

Naturalistic viewing of movies and listening of stories is increasing in popularity. In some approaches, model free analysis are used to make inferences about brain representations and their consistency across subjects (Hasson et al., 2004). This makes it difficult to infer functional interpretations to those data alone. Similarly, recent work has explored the used of movies for the blind for exploring brain representations (Hanke et al., 2014). These data are limited in brain coverage (not whole brain), data quality (unknown functional SNR), number of modalities explored (only audio), and labeled feature spaces (only one rich annotation is). More recently, model-free

methods have been used to recover event temporal structure boundaries from films (Baldassano et al., 2017). However, this is only one aspect of the film content. We have shown that our short film paradigm combined with our modeling framework and data is powerful enough to recover functional maps across multiple feature spaces in individual subjects. Our paradigm has the advantage of being ever-expandable. New feature spaces can be included and compared then to existing ones in terms of prediction performance. This is approach is particularly important because the validity of findings in neuroimaging is increasingly questioned (Goodman et al., 2016).

### 3.5.4 Replicability and generalization in every subject

The voxelwise encoding model approach used in this chapter was developed with a focus on individual subject results (Wu et al., 2006, Naselaris et al., 2011). The logic behind voxelwise modeling is that by analyzing every subject individually and reporting every subjects results, the reader is able to see that the vast majority of the results are replicated for every subject. In this chapter, the rich functional maps that were derived for each subject show a very consistent pattern across all subjects. In effect, every single subjects is a separate individual experiment. Every one of our results is assessed in terms of how well models can generalize and predict a new dataset of brain responses collected while subjects watch completely different stimuli. And every result is evaluated at the individual subject level. For these reasons, our results pass a much higher statistical bar than most fMRI studies that are commonly based on summary statistics computed on a single dataset on group-averaged data.

### 3.5.5 Observations

There are a number of observations that will be pursued in future work.

*Dorsal parietal – auditory features*: A region in dorsal parietal cortex is weakly but significantly predicted by sound features in 4/5 subjects. This region has been reported to be selective for auditory information (Sood and Sereno, 2016) and in the representation of human sounds (Brefczynski-Lewis and Lewis, 2017). Currently, the environmental sound categories are not labeled. A feature space build from environmental sounds will allow us to evaluate whether this region is indeed selective for auditory features or particular classes of environmental sounds.

*Posterior dorsal medial parietal – auditory features*: A region in medial parietal cortex is weakly but significantly predicted by sound features in 7/10 hemispheres. This region lies posterior dorsal to semantically selective regions in medial parietal cortex.

*Anterior precuneus – visual thematics and motion energy features*: A region in anterior precuneus is best predicted by visual thematic features in all subjects. A region immediately posterior to it is best predicted by motion energy features in 7/10 hemispheres. Previous studies have reported a retinotopic region in anterior precuneus (aPCu; Huang and Sereno, 2018). Our preliminary retinotopy analysis could not recover visual angle selectivity in this region. However, our results suggest that the anterior-most region in precuneus might be functionally different from the region immediately posterior to it which is better predicted by motion energy features.

*Posterior central sulcus – visual semantics and thematics*: The regions immediately posterior to primary somatosensory cortex (S1) are well predicted by visual semantics in the majority of subjects. Interestingly the semantic tuning of these regions maps onto the same part of the semantic space derived from a language experiment (Huth et al., 2016). The semantic tuning for visual semantics in these regions is likely more diverse than the language semantic space can captures. A previous exploring semantic representations in silent movies did not find consistent semantic tuning across subjects (Huth2012). This is likely due to the low prediction performance in these region in 3/5 subjects. In future work, we will explore the visual semantic tuning of this region.

*Speech thematics – preliminary observations*: Speech thematics were labeled in the short films but the results are not shown in this chapter. Preliminary observations show that the thematic role labels alone do not significantly outperform other speech models. However, semantic models built from words appearing in a specific thematic role (e.g. constructing a semantic features using only words appearing in the actor role) predict activity in a region of ventrolateral PFC (dorsal anterior to Brocas area) and the other speech features do not. This suggests that semantic representations are affected by the particular thematic role that the words occupy in the sentence.

*Speech syntax*: We have not explored the individual effects of syntactic part of speech or syntactic word dependency labels.

## 3.6 Future directions

### 3.6.1 Multiview auto-encoder

The short films are rich audio-visual stimuli that can be used to recover functional maps from individual subjects, and we have shown here that the associated brain activity is highly similar across subjects. Further, in a separate project, we have built a multiview autoencoder network to estimate this joint brain representation across

subjects. This autoencoder is able to learn a mapping between the activity of the different subjects despite anatomical differences, and between stimulus feature spaces and subjects brain activity. The autoencoder can be used used to predict subjects brain activity from other subjects (functional alignment; Haxby et al., 2011, Bilenko and Gallant, 2016) and also from novel stimulus features corresponding to different experiments. The multiview autoencoder network can predict brain activity in a story listening experiment (Huth et al., 2016), in a silent movie watching experiment (Nishimoto et al., 2011) and in a novel multi-task experiment, after training only on the short films dataset. Short films are therefore rich enough to allow the autoencoder to learn the mapping between subjects brain activity that generalizes to novel experiments in visual, auditory and cognitive modalities. Our next step is to find the minimum subset of short films that is required for a new subject to be integrated in the pre-trained autoencoder. From this minimal subset of data, a mapping of the new subjects brain activity to other subjects will be learned. Then, the new subjects brain activity under all the other experiments can be predicted from the existing subjects. This means that multiple rich functional maps for the new subject can be constructed at the expense of only the minimal set of short films.

### 3.6.2 Human Connectome Project 7T film data

We collaborated with the Human Connectome Project (HCP) and provided visual semantic and stimulus motion-energy feature spaces for one hour of films (different from the ones used in our study). More than 150 subjects watched and listened to these films while their brain activity was recorded with a 7T MRI scanner. In future work, we hope to expand the feature spaces labeled in those films in order to explore our results in a much larger population. In addition, the HCP provides genetic, behavioral, and psychological measures of subjects. This dataset might prove useful in understanding how differences in brain activity across subjects relates to individual subject variation.

### 3.6.3 Visual imagery

In additional work not included in this chapter, we collected data while subjects listened to the audio from the short films with their eyes closed. Subjects were instructed to visualize the video content of eight short films while their brain activity was recorded using fMRI. We then used the same feature spaces that were used in the main viewing condition to predict brain activity. Because the subjects were imagining the films alongside the soundtrack, we could use the features with the same timing as the viewing condition. Preliminary results suggest that anterior visual cortex is

engaged during this task. We are able to predict activity in regions like retrosplineal cortex (RSC) and occipital place area (OPA) from the visual semantic features of the films that were being imagined, without any visual input. However, we were not able to predict activity in early visual cortex from motion-energy nor visual semantic features.

We also found that lateral parietal cortex was well predicted by the visual semantic features during the imagery condition. This is interesting because in the viewing condition (and in previous experiments; Huth2016) these regions are well predicted by speech semantics instead. However, the speech content is the same during both the visual imagery and viewing tasks because the full auditory content of the short films is played to the subjects. These preliminary results (not shown) suggest a possible task dependent recruitment of these association regions. During imagery, the task of imaging the short films video content recruits these association regions and so they become more involved in representing visual semantics. We will pursue this hypothesis in future work.

# Bibliography

E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284, 1985.

G. K. Aguirre, E. Zarahn, and M. D'esposito. The variability of human, BOLD hemodynamic responses. *NeuroImage*, 8(4):360–9, Nov. 1998.

T. D. Albright. Direction and orientation selectivity of neurons in visual area MT of the macaque. *Journal of Neurophysiology*, 52(6):1106–1130, 1984.

D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins. Globally Normalized Transition-Based Neural Networks. 2016. URL `http://arxiv.org/abs/1603.06042`.

Anonymous. Notes of the berlin physiological society (report of h. munk's presentation of july 27, 1883). *Nature*, 28:431–432, 1883.

C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, and K. A. Norman. Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, 95(3):709–721.e5, 2017.

N. Bazargani and A. Nosratinia. Joint maximum likelihood estimation of activation and Hemodynamic Response Function for fMRI. *Medical image analysis*, 18(5): 711–24, jul 2014.

J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.

J. Bergstra, D. Yamins, and D. D. Cox. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. *12th PYTHON IN SCIENCE CONF. (SCIPY 2013)*, (Scipy):13–20, 2013.

N. Y. Bilenko and J. L. Gallant. Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in Neuroinformatics*, 10:49, 2016.

P. Boersma and V. van Heuven. Speak and unSpeak with Praat. *Glot International*, 5(9-10):341–347, 2001.

D. Borcard, P. Legendre, and P. Drapeau. Partialling out the spatial component of ecological variation. *Ecology*, 73(3):1045–1055, 1992.

G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger. Linear systems analysis of functional magnetic resonance imaging in human v1. *Journal of Neuroscience*, 16(13):4207–4221, 1996.

J. A. Brefczynski-Lewis and J. W. Lewis. Auditory object perception: A neurobiological model and prospective review. *Neuropsychologia*, 105:223–242, 2017.

K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. a. Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 12(12):4745–4765, 1992.

P. Broca. Perte de la parole, ramollissement chronique et destruction partielle du lobe antérieur gauche du cerveau. *Bull Soc Anthropol*, 2(1):235–238, 1861a.

P. Broca. Remarques sur le siège de la faculté du langage articulé, suivies dune observation daphémie (perte de la parole). *Bulletin et Memoires de la Societe anatomique de Paris*, 6:330–357, 1861b.

A. Bushuev, S. A. Eugster, J.-B. Mardelle, R. Morton, V. Pinon, and the community. Kdenlive – kde non-linear video editor.

M. B. Cai, N. W. Schuck, J. W. Pillow, and Y. Niv. A bayesian method for reducing bias in neural representational similarity analysis. In *Advances in Neural Information Processing Systems*, pages 4951–4959. 2016.

M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. a. Olshausen, J. L. Gallant, and N. C. Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–97, nov 2005.

R. Casanova, S. Ryali, J. Serences, L. Yang, R. Kraft, P. J. Laurienti, and J. A. Maldjian. The impact of temporal regularization on estimates of the bold hemodynamic response function: a comparative analysis. *NeuroImage*, 40(4):1606–1618, 2008.

T. Çukur, A. G. Huth, S. Nishimoto, and J. L. Gallant. Functional subdomains within scene-selective cortex: Parahippocampal place area, retrosplenial complex, and occipital place area. *Journal of Neuroscience*, 36(40):10257–10273, 2016.

A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999.

J. G. Daugman. Spatial visual channels in the fourier plane. *Vision Research*, 24(9): 891–910, 1984.

V. De Angelis, F. De Martino, M. Moerel, R. Santoro, L. Hausfeld, and E. Formisano. Cortical processing of pitch: Model-based encoding and decoding of auditory fmri responses to real-life sounds. *NeuroImage*, 2017.

E. De Boer and P. Kuyper. Triggered correlation. *IEEE Transactions on Biomedical Engineering*, (3):169–179, 1968.

W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, and F. E. Theunissen. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27):6539–6557, 2017.

J. Diedrichsen and N. Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, 13(4):e1005508, 2017.

D. D. Dilks, J. B. Julian, A. M. Paunov, and N. Kanwisher. The Occipital Place Area Is Causally and Selectively Involved in Scene Perception. *Journal of Neuroscience*, 33(4):1331–1336, 2013.

A. Doicu, T. Trautmann, and F. Schreier. *Numerical regularization for atmospheric inverse problems*. Springer Science & Business Media, 2010.

P. E. Downing, Y. Jiang, M. Shuman, and N. Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–3, 2001.

R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.

S. O. Dumoulin and B. A. Wandell. Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2):647–660, 2008.

J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. 2017. URL `http://arxiv.org/abs/1704.01279`.

R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, apr 1998.

R. R. Fay, A. N. Popper, and D. B. Webster. *The evolutionary biology of hearing*. Springer-Verlag, 1992.

C. J. Fillmore. *The case for case*. 1968.

S. Finger. *Origins of neuroscience: a history of explorations into brain function*. Oxford University Press, USA, 2001.

E. Formisano, D.-S. Kim, F. D. Salle, P.-F. Van De Moortele, K. Ugurbil, and R. Goebel. Mirror-Symmetric Tonotopic Maps in Human Primary Auditory Cortex. *Neuron*, 40:859–869, 2003.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer, New York, NY, 2001.

K. J. Friston, K. J. Worsley, R. Frackowiak, J. C. Mazziotta, and A. C. Evans. Assessing the significance of focal activation using their spatial extent. *Human Brain Mapping*, 1:210–220, 1993.

K. J. Friston, O. Josephs, G. Rees, and R. Turner. Nonlinear event-related responses in fmri. *Magnetic resonance in medicine*, 39(1):41–52, 1998.

J. S. Gao, A. G. Huth, M. D. Lescroart, and J. L. Gallant. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9(September):1–12, 2015.

G. H. Glover. Deconvolution of impulse response in event-related bold fmri. *NeuroImage*, 9(4):416–429, 1999.

G. H. Glover, T. Q. Li, and D. Ress. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic Resonance in Medicine*, 44(1):162–167, 2000.

S. N. Goodman, D. Fanelli, and J. P. A. Ioannidis. What does research reproducibility mean? *Science*, 2016.

C. Goutte, F. A. Nielsen, and K. Hansen. Modeling the hemodynamic response in fmri using smooth fir filters. *IEEE transactions on medical imaging*, 19(12): 1188–1201, 2000.

K. Grill-Spector. The neural basis of object perception. *Current Opinion in Neurobiology*, 13(2):159–166, 2003.

E. Halgren, a. M. Dale, M. I. Sereno, R. B. Tootell, K. Marinkovic, and B. R. Rosen. Location of human face-selective cortex with respect to retinotopic areas. *Human Brain Mapping*, 7(1):29–37, jan 1999.

D. A. Handwerker, J. M. Ollinger, and M. D'Esposito. Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21(4):1639–1651, 2004.

M. Hanke, F. J. Baumgartner, P. Ibe, F. R. Kaule, S. Pollmann, O. Speck, W. Zinke, and J. Stadler. A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie. *Scientific Data*, 1:140003, 2014.

K. A. Hansen, K. N. Kay, and J. L. Gallant. Topographic organization in and near human visual area V4. *Journal of Neuroscience*, 27(44):11896–911, 2007.

P. C. Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, 1998.

U. Hasson, M. Harel, I. Levy, and R. Malach. Large-scale mirror-symmetry organization of human occipito-temporal object areas. *Neuron*, 37(6):1027–1041, 2003.

U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach. Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, 303:1634–1640, 2004.

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

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

R. H. Hess, C. L. Baker, and J. Zihl. The "motion-blind" patient: low-level spatial and temporal filters. *Journal of Neuroscience*, 9(May):1628–1640, 1989.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

R.-S. Huang and M. I. Sereno. *Multisensory and sensorimotor maps*, volume 151. Elsevier B.V., 1 edition, 2018.

A. C. Huk, R. F. Dougherty, and D. J. Heeger. Retinotopy and functional subdivision of human areas MT and MST. *Journal of Neuroscience*, 22(16):7195–205, aug 2002.

C. Humphries, E. Liebenthal, and J. R. Binder. Tonotopic organization of human auditory cortex. *NeuroImage*, 50(3):1202–1211, 2010.

A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

Itseez. Open source computer vision library, 2015. URL `https://github.com/itseez/opencv`.

M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. Fsl. *NeuroImage*, 62(2):782–790, 2012.

N. Kanwisher. Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, 107(25):11163–11170, 2010.

N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–11, jun 1997.

K. N. Kay, S. V. David, R. J. Prenger, K. A. Hansen, and J. L. Gallant. Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fmri. *Human Brain Mapping*, 29(2):142–156, 2008a.

K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008b.

S. M. Khaligh-Razavi, L. Henriksson, K. Kay, and N. Kriegeskorte. Fixed versus mixed RSA:Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology*, 76: 184–197, 2017.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. Extending VerbNet with novel verb classes. *Proceedings of LREC*, 2006(2.2):1, 2006.

N. Kriegeskorte and R. A. Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, 2013.

N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–8, mar 2006.

N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008a.

N. Kriegeskorte, M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P. A. Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008b.

M. D. Lescroart, D. E. Stansbury, and J. L. Gallant. Fourier power, subjective distance, and object categories all provide plausible models of bold responses in scene-selective visual areas. *Frontiers in Computational Neuroscience*, 9, 2015.

J. W. Lewis, J. a. Brefczynski, R. E. Phinney, J. J. Janik, and E. a. DeYoe. Distinct cortical pathways for processing tool versus animal sounds. *Journal of Neuroscience*, 25(21):5148–58, may 2005.

J. W. Lewis, W. J. Talkington, A. Puce, L. R. Engel, and C. Frum. Cortical networks representing object categories and high-level attributes of familiar real-world action sounds. *Journal of Cognitive Neuroscience*, 23(8):2079–101, aug 2011.

J. W. Lewis, W. J. Talkington, K. C. Tallaksen, and C. a. Frum. Auditory object salience: human cortical processing of non-biological action sounds and their acoustic signal attributes. *Frontiers in Systems Neuroscience*, 6(May):27, jan 2012.

E. Maguire. The retrosplenial contribution to human navigation: A review of lesion and neuroimaging findings. *Scandinavian Journal of Psychology*, 42(3):225–238, 2001.

P. Z. Marmarelis and K.-I. Naka. White-noise analysis of a neuron chain: an application of the wiener theory. *Science*, 175(4027):1276–1278, 1972.

G. Marrelec, H. Benali, P. Ciuciu, M. Pelegrini-Issac, and J.-B. Poline. Robust bayesian estimation of the hemodynamic response function in event-related bold fmri using basic physiological information. *Human Brain Mapping*, 19(1):1–17, 2003.

K. McLaren. XIII The Development of the CIE 1976 (L* a* b*) Uniform Colour Space and Colour-difference Formula. *Journal of the Society of Dyers and Colourists*, 92(9):338–341, 1976.

M. M. Merzenich and C. E. Schreiner. *Mammalian Auditory Cortex—Some Comparative Observations*, pages 673–688. Springer New York, New York, NY, 1992. ISBN 978-1-4612-2784-7.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

M. Moerel, F. De Martino, R. Santoro, K. Ugurbil, R. Goebel, E. Yacoub, and E. Formisano. Processing of Natural Sounds: Characterization of Multipeak Spectral Tuning in Human Auditory Cortex. *Journal of Neuroscience*, 33(29):11888–11898, 2013.

M. Moerel, F. De Martino, and E. Formisano. An anatomical and functional topography of human auditory cortical areas. *Frontiers in Neuroscience*, 8(8 JUL):1–14, 2014.

MPlayer Team. Mplayer–the movie player. URL `http:/www.mplayerhq.hu`.

H. Munk. *Über die Functionen der Grosshirnrinde: Gesammelte Mittheilungen aus den Jahren 1877-1880*. August Hirschwald, 1881.

K. Nakamura, R. Kawashima, N. Sato, A. Nakamura, M. Sugiura, T. Kato, K. Hatano, K. Ito, H. Fukuda, T. Schormann, et al. Functional delineation of the human occipito-temporal areas related to face and scene processing: a pet study. *Brain*, 123(9):1903–1912, 2000.

T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant. Encoding and decoding in fmri. *NeuroImage*, 56(2):400–410, 2011.

H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte. A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, 10(4), 2014.

S. Nishimoto and J. L. Gallant. A three-dimensional spatiotemporal receptive field model explains responses of area mt neurons to naturalistic movies. *Journal of Neuroscience*, 31(41):14551–14564, 2011.

S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

S. Norman-Haignere, N. G. Kanwisher, and J. H. McDermott. Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, 88(6):1281–1296, 2015.

T. E. Oliphant. Scipy: Open source scientific tools for python. *Computing in Science and Engineering*, 9:10–20, 2007.

M. Ono, S. Kubik, and C. D. Abernathey. *Atlas of the cerebral sulci*. Tps, 1990.

A. V. Oppenheim, A. Willsky, and I. Young. *Signals and systems.* Englewood Cliffs, NJ: Prentice-Hall, 1983.

S. E. Palmer. *Vision science: Photons to phenomenology*. MIT press, 1999.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

F. Pedregosa, M. Eickenberg, P. Ciuciu, B. Thirion, and A. Gramfort. Data-driven hrf estimation for encoding and decoding models. *NeuroImage*, 104:209–220, 2015.

W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Academic press, 2011.

S. Pitzalis, C. Galletti, R.-S. Huang, F. Patria, G. Committeri, G. Galati, P. Fattori, and M. I. Sereno. Wide-Field Retinotopy Defines Human Cortical Visual Area V6. *Journal of Neuroscience*, 26(30):7962–7973, 2006.

N. Rappin and R. Dunn. *wxPython in Action*. Manning Publications, 2006.

I. E. Richardson. *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*. John Wiley & Sons, 2004.

J. B. Ritchie, S. Bracci, and H. Op de Beeck. Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *NeuroImage*, 148 (December 2016):197–200, 2017.

M. Saenz and D. R. Langers. Tonotopic mapping of human auditory cortex. *Hearing Research*, 307:42–52, 2014.

O. Schoppe, N. S. Harper, B. D. Willmore, A. J. King, and J. W. Schnupp. Measuring the performance of neural models. *Frontiers in Computational Neuroscience*, 10, 2016.

M. Sereno, A. Dale, J. Reppas, K. Kwong, J. Belliveau, T. Brady, B. Rosen, and R. Tootell. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212):889–893, 1995.

M. I. Sereno, A. Lutti, N. Weiskopf, and F. Dick. Mapping the Human Cortical Surface by Combining Quantitative T1 with Retinotopy. *Cerebral Cortex*, 23(9): 2261–2268, 2013.

M. R. Sood and M. I. Sereno. Areas activated during naturalistic reading comprehension overlap topological visual, auditory, and somatotomotor maps. *Human Brain Mapping*, 37(8):2784–2810, 2016.

D. E. Stansbury, T. Naselaris, and J. L. Gallant. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5): 1025–1034, 2013.

T. Talavage, P. Ledden, M. Sereno, B. Rosen, and A. Dale. Multiple phase-encoded tonotopic maps in human auditory cortex. *NeuroImage*, 5(4 PART I), 1997.

T. M. Talavage, M. I. Sereno, J. R. Melcher, P. R. Ledden, B. R. Rosen, and A. M. Dale. Tonotopic Organization in Human Auditory Cortex Revealed by Progressions of Frequency Sensitivity. *Journal of Neurophysiology*, 91(3):1282–1296, 2004.

B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J.-B. Poline, D. Lebihan, and S. Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104–16, Dec. 2006.

B. Thirion, F. Pedregosa, M. Eickenberg, and G. Varoquaux. Correlations of correlations are not reliable statistics: implications for multivariate pattern analysis. In *ICML Workshop on Statistics, Machine Learning and Neuroscience (Stamlins 2015)*, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A. N. Tikhonov, V. I. Arsenin, and F. John. *Solutions of ill-posed problems*, volume 14. Winston Washington, DC, 1977.

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. pages 1–15, 2016. URL `http://arxiv.org/abs/1609.03499`.

D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, and K. Ugurbil. The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80: 62–79, 2013.

A. Walther, H. Nili, N. Ejaz, A. Alink, N. Kriegeskorte, and J. Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 2015.

A. B. Watson and A. J. Ahumada. Model of human visual-motion sensing. *Journal of the Optical Society of America*, 2(2):322–342, 1985.

L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS ONE*, 9(11):e112575, 2014.

K. S. Weiner and K. Grill-Spector. Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. *NeuroImage*, 52(4):1559–1573, 2010.

K. S. Weiner and K. Grill-Spector. Not one extrastriate body area: Using anatomical landmarks, hMT+, and visual field maps to parcellate limb-selective activations in human lateral occipitotemporal cortex. *NeuroImage*, 56(4):2183–2199, 2011.

K. S. Weiner and K. Grill-Spector. Neural representations of faces and limbs neighbor in human high-level visual cortex: Evidence for a new organization principle. *Psychological Research*, 77(1):74–97, 2013.

M. W. Woolrich, T. E. Behrens, and S. M. Smith. Constrained linear basis sets for hrf modelling using variational bayes. *NeuroImage*, 21(4):1748–1761, 2004.

K. J. Worsley, S. Marrett, P. Neelin, and A. C. Evans. Searching scale space for activation in PET images. *Human Brain Mapping*, 4(1):74–90, 1996.

M. C.-K. Wu, S. V. David, and J. L. Gallant. Complete Functional Characterization of Sensory Neurons By System Identification. *Annual Review of Neuroscience*, 29 (1):477–505, Jan. 2006.

J. Yuan and M. Liberman. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878, 2008.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005.

# Chapter 4

# Appendix to spatiotemporal encoding models

## 4.1 Standard form derivation

$$\begin{aligned}
\hat{\beta}_T &= \left(X^\top X + \lambda^2 C^\top C\right)^{-1} X^\top Y \\
C\hat{\beta}_T &= C\left(X^\top X + \lambda^2 C^\top C\right)^{-1} X^\top Y \\
C\hat{\beta}_T &= \left(X^\top X C^{-1} + \lambda^2 C^\top C C^{-1}\right)^{-1} X^\top Y \\
C\hat{\beta}_T &= \left(X^\top X C^{-1} + \lambda^2 C^\top\right)^{-1} X^\top Y \\
C\hat{\beta}_T &= \left(X^\top X C^{-1} + \lambda^2 C^\top\right)^{-1} C^\top C^{-1\top} X^\top Y \\
C\hat{\beta}_T &= \left(C^{-1\top} X^\top X C^{-1} + \lambda^2 I_p\right)^{-1} C^{-1\top} X^\top Y
\end{aligned}$$

Define $A = XC^{-1}$ and $\hat{\beta}_A = C\hat{\beta}_T$. The solution becomes

$$\hat{\beta}_A = \left(A^\top A + \lambda^2 I_p\right)^{-1} A^\top Y,$$

and one can recover the original weights with

$$\hat{\beta}_T = C^{-1}\hat{\beta}_A.$$

There exists an interesting relationship between the prior covariance matrix, $\Sigma$, and the Tikhonov penalty matrix, $C$. When the penalty Gram matrix, $C^\top C$, is full-rank, it is invertible and there exists a corresponding unique prior,

$$\Sigma = \left(C^\top C\right)^{-1}.$$

However, the standard form decouples the two concepts. There exist well-defined Tikhonov penalties for which a prior cannot be expressed. In particular, if the penalty Gram matrix is not positive semi-definite, no inverse exists and therefore the prior cannot be formally expressed. The converse is also true. A rank-deficient prior can be used if the problem is in standard form, yet there is no corresponding penalty matrix. See Doicu et al. (2010) for a full treatment.

## 4.2 Equivalence of FIR models with temporal priors and convolution followed by ridge

Estimating an FIR model with a temporal prior $\Sigma^T = \mathbb{B}\mathbb{B}^\top$

$$Y = X\beta + \epsilon$$

$$\beta_i \sim \mathcal{N}_d\left(0, \lambda^{-2}\mathbb{B}\mathbb{B}^T\right)$$

is equivalent to convolving the features $x_i$ with the columns of $\mathbb{B}$ and estimating the model using ridge regression:

$$Y = \begin{bmatrix} | & | & | & | & | & | & | \\ (x_1 * b_1) & \dots & (x_1 * b_k) & \dots & (x_p * b_1) & \dots & (x_p * b_k) \\ | & | & | & | & | & | & | \end{bmatrix} \beta + \epsilon \qquad (4.1)$$

$$\beta \sim \mathcal{N}_{pk}\left(0, \lambda^{-2} I_{pk}\right)$$

Recall the definition of the standard form transform:

$$\begin{aligned} \Sigma^T &= \left(C^\top C\right)^{-1} \\ A &= XC^{-1}, \end{aligned}$$

where $C^{-1} = \mathbb{B}$ for a temporal prior $\Sigma^T = \mathbb{B}\mathbb{B}^\top$. The standard transform of the FIR model can be written as

$$A = \overbrace{\begin{bmatrix} | & | & | & | & | & | \\ X^{(1)} & X^{(2)} & \dots & X^{(i)} & \dots & X^{(p)} \\ | & | & | & | & | & | \end{bmatrix}}^{X} \overbrace{\begin{bmatrix} b_1(0) & b_2(0) & \dots & b_k(0) \\ b_1(1) & b_2(1) & \dots & b_k(1) \\ \vdots & \vdots & \dots & \vdots \\ b_1(d) & b_2(d) & \dots & b_k(d) \end{bmatrix}}^{\mathbb{B}}$$

where each $X^{(i)} \in \mathbb{R}^{n \times d}$ is a matrix that contains every feature $x_i \in \mathbb{R}^{n \times 1}$ at delays 0 through $d$, and every row of $X^{(i)}$ corresponds to a particular time point $t$:

$$X^{(i)}(t) \;=\; \left[\; x_i(t) \quad x_i(t-1) \quad \dots \quad x_i(t-d) \;\right]$$

We can express every entry of the matrix $A$ as the dot product between $X^{(i)}(t)$ and each column of the temporal basis set, $b_j$ :

$$\begin{aligned}
a_i^{b_j}(t) &= \left[\; x_i\,(t) \quad x_i\,(t-1) \quad \dots \quad x_i\,(t-d) \;\right] \begin{bmatrix} b_j(0) \\ b_j(1) \\ \vdots \\ b_j(d) \end{bmatrix} \\
a_i^{b_j}(t) &= \left\langle \begin{array}{ccccc} [x_i(t), & x_i(t-1), & x_i(t-2), & \dots, & x_i(t-d)], \\ [b_j(0), & b_j(1), & b_j(2), & \dots, & b_j(d)] \end{array} \right\rangle \\
a_i^{b_j}(t) &= \sum_{\delta=0}^{d} x_i\,(t-\delta)\, b_j\,(\delta)
\end{aligned}$$

which is the definition of discrete convolution

$$\begin{aligned}
(x_i * b_j)\,[t] &\equiv \sum_{\delta=0}^{d} x_i(t-\delta) b_j(\delta) \\
a_i^{b_j}(t) &= (x_i * b_j)\,[t]
\end{aligned}$$

Finally, we rewrite $A = X\mathbb{B}$ as the convolution of each feature $i$ with each temporal basis $j$

$$\begin{aligned}
a_i &= \begin{bmatrix} | & | & | & | \\ (x_i * b_1) & (x_i * b_2) & \dots & (x_i * b_k) \\ | & | & | & | \end{bmatrix} \\
A &= \begin{bmatrix} | & | & | & | & | & | \\ a_1 & a_2 & \dots & a_i & \dots & a_p \\ | & | & | & | & | & | \end{bmatrix} \\
A &= \begin{bmatrix} | & | & | & | & | & | & | \\ (x_1 * b_1) & \dots & (x_1 * b_k) & \dots & (x_p * b_1) & \dots & (x_p * b_k) \\ | & | & | & | & | & | & | \end{bmatrix}
\end{aligned}$$

This is exactly Equation 4.1.

## 4.3 Kernel solution to encoding models with spatiotemporal MVN priors

The standard form solution is

$$\hat{\beta}_A = \left(A^\top A + \lambda^2 I\right)^{-1} A^\top Y$$

The kernel solution to the standard form problem becomes

$$\hat{\beta}_A = A^\top \left(AA^\top + \lambda^2 I\right)^{-1} Y$$

Expanding this out using the fact that $A = XC^{-1}$

$$\hat{\beta}_A = C^{-1\top} X^\top \left(XC^{-1}C^{-1\top}X^\top + \lambda^2 I_p\right)^{-1} Y$$

We know $\Sigma = C^{-1}C^{-1\top} = (CC^\top)^{-1}$. Replacing this in

$$\hat{\beta}_A = C^{-1\top} X^\top \left(X\left(C^{-1}C^{-1\top}\right)X^\top + \lambda^2 I_p\right)^{-1} Y$$

To recover the Tikhonov solution, recall that $\hat{\beta}_T = C^{-1}\hat{\beta}_A$. Substituting this in

$$\hat{\beta}_T = C^{-1}C^{-1\top} X^\top \left(X\left(C^{-1}C^{-1\top}\right)X^\top + \lambda^2 I_p\right)^{-1} Y$$

We know $\Sigma = C^{-1}C^{-1\top}$. Replacing this in

$$\hat{\beta}_T = \Sigma X^\top \left(X\Sigma X^\top + \lambda^2 I_p\right)^{-1} Y$$

In the case of spatiotemporal kernels $\Sigma = \Sigma^T \otimes \Sigma^X$. The full solution becomes

$$\hat{\beta}_T = \left(\Sigma^T \otimes \Sigma^X\right) X^\top \left(X\left(\Sigma^T \otimes \Sigma^X\right)X^\top + \lambda^2 I_p\right)^{-1} Y$$

## 4.4 Efficient kernel solution for models with spatiotemporal MVN priors

We now derive a computationally efficient solution for the kernel solution for an encoding model with non-spherical spatiotemporal multivariate normal priors. This formulation makes the estimation of these models computationally tractable.

The spatiotemporal prior is constructed by computing the Kronecker product ($\otimes$) between the feature prior $\Sigma^X \in \mathbb{R}^{p\times p}$ and the temporal prior $\Sigma^T \in \mathbb{R}^{d\times d}$,

$$\Sigma = \Sigma^T \otimes \Sigma^X = \begin{bmatrix} \Sigma^T_{1,1}\Sigma^X & \cdots & \Sigma^T_{1,d}\Sigma^X \\ \vdots & \ddots & \vdots \\ \Sigma^T_{d,1}\Sigma^X & \cdots & \Sigma^T_{d,d}\Sigma^X \end{bmatrix}.$$

The resulting spatiotemporal prior is $\Sigma \in \mathbb{R}^{pd \times pd}$. Notice that when both the feature and the temporal priors are spherical, the spatiotemporal prior is also spherical.

The Tikhonov solution to an encoding model with a spatiotemporal multivariate normal prior $\Sigma^T \otimes \Sigma^X$ can be expressed as (see Appendix 4.3):

$$\hat{\beta}_T = \left(\Sigma^T \otimes \Sigma^X\right) X^\top \underbrace{\overbrace{\left(X \left(\Sigma^T \otimes \Sigma^X\right) X^\top\right)}^{K} + \lambda^2 I\Big)^{-1} Y.}_{\hat{\alpha}}$$

A computationally efficient solution can be derived by re-arranging terms. First, notice that the kernel regression solution to the standard form problem is embedded within the Tikhonov solution above:

$$\hat{\alpha} = \left(X \left(\Sigma^T \otimes \Sigma^X\right) X^\top\right) + \lambda^2 I)^{-1} Y.$$

The term inside the parenthesis is the regularized $n \times n$ kernel matrix $K$ of the standard form transformation:

$$\left(K + \lambda^2 I\right) = \left(X \left(\Sigma^T \otimes \Sigma^X\right) X^\top + \lambda^2 I\right).$$

Recall that $X$ is an $n \times pd$ FIR matrix which includes delayed copies of the linearized stimulus feature matrix. Computing the kernel matrix thus requires the following matrix multiplication

$$K = \begin{bmatrix} X_{\delta(1)} \cdots X_{\delta(d)} \end{bmatrix} \begin{bmatrix} \Sigma^T_{1,1}\Sigma^X & \cdots & \Sigma^T_{1,d}\Sigma^X \\ \vdots & \ddots & \vdots \\ \Sigma^T_{d,1}\Sigma^X & \cdots & \Sigma^T_{d,d}\Sigma^X \end{bmatrix} \begin{bmatrix} X_{\delta(1)} \\ \vdots \\ X_{\delta(d)} \end{bmatrix}$$

Finally, this matrix multiplication can be expressed as a sum of matrix products,

$$K = \sum_j^d \sum_i^d \Sigma^T_{(i,j)} \left(X_{\delta(i)} \Sigma^X X^\top_{\delta(j)}\right).$$

This formulation makes the problem of estimating encoding models with spatiotemporal multivariate normal priors tractable in contexts when $n < p$.

$$\hat{\beta}_T = \left(\Sigma^T \otimes \Sigma^X\right) X^\top \left(\sum_j^d \sum_i^d \Sigma^T_{(i,j)} \left(X_{\delta(i)} \Sigma^X X^\top_{\delta(j)}\right) + \lambda^2 I\right)^{-1} Y$$

## 4.5 Extension to priors on priors: hyper-priors

We have shown the usefulness of imposing various temporal and spatial priors on feature weights to improve predictive models. There exist situations, however, when the expert prior itself needs to be regularized. This can be the case when the expert prior is derived empirically and is noisy, or when the prior can be modified to match the data better. In such cases, we can apply the same principle and impose a prior on the prior—a hyper-prior. We next show an example on how to incorporate hyper-priors to our framework.

In the section on Temporal Priors we described the smoothness prior. Our results show that imposing a smoothness prior on the temporal delays does not improve the prediction performance of the motion-energy model. This is surprising. We expect the haemodynamic response function to be temporally smooth, and so imposing a smoothness prior should improve prediction performance. This intuition, however, ignores the structure of the smoothness prior.

The smoothness prior imposes a strong covariance to delays in the middle of the temporal filter (see Figure 1.3). This is problematic because the goodness of the prior will depend on the number of delays. This is a bad assumption in many cases. In order to avoid this issue, we can impose a spherical prior on the smoothness prior. This can be thought of as trading off between a spherical prior and the smoothness prior, where the tradeoff is controlled by the hyper-prior hyper-parameter.

In general, hyper-priors can be expressed as

$$\beta \sim N_p\left(0, \lambda^{-2}\Sigma\right)$$
$$\Sigma \sim W_p\left(\gamma^{-2}\Lambda_p\right),$$

where $W$ is a Wishart distribution. In the case of the smoothness prior, this results in

$$\Sigma^* = \lambda^{-2}\left(\mathbb{D}^2 + \gamma^2 I_p\right)^{-1},$$

where $\lambda$ and $\gamma$ are hyper-parameters.

Estimating models that include both a prior and a hyper-prior is feasible under our framework (and implemented in the accompanying software). However, this flexibility comes at the cost of computational resources because the hyper-prior hyper-parameter ($\gamma$) needs to be estimated via cross-validation.

## 4.6 Prior covariance matrix and matrix rank

In order for some matrix $\Sigma$ to serve as the covariance matrix for a multivariate normal prior, that matrix should have certain properties that are common across

all covariance matrices. It should be symmetric and positive semi-definite, meaning that all of its eigenvalues should be non-negative. Oddly, the definition of Tikhonov regression seems to require that $\Sigma$ be full rank (and thus positive definite rather than semi-definite), because the penalty matrix, $C$, is related to $\Sigma$ by $\Sigma = (C^\top C)^{-1}$. Unless elements of $C$ are allowed to approach infinity, this relationship requires that $\Sigma$ be invertible. However, if we use the standard form of Tikhonov regression this requirement disappears, since it only depends on $C^{-1}$, which is well-defined as long as $\Sigma$ is finite and positive semi-definite. Thus, if one is using the standard form it seems that there is no requirement that $\Sigma$ be full rank.

Indeed, there are many situations in which $\Sigma$ will be rank-deficient. For example, $\Sigma$ could be constructed using the feature extraction method detailed above, with the number of features being less than the number of channels in the model. In this case the feature matrix $E$ is tall, having more rows than columns, and $EE^\top$ is not full rank. This corresponds to a prior covariance matrix in which some directions have exactly zero variance. Iso-probability curves in the distribution defined by this covariance matrix will have a pancake-like appearance with exactly zero thickness along the null directions.

# Chapter 5

# Appendix to evaluation of RSA

## 5.1 Relationship between RSA and the stimulus triggered average

The stimulus triggered average (STA) is one of the simplest models that is used in neuroscience (De Boer and Kuyper, 1968, Marmarelis and Naka, 1972). The stimulus triggered average (STA) is optimal if there are no correlations between the stimuli or features of interest (i.e. orthogonal design matrix) and if the errors are uncorrelated in time (i.e. iid errors). The weight estimates for an STA model can be computed directly as:

$$\hat{\beta}_{STA} = \frac{X^\top Y}{n}.$$

### 5.1.1 The coefficient of determination for STA

We can evaluate the STA model by computing the amount of variance it explains in the data using the coefficient of determination ($R^2$). This can be achieved by computing the matrix trace between the predicted ($\hat{Y}$) and the actual ($Y$) responses. After some algebra, this can be expressed as

$$\begin{aligned} R^2 &\propto trace\left(Y^\top \hat{Y}\right) \\ \hat{Y} &= X\left(X^\top Y\right) \\ R^2 &\propto trace\left(Y^\top X\left(X^\top Y\right)\right) \\ trace\left(Y^\top X\left(X^\top Y\right)\right) &= trace\left(X\left(X^\top Y\right)Y^\top\right) \\ R^2 &\propto trace\left(XX^\top YY^\top\right). \end{aligned}$$

### 5.1.2 Relationship between RSA and STA

Recall the RSA similarity estimate when using correlation

$$corr\left(triang\left(\frac{XX^\top}{p}\right), triang\left(\frac{YY^\top}{v}\right)\right)$$

In fact, the relationship between the stimulus triggered average $R^2$ defined above and this quantity can be derived exactly (not shown). For this reason, RSA can be understood in terms of an STA encoding model.

In cases where the features are not orthogonal and there is a need to trade-off between the empirical covariance and some regularization term, STA is not a good model. In cases, where ignoring the empirical feature covariance (or when features are orthogonal), STA can be a good model. In either case, STA is a particular type of encoding model that assumes the features are uncorrelated with each other (i.e. are orthogonal). In some cases, that might be a good assumption and in some cases it not. RSA is closely related to the STA and will fail whenever STA is a poor model of the relationship between features and brain responses.

### 5.1.3 Modifying the ridge solution to include STA as a special case

Ridge regression can be modified in order to include this solution as a special case. To see this, notice that we can express the ridge solution as a trade-off between the empirical feature covariance ($X$) and the ridge penalty ($I$). This trade-off is controlled by the regularization parameter $\alpha$:

$$\hat{\beta}_{Ridge} = \left(\alpha X^\top X + (1-\alpha)\, I\right)^{-1} \left(\frac{X^\top Y}{n}\right).$$

Finally, notice that if the regularization parameter is zero ($\alpha = 0$) or if the features are completely uncorrelated (i.e. are orthogonal, $X^\top X = I_p$), the term inside the inverse becomes a diagonal matrix and the ridge weights become proportional to the STA:

$$\hat{\beta}_{Ridge} \propto \hat{\beta}_{STA} = \frac{X^\top Y}{n}$$