An Evaluation of Representational Similarity Analysis for Model Selection and Assessment in Computational Neuroscience

Luke Chen

Department of Electrical Engineering and Computer Sciences
University of California at Berkeley, Berkeley, CA 94720

Abstract

An important goal in neuroscience is to determine what types of information are represented across brain regions. Often, a computational model is used to extract stimulus features that are hypothesized to be represented within a particular brain region. Any particular study tries to assess the relationship between the features extracted by the computational model and the measured activity from a brain region. In recent years, several approaches to studying this relationship have been developed in the field of cognitive neuroscience. A simple and widely used approach is representational similarity analysis (RSA). This approach attempts to quantify similarities between the representational space of a computational model and a set of brain responses. RSA begins with an estimate of the stimulus-by-stimulus representational similarity (or dissimilarity) matrix computed from a set of stimulus-evoked brain responses. Then, a stimulus-by-stimulus representational similarity matrix is obtained from a computational model. RSA computes the similarity of these similarity matrices. However, there exists little work assessing the validity of RSA. In this paper, we show that RSA actually makes very strong assumptions about the relationship between representational spaces and brain responses. When these assumptions are violated, RSA can fail to detect significant relationships. More worryingly, when used for model selection RSA can lead researchers to the wrong answer. In contrast, we show that standard encoding models that use regression methods perform better than RSA.

Keywords: representational similarity analysis, encoding models, functional MRI

1. INTRODUCTION

An important goal of neuroscience is to identify which types of information different brain regions represent. In one strategy for studying brain representations, researchers first record brain responses to different stimuli. Statistical methods are then used to assess the strength of the relationship between stimulus and brain responses. These statistical assessments are then used to make inferences about the representational space encoded in regions of interest. Many statistical techniques are available to cognitive neuroscientists. They range from statistical parametric mapping approaches [12], multivariate pattern analysis (MVPA) techniques [4], and encoding models [16]. A common type of MVPA analysis is representational similarity analysis (RSA; [10]). RSA has been widely adopted in part due to its computational simplicity. However, little work to-date has explored the validity of RSA.

In this paper, we use simulated and real data to evaluate the validity of RSA as an approach to model assessment and model selection, and more generally as a tool for computational neuroscience. Model assessment refers to the ability to to detect a significant relationship between the stimuli and the responses when a

relationship exists. First, we show that RSA is under-powered when used for model assessment. This leads to an increased Type II error (i.e. many false negatives) relative to encoding models using cross-validated regularized regression (CVR). Second, model comparison refers to the ability to correctly adjudicate between multiple candidate representational spaces and choose the correct one. We show that model selection with RSA can in fact fail and lead researchers to incorrect conclusions. This leads to an increased Type I error (i.e. many false positives) relative to CVR. This is particularly problematic given its wide use in the literature. Third, we also show that searchlight RSA imposes a strong spatial prior and can fail to detect the brain regions that encode a representational space. Fourth, the use of searchlight RSA can also lead to wrong inferences when selecting which of a set of regions better encodes a representational space. Fifth, we show that the new incarnation of RSA ("mixed-RSA") can be viewed as a roundabout way of doing linear regression with strong assumptions. Finally, we show that only in very limited cases RSA is a valid inference technique. It is valid for ROI-specific studies where mixed-RSA is used to test a single model. Unfortunately, this is a very rare use case putting into question the validity of RSA.

In order to get a sense of how much fixed RSA is used for model assessment and model selection, we surveyed the literature. We found 1000 papers use the term "representational similarity analysis" in the text. We randomly sampled 100 of those papers. Of the 100 sampled papers, $X\%$ use RSA for model comparison. This means there are likely XXX papers that might need to be re-analyzed in order to determine whether their conclusions are valid or whether they are driven by the assumptions of RSA.

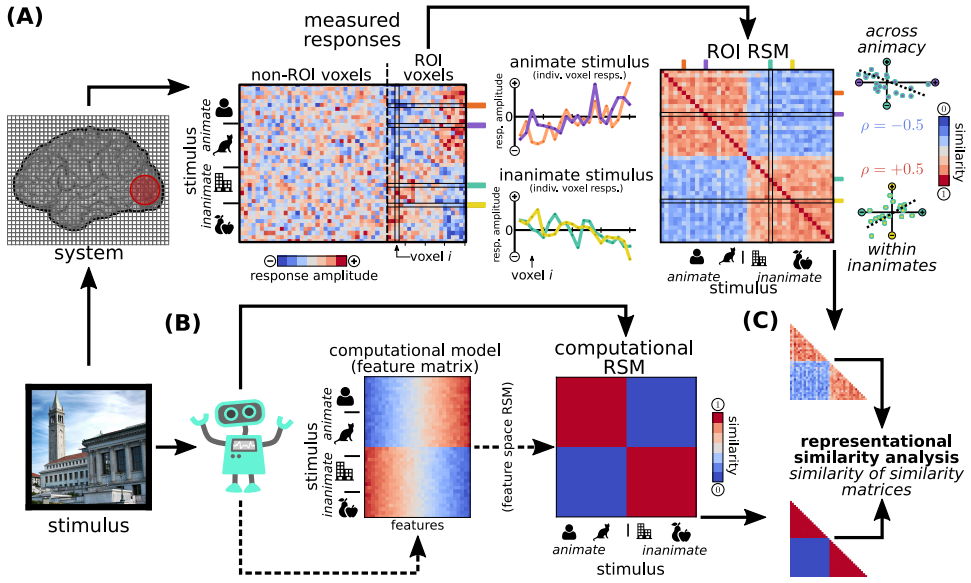

Figure 1: **Description of representational similarity analysis. (A)** A series of stimuli are presented to a subject while their brain activity is measured. In this hypothetical example, the stimuli come from animate (humans and animals) and inanimate (buildings and fruits) object categories. Brain responses are measured from the whole brain but only voxels that fall within the region-of-interest (ROI; red circle) are analyzed. The "measured responses" matrix contains a row per stimulus and a column per voxel. The individual voxel responses for a pair of animate (orange and purple) and inanimate (green and yellow) stimuli are plotted separately. The voxel responses to the two example animate stimuli are very similar (same for the inanimate pair). We then build the ROI representational similarity matrix (RSM) by correlating the voxel responses for each stimulus pair. Each entry in the ROI RSM represents the similarity between each pair of stimuli. In this example, there is a high degree of similarity within animate and within inanimate stimuli, and a low degree of similarity across animate and inanimate categories. **(B)** The stimuli are also shown to a computational model which is used to construct an RSM. Often, a computational model provides a feature representation of the stimuli. The resulting feature matrix is then used to build an RSM. In this example, the computational RSM captures the idea that stimuli within a category should be represented similarly (red), and that stimuli across categories should have low representational similarity (blue). **(C)** The final step in RSA is to compute the similarity between all the unique stimulus pairs from the two similarity matrices.

2. REPRESENTATIONAL SIMILARITY ANALYSIS

In a typical cognitive neuroscience experiment, a subject is presented a set of stimuli. The subject is asked to perform a task, or to passively perceive the stimuli while their brain activity is measured. In neuroimaging, the measured brain responses often consist of u-dimensional "images" recorded at time t every few moments, $y(t) \in \mathbb{R}^u$. With these data in hand, the researcher can use RSA to assesses whether there is a statistical relationship between the stimulus and the measured brain responses (Figure 1).

2.1. ***Brain region representational similarity matrix*** Ω_Y***: a normalized response kernel***

A subset of v measured responses $Y \in \mathbb{R}^{n \times (v<u)}$ is then selected to compute a stimulus-by-stimulus representational similarity matrix (RSM; Figure 1A). One approach to choosing the subset of v measured responses is to have an *a priori* region of interest (ROI). Another approach is to use a spatial window over the brain image (REF searchlight). In either case, the RSM acts as an estimate of the response similarity to the stimuli. The entries in the RSM are obtained by correlating the population responses to each stimulus pair:

$$\Omega_Y = \begin{bmatrix} \rho\left(y^1, y^1\right) & \rho\left(y^1, y^2\right) & \cdots & \rho\left(y^1, y^n\right) \\ \rho\left(y^2, y^1\right) & \ddots & & \rho\left(y^2, y^n\right) \\ \vdots & & \ddots & \vdots \\ \rho\left(y^n, y^1\right) & \rho\left(y^n, y^2\right) & \cdots & \rho\left(y^n, y^n\right) \end{bmatrix},$$

where each $y^j \in \mathbb{R}^v$ is the population response to stimulus j. In the RSA literature $1 - \Omega_Y$ is commonly used and is called the representational dissimilarity matrix (RDM). Throughout this paper, and without loss of generality, we use the representational similarity matrix instead of the RDM.

The RSM is related to the matrix product of the measured responses $YY^\top$ by:

$$\Omega_Y = \frac{YY^\top}{v},$$

where v is the number of units in the population and the rows of Y are zero-mean and unit-norm (i.e. z-scored). The product $YY^\top$ is called the linear "kernel" of Y.

2.2. ***Candidate representational similarity matrix*** Ω_X***: a normalized feature kernel***

Just as the stimuli are shown to the subject, the stimuli are also shown to one or more computational models whose outputs are recorded (Figure 1B). A computational model can be thought of as implementing a hypothesis about the properties of the stimulus that are important for brain responses. In the RSA framework, a stimulus-by-stimulus representational similarity matrix, can be obtained from a computation model by correlating the outputs to (i.e. features of) each stimulus pair[1]:

$$\Omega_X = \begin{bmatrix} \rho\left(x^1, x^1\right) & \rho\left(x^1, x^2\right) & \cdots & \rho\left(x^1, x^n\right) \\ \rho\left(x^2, x^1\right) & \ddots & & \rho\left(x^2, x^n\right) \\ \vdots & & \ddots & \vdots \\ \rho\left(x^n, x^1\right) & \rho\left(x^n, x^2\right) & \cdots & \rho\left(x^n, x^n\right) \end{bmatrix}.$$

This is equivalent to $\frac{XX^\top}{p}$, where $X \in \mathbb{R}^{n \times p}$ is the stimulus feature matrix and the rows are normalized to have zero-mean and unit-norm (i.e. z-scored). Also, $XX^\top$ is the linear kernel of X.

[1]Some computational models may output an RSM directly. In such cases, the RSM can be decomposed to recover a feature representation (e.g. via singular value decomposition).

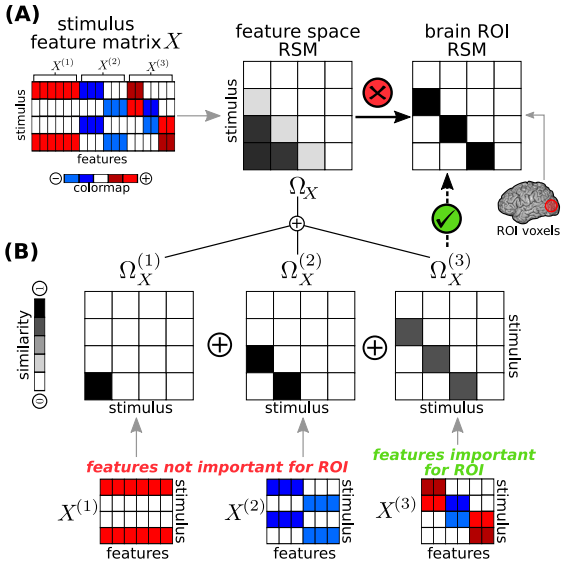

Figure 2: **RSA fails in model assessment when only a subset of features are important.** A simple hypothetical example where a subset of the features within a feature space are encoded in a region of interest (ROI). RSA fails to find a relationship between the feature space and the brain region because the relevant features get "washed out" by unimportant features. **(A)** The stimulus feature matrix X can be divided into three sets of features. However, only a subset of these features ($X^{(3)}$) is important in driving brain activity in the ROI. **(B)** However, because the feature space RSM combines all features equally, the unimportant features will make the feature space RSM (Ω_X) very different from the ROI RSM, even if the ROI encodes those features. This results in a statistical power decrease of RSA and the relationship will not be found (red x-mark). If we somehow knew a priori what the relevant features were, we could construct an RSM from the important features ($\Omega_X^{(3)}$) and RSA would be able to detect the relationship with the brain RSM (green check-mark, dashed arrow). However, in general, it is unfeasible to know a priori what exact features a brain region represents.

2.3. ***Similarity of similarity matrices: a scaling of the matrix trace***

Given the measured brain responses and a computational model, the next step in RSA is to estimate the similarity between the representational similarity matrices (Figure 1C). The RSA similarity estimate is computed using only the upper (or lower) triangular entries of the RSMs:

$$RSA\left(\Omega_Y, \Omega_X\right) \equiv similarity\left(triang(\Omega_Y), triang(\Omega_X)\right)$$

Typically, some form of correlation is used as the similarity function [9]. Throughout the paper, we consider only the Pearson correlation due to its mathematical simplicity and its widespread use in the RSA literature. In fact, when the Pearson correlation is used, the RSA similarity estimate is closely related to the trace of the product of the response and feature linear kernels ($trace(YY^\top XX^\top)$; see Appendix A).

3. FAILURE CASES OF RSA

Previous work has shown that the matrix trace can be used to assess the statistical relationship between (non-linear) kernels [3]. However, RSA similarity estimates differ from the trace in various ways (Appendix A). In practice, these differences are enough to make RSA similarity estimates unreliable statistics for inference [14]. We present more cases where RSA can fail.

3.1. ***RSA can fail when only a subspace of the feature space is important***

RSA can fail to detect a significant relationship between features and brain responses when only a subset of features are important for an ROI (Figure 2). This occurs because in RSA all the features of a computational model are considered equally important for the ROI. For example, if the computational model is a set of Gabor wavelets, the representational similarity between two stimuli will contain information about left and right visual fields. This is appropriate for bilateral visual regions of interest. However, it is not appropriate to include features from both visual fields when analyzing individual hemisphere ROIs. That is because each hemisphere processes information from only one visual field. This lack of feature selectivity is at the heart of the problem. RSA can fail to detect a significant relationship between features and brain responses whenever irrelevant features "wash out" the features that are important for the ROI (Figure 3B).

More problematic is the use of RSA for adjudicating between candidate computational models (Figure 3). This is more commonly referred to as model selection [2]. Unfortunately, RSA can give the wrong answer when selecting which computational model is a better representational space for an ROI. A wrong answer can occur when the assumptions implicit in RSA are better met by the wrong representational space.

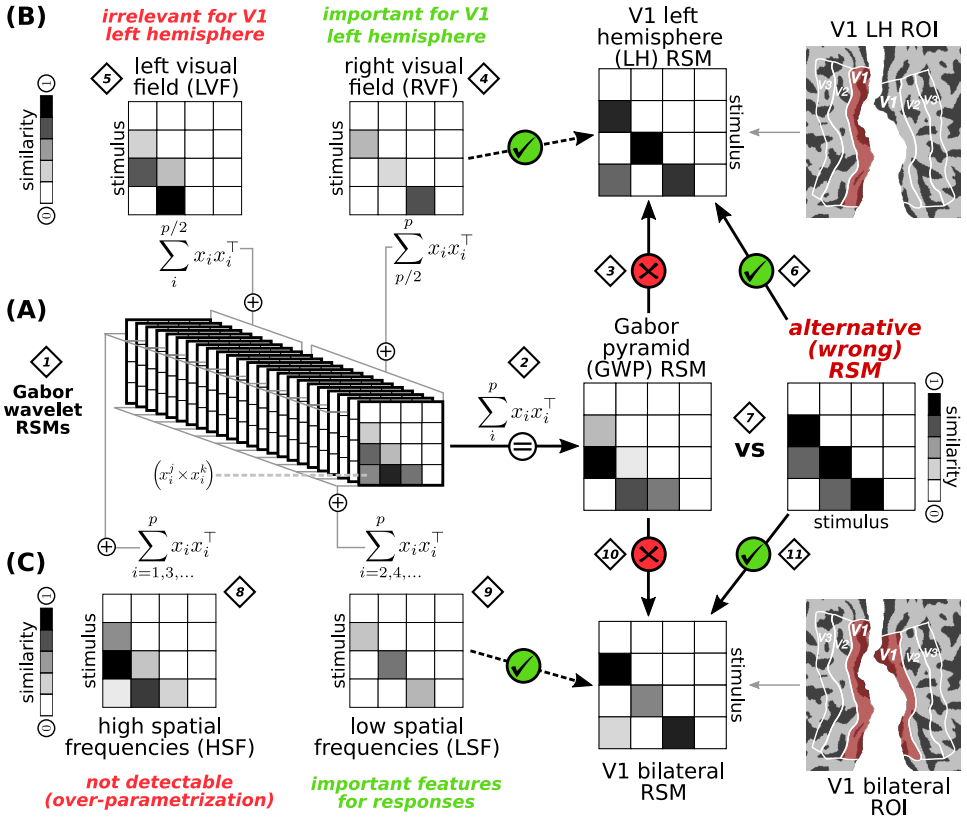

Figure 3: **RSA can easily fail to find the correct answer when comparing representational spaces.** In a typical study, a researcher seeks to find out which of two or more candidate feature spaces are represented in a given brain region. RSA is commonly used for these types of comparisons. In this hypothetical example, we want to find out whether V1 representations are better captured by a Gabor wavelet pyramid or some other alternative model. **(A)** First, a set of p Gabor wavelets are used to compute an RSM across all stimuli (**1**). (**2**) The Gabor wavelet pyramid (GWP) RSM is the sum of all p individual Gabor wavelet RSMs. (**3**) We compare the Gabor wavelet pyramid RSM against the left hemisphere (LH) V1 ROI. We can see that the GWP RSM is not similar to the LH V1 RSM (red x-mark). **(B)** However, we know from neuroanatomy that the left hemisphere only processes information from the right visual field (**4**). If we select all the Gabor wavelet RSMs that correspond to the right visual field, we can see that the resulting RSM is very similar to the LH V1 RSM (green check-mark, dashed line). The GWP RSM is not similar to the LH V1 RSM because the relevant RFV features get "washed out" by the irrelevant LFV features (**5**). (**6**) This issue becomes especially problematic when using RSA for model comparison. In our example, the alternative (wrong) RSM is similar to the LH V1 RSM by chance (green check-mark). (**7**) When testing whether GWP or the alternative feature space is a better model for the representations of LH V1, RSA chooses the alternative (wrong) model. This is a bad property of RSA. (**8**) The same thing can happen if our Gabor wavelet pyramid is over-parametrized (e.g. by including very high spatial frequencies not detectable at the resolution of fMRI). As in **(B)**, the relevant features (i.e. low frequency Gabor wavelets, (**9**)) get "washed out" and RSA will fail to detect the GWP RSM similarity with V1 (**10**). (**11**) Again this can easily lead to incorrect inferences if the alternative model is by chance similar to the V1 RSM. This can in fact happen with real data (see Section 5.1 and Figure 6).

*3.2. **RSA can fail when only a sub-region of the ROI is important***

RSA can also fail to detect a significant relationship between a representational space and brain responses when only a sub-region of the ROI encodes the representational space (Figure 4). This is not a problem if the ROI is hypothesis-driven and well-specified. The candidate representational space does not match the well-specified ROI RSM and that is that. However, if the ROI is not well-specified (e.g. is derived from an atlas) this conclusion might be wrong because a sub-population of responses might in fact be a better match to the candidate representational space (Figure 4A). This problem is not ameliorated by the use of a searchlight (REF). In fact, the searchlight size is rarely explored in the literature and thus suffers from issues typical to

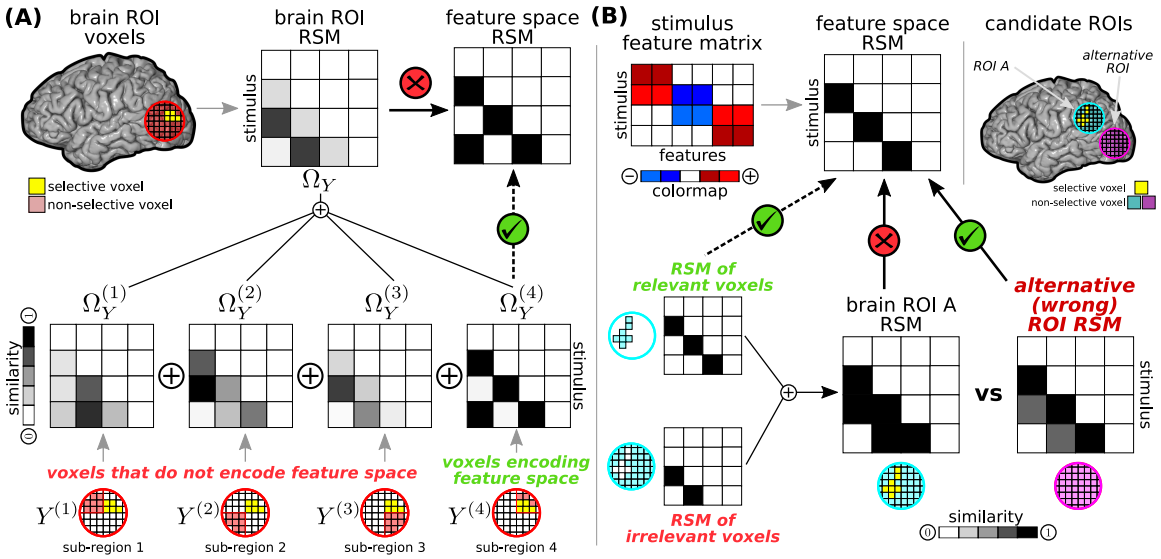

Figure 4: **RSA can fail to find the brain region encoding a representational space and give the wrong answer. (A)** In a typical searchlight analysis, the voxels contained within a sphere (red circle) are used to construct a region-of-interest (ROI) representational similarity matrix (RSM). It can be the case that only a small subset of voxels within the ROI are selective for the representational space of interest (yellow voxels). In such cases, RSA will fail to find a significant relationship between the RSMs (red x-mark). However, by dividing the ROI into sub-regions RSA can find the relationship. This shows that the size of the searchlight is important for detecting an effect (green check-mark). **(B)** More worryingly, when comparing which of two ROIs encodes a representational space, RSA can lead to the wrong answer. This applies to searchlight and non-searchlight RSA analyses. To see this, again assume that only a subset of voxels are relevant for the representational space (yellow voxels in cyan ROI). The irrelevant voxels will "wash out" the effect of the relevant voxels and no effect will be found (red x-mark). By chance, an alternative ROI RSM (magenta ROI) can more closely resemble the representational space of interest (green check-mark). In this case, RSA will incorrectly show that the wrong ROI encodes the representational space.

using a non-optimal filter (REF). More worryingly, an alternative ROI might be significantly more similar to the candidate representational space by chance. In such cases, RSA will yield the wrong conclusion when comparing ROIs (Figure 4B). These issues apply to both mixed- and fixed-RSA, and whenever arbitrary spatial smoothing is performed.

3.3. Intuition as to why RSA can fail

A formal analysis of RSA is difficult because it is not a theoretically grounded technique. Previous work has shown that the RSA similarity estimate exhibits a tenuous relationship with R^2. It turns out that whereas R^2 is a reliable statistic the RSA similarity estimate is not [14]. The differences between R^2 and the RSA similarity estimate are mostly normalization factors, which are the result of RSA using correlation, and they can be enough to make RSA unreliable. Other work exists showing other limitations of RSA (REF: nips paper, op de beck paper). On the other hand, recent work has provided empirical evidence to support RSA as a valid technique (REF hoern). Unfortunately, that work assumes that the representational space is known, which is rarely the case.

In this sub-section, we draw links between encoding models and RSA in order to gain an intuition as to the conditions under which RSA can fail. We show that (i) RSA shares some similarities with STA models and so might be appropriate when features implicit in the representational space are orthogonal (though it can fail in such cases too [14]). When the features are not orthogonal, then (ii) RSA implicitly assumes that all the features matter equally to the population. In the case of linear models, this implies that the *feature weights* are orthogonal in the population. These are strong assumptions about how features are encoded by the population. The more the data diverges from these assumptions, the more RSA is likely to provide incorrect results. Instead of making strong assumptions about how the features are encoded, we can simply estimate how they are encoded in the population. Encoding models are flexible enough to capture the two

scenarios where RSA might be appropriate and many more. Furthermore, encoding models are grounded on standard statistical learning techniques [16, 2].

In linearized encoding models, the population feature weights capture the relative importance of the features for (each unit in) the population [16]. Under certain conditions, we can use Tikhonov regression to obtain an estimate for the population feature weights for a linear model with a multivariate normal prior [15], (Appendix for gaussian stuff). When the identity matrix is used as a prior this is called "ridge regression" and is widely used [5]. In practice, the scale on the prior covariance needs to be estimated via cross-validation. We can express the ridge solution as a trade-off between the empirical covariance ($X^\top X$) and the ridge penalty (I):

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

where $\alpha \in [0, 1]$ is the regularization coefficient that controls this trade-off. Note that when the feature matrix is orthogonal, the term inside the inverse becomes an identity matrix. In such cases, the ridge solution becomes the stimulus triggered average (STA)

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

$$R^2 \propto trace\left(XX^\top YY^\top\right). \tag{1}$$

However, if the features are not orthogonal and a trade-off between the prior exists (i.e. $\alpha \in [0, 1)$), STA is not a good model and it will provide suboptimal answers to scientific questions. Under certain conditions (Appendix A), the ridge estimate will provide better answers (Eq. 3). In cases where some features are completely irrelevant for population responses, LASSO or elastic-net might be better models since they can achieve feature selection.

We can evaluate the ridge regression model by computing the amount of variance it explains in the data using R^2. This can be achieved by computing the matrix trace between the predictions and the actual responses

$$R^2 \propto trace\left(Y^\top \hat{Y}\right) \tag{2}$$

$$trace\left(Y^\top \hat{Y}\right) = trace\left(X\left((1-\alpha)\, X^\top X + \alpha I\right)^{-1} X^\top YY^\top\right). \tag{3}$$

It turns out that RSA is more like an STA model (Eq. 1) than a ridge regression model (Eq. 3). In many RSA studies (REFs), the implicit features in the representational similarity matrix turn out to be orthogonal ($X^\top X = I$). In such cases, the STA solution is appropriate and so RSA can find a relationship between the representational space and the brain responses.

3.4. ***When RSA works***

The RSA similarity estimate is sensible whenever the optimal solution is given by the STA. As illustrated above, this can be the case when the features are orthogonal. It can also occur when the empirical feature covariance is non-orthogonal but an identity prior provides a better model (i.e. $\alpha = 1, \Sigma_\beta = I$). In both these cases, the result is to ignore the feature covariance when estimating the feature weights.

We can certainly explore the cases where RSA will be valid when the solution is approximately given by the STA (whenever we have an a priori ROI, AND the features are orthogonal or are better modeled as orthogonal). However these are very difficult conditions to be met.

3.5. ***Mixed-RSA is roundabout regression with a spatial prior***

In recent work, RSA has incorporated the idea of model estimation into its framework. This is referred to as "mixed-RSA" [8]. This is a step forward for RSA but it is not a novel approach. In effect, the mixed-RSA

similarity estimate is not practically different from taking the mean of the prediction performance from a model that has been estimated. In general, the mixed-RSA similarity estimate can be expressed as

$$RSA\left(\Omega_Y, \Omega_{\hat{Y}}\right) \propto trace\left(YY^\top \hat{Y}\hat{Y}^\top\right),$$

where $Y \in \mathbb{R}^{n\times v}$ are the actual responses and $\hat{Y} \in \mathbb{R}^{n\times v}$ are the predicted responses. Note that the term in the middle ($Y^\top\hat{Y}$) is the dot product of every measured response $y_i \in \mathbb{R}^n$ with every predicted response $\hat{y}_i \in \mathbb{R}^n$. In fact, if the columns of Y and $\hat{Y}$ are zero-mean and unit variance, we can express the mean of the actual and predicted response correlations for all v measured responses as

$$\frac{1}{v}\sum_i^v corr\left(y_i, \hat{y}_i\right) \quad = \quad \frac{1}{v\times n} trace\left(Y^\top\hat{Y}\right)$$

After some algebra, we can see that mixed-RSA has a straight forward relationship to the mean squared prediction performance of the population (see Appendix B):

$$RSA\left(\Omega_Y, \Omega_{\hat{Y}}\right) \propto \frac{1}{v}\sum_i^v corr\left(y_i, \hat{y}_i\right)^2.$$

The main benefit of mixed-RSA is spatial pooling. When used in combination with searchlight, this achieves spatial pooling with a sphere instead of Gaussian blurring as it's usually done in fMRI. It is beyond the scope of this paper to explore the cases where mixed-RSA might provide benefits above and beyond standard spatial smoothing. If such cases exist, they would be interesting to examine. However, spatial pooling comes with its own set of issues (Section 3.2).

4. MODEL ASSESSMENT WITH RSA

RSA assumes that the population weights are orthogonal for each pair of features. When this assumption is met, RSA can assess whether there is a relationship between the representational space and the brain responses. However, if this assumption is not true and the population weights are far from orthogonal, then RSA can lead to incorrect results.

4.1. ***RSA fails when its assumptions are violated***

In order to evaluate the use of RSA for model assessment, we simulated 2000 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 1000 simulations, the weight feature weight matrix was approximately orthogonal ($\Lambda_\beta \approx I$). In the other 1000 simulations, the matrix was very far from orthogonal and all the units in the population had approximately the same weight vector (rank-one). After generating the data, we conducted RSA as described in Section 2. For each simulation, we assessed the significance of the relationship by shuffling the RSM matrix 10^3 times [9].

When the feature weight matrix was close to orthogonal, RSA detected the statistical relationship between the representational space and brain responses (all 1000 $p = 10^{-3}$; Figure 5A). This is expected because the RSA assumption is met ($\Lambda_\beta \approx I$). However, when this assumption is violated and the 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 it assumes that all the features will be equally useful in driving the population responses, which is not the case when the weight matrix is far from orthogonal.

In contrast, regression models explicitly estimate weights for each unit in the population and can therefore reliably find a significant relationship in both cases (all 2000 simulations $p < 0.05$, not shown).

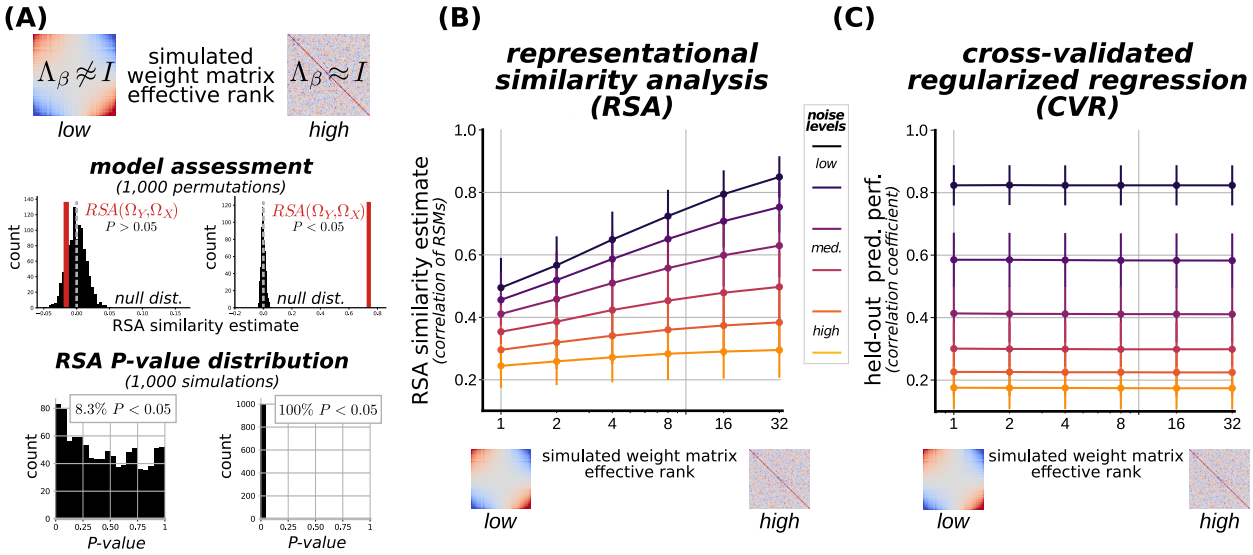

Figure 5: **RSA estimates are affected by how orthogonal the population weight vectors are, encoding model estimates are not**. Simulations show that the ability of RSA to detect a relationship between a representational space and the measured brain responses is affected by the structure of the covariance in population weights. In particular, RSA does not work well when the population weights are far from orthogonal. We simulated data using simple linear models ($Y = X\beta + \epsilon$). For each simulated population of responses, we varied how orthogonal the weight vectors were. This ranged from highly orthogonal to not very orthogonal. This was achieved by varying the "effective rank" (i.e. the skew of the eigenspectrum) of the population weight matrix covariance. A population weight matrix that is far from orthogonal implies that only a subspace of the representational space is important to the measured responses. Estimating the weight matrix is important in such cases. **(A)**. In this simulation, population responses were generated from sampled weights from a low (rank-one) and a high rank (close to identity) covariance matrix (Λ_β). RSA was used to assess the statistical relationship between the features and responses RSMs. The ground-truth is that there is a significant relationship. Significance was assessed by comparing the RSA similarity against the null distribution computed by shuffling the feature RSM 1,000 times. We repeated this simulations 1,000 times. The P-value distribution shows that when the weight matrix rank is low, RSA fails to find a significant relationship (only 83/1,000 $P < 0.05$; CVR all 1,000 $Ps < 0.05$, not shown). **(B)** As the population weight matrix becomes less orthogonal, RSA yields a smaller estimate. This leads to a decrease in statistical power. In low noise regimes, there is a large difference in the RSA estimates between orthogonal and far from orthogonal population weights. (Error bars indicate standard deviation). **(C)** Cross-validated regularized regression does not suffer from this issue because the population weight matrix is estimated. The prediction performance is dominated by noise and not the structure of the population weight covariance. Y-axes not comparable across panels.

4.2. RSA similarity estimates depend on weight matrix orthogonality

We next evaluated how RSA and CVR models are affected as the feature weight matrix varies from orthogonal to very far from orthogonal. This was achieved by simulating population weight matrices with varying levels of effective rank (1, 2, 4, 8, 16, 32). We ranged the number of stimuli (100, 1000), features (100, 1000), voxels (128, 256, 512), noise levels (1, 2, 3, 4, 5, 6; iid Gaussian s.d.), and feature matrix effective rank (1, 5, 10, 20). The population responses to the stimuli were generated using a linear model (see Appendix E). This resulted in a total of 4,350 simulations for each of the six noise levels.

The RSA similarity estimate is strongly affected by how orthogonal the feature weight matrix is and the noise level (Figure 5B). For any one simulation, all else being equal, the ability of RSA to detect a relationship is related to the similarity estimate. As the effective rank of the weight matrix decreases, the similarity estimate and the ability of RSA to detect a significant relationship decreases. An RSA similarity estimate under high noise and high weight matrix rank conditions may be similar to a low noise estimate with low weight matrix rank. In contrast, cross-validated regularized regression estimates depend little on the effective rank of the weight matrix. Instead CVR estimates are mainly affected by noise (Figure 5C).

In summary, the ability of RSA to detect significant relationships depends on the effective rank of the population feature weights. This affects the likelihood of detecting a relationship between a feature space and

brain responses. 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. A big concern, however, is the use of RSA for model selection.

5. MODEL SELECTION WITH RSA

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

5.1. ***RSA fails to select Gabor features as representational space for V1***

We used V1 data from a vision experiment to evaluate the use of RSA for model selection [13]. We tested whether V1 representations are better captured by Gabor wavelets computed on (i) luminance images, or (ii) object silhouette segmentations (red and blue, respectively, Figure 6). A wealth of evidence has shown that Gabor wavelets computed on natural images are a good model of V1 in neurophysiology (REF) and fMRI (REF). While Gabor wavelets are not the "ground-truth" representational space for V1, they are a good approximation. Thus, there is strong *a priori* expectation that Gabor wavelet features computed on images should capture V1 representations more accurately than those computed on object silhouettes.

A total of 1,260 natural images were shown to two subjects while BOLD responses were recorded with fMRI [13]. The hemodyamic response function and the response to each stimulus was estimated for each voxel separately using generalized least squares [13]. The silhouettes of each object in each stimulus image was drawn by hand and the resulting segmented image was binarized. These silhouette images were used to extract object silhouette features. We also extracted luminance images from the original RGB image from the LAB color space (REF). These luminance images were used to extract image Gabor features.

We used two Gabor wavelet pyramids to extract feature matrices for each of the (i) image Gabor and (ii) object silhouette feature spaces. One pyramid was small and the other was large. This yielded a total of four feature matrices: (i) small and large object silhouette feature matrices, and (ii) small and large Gabor feature matrices. 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 BOLD V1 responses. These large versions of the feature spaces can be thought of as an over-parametrization of the feature spaces.

We tested whether V1 representations are more similar to image Gabor RSMs or object silhouette RSMs using RSA (REF). We bootstraped the difference in similarity to the V1 RSM 10^3 times and computed p-values from this distribution. We also estimated an encoding model for each voxel separately using cross-validated regularized regression [5]. We measured prediction performance by computing the correlation coefficient between predicted and actual voxel responses to 126 held-out images. We boostraped the difference in prediction performance 10^3 times and selected the better feature space based on the mean difference. We also evaluated the effect that using a different number of stimuli has on these analyses (Figure 6C).

Results

RSA yields the expected result when we compare the small feature spaces (Figure 6B, red). Here we see that as the number of stimuli increases, the RSA comparison remains stable. However, when we compare the large feature spaces, RSA gives the opposite answer: representations in V1 appear to be better captured by the silhouette feature space (Figure 6C, orange). As we increase the number of stimuli, the difference between the silhouette and Gabor estimates approaches zero. While we cannot say that this is the "wrong" answer, it certainly goes against expectations. We can say that RSA can give different answers for different parameterizations of the same feature space.

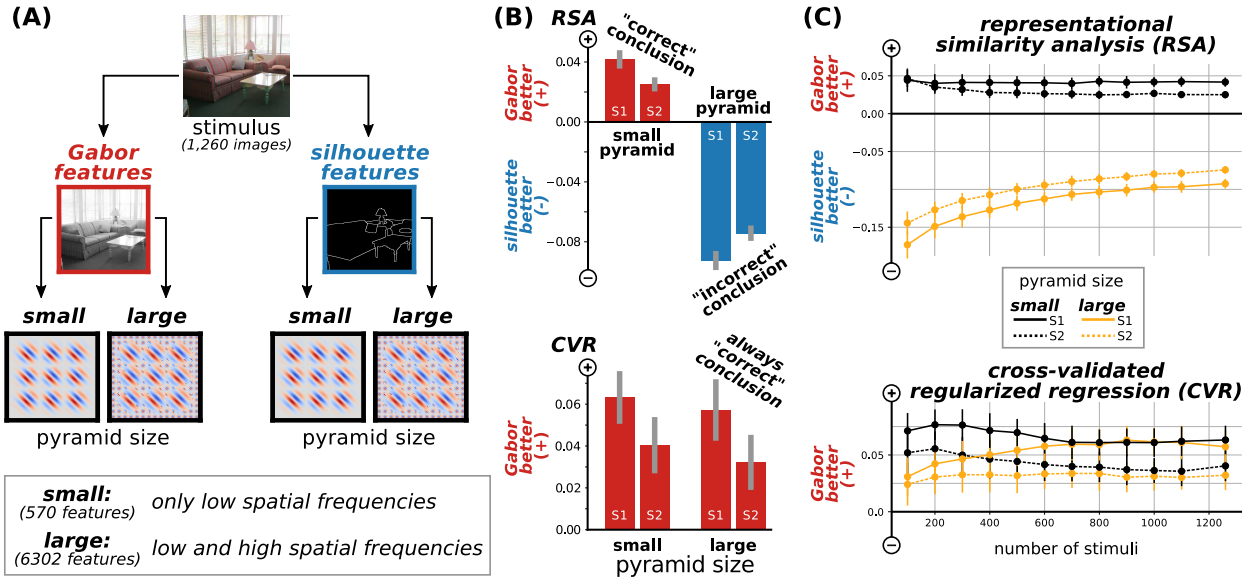

Figure 6: **RSA can lead researchers to the wrong conclusion when used for model selection.** RSA is commonly used to infer which of two or more representational spaces a brain region represents. We used fMRI data collected from two subjects while they viewed 1,260 natural images [13] to test candidate representational spaces for V1. RSA and cross-validated regularized regression (CVR) were used separately to select between representational spaces. **(A)** Candidate representational spaces were constructed from Gabor features (cyan) and object silhouette features (red). We constructed two versions of each feature space: One small that contained only low spatial frequencies, and one large that contained high spatial frequencies in addition to the low spatial frequencies. **(B)** When computing the Gabor RSM the high spatial frequency features "wash out" the contribution of low spatial frequencies that dominate the measured V1 voxel responses at the resolution of fMRI. For this reason, RSA gives a surprising likely incorrect result: silhouette features are a better representational space for V1. This goes against empirical evidence, yet it has been reported in the RSA literature before [9]. In contrast, CVR consistently shows that the Gabor features provide a better predictive model of V1 responses to novel stimuli than silhouette features. **(C)** The analyses were conducted on different number stimuli. The stimuli samples were drawn with replacement and the RSA and CVR differences between feature spaces were computed. RSA consistently gives the wrong answer. The CVR difference estimates for small and large pyramid sizes converge to similar values as more stimuli are included.

In contrast, encoding models estimated with cross-validated L2 regularized regression give consistent results for each subject and feature space size (Figure 6C). We see that as the number of stimuli increases the difference in performance between the silhouette and Gabor feature spaces increases and asymptotes. The statistical power is lower when using the large version relative to the small version of the feature space. This is expected. It is well-known that increasing the number of features decreases statistical power, especially if the features are not useful. Nevertheless, the difference estimate of the regression model remains positive in all the comparisons (Gabor > silhouette).

Under some circumstances, RSA finds that the silhouette feature space is better than the Gabor feature space at describing V1 representations. This is contrary to earlier electrophysiology and fMRI results, which suggest that Gabor wavelets better capture activity V1. The seemingly backwards RSA result only appears when the feature space is over-parametrized. Conducting the same analysis with smaller feature spaces flips the RSA estimates, yielding the expected result. This suggests that RSA does not handle noisy features or high dimensional feature spaces well.

5.2. RSA has lower statistical power than regression for model selection

In the previous experiment, we did not have access to ground-truth features that drive brain responses, nor how they would relate to measured BOLD responses. Thus, we cannot conclude that RSA gave the incorrect answer. To determine the conditions under which RSA can give the wrong answer we performed a series of simulations where the ground-truth model is known (Figure 7A; Appendix F).

We simulated population responses to stimuli as a linear combination of ground-truth features plus noise. We sampled stimulus feature representations that were similar either to the ground-truth features or that were similar to the empirical stimulus-by-stimulus response covariance (X and Z, respectively). The data were generated by simulations that varied in the number of stimuli (100, 300), features (100, 1000), voxels (128, 256, 512), noise levels (iid Gaussian with 1, 2, 3, 4, 5, or 6 s.d.), weight matrix effective rank (1, 3, 5, 7, ..., 32), similarity between the "candidate" feature space X and the ground-truth feature space (10^{-3} to 1; 14 log-spaced samples), and the similarity between the "alternative" feature space Z and the empirical voxel responses (10^{-5} to 1; 10 log-spaced samples). A total of 25,000 simulations for each of the six noise levels were performed. The data were then used to assess the statistical power of RSA for model selection.

In each simulation we tested whether the candidate feature space X was found to be better than the alternative feature space Z at capturing the generated population responses (Figure 6B). The p-value of the difference between the feature spaces was assessed from the empirical distribution estimated from 10^3 bootstrap samples. We quantified the statistical power of RSA and cross-validated regularized regression for model selection by counting the number of times the candidate feature space X was found to be better than the alternative feature space Z for every p-value significance threshold ($p < \alpha$).

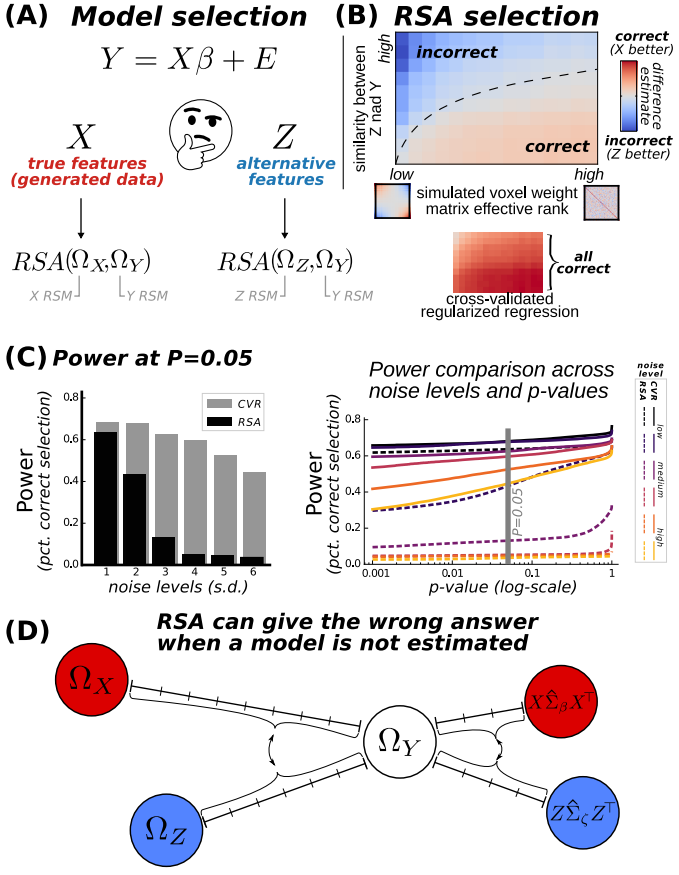

Figure 7: **Comparison of power analysis for model selection between RSA and cross-validated regularized regression.** Model comparisons for a simulation with 300 stimuli, 1,000 features, and 2 s.d. Gaussian noise **(A-B)**[[TOP LEFT AND RIGHT]] RSA yields the wrong answer ($Z > X$) when the weight matrix is low rank and when the misleading feature space (Z) is similar to the measured brain responses (Y). CVR gives the correct answer ($X > Z$) except in cases where the alternate feature space is very similar to the measured brain responses. **(C)**[[BOTTOM LEFT]] The p-value for each difference estimate test ($X > Z$) is computed via permutation (10^3 shuffles). **(D)** Histogram of p-values for simulations with 2 s.d. Gaussian noise with varying number of stimuli (100, 300) and features (100, 1000). RSA detects the correct relationship less than CVR across this wide range of simulations. A total of 25,000 p-values are plotted. *CVR*: cross-validated L2 regularized regression. *RSA*: representational similarity analysis

Results

Figure 7C plots the statistical power as a function of p-value threshold. RSA has less statistical power than cross-validated regularized regression at every noise level. This is particularly evident for medium to high levels of noise. At the typical $p < 0.05$ threshold, it is clear that RSA has less statistical power than cross-validated regularized regression (Figure 7C).

The reason for the low statistical power of RSA is illustrated in Figure 7D. The similarity between the observed responses stimulus-by-stimulus covariance ($YY^\top$) and the ground-truth features ($\Phi\Phi^\top$) differs by the amount of noise (σ) and how orthogonal the ground-truth population feature weight covariance (Σ_ω) is. The latter can be thought of as a distance to an identity matrix (i.e. $d(\Sigma_\omega, I)$). In the same way, the similarity between the candidate feature space stimulus-by-stimulus covariance ($XX^\top$) and the observed responses covariance ($YY^\top$) depends on the noise level (σ), the distance to the ground-truth feature space ($d(\Phi, X)$), and the orthogonality of the empirical feature weight covariance matrix ($d(\hat{\Sigma}_\beta, I)$). All else being equal, the distance between a misleading feature space (Z) and the observed responses might be small by chance. Ultimately, RSA can yield the wrong conclusion about brain representations because of its assumption of orthogonality of the population feature weights covariance.

6. DISCUSSION

We have shown that RSA makes a strong assumption about the population feature weight matrix, namely that it is close to orthogonal. When this assumption is met, RSA is able to detect relationships between representational spaces and brain responses. However, when this assumption is violated, RSA can fail to find the statistical relationship. This is particularly worrisome when RSA is used to compare representational spaces. In such cases, when the assumption is better for one feature space than another, it can lead to the wrong conclusion.

6.1. ***Encoding models provide a direct answer to the first order question***

Encoding models explicitly test the relationship between feature spaces and brain responses. Encoding models learn the representational space relevant for any population of responses by estimating a the features that are important for individual responses. Researchers can directly assess what features are represented in which units within the population, and construct a representational space from the population of interest [6, 7]. The voxel-wise encoding model paradigm is a powerful technique that avoids spatial priors and allows inference at both the individual or population voxel level [16, 11, 6].

An important part of exploring brain representations is assessing which specific features are represented in a given brain region. Classically, this is referred to as "tuning." RSA cannot provide an answer to this question and is a key limitation. RSA can only state whether the unweighted candidate representational space *as a whole* can be said to be capture the stimulus-by-stimulus covariance of the measured responses. As we show in our simulations, has low statistical power and can yield misleading answers when comparing representational spaces. An encoding model approach gives us weight estimates which can be directly interpreted as a measure of feature importance in a given voxel. By analyzing weights from any subset of the population, we can inspect the model and make inferences about how the feature space is represented in the brain [6, 1]. That is, we can learn the representational space from the brain responses.

Furthermore, voxel-wise encoding models explicitly state the assumptions made. When using L2 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 [7]. These priors can be compared using standard statistical learning techniques [2] or Bayesian approahces (REF bayes factor). The voxel-wise encoding model is truly data-driven. One limitation is that voxel-wise encoding models require much more data than is common for a typical cognitive neuroscience experiment. However, larger high quality datasets are worth the cost. Inferring representational spaces from small stimulus sets and unstated assumptions is a risky endeavour. We hope our work shows that making inferences about representational spaces with RSA should be taken with caution.

Appendix A. Relationship between matrix trace and RSA

We begin by expanding out the definition of $RSA(\Omega_Y, \Omega_X)$ when Y and X are row-wise z-scored. This becomes the pearson correlation of the triangle matrices where each entry is the dot-product of the voxel response stimulus pairs and feature vector stimulus pairs

$$RSA_\rho(\Omega_Y, \Omega_X) =$$

$$\rho\left(\begin{bmatrix} \frac{\langle y^1, y^2\rangle}{v} & \frac{\langle y^1, y^3\rangle}{v} & \cdots & \frac{\langle y^1, y^{n-1}\rangle}{v} & \frac{\langle y^1, y^n\rangle}{v} \\ & \frac{\langle y^2, y^3\rangle}{v} & \cdots & \frac{\langle y^2, y^{n-1}\rangle}{v} & \frac{\langle y^2, y^n\rangle}{v} \\ & & \cdots & \vdots & \vdots \\ & & & \frac{\langle y^{n-2}, y^{n-1}\rangle}{v} & \frac{\langle y^{n-2}, y^n\rangle}{v} \\ & & & & \frac{\langle y^{n-1}, y^n\rangle}{v} \end{bmatrix}, \begin{bmatrix} \frac{\langle x^1, x^2\rangle}{p} & \frac{\langle x^1, x^3\rangle}{p} & \cdots & \frac{\langle x^1, x^{n-1}\rangle}{p} & \frac{\langle x^1, x^n\rangle}{p} \\ & \frac{\langle x^2, x^3\rangle}{p} & \cdots & \frac{\langle x^2, x^{n-1}\rangle}{p} & \frac{\langle x^2, x^n\rangle}{p} \\ & & \cdots & \vdots & \vdots \\ & & & \frac{\langle x^{n-2}, x^{n-1}\rangle}{p} & \frac{\langle x^{n-2}, x^n\rangle}{p} \\ & & & & \frac{\langle x^{n-1}, x^n\rangle}{p} \end{bmatrix}\right),$$

where $y^j \in \mathbb{R}^v$ and $x^j \in \mathbb{R}^p$ are the voxel and feature representations for stimulus $j \in n$, respectively. Next we define the mean and standard deviation of the similarity matrices upper triangles omitting the diagonal:

$$\begin{aligned} \bar{\Omega}_Y &= mean\left(triang\left(\Omega_Y\right)\right) \\ \Omega_Y^\sigma &= S.D.\left(triang\left(\Omega_Y\right)\right) \end{aligned}$$

We can express the RSA estimate as

$$\begin{aligned} RSA(\Omega_Y, \Omega_X) &= \left(\frac{n^2-n}{2}\right)\frac{1}{2}\Bigg[\sum_{j,k}^{n}\left(\frac{\langle y^j, y^k\rangle - \bar{\Omega}_Y}{\Omega_Y^\sigma}\right)\left(\frac{\langle x^j, x^k\rangle - \bar{\Omega}_X}{\Omega_X^\sigma}\right) \\ &\quad - \sum_{j}^{n}\left(\frac{\langle y^j, y^j\rangle - \bar{\Omega}_Y}{\Omega_Y^\sigma}\right)\left(\frac{\langle x^j, x^j\rangle - \bar{\Omega}_X}{\Omega_X^\sigma}\right)\Bigg] \\ &= \left(\frac{n^2-n}{2}\right)\frac{1}{2}\Bigg[\sum_{j,k}^{n}\left(\frac{\left(YY^\top\right)_{j,k} - \bar{\Omega}_Y}{\Omega_Y^\sigma}\right)\left(\frac{\left(XX^\top\right)_{j,k} - \bar{\Omega}_X}{\Omega_X^\sigma}\right) \\ &\quad - \sum_{j}^{n}\left(\frac{\left(YY^\top\right)_{j,j} - \bar{\Omega}_Y}{\Omega_Y^\sigma}\right)\left(\frac{\left(XX^\top\right)_{j,j} - \bar{\Omega}_X}{\Omega_X^\sigma}\right)\Bigg] \end{aligned}$$

Because the matrix diagonal is identity, the expression simplifies to

$$\begin{aligned} RSA(\Omega_Y, \Omega_X) &= \left(\frac{n^2-n}{2}\right)\frac{1}{2}\Bigg[\sum_{j,k}^{n}\left(\frac{\left(YY^\top\right)_{j,k} - \bar{\Omega}_Y}{\Omega_Y^\sigma}\right)\left(\frac{\left(XX^\top\right)_{j,k} - \bar{\Omega}_X}{\Omega_X^\sigma}\right) \\ &\quad - n\left(\frac{1-\bar{\Omega}_Y}{\Omega_Y^\sigma}\right)\left(\frac{1-\bar{\Omega}_X}{\Omega_X^\sigma}\right)\Bigg] \end{aligned}$$

Rearranging terms

$$\begin{aligned} \left(\frac{4}{n^2-n}\right)RSA(\Omega_Y, \Omega_X) + n\left(\frac{1-\bar{\Omega}_Y}{\Omega_Y^\sigma}\right)\left(\frac{1-\bar{\Omega}_X}{\Omega_X^\sigma}\right) &= \sum_{j,k}^{n}\left(\frac{\left(YY^\top\right)_{j,k} - \bar{\Omega}_Y}{\Omega_Y^\sigma}\right)\left(\frac{\left(XX^\top\right)_{j,k} - \bar{\Omega}_X}{\Omega_X^\sigma}\right) \\ &= \frac{\sum_{j,k}^{n}\left(YY^\top\right)_{j,k}\left(XX^\top\right)_{j,k} - n\bar{\Omega}_Y\bar{\Omega}_X}{n\Omega_Y^\sigma\Omega_X^\sigma} \end{aligned}$$

$$\left[\left(\frac{4}{n^2-n}\right) RSA\left(\Omega_Y,\Omega_X\right) + n\left(\frac{1-\bar{\Omega}_Y}{\Omega_Y^\sigma}\right)\left(\frac{1-\bar{\Omega}_X}{\Omega_X^\sigma}\right)\right] n\Omega_Y^\sigma\Omega_X^\sigma + n\bar{\Omega}_Y\bar{\Omega}_X = \sum_{j,k}^{n}\left(YY^\top\right)_{j,k}\left(XX^\top\right)_{j,k}$$

$$\left[\left(\frac{4}{n^2-n}\right) RSA\left(\Omega_Y,\Omega_X\right) + n\left(\frac{1-\bar{\Omega}_Y}{\Omega_Y^\sigma}\right)\left(\frac{1-\bar{\Omega}_X}{\Omega_X^\sigma}\right)\right] n\Omega_Y^\sigma\Omega_X^\sigma + n\bar{\Omega}_Y\bar{\Omega}_X = trace\left(YY^\top XX^\top\right)$$

However, even these scaling differences are problematic and can cause RSA estimates to be unreliable when used to infer brain representations [14].

Appendix B. Mixed-RSA

Appendix B.1. Mixed-RSA with STA estimates

There are many ways to estimate the population weights $\hat{\beta} \in \mathbb{R}^{p\times v}$. For linear models, one of the simplest solutions is the stimulus triggered average (STA). This can be expressed as

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

and the columns of X have mean zero. We can ignore the $\frac{1}{N}$ factor for simplicity without loss of generality. We can then expand the terms necessary to compute the trace approximation of the RSA estimate

$$\begin{aligned}
\hat{Y} &= X\hat{\beta}_{STA} \\
\hat{Y} &= X\left(X^\top Y\right) \\
\hat{Y}\hat{Y}^\top &= \left(X\left(X^\top Y\right)\right)\left(\left(Y^\top X\right)X^\top\right) \\
YY^\top\hat{Y}\hat{Y}^\top &= YY^\top\left(X\left(X^\top Y\right)\right)\left(\left(Y^\top X\right)X^\top\right) \\
YY^\top\hat{Y}\hat{Y}^\top &= \left(YY^\top XX^\top\right)\left(YY^\top XX^\top\right)
\end{aligned}$$

In "mixed RSA", when the feature weights are estimated via STA the estimate becomes

$$\begin{aligned}
RSA\left(YY^\top,\hat{Y}\hat{Y}^\top\right) &\propto trace\left(YY^\top\hat{Y}\hat{Y}^\top\right) \\
RSA\left(YY^\top,\hat{Y}\hat{Y}^\top\right) &\propto trace\left(YY^\top XX^\top YY^\top XX^\top\right) \text{ STA predictions} \\
RSA\left(YY^\top,\hat{Y}\hat{Y}^\top\right) &\propto trace\left(\left(YY^\top XX^\top\right)\left(YY^\top XX^\top\right)\right) \text{ assoc} \\
RSA\left(YY^\top,\hat{Y}\hat{Y}^\top\right) &\propto trace\left(\left(YY^\top XX^\top\right)^2\right) \text{ matrix power}
\end{aligned}$$

By trace properties, this is equivalent to

$$RSA_{STA}\left(YY^\top,\hat{Y}\hat{Y}^\top\right) \propto \sum_i \lambda_i^2 \text{ trace of matrix powers}$$

where λ_i is again the ith eigenvalue of

$$YY^\top XX^\top = Q\ \mathrm{diag}\left\{\lambda_1,\ldots,\lambda_m\right\}Q^\top$$

This leads to an interesting observation in the case of STA, that the "mixed RSA" and the "fixed RSA" estimates are simply related by their square. It seems unlikely that this difference is useful for statistical inteference.

Appendix B.2. Mixed-RSA for arbitrary estimates

In general, we can express the similarity matrix of the predictions from a linear model and the mixed RSA estimate as

$$\begin{aligned} \hat{Y} &= X\hat{\beta} \\ \hat{Y}\hat{Y}^\top &= X\hat{\beta}\hat{\beta}^\top X^\top \\ RSA\left(YY^\top, \hat{Y}\hat{Y}^\top\right) &\propto trace\left(YY^\top\hat{Y}\hat{Y}^\top\right) \end{aligned}$$

$$\begin{aligned} trace\left(YY^\top\hat{Y}\hat{Y}^\top\right) &= trace\left(Y\left(Y^\top\hat{Y}\right)\hat{Y}^\top\right) \text{ assoc prop} \\ trace\left(YY^\top\hat{Y}\hat{Y}^\top\right) &= trace\left(\left(Y^\top\hat{Y}\right)\left(\hat{Y}^\top Y\right)\right) \text{ cyclic prop} \\ trace\left(YY^\top\hat{Y}\hat{Y}^\top\right) &= trace\left(\left(Y^\top\hat{Y}\right)\left(Y^\top\hat{Y}\right)^\top\right) \text{ transpose of prods} \\ trace\left(YY^\top\hat{Y}\hat{Y}^\top\right) &= trace\left(\left(Y^\top\hat{Y}\right)^2\right) \text{ matrix squared} \\ trace\left(YY^\top\hat{Y}\hat{Y}^\top\right) &= \sum_i \lambda_i^2 \text{ trace of matrix powers} \end{aligned}$$

where λ_i is the ith eigenvalue of from the eigen-decomposition

$$Y^\top\hat{Y} = Q \text{ diag}\left\{\lambda_1, \ldots, \lambda_v\right\} Q^\top.$$

Assuming Y and $\hat{Y}$ are column-wise zero-mean and unit variance, we can express the mean of the actual and predicted response correlations for all v voxels as

$$\frac{1}{v}\sum_i^v corr\left(y_i, \hat{y}_i\right) = \frac{1}{v}\sum_i^v \frac{1}{n}(Y^\top\hat{Y})_{i,i}$$

We can see that this equivalent to computing the matrix trace

$$\begin{aligned} \frac{1}{v}\sum_i^v corr\left(y_i, \hat{y}_i\right) &= \frac{1}{v \times n} trace\left(Y^\top\hat{Y}\right) \\ \frac{1}{v}\sum_i^v corr\left(y_i, \hat{y}_i\right) &= \frac{1}{v \times n}\sum_i^v \lambda_i \end{aligned}$$

where again λ_i is the ith eigenvalue of $Y^\top\hat{Y}$. And so we have that mixed RSA has a very straight forward relationship to the mean of the population of the squared correlation

$$RSA\left(YY^\top, \hat{Y}\hat{Y}^\top\right) \propto \frac{1}{v}\sum_i^v corr\left(y, \hat{y}_i\right)^2$$

Appendix C. Generative model of stimulus-evoked responses

We now explore the conditions under which RSA can work, and in which cases it can fail (Figure **??**).

$$\hat{Y}\hat{Y}^\top = X\hat{\Lambda}_\beta X^\top,$$

where $\hat{\Lambda}_\beta = \hat{\beta}\hat{\beta}^\top$. We can think of the true feature weight covariance matrix Λ_β as being sampled from some distribution. We can model the covariance of the population weights $\Lambda_\beta \in \mathbb{R}^{p\times p}$ as a positive semidefinite real matrix. Without loss of generality, we can assume that Λ_β is sampled from a Wishart distribution

$$\Lambda_\beta \sim \mathcal{W}_p\left(V_\beta, N\right).$$

This means that the population weight covariance $\Lambda_\beta \in \mathbb{R}^{p\times p}$ is sampled from an underlying true population covariance matrix $V_\beta \in \mathbb{R}^{p\times p}$.

The individual population weights $\beta_i \in \mathbb{R}^p$ are in turn sampled from a multivariate distribution with covariance equal to Λ_β. The most widely assumed distributions are the Laplace and Gaussian distributions. For simplicity and without loss of generality, we will assume a zero-mean multivariate normal Gaussian distribution. This imposes a multivariate normal distribution on the individual population weights $\beta_i \in \mathbb{R}^p$

$$\beta_i \sim \mathcal{N}_p\left(0_p, \Lambda_\beta\right)$$

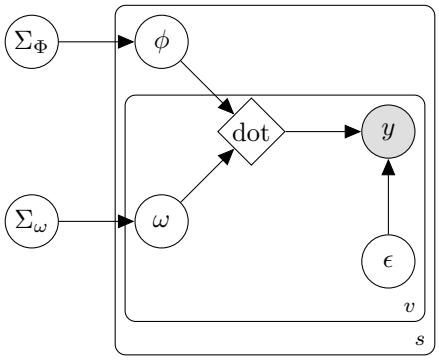

Figure C.8: **A simple generative model of voxel responses.** The voxel response to a stimulus is generated as a linear combination of the ground-truth features plus Gaussian noise, $y = \phi\omega + \epsilon$. We further assume that the stimulus ground-truth features, ϕ, and the voxel feature weights, ω, are drawn from multivariate normal distributions.

In order to clearly state the assumptions of RSA, we first present a simple generative model of voxel responses (Fig. C.8). We assume that the response of voxel i is a function, $\mathcal{F}$, of the stimulus, s, other non-stimulus driven factors, C, and noise, ϵ_i:

$$y_i^s = \mathcal{F}\left(j, C\right) + \epsilon_i.$$

For the rest of the paper we will consider only the stimulus-driven part of the evoked brain responses.

In this framework, the stimulus-response model becomes

$$y_i^s = f\left(s\right) + \epsilon_i.$$

We then assume that f is a linearizable function, $\mathbb{L}$, such that $\mathbb{L}(s) = \phi^s \in \mathbb{R}^p$. We can then model responses as a linear combination of the stimulus features:

$$y_i^s = \phi^s\omega_i + \epsilon_i,$$

where $\omega_i \in \mathbb{R}^p$ is the vector of feature weights for voxel i. These are known as linearized encoding models [16].

In what follows we assume this simple generative model. Stimulus-evoked responses are generated from a linear combination of the ground-truth stimulus features $\Phi \in \mathbb{R}^{n\times p}$

$$y_i = \Phi\omega_i + \epsilon_i.$$

One caveat is that the experimenter does not have access to the ground-truth features, Φ, nor their covariance, Λ_Φ. Instead, the experimenter only has a guess as to what these features might be.

Appendix D. Implicit RSA assumption on feature weights

The ground-truth stimulus-by-stimulus second moment $YY^\top \in \mathbb{R}^{n \times n}$ for a set of voxel responses $Y \in \mathbb{R}^{n \times v}$ becomes

$$YY^\top = \Phi \Lambda_\omega \Phi^\top + EE^\top, \tag{D.1}$$

where $\Lambda_\omega = \omega\omega^\top$ is the covariance of the feature weights across the v voxels used in the analysis and E is i.i.d. noise. Recall the relationship between the representational similarity matrix Ω_Y and the row-size z-scored second moment of the Y responses:

$$\Omega_Y = \frac{YY^\top}{v}$$

This equation illustrates the main point of this paper: The strength of the relationship between the brain response similarity matrix, Ω_Y, and the feature space, Φ, depends on the underlying feature weight covariance matrix Λ_ω.

In RSA, feature weights are not estimated. Instead, brain response similarity is compared directly to feature space similarity, $\Sigma_\Phi = \Phi\Phi^\top$. From Equation D.1 we see that this comparison assumes that the feature weight covariance matrix is identity, $\Sigma_\omega = I_p$, leaving $\Phi\Sigma_\omega\Phi^\top = \Phi\Phi^\top = \Sigma_\Phi$. For this assumption to be true, the vector of weights for each feature across voxels $\omega^j \in \mathbb{R}^v$ must be orthogonal to that for every other feature:

$$\left\langle \omega^j, \omega^{k \neq j} \right\rangle = 0.$$

This assumption flies in the face of empirical findings that feature weight matrices tend to be dominated by low rank structure in a variety of settings (REF: Huth2012,2016). When the feature weight matrix has low rank structure, the statistical power of RSA will be reduced relative to when the feature weight vectors are orthogonal.

Appendix E. Effect of feature weight matrix effective rank on RSA

To illustrate the effect of weight matrix rank on RSA statistical power, we conduct a number of simulations (Fig. E.9). First, we draw stimulus features from a multivariate normal distribution with a specified low effective rank feature covariance:

$$\phi^j \sim \mathcal{N}_p\left(0, \Sigma_\Phi\right).$$

We choose low rank features because in most experimental designs with naturalistic stimuli, there are very strong correlations among features (REF Lescroart). This particular choice does not affect our findings however (see Figure 1; TODO). Voxel responses to the stimuli are then generated from a linear combination of the ground-truth stimulus features determined by the voxel's feature weights, ω_i

$$y_i = \Phi\omega_i + \epsilon_i.$$

The feature weights for each voxel are also sampled using a multivariate normal distribution

$$\omega_i \sim \mathcal{N}_p\left(0, \Sigma_\omega\right)$$

In a Bayesian framework, we can think of the feature weight matrix covariance, Σ_ω, as being drawn from a Wishart distribution. In our simulations, we sample Σ_ω from a Wishart distribution with covariance $\Lambda_\omega \in \mathbb{R}^{p \times p}$. We generate a range of Λ_ω matrices that vary in effective rank.

$$\Sigma_\omega \sim W_p\left(\Lambda_\omega\right)$$

We achieve low effective rank by skewing the distribution of the singular values. We generate singular values as (REF sklearn):

$$SV_i = (1-\alpha)\, e^{-\gamma\left(\frac{i-1}{rank}\right)^2} + \alpha e^{-\gamma\left(\frac{i-1}{rank}\right)},$$

where i is the singular value index, $rank$ controls the effective rank, γ the decay of the singular values, α relative importance of the tiny singular values.

This allows us to precisely explore how the statistical power of RSA is affected by the rank of the voxel weight matrix.

Finally, we sample a feature space stimulus representation, x^j, by first sampling a covariance matrix, Σ_X, from a Wishart distribution centered on the ground-truth feature space covariance, Σ_Φ:

$$\Sigma_X \sim W_p\left(\Sigma_\Phi\right)$$

$$x^j \sim \mathcal{N}_p\left(0, \Sigma_X\right)$$

We take this approach because in a typical experiment the researcher does not have access to the ground-truth features. Rather, the researcher has a computational model that has some (unknown) amount of similarity to the ground-truth.

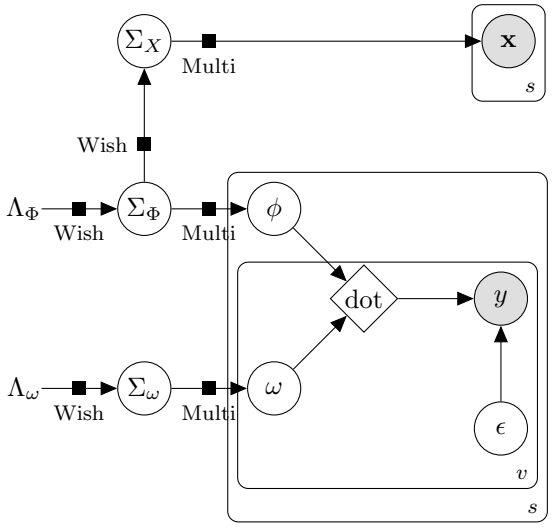

Figure E.9: **Generative model used to explore the effect of weight matrix rank on RSA.** The voxel response to a stimulus is generated as a linear combination of the ground-truth features plus Gaussian noise, $y = \phi\omega + \epsilon$. We further assume that the researcher does not have access to the ground-truth features, ϕ. Instead, the researcher has a candidate feature space, X, (e.g. a computational model) that has some similarity to the ground-truth feature space. This candidate feature space is drawn from a Wishart distribution with the same covariance as the ground-truth feature space covariance, Σ_Φ. The ground-truth feature weights, ω, are sampled from a multivariate normal distribution with an effective low rank covariance matrix. This reduces the number of important feature space dimensions that matter to a voxel. We explore how RSA is affected as less and less ground-truth feature space dimensions are important.

After generating the data, we conduct representational similarity analysis as described in Section SEC-NUM. For each simulation, we assessed the statistical significance of RSA by shuffling the RSM matrix 10^3 times (REF RSA pappers). We boostrap the RSM correlation estimate and compare the bootstrap mean against the null distribution to obtain a p-value. Our simulations span a wide set of parameters. We ranged number of stimuli (100, 1,000), features (100, 1,000), voxels (128, 256, 512), noise levels (1, 2, 3, 4, 5, 6; Gaussian s.d.), feature matrix effective rank (1, 5, 10, 20), and weight matrix effective rank (1, 2, 4, 8, 16, 32). A total of 4,350 simulations for each of the six noise levels were performed.

Appendix F. Simulations to assess the statistical power of feature space comparisons

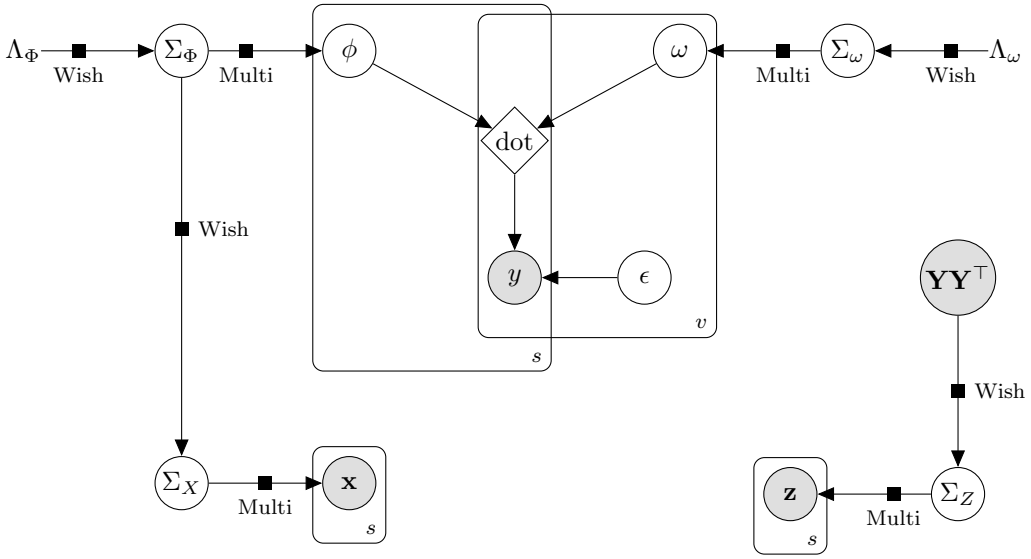

Figure F.10: **Graphical model used to generate data for model comparison power analysis.** We generate a voxel's response to a stimulus as linear combination of the ground-truth features, ϕ. We then sample stimulus feature representations that are similar either to the ground-truth features, x, or to the empirical stimulus-by-stimulus response covariance, z. These generated data were used to assess the statistical power of RSA and cross-validated L2-regularized regression.

Next, we use simulations to assess the statistical power of RSA and encoding models in a model comparison setting (Fig. F.10). We simulate voxel responses, $Y \in \mathbb{R}^{s \times v}$, from a linear combination of ground-truth features, $\Phi \in \mathbb{R}^{s \times p}$, determined by the feature weight matrix $\boldsymbol{W} \in \mathbb{R}^{p \times v}$:

$$\boldsymbol{Y} = \boldsymbol{\Phi\omega} + \boldsymbol{E},$$

where $E \in \mathbb{R}^{s \times v}$ is zero mean iid Gaussian noise. Wwe again sample the voxel feature weights, $\omega_i \in \mathbb{R}^p$, from a multivariate normal distribution centered on a covariance matrix $\Sigma_\omega \in \mathbb{R}^{p \times p}$ of low effective rank:

$$\omega_i \sim \mathcal{N}_p\left(0, \boldsymbol{\Sigma_\omega}\right).$$

This is important because the distance between Σ_Y and Σ_Φ is in large part determined by Σ_ω.

We generate a candidate feature space $X \in \mathbb{R}^{s \times p}$ by first sampling a feature covariance matrix $\Sigma_X \in \mathbb{R}^{p \times p}$ from a Wishart distribution centered around the ground-truth feature covariance, $\Sigma_\Phi \in \mathbb{R}^{p \times p}$:

$$\boldsymbol{\Sigma_X} \sim W_p\left(\gamma_x \boldsymbol{\Sigma_\Phi}\right),$$

where γ_x controls the similarity to Σ_Φ. We then use Σ_X to sample feature representations $x^j \in \mathbb{R}^p$ for each stimulus j from a multivariate normal distribution:

$$x^j \sim \mathcal{N}_p\left(0, \boldsymbol{\Sigma_X}\right).$$

For the alternate feature space $Z \in \mathbb{R}^{s \times p}$, we sample a stimulus-by-stimulus covariance matrix $\Sigma_Z \in \mathbb{R}^{s \times s}$ that is similar to the empirical response covariance $\Sigma_Y \in \mathbb{R}^{s \times s}$

$$\boldsymbol{\Sigma_Z} \sim W_s\left(\gamma_z \boldsymbol{\Sigma_Y}\right),$$

where γ_z controls the similarity. This captures how misleading the alternate feature space is. Finally, we sample stimulus representations $z_k \in \mathbb{R}^s$ with stimulus-by-stimulus covariance Σ_Z:

$$z_k^\top \sim \mathcal{N}_s\left(0, \boldsymbol{\Sigma_Z}\right).$$

The data generated by the simulations vary in the number of stimuli (100, 300), features (100, 1000), voxels (128, 256, 512), noise level (1, 2, 3, 4, 5, 6), weight matrix effective rank (1, 3, 5, 7, ..., 32), similarity between the ground-truth feature space, ϕ, and the candidate feature space, x, (10^{-3} to 1; 14 log-spaced samples), and the similarity between the alternate feature space, z, and the empirical voxel responses $YY^\top$ (10^{-5} to 1; 10 log-spaced samples). A total of 25,000 simulations for each of the six noise levels were run. Statistical significance for both RSA and encoding models was assessed as in the fMRI experiment.

References

[1] Tolga Çukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature neuroscience*, 16(6):763–770, 2013.

[2] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[3] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.

[4] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[5] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[6] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack 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.

[7] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[8] Seyed-Mahdi Khaligh-Razavi, Linda Henriksson, Kendrick Kay, and Nikolaus 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.

[9] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008.

[10] Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008.

[11] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[12] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Academic press, 2011.

[13] Dustin E Stansbury, Thomas Naselaris, and Jack L Gallant. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5):1025–1034, 2013.

[14] Bertrand Thirion, Fabian Pedregosa, Michael Eickenberg, and Gaël 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.

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

[16] Michael C-K Wu, Stephen V David, and Jack L Gallant. Complete Functional Characterization of Sensory Neurons By System Identification. *Annual Review of Neuroscience*, 29(1):477–505, January 2006. ISSN 0147-006X.