Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis

Jörn Diedrichsen[1] & Nikolaus Kriegeskorte[2]

1. *Brain and Mind Institute, Department for Computer Science, Department for Statistical and Actuarial Science, Western University, Canada*
2. *Cognitive and Brain Sciences Unit, Cambridge University, UK*

Address correspondence:

Jörn Diedrichsen

Brain Mind Institute

Natural Science Center

Western University

London, Ontario, N6A 5B7

Canada

Email: jdiedric@uwo.ca

Abstract

Representational models specify how activity patterns in populations of neurons (or, more generally, in multivariate brain-activity measurements) relate to sensory stimuli, motor responses, or cognitive processes. In an experimental context, representational models can be defined as hypotheses about the distribution of activity profiles across experimental conditions. Previous studies have used three different methods to test such hypotheses: encoding analysis, pattern component modeling (PCM), and representational similarity analysis (RSA). Here we develop a common mathematical framework for understanding the relationship of these three methods, which all share one core commonality: all three evaluate the second moment of the distribution of activity profiles, which determines how well any feature can be linearly decoded from population activity. Using simulated data for three different experimental designs, we compare the power of the methods to adjudicate between competing representational models. PCM implements a likelihood-ratio test and therefore provides the most powerful test if its assumptions hold. However, the other two approaches – when conducted appropriately – can perform similarly. In encoding analysis, the linear model needs to be appropriately regularized, which effectively imposes a prior on the activity profiles. With such a prior, an encoding model specifies a well-defined distribution of activity profiles. In RSA, the unequal variances and statistical dependencies of the dissimilarity estimates need to be taken into account to enable near-optimal inference. The three methods render different aspects of the information explicit (e.g. single-response tuning in encoding analysis and population-response representational dissimilarity in RSA) and have specific advantages in terms of computational demands, ease of use, and extensibility. The three methods are properly construed as complementary components of a comprehensive data-analytical toolkit for understanding neural representations on the basis of multivariate brain-activity data.

Author Summary

Modern neuroscience can measure activity of many neurons or the local blood oxygenation of many brain locations simultaneously. As the number of simultaneous measurements grows, we can better investigate how the brain represents and transforms information, to enable perception, cognition, and behavior. Recent studies go beyond showing *that* a brain region is involved in some function. They use representational models that specify *how* different perceptions, cognitions, and actions are encoded in brain-activity patterns. In this paper. we provide a general mathematical framework for such representational models, which clarifies the relationships between three different methods that are currently used in the neuroscience community. All three methods evaluate the same core feature of the data, but each has distinct advantages and disadvantages. Pattern component modelling (PCM) implements the most powerful test between models, and is analytically tractable and expandable. Representational similarity analysis (RSA) provides a highly useful summary statistic (the dissimilarity) and enables model comparison with weaker distributional assumptions. Finally, encoding models characterize individual responses and enable the study of their layout across cortex. We argue that these methods should be considered components of a larger toolkit for testing hypotheses about the way the brain represents information.

Introduction

The measurement of brain activity is rapidly advancing in terms of spatial and temporal resolution, and in terms of the number of responses that can be measured simultaneously [1]. Modern electrode arrays and calcium imaging enable the recording of hundreds of neurons in parallel. Electrophysiological signals that reflect summaries of the population activity can be recorded using both invasive (e.g. the local field potential, LFP) and non-invasive techniques (e.g. scalp electrophysiological measurements) at increasingly high spatial resolution. Modern functional

magnetic resonance imaging (fMRI) enables us to measure hemodynamic activity in hundreds of thousands of voxels across the entire human brain at sub-millimeter resolution.

In order to translate advances in brain-activity measurement into advances in computational theory [2], researchers increasingly seek to test representational models that capture both what information is represented in a population of neurons, and how it is represented. Knowing the content and format of representations provides strong constraints for computational models of brain information processing. We refer to hypotheses about the content and format of brain representations as *representational models*. We address here the important methodological question of how to best test representational models.

Referring to an activity pattern as a "representation" of some variable (such as a perceptual property, some cognitive content, or an action parameter) implies (1) that there is information about the variable present in the pattern of activity and (2) that this information *serves a functional purpose* for further processing [3]. Even though observational methods such as fMRI or single-cell recording can generally not be used to test (2) directly, we can test whether the variable is encoded in such a form that makes it easily accessible to read out. This criterion leads to our working definition of an *explicit neural representation*, on which our specific formulation of a "representational model" is based: A variable is *explicitly represented* in an area, if it can be *linearly* decoded from the neural activity pattern in the area [4]. To understand why the limitation to linear decoding makes sense, consider the following example: When looking at a face, the firing pattern of retinal neurons contains *information* about the identity of the person, as it can be read out from using a highly non-linear decoder. However, the face identity is *implicit* in the retinal representation, requiring the decoder, in essence, to perform face recognition. For a population code to constitute an *explicit representation*, another area must be able to read out the represented variable directly, for example as a weighted sum of the neural activity pattern across neurons in the region [2, 4, 5]. Note that this definition does not restrict representations to highly localized codes,

such as the "grandmother neuron" [6], but encompasses any distributed code, that can be read out by computing a linear combination of neuronal activities.

This definition of explicit representation has motivated researchers to use linear decoding methods to study neural representations [7-9]. Decoding analysis can establish that a given feature is explicitly represented in the activity pattern. Representational models, as considered here, go one step further: they fully characterize the representational geometry, which defines *all* explicitly represented features in a region, how strongly each of them is represented (relative signal to noise ratio), and how the activity patterns associated with different features relate to each other. Representational models therefore fully specify the explicit representational content of an area.

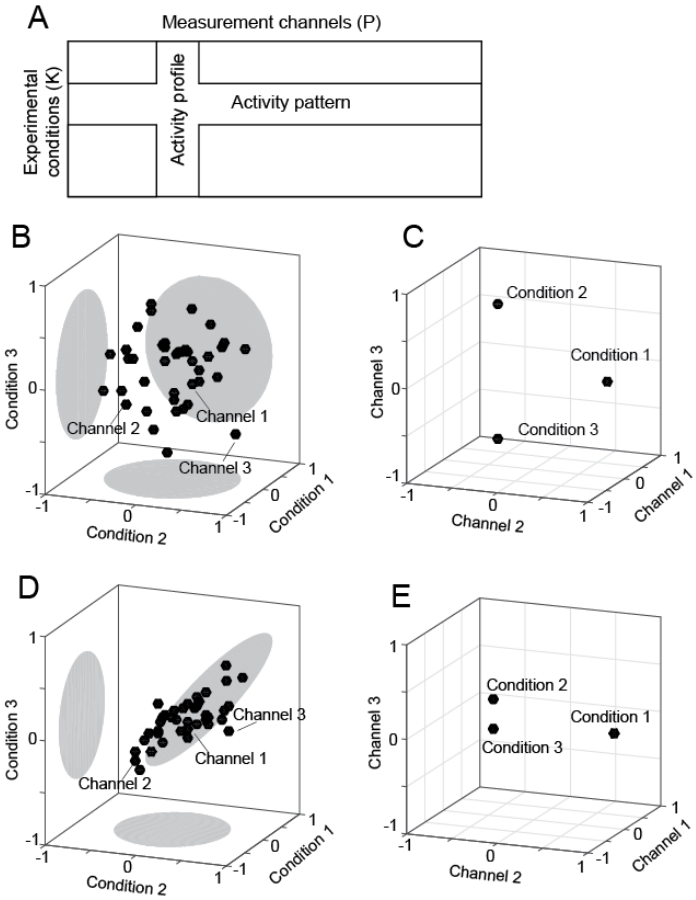

Figure 1. Two complementary perspectives on population activity. *(**A**) The multivariate activity data can be viewed as a set of activity profiles (columns) or as a set of activity patterns (rows). An activity profile is a vector of responses of a single channel across experimental conditions. An activity pattern is a vector of responses across all channels for a single condition. Activity data can be visualized by plotting activity profiles as points in a space defined by the experimental conditions (B,D), or by plotting the activity patterns as points in a space defined by the measurement channels (C,E). (**B**) If the activities are uncorrelated between conditions, then (**C**) the corresponding activity patterns of all three conditions are equidistant to each*

*other, and can be equally well distinguished. (**D**) If the activities are positively correlated for two conditions (conditions 2 and 3 here), then (**E**) the activity patterns for these conditions are closer to each other and can be less well distinguished.*

To define representational models formally, we need to consider two complementary perspectives on activity data, as illustrated in Fig. 1. The activity of many neurons, or more generally *measurement channels* (neurons, electrodes, or fMRI voxels), can be measured across a range of *experimental conditions* (stimuli, movements, or tasks). Thus, each channel will have an *activity profile,* which can be plotted as a point in the space spanned by the experimental conditions (Fig. 1b). A representational model specifies a *probability distribution of activity profiles* in the space spanned by the experimental conditions – i.e. it treats the true activity profiles as a random variable. In other words, a representational model predicts, for each possible activity profile, the probability of observing a measurement channel exhibiting that profile. However, it does not predict the activity profile for each individual channel actually measured. The motivation for this approach derives from the idea that the computational function of a region does not depend on specific neurons having specific response properties, but on the fact that certain features can be read out from the population by downstream neurons. The probability distribution over activity profiles determines which features can be linearly read out from the code and the signal-to-noise ratio of the readout. By basing further analyses on the probability distribution of the activity profiles, we are disregarding three aspects of the code: (1) which neuron fulfills which function, (2) where neurons are located on the cortical sheet, and (3) the degree to which the information about a given represented feature is concentrated in a few neurons (as in single-cell selectivity for a represented feature) or spread out over the population. Ignoring these aspects may be viewed as an advantage or a disadvantage, depending on the level of description that a researcher is interested in. We argue that treating activity profiles as random vectors is a simplification that is useful for drawing computational insights from population activity measurements.

In this paper, we show that the signal-to-noise ratio with which any given feature can be linearly decoded is determined by a specific aspect of the distribution of the activity profiles, namely its multivariate *second moment*. We discuss three established methods for adjudicating between representational models: encoding analysis, pattern-component modeling (PCM) and representational similarity analysis (RSA, see Table 1). We show that these three techniques all exclusively rely on information contained in the second moment of the distribution of activity profiles. This core commonality enables us to consider them in the same formal framework.

Table 1: *Comparison of encoding analysis with regularization, pattern component modelling (PCM), and representational dissimilarity analysis (RSA).*

	Encoding analysis	PCM	RSA
Model definition	Model-feature matrix **M**, regularization / prior	Predicted second-moments matrix (**G**)	Representational dissimilarity matrix (RDM)
First-level parameters (characterizing individual activity profiles)	One weight per feature and measurement channel	None; integrated out in the likelihood	None; integrated out when calculating dissimilarities
Second-level parameters (characterizing the distribution of activity profiles)	Regularization / Ridge coefficient (determined by noise / signal ratio)	Scale parameter s Noise variance	Scaling between predicted and observed distances (s)
Prediction target	Responses to test conditions	Distribution of measurement channels in activity-profile space	Dissimilarities among activity patterns
Training data required	always	not for fixed models, only if additional second-level parameters are to be fitted	not for fixed models, only if additional second-level parameters are to be fitted
Explicit likelihood for fitting additional model parameters	No – need to do nested within crossvalidation	Yes	Yes
Fitting algorithms for model parameters	-	EM Gradient descent Newton-Raphson	Linear and non-negative regression IRLS

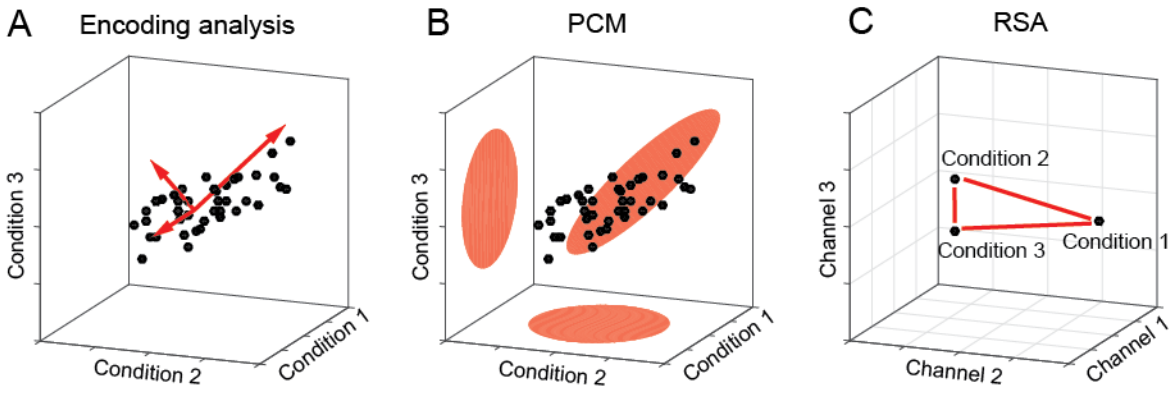

Figure 2. Three approaches to testing representational models. *(**A**) In encoding analysis, the distribution of activity profiles is described by the underlying features (red vectors). The direction of feature vector determines the associated activity profile, and the length the strength of the feature encoding in the representation. (**B**) PCM models the distribution of the activity profiles as a multivariate Gaussian. This model is parametrized by the second moment of the activity profiles, which determines at what signal-to-noise ratio any feature is linearly decodable from the population. (**C**) RSA uses the representational distances (or, more generally, dissimilarities) between activity patterns as a summary statistic to describe the underlying distribution. When the representational distances are estimates of the Mahalanobis distances (normalized by the error covariance), the distance matrix, like a second moment of the activity profiles, is a sufficient statistic that captures the linear decodability of any feature.*

In *encoding analysis* [10, 11], representational models are defined in terms of the underlying *features* (Fig. 2A). Each activity profile can be characterized by a linear combination of such features. Examples include Gabor filters [12] (a model of low-level visual representation), abstract semantic dimensions [13] (a higher-level cognitive representation), and force, direction or hand position [14-16] (a movement representation). The prevalence of each feature in each channel is measured by a feature weight. Feature weights are considered first-level parameters in our framework, as they describe the individual activity profiles, as opposed to second-level parameters that describe the distribution of the activity profiles (Table 1). The large number of parameters

(number of features in the model times number of channels in the measurements) engenders a danger of over-fitting. To account for effects of overfitting, encoding models are commonly evaluated using cross-validation: The feature weights are estimated on a training set, and the model is evaluated in terms of its performance at predicting left-out data [10]. The test data may consist in a sample of experimental conditions not used in training, so as to test the model's generalization performance [11, 12]. While many studies use simple linear regression to estimate the weights [11, 17], it is increasingly common to introduce a regularization penalty in the estimation of model weights (for example the L2 norm of the vector of weights) [12, 13]. We will see that regularization is not merely a technical trick used in fitting a given model. Instead, the regularization (and its implicit distributional assumptions) are an essential part of the representational hypothesis that is tested. Without it, encoding models do not specify a probability distribution with a defined second moment and thus do not define the linear decodability of different features.

Pattern component modeling [18] is based on an explicit generative model of the process that produced the data and can be considered a Bayesian approach. The true activity profiles are assumed to have a multivariate Gaussian distribution in the space spanned by the experimental conditions (Fig. 2B). The activity measurements are assumed to be corrupted by zero-mean additive Gaussian noise. This formulation enables us to evaluate the marginal likelihood of the observed activity profiles under the probability distribution specified by the model. Thus, we do not fit any first-level parameters (feature weights) and hence reduce the risk of overfitting. It thus enables us to compare models with different numbers of features without having to correct for model complexity. If the assumptions of the generative model hold, PCM implements the likelihood-ratio test between models [19], which by the Neyman-Pearson lemma [20], is the most powerful test of its size. In theory, therefore, PCM should yield more accurate inferences than any of its competitors, that is it should be able to more sensitively adjudicate among competing models.

Finally, *representational similarity analysis* (RSA [21-23]) approaches the problem from a complementary perspective. Rather than considering the activity profiles of the measurement

channels as points in the space spanned by the conditions (Fig. 1B,D), it considers the activity patterns associated with the experimental conditions as points in the space spanned by the measurement channels (Fig. 1C,E). RSA then uses the representational distances (Fig. 2C) between the conditions as a summary statistic. We will see that these distances again exclusively depend on the second moment of the distribution of activity profiles. Having obtained a matrix of dissimilarities between activity patterns (the representational dissimilarity matrix, RDM), RSA then tests models by comparing the observed distances to the distances predicted by each representational model. In this paper, we consider different methods to perform this comparison, including rank-based correlations [24], Pearson correlations [25], and a novel likelihood-based approach that uses a multivariate normal approximation to the joint distribution of the cross-validated Mahalanobis distances [26, 27]. As shown in our simulations here, this latter technique provides a particularly powerful means to adjudicate between representational models.

In the remainder of the paper, we first introduce the second moment of the activity profiles and explain why it is the sufficient statistic of linear decodability. We then define the three methods in detail, using a common mathematical notation. Finally, using simulated data and models taken from our fMRI work, we assess the statistical efficiency, i.e. how well these methods adjudicate between two or more competing representational models given limited data. We also compare the methods in terms of their computational efficiency.

Materials and Methods

Basic definitions

All symbols used in the following derivations are summarized in Table 2. First, we define $\mathbf{U}$ to be the matrix of noiseless activity profiles with K (number of experimental conditions) rows and P (number of measurement channels) columns. Each row of this matrix is an activity pattern, the response of the whole population to a single condition. Each column of this matrix is an activity profile (Fig. 1a).

Table 2: Notation used. For non-scalars, the second column indicates the vector / matrix size.

K:		Number of conditions
M:		Number of independent partitions of the data (imaging runs)
P:		Number of measurement channels (voxels, electrodes, neurons)
N:		Overall number of measurements ($N_m \times M$)
Q:		Number of features in model
$\mathbf{U}$:	$K \times P$	Matrix of true activation patterns
$\boldsymbol{u}_{i,\cdot}$:	$1 \times P$	Activation pattern for condition i; i^{th} row of $\mathbf{U}$
$\boldsymbol{u}_{\cdot,j}$:	$K \times 1$	Activation profile for measurement channel j; j^{th} column of $\mathbf{U}$
$\hat{\mathbf{U}}^{(m)}$:	$K \times P$	Matrix of estimated activity patterns, based on data from partition m
$\tilde{\mathbf{U}}^{(\sim m)}$:	$K \times P$	Model prediction for activity patterns, based on data independent of m
$\mathbf{M}$:	$K \times Q$	Matrix of model features for all condition
$\mathbf{W}$:	$Q \times P$	Matrix of voxel weights for each feature
$\mathbf{Y}$:	$N \times P$	Matrix of brain measurements, concatenated activity estimates or time series data
$\mathbf{Z}$:	$N \times K$	Design matrix, indicating how measurements relate to activity patterns
$\mathbf{X}$:	$N \times R$	Design matrix containing n regressors of no-interest
$\mathbf{G}$:	$K \times K$	Second moment of $\mathbf{U}$
$d_{i,k}$:		Distance between condition i and k
J:		Number of distances, normally $K(K-1)/2$
$\mathbf{D}$:	KxK	Representational dissimilarity matrix of all pairwise distances
$\mathbf{d}$:	$Jx1$	Vector of all pairwise distances
$\tilde{\mathbf{d}}$:	$Jx1$	Vector of predicted distances
$\mathbf{C}$:	JxK	Contrast matrix, defining the J pairwise differences between conditions
$\mathbf{\Sigma}_P$:	PxP	Variance-covariance matrix between the P voxels
$\mathbf{\Sigma}_K$:	KxK	Variance-covariance matrix of the columns of $\hat{\mathbf{U}}^{(m)}$
$\mathbf{V}$:	NxN	Variance-covariance matrix of $\mathbf{Y}$
$\mathbf{S}$:	JxJ	Variance-covariance matrix of all pair-wise distances

Because we are interested in the distribution of activity profiles, but not in the activity profiles per se, we consider the columns of $\mathbf{U}$ to be a random variable. This is an essential step underlying our common framework, which is justified by the fact that, for the purpose of reading out information, the different measurement channels are exchangeable (see introduction). We assume that the activity profiles are repeatedly measured, with the data consisting of M independent partitions, each containing at least one activity measurement for each condition and measurement

channel. In the context of fMRI, a partition will consist of a separate phase of data acquisition, e.g. a scanner run. The activity estimates $\widehat{\mathbf{U}}^{(m)}$ of partition *m* are the true patterns **U** plus noise $\mathbf{E}^{(m)}$. The noise captures both neural trial-by-trial variability of the activity pattern in a single condition, as well as measurement noise.

$$\widehat{\boldsymbol{U}}^{(m)} = \boldsymbol{U} + \boldsymbol{E}^{(m)}. \qquad \text{(Eq. 1)}$$

For the purposes of this paper, we assume that the noise is Gaussian, and independent and identically distributed (i.i.d.) across conditions and partitions (homoscedasticity). Possible dependence within each partition, however, can be easily accounted for [26, 28].

Dependence between measurement channels

The discussion below further assumes that the noise is also i.i.d. across different measurement channels (isotropicity). However, noise in fMRI, MEG, and even invasive electrophysiology exhibits strong correlations between neighboring locations in the brain. To account for these dependencies, we employ multivariate noise normalization (i.e. spatial prewhitening), which has been shown to increase the reliability of inference [29]. Across all measurement channels, we estimate the *P*×*P* variance-covariance matrix across trials, $\boldsymbol{\Sigma}_P$ and then regularize the estimate by shrinking it towards a diagonal matrix [30]. In the context of fMRI, we can use the residual time series from the fitting of the time-series model to estimate noise covariance [29] [31]. We then post-multiply our activity estimates by $\widehat{\boldsymbol{\Sigma}}_P^{-1/2}$, rendering the model errors in the channels approximately uncorrelated. If multivariate noise normalization is not performed or is incomplete, inference will be suboptimal in all three methods (for details see [26]).

Second moment matrix and linear decodeability

In this section, we show that the *second moment* of the activity profiles fully characterizes the linear decodability of any feature in the space spanned by the experimental conditions. This provides a motivation, from the perspective of brain computation, for using the second moment matrix as a summary statistic. While higher statistical moments may reveal some computationally

important features of the population activity (a point we will return to in the Discussion), we believe that focusing on linear encoding (and hence the second moment) constitutes the natural starting point for adjudicating among representational models.

The n[th] moment of a scalar random variable u is $E(u^n)$, where E() denotes the expected value. Here we use a multivariate extension of the concept, with the second moment of the random vector $\mathbf{u}$ defined as the matrix $E(\mathbf{u}\mathbf{u}^T)$, the expected outer product of the activity profiles, where the expectation is across the measurement channels. The second-moment matrix of the activity profiles is given by

$$\mathbf{G} \equiv \sum_{j=1}^{P} \mathbf{u}_{\cdot,j}\mathbf{u}_{\cdot,j}^T / P = \mathbf{U}\mathbf{U}^T / P \ . \qquad \text{(Eq. 2)}$$

Thus, each cell of this matrix contains the normalized inner product of two activity patterns.

Before calculating $\mathbf{G}$, some investigators subtract the mean activity across measurement channels for each condition from the data. In this case, Eq. 2 becomes the variance-covariance matrix of the activity profiles –the second moment around the mean activity profile. Here we do not remove the mean, but use the second moment around zero. From the perspective of a neuron that reads out the activity pattern of an area, any difference between activity patterns across conditions can be used to encode information. Some features (for example, stimulus intensity) may be encoded in the mean activity over all measurement channels. Other properties (for example, stimulus identity) may be encoded in relative activity differences, with some measurement channels responding more to one condition, and others to a different condition. The second moment (around zero) captures both of these potentially meaningful types of difference.

To obtain a read-out for a certain feature y for condition i, for example a stimulus property represented in the code, we would weight each channel's observed activity during condition i using the $P \times 1$ read-out vector $\mathbf{v}$,

$$\hat{y}_i = \hat{\mathbf{u}}_{i,\cdot}\mathbf{v}. \qquad \text{(Eq. 3)}$$

We would now like the estimate $\hat{y}$ to have very different values for two conditions that differ in the feature being read out, while showing small differences for trials that have the same feature value.

The feature to be read out can be defined by a specific contrast on the experimental conditions, defined by a *K×1* vector **f**. We are looking for the readout weight vector **v** that maximizes the ratio between feature variance and error variance, and thus the signal-to-noise ratio (S), of the readout:

$$S = \frac{\mathbf{v^T U^T f f^T U v}}{\mathbf{v^T E^T f f^T E v}} \quad . \qquad \text{(Eq. 4)}$$

The solution to this equation is commonly known as Fisher's linear discriminant [32], which, under the assumption of homoscedastic Gaussian noise, is the best achievable linear decoder. If the noise is isotropic (or the data is adequately pre-whitened), then $\mathbf{E^T f f^T E} = \mathbf{I}b$, where *b* is a constant. The denominator then depends only on the norm of the read-out vector **v**, not on its direction, and can be ignored when **v** is constrained to have a norm of 1. The best readout vector **v** is then given by the first eigenvector of the matrix $\mathbf{U^T f f^T U}$, and the quality of the best readout is determined by the largest eigenvalue.

The non-zero eigenvalues (*eig*) of a square matrix are invariant to cyclic permutations of the product order:

$$eig(\mathbf{U^T f f^T U}) = eig(\mathbf{f^T U U^T f}) = P\ eig(\mathbf{f^T G f}) \ . \qquad \text{(Eq. 5)}$$

Therefore, the quality of the best linear decoder for *any* feature (as defined by **f**) is fully characterized by the second moment matrix **G** of the pre-whitened activity patterns.

Representational analysis in the context of fMRI

The methods in this paper were first developed in the context of fMRI data analysis, and our examples will come from this domain. A simple way to apply the analyses to fMRI data is to use as activity estimates ($\widehat{\mathbf{U}}^{(m)}$) the regression coefficients, or "beta"-weights, from a first-level time series analysis [33, 34]. The time-series model accounts for the hemodynamic lag and the temporal autocorrelation of the noise. The activity estimates usually express the difference in activity during a condition relative to rest. Activity estimates commonly co-vary together across fMRI imaging runs, because all activity estimates within a partition are measured relative to the same resting baseline. This positive correlation can be reduced by subtracting, within each partition, the mean

activity pattern (across conditions) from each activity pattern. This makes the mean of each measurement channel (across condition) zero and thus centers the ensemble of points in activity-pattern space that is centered on the origin.

Rather than using the concatenated activity estimates from different partitions, encoding analysis and PCM can also be applied directly to time series data. As a universal notation that encompasses both situations, we can use a standard linear mixed model [35]:

$$\boldsymbol{Y} = \boldsymbol{ZU} + \boldsymbol{XB} + \boldsymbol{E}\,, \qquad \text{(Eq. 6)}$$

where **Y** is an $N \times P$ matrix of all activity measurements, **Z** the $N \times K$ design matrix, which relates the activity measurements to the K experimental conditions, and **X** is a second design matrix for nuisance variables. **U** is the $K \times P$ matrix of activity patterns (the random effects), **B** are the regression coefficients for these nuisance variables (the fixed effects), and **E** is the matrix of measurement errors. If the data **Y** are the concatenated activity estimates, the nuisance variables typically only model the mean pattern for each run. If **Y** consists of time-series data, the nuisance variables typically capture additional effects such as time-series drifts and residual head-motion-related artifacts.

Representational analysis in the context of neurophysiological recordings

All three methods can also be applied to recordings of single cell activity or neurophysiological potentials [21, 22]. The activity estimates can then be firing rates estimated over a temporal window for each trial, or other multivariate summaries of the spatiotemporal activity patterns measured, e.g. the time-frequency response pattern. Because the trial-by-trial variability of firing rates will usually increase with the mean firing rate, it is advisable to use the square root of firing rates to make the data conform better to the assumption that the variance of the noise is independent of the signal [36].

Here we focus on models that treat the activity patterns **U** as static snapshots. To exploit the temporal detail provided by electrophysiological recordings, the analyses can be either performed

using a sliding window over the time course of the trial [37-39], or by "stacking" the time series and conditions, resulting in T (timepoints) × K (condition) by P (neurons) activity matrix [e.g., 40].

Encoding analysis

An encoding model characterizes the structure of the representation in terms of a set of features [10-13]. The value of each feature for each experimental condition is coded in the *model matrix* **M** (K conditions by Q features). The *feature weight matrix* **W** (Q features by P channels) then determines how the different model features contribute to the activity profiles of different measurement channels to produce the predicted activity patterns **U**:

$$\mathbf{U}=\mathbf{MW}. \qquad \text{(Eq. 7)}$$

Geometrically, we can think of the features as the basis vectors of the subspace, in which the activity profiles reside (Fig. 2A).

Encoding analysis without regularization

To adjudicate among encoding models of different numbers of features – and hence different numbers of free first-level parameters - most researchers use independent test sets [11-13]. A training data set is used to estimate the feature weights for each channel, and the resulting prediction is then evaluated on a held-out test data set. This can be implemented in a statistically efficient manner by using cross-validation, which is usually performed by holding out a single partition (e.g. fMRI imaging run) as a test set, and using the remaining M-1 partitions as the training set. Each partition is held out as the test set once and prediction performance is averaged across the M folds of cross-validation. Encoding models can also make predictions about conditions that are not in the training set (Discussion). However, we focus our simulations on cases, in which training and test sets include the same experimental conditions.

The weights can be chosen to minimize the sum of squared errors on the training data, i.e. using linear regression:

$$\widehat{\mathbf{W}} = (\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\widehat{\mathbf{U}}^{(\sim \mathrm{m})}, \qquad \text{(Eq. 8)}$$

where we define $\widehat{\mathbf{U}}^{(\sim m)}$ to be the average activity estimates from all partitions except *m*. The prediction for the left-out test data of run *m* is

$$\widetilde{\mathbf{U}}^{(\sim \mathrm{m})} = \mathbf{M}\widehat{\mathbf{W}}. \qquad \text{(Eq. 9)}$$

The accuracy of the prediction can be assessed by relating the residual sums-of-squares (SSR) of the prediction to the total sums-of-squares (SST) of the observed activities, summed over all partitions, conditions, and voxels

$$R_{cv}^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{m,i,j}\left[\widehat{\mathbf{U}}_{i,j}^{(m)} - \widetilde{\mathbf{U}}_{i,j}^{(\sim m)}\right]^2}{\sum_{m,i,j}\widehat{\mathbf{U}}_{i,j}^{(m)2}}. \qquad \text{(Eq. 10)}$$

Alternatively, we can evaluate the prediction by correlating the predicted and observed activity patterns across all conditions and channels. Assuming that the mean of each channel across all conditions is zero (given mean pattern subtraction), the correlation is given by

$$r = \frac{\sum_{m,i,j}\widehat{\mathbf{U}}_{i,j}^{(\boldsymbol{m})} \cdot \widetilde{\mathbf{U}}_{i,j}^{(\sim \boldsymbol{m})}}{\sqrt{\sum_{m,i,j}\widehat{\mathbf{U}}_{i,j}^{(\boldsymbol{m})^2} \cdot \sum_{m,i,j}\widetilde{\mathbf{U}}_{i,j}^{(\sim \boldsymbol{m})^2}}} \qquad \text{(Eq. 11)}$$

The correlation introduces an arbitrary scaling factor between prediction and observations and, in contrast to Eq. 10, allows the model to over- or under-predict the data by a scalar factor without penalty. Encoding analysis can also be applied directly to the time-series data without an intervening model (Eq. 6). In this case, the design matrix for the estimation of the weights (Eq. 8) and for the prediction of the left-out data (Eq. 9) becomes the product of the original first-level design matrix **Z** and the model feature matrix **M**.

To understand how encoding analysis adjudicates between models, consider the graphical representation of the estimation process for a single measurement channel in Figure 3. The training data set (black cross) is the activity profile of the measurement channel, which can be visualized as a point in activity-profile space. Regression analysis can be understood as the orthogonal projection of the measured activity profile onto the linear subspace spanned by the features of the model. The two models depicted in Fig. 3A and Fig. 3B have different features (blue arrows) that define different subspaces (planes with blue outlines). Therefore, the training data is projected onto two

different planes and the prediction for the test data differs between the two models. The model with a subspace that better describes the cloud of activation profiles will make better predictions overall across the measurement channels, show lower cross-validation error, and will hence be more likely selected as the winning model.

Importantly, encoding analysis without regularization compares the subspaces of the competing models, but not their probability distributions. For example, the model depicted in Fig. 3C predicts a different distribution than the one in Fig. 3A. The features of these two models, however, span the same subspace. Therefore, without regularization, the predictions of these two models are identical (black dots) and the models indistinguishable.

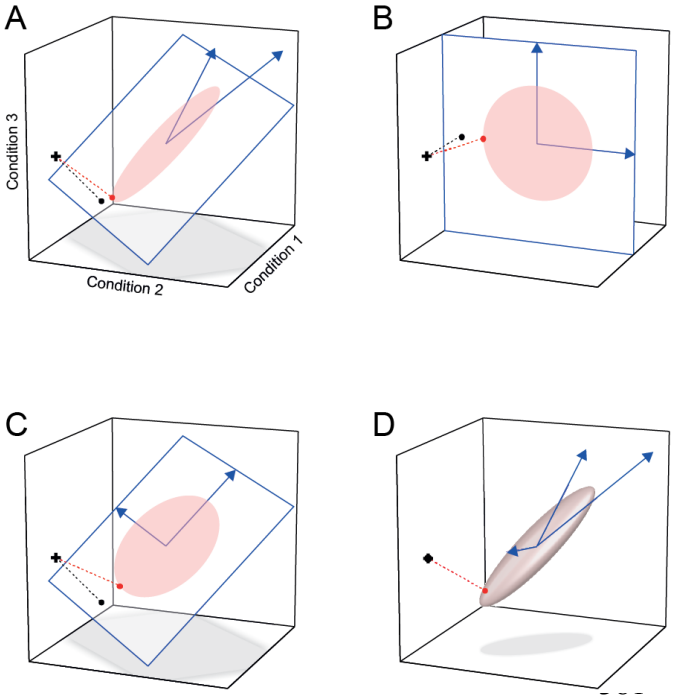

Figure 3. Adjudicating between encoding models. *The axes of the three-dimensional space are formed by the response to three experimental conditions. The activity profile of each unit defines a point in this space. Models are defined by their features (blue arrows) and a prior distribution of the weights for these features. The features and the prior, together, define a distribution of activity profiles (ellipsoids indicate an iso-probability-density contours of the Gaussian distributions). To predict the activity profile of a single measurement channel, the model is fitted to the training data set (cross). Simple regression finds the shortest projection (black dot) onto the subspace defined by the features, whereas regression with regularization (red dot) biases the prediction towards the predicted distribution. Two models (**A, B**) with features that span different model subspaces are distinguishable using regression without regularization. (**C**) This model spans the same subspace as model A. Unregularized*

*regression results in the same projection as for model A, whereas regression with regularization leads to a different projection. (**D**) A saturated model with as many features as conditions. Unregularized regression can perfectly fit any data point (cross and black dot coincide). With regularization, the prediction is biased towards the predicted distribution (iso-probability-density ellipsoid).*

Encoding analysis with regularization

When using regularized regression, encoding analysis evaluates models according to their predicted distribution of activity profiles. From a Bayesian perspective, regularization can be motivated by assuming a prior probability distribution on the weight vectors $\mathbf{w}_{.,i}$ the columns of $\mathbf{W}$. Specifically, L2-norm (Tikhonov) regularization can be derived by assuming a multivariate Gaussian prior with zero mean and variance-covariance matrix $\boldsymbol{\Omega}$. Under this assumption, the predicted second moment of the activity profiles is given by

$$\mathbf{G} = \mathbf{M}\mathbf{W}\boldsymbol{W}^T\boldsymbol{M}^T/P$$

$$= \boldsymbol{M}\boldsymbol{\Omega}\boldsymbol{M}^T\,. \qquad \text{(Eq. 12)}$$

Thus, the model features together with the prior distributional assumption on the feature weights define a probability distribution over activity profiles. For example, a representational model of motor cortical activity could be defined by assuming that the features are individual units with cosine-tuning for different movement directions [14], and that (as a prior) the preferred directions of the units are uniformly distributed.

In practice, we allow a scalar factor, s, between the predicted and measured second moment. This accounts for the fact that different subjects or regions will have different signal levels and that hence the distribution of activity profiles have different widths. Under the assumption that the feature weights come from a multivariate Gaussian distribution with variance $\boldsymbol{\Omega} \cdot s$, the best linear unbiased predictor [BLUP, 41], i.e. the predictor that minimizes the squared error on the held-out data is:

$$\widehat{\mathbf{W}} = (\mathbf{M}^{\mathrm{T}}\mathbf{M} + \boldsymbol{\Omega}^{-1}s^{-1}\sigma_{\epsilon}^{2})^{-1}\mathbf{M}^{\mathrm{T}}\widehat{\mathbf{U}}^{(\sim \mathrm{m})}\,, \qquad \text{(Eq. 13)}$$

where σ_{ϵ}^{2} is the noise variance on the observations [42]. The strength of regularization is determined by the ratio of this noise variance and the variance of the signal $\boldsymbol{\Omega} \cdot s$, consistent with Bayesian inference of the weights on the basis of the prior and the data.

After assuming a prior on the model weights, the two models depicted in Fig. 3A and 3C predict different distributions of the activation profiles. When estimating the weights (Eq. 13), the activity profiles are projected onto the space spanned by $\mathbf{M}$, but this time biased (red dot) towards the denser part of the model-predicted distribution of activity profiles. As a result, the two models make different predictions. An accurate prior will help the model generalize to the held-out data; an inaccurate prior will hurt generalization performance. The model with the distribution that is closest to the true distribution of activity profiles will yield the best cross-validation performance (as measured by R^2 or r). When using regularized regression (Eq. 13), models can also have as many features as conditions (Fig. 3D), or even more features than conditions. When using unregularized regression, such *saturated* models are indistinguishable from each other. They become distinct only after adding weight-distributional priors.

Because regularization is equivalent to imposing a prior on the feature weights, it is not just a technical trick for estimation. Instead the prior is an integral part of the hypothesis being evaluated as it co-determines (together with the features) the probability distribution over activity profiles that the model predicts. Therefore, we will refer to encoding models evaluated using regularized regression analysis in the following as "encoding models with a prior".

One important consequence of Eq. 12 is that the same representational model can be defined using different feature sets. Because a representational model is defined by its second moment, two sets of features $\mathbf{M}_1$ and $\mathbf{M}_2$, combined with corresponding second moment matrices of the weights, $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$, define the same representational model, if

$$\boldsymbol{G} = \boldsymbol{M}_1\boldsymbol{\Omega}_1\boldsymbol{M}_1^{T} = \boldsymbol{M}_2\boldsymbol{\Omega}_2\boldsymbol{M}_2^{T}. \qquad \text{(Eq. 14)}$$

Thus, an important caveat when using encoding models is that one does not compare different feature sets per se – but rather different distributions (when using regularization) or different subspaces of activity profiles (when not using regularization). The winning model in either case can be equivalently re-expressed using a different feature set. Interpretation, therefore, must consider the model-predicted distributions or subspaces of activity profiles, not the particular feature basis set chosen (as the latter is not unique for any given representational model).

Technically, this also means that regression with a Gaussian prior can be implemented using ridge regression [42]. The equivalence is established by scaling and rotating the model matrix **M** in such a way that **Ω** becomes the identity matrix. Any representational model can be brought into this diagonal form by setting the columns of **M** to the eigenvectors of **G**, each one multiplied by the square root of the corresponding eigenvalue:

$$\mathbf{M} = \left[\mathbf{v_1}\sqrt{\lambda_1} \quad \ldots \mathbf{v_2}\sqrt{\lambda_2}\right] \qquad \text{(Eq. 15)}$$

$$\mathbf{G} = \mathbf{M}\mathbf{M}^{\mathbf{T}}.$$

The strength of the regularization is determined by a scalar ridge coefficient defined by $s^{-1}\sigma_\varepsilon^2$. For an encoding model with regularization, the ridge coefficient still needs to be determined for each cross-validation fold. This can be done again by nested cross-validation [12], generalized cross-validation [43], or restricted maximum-likelihood estimation (Eq. 18). To save time, it is also possible to use a constant regularization coefficient. For our simulations, we estimated the optimal $s^{-1}\sigma_\varepsilon^2$ by maximizing Eq. 18 for the training set (across all voxels). Generalized cross-validation [43] yielded very similar results.

Pattern component modeling

An alternative to cross-validation is to evaluate the likelihood of the measured activity profiles under the representational model. This approach is taken in pattern-component modeling [18]. We start with a generative model of the activity profiles (Eq. 6). We consider the activity profiles (columns of **U**) to come from a multivariate Gaussian distribution with zero mean and second-moment matrix **G**. To account for other nuisance effects (mean activity for each partition, low-

frequency drift, etc), the model also has some fixed-effects regressors (**B**). We are not interested in fitting **U** per se, but simply want to evaluate the likelihood of the data under different models, marginalized over all possible values of **U**. The marginal distribution for each channel (columns of matrix **Y**) takes the form of a multivariate normal:

$$\mathbf{y}_{.,j} \sim \mathrm{N}\left(\mathbf{X}\mathbf{b}_{.,j}, \mathbf{V}(\boldsymbol{\theta})\right)$$

$$\mathbf{V}(\boldsymbol{\theta}) = \mathbf{Z}\mathbf{G}s\mathbf{Z}^{\mathbf{T}} + \mathbf{I}\sigma_{\epsilon}^{2} \qquad \text{(Eq. 16)}$$

$$\boldsymbol{\theta} = \{s, \sigma_{\epsilon}^{2}\}.$$

The predicted covariance matrix of the activity measurements for each person is the function of the model (as encoded in the second-moment matrix **G**) and two second-level parameters (**θ**): one that determines the strength of the signal (s) and one that determines the variance of the noise (σ_{ϵ}^{2}). In determining the likelihood, we remove the fixed effects using the residual forming matrix

$$\mathbf{R} = \mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1} . \qquad \text{(Eq. 17)}$$

We need to then account for the removal of these fixed effects by evaluating the restricted likelihood l(**Y**|**G**,**θ**) [44]:

$$\mathrm{l}(\mathbf{Y}|\mathbf{G}, \boldsymbol{\theta}) = -\frac{NP}{2} log(2\pi) - \frac{P}{2} log|\mathbf{V}|$$

$$-\frac{1}{2} trace(\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{R}\mathbf{Y}) - \frac{P}{2} log|\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}|. \qquad \text{(Eq. 18)}$$

To evaluate the fit of a model, the scaling and noise parameters need to be determined. For fMRI data, these two parameters can vary widely between different brain regions and individuals, and are not meaningful in themselves. We therefore replace **θ** with point estimates that maximize Eq. 18 – i.e., the approach uses Empirical Bayes, or Type-II maximum likelihood for model comparison [42]. Because every model has the same two free second-level parameters, even models that are based on different numbers of features can be compared directly. An efficient implementation of this algorithm can be found in the open-source Matlab package for PCM [45].

Representational similarity analysis

Relationship between representational dissimilarities and second-moment matrices

In RSA, representational models are conceptualized in terms of the dissimilarities between the activity patterns elicited across channels by the experimental conditions (Fig. 2C). One important dissimilarity measure is the Euclidean distance, which is closely related to the second-moment matrix **G**. The squared Euclidean distance between the true activity patterns for condition *i* and *k* (normalized by the number of measurement channels) is

$$d_{i,k} = \left(\mathbf{u}_{i,\cdot} - \mathbf{u}_{k,\cdot}\right)\left(\mathbf{u}_{i,\cdot} - \mathbf{u}_{k,\cdot}\right)^{T}/P = \mathbf{G}_{i,i} - 2\mathbf{G}_{i,k} + \mathbf{G}_{k,k} \,. \qquad \text{(Eq. 19)}$$

The Euclidean distance matrix is therefore a function the second moment of the activity profiles. The generalization of the Euclidean distances to non-isotropic noise is the Mahalanobis distance (see below). Correlation distances, another class of popular dissimilarity measures, can also be computed from the second-moment matrix. The cosine angle distance is defined as

$$\mathrm{d}_{i,k} = 1 - \frac{\mathbf{u}_k \mathbf{u}_i^T}{\sqrt{\mathbf{u}_i \mathbf{u}_i^T \mathbf{u}_k \mathbf{u}_k^T}} = 1 - \frac{\mathbf{G}_{k,i}}{\sqrt{\mathbf{G}_{i,i}\mathbf{G}_{k,k}}} \,. \qquad \text{(Eq. 20)}$$

Here we focus on Euclidean and Mahalanobis distances, as they are independent of the resting baseline and generally easier to interpret [29].

In the following, we either represent these distances as a *K*×*K* representational dissimilarity matrix (RDM) **D**, or a *K(K-1)/2* vector **d** that contains all unique pairwise dissimilarities (the lower triangular entries of **D**). The vector of all pairwise dissimilarities can be obtained from **G** by defining a contrast matrix **C**, with each row encoding one of the pairwise contrasts, with a 1 and a -1 for the contrasted conditions and zeros elsewhere:

$$\mathbf{d} = \mathrm{diag}(\mathbf{C}\mathbf{G}\mathbf{C}^{\mathbf{T}}) \quad . \qquad \text{(Eq. 21)}$$

The distances contain the same information as the second moment matrix – however, we are losing the distance of each pattern to the baseline, which was encoded on the diagonal of **G**. Thus, in order to go from a distance matrix to a second-moment matrix, we need to re-set the origin of the coordinate system. An obvious choice is to define the mean activity pattern across all conditions to

be the baseline. This is equivalent making the sum of all rows and columns of **G** zero, which can be achieved by defining the centering matrix $\mathbf{H} = \mathbf{I}_K - \mathbf{1}_K/K$, with $\mathbf{1}_K$ being a square matrix of ones. Under these conditions, **G** can be computed from **D** as

$$\mathbf{G} = -\frac{1}{2}\mathbf{HDH}\,. \qquad \text{(Eq. 22)}$$

This yields the **G** that would be obtained if the patterns in **U** were centered about the origin, as can be achieved by subtracting the mean pattern from each pattern.

Multivariate noise normalization and cross-validation: the crossnobis estimator

A particularly useful dissimilarity measure is the cross-validated, squared Mahalanobis distance estimator (or crossnobis estimator for short). This estimator has superior characteristics in terms of reliability and interpretability as compared to other dissimilarity measures [29].

The crossnobis estimator uses multivariate noise normalization (see section Spatial dependence) to make the errors of different measurement channels approximately independent of each other. Euclidean distances (Eq. 19) computed on these pre-whitened activity estimates are equivalent to the Mahalanobis distance defined by the error-covariance matrix between channels (for details see [26, 29]).

The crossnobis estimator is cross-validated to yield an unbiased estimate of the Mahalanobis distance (assuming that the error covariance is correctly estimated). Conventional distances, which are non-negative by definition, are positively biased when estimated on noisy data: When one replaces the true activity patterns in Eq. 19 with their noisy estimates, the expected value of the Euclidean distance will be always higher than the true distances, because the noise terms are squared and summed. We can obtain an unbiased estimate of the true distance by computing the difference vectors between the two activity patterns from two independent data partitions and taking the inner product of the difference vectors. Thus, if we have *M* independent partitions, the crossnobis estimator can be computed using a leave-one-out cross-validation scheme:

$$d_{i,k} = 1/M \sum_{m=1}^{M}\left(\hat{\boldsymbol{u}}_{i,\cdot}^{(m)} - \hat{\boldsymbol{u}}_{k,\cdot}^{(m)}\right)\left(\hat{\boldsymbol{u}}_{i,\cdot}^{(\sim m)} - \hat{\boldsymbol{u}}_{k,\cdot}^{(\sim m)}\right)^T/P\,, \qquad \text{(Eq. 23)}$$

where $\hat{\mathbf{u}}_{i,\cdot}^{(m)}$ is the prewhitened pattern for condition *i* measured on partition *m*, and $\hat{\mathbf{u}}_{i,\cdot}^{(\sim m)}$ is same activity pattern determined from the data of all other partitions. The expected value of this estimator matches the true Mahalanobis distance [26, 29] (except for a multiplicative bias caused by inaccuracies of the error covariance). In particular, if the patterns of two conditions only differ by noise, then the expected value of this measure will be zero. We will see below that the interpretable zero point can be advantageous for adjudicating among representational models.

Model comparison

In RSA, different representational models are evaluated by comparing the predicted to the observed dissimilarities. The overall magnitude of the Mahalanobis distances can vary considerably between subjects. The inter-subject variation is caused by differences in physiological responsiveness, physiological noise, and head movements – in short, by all the factors contributing to signal strength or the noise distribution, by which the Mahalanobis distance is scaled. Therefore, it is advisable to introduce a subject-specific scaling factor between observed and predicted distances, relying on the ratios between distances to distinguish models.

The unknown scaling of the observed dissimilarities is usually accounted for by calculating the correlation between the predicted and observed representational dissimilarity vectors (not to be confused with the use of correlation distance as an activity-pattern dissimilarity measure, Eq. 20).

The most cautious approach is to assume that we can only predict the rank ordering of distances [22]. It is then only appropriate to use Spearman correlation, or (in the case any of the models predict equal ranks for different pairs of conditions) Kendall's τ_a [24]. Evaluating models based on their ordinal dissimilarity predictions is conservative in terms of assumptions. However, the lesser reliance on assumptions comes at the cost of reduced sensitivity to certain differences between models. For more quantitative models, it may be appropriate to assume that distance predictions can be made on an interval scale. The assumption of a linear relationship between the predicted and measured distances motivates the use of Pearson correlation [25]. It may be justifiable in certain cases and can increase our sensitivity to differences between representational models.

Both rank-based and linear correlation coefficients not only allow an arbitrary scaling factor between observed and predicted distances, but also an arbitrary additive constant due the intercept of regression. However, the crossnobis estimator has an interpretable zero point: If a model predicts a zero distance for two conditions, then a brain region explained by the model should not be sensitive to the difference between the two conditions. This is a very meaningful prediction, which we can exploit to discriminate among models. Pearson and rank-based correlation coefficients discard this information. This suggest the use of a normalized inner product, a quantity analogous to a correlation coefficient, but in which the predictions and the data are not centered about their mean:

$$r_n = \mathbf{d}^{\mathbf{T}}\tilde{\mathbf{d}}/\sqrt{\tilde{\mathbf{d}}^{\mathbf{T}}\tilde{\mathbf{d}}\mathbf{d}^{\mathbf{T}}\mathbf{d}} \qquad \text{(Eq. 24)}$$

This amounts to a linear regression model between the predicted and observed distances, where the regression line is constrained to pass through the origin [46]:

$$\mathbf{d} = \tilde{\mathbf{d}} \cdot s. \qquad \text{(Eq. 25)}$$

Here s is a scaling factor that is estimated from the data by minimizing the sum-of-squared errors between predicted and observed values.

Eq. 24 would provide optimal inference, if all distances estimates were independent and of equal variance. However, for the crossnobis estimator (and for most other dissimilarity measures), the assumptions of independence and equal variance are both violated. Estimated squared distances with larger true values are estimated with higher variability. Furthermore, the estimated distance between conditions A and B is not independent from the estimated distances between A and C [26]. To account for these factors, we need to know the predicted probability distribution of representational dissimilarity matrix estimates given a model. While the exact distribution of the vector of *K(K-1)/2* crossnobis estimates is difficult to obtain, we have shown that their distribution is well approximated by a multivariate normal distribution [26]:

$$\boldsymbol{d} \sim N\left(\tilde{\boldsymbol{d}}s, \boldsymbol{S}\left(\tilde{\boldsymbol{d}}s\right)\right). \qquad \text{(Eq. 26)}$$

The mean of the distribution is the true distance matrix, scaled by a parameter relating to the signal strength in this subject (s). In [26], we showed that that the variance-covariance matrix of $\mathbf{d}$ is given by

$$\mathbf{S}(\mathbf{G}, s, \mathbf{\Sigma}_{\mathrm{K}}, \mathbf{\Sigma}_{\mathrm{P}}) = \left[4\frac{[s\mathbf{C}\mathbf{G}\mathbf{C}^{\mathbf{T}}]\circ[\mathbf{C}\mathbf{\Sigma}_{\mathrm{K}}\mathbf{C}^{\mathbf{T}}]}{M} + 2\frac{[\mathbf{C}\mathbf{\Sigma}_{\mathbf{K}}\mathbf{C}^{\mathbf{T}}]\circ[\mathbf{C}\mathbf{\Sigma}_{\mathbf{K}}\mathbf{C}^{\mathbf{T}}]}{M(M-1)}\right] \cdot \frac{\mathrm{trace}(\mathbf{\Sigma}_{\mathrm{P}}\mathbf{\Sigma}_{\mathrm{P}})}{P^2}. \quad \text{(Eq. 27)}$$

Where $\mathbf{G}$ is the predicted second-moment matrix of the patterns, $\mathbf{C}$ the contrast matrix that transforms the second-moment matrix into distances, and $\circ$ refers to the element-by-element multiplication of two matrices. $\mathbf{\Sigma}_K$ is the condition-by-condition covariance matrix of the estimates of the activation profiles across partitions, which can be estimated from the variability of the activity patterns around their mean ($\overline{\mathbf{U}}$):

$$\widehat{\mathbf{\Sigma}}_K = \frac{1}{M-1}\sum_{\mathrm{m}}\left(\widehat{\mathbf{U}}^{(m)} - \overline{\mathbf{U}}\right)\left(\widehat{\mathbf{U}}^{(m)} - \overline{\mathbf{U}}\right)^{\mathrm{T}}/P \quad \text{(Eq. 28)}$$

$\mathbf{\Sigma}_P$ is the voxel-by-voxel correlation matrix of the activation estimates. If multivariate noise-normalization [29] was completely successful, then this would be the identity matrix. However, given the shrinkage of the noise-covariance matrix used for noise-normalization, some residual correlations will remain; for accurate predictions of the variance, these must be estimated and accounted for [26].

Based on this approximation we can now express the log-likelihood of the measured distances $\mathbf{d}$ under the model predictions $\tilde{\mathbf{d}}$.

$$l\left(\mathbf{d}|\tilde{\mathbf{d}}\cdot s\right) = -\frac{D}{2}log(2\pi) - \frac{1}{2}log\left|\mathbf{S}\left(\tilde{\mathbf{d}}\mathrm{s}\right)\right| - \frac{1}{2}\left(\mathbf{d} - \tilde{\mathbf{d}}\mathrm{s}\right)^{\mathbf{T}}\mathbf{S}\left(\tilde{\mathbf{d}}\mathrm{s}\right)^{-\mathbf{1}}\left(\mathbf{d} - \tilde{\mathbf{d}}\mathrm{s}\right) \quad \text{(Eq. 29)}$$

To evaluate the likelihood, we first need to estimate the scaling coefficient between predicted and observed distances by choosing s to maximize the likelihood. This can be done efficiently using iteratively-reweighted least squares (IRLS): Given a starting estimate of $\mathbf{S}$, we can obtain the generalized least squares estimate of s,

$$s = \left(\tilde{\mathbf{d}}^{\mathbf{T}}\mathbf{S}^{-\mathbf{1}}\tilde{\mathbf{d}}\right)^{-1}\tilde{\mathbf{d}}^{\mathbf{T}}\mathbf{S}^{-\mathbf{1}}\mathbf{d}, \quad \text{(Eq. 30)}$$

re-estimate $\mathbf{S}$ according to Eq. 27, and iterate until convergence.

Simulated example data sets

We use simulated data sets here to evaluate and compare the three analysis techniques in a situation where the ground-truth is known. The three simulated example data sets are inspired by real fMRI studies. The first two examples are motivated by a paper investigating the representational structure of finger movements in primary motor and sensory cortex [25]. The structure of the empirically measured distances between movements of the five fingers was highly reliable across different individuals. The main question was whether this invariant structure is best explained by the correlations of finger movements in every-day life – i.e. the natural statistics of movement [47], or by the patterns of muscle activity required for these movements. Rather than hypothesizing that certain features form the basis set generating the activity profiles distribution, we could directly predict the second-moment matrices, and hence the RDMs, from the correlations between naturally occurring movements, or the correlations of muscle activity patterns. The predicted RDM for individuated movements of the five fingers (Exp. 1) is shown in Fig. 3A, B. The second example comes from experiment 3 in the same paper, this time looking at 31 different finger movements, which span the whole space of possible "piano-chord" combinations (Fig. 3C, D).

The third example uses an experiment investigating the response of the human inferior temporal cortex to 96 images, including animate and inanimate objects [21]. The model predictions are derived from a convolutional deep neural network model – with each of the 7 layers providing a separate representational model [48].

All data sets where simulated with 8 runs, 160 voxels, and independent noise on the observations. The noise variance was set to $\sigma^2 = 1$. We first normalized the model predictions, such that the norm of the predicted squared Euclidean distances was 1. We then derived the second moment matrix ($\mathbf{G}$) for each model using Eq. 22 and created true activity patterns that were normally distributed with second moment $\mathbf{U}\mathbf{U}^{\mathbf{T}}/P = \mathbf{G} \cdot s$. The signal-strength parameter s was varied systematically starting from 0 (pure noise data).

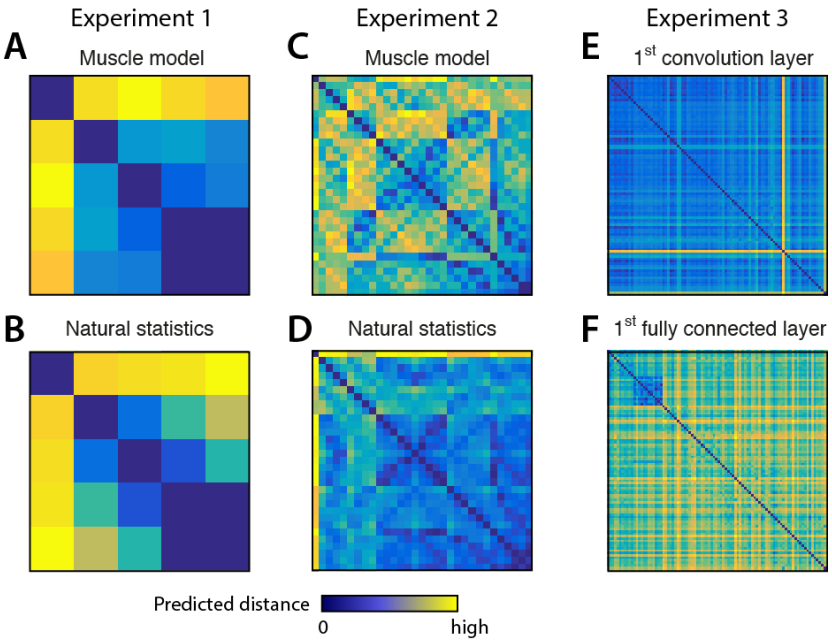

Figure 4. Representational dissimilarity matrices (RDMs) for the models used in simulation. *Each entry of an RDM shows the dissimilarity between the patterns associated with two experimental conditions. RDMs are symmetric about a diagonal of zeros. Note that while zero is meaningfully defined (no difference between conditions), the scaling of the distances is arbitrary. For Experiment 1, the distance between the activity patterns for the five fingers are predicted from the structure of (**A**) muscle activity and (**B**) the natural statists of movement. In Experiment 2 (**C, D**) the same models predict the representational dissimilarities between finger movements for 31 piano-like chords. For Experiment 3 (**E, F**), model predictions come from the activity of the seven layers of a deep convolutional neural network in response to 96 visual stimuli. The 1st convolutional layer and the 1st fully connected layer are shown as examples.*

We generated 3,000 data sets for each experiment, parameter setting, and model. Each data set was generated by one model (ground truth) and was analyzed so as to infer the data-generating model, using each of the inference methods. To evaluate how well the methods adjudicated between the models, we compared the fit of the true model (i.e. the model that generated that particular data set) with each alternative model by counting the number of instances, in which the method decided

in favor of the correct model. Thus, even though there were 7 alternative models in Experiment 3, chance performance for the pairwise comparisons was always 50%. The percentage of correct decisions over all possible model pairs and simulation was used as a measure of model-selection accuracy.

Results

Encoding analysis without regularization

When evaluating encoding models without using regularization, one compares the subspaces spanned by the respective model features. To make different models distinguishable, one typically needs to reduce the dimensionality of the model matrix **M,** for example by using only the eigenvectors with the *n* highest eigenvalues of the predicted second-moment matrix. The decision to use a given number regressors is somewhat arbitrary: For example, Leo et al. [17] used 5 "synergies" (i.e. principal components of the kinematic data of 20 movements), as these explained 90% of the variance of the behavioral data.

Here we explore systematically how the number of principal components influences model selection. For each experiment, we simulated data sets with a fixed signal-to-noise ratio (Exp. 1 and Exp. 3: $s = 0.3$, Exp 2: $s = 0.1$; $\sigma_\varepsilon^2 = 1$), and compared model selection accuracies using a number of principal components ranging between one and the maximum number. We used both cross-validated performance measures, R_{cv}^2 (Eq. 10) and r (the correlation between predicted and observed values; Eq. 11) to perform model selection.

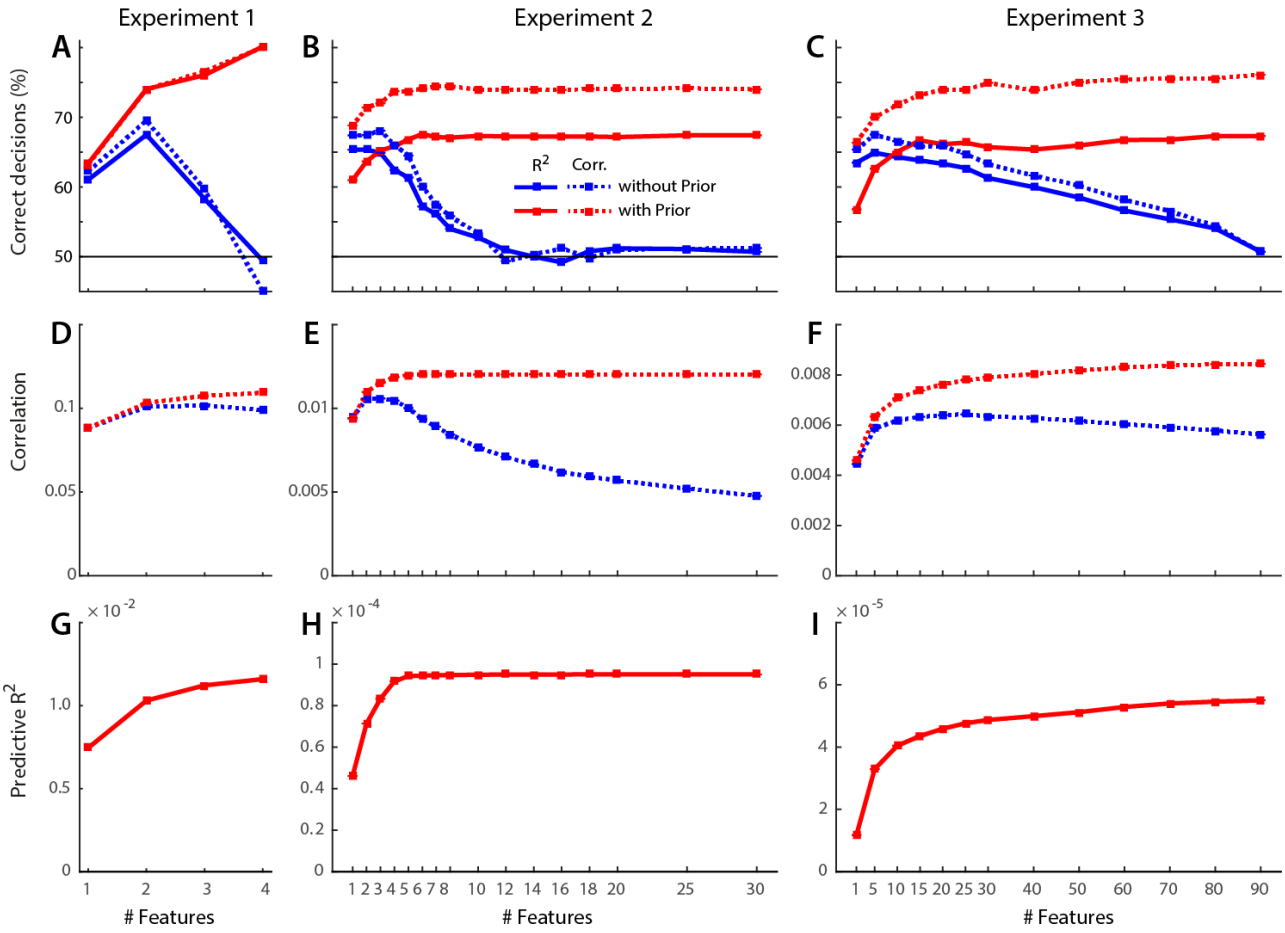

Figure 5. Dependence of encoding model analysis on regularization and the number of included model features. *(**A-C**) Percent correct model selections using either R^2_{cv} (solid line) or correlation (dashed line) for encoding models without a prior (blue lines) and with a prior (red line). (**D-F**) Correlation between predicted and observed patterns. (**G-I**) Predictive R^2_{cv} for the encoding models with prior. All R^2_{cv} values for models without prior are negative, and therefore not visible.*

Fig. 5A-C shows the percentage of correct model selections for Experiments 1-3. Results for encoding analysis without regularization are shown in blue. The dimensionality that differentiated best between competing models was 2, 3, and 5 features, respectively. As more features were included, the number of correct model selections declined. When the number of features was the same as the number of conditions minus 1 (due to the mean subtraction) , i.e. the models became

saturated, model selection accuracy fell to chance. This is expected, as two saturated models span exactly the same subspace and hence make identical predictions (Fig. 3D).

Using correlations as selection criterion led to more accurate decisions than using R^2_{cv}. Correlations (Fig. 5D-F, blue lines) were generally positive and peaked at a number of features that was slightly higher than the optimal dimensionality for model selection. R^2_{cv} values for encoding without a prior were all negative (and are therefore not visible), because the approach does not account for the noise in the data and hence leads to predictions that are too extreme – i.e. the approach over-predicts the scale of the data. Correlations are insensitive to this problem as they allow for arbitrary scaling between predicted and observed values.

Encoding approaches with regularization

From a Bayesian perspective, employing regularization (Eq. 13) is equivalent to adding a prior to the feature weights. Note that this changes the representational hypotheses tested. For example, the models for Experiment 3, based on the neural network representations, now predicted not only that some weighted combination of the neural network features can account for the data, but more specifically that the weights should be small and of similar magnitude, preserving the representational geometry of the original neural network representations. In the model matrix, we scaled each principal component of **G** with the square root of the eigenvalue (Eq. 15), such that we could employ ridge regression to obtain the best linear unbiased predictor for the held-out data patterns.

For encoding models with a prior, model selection performance increased with increasing number of features (red lines, Fig. 5A-C). Thus, dimensionality reduction of the model is not necessary here. Furthermore, model selection was always more powerful with than without a prior when correlation was used for model selection. This reflects the fact that the prior provides additional information about the models to be compared. It enables us to compare well-defined distributions of activity profiles instead of just subspaces.

For Experiments 2 and 3, the R^2_{cv} criterion performed substantially worse than the correlation between predicted and observed activity patterns. The difference between the two criteria arises from the fact that correlations allow for an arbitrary scaling between predicted and observed activity patterns, whereas R^2_{cv} penalizes deviation in scale. The scaling of the prediction in turn strongly depends on the choice of the scalar regularization coefficient. This fact is illustrated in Fig. 6, where we simulated data from Exp. 2 with a fixed noise and signal strength, and varied the regularization coefficient systematically. While R^2_{cv} is highly sensitive to the choice of the regularization coefficient, the correlation criterion is not. Because the regularization coefficient is determined separately for each cross-validation fold and model, deviations from the optimal ridge will decrease model selection accuracy for the cross-validated R^2_{cv} criterion, but not for the correlation criterion.

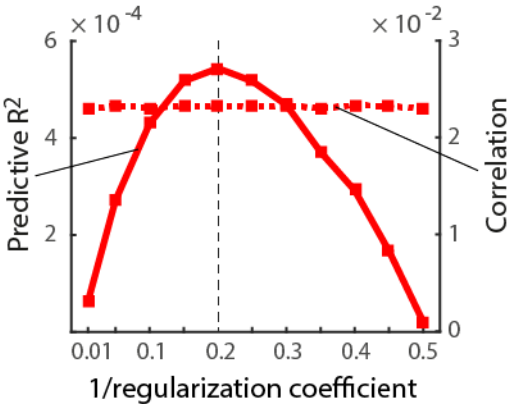

Figure 6. Sensitivity of the R^2_{cv} (solid line) and correlation (dashed line) to the choice of the regularization coefficient. *Simulations come from Experiment 2 with a true signal strength of s=0.2 and a noise variance of 1. For this combination, the optimal regularization coefficient is $s^{-1}\sigma^2_\varepsilon$ (dashed vertical line). The correlation criterion is generally robust against non-optimal choice of regularization coefficient.*

In sum, using regularization improves model selection performance, even if the encoding model has fewer features than conditions or measurements. Rather than just comparing subspaces, the implicit prior on the weights means that a more specific hypothesis is being tested. From this perspective, it is unsurprising that we can adjudicate between these hypotheses with greater accuracy. Furthermore, the use of correlation instead of the predictive R^2_{cv} makes model selection more robust against variations in the regularization coefficient.

Representational similarity analysis

When evaluating models with RSA, there is no need to restrict the model to a specific number of features – the second-moment matrix from all features can determine the predicted distances. As an empirical dissimilarity measure, we used the crossnobis estimator [29] and compared the predicted to the measured RDM. To select the winning model, we used rank-based correlation of dissimilarities [24], Pearson correlation, correlation with a fixed intercept (Eq. 24), and the likelihood of the observed distances under the normal approximation (Eq. 26) using the full variance-covariance matrix of the estimated dissimilarities.

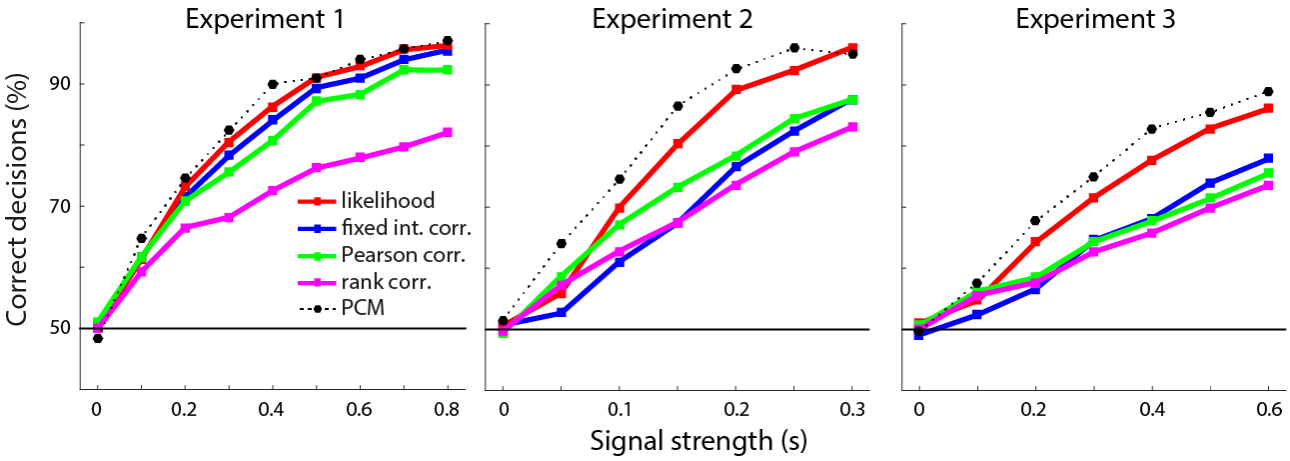

Figure 7. RSA model selection accuracies for different criteria of RDM fit. *Data sets for all three experiments were generated with varying signal strength (horizontal axis). The percentage of correct decisions using different criteria is shown (dotted line). Models were selected based on the Spearman rank correlation (purple), Pearson correlation (green), fixed intercept correlation (blue) or likelihood under the multinormal approximation (red). For comparison, the model selection accuracy for PCM is shown in the dotted line.*

For Experiment 1 (Fig. 7), rank-based correlation performed substantially worse than the other criteria. The lower performance of rank correlation may have been exacerbated here by the fact that the two models predict relatively similar dissimilarity ranks. However, we expect lower performance for rank correlation in general, because this approach does not use all the information in the measured RDMs. It forgoes the assumption of a linear relationship between predicted and

measured dissimilarities and therefore does not exploit the information in the continuous magnitudes of the dissimilarities. Likelihood-based RSA yielded the best decisions; slightly better than Pearson correlation and fixed-intercept correlation.

The advantage of the likelihood-based approach was clearer for Exp. 2 and 3. Here, it led to about 10 percentage points greater accuracy of the decisions than the next-best RSA approach. This advantage is likely due to the fact that Pearson correlations and especially fixed-intercept correlations (Eq. 24) are sensitive to the observed value for the largest predicted dissimilarities, as these data points have a large leverage on the estimated regression line. Indeed, some of the models for Exp. 2 and 3 contain a few especially large dissimilarities, which will influence the model fit strongly. The likelihood-based approach incorporates the knowledge that large dissimilarities are measured with substantially larger variability [26], and hence discounts their influence. Notably, rank-based correlation performed relatively well on these models as compared to Pearson correlation, likely because rank correlation is robust to outliers and less dominated by the large predicted distances.

In sum, these simulations show that it is advantageous to take the covariance structure of the measured dissimilarities into account whenever the additional assumptions this requires are justified.

Pattern component modeling

In the same simulations, we also applied the direct likelihood-ratio test, as implemented by PCM. As all the assumptions of the generative model behind PCM are met in the simulation, we would expect, by the Neyman-Pearson lemma [20], that this method should provide us with highest achievable model selection accuracy. Model selection performance (dotted line in Fig. 7) was indeed systematically higher than for the best RSA-based method. For direct comparison of the so far best methods – PCM, likelihood-based RSA, and encoding analysis with regularization (using correlations as a model selection criterion) – we simulated the three Experiments at a single signal strength (Fig. 8).

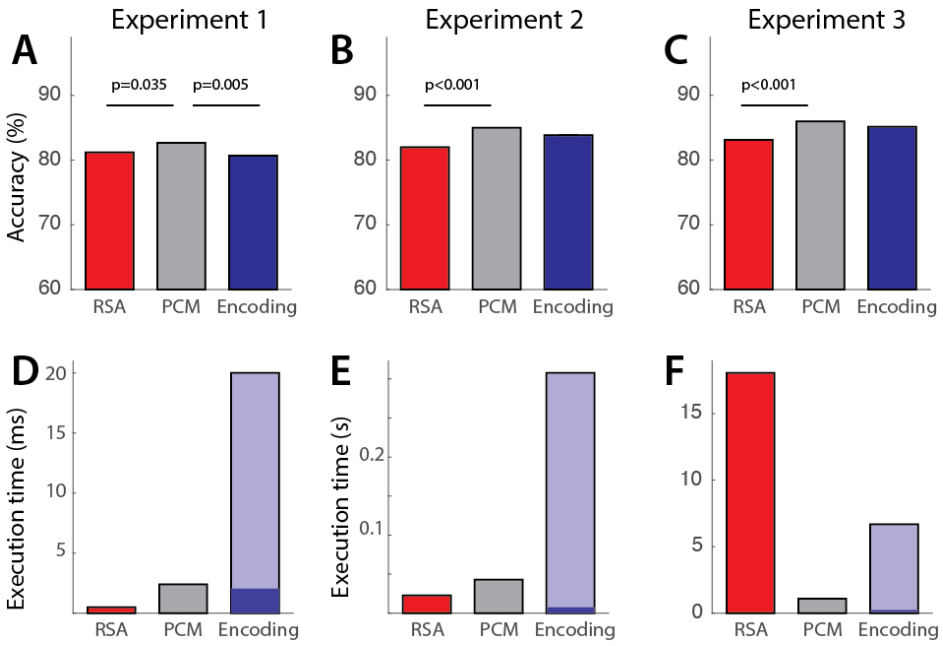

Figure 8. Model selection accuracy and execution time for likelihood-based RSA, PCM, and encoding analysis with regularization. *(**A-C**) Model-selection accuracy was inferentially compared between the three techniques on the basis of N=3,000 simulations, using a likelihood-ratio test of counts of correct model decisions [49]. The signal-strength parameter for the simulation was set to s = 0.3 for Exp. 1, s = 0.15 for Exp. 2, and s = 0.5 for Exp. 3. All resulting significant differences (two-tailed, p<0.01, uncorrected) are indicated by a horizontal line above the bars. (**D-F**) Execution times for the evaluation of a single data set under a single model. For encoding, the time is split into the time required to estimate regression coefficients (dark blue) and the time to determine the regularization constant (light blue).*

In this simulation, PCM resulted in 1.48, 3.01 and 2.86 percentage points (for Exp. 1-3, respectively) better model selection accuracy than likelihood-based RSA, and 1.98, 1.17 and 0.85 percentage points higher model selection accuracies than an encoding analysis using correlations. PCM never performed worse than another method and performed significantly better than the other two approaches in 4 of 6 total comparisons across the three experiments (Fig. 8). There were no

798 significant performance differences between RSA and encoding analysis. Overall, however, all
799 three methods were very close in performance.

800 ***Computational cost***

801 A practical concern is the speed at which the model comparison can be performed. This is usually
802 not important when evaluating the model fit on a small number of participants or ROIs. However, if
803 a larger number of models is evaluated continuously over the cortical surface using a searchlight
804 approach [50, 51], or in data sets with large numbers of participants, computational cost becomes a
805 practical issue. While we cannot treat this issue exhaustively, we provide here a brief overview over
806 the computation time required for the three methods for our specific examples and implementation.
807 In general, the computation time will of course depend on the number of conditions, the number of
808 channels, the exact variant of each technique. Our goal here is simply to give the reader a starting
809 point for making a choice for a particular application, trading off computational and statistical
810 efficiency.

811 Both RSA and PCM operate on the inner product matrix of the activity estimates, thus the
812 computational costs for these approaches is virtually independent of the number of voxels. PCM
813 works on the $MK \times MK$ inner product matrix of the activity estimates, whereas RSA operates on a K
814 $\times K$ matrix of distances between conditions. For a small number of conditions, this explains the
815 favorable computational cost of RSA. However, when using likelihood-based RSA, the covariance
816 matrix of the distances needs to be calculated and inverted. The size of this matrix is $(K(K\text{-}1)/2)^2$
817 and it therefore grows with the 4^{th} power of the number of conditions K. For Exp. 3 (Fig. 8F) with K
818 = 96, this is computationally costly, whereas PCM still only needs to invert matrices of size $(MK)^2$.
819 Using RDM-correlation-based model selection (whether rank, Pearson or fixed-intercept), RSA is
820 much more computationally efficient (not shown).

821 For encoding models, conducting the actual ridge regression for each cross-validation fold
822 (dark blue area) is extremely fast and efficient. The main cost of the technique lies in the
823 determination of the optimal ridge coefficient (light blue area). In our simulations, we use restricted

maximum likelihood estimation (Eq. 18) to do so – therefore this cost is always *M* times higher than for PCM alone. Depending on the implementation, generalized cross-validation [43] may offer a considerable speed-up. If very high speeds are required, one could use a constant ridge coefficient and accept the possible loss in model selection accuracy. In sum, while PCM is computationally feasible across the three experiments, encoding models were less efficient in the present implementation and likelihood-based RSA was less efficient than PCM for the condition-rich scenario of Experiment 3. Alternative variants of encoding models (with fixed ridge coefficient) and RSA (with correlation-based model selection) are less statistically efficient, but beat PCM in terms of computational efficiency.

Discussion

In this paper, we defined representational models as formal hypotheses about the distribution of the activity profiles in the space defined by the experimental conditions. That is, a representational model specifies, which features are represented in a brain region, and how strongly they are represented. The "strength" of representation of a feature has two aspects: the number of responses (e.g. neurons) dedicated to a feature and the scaling of their activity profiles relative to the noise. The second-moment matrix of the activity profiles captures the combined effect of these aspects of feature strength. Two distinct representations with identical second-moment matrices therefore support linear decoding of any given feature at the same signal-to-noise ratio. This holds independent of the question whether the distribution of activity profiles is Gaussian. It motivates using the second moment as a summary statistic for characterizing representations. RSA, PCM and encoding models offer different tests of representational models, but all three depend, explicitly or implicitly, on the second-moment matrix to characterize each representational hypothesis. Thus, these methods are deeply related and should be understood as part of the same multivariate toolbox. The main characteristics of the three methods are summarized in Table 1.

Encoding models without prior define subspaces, not distributions of activity profiles

There is a fundamental difference between encoding models with and without weight priors. Without a prior on the feature weights, encoding models test how well the subspace spanned by the model features captures the observed activity profiles. For models to be discriminable, the dimensionality (i.e. the number of features) of each model must be substantially lower than the number of experimental conditions. As the number of model dimensions increases, the subspaces of competing models increasingly overlap. Once the number of features matches the number of experimental conditions, their subspaces comprise the entire space of activity profiles, each perfectly fits the training data, and their predictions for unseen data become identical.

A subspace specifies what activity profiles are possible and what activity profiles are impossible (though they might still arise as estimates because of the noise). A subspace might be conceptualized as an infinite flat distribution over the subspace dimensions, with 0 probability outside the subspace. However, a uniform distribution on an infinite interval has an infinite second moment and hence does not specify the neural representation uniquely.

L2-norm regularization (i.e. ridge regression) is equivalent to imposing a Gaussian prior on the regression weights. With such a prior, the representational model specifies a probability distribution with a finite second moment. When changing the form of regularization, one also changes the implicit prior, and hence the representational model that is being tested. Thus, regularization is not simply a trick for stabilizing the fit. Instead, the weight prior forms an integral part of the model, which determines the strength with which each feature is encoded according to the model. Choosing a specific form of regularisation therefore constitutes a decision about the neuroscientific hypothesis to be tested rather than a methodological consideration.

Encoding models tests hypotheses about activity profile distributions, not features sets

Encoding models do not support inferences about the particular feature set generating a representation, because infinitely many feature sets can span the same space. Even when using a prior, the feature set that characterizes a given representational model is not unique. Features should not in general be constrained to be orthogonal in the space of experimental conditions, because the structure of the model is not usually meant to depend on the experimental conditions chosen. Whether the features chosen are orthogonal or not, there is an infinite number of basis sets of features that express the same representational model (inducing the same second moment of activity profiles, Eq. 3). For example, two equally long correlated feature vectors can equally well describe a distribution with elliptical isoprobability-density contours (Fig. 3A) as two orthogonal features, with one vector longer than the other. Thus, when one representational model is shown to be superior to others, it does not imply anything special about the feature set chosen to express that model. These complications need to be kept in mind in the interpretation of the results of encoding model analyses. It is very tempting to attribute meaning to the particular feature basis chosen, especially when they are mapped onto the cortical surface [13, 17]. When interpreting these maps, one needs to remember that a feature set only describes a distribution of activity profiles, and that very different maps can emerge when the same distribution is described by a rotated set. In PCM and RSA, the equivalence of different feature sets is made explicit, as they lead to the same second-moment and representational dissimilarity matrices.

Likelihood-based RSA is more sensitive than correlation-based RSA

When using RSA to test representational models, the crossnobis estimator provides a highly reliable measure of dissimilarity with the added advantage of having an interpretable zero-point [29]. Rank-based, Pearson, and fixed-intercept correlation provide fast and straightforward ways of measuring the correspondence between predicted and observed distances, so as to select the representational model most consistent with the data. However, using simple correlations ignores

the dependence of the distance estimates, as well as their unequal variances. In other words, the sampling distribution of the estimated RDM in the space spanned by the dissimilarities (one dimension per pair of conditions) is not isotropic. This problem is addressed in likelihood-based RSA, which uses a multivariate-normal approximation to the sampling distribution of the crossnobis RDM estimate [26]. The approximation provides an analytical expression for the statistical dependency of distance estimates, as well as their signal-dependent variances. In the simulations, likelihood-based RSA was shown to be more powerful than correlation-based RSA. Its model-selection accuracy was only slightly below the theoretical upper bound, as established by PCM. Likelihood-based RSA might therefore become the approach of choice when comparing representational models using crossnobis estimates.

There are situations, however, in which the models are not specific enough to support ratio-scale predictions of representational dissimilarities. Moreover, for measurement modalities like fMRI, it might be undesirable to assume a linear relationship between predicted and measured representational dissimilarities. Rank-correlation-based RSA [22, 24] provides a robust method that is not dependent on the assumption of a linear reflection of the underlying neural dissimilarities in the data RDM. It is also more computationally efficient in the context of condition-rich designs. Likelihood-based RSA becomes computationally expensive as the number of conditions increases. A practical compromise might be to only use the diagonal of the variance-covariance matrix, which would dramatically reduce computational complexity at the expense of neglecting dependencies among dissimilarity estimates.

Which method is best?

For all simulations, model selection using PCM [18] was better than competing methods. This is not surprising, as the data were simulated exactly according to the generative model underlying this approach (Gaussian distribution of noise and signal, independence across voxels). In this case, PCM implements the likelihood-ratio test, which by the Neyman-Pearson lemma [20] is the most powerful test. Beyond confirming what we know from theory, the simulations were important

because they revealed how close the other two techniques come to the theoretical upper bound established by PCM. Results showed that encoding models with a prior and likelihood-based RSA perform near-optimally. In practice, we therefore expect these three techniques to provide similar answers. When its assumptions hold, PCM has clear advantages for model comparison, providing optimal power at reasonable computational cost. However, the other two techniques have other advantages that make them attractive for specific applications.

RSA using RDM correlation for model selection gives up statistical efficiency for computational efficiency, and beats PCM at the latter. When rank correlation is used to compare RDMs, the inference does not rely on a linear relationship between the true dissimilarities and the estimated dissimilarities, an assumption that might be violated in many contexts. RSA also provides readily interpretable intermediate statistics (cross-validated distances), which are closely related to linear decoders for all pairs of stimuli. These statistics can be used to test whether two conditions have different activity patterns [24, 26], or whether the dissimilarity is larger for one pair than for another pair of conditions. Multidimensional scaling of the stimuli on the basis of their representational dissimilarities also provides an intuitive visualization of the representational structure [22], which can be very helpful in the generation of novel representational hypotheses.

In contrast, PCM sometimes demands complicated approaches to answer simple questions: For example, to test the hypothesis that a difference between two conditions is encoded, one would need to fit one model that allows for separate patterns and one model that does not – and then compare the marginal likelihood of these models. Furthermore, PCM requires the noise to be explicitly modeled, whereas RSA removes the bias arising from noise through cross-validated distances.

Encoding analysis explicitly estimates the first-level parameters that describe the response for each individual voxel. This enables the mapping of the estimated features onto the cortical surface to study their spatial distribution [13, 17].

In sum, the three methods are deeply related in that they test hypotheses about the second moment of the activity profiles. However, each method constitutes a unique perspective on the data and supports different kinds of exploratory analyses. We view the methods as complementary tools that are part of a single coherent toolkit for analyzing representations.

Single-voxel vs. multi-voxel inference

An important issue, which we have not touched upon so far, is whether to perform model comparison on single or multiple voxels. While RSA and PCM are usually applied to groups of voxels (such as for ROIs or searchlights), encoding models are often compared on the single-voxel level. This tendency, however, is not strictly inherent in methodological constraints: The searchlight approach for RSA and PCM can be reduced to single voxels, and encoding models can be combined with multi-voxel searchlights. Analyses with courser granularity give up some spatial precision of the map in exchange for greater statistical power. Searchlight mapping boosts power (1) by locally combining the evidence, (2) by enabling the use of a multivariate noise normalization, and (3) by reducing the effective number of multiple comparisons [52]. There is no reason to assume that a single-voxel searchlight is always the optimal choice when balancing spatial precision and power. Based on our previous results [29], we expect that ignoring voxel dependencies will entail a loss of sensitivity when making inferences on representational models for regions of interest comprising multiple responses.

Testing models without overfitting to the noise and to the sample of experimental conditions

Whenever a model is fitted using experimental data, its parameters will necessarily be overfitted to the data to some extent. Assessing the performance of a fitted model therefore requires independent test data. An important question is whether the test data should consist in independent measurements for the same experimental conditions or in measurements for a fresh sample of experimental conditions (e.g. a different sample of visual images). The simple answer is that it depends on the inference we would like to make. If our hypothesis is restricted to the present set of

conditions (e.g. five finger movements), we need only account for overfitting to the noise in the data and require different measurements for the same conditions. If our hypothesis is about a population of conditions (e.g. all natural images), we need to account for overfitting to the condition sample and require measurements for an independent random sample of conditions from the population of conditions covered by our hypothesis.

However, overfitting only needs to be accounted for when the model being tested had parameters fitted in the first place. Encoding models always require independent test sets to account for the over-fitting of the first-level parameters of the representational model (feature weights). RSA and PCM, by contrast, rely explicitly on summary statistics of the responses. Therefore, only second-level parameters related to the strength of the signal and noise need to be fit (see Table 1). The representational models considered here had the same number of such second-level parameters and could therefore be compared directly. Even in this context, however, cross-validation across stimuli can provide valuable information about the generalizability of the model.

What about decoding approaches?

Linear decodeability is central to our definition of a neural representation. Correspondingly, decoding is widely used in multivariate analysis of brain imaging data [7-9]. Can it serve us also as a tool for comparing representational models? While one can use standard decoding approaches to determine whether specific features are represented in an area or not, it does not lend itself to the comparison of full representational models (as defined here). Representational models determine (via the second moment matrix) the decodability of any linear feature, and therefore constitute a hypothesis-driven generalization of the concept of decoding. This is most obvious in RSA, where the RDM assembles all pairwise condition discriminabilities. It is of course possible to use decoding in the context of the methods considered here. For example, some studies have used encoding models to decode stimuli [11, 12, 17]. Decoding accuracy can then serve, instead of correlation or R^2_{cv}, to evaluate the performance of an encoding model on held-out data. While this approach is motivated by the intuitive demonstration of mind reading it does not provide a

particularly natural or powerful approach to adjudicating between representational models. Alternatively, we could use classification accuracy as a measure of dissimilarity between two conditions in the context of RSA [53]. However, classification essentially converts a continuous measure of dissimilarity into a binary label of correct / incorrect. It is therefore expected to be less informative than the underlying continuous measure, and we have shown previously that this entails a loss of sensitivity in practice [29]. In sum, decoding is not particularly useful for the evaluation of representational models [10, 19] and should therefore be limited to situations, in which the quality of the decoding itself is the measure of interest.

Flexible representational models

All models considered here were "fixed", i.e., they did not include free parameters that would change the predicted second-moment matrix. In many applications, however, the relative importance of different features (for example encoding strength for orientation and color) are unknown. In this case, the predicted second moment can be expressed as the weighted sum of different pattern components, i.e. $\mathbf{G} = \sum_i \omega_i \mathbf{C_i}$ [18, 54-56], with the weights being free second-level parameters. In other situations, **G** is a nonlinear function of free model parameters: For example, **G** depends non-linearly on the spatial tuning width in population receptive field modeling [57]. Both RSA and PCM already provide a mechanism to estimate such parameters, as both approaches already need to estimate the signal strength parameters *s* by maximizing the respective likelihood function (Eq. 17, 28) – and the analytical derivatives of the likelihood (Eq. 17, 28) with respect to the parameters are easily obtained. In the context of encoding approaches using ridge regression, free model parameters that change the model structure would result in independent scaling of different features, rotations, or extensions of the model matrix **M**. At the time of writing there are no published examples of such parameter optimization in the context of cross-validated encoding models that we know of.

The inclusion of free parameters into the model also enables the specification of measurement models. Representational models ideally test hypotheses about the distribution of

activation profiles of the core computational elements – i.e. neurons. When using indirect measures of brain activity such as fMRI or MEG, the distribution of activity profiles across measurement channels is also influenced by the measurement process, which samples and mixes neuronal activity signals in the measurement channels [27, 58-61]. As the underlying brain computational models become more specific and detailed, the corresponding measurement models will also have to be improved.

Non-Gaussian representational models

A second restriction of the present exploration is that we focused on approaches that characterize the distribution of activity profiles by its second moment. If the true distribution of the activity profiles is a multivariate Gaussian, then this is a fully sufficient approach. However, a representational hypothesis may not only predict that the response for condition A is uncorrelated to the response for condition B, but, for example, that channels either respond to A or B, but not to both A and B. Such tuning is for example prevalent in primary visual cortex, where neurons (and voxels) respond to a stimulus in a *one* specific part of the visual field, but less often two or more disparate locations [57]. This would correspond to a non-Gaussian prior on the feature weights. In a recent publication, Norman-Haignere and colleagues [62] suggested a likelihood-based method, in which the Gaussian prior on the feature weights **W** is replaced with a Gamma distribution, essentially providing a non-Gaussian extension of PCM. It will be interesting to determine to what degree such non-Gaussian distributions are present in fMRI or single-cell data, and what computational function these may play.

It is important to stress that the approaches considered here are still appropriate when the distribution of activity profiles is truly non-Gaussian. Even in the non-Gaussian case, the second moment determines the linear decodability of all possible features, and hence what is explicitly represented, and thus remains essential for characterizing the representation. Taking into account higher moments of the activity profile distribution would enable us to distinguish between

representations that afford the same linear readout of features, but achieve this by distinct population codes.

Conclusions

If advances in brain-activity measurements are to yield theoretical insights into brain computation, they need to be complemented by analytical methods to test computational models of information processing [63]. The main purpose of this paper was to provide a clear definition of one important class of models – representational models – and to compare three important approaches of testing these. We have shown that PCM, RSA and encoding analysis are all closely related, testing hypotheses about the distribution of activity profiles. Moreover, all three approaches, in their dominant implementations, are sensitive only to distinctions between representations that are reflected in the second moment of the activity profiles. Thus, these three methods are properly understood as components of a single analytical framework. Each of the three methods has particular advantages and disadvantages and preferred areas of application.

1. PCM provides an analytic expression for the marginal likelihood of the data under the model, and therefore constitutes the most powerful test for adjudicating between representational models if the assumptions hold. Its analytical tractability and relative computational efficiency are further attractive features, especially when considering models with increasing numbers of free parameters.
2. RSA provides highly interpretable intermediate statistics and is therefore ideally suited for the visualization and exploratory analysis. Furthermore, simple models are often more easily tested than with PCM. The normal approximation to the distribution of estimated distances enables inference that is nearly as powerful as the likelihood-ratio test provided by PCM. Finally, dissimilarity-rank-based RSA, though less sensitive, provides a means of inference that does not rely on the assumption of a linear relationship between predicted and measured dissimilarities and is computationally efficient even for condition-rich designs.

3. Encoding approaches enable the voxel-wise mapping of model features onto the cortical surface. They therefore are the natural choice when the spatial distribution of features or the voxel-wise comparison of representational models is the main interest.

We hope that the general framework presented here will enable researchers to combine these approaches to make progress revealing the computational mechanisms of biological brains.

Bibliography

1. Stevenson IH, Kording KP. How advances in neural recording affect data analysis. Nat Neurosci. 2011;14(2):139-42. doi: 10.1038/nn.2731.
2. DiCarlo JJ, Zoccolan D, Rust NC. How does the brain solve visual object recognition? Neuron. 2012;73(3):415-34. doi: 10.1016/j.neuron.2012.01.010.
3. deCharms RC, Zador A. Neural representation and the cortical code. Annu Rev Neurosci. 2000;23:613-47. doi: 10.1146/annurev.neuro.23.1.613.
4. Kriegeskorte N. Pattern-information analysis: from stimulus decoding to computational-model testing. Neuroimage. 2011;56(2):411-21. doi: 10.1016/j.neuroimage.2011.01.061.
5. DiCarlo JJ, Cox DD. Untangling invariant object recognition. Trends Cogn Sci. 2007;11(8):333-41. doi: 10.1016/j.tics.2007.06.010.
6. Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I. Invariant visual representation by single neurons in the human brain. Nature. 2005;435(7045):1102-7. doi: 10.1038/nature03687.
7. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 2001;293(5539):2425-30. Epub 2001/09/29. doi: 10.1126/science.1063736.
8. Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci. 2006;10(9):424-30. Epub 2006/08/11. doi: 10.1016/j.tics.2006.07.005 [doi].
9. Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage. 2009;45(1 Suppl):S199-209. Epub 2008/12/17. doi: 10.1016/j.neuroimage.2008.11.007.
10. Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. Neuroimage. 2011;56(2):400-10. Epub 2010/08/10. doi: 10.1016/j.neuroimage.2010.07.073.
11. Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, et al. Predicting human brain activity associated with the meanings of nouns. Science. 2008;320(5880):1191-5. doi: 10.1126/science.1152876.
12. Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008;452(7185):352-5. Epub 2008/03/07. doi: 10.1038/nature06713.
13. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature. 2016;532(7600):453-8. doi: 10.1038/nature17637.
14. Georgopoulos AP, Schwartz AB, Kettner RE. Neuronal population coding of movement direction. Science. 1986;233(4771):1416-9.
15. Sergio LE, Hamel-Paquet C, Kalaska JF. Motor cortex neural correlates of output kinematics and kinetics during isometric-force and arm-reaching tasks. J Neurophysiol. 2005;94(4):2353-78.

16. Sergio LE, Kalaska JF. Systematic changes in directional tuning of motor cortex cell activity with hand location in the workspace during generation of static isometric forces in constant spatial directions. J Neurophysiol. 1997;78(2):1170-4. Epub 1997/08/01.
17. Leo A, Handjaras G, Bianchi M, Marino H, Gabiccini M, Guidi A, et al. A synergy-based hand control is encoded in human motor cortical areas. Elife. 2016;5. doi: 10.7554/eLife.13420.
18. Diedrichsen J, Ridgway GR, Friston KJ, Wiestler T. Comparing the similarity and spatial structure of neural representations: a pattern-component model. Neuroimage. 2011;55(4):1665-78. Epub 2011/01/25. doi: 10.1016/j.neuroimage.2011.01.044.
19. Friston K, Chu C, Mourao-Miranda J, Hulme O, Rees G, Penny W, et al. Bayesian decoding of brain images. Neuroimage. 2008;39(1):181-205. Epub 2007/10/09. doi: 10.1016/j.neuroimage.2007.08.013.
20. Neyman J, Pearson ES. On the problem of the most efficient test of statistical hypotheses. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 1933;231:289–337. doi: doi:10.1098/rsta.1933.0009.
21. Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, et al. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron. 2008;60(6):1126-41. Epub 2008/12/27. doi: 10.1016/j.neuron.2008.10.043 [doi].
22. Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci. 2008;2:4. Epub 2008/12/24. doi: 10.3389/neuro.06.004.2008.
23. Kriegeskorte N, Kievit RA. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci. 2013;17(8):401-12. Epub 2013/07/24. doi: 10.1016/j.tics.2013.06.007.
24. Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. A toolbox for representational similarity analysis. PLoS Comput Biol. 2014;10(4):e1003553. Epub 2014/04/20. doi: 10.1371/journal.pcbi.1003553.
25. Ejaz N, Hamada M, Diedrichsen J. Hand use predicts the structure of representations in sensorimotor cortex. Nat Neurosci. 2015;18(7):1034-40. Epub 2015/06/02. doi: 10.1038/nn.4038.
26. Diedrichsen J, Zareamoghaddam H, Provost S. The distribution of crossvalidated mahalanobis distances. ArXiv. 2016.
27. Kriegeskorte N, J. D. Inferring brain-computational mechanisms with models of activity measurements. Proceedings of the Royal Society. 2016.
28. Cai MB, Schuck NW, Pillow J, Niv Y. A Bayesian method for reducing bias in neural representational similarity analysis. BioRxiv. 2016. doi: 10.1101/073932.
29. Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J. Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage. 2016;137:188-200. doi: 10.1016/j.neuroimage.2015.12.012.
30. Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance. 2003;10(5)(603–621).
31. Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage. 2010;53(1):103-18. Epub 2010/06/29. doi: 10.1016/j.neuroimage.2010.05.051.
32. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936;7(2):179-88.
33. Friston KJ, Holmes AP, Poline JB, Grasby PJ, Williams SC, Frackowiak RS, et al. Analysis of fMRI time-series revisited. Neuroimage. 1995;2(1):45-53.
34. Worsley KJ, Friston KJ. Analysis of fMRI time-series revisited--again. Neuroimage. 1995;2(3):173-81.
35. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;38(4):963-74.

36. Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. J Neurophysiol. 2009;102(1):614-35. doi: 10.1152/jn.90941.2008.
37. Carlson T, Tovar DA, Alink A, Kriegeskorte N. Representational dynamics of object vision: the first 1000 ms. J Vis. 2013;13(10). doi: 10.1167/13.10.1.
38. Wardle SG, Kriegeskorte N, Grootswagers T, Khaligh-Razavi SM, Carlson TA. Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. Neuroimage. 2016;132:59-70. doi: 10.1016/j.neuroimage.2016.02.019.
39. Cichy RM, Pantazis D, Oliva A. Resolving human object recognition in space and time. Nat Neurosci. 2014;17(3):455-62. doi: 10.1038/nn.3635.
40. Kobak D, Brendel W, Constantinidis C, Feierstein CE, Kepecs A, Mainen ZF, et al. Demixed principal component analysis of neural population data. Elife. 2016;5. doi: 10.7554/eLife.10989.
41. Robinson GK. That BLUP is a Good Thing: The Estimation of Random Effects. Statistical Science. 1991;6(1):15–32. doi: 10.1214/ss/1177011926.
42. Murphy KP. Machine Learning: A probabilistic perspective. Cambridge, MA: MIT press; 2012.
43. Golub GH, Heath M, Wahba G. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. Technometrics. 1979;21(2):215-23. doi: 10.1080/00401706.1979.10489751.
44. Harville DA. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association. 1977;72(358):320-38. doi: 10.1080/01621459.1977.10480998.
45. Diedrichsen J, Yokoi A, Arbuckle S. Pattern component modeling toolbox. 2016.
46. Eisenhauer JG. Regression through the origin. Teaching Statistics. 2003;25(3):76-80.
47. Ingram JN, Kording KP, Howard IS, Wolpert DM. The statistics of natural hand movements. Exp Brain Res. 2008;188(2):223-36. Epub 2008/03/29. doi: 10.1007/s00221-008-1355-3.
48. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. NIPS; Lake Tahoe, Nevada2012.
49. Sokal RR, Rohlf FJ. Biometry: the principles and practice of statistics in biological research. 2nd ed. San Fransisco: W. H. Freeman; 1981.
50. Oosterhof NN, Wiestler T, Downing PE, Diedrichsen J. A comparison of volume-based and surface-based multi-voxel pattern analysis. Neuroimage. 2011;56(2):593-600. Epub 2010/07/14. doi: 10.1016/j.neuroimage.2010.04.270.
51. Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. Proc Natl Acad Sci U S A. 2006;103(10):3863-8.
52. Kriegeskorte N, Bandettini P. Analyzing for information, not activation, to exploit high-resolution fMRI. Neuroimage. 2007;38(4):649-62. doi: 10.1016/j.neuroimage.2007.02.022.
53. O'Toole AJ, Jiang F, Abdi H, Haxby JV. Partially distributed representations of objects and faces in ventral temporal cortex. J Cogn Neurosci. 2005;17(4):580-90. doi: 10.1162/0898929053467550.
54. Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. 2014;10(11):e1003915. doi: 10.1371/journal.pcbi.1003915.
55. Jozwik KM, Kriegeskorte N, Mur M. Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. Neuropsychologia. 2016;83:201-26. doi: 10.1016/j.neuropsychologia.2015.10.023.
56. Khaligh-Razavi SM, Henriksson L, Kay K, Kriegeskorte N. Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models BioRxiv. 2016.

57. Dumoulin SO, Wandell BA. Population receptive field estimates in human visual cortex. Neuroimage. 2008;39(2):647-60. doi: 10.1016/j.neuroimage.2007.09.034.
58. Kriegeskorte N, Cusack R, Bandettini P. How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter? Neuroimage. 2010;49(3):1965-76. doi: 10.1016/j.neuroimage.2009.09.059.
59. Kamitani Y, Tong F. Decoding the visual and subjective contents of the human brain. Nat Neurosci. 2005;8(5):679-85. Epub 2005/04/27. doi: 10.1038/nn1444.
60. Chaimow D, Yacoub E, Ugurbil K, Shmuel A. Modeling and analysis of mechanisms underlying fMRI-based decoding of information conveyed in cortical columns. Neuroimage. 2011;56(2):627-42. doi: 10.1016/j.neuroimage.2010.09.037.
61. Ramirez FM, Cichy RM, Allefeld C, Haynes JD. The neural code for face orientation in the human fusiform face area. J Neurosci. 2014;34(36):12155-67. doi: 10.1523/JNEUROSCI.3156-13.2014.
62. Norman-Haignere S, Kanwisher NG, McDermott JH. Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. Neuron. 2015;88(6):1281-96. Epub 2015/12/22. doi: 10.1016/j.neuron.2015.11.035.
63. Kording K, Jonas E. Could a neuroscientist understand a microprocessor? . BioRxiv. 2016. doi: 10.1101/055624.