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 & Departments for Statistics and Computer 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 explain how activity patterns in populations of neurons (or, more generally, in multivariate brain activity measurements) relate to sensory stimuli, motor actions, or cognitive processes. In an experimental context, representational models can be defined as probabilistic hypotheses about what activation profiles across experimental conditions are likely to be observed. We describe three methods to test such models – encoding approaches, pattern component modeling (PCM), and representational similarity analysis (RSA). We show that these methods are closely related in that they evaluate the statistical second moment of the activity profile distribution. Using simulated data from three different fMRI experiments, we compare the power of the approaches to adjudicate between competing representational models. PCM implements a likelihood-ratio test and therefore constitutes the most powerful test if its assumptions hold. However, the other two approaches – when conducted appropriately – can perform similarly. In encoding approaches, the linear model needs to be appropriately regularized, which imposes a prior on the activity profiles. Without such a prior, encoding approaches do not test well-defined representational models. In RSA, the unequal variances and dependencies of the distance measures need to be taken into account to enable near-optimal inference. The three techniques render different aspects of the information explicit (e.g. single response tuning in encoding approaches and population representational dissimilarity in RSA) and have specific advantages in terms of computational demands, ease of use, and extensibility. We argue that they constitute complementary parts of the same computational toolkit aimed at understanding neural representations on the basis of multivariate brain-activity data.

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 recorded simultaneously [1]. Electrode arrays and calcium imaging enable the recording of hundreds or many thousands 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 at sub-millimeter resolution.

In order to translate advances in brain-activity measurement into advances in computational theory, researchers increasingly seek to test representational models that capture both what information is represented in a population of neurons and in what format 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.

We use the term *representation* in a general sense, implying that the represented variable can be linearly decoded by a downstream area – i.e. read out by weighting and summing the elements of the ongoing neural activity. That is, we do not subscribe to the narrow meaning of representation as referring to a localized code, in which the neural activity can be fully explained by a physically or semantically meaningful variable.

The vector of responses of one *measurement channel* (neurons, electrodes, or fMRI voxels) across the *experimental conditions* (stimuli, movements, or tasks) is referred to as an *activity profile*. A representational model predicts a probability distribution over the space of activity profiles (Fig. 1). The measurement channels are considered to be samples from this distribution.

Activity profiles can (but do not have to be) explained by one or more *features* that characterize the experimental conditions. Examples of features are the color of a visual stimulus, the meaning of a sound, the direction of a movement, or the response of one element of a computational model (e.g. a unit in a neural network that processes the stimuli). In the special case that only one feature is varied, an activity profile is also called a tuning curve.

A representational model specifies which features are reflected in the measurements, and how strongly they are reflected. A representational model does not, however, predict the activity profile for each individual channels. Instead it predicts how likely it is to observe a specific activity profile in the population of measurement channels. Therefore, a representational model is a hypothesis about the probability distribution over the space of possible activity profiles (Fig. 1A,C). This definition of representational model is helpful because the computational function of a region does not depend on specific neurons having specific response properties, but on the fact that specific features can be read out linearly from the population, even if the information is spread out over multiple neurons. When analyzing measurements of local summaries of neuronal activity, such as fMRI voxels or LFP recordings, the representational model should also include a measurement model [2] that accounts for the effect of the local averaging on the distribution of activity profiles.

How should we evaluate how well the distribution of measured activity profiles matches the prediction? From the perspective of a neuron that reads out the activity of an area, i.e. a linear decoder, any difference between activity patterns across conditions is equally meaningful. Some features (for example, stimulus intensity) may be encoded in the mean response, with overall higher

activity for condition 1 than 2. Other properties (for example, stimulus identity) may be encoded in relative activity differences, with some measurement channels responding more to condition 1 and others more to 2. If we want to summarize what information can be linearly decoded from population activity, we require a measure that captures both of these scenarios. It turns out that the *second moment matrix* of the activity profiles provides a sufficient statistic for characterizing linear decodability of any arbitrary feature.

More formally, we define **U** to be the matrix of true 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. The second moment[1] matrix of the activity profiles is defined as

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

The variance-covariance matrix of the activity profile is a special case of Eq. 1 in which the mean activity profile is first subtracted from each activity profile. For this reason, the variance-covariance is often call the second moment around the mean. In contrast, the second moment around zero captures any activity differences between experimental conditions – i.e. it depends on both the mean and variance of the distribution of activity profiles.

Assuming Gaussian noise that is independent and identically distributed both across measurement channels (isotropic noise) and across conditions (homoscedastic noise), the second moment matrix determines how well any feature can be linearly decoded from the activity patterns (see Methods). This provides a motivation, from the perspective of brain computation, for using the second moment matrix as a summary statistic. While higher statistical moments of the distribution of activity profiles may reveal some computationally important features of the population activity (see Discussion), we believe that focusing on linear encoding (and hence the second moment) constitutes the natural starting point for testing representational models. We will see below that the techniques discussed in this paper (i.e. encoding approaches, PCM and RSA) all rely exclusively on information contained in the second moments. This core commonality enables us to consider these methods in the same formal framework.

Three different approaches are currently being used to compare representational models on multivariate data (for comparison, see Table 1). In *encoding approaches* [3, 4], the representational models is defined in terms of the underlying *features*. Features for models of low-level visual representations may include Gabor filters [5], whereas features for cognitive representations may include abstract semantic dimensions [6]. The value of each feature for each experimental condition is coded in the model feature 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:

$$\mathbf{U}=\mathbf{M}\mathbf{W}. \qquad \text{(Eq. 2)}$$

[1] The nth moment of a random variable x is $E(x^n)$. Here we use a multivariate extension of the concept, with the second moment of the random column vector **x** defined as the matrix $E(\mathbf{x}\mathbf{x}^{T})$, the expected outer product of the activity profiles, where the expectation is across the measurement channels. Equivalently, each cell of this matrix contains the normalized inner product of two activity patterns.

Geometrically, we can think of the features as the basis vectors of the subspace, in which the activity profiles reside (Fig. 2). In this notation, the second moment of the activity profiles is given by

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

$$= \mathbf{M}\boldsymbol{\Omega}\mathbf{M}^{\mathbf{T}}\,, \qquad \text{(Eq. 3)}$$

where $\boldsymbol{\Omega}$ is the second moment of the distribution on the weights **W**. Thus, the model features together with a distributional assumption (i.e. a prior) on the hidden weights define a probability distribution over activity profiles, and thus specify a representational model.

For example, the second-moment matrix for motor cortical activity could be predicted by assuming that the features are individual units with cosine-tuning for different movement directions [7], and that (as a prior) the preferred directions of the units are uniformly distributed. While it is often helpful to formulate a representational model in terms of its features, it is not always necessary: In a recent study we predicted the second-moment matrix for an experiment studying hand movements directly (i.e. without specifying features), using the natural correlations between finger movements in everyday life [8]. Indeed, one caveat when specifying representational models in terms of features, is the fact that there are there are many different feature sets that define the same representational model. Because a representational model is defined by its second moment, two sets of features $\mathbf{M}_1$ and $\mathbf{M}_2$, combined with their corresponding priors $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$, define the same representational model, if

$$\mathbf{G} = \mathbf{M}_1\boldsymbol{\Omega}_1\mathbf{M}_1^{\mathbf{T}} = \mathbf{M}_2\boldsymbol{\Omega}_2\mathbf{M}_2^{\mathbf{T}}. \qquad \text{(Eq. 4)}$$

Because the matrix **W** contains a large number of free parameters, encoding models are usually evaluated using crossvalidation: The weights are estimated on a training set, and the fitted model is evaluated in terms of its performance at predicting left-out data [3]. The left-out data may consist in novel experimental conditions, so as to test the fitted model's generalization performance [4, 5]. In encoding approaches, the prior on **W** typically enters as a regularization penalty in the estimation of model weights (e.g. a penalty on the L2 norms of the columns of **W**). However, encoding models are also often fitted without regularization [4, 9]. We will see that this approach does not test to a well-defined distribution of activity profiles, but rather evaluates the subspaces defined by the features. Note that although the encoding approach does not involve explicit computation of **G**, it centrally depends on **G** via Eq. 3, 4.

In *pattern component modeling* (PCM [10]), the activity patterns are assumed to have a multivariate normal distribution and the representational model is evaluated by calculating the marginal likelihood of the observed activity pattern under the model. The marginal likelihood explicitly integrates over all possible values of **W**, i.e. it computes the expected likelihood under the assumed prior distribution of the model weights. This approach prevents overfitting and therefore makes crossvalidation unnecessary. If the assumptions of this method are met, PCM provides the likelihood-ratio test between models [11], which by the Neyman-Pearson lemma [12] is the most powerful test of its size. That is, in theory this approach should yield more accurate inferences that any of its competitors.

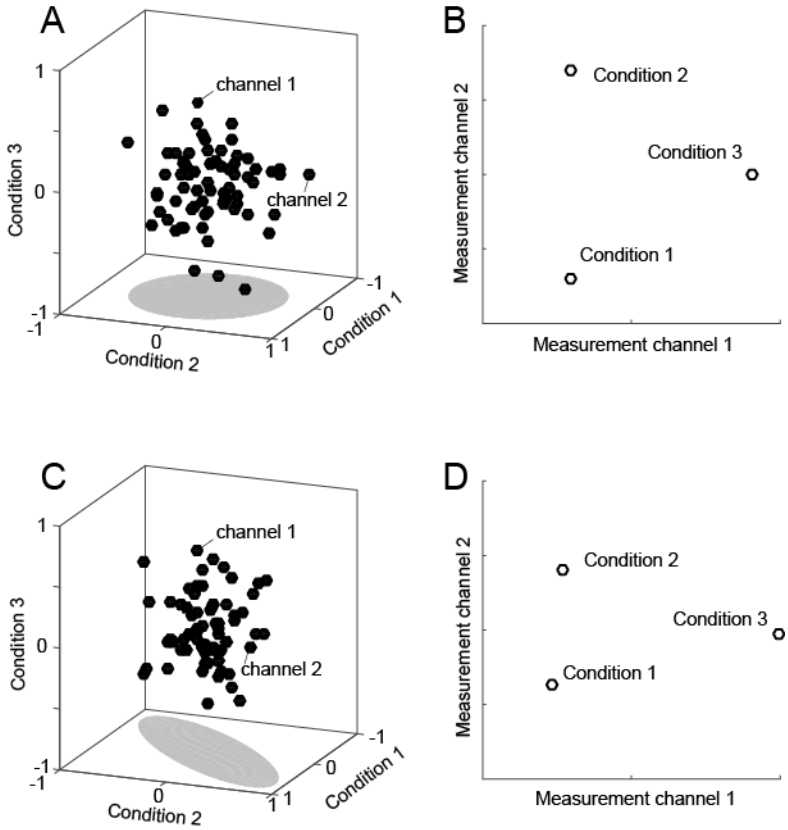

Figure 1. Two complementary perspectives on the population activity across experimental conditions. *(**A**, **C**) The condition-by-measurement-channel matrix can be visualised by plotting the measurement channels as single points in a space defined by the experimental conditions. The coordinate of each measurement channel reflects its activity profile across conditions. (**B**, **D**) The experimental conditions are plotted as points in the space defined by the measurement channels. The coordinates of each condition reflects the activity pattern across measurement channels. (**A**) The activity profiles have spherical distribution. (**B**) The corresponding activity patterns of all three conditions are equidistant to each other. (**C**) The activity for condition 1 is positive correlated with the activity for condition 2. (**D**) The activity patterns for condition 1 and 2 are closer to each other than to condition 3.*

Finally, *representational similarity analysis* (RSA [13-15]) 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. 1A,C), one can consider the activity patterns of the conditions as points in the space spanned by the measurement channels (Fig. 1B,D). The distances between these points are determined by the corresponding distribution of activity profiles. Indeed, in the case of Euclidean distances, the arrangement of the conditions is fully determined by the eigenvectors of the second-moment matrix **G**. 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 the representational model. In this paper, we evaluate different methods to perform this comparison, including rank-based correlations [16], Pearson correlations [8], and a novel likelihood-based approach that uses a multivariate normal approximation to the joint distribution of the cross-validated Mahalanobis distances [17]. As shown in the simulations, this latter technique provides a powerful means to adjudicate between representational models.

In the methods, we will describe the different approaches in detail and clarify their relationship. It will become apparent that the encoding approach with a Gaussian prior (implemented as L2-norm or Tikhonov regularization), PCM, and RSA all test the same underlying probabilistic representational model. We will then use three simulated scenarios inspired by our fMRI work to assess how efficiently these approaches select between two or more competing models. Because fMRI does not resolve the dynamics of representations with high temporal resolution, we restrict ourselves to models that treat the activity patterns as static snapshots. However, our considerations here also provide a foundation for testing models that incorporate dynamics.

Methods

Basic definitions

The methods in this paper were first developed in the context of multi-voxel analysis of fMRI data. This is reflected in the simulated scenarios chosen. However, the ideas in this paper equally apply to other modalities of brain-activity measurement. We assume that the main data consist of M independent partitions, each containing at least one activity measurement for each condition and measurement channel. Typically, each partition corresponds to a separate phase of data acquisition, e.g. a scanner run in fMRI. Estimates from different partitions are assumed to be independent. In general, the activity estimates $\hat{\mathbf{U}}^{(m)}$ ($K \times P$) of partition m are a function of the true patterns $\mathbf{U}$ plus estimation noise $\mathbf{E}^{(m)}$

$$\hat{\mathbf{U}}^{(m)} = \mathbf{U} + \mathbf{E}^{(m)}. \qquad \text{(Eq. 5)}$$

In fMRI, the patterns estimates are the regression coefficients, or "beta"-weights, from a first-level time series analysis [18, 19], which accounts for the hemodynamic lag and the temporal autocorrelation of the noise. The activity estimates express the difference in activity during a condition relative to rest. Patterns measured in the same fMRI imaging run are usually correlated. One cause for this correlation is that all activity estimates within a partition are measured relative to the same resting baseline. To correct for this, the mean activity pattern (across conditions) can be subtracted 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 on the origin.

Encoding approaches and PCM can be applied either to the concatenated activity estimates from different partitions or directly to time series data. As a universal notation that encompasses both situations, we can write:

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

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

Spatial dependence

Noise in fMRI is spatially correlated. To test representational models, we therefore use multivariate noise normalization (i.e. spatial prewhitening), which has been shown to increase the reliability of inference [20]. Within each ROI or searchlight [20], we estimate the $P \times P$ variance-covariance matrix $\mathbf{\Sigma}_P$ between the residual time series from the first-level model, regularize it by shrinking it slightly towards a diagonal matrix [21]. We then post-multiply our activity estimates by $\mathbf{\Sigma}_P^{-1/2}$, rendering the model errors in the voxels approximately independent. For the simulations, we assume spatially independent errors from the beginning. If multivariate noise normalization is not performed or is incomplete, inference will be suboptimal in all three methods – for details see [17].

Linear decoding and second moment matrices

Based on our definition, a feature is represented in an area if it can be linearly read out from the activity. For example, a feature of interest may be the contrast between two stimuli. To obtain a read-out, we would weight each channel's observed activity using the $P \times 1$ read-out vector $\mathbf{v}$. This results in the read out variable

$$\hat{y}_i = \hat{\mathbf{u}}_{i,.}\mathbf{v}. \qquad \text{(Eq. 7)}$$

We would now like the variable y to have very different values for two stimuli, while showing low variability for different trials of the same stimulus. More generally, we want a specific contrast on the experimental conditions, defined by a $K \times 1$ vector $\mathbf{f}$, to have maximal variability when the true patterns are used as input, and minimal variability if only noise is used as input. Thus, we are searching for a $\mathbf{v}$ that maximizes the signal-to-noise ratio S of the decoded variable:

$$S = \frac{\mathbf{v}^\mathbf{T}\mathbf{U}^\mathbf{T}\mathbf{f}\mathbf{f}^\mathbf{T}\mathbf{U}\mathbf{v}}{\mathbf{v}^\mathbf{T}\mathbf{E}^\mathbf{T}\mathbf{f}\mathbf{f}^\mathbf{T}\mathbf{E}\mathbf{v}} \ . \qquad \text{(Eq. 8)}$$

Under isotropic noise, $\mathbf{E}^\mathbf{T}\mathbf{f}\mathbf{f}^\mathbf{T}\mathbf{E} = \mathbf{I}b$, with b being a constant. Therefore, the denominator is the scaled norm of the read-out vector. We can therefore can get rid of the denominator by constraining the length of $\mathbf{v}$ to 1. The best vector is then given by the first eigenvector of the matrix $\mathbf{U}^\mathbf{T}\mathbf{f}\mathbf{f}^\mathbf{T}\mathbf{U}$, and the quality of the best readout is determined by the largest eigenvalue.

The vector of non-zero eigenvalues (*eig*) is invariant to any allowed rotation of the elements of the matrix product:

$$eig(\mathbf{U}^\mathbf{T}\mathbf{f}\mathbf{f}^\mathbf{T}\mathbf{U}) = eig(\mathbf{f}^\mathbf{T}\mathbf{U}\mathbf{U}^\mathbf{T}\mathbf{f}) = eig(\mathbf{f}^\mathbf{T}\mathbf{G}\mathbf{f})\mathrm{P} \ . \qquad \text{(Eq. 9)}$$

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

Encoding approaches

Encoding approach without a prior

In encoding approaches, a training data set is used to estimate the channel weights, and the resulting prediction is then evaluated on a left-out test data set. This cross-validation is usually performed by leaving a single partition (e.g. fMRI imaging run) out 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. While encoding models can make predictions about conditions that are not in the training set (see Discussion), we focus our simulations on cases, in which training and test sets include the same experimental conditions.

A simple way to estimate weights $\mathbf{W}$ is through linear regression:

$$\widehat{\mathbf{W}} = (\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\widehat{\mathbf{U}}^{(\sim \mathrm{m})}, \qquad \text{(Eq. 10)}$$

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

$$\widetilde{\mathbf{U}}^{(\sim \mathrm{m})} = \mathbf{M}\widehat{\mathbf{W}}. \qquad \text{(Eq. 11)}$$

The quality of the prediction can be assessed using the proportion of activity variance that could be predicted across all conditions and channels

$$R^{\mathbf{2}} = 1 - \frac{1}{M}\sum \frac{\boldsymbol{trace}\left(\left(\widehat{\mathbf{U}}^{(m)} - \widetilde{\mathbf{U}}^{(\sim m)}\right)\left(\widehat{\mathbf{U}}^{(m)} - \widetilde{\mathbf{U}}^{(\sim m)}\right)^{\mathbf{T}}\right)}{\boldsymbol{trace}\left(\widehat{\mathbf{U}}^{(m)}\widehat{\mathbf{U}}^{(m)T}\right)}. \qquad \text{(Eq. 12)}$$

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

$$r = \frac{1}{M}\sum \frac{\boldsymbol{trace}\left(\widehat{\mathbf{U}}^{(\mathrm{m})}\widetilde{\mathbf{U}}^{(\sim \mathrm{m})\mathrm{T}}\right)}{\sqrt{\boldsymbol{trace}\left(\widehat{\mathbf{U}}^{(\mathrm{m})}\widehat{\mathbf{U}}^{(\mathrm{m})\mathrm{T}}\right)\boldsymbol{trace}\left(\widetilde{\mathbf{U}}^{(\sim \mathrm{m})}\widetilde{\mathbf{U}}^{(\sim \mathrm{m})\mathrm{T}}\right)}} \qquad \text{(Eq. 13)}$$

The correlation introduces an arbitrary scaling factor between prediction and observations and, in contrast to Eq. 12, allows the model to over- or under-predict the data by a scalar factor without penalty. Encoding models can also be applied to time-series data (Eq. 6), in which case the regression matrix for the estimation of the weight (Eq. 10) and for the prediction of the left-out data (Eq. 11) becomes the product of the design matrix **Z** and the model feature matrix **M**.

To understand how encoding approaches adjudicate between models, consider the graphical representation of the estimation process in Figure 2. The training data point (black cross) is the activity profile of a single measurement channel, which can be visualized as a point in activity-profile space. Regression analysis can be understood as the orthogonal projection of the data point onto the linear subspace spanned by the features of the model matrix. The two models depicted in Fig. 2A and Fig. 2B 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 whose features span a subspace that better describes the cloud of activation profiles will make better predictions and show lower cross-validation error.

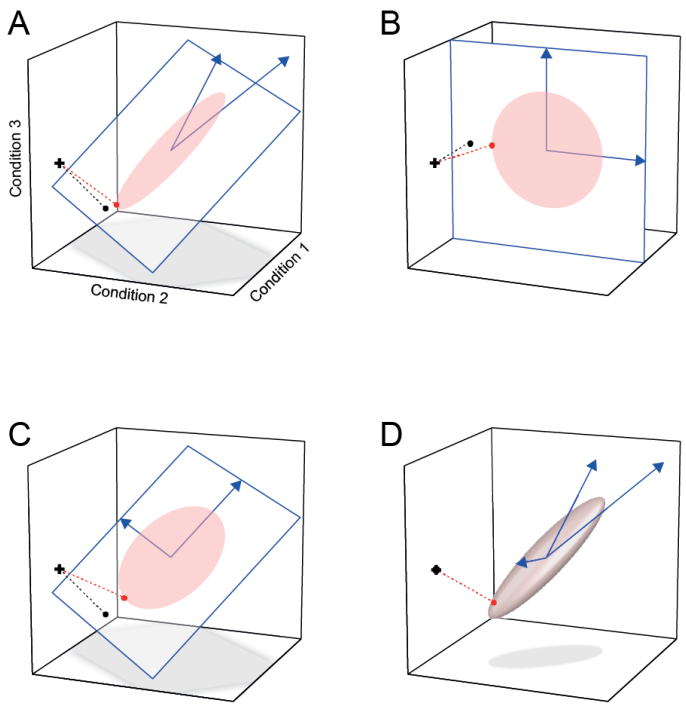

Figure 2. Model comparison using encoding approaches. *The axes of the three-dimensional space are formed by the response to three experimental conditions. The activity profile of each unit defines a point in this space. Models are defined by their features (blue arrows) and a prior distribution of the weights for these features – which together define a distribution of activity profiles (ellipsoids indicate Gaussian iso-probability-density contours). To predict the activity profile of a single measurement channel, the training data (cross) is fitted by the model. Simple regression finds the shortest projection (black dot) onto the subspace defined by the features, whereas regression with a prior (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 a prior leads to a different projection. (**D**) A saturated model with as many features as conditions. Unregularized regression can perfectly fit any data point. Regression with a prior predicts activity profiles more consistent with the prior – here shown as an iso-probability-density contour.*

Encoding approach with prior

Encoding approaches using unregularized linear regression test for subspaces, but not for a well-defined *probability distribution* of activity profiles. For example, the predicted distribution (pink ellipse) of the model depicted in Fig. 2C is distinct from the one in Fig. 2A. The features of these two models, however, span the same subspace. Therefore, the prediction of these two models using unregularized regression (black dots) are identical and the models indistinguishable. However, if we assume a standard normal prior over the weights **W**, then the feature basis in **M** matters, and the two models predict very different distributions of the activation profiles.

The hypothesis about the distribution of the activation profiles can be included in the regression approach as a prior on the **W**. Specifically, we can impose a Gaussian prior with zero mean and variance-covariance matrix $\mathbf{\Omega}s$, where s is a scalar scaling factor that accounts for the different signal levels in different brain regions or individuals. Under this assumption, the best linear unbiased predictor for the unit weights is

$$\widehat{\mathbf{W}} = (\mathbf{M}^{\mathrm{T}}\mathbf{M} + \mathbf{\Omega}^{-1}s^{-1}\sigma_{\epsilon}^{2})^{-1}\mathbf{M}^{\mathrm{T}}\widehat{\mathbf{U}}^{(\mathrm{m})}\,, \qquad \text{(Eq. 14)}$$

where σ_ϵ^2 is the noise variance on the observational units [22]. Linear regression with a prior can also be viewed as a projection of the activity profile into the space spanned by **M**, this time, however, biasing the projection towards the assumed distribution of **W** on the plane (red dot), i.e. towards activity profiles that the model considers to be more probable. Thus, the two models in Fig. 2A and Fig. 2C now make different predictions. The model in which the assumed distribution corresponds more tightly to the data distribution will tend to yield a smaller cross-validation error.

When using a prior, models can also have as many features as conditions, or even more features than conditions (Fig. 2D). Such saturated models would be indistinguishable using unregularized regression. Adding weight-distributional priors can render such models distinct. Thus, using a Bayesian prior is not just a technical trick for estimation, but is an integral part of the hypothesis being tested, as it specifies the assumed probability distribution over activity profiles.

Technically, regression with a Gaussian prior can be implemented using Tikhonov regularization or ridge-regression [22]. The equivalence is established by scaling and rotating the model matrix **M** in such a way that **Ω** becomes the identity matrix. In this context, the strength of the regularization is determined by a scalar ridge coefficient defined by $s^{-1}\sigma_\varepsilon^2$. 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

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

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

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 [5], or using the maximum-likelihood estimate from the training data (Eq. 17). To save time, it is also possible to use a constant regularization coefficient. In our simulations below, we used PCM to estimate $s^{-1}\sigma_\varepsilon^2$ on each training set (across all voxels) and then apply it to make a prediction for the left-out test set.

Pattern component modeling

An alternative to cross-validation is to evaluate the likelihood of the observed activity estimates under the representational model. This approach is taken in pattern-component modeling [10]. We start with a generative model of the observed activity values (Eq. 6). We consider the patterns **U** to be Gaussian random variables, where the columns are distributed to give the second-moment matrix **G**. The nuisance regressors **B** are fixed effects, typically encoding the mean activity of each voxel across conditions within each partition. As in the other analysis approaches, we are not interested in the patterns per se, but want to evaluate their likelihood under different models. The activity measurements for each measurement channel (columns of matrix **Y**) are therefore assumed to be independently distributed as a multivariate normal:

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

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

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

The predicted covariance matrix of the activity measurements for each person is the function of the model (as encoded in the second-moment matrix) and two nuisance parameters: 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

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

and then account for the removal of these fixed effects by evaluating the restricted likelihood [23], i.e. the probability of the data given the model, where the data are activity pattern estimates and the model is defined by second-moments matrix **G** and scalar signal and noise amplitudes $\boldsymbol{\theta}$

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

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

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 set $\boldsymbol{\theta}$ to maximize the likelihood before evaluating the model for each simulation. Because we can obtain closed-form first- and second-order derivatives of Eq. 18 with respect to $\boldsymbol{\theta}$, this can be done efficiently using gradient descent or the Newton-Raphson algorithm. To prevent variance estimates from becoming negative, the optimization is performed on $\mathbf{h} = log(\boldsymbol{\theta})$. Efficient implementations of this algorithm can be found in the open-source Matlab package for PCM [24]. Because every model uses the same two nuisance parameters, models can be compared using the maximal log-likelihood**.**

Representational similarity analysis

Distances and second-moment matrices

In RSA, representational hypotheses are conceptualized in terms of the dissimilarities between the activity patterns (Fig. 2). One important dissimilarity measure is the Euclidean distance, which has a very close relationship 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

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

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 (see below) is the Mahalanobis distance. Correlation distances, another class of popular dissimilarity measures, can also be computed from the second-moment matrix. The cosine angle distance is defined as

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

Here we focus on Euclidian and Mahalanobis distances, as they are independent of the resting baseline and generally easier to interpret [20].

In the following we either represent these distances as a *K×K* representational dissimilarity matrix **D**, or a *K(K-1)/2* vector **d** that contains all pair-wise distances. The vector of all pairwise distances 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, except for the relationship of the activity patterns to the baseline, which is not represented in the distance matrix. 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

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

Multivariate noise normalization and cross-validation: the crossnobis distance

Here we focus on one particular distance measure: the cross-validated squared Mahalanobis distance (or crossnobis distance for short), which has superior characteristics in terms of reliability and interpretability as compared to other distance measures [20].

The crossnobis distance uses multivariate noise normalization (see 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 voxels (for details see [17, 20]).

Furthermore, the crossnobis distance is cross-validated: When one replaces the true activity patterns in Eq. 19 with their noisy estimates, the expected value of the resultant Euclidean distance will be always higher than the true distances, as the noise terms are squared and summed. The Euclidean distance between two pattern estimates, thus, is a positively biased estimator of the true distance. We can obtain an unbiased estimate of the true distance by computing the 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 distance can be computed using a leave-one-out cross-validation scheme:

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

where $\hat{\mathbf{u}}_{\mathrm{i},\cdot}^{(\mathrm{m})}$ is the prewhitened pattern for condition *i* measured on partition *m*, and $\hat{\mathbf{u}}_{\mathrm{i},\cdot}^{(\sim\mathrm{m})}$ is same activity pattern determined from the data of all other partitions. The expected value of this estimator matches the true Mahalanobis distance. 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 testing representational models.

Model comparison

In RSA, different representational models are evaluated by comparing the predicted to the observed distances. The 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 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.

Such arbitrary scaling can be accounted for by calculating the correlation between the predicted and observed distance vectors (not to be confused with the use of correlation distance as a activity-pattern dissimilarity measure, Eq. 20). The most cautious approach is to assume that we can only predict the rank ordering of distances [14]. It is then only appropriate to use Spearman

correlation, or (in the case of split ranks) Kendall's tau 2 [16]. For more quantitative models, it may be appropriate to assume that distance predictions can be made on interval scale rather than an ordinal scale, suggesting the use of the Pearson-correlation coefficient for model comparison [8]. The assumption of a linear relationship between the predicted and measured distances may be justifiable and can increase our sensitivity to differences between representational models.

Both rank-based and linear correlation coefficients are invariant not only with repect to the scaling factor, but also with respect to the intercept of regression. However, the cross-validated Mahalanobis distance has an interpretable zero point, indicating that two activity patterns are not different. 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 correlation coefficient, in which the predictions and the data are not centered about their mean:

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

This amounts to a linear regression model between the predicted and observed distances, where the regression line is constrained to pass through the origin [25]:

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

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 were independent and of equal variance. However, for the crossnobis distance (and for most other distance measures), the assumptions of independence and equal variance are both violated. 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 [17]. To account for these factors, we need to know the predicted probability distribution of the distance 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 [17]

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

The mean of the distribution are the true distances, scaled by a parameter relating to the signal strength in this subject (s). In a separate paper [17], we show that that the exact variance-covariance matrix of **d** is given by

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

Where **G** is the predicted second-moment matrix of the patterns, **C** the contrast matrix that transforms the second-moment matrix into distances, and ∘ 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 activation patterns around their mean ($\overline{\mathbf{B}}$):

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

$\mathbf{\Sigma}_P$ is the voxel-by-voxel correlation matrix of the activation estimates. If multivariate noise-normalization [20] 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 [17].

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

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

To evaluate the likelihood, we first need to estimate the scaling coefficient between predicted and observed distances by maximizing this expression in respect to *s*. This can be done efficiently using iteratively-reweighted least squares (IRLS): Given a starting estimate of **S**, we can obtain the generalized least squares estimate of s,

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

re-estimate **S** according to Eq. 27, and iterate until convergence.

Simulation examples

In this paper we use 3 simulation examples inspired by real fMRI studies. The first two examples come from a paper investigating the representational structure of finger movements in primary motor and sensory cortex [8]. The representational structure of these activity patterns is 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 [26], or by the patterns of muscle activity required for these movements. The predicted representational dissimilarity matrices for individuated movements of the five fingers (Exp. 1) are shown in Fig. 3A,B. The second example comes from experiment 3 from 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 [13]. The model predictions are derived from a convolutional deep neural network model – with each of the 7 layers providing a separate prediction [27].

All data sets where simulated with 8 runs, 160 voxel, 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 (**G**) for each model using Eq. 22 and created true activity pattern that were normally distributed with second moment $\mathbf{UU}^{\mathbf{T}}/P = \mathbf{G}s$. The signal-strength parameter *s* was varied systematically starting from 0 (pure noise data).

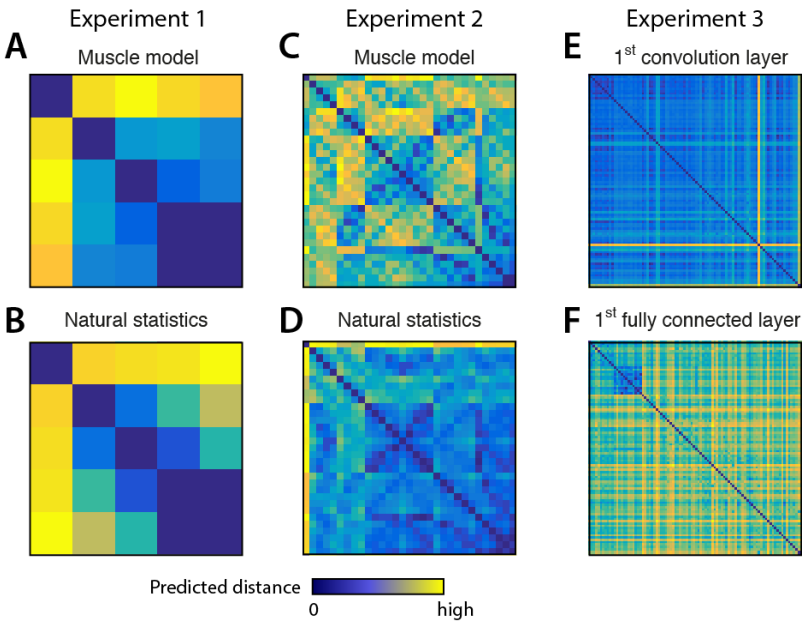

Figure 3. Representational dissimilarity matrices (RDMs) for the models used in simulation. *Each entry of an RDM shows the dissimilarity between the patterns associated with two experimental conditions. RDMs are symmetric about a diagonal of zeros. Note that while zero is meaningfully defined (no difference 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 (either 2 or 7). 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 approaches without a prior on the weights

When employing encoding approaches without a prior, one typically needs to reduce the dimensionality of the model matrix **M**. This can be done by using only the eigenvectors with the *n* highest eigenvalues of the predicted second moment matrix. How many regressors to include is a somewhat arbitrary decision: For example, Leo et al. [9] 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 & Exp 3: $s = 0.3$, Exp 2: $s = 0.1$, $\sigma_\varepsilon^2 = 1$), and compared model selection accuracies using between one and the maximal number of principal components. We used both cross-validated R^2 (Eq. 12) and the correlation between predicted and observed values (Eq. 13) to perform model selection.

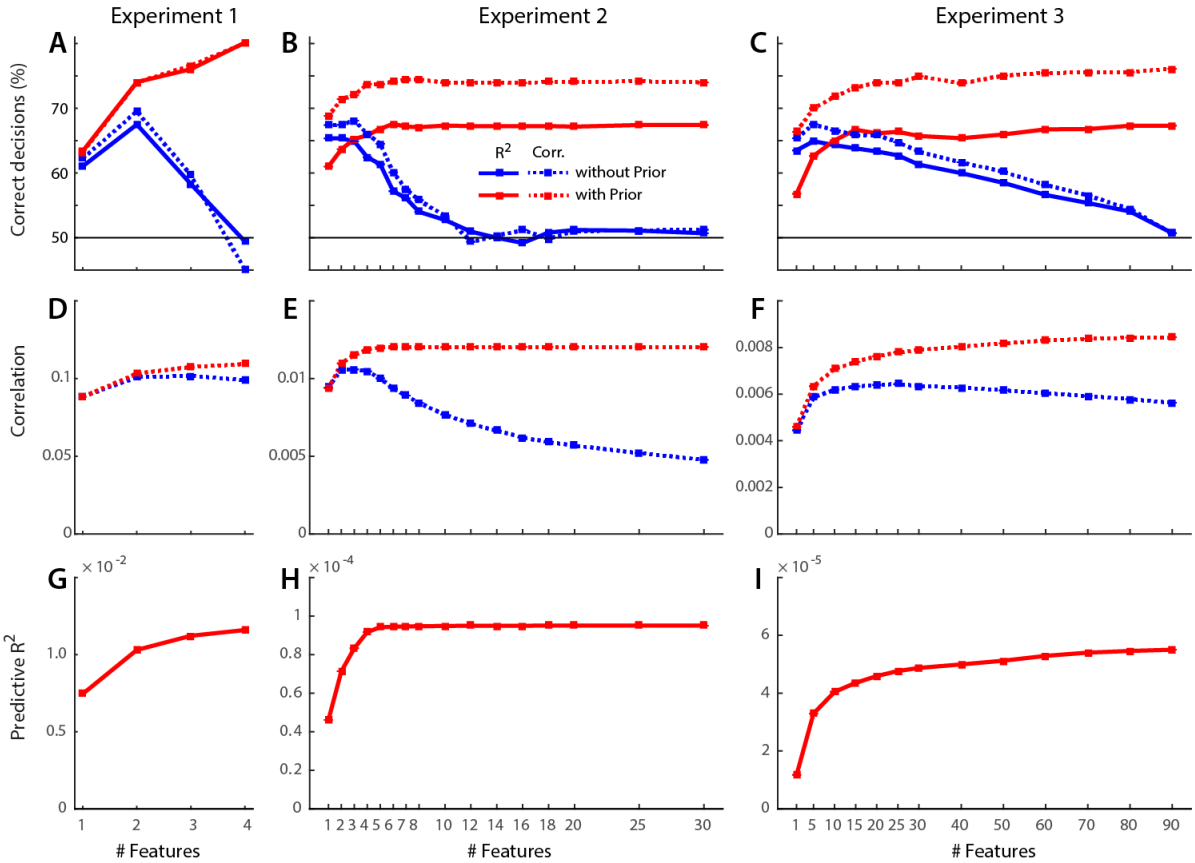

Figure 4. *Dependence of encoding approach without (blue lines) and with a prior (red line) on the number of included features (x-axis).* ***(A-C)*** *Percent correct model selections using either R^2 (solid line) or correlations (dashed line).* ***(D-F)*** *Correlation between predicted and observed patterns (Eq. 13).* ***(G-I)*** *Predictive R2 (Eq. 12) for the encoding approach with prior. All R^2 values without prior are negative.*

Fig. 4A-C shows the percentage of correct model selections for Experiments 1-3. Results obtained without a prior on the weights 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, 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. 2D).

Using correlations as selection criterion led to more accurate decisions than using R^2. Correlations (Fig. 4D-F, blue lines) were generally positive and peaked at a number of features that was slightly higher than the optimal dimensionality for model selection. Predictive R^2 values for encoding without a prior were all negative (and therefore not shown), 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 a prior on the weights

We then fit the same simulated data using an encoding approach with prior (Eq. 14). 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 left-out data patterns. To determine the optimal ridge coefficient, we estimated the signal and noise level on each test data set using PCM.

With a prior, model selection performance increased with increasing number of features (red lines, Fig. 4A-C). Thus, dimensionality reduction of the model is not necessary here. Furthermore, model selection was always more powerful with than without a prior. 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 Exp. 2-3, the R^2 criterion performed substantially worse than the correlation between predicted and observed activity patterns. The difference between the two criteria arises from the fact correlations allow for an arbitrary scaling between predicted and observed activity patterns, whereas R^2 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. 5, where we simulated data from Exp. 2 with a fixed noise and signal strength, and varied the regularization coefficient systematically. While R^2 is highly sensitive to the choice of the regularization coefficient, the correlation criterion is not. Thus, different deviations from the optimal ridge coefficient for different models will decrease model selection accuracy for the cross-validated R^2 criterion, but not for the correlation criterion.

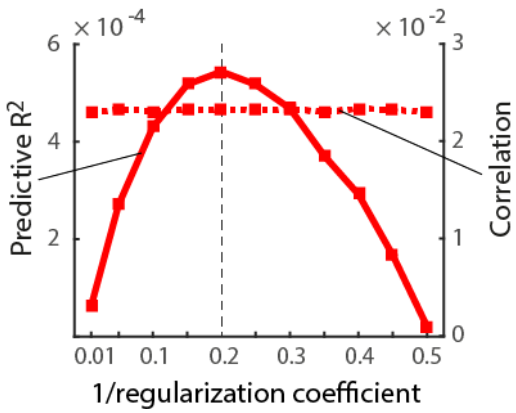

Figure 5. *Sensitivity of the predictive R^2 (solid line) and correlation criterion (dashed line) to the choice of the regularization coefficient. Simulations come from Experiment 2 with a true signal strength of s=0.2 and a noise variance of 1. For this combination the optimal regularization coefficient is $s^{-1}\sigma_\varepsilon^2$ (dashed vertical line). The correlation criterion is generally robust against the non-optimal choice of regularization coefficient.*

In sum, the use of an encoding approach with a prior yields superior model selection performance, even if the model has fewer features than conditions. However, it is important to note that adding a prior changes the hypothesis being tested. Rather than just comparing subspaces, adding a prior on the weights means that more specific hypotheses are being compared. From this perspective it is unsurprising that we can adjudicate between these hypotheses with greater accuracy. Furthermore, the use of correlation instead of predictive R^2 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 whole second-moment matrix will determine the predicted distances. As an empirical distance measure we calculated the cross-validated squared Mahalanobis distance [20] and then compared the predicted to the measured distances. This can be accomplished using a number of different approaches: We employed rank-based correlation of distances [16], 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 distances.

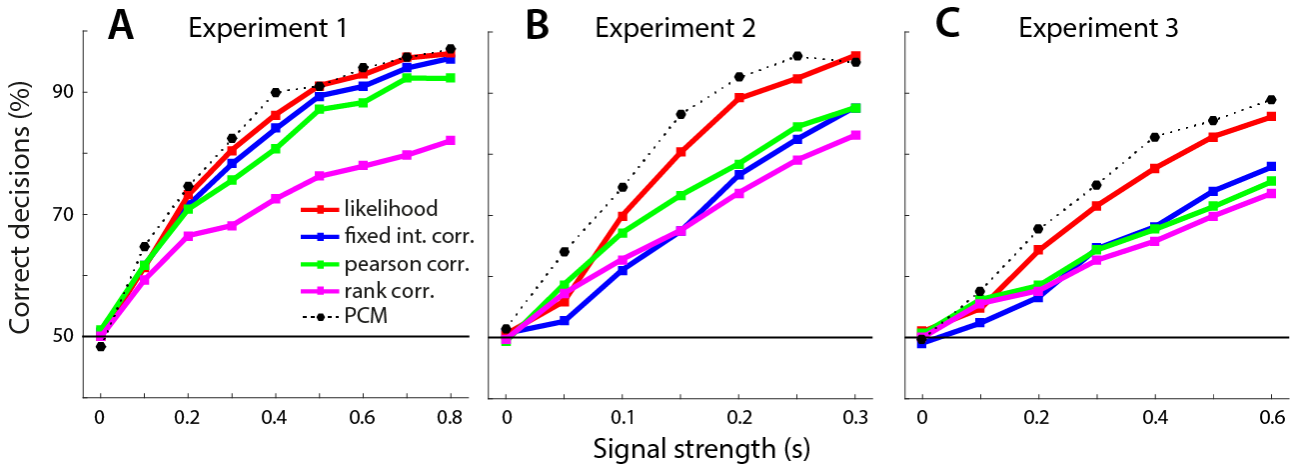

Figure 6. Model selection accuracies for different ways of testing representational models using RSA. *Data sets for all three experiments were generated with varying signal strength (x-axis). The percentage of correct decisions using different criteria is shown (dotted line). Models were selected based on the rank-correlation (purple), Pearson correlation (green), fixed intercept correlation (blue) or likelihood under the multi-normal approximation (red). For comparison the model selection accuracy for PCM is shown in the dotted line.*

For Experiment 1 (Fig. 6A), rank-based correlations performed worst. This decrement in performance may have been exacerbated here by the fact that the distance rank-structure under the two models is relatively similar. However, we expect lower performance in general for rank correlation, because this approach does not rely on the assumption of a linear relationship between predicted and measured dissimilarities. Among the other approaches, likelihood-based RSA yielded the best decisions; slightly better than Pearson and fixed intercept correlations.

The advantage of the likelihood-based approach was clearer for Exp. 2 and 3. Here, it led to about 10% 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 distances, 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 distances, which will influence the model fit strongly. The likelihood-based approach incorporates the knowledge that large distances are measured with substantially larger variability, and hence discounts their influences. Notably, rank-based correlations perform relatively well on these models as compared to Pearson correlations, likely because they are more robust against outlying values for the large predicted distances.

In sum, these simulations show convincingly that when disambiguating between different representational models it is highly advantageous to take the covariance structure of the measured distances 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 [12] that this method should provide us with highest achievable model selection accuracy.

Model selection performance (dotted line in Fig. 6) 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 with a prior - we simulated the three Experiments at a single signal strength (Fig. 7).

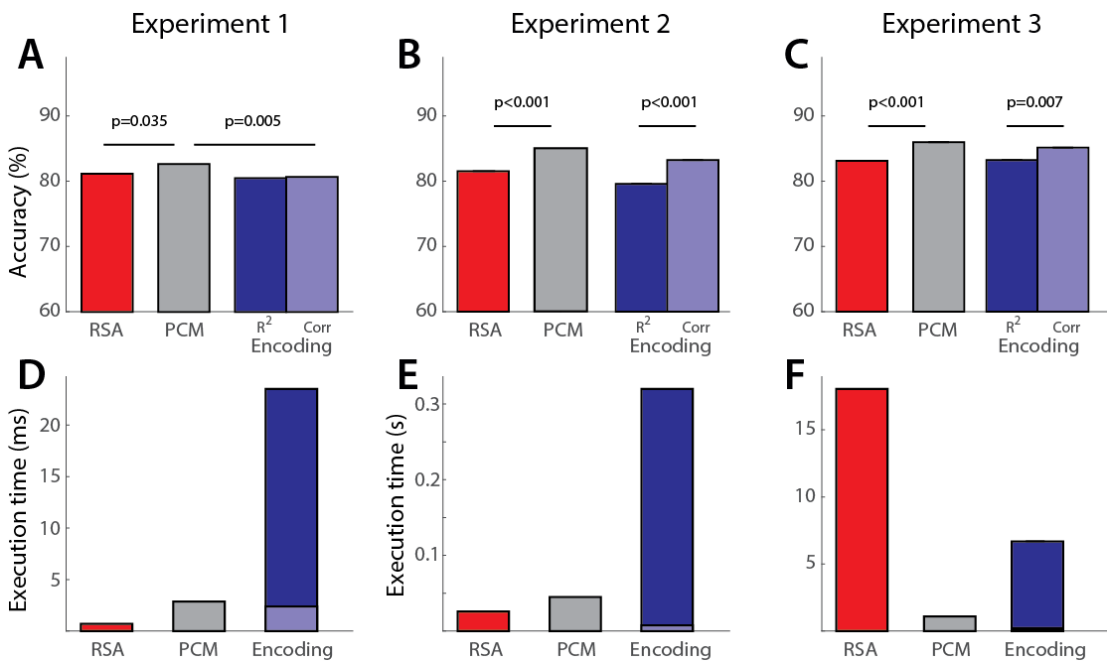

Figure 7. *Model selection accuracy and execution time of likelihood-based RSA, PCM, and encoding approaches using either R^2 (Eq. 12) or correlation (Eq. 13) as a fitting criterion. (**A-C**) Significant differences in model selection accuracy were tested for N=3,000 simulations using a likelihood-ratio test of counts of correct model decisions [28]. 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. (**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 (light blue) and the time to determine the regularization constant (dark blue).*

In this simulation, PCM resulted in 1.48%, 3.01% and 2.86% (for Exp. 1-3, respectively) better model selection accuracy than likelihood-based RSA, and 1.98%, 1.17% and 0.85% higher model selection accuracies than an encoding approach using correlations. In many cases, PCM performed significantly better than the other two approaches (Fig. 7). There were no significant performance differences between RSA and encoding methods. Among encoding approaches, the simulations again showed the advantage of using correlations, rather than R^2 as an evaluation criterion for encoding models. Overall, despite a significant slight advantage for PCM, all three methods were very close in performance.

Computational cost

A practical concern is the speed at which the model comparison can be performed. This is usually not an important constraint when evaluating the model fit on a small number of participants or ROIs. However, if models are evaluated continuously over the cortical surface using a searchlight approach [29, 30], or in data sets with large numbers of participants, computational speed becomes a practical issue.

Both RSA and PCM approaches operate on the inner product matrix of the activity estimates, thus the computational costs for these approaches is virtually independent from the number of voxels. PCM works on the *MK* x *MK* inner product matrix of the activity estimates, whereas RSA operates on a *KxK* matrix of distances between conditions. For small number of conditions, this explains the relatively favorable computational costs for RSA. However, when using likelihood-based RSA, the covariance matrix of the distances needs to be calculated and inverted. The size of this matrix is $(K(K\text{-}1)/2)^2$ and it therefore grows with the 4th power of the number of conditions. For Exp. 3 (Fig 7F) with K = 96, this leads to rather slow performance, whereas PCM still needs to only invert matrices of size MK^2.

For encoding approaches, the optimal ridge coefficient needs to be determined for each cross-validation fold. In our simulations, we used PCM to do so – therefore the computational cost for this task (Fig. 7D-F, dark blue area) is always *M* times higher than for PCM alone. Conducting the actual ridge regression (light blue area) is comparably fast and efficient. Thus, if high speeds are required, one could use a constant ridge coefficient and accept the possible loss in model selection accuracy.

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 both of these aspects of feature strength. Two distinct representations with identical second-moment matrices therefore support decoding of any given feature at the same signal-to-noise ratio. This motivates using the second moment as a summary statistic for characterizing representations.

RSA, PCM and encoding approaches offer different tests of representational models, but all rely on the second moment matrix to characterize the representational hypotheses. The main characteristics of the three methods are summarized in Table 1. Thus, these multivariate methods are highly related and should indeed be understood as part of the same multivariate toolbox.

Encoding approaches without prior test models about subspaces, not distributions

There is a fundamental difference between encoding approaches with and without a prior. Without a prior – when using standard regression analysis to estimate the response weights – encoding approaches test how well the subspace spanned by the model features describes the subspace occupied by the observed activation profiles. This necessitates the dimensionality (i.e. the number of features) of each model to be substantially lower than the number of tested conditions. As the number of model dimensions increases, the subspaces of competing models increasingly overlap. Once two models have as many features as stimuli their subspaces and hence their prediction for unseen data become the same.

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 over the space of possible activity profiles, rather than just a subspace. 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 to make regression more stable, but should be considered an integral part of each representational model. The choice of the prior is therefore a neuroscientific rather than methodological consideration: It specifies what representational hypothesis is to be tested.

Encoding models do not support inferences about the particular model-defining feature set

Even when using a prior, the feature sets that characterize encoding models are not unique. Features should not be artificially constrained to be orthogonal in the space of experimental conditions, as otherwise the structure of the model would depend on the experiment conducted. Without this orthogonality constraint, there are 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. 2A) 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 describe that model. Rather, it is the feature space in conjunction with the prior that determine the representational model. These complications need to be kept in mind in the interpretation of the results of encoding model analyses. It is very tempting to assign meaning to the particular feature basis chosen, especially when they are mapped onto the cortical surface [6, 9]. 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 cross-validated Mahalanobis distance provides a highly reliable measure of dissimilarity with the added advantage of having an interpretable zero-point [20]. Rank-based, Pearson, and fixed-intercept correlations provide a 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. This problem is addressed in likelihood-based RSA, which uses a multivariate-normal approximation to the sampling distribution of the crossnobis distance [17]. 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 performance was only slightly below the theoretically best achievable level, as established by PCM.

We therefore think that likelihood-based RSA might become the approach of choice when comparing representational models using crossnobis distances. However, there are situations, 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. In these cases, rank-correlation-based RSA [16] should remain in our toolkit as a safe, though perhaps not optimally sensitive, method of inference. Likelihood-based RSA becomes computationally expensive as the number of conditions increases. In this case a practical approach is 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 [10] 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 [12] is the most powerful test. The simulations, however, also show that compared to this theoretical ceiling, encoding models with a prior and likelihood-based RSA perform near-optimally. In practice, we therefore expect these three approaches to provide similar answers. While PCM has clear advantages for model comparison in terms of power, computational speed, and analytical tractability, the other two approaches have other distinct advantages that make them attractive choices for specific applications.

RSA 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 [16, 17], or whether the dissimilarity is larger between one pair than between another pair of conditions. Multidimensional scaling of the stimuli on the basis of their representational dissimilarities also provides a intuitive visualization of the representational structure [14], 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 effect of noise through cross-validated distances.

Encoding approaches explicitly estimate the 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 [6, 9]. Furthermore, the model fit can be assessed for each individual measurement channels, rather than for groups of voxels (such as for ROIs or searchlights). Latter advantage, however, comes at the cost that multivariate noise normalization cannot be performed on single voxels. Note also that the searchlight approach for RSA and PCM can be reduced to single voxels, if this is preferred. However, larger searchlights enable us to take noise covariance into account and to pool the local evidence, supporting more robust statistical inferences. Based on our previous results [20], 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.

Many studies using encoding models have tested whether a representational model generalizes to new experimental conditions (e.g. a different sample of visual stimuli). This requires a test set with experimental conditions that are not included in the training set. The features (and their prior) provide a straightforward way of making predictions for such unobserved conditions. Generalization is an especially useful test for representational models with free parameters, which can otherwise overfit of the data. Although used much less frequently, generalization to new conditions can also be assessed within RSA or PCM. Here one simply leaves out part of the second-moment matrix for parameter estimation and then evaluates the model likelihood only for the left-out components [31].

Note that encoding models strictly require independent test sets to account for the over-fitting of the weights. Weight fitting cannot be avoided in encoding models, because individual responses are to be predicted. RSA and PCM rely explicitly on summary statistics of the responses. This renders weight fitting optional. RSA and PCM can therefore compare fixed representational models without cross-validation.

What about decoding approaches?

While decoding approaches are widely used in multivariate analysis of brain imaging data [32-34], we have not considered them in the context of model comparison. It is possible to use classification accuracy instead of correlation or R^2 to evaluate the fit of an encoding model on left-out data [4, 9] – or to use classification accuracy as a measure of dissimilarity between two conditions [35]. 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 that this entails a loss of sensitivity in practice [36]. Hence, decoding approaches are not particularly useful for the evaluation of representational models [3, 11], and should therefore be limited to situations, in which the quality of the decoding is in itself the measure of interest.

Extensions of the present exploration

Our simulations here have focused on a particular class of representational model. There are, however, a number of immediate extensions to our general framework; and it is important to consider how easily each approach can be extended to encompass a broader class of models.

The first major restriction is that 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. $\mathbf{G} = \sum_i \omega_i \mathbf{C_i}$ [10, 31], with the weights being free 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 [37]. 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). Thus, for these methods the extension to parameterized models does not require new algorithms.

In encoding approaches using ridge regression, free model parameters that change the model structure would result in scaling and rotations of the model matrix **M**. Each of these possible changes needs to be evaluated using cross-validation. Hence optimization of parameters will be generally more cumbersome than in the other two approaches, for which analytical derivatives of the likelihood (Eq. 17, 28) in respect to the parameters are easily obtained.

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 [2]. As the underlying brain computational models become more specific and detailed, the corresponding measurement models will also have to be improved.

A final important extension of representational models concerns the distributional assumptions of the true activity profiles of the measurement units. In this paper, we focused on approaches that characterize the distribution 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 across channels to condition A is uncorrelated to the response to condition B, but, for example, that channels either respond to A or B, but with much lower probability to both A and B. Such tuning is for example prevalent in primary visual cortex, where neurons (and voxels) respond a stimulus in a *one* specific part of the visual field, but less often two or more disparate locations [37]. This would correspond to a non-Gaussian prior on the feature weights. In a recent publication, Norman-Haignere and colleagues [38] suggested a likelihood-based method, in which the Gaussian prior on the feature weights **W** is

replaced with a Gamma distribution, 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 presence of non-Gaussian distributions does not invalidate the approaches considered here. Even in this case the second-moment remains a very important characteristic of the activity patterns; it determines the linear decodability, and hence the representation, of all possible features. Taking into account higher moments of the activity profiles would enable us to distinguish between representations that afford the same linear readout of features, but achieve this by distinct distributions of activity profiles.

Summary

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 [1]. 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 approaches are all closely related, testing probabilistic hypotheses about the distribution of activity profiles. Moreover, all three approaches, in their dominant implementations, are sensitive only to distinctions between representations that are reflects 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 between representational models. Its computational efficiency and analytical tractability further makes it the best method to test representational models, 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.
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 general framework presented here will enable researchers to optimally combine these approaches to make progress in uncovering the nature of information processing in the brain.

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. Kriegeskorte N, J. D. Inferring brain-computational mechanisms with models of activity measurements. Proceedings of the Royal Society. 2016.
3. 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.
4. 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.
5. 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.
6. 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.
7. Georgopoulos AP, Schwartz AB, Kettner RE. Neuronal population coding of movement direction. Science. 1986;233(4771):1416-9.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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].
14. 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.
15. 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.
16. 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.
17. Diedrichsen J, Zareamoghaddam H, Provost S. The distribution of crossvalidated mahalanobis distances. ArXiv. 2016.
18. 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.
19. Worsley KJ, Friston KJ. Analysis of fMRI time-series revisited--again. Neuroimage. 1995;2(3):173-81.
20. Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J. Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage. 2015. doi: 10.1016/j.neuroimage.2015.12.012.
21. 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).
22. Murphy KP. Machine Learning: A probabilistic perspective. Cambridge, MA: MIT press; 2012.

23. 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: doi: 10.1080/01621459.1977.10480998.
24. Diedrichsen J, Yokoi A, Arbuckle S. Pattern component modeling toolbox. 2016. Available from: https://github.com/jdiedrichsen/pcm_toolbox.
25. Eisenhauer JG. Regression through the origin. Teaching Statistics. 2003;25(3):76-80.
26. 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.
27. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. NIPS; Lake Tahoe, Nevada2012.
28. Sokal RR, Rohlf FJ. Biometry: the principles and practice of statistics in biological research. 2nd ed. San Fransisco: W. H. Freeman; 1981.
29. 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.
30. Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. Proc Natl Acad Sci U S A. 2006;103(10):3863-8.
31. 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.
32. 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.
33. 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].
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.

Appendix

Notation

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
N_m:		Number of measurements for partition m
$\mathbf{U}$:	KxP	Matrix of true activation patterns
$\boldsymbol{u}_{i,\cdot}$:	$1xP$	Activation pattern for condition i; i^{th} row of $\mathbf{U}$
$\boldsymbol{u}_{\cdot,j}$:	$Kx1$	Activation profile for measurement channel j; j^{th} column of U
$\widehat{\mathbf{U}}^{(m)}$:	KxP	Matrix of estimated activity patterns, based on data from partition m
$\widetilde{\mathbf{U}}^{(\sim m)}$:	1xP	Model prediction for activity patterns, based on data independent of m
$\mathbf{M}$:	KxQ	Matrix of model features for all condition
$\mathbf{W}$:	QxP	Matrix of voxel weights for each feature
$\mathbf{Y}$:	$N\ x\ P$	Matrix of brain measurements, concatenated activity estimates or time series data
$\mathbf{Z}$:	$N\ x\ K$	Design matrix, indicating how measurements relate to activity patterns
$\mathbf{X}$:	$N\ x\ Q$	Design matrix containing n regressors of no-interest
$\mathbf{G}$:	KxK	Second moment of $\mathbf{U}$
d_j:		j^{th} distance, defined between condition i and k
J:		Number of distances, normally $K(K\text{-}1)/2$
$\mathbf{D}$:	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
$\boldsymbol{\Sigma}_P$:	PxP	Variance-covariance matrix between the P voxels
$\boldsymbol{\Sigma}_K$:	KxK	Variance-covariance matrix of the columns of $\widehat{\mathbf{U}}^{(m)}$
$\mathbf{V}$:	NxN	Variance-covariance matrix of $\mathbf{Y}$
$\mathbf{S}$:	JxJ	Variance-covariance matrix of all pair-wise distances

Tables

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

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