NeuroImage 180 (2018) 119–133

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Pattern component modeling: A flexible approach for understanding the representational structure of brain activity patterns

Jörn Diedrichsen [a,b,c,*], Atsushi Yokoi [a,d], Spencer A. Arbuckle [a,e]

[a] Brain and Mind Institute, Western University, Canada
[b] Department of Statistical and Actuarial Sciences, Western University, Canada
[c] Department of Computer Science, Western University, Canada
[d] Graduate School of Frontier Biosciences, Osaka University, Japan
[e] Department of Neuroscience, Western University, Canada

ARTICLE INFO

Keywords:
Multi-voxel pattern analysis
fMRI
Bayesian models
Motor representations

ABSTRACT

Representational models specify how complex patterns of neural activity relate to visual stimuli, motor actions, or abstract thoughts. Here we review pattern component modeling (PCM), a practical Bayesian approach for evaluating such models. Similar to encoding models, PCM evaluates the ability of models to predict novel brain activity patterns. In contrast to encoding models, however, the activity of individual voxels across conditions (activity profiles) are not directly fitted. Rather, PCM integrates over all possible activity profiles and computes the marginal likelihood of the data under the activity profile distribution specified by the representational model. By using an analytical expression for the marginal likelihood, PCM allows the fitting of flexible representational models, in which the relative strength and form of the encoded feature spaces can be estimated from the data. We present here a number of different ways in which such flexible representational models can be specified, and how models of different complexity can be compared. We then provide a number of practical examples from our recent work in motor control, ranging from fixed models to more complex non-linear models of brain representations. The code for the fitting and cross-validation of representational models is provided in an open-source software toolbox.

1. Introduction

The study of brain representations aims to illuminate the relationship between complex patterns of activity in the brain and "things in the world" - be it objects, actions, or abstract concepts. By understanding internal syntax of brain representations, and especially how the structure of representations varies across different brain regions, we ultimately hope to gain insight into the way the brain processes information.

Central to the definition of representation is the concept of decoding (deCharms and Zador, 2000). A feature (i.e. a variable that describes some aspect of the "things in the world") that can be decoded from the ongoing neural activity in a region may be said to be represented there. For example, a feature could be the direction of a movement, the orientation and location of a visual stimulus, or the semantic meaning of a word. Of course, if we allow the decoder to be arbitrarily complex, we would use the term representation in the most general sense. For example, using a computer vision algorithm, one may be able to identify objects based on activity in primary visual cortex. However, we may not necessarily conclude that object identity is represented in V1 - at least not explicitly. Therefore, it makes sense to restrict our definition of a representation explicitly to features that can be linearly decoded from the population activity (DiCarlo et al., 2012; DiCarlo and Cox, 2007; Kriegeskorte, 2011; Diedrichsen and Kriegeskorte, 2017).

While decoding approaches are very popular in the study of multi-voxel activity patterns (Haxby et al., 2001; Norman et al., 2006; Pereira et al., 2009), they are not particularly useful when making inferences about the nature of brain representations. The fact that we can decode feature X well from region A does not imply that the representation in A is well characterized by feature X - there may be many other features that better describe the activity patterns in this region.

Encoding analysis offers a solution to this problem, as it characterizes how well the activities in a specific region can be explained by a given set of features. Each column in the data matrix (Fig. 1a) is an activity profile of a single voxel. Note that we will in the following use the term voxel

* Corresponding author. Brain and Mind Institute, Natural Sciences Centre, Western University, London, Ontario, N6A 5B7, Canada.
E-mail address: jdiedric@uwo.ca (J. Diedrichsen).

http://dx.doi.org/10.1016/j.neuroimage.2017.08.051
Received 21 March 2017; Received in revised form 21 July 2017; Accepted 18 August 2017
Available online 24 August 2017

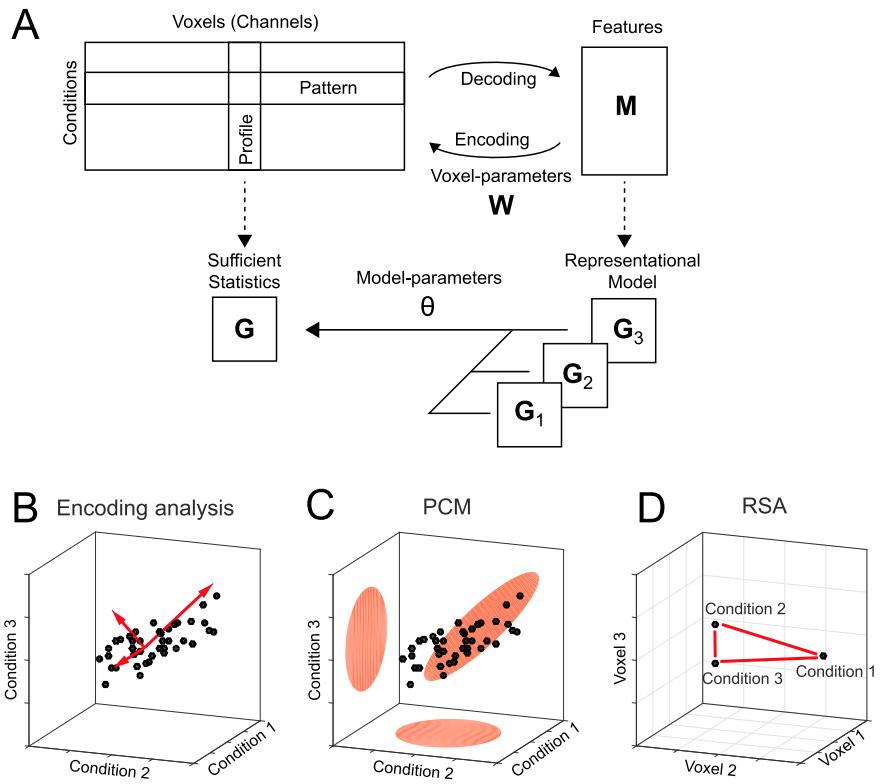

Fig. 1. Decoding, encoding, and representational models. (**A**) Activity data is measured across a number of conditions and voxels (or more generally, measurement channels). The data can be viewed in terms of activity patterns (rows) or activity profiles (columns). The data can be used to decode specific features that describe the experimental conditions (decoding). Alternatively, a set of features can be used to predict the activity data (encoding). Rather than fitting voxel-parameters, representational models work at the level of a sufficient statistics (the second moment, **G**) of the activity profiles. Models are also formulated in terms of their predicted second moment. Different feature models can be flexibly combined using second-level parameters (θ). (**B**) In Encoding analysis, the activity profile of each voxel is viewed as a points in the space of the experimental conditions. The distribution of activity profiles is then described by a set of features (red vectors). (**C**) In PCM, the distribution of activity profiles is directly fitted using a multivariate normal distribution.(**D**) Representational similarity analysis (RSA) provides an alternative view by plotting the activity patterns in the space defined by different voxel activities. The distances between activity patterns serves here as the sufficient statistic, which is fully defined by the second moment matrix.

interchangeably with the more general term of measurement channel, which could, depending on the measurement modality, refer to a single neuron, an electrode, or sensor. In encoding analyses each activity profile is modeled as the linear combination of a set of features. Each voxel therefore has its own set of parameters (**W**) that determine the weight of each feature. This approach can visualized by plotting the activity profile of each voxel into the space spanned by the experimental conditions (Fig. 1b). Each dot refers to the activity profile, indicating how strongly the voxel is activated by each condition. Estimating the weights is equivalent to a projection of each of the activity profiles onto the feature vectors. The quality of the model can then be evaluated by determining how well unseen activity data can be predicted. When estimating the weights, encoding models often use some form of regularization, which essentially imposes a prior on the feature weights. This prior is an important component of the model. It determines a predicted distribution of the activity profiles (Diedrichsen and Kriegeskorte, 2017). An encoding model that matches the real distribution of activity profiles best will show the best prediction performance.

The interpretational problem for encoding models is that for each feature set that predicts the data well, there is an infinite number of other (rotated) features sets that describe the same distribution of activity profiles and, hence, predict the data equally well. The argument may be made that to understand brain representations, we should not think about specific features that are encoded, but rather about the distribution of activity profiles. This can be justified by considering a read-out neuron that receives input from a population of neurons. From the standpoint of this neuron, it does not matter which neuron has which activity profile (as long as it can adjust input weights) and which features were chosen to describe these activity profiles - all that matters is the information can be read out from the code. Thus, from this perspective it may be argued that the formulation of specific feature sets and the fitting of feature weights for each voxel are unnecessary distractions in understanding the information content of the activity patterns.

Therefore, the approach of pattern component modeling (PCM) abstracts from specific activity patterns. This is done by summarizing the data using a suitable summary statistic (Fig. 1a) that describes the shape of the activity profile distribution (Fig. 1c). This critical characteristic is the covariance matrix (or more generally, the second moment) of the activity profile distribution. The second moment determines how well we can linearly decode any feature from the data. If, for example, activity measured for two experimental conditions is highly correlated in all voxels, then the difference between these two conditions will be very difficult to decode. If however, the activities are uncorrelated, then decoding will be very easy. Thus, the second moment is a central statistical quantity that determines the representational content of the brain activity patterns of an area (Diedrichsen and Kriegeskorte, 2017).

Similarly, a representational model is formulated in PCM not by its specific feature set, but by its predicted second moment matrix. If two feature sets have the same second moment matrix, then the two models are equivalent. Thus, PCM makes hidden equivalences between encoding models explicit. To evaluate models, PCM simply compares the likelihood of the data under the distribution predicted by the model. To do so, we rely on a generative model of brain activity data, which fully specifies the distribution and relationship between the random variables. Specifically, the true activity profiles are assumed to have a multivariate Gaussian distribution and the noise is also assumed to be Gaussian with known covariance structure. Having a fully-specified generative model allows us to calculate the marginal likelihood of data under the model,

averaged over all possible values of the feature weights. This results in the so-called model evidence, which can be used to compare different models directly, even if they have different numbers of features. In summarizing the data using a sufficient statistic, PCM is closely linked to representation similarity analysis (RSA), which characterizes the second moment of the activity profiles in terms of the distances between activity patterns (Fig. 1d). In a previous paper we have extensively compared these three approaches (Diedrichsen and Kriegeskorte, 2017) in their ability to infer on so-called fixed representational models, i.e. models that consist of one specific feature set or predicted representational structure. We showed that, if assumptions of the generative model are met, PCM provides the likelihood-ratio test between different models and therefore outperforms, as predicted by the Neyman-Pearson Lemma (Neyman and Pearson, 1933), both encoding analysis and RSA in the number of correct model inferences.

In this paper, we focus on a different advantage of PCM, namely that it removes the requirement to fit and cross-validate individual voxel weights by providing a single analytical expression for the marginal likelihood of the data under the model. This enables the user to easily fit second-level parameters, namely model parameters that determine the shape of the distribution of activity profiles. From the perspective of encoding models, these would be parameters that change the form of the feature matrix. For example, we can fit the distribution of activity profiles using a weighted combination of 3 different feature sets (Fig. 1a). Such component models (see section 2.2.2) are extremely useful if we hypothesize that a region cares about different groups of features (i.e. colour, size, orientation), but we do not know how strongly each feature is represented. In encoding models, this would be equivalent to providing separate scaling factors for different parts of the feature matrix (de Heer et al., 2017).

In section 2, we will present the fundamentals of the generative approach taken in PCM and outline different ways in which flexible representational models with free parameters can be specified. We will then discuss methods for model fitting and model evaluation. In section 3, we provide three illustrative examples from our work on finger representations in primary sensory and motor cortices, highlighting different representational models and ways of providing inferences on them. We also show that PCM continues to make stable inferences when the assumption of Gaussianity is violated. The methods presented in this paper are available as an open-source Matlab toolbox (Diedrichsen et al., 2016a).

2. Methods

2.1. Generative model

Central to PCM is a generative model of the measured brain activity data **Y**, a matrix of *NxP* activity measurements, referring to *N* time points (or trials) and *P* voxels (for notation see appendix A). The data can refer to the minimally preprocessed raw activity data, or to already deconvolved activity estimates, such as those obtained as beta weights from a first-level time series model. **U** is the matrix of true activity patterns (a number of conditions x number of voxels matrix) and **Z** the design matrix. Also influencing the data are effects of no interest **B** and noise:

$$\mathbf{Y} = \mathbf{ZU} + \mathbf{XB} + \boldsymbol{\varepsilon} \tag{1}$$

$$\mathbf{u}_p \sim N(0, \mathbf{G}) \tag{2}$$

$$\boldsymbol{\varepsilon}_\mathbf{p} \sim N\left(0, \mathbf{S}\sigma^2\right) \tag{3}$$

There are a five assumptions in this generative model. First, the activity profiles ($\mathbf{u}_p$, columns of **U**) are considered to be a random variable. Representational models therefore do not specify the exact activity profiles of specific voxels, but simply the characteristics of the distribution from which they originate. Said differently, PCM is not interested in which voxel has which activity profiles - it ignores their spatial arrangement. This makes sense considering that activity patterns can vary widely across different participants (Ejaz et al., 2015) and do not directly impact what can be decoded from a region. For this, only the distribution of these activity profiles in this region is important. The decision to treat $\mathbf{u}_p$, as a random variable makes PCM different from classical linear multivariate models, such as MANOVA or MANCOVA. Rather, the imposition of a prior on $\mathbf{u}_p$, puts PCM in the class of hierarchical Bayesian linear models.

The second assumption is that the mean activity pattern (the mean activity of each voxel across conditions) is modeled using the effects of no interest. Therefore, **X** should include an intercept term that models the condition-independent mean of each activity profile. While one could also artificially remove the mean of each condition across voxels (Walther et al., 2016), this approach would remove differences that, from the perspective of decoding and representation, are highly meaningful (Diedrichsen and Kriegeskorte, 2017).

The third assumption is that both the true activity profiles and the noise come from a multivariate Gaussian distribution. This assumption is common to many analysis approaches to fMRI data, and enables here the analytical derivation of the marginal likelihood. The main consequence of the Gaussian assumption is that it causes PCM to focus on the second moment as the sufficient statistic, as the first two moments completely determine the multivariate Gaussian distribution. We believe that this assumption is sensible for 3 reasons. Even if the true distribution of the activity profiles is better described by a non-Gaussian distribution, the focus on the second moment makes sense, as it characterizes the linear decodability of any feature of the stimuli (Diedrichsen and Kriegeskorte, 2017). Secondly, while real fMRI data typically show some deviations from Gaussianity, these violations are usually mild due to the large-scale averaging inherent in the methods. Finally, inferences with PCM remain robust and superior relative to other competing methods, even when the data is non-Gaussian (see section 3.4).

Fourthly, the model assumes that different voxels are independent from each other. If we used raw data, this assumption would clearly be violated, given the strong spatial correlation of noise processes in fMRI data. To reduce these dependencies we typically uses spatially pre-whitened data, which is divided by an estimate of the spatial covariance matrix (Walther et al., 2016; Diedrichsen and Kriegeskorte, 2017). One complication here is that spatial pre-whitening usually does not remove spatial dependencies completely, given the estimation error in the spatial covariance matrix. While residual dependencies can be taken into account when calculating the likelihood (Diedrichsen et al., 2016b), this will not alter the rank-ordering of different model likelihoods, but simply their scaling.

Finally, we assume that the temporal covariance of the activity data is known up to a constant term σ^2. The original formulation of PCM used a model which assumed that the noise is also temporally independent and identically distributed (i.i.d.) across different trials, i.e. $\mathbf{S} = \mathbf{I}$. However, as pointed out recently (Cai et al., 2016), this assumption is often violated in non-random experimental designs with strong biasing consequences for estimates of the covariance matrix. PCM therefore allows the user to provide an estimate of the true covariance structure of the noise (**S**), or derive parts of the noise structure within the model-fitting procedure (see section 2.3).

2.2. Representational model types

Given our definition (in section 2.1), a representational model is fully specified by the second-moment matrix of the activity profiles (**G**). Many models, including encoding models with a fixed feature set, predict a specific structure of this matrix (fixed models). In other situations we may wish to estimate the most likely structure of **G** from the data without constraints (free models). The most interesting cases, however, are models that impose some constraints on the possible structure of **G**, with the exact form depending on some additional model parameters θ.

2.2.1. Fixed models

In fixed models, the second moment matrix **G** is exactly predicted by the model. The simplest example is the Null model, which states that $\mathbf{G} = \mathbf{0}$. This is equivalent to assuming that there is no difference between the activity patterns measured under any of the conditions. The Null-model is useful if we want to test whether there are any differences between experimental conditions.

Fixed models also occur when the representational structure is predicted from independent data. An example for this is shown in section 3.1, where we predict the structure of finger representations directly from the correlational structure of finger movements in every-day life (Ejaz et al., 2015). Importantly, fixed models only predict the second moment matrix up to a proportional constant. The width of the distribution will vary with the overall signal-to-noise-level (assuming we use pre-whitened data). Thus, when evaluating fixed models we allow the predicted second moment matrix to be scaled by an arbitrary positive constant, $exp(\theta_s)$. Encoding models with a specific feature set are also equivalent to a fixed PCM model. The estimation of the common ridge coefficient in encoding analysis takes the place of the determination of the signal (θ_s) and noise (θ_ε) parameter in PCM.

2.2.2. Component models

A more flexible model is to express the second moment matrix as a linear combination of different components. For example, the representational structure of activity patterns in the human object recognition system in inferior temporal cortex can be compared to the response of a convolutional neural network that is shown the same stimuli (Khaligh-Razavi and Kriegeskorte, 2014). Each layer of the network predicts a specific structure of the second moment matrix and therefore constitutes a fixed model. However, the real representational structure seems to be best described by a mixture of multiple layers. In this case, the overall predicted second moment matrix is a weighted sum of component matrices:

$$\mathbf{G} = \sum_h \exp(\theta_h)\mathbf{G}_h. \tag{4}$$

The weights for each component need to be positive - allowing negative weights would not guarantee that the overall second moment matrix would be positive definite. Therefore we use the exponential of the weighing parameter here, such that we can use unconstrained optimization to estimate the parameters.

2.2.3. Feature models

A representational model can be also formulated in terms of the features that are thought to be encoded in the voxels, as done in encoding approaches. Features are hypothetical tuning functions, i.e. models of what activation profiles of single neurons could look like. Examples of features would be Gabor elements for lower-level vision models (Kay et al., 2008), elements with cosine tuning functions for different movement directions for models of motor areas (Eisenberg et al., 2010), and semantic features for association areas (Huth et al., 2016). The actual activity profile of each voxel is a weighted combination of the feature matrix $\mathbf{u}_p = \mathbf{M}\mathbf{w}_p$. To derive the predicted second moment matrix of the activity profiles from the feature set, we need to make an assumption about the distribution of $\mathbf{w}_p$. Analogous to the use of ridge-regression in encoding analysis (Diedrichsen and Kriegeskorte, 2017), one option is to assume that all features are equally strongly and independently encoded, i.e. $E(\mathbf{w}_p\mathbf{w}_p^T) = \mathbf{I}$. Under this assumption the second moment becomes $\mathbf{G} = \mathbf{M}\mathbf{M}^T$. A feature model can now be flexibly parametrized by expressing the feature matrix as a weighted sum of linear components.

$$\mathbf{M} = \sum_h \theta_h \mathbf{M}_h \tag{5}$$

Each parameter θ_h determines how strong the corresponding set of features is represented across the population of voxels. Note that this parameter is different from the actual feature weights **W**. Under this model, the second moment matrix becomes

$$\mathbf{G} = \mathrm{U}\mathrm{U}^T/P = \frac{1}{P}\sum_h \theta_h^2 \mathbf{M}_h\mathbf{M}_h^T + \sum_i \sum_j \theta_i\theta_j \mathbf{M}_i\mathbf{M}_j^T. \tag{6}$$

From the last expression we can see that, if features that belong to different components are independent of each other, i.e. $\mathbf{M}_i\mathbf{M}_j = \mathbf{0}$, then a feature model is equivalent to a component model with $\mathbf{G}_h = \mathbf{M}_h\mathbf{M}_h^T$. The only technical difference is that we use the square of the parameter θ_h, rather than its exponential, to enforce non-negativity. Thus, component models assume that the different features underlying each component are encoded independently in the population of voxels - i.e. knowing something about the tuning to feature of component A does not tell you anything about the tuning to a feature of component B. If this cannot be assumed, then the representational model is better formulated as a feature model.

In summary, feature and component models test how well combination of feature set or feature spaces can explain the neuronal data, very similar to the variance partitioning approach for encoding models (de Heer et al., 2017; Lescroart et al., 2015). The advantage of performing this analysis in PCM is that the optimal weighting of each feature space in the combination can be easily determined. Note also that MANOVA models with factorial designs (Allefeld and Haynes, 2014) can be written as component or feature models, with one component for each main effect or interaction.

2.2.4. Nonlinear models

The most flexible way of defining a representational model is to express the second moment matrix as a non-linear (matrix valued) function, $\mathbf{G} = F(\boldsymbol{\theta}_m)$. While often a representational model can be expressed as a component or feature model, sometimes this is not possible. One example is a representational model in which the width of the tuning curve (or the width of the population receptive field) is a free parameter (Dumoulin and Wandell, 2008). Such parameters would influence the features, and hence also the second-moment matrix in a non-linear way. Computationally, such non-linear models are not much more difficult to estimate than component or feature models, assuming that one can analytically derive the matrix derivatives $\partial\mathbf{G}/\partial\theta$ for each non-linear parameter. We provide an example of a very useful non-linear model below (see section 3.3).

2.2.5. Free models

The most flexible representational model is the free model, in which the predicted second moment matrix is unconstrained. Thus, the estimation of this model simply derives the maximum-likelihood estimate of the second-moment matrix in the presence of noise. A free model can be useful for a number of reasons. First, we may want to estimate the likelihood of the data under a free model to obtain a noise ceiling - i.e. an estimate of how well the most flexible model can fit the data (see section 2.8). Secondly, we may want an estimate of the second moment matrix to derive the corrected correlation between different patterns, which is less influenced by noise than the simple correlation estimate (Cai et al., 2016; Diedrichsen et al., 2011). Note, however, that this estimator still has the problem of being biased in small samples (see 3.2). In estimating an unconstrained **G**, it is important to ensure that the estimate will still be a positive definite matrix. For this purpose, we express the second moment as the square of an upper-triangular matrix, $\mathbf{G} = \mathbf{A}\mathbf{A}^T$ (Cai et al., 2016; Diedrichsen et al., 2011). The parameters $\boldsymbol{\theta}_m$ are then simply all the upper-triangular entries of **A**.

2.3. Noise assumptions

To complete our generative model, we need to introduce some assumptions about the distribution of the noise. In general, we assumed that the noise comes from a multivariate normal distribution with

covariance matrix $\mathbf{S}\sigma^2$ (see section 3.4 for violations of this assumption). What is a reasonable noise structure to assume? First, the data can usually be assumed to be independent across imaging runs. If the data are regression estimates from a first-level model, and if the design of the experiment is balanced, then it is usually also permissible to make the assumption that the noise is independent within each imaging run $\mathbf{S} = \mathbf{I}$, (Diedrichsen et al., 2011). Usually, however, the regression coefficients from a single imaging run show positive correlations with each other. This is due to the fact that the regression weights measure the activation during a condition as compared to a resting baseline, and the resting baseline is common to all conditions within the run (Diedrichsen et al., 2011). To account for this, one can model the mean activation (across conditions) for each voxel with a separate fixed effect for each run. This effectively accounts for any uniform correlation.

Usually, assuming equal correlations of the activation estimates within a run is only a rough approximation to the real covariance structure. A better estimate can be obtained by using an estimate derived from the design matrix and the estimated temporal autocorrelation of the raw signal. As pointed out recently (Cai et al., 2016), the structure of some experimental designs can have substantial biasing influence on the estimation of the second moment matrix. This is especially evident in cases where the sequence of trials is not random, but in which one specific condition is more likely followed by another specific condition. In these cases, the accuracy of inference hinges critically on the quality of our estimate of the temporal auto-covariance structure of the true noise **S**. Note that it has been recently demonstrated that especially for high sampling rates, a simple autoregressive model of the noise is insufficient (Eklund et al., 2012).

The last option is to estimate the covariance structure of the noise from the data itself within the PCM model. This can be achieved by introducing random effects into the generative model (Eq. 1) that account for the noise covariance structure. One example used here is to assume that the data are independent within each imaging run, but share an unknown covariance, which is then estimated as a part of the covariance matrix (Diedrichsen et al., 2011). While this approach is similar to just removing the run mean from the data as a fixed effect (see above) it is a good strategy if we actually want to model the difference of each activation pattern against the resting baseline. When treating the mean activation pattern in each run as a random effect, the algorithm finds a compromise between how much of the shared pattern in each run to ascribe to the random run-to-run fluctuations, and how much to ascribe to a stable mean activation. This approach is use in example 3.3.

2.4. Likelihood and optimization

Given our generative model we can derive the likelihood of the data under the model. Importantly, we do not want the likelihood for specific values of the estimates of the true activity patterns **U**. This is a difference to encoding approaches, in which we would estimate the values of **U** by estimating the feature weights **W**. In PCM, we want to assess how likely the data is under any possible value of **U**, as specified by the prior distribution. Thus, we wish to calculate the marginal likelihood

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \int p(\mathbf{Y}|\mathbf{U},\boldsymbol{\theta}) p(\mathbf{U}|\boldsymbol{\theta}) d\mathbf{U}. \tag{7}$$

Given that all the involved distributions are normal, this marginal distribution of a single measured activity profile $\mathbf{y}_i$ (given the fixed effects estimates $\mathbf{b}_i$) can be derived as

$$\begin{aligned} p(\mathbf{y}_i|\boldsymbol{\theta},\mathbf{b}_i) &= N(\mathbf{X}\mathbf{b}_i, \mathbf{V}(\boldsymbol{\theta})) \\ \mathbf{V}(\boldsymbol{\theta}) &= \mathbf{Z}\mathbf{G}(\boldsymbol{\theta})\mathbf{Z}^T + \mathbf{S}\sigma_\varepsilon^2 \\ \sigma_\varepsilon^2 &= \exp(\theta_\varepsilon) \end{aligned} \tag{8}$$

To make our marginal likelihood unconditional on the fixed effects, we need to take into account that we estimate $\mathbf{b}_i$ from the data and then use the residuals to calculate the likelihood. This removal can be written as a pre-multiplication with the residual-forming matrix

$$\mathbf{R}(\boldsymbol{\theta}) = \mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}(\boldsymbol{\theta})^{-1}, \tag{9}$$

with the residual after removing the fixed effects being **RY**. With this in hand, our final marginal log-likelihood of all data (assuming independence of the voxels) is

$$\begin{aligned} \log p(\mathbf{Y}|\boldsymbol{\theta}) &= -\frac{NP}{2} ln(2\pi) - \frac{P}{2}\ln(|\mathbf{V}(\boldsymbol{\theta})|) \\ &\quad -\frac{1}{2} trace(\mathbf{Y}\mathbf{Y}^T\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{R}(\boldsymbol{\theta})) - \frac{P}{2}\ln|\mathbf{X}^T\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}|, \end{aligned} \tag{10}$$

where the last term accounts for the removal of the fixed effects. The interested reader can find a full derivation of the log likelihood and its derivatives in respect to the parameters in the manual accompanying our PCM toolbox. For maximum likelihood estimation, we can use conjugate gradient descent, Newton-Raphson (Lindstrom and Bates, 1988), or Expectation-Maximization (McLachlan and Krishnan, 1997). Fast implementations of these algorithms are available and described in the toolbox (Diedrichsen et al., 2016a).

2.5. Cross-validated estimate of the second moment matrix

Instead of the full optimization of Eq. 10, we can also obtain simpler estimates of the second moment matrix. The first option is to first estimate the average activity patterns for each condition using linear regression, and then calculate the second moment matrix of these activity patterns.

$$\widehat{\mathbf{U}} = \left(\mathbf{Z}^T\mathbf{S}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{S}^{-1}\mathbf{Y} \tag{11}$$

$$\widehat{\mathbf{G}}_{simple} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^T/P \tag{12}$$

The problem with this naive estimate is that it is positively biased by the measurement noise. In the absence of fixed effects, the expected value of this estimate is

$$E\left(\widehat{\mathbf{G}}_{simple}\right) = \mathbf{G} + \left(\mathbf{Z}^T\mathbf{S}^{-1}\mathbf{Z}\right)^{-1}\sigma_\varepsilon^2. \tag{13}$$

An unbiased estimate of G can be obtained by cross-validation. For each of the independent partitions of the data *m*, we estimate the activity patterns on that run, $\widehat{\mathbf{U}}^{(m)}$ and on all other runs, $\widehat{\mathbf{U}}^{(\sim m)}$ separately. Then, we can calculate the average matrix product over these independent partitions of the data

$$\widehat{\mathbf{G}}_{cv} = \frac{1}{M}\sum_m \widehat{\mathbf{U}}^{(m)}\widehat{\mathbf{U}}^{(\sim m)T}/P. \tag{14}$$

Because in each product we multiply noise terms of independent partitions of the data, we obtain an unbiased estimate of the second moment matrix

$$E\left(\widehat{\mathbf{G}}_{cv}\right) = \mathbf{G}. \tag{15}$$

While this estimate of the second moment matrix is unbiased, it is not guaranteed to be positive definite. Thus it is not equivalent to the maximum-likelihood estimate obtained by maximizing Eq. 10. For data visualization and starting values, however, this estimate is extremely useful. Furthermore, distances calculated from this estimate are equivalent to the crossnobis distance estimator, described in detail elsewhere (Diedrichsen and Kriegeskorte, 2017; Walther et al., 2016; Diedrichsen et al., 2016a,b).

2.6. Visualization

After fitting a representational model to the data, it is important to

visualize the result. The first obvious visualization is to plot the values of the model parameters $\boldsymbol{\theta}_m$. How important are different feature components to explain the representation in an area?

The second, more comprehensive approach is to visualize the second-moment matrix itself. Specifically, we want to compare the fitted second moment matrix to a direct estimate of the second moment matrix from the data. For the latter, we can either use the estimate from a free model, or we can use the cross-validated estimate (Eq. 14). A very powerful way to visualize the second moment matrix is to plot the different conditions into a space spanned by the most important eigenvectors of the second moment matrix (for an example, see Fig. 3b, Fig. 5d–i). The distances between different conditions will then show the approximate distances between the patterns. Note that this yields exactly the same visualization as when performing classical multi-dimensional scaling (Borg and Groenen, 2005) on the Euclidean distances between activity patterns (Diedrichsen and Kriegeskorte, 2017).

Finally, we can also visualize the actual estimated activity patterns for each condition across different voxels. This is usually done in encoding approaches, to inspect the spatial distribution of activity profile across the cortical surface. In PCM, the true activation profiles (the random effects) are never explicitly calculated, as they are integrated over when calculating the marginal likelihood (Eq. 7). However, the estimates can be derived as:

$$\widehat{\mathbf{U}} = \mathbf{G}(\boldsymbol{\theta})\mathbf{Z}^T\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{R}(\boldsymbol{\theta})\mathbf{Y}. \tag{16}$$

Similarly, for a feature model, the voxel-feature weights can be calculated as [1]

$$\widehat{\mathbf{W}} = \mathbf{M}(\boldsymbol{\theta})^T\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{R}\mathbf{Y}. \tag{17}$$

2.7. Inference

While visualization is important to further understand the structure of neuronal representations, the main purpose of applying models to data is statistical inference. PCM leaves open a number of possibilities here. How to exactly perform inference on a representational model, especially for groups of participants, is still a subject of debate and development. As we will see, the discussion shares a lot of the problems and issues with model inference for other multivariate fMRI models, such as DCM (Penny et al., 2004).

Principally, there are two ways of performing inferences from the fit of a representational model. First, we can perform inference on the parameter estimates. Alternatively, we can use the marginal likelihood of the data (Eq. 10) as an approximation of the model evidence to compare which of our candidate models provides the most appropriate description of our data.

2.7.1. Inference based on individual parameter estimates

First we may make inferences based on the parameters of a single fitted model. The parameter may be the weight of a specific component or another metric derived from the second moment matrix. For example, the estimated correlation coefficient between condition 1 and 2 would be $r_{1,2} = \mathbf{G}_{1,2}/\sqrt{\mathbf{G}_{1,1}\mathbf{G}_{2,2}}$. We may want to test whether the correlation between the patterns is larger than zero, or whether a parameter differs between two different subject groups, two different regions, or whether they change with experimental treatments.

The simplest way of testing parameters would be to use the point estimates from the model fit from each subject and apply frequentist statistics to test different hypotheses, for example using a t- or F-test. Alternatively, one can obtain estimates of the posterior distribution of the parameters using MCMC sampling (Murphy, 2012) or Laplace approximation (Friston et al., 2007). This allows the application of Bayesian inference, such as the report of credibility intervals.

One important limitation to keep in mind is that parameter estimates from PCM are not unbiased in small samples. This is caused because estimates of $\mathbf{G}$ are constrained to be positive definite. This means that the variance of each feature must be larger or equal to zero. Thus, if we want to determine whether a single activity pattern is different from baseline activity, we cannot simply test our variance estimate (i.e. elements of $\mathbf{G}$) against zero - they trivially will always be larger, even if the true variance is zero. Similarly, another important statistic that measures the pattern separability or classifiability of two activity patterns, is the Euclidean distance, which can be calculated from the second moment matrix as $d = \mathbf{G}_{1,1} + \mathbf{G}_{2,2} - 2\mathbf{G}_{1,2}$. Again, given that our estimate of $\mathbf{G}$ is positive definite, any distance estimate is constrained to be positive. To determine whether two activity patterns are reliably different, we cannot simply test these distances against zero, as the test will be trivially larger than zero. A better solution for inferences from individual parameter estimates is therefore to use a cross-validated estimate of the second moment matrix (Eq. 14) and the associated distances (Walther et al., 2016; Diedrichsen et al., 2016b). In this case the expected value of the distances will be zero, if the true value is zero. As a consequence, variance and distance estimates can become negative. These techniques, however, take us out of the domain of PCM and into the domain of representational similarity analysis (Kriegeskorte et al., 2008; Diedrichsen and Kriegeskorte, 2017).

2.7.2. Inference using model evidence

As an alternative to parameter-based inference, we can fit multiple models and compare them according to their model evidence; the likelihood of the data given the models (integrated over all parameters). In encoding models, the weights $\mathbf{W}$ are directly fitted to the data, and hence it is important to use cross-validation to compare models with different numbers of features. The marginal likelihood (Eq. 10) already integrates all over all likely values of $\mathbf{U}$, and hence $\mathbf{W}$, thereby removing the bulk of free parameters. Thus, in practice the marginal likelihood will be already close to the true model evidence.

Our marginal likelihood (Eq. 10), however, still depends on the free parameters θ. So, when comparing models, we need to still account for the risk of overfitting the model to the data. For fixed models, there are only two free parameters: one relating to the strength of the noise (θ_ε) and one relating to the strength of the signal (θ_s). This compares very favorably to the vast number of free parameters one would have in an encoding model, which is the size of $\mathbf{W}$, the number of features x number of voxels. However, even the fewer model parameters still need to be accounted for. We consider here four ways of doing so.

The first option is to use empirical Bayes or Type-II maximal likelihood. This means that we simply replace the unknown parameters with the point estimates that maximize the marginal likelihood. This is in general a feasible strategy if the number of free parameters is low and all models have the same numbers of free parameters, which is for example the case when we are comparing different fixed models. The two free parameters here determine the signal-to-noise ratio. For models with different numbers of parameters we can penalize the likelihood by $\frac{1}{2}d_\theta \log(n)$, yielding the Bayes information criterion (BIC) as the approximation to model evidence.

As an alternative option, we can use cross-validation within the individual (Fig. 2a) to prevent overfitting for more complex flexible models, as is also currently common practice for encoding models (Naselaris et al., 2011). Taking one imaging run of the data as test set, we can fit the parameters to data from the remaining runs. We then evaluate the likelihood of the left-out run under the distribution specified by the estimated parameters. By using each imaging run as a test set in turn, and

[1] It is important to keep in mind that the second moment matrix (and hence correlations or distances) calculated from the pattern estimates $\hat{\mathbf{U}}$ do not equal those derived directly from our estimate $\mathbf{G}(\boldsymbol{\theta})$. This is because the former do not reflect the uncertainty in the activation estimates, which plays a role in determining the second moment. In short, the second moment of the mean estimates does not equal the mean second moment estimate. Therefore, the activity estimates should only be used to visualize the spatial arrangement of activity patterns, not to make inferences about the representational structure.

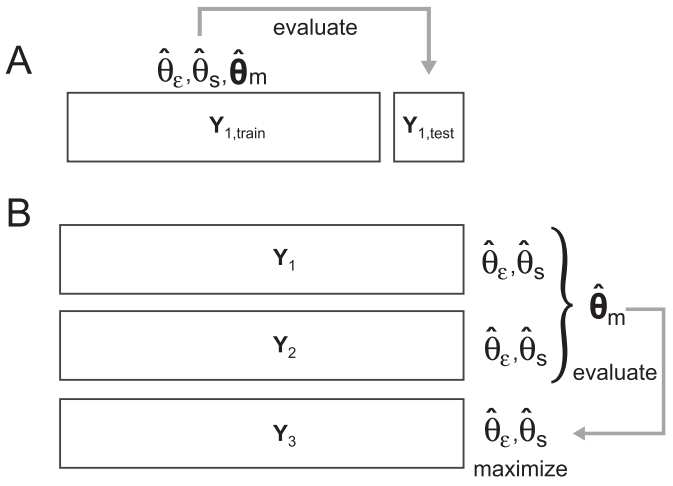

Fig. 2. Cross-validation for testing representational models. (**A**) Within-subject cross-validation. Both the model parameters ($\boldsymbol{\theta}_m$), as well as the noise (θ_ε), and the signal parameter (θ_s) are fitted on the training data set. The likelihood of the test data ($\mathbf{Y}_{1,test}$) is then evaluated under those parameters. (**B**) Group cross-validation, here shown with the data from three subjects ($\mathbf{Y}_{1-3}$). The model parameters are commonly fit to all but one subject, allowing each subject a separate noise and signal parameter. The likelihood of the data from the left-out subject is then evaluated under the group model parameters after optimizing the individual's noise and signal parameters.

adding the log-likelihoods (assuming independence across runs), we thus can obtain an approximation to the model evidence. Note, however, that for a single (fixed) encoding model, cross-validation is not necessary under PCM, as the activation parameters for each voxel (**W** or **U**) are integrated out in the likelihood. Therefore, it can be handled with the first option we described above.

For the third option, if we want to test the hypothesis that the representational structure in the same region is similar across subjects, we can perform cross-validation across participants (Fig. 2b). We can estimate the parameters that determine the representational structure using the data from all but one participant and then evaluate the likelihood of data from the left-out subject under this distribution. When performing cross-validation within individuals, a flexible model can fit the representational structure of individual subjects in different ways, making the results hard to interpret. When using the group cross-validation strategy, the model can only fit a structure that is common across participants. Different from encoding models, representational models can be generalized across participants, as we do not fit the actual activity patterns, but rather the representational structure. In a sense, this method is performing "hyper alignment" (Guntupalli et al., 2016) without explicitly calculating the exact mapping into voxel space (Eq. 17). When using this approach, we still allow each participant to have its own signal and noise parameters, because the signal-to-noise ratio is idiosyncratic to each participant's data. When evaluating the likelihood of left-out data under the estimated model parameters, we therefore plug in the ML-estimates for these two parameters for each subject.

Finally, a last option is to implement a full Bayesian approach and to impose priors on all parameters, and then use a Laplace approximation to estimate the model evidence (Kass and Raftery, 1995; Friston et al., 2007). While it certainly can be argued that this is the most elegant approach, we find that cross-validation at the level of model parameters provides us with a practical, straightforward, and transparent way of achieving a good approximation.

Each of the inference strategies supplies us with an estimate of the model evidence. To compare models, we then calculate the log Bayes factor, which is the difference between the log model evidences.

$$\log B_{12} = \log \frac{p(\mathbf{Y}|M_1)}{p(\mathbf{Y}|M_2)} = \log p(\mathbf{Y}|M_1) - \log p(\mathbf{Y}|M_2) \quad (18)$$

Log Bayes factors of over 1 are usually considered positive evidence and above 3 strong evidence for one model over the other (Kass and Raftery, 1995).

2.7.3. Group inference

How to perform group inference in the context of Bayesian model comparison is a topic of ongoing debate in the context of neuroimaging. A simple approach is to assume that the data of each subject is independent (a very reasonable assumption) and that the true model is the same for each subject (a maybe less reasonable assumption). This motivates the use of log Group Bayes Factors (GBF), which is simple sum of all individual log Bayes factor across all subjects n

$$\log GBF = \sum_n logB_n. \quad (19)$$

Performing inference on the GBF is basically equivalent to a fixed-effects analysis in neuroimaging, in which we combine all time series across subjects into a single data set, assuming they all were generated by the same underlying model. A large GBF therefore could be potentially driven by one or few outliers. We believe that the GBF therefore does not provide a desirable way of inferring on representational models - even though it has been widely used in the comparison of DCM models (Friston et al., 2003).

At least the distribution of individual log Bayes factors should be reported for each model. When evaluating model evidences against a Bayesian criterion, it can be useful to use the average log Bayes factor, rather than the sum. This stricter criterion is independent of sample size, and therefore provides a useful estimate or effect size. It expresses how much the favored model is expected to perform better on a new, unseen subject. We can also use the individual log Bayes factors as independent observations that are then submitted to a frequentist test, using either a t-, F-, or nonparametric test. This provides a simple, practical approach that we will use in our examples here. Note, however, that in the context of group cross-validation, the log-Bayes factors across participants are not strictly independent.

Finally, it is also possible to build a full Bayesian model on the group level, assuming that the winning model is different for each subject and comes from a multinomial distribution with unknown parameters (Stephan et al., 2009).

2.8. Noise ceilings

Showing that a model provides a better explanation of the data as compared to a simpler Null-model is an important step. Equally important, however, is to determine how much of the data the model does not explain. Noise ceilings (Nili et al., 2014) provide us with an estimate of how much systematic structure (either within or across participants) is present in the data, and what proportion is truly random. In the context of PCM, this can be achieved by fitting a fully flexible model, i.e. a free model in which the second moment matrix can take any form. The non-cross-validated fit of this model provides an absolute upper bound - no simpler model will achieve a higher average likelihood. As this estimate is clearly inflated (as it does not account for the parameter fit) we can also evaluate the free model using cross-validation. Importantly, we need to employ the same cross-validation strategy (within/between subjects) as used with the models of interest. If the free model performs better than our model of interest even when cross-validated, then we know that there are definitely aspects of the representational structure that the model did not capture. If the free model performs worse, it is overfitting the data, and our currently best model provides a more concise description of the data. In this sense, the performance of the free model in the cross-validated setting provides a "lower bound" to the noise ceiling. It still may be the case that there is a better model that will beat the currently best model, but at least the current model already provides an adequate description of the data. Because they are so useful,

noise ceilings should become a standard reporting requirement when fitting representational models to fMRI data, as they are in other fields of neuroscientific inquiry already. The Null-model and the upper noise ceiling also allow us to normalize the log model evidence to be between 0 (Null-model) and 1 (noise ceiling), effectively obtaining a Pseudo-R^2.

3. Examples and results

In this section we provide examples of how to build and evaluate representational models from our recent studies on movement representations. We did not always use PCM in the original papers - partly because some of the methods were not fully developed at this point - partly because other strategies of getting similar results (i.e. using RSA with cross-validated distances) were at this point more common and easier to communicate. We show here how to perform these analyses using PCM in a concise manner - often allowing for more powerful inferences.

3.1. Fixed and component models: representational structure of finger movements

In a recent paper (Ejaz et al., 2015), we studied the activity patterns associated with single finger movements in primary sensory and motor cortices. The actual patterns of neural activity were quite variable across different participants. However, when we studied the second moment matrix of the activity profile distribution (Fig. 3a), we found that they were highly correlated across different individuals. As can be seen from a multi-dimensional scaling of the activity patterns (Fig. 3b), the thumb (d1) showed the most unique pattern, while the patterns of other fingers (d2-d5) were arranged according to their neighborhood relationship. Interestingly, the pattern for the little finger (d5) was already more similar to the thumb pattern than the middle or ring finger - thus the structure was not well explained by a simple somatotopic ordering of the fingers on the cortical sheet.

What can explain this highly invariant representation? We tested a number of hypotheses. First, we considered the possibility that two cortical activity patterns could also be more similar, because the two movements involved similar sets of muscles. We therefore measured the muscle activity involved in making the single finger movements and predicted the second-moment matrix of the patterns directly from the correlation matrix of the measured muscle patterns (Fig. 3c). Alternatively, we measured the movements of the fingers during normal everyday activities (Ingram et al., 2008), hypothesizing that two fingers that commonly move together would also exhibit similar activity pattern. Thus, we predicted the second-moment matrix of the patterns directly from the correlation matrix of the natural statistics of finger movements (Fig. 3d).

In the original paper (Ejaz et al., 2015), we used RSA to compare the models. However, we have recently shown that the same model comparison can be performed in a more powerful fashion through PCM (Diedrichsen and Kriegeskorte, 2017). The two models give us a classic example of "fixed" models, in which we fully predict the distribution of activity profiles (up to a signal scaling constant) using external data. Per subject, we therefore only estimated the signal and noise parameter to optimize the likelihood (Eq. 10). Because both models had only those two free parameters, we used Type-II maximum likelihood (empirical Bayes) as an estimate of model evidence.

To scale the model evidence we also fit a null model. In this case the null model predicted that the second moment matrix would be identity - that is, all finger patterns are equally far apart from each other. We then expressed each subject's model evidence relative to the null model by taking the difference between log model evidences (see section 2.7.2). Note that we also could have used the null model that there are no differences between activity patterns. This would have given us much larger log-Bayes factors against the null model, but would have left the differences between models unchanged.

We also calculated a noise ceiling by fitting a free model of the second moment matrix either to all the subjects (upper bound) or to all participants, excluding the one that we evaluated the fit on (lower bound, see section 2.8).

We can then display the model evidence in the form of a "Pseudo-R2", with zero referring to the null model and 1 to the upper noise ceiling. As

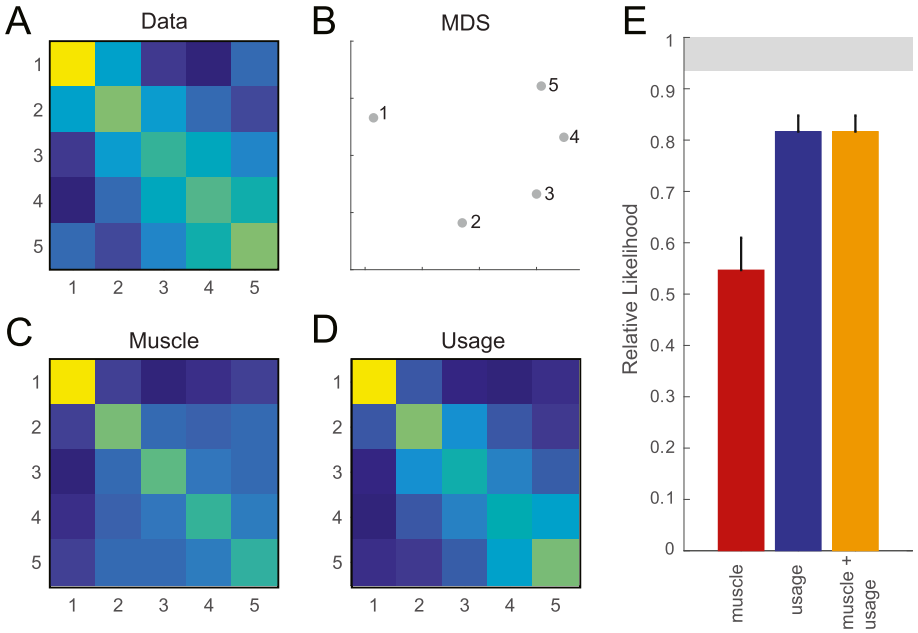

Fig. 3. Comparison of representational models of the structure of finger representations using PCM. (**A**) A cross-validated estimate of the second moment of the activity profile distribution. The rows and columns of the matrix refer to digit 1–5. (**B**) Classical multi-dimensional scaling of the representational structure. Shown are the finger patterns projected on the eigenvectors of the second moment matrix (**C**) Predicted second-moment matrix under the muscle model. (**D**) Predicted second-moment matrix under the natural statistics of hand movements. (**E**) Pseudo-R2 based on the log-Bayes factor for the models against a null model. The gray bar indicates the area between the lower and upper noise ceiling.

can be seen from Fig. 3e, the usage model was consistently associated with a larger evidence than the muscle model. This was the case in all of the seven participants. The average log-Bayes factor for a single individual was 81.56 (SE = 24.801), suggesting overwhelming evidence for the usage over the muscle model. The usage model did not, however, reach the lower noise ceiling, indicating that there is still systematic structure in the representations that is unexplained by the model.

We then explored whether a combination of the two models would explain the data better than the hand usage model alone. Thus, we built a component model, where the two component matrices (Eq. 4) corresponded to predicted second moment matrices from the muscle and usage model, respectively. Such a model tests the idea that some voxels in the population may be better described as tuned to muscle activations independent of natural usage, whereas others reflect the natural statistics, with the exact proportions of these populations unknown. While we here only combine 2 components, in ongoing work we are using many more components to describe the representation of movement sequences.

Adding more possible components will of course lead to a better fit to the data. Thus, to evaluate the model evidence, we are using here cross-validation across subjects (Fig. 2b), estimating the mixture proportions on data from all but one subject and then evaluating the fit on the left-out subject (only allowing free scaling and noise parameters, as for the other models). In this particular case, the combination model did not perform better than the usage model alone. The weight of the muscle model was usually very low, and the log Bayes factor between the two models was ever so slightly in favour of the simpler usage model, indicating that the combination model overfit the data.

3.2. Feature models: correlation between ipsilateral and contralateral finger presentations

A common application of PCM (and the one that it was originally developed for) is to estimate the true correspondence between two activity patterns. For example in M1, both movements of the contralateral and ipsilateral fingers evoke activity patterns that allow decoding of which finger moved (Diedrichsen et al., 2013). The patterns evoked by ipsilateral movement are weaker than those elicited by contralateral movement, but they appear to match on a finger-by-finger basis.

While it is easy to establish that the patterns are more correlated than expected by chance ($r > 0$), the true extent of the correspondence is notoriously hard to estimate. This is because correlation estimates are biased by noise. From a neuroscientific perspective, however, it is important to know if the true correlation is $r = 1$, as this would indicate

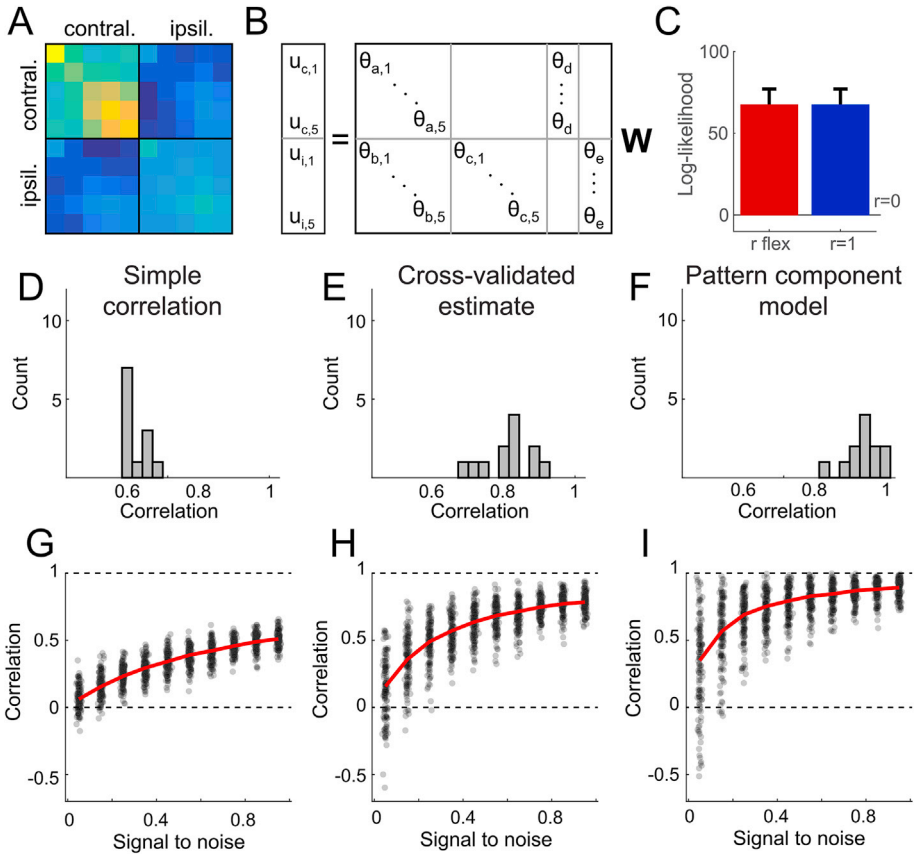

Fig. 4. Determining the true correlation between activity patterns. (**A**) Cross-validated estimate of the covariance matrix of fingers of the contra- and ipsilateral hands. The upper right 5 × 5 block of the matrix indicates the co-variances of finger patterns across hands. The fact that the diagonal of this block is higher than the off-diagonal, indicates a positive correlation of matching finger pairs. (**B**) Feature model, expressing the patterns of the contralateral fingers ($u_{c,1} - u_{c,5}$) and ipsilateral fingers ($u_{i,1} - u_{i,5}$) as a combination of the shared finger pattern (weighted by $\theta_{a,i}$ and $\theta_{b,i}$, respectively) and one that is idiosyncratic to the ipsilateral hand, weighted by $\theta_{c,i}$. Empty parts of the matrix are zero. To complete the model, we also added a hand-specific pattern for the contra and ipsilateral hand, weighted by θ_d and θ_e respectively. (**C**) Marginal log-likelihood of the flexible model (red) and the model constraining the correlation to one (blue, $\theta_{c,i} = 0$), relative to the log-likelihood of the zero-correlation model. (**D**) Distribution of simple correlation coefficients for the 12 hemispheres. (**E**) Distribution of the correlation calculated from a cross-validated estimate of the covariance matrix. (**F**) Distribution of correlation coefficients derived from the PCM model. (**G**) Simulation assuming that the contra- and ipsilateral patterns are perfectly correlated (r = 1) with unequal signal strength forms hands and fingers are uneven. Correlation estimates approach 1 as signal-to-noise ratio (x-axis) increases. (**H**) The same simulation for finger-specific feature model and (**I** for cross-validated estimates. The dots reflect the estimated correlation from single simulations. Red lines indicate the average.

that the ipsilateral movements reactivate the exact patterns of the contralateral movements. In contrast, a slightly lower correlation, i.e. $r = 0.8$ would provide evidence for that there exists a neuronal population that uniquely encodes ipsilateral finger movements. This in turn would suggest that both contra- and ipsilateral M1 have a function in controlling fine finger movements.

Naively, we could just estimate the correlation between finger patterns from the covariance between the pattern for the contra (c) and ipsilateral (i) fingers

$$r = \frac{cov(u_c, u_i)}{\sqrt{var(u_c)var(u_i)}}. \tag{20}$$

For such a correlation to be informative, we first need to subtract the mean activity patterns for all fingers of each hand, to ensure that the correlations are not simply driven by a correlation between the overall activity patterns. Using this simple covariance estimate (Eq. 11), we obtain an average correlation between finger pairs of $r = 0.28(+0.02)$ (Fig. 4d), which is significantly larger than zero. However, this estimate is severely biased by noise. To show this, we simulated data using a true correlation of 1 between control and ipsilateral patterns. To make the simulation realistic, we used different signal-to-noise levels for the contralateral fingers (1, 0.5, 0.1), and strongly reduced signal to noise (0.2) for the ipsilateral hand. Even with these relatively low levels of measurement noise, the simple correlation estimates are strongly biased towards 0 (Fig. 4g).

To obtain a better estimate, we could use a cross-validated estimate of the covariance matrix (Eq. 14). This estimate of the covariance matrix is unbiased, but is not guaranteed to be positive definite (see section 2.5). When calculating correlations based on such covariance estimates, one can encounter a number of problems. If one of the variance estimates (diagonal entries in the matrix) is negative, the correlation is undefined. Even if the variance estimates are forced to be positive, the denominator of Eq. 20 may become very small, leading to correlations far outside the interval of $[-1; 1]$. To make the correlation estimate more stable, we can force the matrix to be semi-positive definite by removing eigenvectors with negative eigenvalues. Using this technique, we obtain a higher estimates of the correlation (Fig. 4e). Simulations, however, demonstrate that this correlation is still strongly biased (Fig. 4h).

Using PCM, we can model a finger-by-finger correlation using a feature model with three parameters per finger (Fig. 4b). The patterns for the contralateral fingers are modeled by a finger-specific pattern $\mathbf{w}_{1-5}$ with variance of 1, multiplied by a finger-specific parameter $\theta_{a,i}$. Thus,

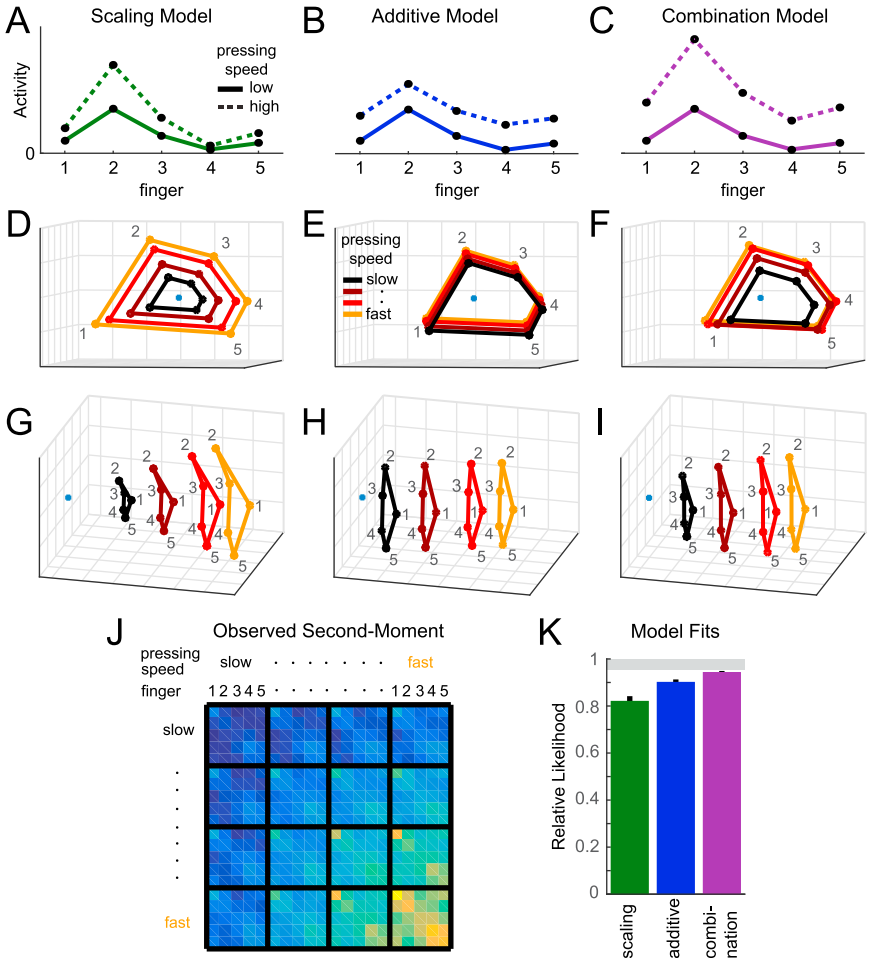

Fig. 5. Scaling of finger activity patterns with movement frequency. (**A–C**) Example voxel tuning curves for the 5 individuated finger movements under the three models: Scaling, Additive, and Combination models. For a voxel with the same baseline tuning (dashed line), each model predicts a different change in tuning as finger movement frequency increases (solid line). (**D–I**) Assuming the same behavior across all voxels, the scaling model (**D, G**), the additive model (**E, H**), and the combination model (**F, I**) predict the a particular arrangement of the patterns in representational space. The space is defined as the first three principal components that best capture the overall representational structure. Conditions with the same movement frequency are connected by lines. The blue crosses indicate resting baseline. Model predictions are made for the set of best-fitting parameters from the group fit. (**J**) The overall estimate of the second moment matrix. (**K**) We then assessed the likelihood of the data under the three models. The combination model yielded the greatest likelihood of the three models, and performed nearly as well as the unconstrained, free model. Plotted here is the cross-validated marginal log-likelihood, standardized by the upper noise ceiling.

the predicted variance (Fig. 4c) of each finger pattern is $\theta^2_{a,i}$. The patterns for the ipsilateral fingers are modeled as the sum of the contralateral pattern, weighted by $\theta_{b,i}$, and an independent ipsilateral finger-specific pattern ($\mathbf{w}_{6-10}$), weighted by $\theta_{c,i}$. Accordingly, the variance of the ipsilateral patterns will be $\theta^2_b + \theta^2_c$ and the covariance $\theta_a\theta_b$. If $\theta_{c,i} = 0$, then the correlation is 1 and the ipsilateral patterns are only a scaled version of the contralateral patterns. If $\theta_{b,i} = 0$, then the correlation is zero. The estimated model correlations are higher than the ones obtained with cross-validation (Fig. 4f), but are still clearly biased (Fig. 4i). Note that this is equally the case when using an unconstrained covariance model, as has been suggested by Cai et al. (2016).

PCM, however, offers the possibility to test whether contra- and ipsilateral patterns correspond perfectly ($r = 1$), or whether there is only partial correspondence ($r < 1$) by performing formal model comparison A model for which all $\theta_{c,i}$ are forced to be zero restricts the correlation to be $r = 1$. We can compare this to a model in which the correlation is free to vary. To account for the added parameters in the second model, we again can perform cross-validation across participants. As a baseline we also consider the null model with $\theta_{b,i} = 0$, i.e. $r = 0$.

As can be seen from the model evidence, there is overwhelming evidence for both main models over the null-model, with an average log BF of 67.4 (SD = 33.6), confirming the result already obtained using simple correlations. Using PCM can now compare the $r = 1$ model to the $r < 1$ model. While latter model has a slight advantage over the more constrained model, the average log-BF was close to zero. Thus, our data provides no notable evidence that the correlation is smaller than 1. We can therefore conclude that in M1 proper, ipsilateral finger movements appear to reactivate the patterns associated with the corresponding contralateral finger. Using fMRI at the current resolution (2.3 *mm*), we could not detect a unique population code for the ipsilateral movement. Importantly this conclusion is only made possible using model comparison with PCM - no parameter-based approach that we know of would enable such a conclusion.

3.3. Nonlinear model: scaling of finger patterns with movement speed

PCM is also helpful in addressing another important question that is often asked in the context of fMRI analysis: How do tuning curves of individual voxels change when a modulatory factor (attention, visual contrast, movement speed, amount of training, etc.) is altered? Often, the average regional activity changes substantially over different levels of these factors. But what happens to the activity patterns/activity profiles themselves?

Here we answer this question in the context of the tuning of voxels for individual finger movements in M1, and how these tuning curves change with increasing movement frequency. It is known that overall BOLD signal increases non-linearly with movement frequency, and that these non-linearities are at least to a large part caused by neuronal factors (Hermes et al., 2012; Siero et al., 2013), namely that the activity elicited by a finger press declines the shorter the time since the last press was. However, we currently do not know how individual tuning functions of the voxels change with increasing movement frequency. We therefore studied the BOLD signal patterns in M1 during tapping of the 5 fingers of the right hand at 2, 4, 8 or 16 presses over 6s.

We then defined four models of how voxel tuning curves to individual finger movements change with increasing movement frequency. The first possibility is that the voxel activities scale with the same (arbitrary) factor for all fingers (Fig. 5a). This would mean that the entire pattern of BOLD activity for each finger remains structurally the same, simply being scaled by a single common factor for each increase in movement frequency. Using PCM, we can formalize this "scaling model" by

$$\mathbf{y}_{i,j} = \theta^s_j \mathbf{f}_i + \boldsymbol{\varepsilon} \tag{21}$$

$$\mathbf{f}_i \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_f\right) \tag{22}$$

$$\boldsymbol{\Sigma}_f = \mathbf{A}\left(\boldsymbol{\theta}^f\right)\mathbf{A}\left(\boldsymbol{\theta}^f\right)^T, \tag{23}$$

where the activity pattern $\mathbf{y}$ for finger i at movement frequency j is a product of the speed scaling parameter θ^s_j and the finger pattern $\mathbf{f}_i$, plus Gaussian noise ε. The finger patterns are distributed according to an arbitrary second moment matrix $\boldsymbol{\Sigma}_f$, which is fit the same way as a free model, i.e. as the square of an upper-triangular matrix $\mathbf{A}$. Thus, the parameters determining the distribution of finger tuning functions ($\boldsymbol{\theta}^f$) interact multiplicatively with the parameters that determine the common scaling of activity with frequency ($\boldsymbol{\theta}^s$). This makes the model inherently non-linear. When plotting this prediction in representational space (Fig. 5d and e), one can see that the distances between finger patterns scale by a certain amount with each increase in movement frequency, but this scaling remains linear from the rest activity (cross).

Secondly, the increase of activity with movement frequency could be related to an independent additive background pattern (Fig. 5b). The presence of such a component could indicate that there is some non-finger specific input to M1 increases with increasing movement frequency (e.g. attentional modulation, etc). Alternatively, such an additive pattern could be caused by a non-specific spreading of the BOLD signal with increasing neural activity. Under this "additive model", the finger patterns are modeled as

$$\mathbf{Y}_{i,j} = \mathbf{f}_i + \theta^a_j \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{24}$$

where the activity pattern $\mathbf{y}$ of finger i at movement frequency j is the sum of the finger pattern $\mathbf{f}$ for finger i, an additive background pattern α that is scaled by parameter θ_a, and Gaussian noise ε. Here we defined the background pattern to be the mean activity pattern across fingers at the same movement frequency. Under this model, the distances between finger patterns in representational space would remain the same across movement frequencies. Importantly, the additive pattern simply shifts the finger patterns away from baseline in this representational space-voxels become more tuned to all finger presses in a non-specific fashion (Fig. 5e and H).

Of course it is also possible to combine the above two models (Fig. 5c). In this case the activity of each voxel for each finger is scaled up (such as in the scaling model) and the activity for each voxel would increase by a set amount for all fingers (as in the additive model):

$$\mathbf{Y}_{i,j} = \theta^s_j \mathbf{f}_i + \theta^a_j \boldsymbol{\alpha} + \boldsymbol{\varepsilon}. \tag{25}$$

The predicted representational structure is then a combination of the additive and scaling effects (Fig. 5f and i). Importantly, this model makes no constraints on the contributions of the scaling and additive components for each movement frequency. Therefore, it is for example possible for pattern changes in low movement frequency to be driven by scaling, whereas changes for high frequencies being better described by an additive component.

Finally, the representational structure could also change in more complex ways, deforming with increasing movement frequency. We would expect such behavior if different voxels showed different points of BOLD saturation or other non-linear effects distortion. Such complex behavior would make the interpretation the representational structure as measured with fMRI difficult, as it would clearly indicate that it can change with the overall amount of activity in a region. For this possibility we considered a free, unconstrained model, which also serves as the noise-ceiling. Thus, this model can capture any systematic change of population tuning with movement frequency.

With PCM, we are effectively assessing the likelihood of the

collection of all measured voxel tuning curves in M1 under each of these models. As before, we are using between-subject cross-validation to compare models with different numbers of parameters. The marginal likelihoods (Fig. 5k) show that the additive model is better than the scaling model, but that both models are clearly outperformed by the noise-ceiling model. The comparison to the free (noise-ceiling) model indicates that the representational structure is nearly completely modeled using the combination of scaling and additive components. Although there is a significant difference between the combination and the noise ceiling model ($t = 3.31$, $p < 0.01$), the average Bayes factor separating these models was very small (0.014). This result indicates that despite some systematic distortion, the changes in the population of voxel tuning curves in M1 are well described by a common overall scaling factor plus a finger-independent background pattern. Whether this additive pattern reflects increased general neuronal processing, or whether it is caused by an unspecific spatial spread of the BOLD signal is currently not clear.

This example should hopefully demonstrate the power of PCM to test well-specified and complex model of population tuning. Such models can help us understand more clearly how attention or repetition influences individual voxel tuning. In the context of repetition suppression, fatigue, sharpening, and facilitation models have been discussed (Grill-Spector et al., 2006), which are structurally very similar to the scaling and additive models presented here. PCM allows for a robust, practical and powerful way of testing such models on fMRI data.

3.4. Sensitivity to the Gaussian-Gaussian assumption

After three example that demonstrate the power and flexibility of PCM, we turn to an important additional question that has not be addressed in previous papers (Diedrichsen and Kriegeskorte, 2017; Diedrichsen et al., 2011): What happens to PCM inferences if the assumption that both the signal and the noise come from a Gaussian distribution is violated? Typically, both raw fMRI data and activity estimates from a first-level GLM show slightly heavier tails (more outlying values) than predicted by the Gaussian distribution. To test how this influences model selection performance of PCM, we simulated data either according to the muscle or the usage model as described in section 3.1. To vary the non-Gaussianity of the data we drew the signal or noise term not from a Gaussian distribution, but from a Student t-distribution with degree of freedom (df) between 3 (very heavy tails) and 100 (practically Gaussian). In all simulations we scaled the data, such that the second moment of the signal and noise distributions remained constant even when the distributional shape changed. Inspection of the empirical data underlying example 1 showed that a t-distribution with 10–20 df was appropriate range to describe the positive kurtosis (heavier tails) of the activity estimates that form the input data to PCM.

We then used either PCM, voxel-wise encoding models, or 3 different variants of RSA to compare which of the two models provided the better explanation for the data. For each approach, the proportion of correct model decisions was recorded. The implementation details of the different methods, and the results for Gaussian data can be found in Diedrichsen and Kriegeskorte (2017). For nearly Gaussian data ($df = 100$) PCM was the method with the largest proportion of correct model choices (Fig. 6). This is expected, as - for Gaussian data - PCM implements the likelihood-ratio test between models, which by the Newman-Pearson Lemma constitutes a theoretical upper bound of the best achievable model selection accuracy (if the two models are equally likely a-priori). However, even when we increase the non-Gaussianity of the noise (Fig. 6a), or the signal (Fig. 6b), PCM robustly remained the best method.

This result may seem counterintuitive at first, as PCM is the only method that explicitly assumes a Gaussian distribution of the data. However, note that this assumption merely leads PCM to focus on the second moment as a sufficient statistic of the data - a characteristic that it shares with all other methods tested here (Diedrichsen and Kriegeskorte, 2017). While our simulations changes higher-order moments of the noise and signal distribution, the second moment was kept constant. Therefore, the inference using all methods was basically blind to these changes. In summary, our simulations show that PCM is not only a very powerful method when conditions are met, but that it still can provide exact (and, compared to other competing methods, superior) model inferences on more realistic data distributions.

4. Discussion

In this paper we have presented a statistical approach for fitting and comparing representational models. PCM is conceptually very closely linked to encoding models, as well as to RSA (Diedrichsen and Kriegeskorte, 2017). Because it implements a likelihood-ratio test between models (Neyman and Pearson, 1933), it can be shown to be statistically more powerful than either competing approach (Diedrichsen and Kriegeskorte, 2017). More importantly, the analytical tractability of PCM makes it especially attractive for the fitting and comparison of flexible

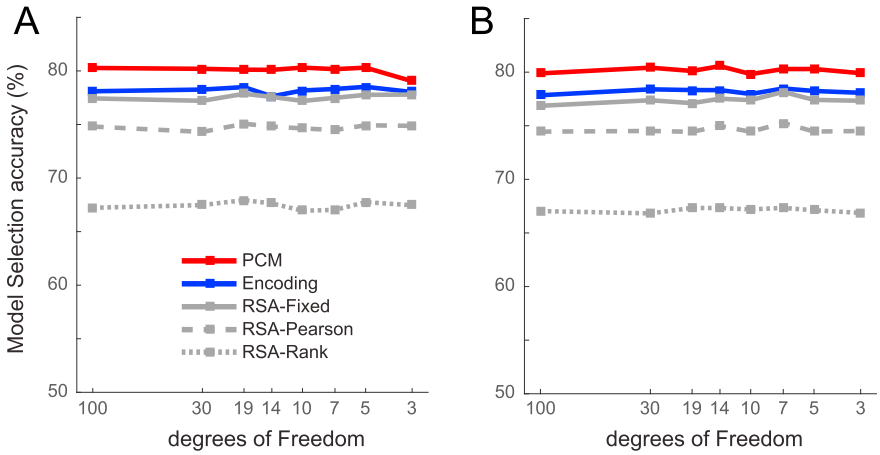

Fig. 6. Robustness of model selection accuracy of PCM, Encoding analysis and RSA for violations of Gaussianity. (**A**) Model selection accuracy for simulated data coming from the muscle and usage model of example 1. The noise was drawn from a student t-distribution with $df = 3 - 100$. RSA models were evaluated using a correlation without intercept, Pearson correlation or Spearman rank-correlation (Diedrichsen and Kriegeskorte, 2017). (**B**) Model selection accuracy for simulated data where the signal was drawn from a t-distribution.

representational models.

In encoding models, a separate weight of each voxel and feature is explicitly fitted. To evaluate model fit, the prediction accuracy for left-out data is used. In contrast, PCM evaluates the marginal likelihood of the data under the distribution predicted by the model. In this expression, the unknown activity profiles for each voxel are integrated out (Eq. 7). This means that we can evaluate different encoding models, even with different numbers of features, by simply comparing their marginal likelihoods. The step of integrating over the actual activity profiles is justified under the assumption that the exact spatial arrangement of the different activity profiles does not matter for the representational content of a brain area - even if the neurons were differently arranged, a fully connected read-out neuron could extract the same information. Indeed, it can be shown that the spatial arrangement of patterns associated with finger movements is variable in primary motor cortex, whereas the representational structure - and hence likely the function - is remarkably invariant across participants (Ejaz et al., 2015).

PCM uses the second moment as a summary statistics for observed data. Analogously, it also uses the second moment to specify the representational model. It therefore constitutes an abstraction from the use of concrete feature sets. Using features to describe a population of activity profiles has a long tradition in neuroscience. In motor control, the response properties of single neurons in primary motor cortex have often been characterized as being tuned to movement direction (Georgopoulos et al., 1986). However, many studies have made it clear that representations in M1 contain a complex mixture of features, including position (Sergio and Kalaska, 1997), direction (Schwartz et al., 1988), force (Sergio et al., 2005), or muscle activity (Kakei et al., 1999), often with non-linear interactions between them (Sergio and Kalaska, 1997). Indeed, it has been questioned whether a description of population activity in terms of semantically definable features is at all sensible (Shenoy et al., 2013).

This is also one of the key lessons from the emerging field of deep learning algorithms as a model of neural processing (Marblestone et al., 2016). Most often, the learned features found in the hidden layers defy simple semantic description. This means that we need tools to formulate and test hypotheses about neural representations that describe the representational space itself, without getting distracted by the relatively superficial issue of which axes should be used to describe the space. Indeed, the representational structure found in different layers of a deep neural networks can be used as a model for representational structures found in the human brain (Khaligh-Razavi and Kriegeskorte, 2014). Even if it turns out that real neural representations are distinctively different from the activation states found in artificial neural networks, deep nets form an ideal testing ground for the methods we employ to understand biological systems. If our methods cannot provide insight into the transformations occurring between the layers of these still relatively simple networks, they most likely will tell us very little about what happens in the brain (Jonas and Kording, 2017).

PCM has a very tight relationship to RSA, which also uses a summary statistics to compare data and model predictions. In the case of RSA, the summary statistics are the distances between all pairs of neural activity patterns. These distances usually contain the same information as the second moment matrix. However, to approach the statistical power in disambiguating models that is inherent in PCM, one needs to takes into account the co-dependences between these distance estimates (Diedrichsen and Kriegeskorte, 2017).

The key advantage of PCM, however, is having an analytical expression for the marginal likelihood. This frees the user from having to perform fitting and cross-validation of voxel-feature weights at every step, and makes it easier to explore a larger space representational models. Note that an encoding model with a specific feature set (despite all the free voxel-feature weight parameters) is a fixed PCM model, in which the only signal and noise parameter (corresponding to the ridge coefficient) is free to vary. More complex models, in which the second moment matrix has some degree of flexibility, would refer to extension, differential scaling, combination or nonlinear changes of the model features. Having an analytical expression for the marginal likelihood allows us to directly estimate these parameters. For model comparison, we can use simple cross-validation (rather than the double cross-validation necessary in encoding approaches). It is also possible to avoid cross-validation completely by using a fully Bayesian approach, in which a normal prior is imposed on the parameters θ. Here we chose cross-validation as a practical and robust approach for estimating the model evidence.

Despite these practical advantages, there are a number of drawbacks and limitations to our approach. The reader should be cautioned that, although PCM attempts to separate the contribution of signal and noise to the second moment of the data, the estimate of the true second moment matrix is not unbiased when using small samples. This is due to the fact that the estimate of **G** is constrained to be positive definite. Hence, when the true signal variance is zero, the variance estimates will be zero or positive, causing the mean estimates to be larger than zero. For unbiased results, one would need to use cross-validation (Eq. 14), which sometimes also results in negative variance estimates. Note, however, that although the cross-validated second moment matrix estimate is unbiased, correlation coefficients derived from this estimate are not. In general, the restriction to positive definite estimates prevents us from being able to determine, based on the second moment matrix alone, whether a certain feature is encoded above chance. For this, we would need to compare a representational model with and without that feature encoded (see section 3.1).

Another limitation of PCM arises from the fact that it uses only the second moment of the data as a sufficient statistic. In this, PCM ignores certain aspects of the neural code. First, it neglects the spatial arrangement of the different activity profiles on the cortical sheet. One of the coding principles in the neocortex is that nearby locations also have similar activity profiles (Graziano and Aflalo, 2007). Therefore, the exact distribution of activity profiles across the cortical surface may contain some important insights into information coding. On the other hand, there is evidence that the exact arrangement of activity patterns on the cortical surface may reflect mostly random biological variation that in itself does not speak to the computational function of the region (Ejaz et al., 2015).

Similarly, PCM ignores all higher statistical moments of the activity profile distribution. In primary visual cortex, for example, many voxels are tuned to one specific stimulus locations in the visual field, but very few respond to two separate locations (Dumoulin and Wandell, 2008). Thus, when we use stimulus location as our axes, the activity profiles will cluster around the axes, and their distribution will be distinctly non-Gaussian. This does not mean that assumption of Gaussian distributed activity profiles (as made in PCM) would automatically lead to invalid inference (see also section 3.4). Independent of the shape of the distribution, the second moment of the activity profiles determines which stimulus features can be read out from the neuronal code by a fully connected neuron. The question of whether we can find evidence for clear non-Gaussian distributions of activity profiles, and what these may imply for the neural representations in these regions, is in highly interesting topic that warrants further study (Norman-Haignere et al., 2013).

Funding

This work was supported by a Scholar award of the James McDonnell foundation, and an NSERC discovery grant, RGPIN-2016-04890, both to JD, a JSPS Postdoctoral Fellowship (#15J03233) to AY.

Appendix A. Notation

Table 1
For non-scalars, the second column indicates the vector / matrix size.

Symbol	Size	Meaning
K		Number of conditions
P		Number of measurement channels (voxels, electrodes, neurons)
N		Overall number of measurements
Q		Number of features in model
M		Number of independent partitions of the data
$\mathbf{B}$	$J \times P$	Regression coefficients for the J regressors of no-interest
$\mathbf{G}$	$K \times K$	Second moment of $\mathbf{u}_p$
$\mathbf{M}$	$K \times Q$	Matrix of model features for all condition
$\mathbf{R}$	$N \times N$	Residual forming matrix for removal of fixed effects
$\mathbf{S}$	$N \times N$	Temporal variance-covariance structure of the noise
$\mathbf{U}$	$K \times P$	Matrix of true activation patterns
$\mathbf{u}_p$	$K \times 1$	True activity profile for voxel p
$\widehat{\mathbf{U}}^{(m)}$	$K \times P$	Matrix of estimated activity patterns, based on data from partition m
$\widehat{\mathbf{U}}^{(\sim m)}$	$K \times P$	Matrix of estimated for activity patterns, based on data independent of m
$\mathbf{V}(\boldsymbol{\theta})$	$N \times N$	Predicted variance-covariance matrix of the data
$\mathbf{W}$	$Q \times P$	Matrix of all voxel-feature weights
$\mathbf{w}_p$	$Q \times 1$	Vector of voxel-feature weights for voxel p
$\mathbf{X}$	$N \times J$	Design matrix containing J regressors of no-interest
$\mathbf{Y}$	$N \times P$	Matrix of brain measurements, concatenated activity estimates or time series data
$\mathbf{Z}$	$N \times K$	Design matrix, indicating how measurements relate to activity patterns
θ_ε		Second-level parameter for noise variance
θ_h		Second-level parameter for signal strength of component h
$\boldsymbol{\theta}_m$		Vector of all second-level parameters that determine the structure of $\mathbf{G}$
$\boldsymbol{\theta}$		Vector of all second-level parameters

References

Allefeld, C., Haynes, J.D., 2014. Searchlight-based multi-voxel pattern analysis of fmri by cross-validated manova. Neuroimage 89, 345–357. http://www.ncbi.nlm.nih.gov/pubmed/24296330.

Borg, I., Groenen, P., 2005. Modern Multidimensional Scaling: Theory and Applications, 2nd Edition. Springer-Verlag, New York.

Cai, M.B., Schuck, N.W., Pillow, J., Niv, Y., 2016. A Bayesian Method for Reducing Bias in Neural Representational Similarity Analysis, pp. 4952–4960.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. J. Neurosci. 37 (27), 6539–6557. https://www.ncbi.nlm.nih.gov/pubmed/28588065.

deCharms, R.C., Zador, A., 2000. Neural representation and the cortical code. Annu Rev. Neurosci. 23, 613–647. https://www.ncbi.nlm.nih.gov/pubmed/10845077.

DiCarlo, J.J., Cox, D.D., 2007. Untangling invariant object recognition. Trends Cogn. Sci. 11 (8), 333–341. https://www.ncbi.nlm.nih.gov/pubmed/17631409.

DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How does the brain solve visual object recognition? Neuron 73 (3), 415–434. https://www.ncbi.nlm.nih.gov/pubmed/22325196.

Diedrichsen, J., Kriegeskorte, N., 2017. Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS Comput Biol. 13 (4), e1005508. https://www.ncbi.nlm.nih.gov/pubmed/28437426.

Diedrichsen, J., Yokoi, A., Arbuckle, S., 2016a. Pattern Component Modelling Toolbox. https://github.com/jdiedrichsen/pcm_toolbox.

Diedrichsen, J., Ridgway, G.R., Friston, K.J., Wiestler, T., 2011. Comparing the similarity and spatial structure of neural representations: a pattern-component model. Neuroimage 55 (4), 1665–1678. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=21256225.

Diedrichsen, J., Wiestler, T., Krakauer, J.W., 2013. Two distinct ipsilateral cortical representations for individuated finger movements. Cereb. Cortex 23 (6), 1362–1377. http://www.ncbi.nlm.nih.gov/pubmed/22610393.

Diedrichsen, J., Zareamoghaddam, H., Provost, S., 2016b. The Distribution of Crossvalidated Mahalanobis Distances. ArXiv.

Dumoulin, S.O., Wandell, B.A., 2008. Population receptive field estimates in human visual cortex. Neuroimage 39 (2), 647–660. http://www.ncbi.nlm.nih.gov/pubmed/17977024.

Eisenberg, M., Shmuelof, L., Vaadia, E., Zohary, E., 2010. Functional organization of human motor cortex: directional selectivity for movement. J. Neurosci. 30 (26), 8897–8905. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20592212.

Ejaz, N., Hamada, M., Diedrichsen, J., 2015. Hand use predicts the structure of representations in sensorimotor cortex. Nat. Neurosci. 18 (7), 1034–1040. http://www.ncbi.nlm.nih.gov/pubmed/26030847.

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., Knutsson, H., 2012. Does parametric fmri analysis with spm yield valid results? an empirical study of 1484 rest datasets. Neuroimage 61 (3), 565–578. http://www.ncbi.nlm.nih.gov/pubmed/22507229.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the laplace approximation. Neuroimage 34 (1), 220–234. http://www.ncbi.nlm.nih.gov/pubmed/17055746.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19 (4), 1273–1302. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12948688.

Georgopoulos, A.P., Schwartz, A.B., Kettner, R.E., 1986. Neuronal population coding of movement direction. Science 233 (4771), 1416–1419.

Graziano, M.S., Aflalo, T.N., 2007. Mapping behavioral repertoire onto the cortex. Neuron 56 (2), 239–251. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17964243.

Grill-Spector, K., Henson, R., Martin, A., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn. Sci. 10 (1), 14–23. https://www.ncbi.nlm.nih.gov/pubmed/16321563.

Guntupalli, J.S., Hanke, M., Halchenko, Y.O., Connolly, A.C., Ramadge, P.J., Haxby, J.V., 2016. A model of representational spaces in human cortex. Cereb. Cortex 26 (6), 2919–2934. http://www.ncbi.nlm.nih.gov/pubmed/26980615.

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

Hermes, D., Siero, J.C., Aarnoutse, E.J., Leijten, F.S., Petridou, N., Ramsey, N.F., 2012. Dissociation between neuronal activity in sensorimotor cortex and hand movement revealed as a function of movement rate. J. Neurosci. 32 (28), 9736–9744. https://www.ncbi.nlm.nih.gov/pubmed/22787059.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532 (7600), 453–458. http://www.ncbi.nlm.nih.gov/pubmed/27121839.

Ingram, J.N., Kording, K.P., Howard, I.S., Wolpert, D.M., 2008. The statistics of natural hand movements. Exp. Brain Res. 188 (2), 223–236. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18369608.

Jonas, E., Kording, K.P., 2017. Could a neuroscientist understand a microprocessor? PLoS Comput Biol. 13 (1), e1005268. https://www.ncbi.nlm.nih.gov/pubmed/28081141.

Kakei, S., Hoffman, D.S., Strick, P.L., 1999. Muscle and movement representations in the primary motor cortex. Science 285 (5436), 2136–2139. http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=10497133.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90 (430), 773–795. http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572.

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452 (7185), 352–355. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18322462.

Khaligh-Razavi, S.M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain it cortical representation. PLoS Comput Biol. 10 (11), e1003915. http://www.ncbi.nlm.nih.gov/pubmed/25375136.

Kriegeskorte, N., 2011. Pattern-information analysis: from stimulus decoding to computational-model testing. Neuroimage 56 (2), 411–421. https://www.ncbi.nlm.nih.gov/pubmed/21281719.

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 4. http://

www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19104670.
Lescroart, M.D., Stansbury, D.E., Gallant, J.L., 2015. Fourier power, subjective distance, and object categories all provide plausible models of bold responses in scene-selective visual areas. Front. Comput Neurosci. 9, 135. https://www.ncbi.nlm.nih.gov/pubmed/26594164.
Lindstrom, M.J., Bates, M.B., 1988. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. J. Am. Stat. Assoc. 83 (404), 1014–1022.
Marblestone, A.H., Wayne, G., Kording, K.P., 2016. Toward an integration of deep learning and neuroscience. Front. Comput Neurosci. 10, 94. https://www.ncbi.nlm.nih.gov/pubmed/27683554.
McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley Series in Probability and Statistics. Wiley, New York.
Murphy, K.P., 2012. Machine Learning: a Probabilistic Perspective. MIT press, Cambridge, MA.
Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fmri. Neuroimage 56 (2), 400–410. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20691790.
Neyman, J., Pearson, E.S., 1933. On the problem of the most efficient test of statistical hypotheses. Philosophical Trans. R. Soc. A: Math. Phys. Eng. Sci. 231, 289–337.
Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput Biol. 10 (4), e1003553. http://www.ncbi.nlm.nih.gov/pubmed/24743308.
Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fmri data. Trends Cogn. Sci. 10 (9), 424–430.
Norman-Haignere, S., Kanwisher, N., McDermott, J.H., 2013. Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. J. Neurosci. 33 (50), 19451–19469. http://www.ncbi.nlm.nih.gov/pubmed/24336712.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. Neuroimage 22 (3), 1157–1172. https://www.ncbi.nlm.nih.gov/pubmed/15219588.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: A tutorial overview. Neuroimage 45 (1 Suppl). S199–209. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19070668.
Schwartz, A.B., Kettner, R.E., Georgopoulos, A.P., 1988. Primate motor cortex and free arm movements to visual targets in three-dimensional space. i. relations between single cell discharge and direction of movement. J. Neurosci. 8 (8), 2913–2927. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3411361.
Sergio, L.E., Hamel-Paquet, C., Kalaska, J.F., 2005. Motor cortex neural correlates of output kinematics and kinetics during isometric-force and arm-reaching tasks. J. Neurophysiol. 94 (4), 2353–2378. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15888522.
Sergio, L.E., Kalaska, J.F., 1997. Systematic changes in directional tuning of motor cortex cell activity with hand location in the workspace during generation of static isometric forces in constant spatial directions. J. Neurophysiol. 78 (2), 1170–1174. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9307146.
Shenoy, K.V., Sahani, M., Churchland, M.M., 2013. Cortical control of arm movements: a dynamical systems perspective. Annu Rev. Neurosci. 36, 337–359. https://www.ncbi.nlm.nih.gov/pubmed/23725001.
Siero, J.C., Hermes, D., Hoogduin, H., Luijten, P.R., Petridou, N., Ramsey, N.F., 2013. Bold consistently matches electrophysiology in human sensorimotor cortex at increasing movement rates: a combined 7t fmri and ecog study on neurovascular coupling. J. Cereb. Blood Flow. Metab. 33 (9), 1448–1456. https://www.ncbi.nlm.nih.gov/pubmed/23801242.
Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46 (4), 1004–1017. https://www.ncbi.nlm.nih.gov/pubmed/19306932.
Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., Diedrichsen, J., 2016. Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage 137, 188–200. http://www.ncbi.nlm.nih.gov/pubmed/26707889.