Representational Models

Jörn Diedrichsen

Introduction

Discovering how the human brain gives rise to intelligent behaviour, complex thought, creativity, and our rich conscious experiences is one of the most daunting scientific challenges of our time. How do neurons represent and process information? In trying to answer this question, an experimental approach inspired by system identification (Marmarelis and Marmarelis, 1978) has been long dominant across many fields of neuroscience. In this approach the researcher independently varies the stimuli presented to the nervous system and repeatedly measures the activity of a neuron under these conditions. The collected responses of the neuron are analyzed as a tuning function or activity profile across the experimental conditions. The goal of the approach is then to discover the mapping between the stimulus characteristics and the neuronal response.

Of course, neurons do not work in isolation. Rather than analyzing individual neurons, we therefore need to study the population code, that is, the activity of groups of neurons. Models of population codes usually specify a class of activity profiles (Wu et al., 2006). For example, the responses of neurons of primary motor cortex have been characterized using cosine tuning to movement direction (Georgopoulos et al., 1988). While each neuron would have a different preferred direction (fire maximally for a different movement direction) and a different tuning strength, the underlying coding principle is the same across neurons. Together, the population codes for movement direction in a distributed fashion (Georgopoulos et al., 1986). In this chapter we are concerned with such models of population activity, and will review current methods that can be used to specify and compare different models.

Measuring representations with fMRI

While many of the techniques discussed here are inspired by sensory neurophysiology (Wu et al., 2006), we will concentrate on their application to functional magnetic resonance imaging (fMRI) data. The idea of a tuning curve or activity profile can be generalized to the main unit of measurement in fMRI, the voxel. The collection of activity profiles of a group of neighbouring voxels can be analyzed as a “population code”, with
the hope that this analysis will reveal something about the underlying neuronal representations. A number of different analysis methods have been proposed to achieve this goal. We are referring to these methods here collectively as representational fMRI analysis to distinguish them from the classical fMRI approach, which simply asks whether a specific task changes the overall activity of a region relative to a well-defined control task.

One may say that analyzing a “population code” across voxels in the hope to learn something about the real underlying neuronal population code is simply one step too far. The activity measured in voxels combines the activity of millions of neurons; any information encoded in the activity differences within a voxel will be lost (Kriegeskorte and Diedrichsen, 2016; Swisher et al., 2010). Furthermore, the hemodynamic response acts as a strong low-pass filter, removing nearly all useful temporal information from the signal. Finally, fMRI does not measure neural activity directly, but through the indirect lens of the vascular response, with many possible non-linearities and irregular spatial spread (O’Herron et al., 2016). Despite these severe limitations, representational fMRI analysis has been successfully used to uncover many known characteristics of basic sensory and motor representations (Ejaz et al., 2015; Kamitani and Tong, 2006; Kay et al., 2008; Norman-Haignere et al., 2015). Importantly, the same techniques are now being applied to study representations in higher-order regions, revealing the nature of spatial (Kim et al., 2017), sequential (Kornysheva and Diedrichsen, 2014; Yokoi et al., 2018), categorical (Kriegeskorte, Mur, Ruff, et al., 2008) and semantic (Huth et al., 2016) representations in the human brain. It is here that the power of representational fMRI analysis lies, as many of these concepts would not be easily accessible in animal models.

Describing activity profiles using features

The basic goal of representational fMRI analysis is to build models of groups of activation profiles — the activity of many voxels across an ideally rich set of experimental conditions. An intuitive way to characterize activity profiles is to describe them using a flexible combination of features. This approach is taken in so-called encoding models (Naselaris et al., 2011). For example, the responses of voxels in V1 to natural stimuli can be captured using a feature set of Gabor functions with different location, frequency and orientation (Kay et al., 2008). The responses of voxels in primary auditory for complex sounds are well predicted by a model that uses the power in specific frequency bands as features (De Angelis et al., 2017).

To understand the relationship between different representational analysis methods, it is useful to take a multi-level modelling view (Figure 1). On the level of the data, the fMRI time series of a single voxel \( y_p \) is expressed as a function of a design matrix \( Z \) that captures the temporally delayed and smoothed response to each event or condition, and an activity profile \( u_p \), which determines the size of the response to each of the experimental condition. In encoding models, the average response is then modelled by a linear combination of a set of features. Each column of the matrix \( M \) corresponds to a feature. The feature weights for each
Figure 1: Multiple levels of modeling representations. The data ($y_p$) are either a collection of fMRI time-series or activity estimates from a lower-level analysis. Encoding models (left column) model the data using a set of features $M$ and feature weights $w_p$ for each voxel. Second-level parameters determine the distribution of the feature weights and noise across voxels. In Pattern-component modelling (right column) the distribution is directly specified on the activity profiles $u_p$. The marginal likelihood of the data under a given model can be directly determined by integrating out of first-level parameters (arrow).
voxel, \( w_p \), determine to what degree the voxel responds to each feature. Each feature set spans a specific subspace of the activation profiles that are “allowed” under this model. Any deviation from these permissible profiles cannot be modelled and hence reduces the measure of fit, \( R^2 \). Early encoding models used simple regression approaches to estimate the feature weights (Mitchell et al., 2008), followed by either classical statistical approaches or decoding approaches to determine the quality of the fit.

Using simple regression in encoding models, however, has severe limitations. When the number of features approaches the number of distinct experimental conditions, all models will fit the data equally well. Therefore, it is common practice to introduce a second level of parameters into the modelling, which specify a prior distribution on the voxels weights \( p(w_p) \). The prior distribution of the feature weights is often assumed to be Gaussian with mean \( 0 \) and (co)variance \( \Omega \theta_s \), where the matrix \( \Omega \) determines the shape of the distribution and the scalar \( \theta_s \) simple scales the overall variance of the signal. Together with the feature matrix this prior specifies which activity profiles are more or less likely under a specific model. Combined with the assumption that the variance of the estimated activity profiles \( \hat{u}_p \) is \( \theta \epsilon \), the best linear predictor of \( w_p \) is:

\[
\hat{w}_p = (M^T M + \Omega^{-1} \theta_s / \theta_s)^{-1} M^T \hat{u}_p.
\]

The term \( \Omega^{-1} \theta_s / \theta_s \) “shrinks” the estimates towards the more likely regions of the prior distribution. If the variances of all feature weights across voxels are assumed to be equal and their covariance to be zero, i.e. \( \Omega = I \), the above equation simplifies to Ridge regression (Murphy, 2012).

The “fit” of such model is usually compared using cross-validation. Leaving out a small subset of the data (~10%), the weights for each voxel are estimated from the remaining training set. The prediction of the left-out data is then assessed using the crossvalidated \( R^2 \) or the correlation between measured and predicted voxel activities. Crossvalidation automatically penalizes model complexity, making it possible to compare models with different number of feature weights directly. In essence, crossvalidation assesses the fit of the activity profiles under the hypothesized distribution of the activity profiles, independent of the actual features or voxel-feature weights. Under the assumptions made above, this activity profile distribution is also a Gaussian with zero mean and a (co)variance matrix that depends on both the features and the prior on the feature weights, \( G = M\Omega M^T \). This matrix therefore fully specifies the representational model.

**The feature fallacy**

This observation has an important consequence: Two representational models are identical if the predicted (co)variance matrix \( G \) is identical. This means that when we find a well-fitting feature model, we need to consider that there is many other feature models that predict the same activity profile distribution and therefore describe and predict the data equally well. That is, we need to be aware that features sets are
tools to describe the distribution of activity profiles, but that there is no special meaning in the features themselves. With the term *feature fallacy* I am referring here to the common mistake of confusing the tools that we use to describe the data and the very thing that we seek to describe.

![Image](image_url)

**Figure 2**: Three equivalent feature sets describe the tuning to finger movements in primary motor and somatosensory cortex equally well. (a) Distribution of voxels in the space of three of the experimental conditions. Each dot represents a voxel’s activity profile across thumb, middle and ring finger. The cloud of dots can be described by using individual fingers as features. (b) Surface map of the human S1 and M1 with voxels coloured according to which finger they are most activated by with thumb = red, index = yellow, middle = green, ring = blue and little = purple. The colour saturation reflects the strength of the tuning, with grey areas showing no preference for any finger. Dotted line indicates the fundus of the central sulcus. Due to distortions connected to the unfolding, the length scale is approximate. (c) The distribution can also be described using the principle components of the natural statistics of finger movements (synergies). (d) Surface map of feature weights for the synergy model. Voxels scoring high on factor 1 (positive=red, negative=yellow), factor 2 (positive=green, negative=blue) or factor 3 (positive=purple). (e) For demonstration purposes I also projected the data onto 5 random vectors. When adjusting the prior of the feature weights accordingly, these vectors explain the data equally well as the other models, but (f) lead to a different feature map.

As an example, consider the representation of finger movements in M1 and S1 (Diedrichsen et al., 2013). In this experiment participants produced isometric finger presses with one of the fingers of the contralateral hand. In a restricted area around the central sulcus, voxels vary their activity systematically with the finger used. Fig. 2a shows the activity a set of these voxel, plotted into the space of three of the fingers. As we can see, the relative amount of activity for middle and ring fingers are highly correlated across voxels, whereas the activity for a thumb movement is relatively independent. The covariance matrix of the activity profiles
is well-preserved across participants and coincides tightly with the covariance matrix of finger movements in every-day life (Ejaz et al., 2015). The findings suggest that the finger movement representation are shaped by every-day structure of such movements.

We can model this distribution of voxels using the 5 fingers as features. The crossvalidated $R^2$ will be especially high when the prior variance $\Omega$ is set to the covariance matrix of natural movements. The estimated tuning for each voxel for each finger can then be visualized in a winner-take-all map on a surface reconstruction of S1 and M1 (Fig. 2b). This plot reveals an orderly representation of individual digits.

Alternatively, we can describe the distribution using the principal components of the covariance matrix of the natural statistics of finger movements (Fig. 2c). In the motor control literature, the components underlying natural behaviours are termed “synergies” (d’Avella et al., 2006). In this coordinate systems, the weights for one feature are approximately uncorrelated those of another feature. When displaying the synergy preference of each voxel on the flatmap, are clear ordered “synergy representation” (Leo et al., 2016) emerges (Fig. 2d).

Finally, we could equally describe the distribution with random features, as displayed in Fig. 2e. As long as the prior (co)variance matrix (and hence the regularization) is adjusted correctly the crossvalidated accuracy of this model remains the same. Again, a convincing looking map of the random features can be produced (Fig. 2f).

The deeper point here is that all three feature sets are equally good descriptors of the data and all would result in the same crossvalidated prediction accuracy. With equally strong conviction we can therefore conclude that motor and somatosensory cortex represents fingers, natural synergies, or random features. Debates about which of these feature sets is more appropriate therefore misses the point. The main finding is that the distribution of activity profiles is highly structured, that this structure is preserved across individuals, and that it relates in systematic fashion to the correlation of these movements in everyday life. The feature labels we use to describe this distribution maybe secondary.

This is not to say that we shouldn’t be allowed to think about neural activity in terms of underlying features. Features can provide a semantically meaningful description of population codes and representational spaces (see below). Taking the representation of features too literally, however, constitutes an intellectual dead end. For example, in the study of response properties of primary motor cortex, decades of neurophysiological research has striven to determine if the firing rate of neurons is better describe in terms of muscle activities (Sergio et al., 2005), extrinsic movement direction (Georgopoulos et al., 1986), or synergies(Overduin et al., 2012). The ultimate answer has been that none of these describes the population especially well. Instead it is commonly found M1 neurons exhibit “mixed selectivity”. This means that motor cortex represents movement not according to any particular feature set, but rather in a latent space that represents the context dependence of complex movement (Griffin et al., 2015), while at the same time producing the dynamics necessary to generate the required patterns of muscle activity (Churchland et al., 2012). Rather than getting stuck in
the search for the underlying features, we should compare models that make testable prediction about the 
distribution of activity profiles.

**Pattern component modelling**

This insight motivated the development of pattern component modelling (PCM), which seeks to evaluate the 
fit of the model to the observed activity profiles directly. In the multi-level view of representational models 
(Fig. 1), PCM is functionally equivalent to an encoding model with a Gaussian prior. However, instead of 
estimating the feature weights and then using cross-validation to evaluate their predictive power, it evaluates 
the marginal likelihood of the data under the model (and the second-level parameters) directly.

\[
p(y_p|\theta) = \int p(y_p|u_p)p(u_p|G(\theta))du_p
\]

The first term in the integral is the conditional probability — the probability of the data given the activity 
profiles. The second term represents the prior probability of the activity profiles under the model. The 
evaluation of the integral in analytical form is possible, as both noise and the signal are assumed to come 
from Gaussian distributions. The result is the marginal likelihood, the probability of the data under the 
assumed distribution, independent of the actual value of individual activity profiles or voxel-feature weights. 
The marginalization achieves the same as the cross-validation employed in encoding models; it removes the 
first-level parameters and lets us judge the model fit accounted for the number and complexity of the features 
used.

To compare models that predict a fixed representational structure, we therefore can simply use this marginal 
likelihood as an approximation of the model evidence - the probability of the data given the model - to 
determine which models provides the most appropriate description of the data. This criterion can be shown 
to provide more powerful model comparison than cross-validation using encoding models or model comparison 
using representational similarity analysis (Diedrichsen and Kriegeskorte, 2017). Indeed, if the assumption of 
the generative model are met, PCM implements a likelihood-ratio test between models, provenly the most 
powerful test possible (Neyman and Pearson, 1933).

All second-level parameters can be efficiently optimized as analytical derivatives of the marginal likelihood in 
respect to these parameters can be relatively easily derived. A Matlab implementation of the corresponding 
algorithms are openly available (2016), and the underlying theory is described in (Diedrichsen et al., 2017). 
This provides a useful tool for model comparison with special advantages for more complex representation 
models (see below).

In sum, the core idea behind PCM is to abstract from the actual activity patterns, as well as from the features
that are used to describe them, to be able to make inferences on representational models more directly and powerfully.

**Representational similarity analysis and representational spaces**

The abstraction from activity patterns and features is shared with a number of other representational analysis techniques, first and foremost representational similarity analysis (RSA; (Kriegeskorte, Mur, and P. Bandettini, 2008; Singh and Scott, 2003)). A central concept in this approach is the notion of a representational space (Guntupalli et al., 2016; Haxby et al., 2014; Kriegeskorte and Kievit, 2013). Instead of thinking about voxel activity profiles as points in the space of experimental conditions (Fig. 3b), we can think about the conditions as points in the space of voxel activities (Fig. 3c). What is important in this representation are not so much the value on each of the axes (i.e., the activity pattern), but rather the relationship between the different activity patterns. In RSA, this relationship is quantified through a dissimilarity measure, a value that is higher, the more distinct the activity patterns are. Examples for dissimilarity measures are the Euclidean distance (and its generalization, the Mahalanobis distance) or the correlation distance (Nili et al., 2014; Walther et al., 2016). The whole representational geometry can be summarized succinctly by the matrix of all pairwise dissimilarities, the representational dissimilarity matrix (RDM).

The intuitive appeal of this approach is that the dissimilarity measure reflects how well the distinction between two condition is represented in an area. That is, it tells us how well a read-out neuron that has access to the whole population code, could distinguish the two conditions. More generally, the representational geometry determines how well any feature that describes the underlying conditions could be decoded from the population code (Diedrichsen and Kriegeskorte, 2017). An especially useful dissimilarity measure in this context is the crossvalidated estimate of the Mahalanobis distance (crossnobis for short, (Diedrichsen et al., 2016; Walther et al., 2016)). This distance estimate is unbiased, i.e., the expected value of the dissimilarity is zero if two activity patterns only differ by noise. Thus, like crossvalidated decoding performance (Pereira et al., 2009), the crossnobis estimate can be used to quickly assess whether there is a true differences between two activity patterns and how strong this difference is.

RSA, PCM and encoding models are tightly related. This relationship is due to the fact that all three approaches assess the correspondence of data and model by comparing the respective second-moment matrices of the activity profile distributions. The second moment of a group of $P$ activity profiles is defined as

$$G = \sum_p u_p u_p^T / P.$$ 

Both Euclidean $(d_{1,2} = \sqrt{G_{1,1} - 2G_{1,2} + G_{2,2}})$ and correlation distances $(d_{1,2} = 1 - G_{2,1} / \sqrt{G_{1,1}G_{2,2}})$ are direct functions of this matrix. Similarly to the RDM, the second-moment matrix therefore determines the
Figure 3: Representational spaces. (a) The data consist of repeated measures of the same set of voxels across a range of conditions. Each column of the matrix constitutes an activity profile, each row an activity pattern across voxels. (b) The activity profiles can be plotted in the space of the experimental conditions. Representational models specify a distribution of activity profiles. (c) The activity patterns can be plotted in the space spanned by the voxels. The relationships between activity patterns in this high-dimensional space define the representational geometry (red lines). (d) Two views of a low-dimensional projection of the representational geometry of individual finger movements in primary motor cortex (1:thumb - 5: little finger) at 4 different movement speed (black: slow - yellow: fast). Resting baseline is indicated by cross.*
representational geometry completely, with the only difference that it also specifies the distance of each
condition from the baseline activity (usually rest in fMRI).

The definition of the second-moment matrix is nearly identical to the definition of the (co)variance matrix
of the activity profiles, the central statistics in encoding and PCM models. The only difference is that for a
(co)variance matrix, the mean activity profile (across voxels) is subtracted before applying Eq. 3. However,
in the context of representational analysis, differences in mean activity between conditions are meaningful —
indeed many important dimensions, such as stimulus intensity, speed or force of action, are encoded in the
overall strength of the activity. Subtracting the mean activity profile (column mean in Fig. 3a) would remove
such difference between conditions and severely distorts the representational geometry. Although sometimes
performed, such subtraction cannot be considered meaningful from a representational standpoint. In PCM
and encoding approaches the mean of the Gaussian distribution is assumed to be zero — so here as well,
we do not remove the empirical mean activity profile from the data. Therefore, all three approaches define
models by the second moment of the predicted activity profile distribution. The only technical difference is
that the loss or cost function, which measures the mismatch between the empirical and the predicted second
moment matrix has slightly different characteristics (Diedrichsen and Kriegeskorte, 2017).

An additional important use of representational spaces is to visualize (without the need to define a-priori
models) the representational geometry of a population code. Because we usually consider patterns with
many voxels, we need to reduce the dimensionality of the space to be able to generate graphs in 2 or 3
dimensions. A common approach here is to use the first 3 eigenvectors of the second-moment matrix, i.e.,
the 3 patterns that best differentiate between conditions. However, other views of the representational space
are possible. An exploration across multiple views can be for example guided by picking dimensions that
maximize specific experimental contrasts of interest (Diedrichsen et al., 2017; Kobak et al., 2016). While
such visualizations can be very powerful (Fig. 3d), visual representations of representational spaces need to
be treated with some caution, as different views of the same space sometimes can tell very different visual
stories. This emphasizes the importance of formal model comparison, which always should be conducted in
the full, rather than the reduced, visualized, representational space.

In summary, the representational space is a very important concept for cognitive neuroscience. Representa-
tional spaces abstract from the underlying spatial activity pattern and from hypothetical features that maybe
be used to describe them. Indeed, it is often observed that the specific activity patterns are quite variable
across individuals, whereas the structure of the representation is well preserved (Ejaz et al., 2015; Guntupalli
et al., 2016). The concept of a representational space asserts that what matters is the representational con-
tent of a population code, but not the details of how it is laid out on the cortical sheet (Graziano and Aflalo,
2007). The feasibility of functional hyper-alignment between subjects (see chapter X) is a consequence of
the functional invariance of representational spaces.
Flexible representational models

Comparison of representational models can be performed using encoding approaches, RSA or PCM. While PCM offers slightly more powerful inferences than its alternatives (Diedrichsen and Kriegeskorte, 2017), the small increase in power alone may not constitute an overwhelming argument for the use of PCM. The strength of the technique, however, becomes apparent when considering more complicated or flexible representational models.

The representational models considered so far have been “fixed”. The crossvalidation in encoding models, the calculation of the marginal likelihood in PCM, and the calculation of distances in RSA, effectively “integrates out” the first-level parameters (Fig. 1). That is, even with thousands of free voxel-feature weights, encoding models that use a single ridge-coefficient predict a fixed distributional shape of activity profiles. The only remaining flexibility is the noise and signal variance, which together determine the ridge coefficient. Therefore, even encoding models with thousands of voxel-feature weights can be considered to be fixed models.

It is, however, a rare stroke of luck if we can explain the neural activity patterns in a specific region fully using a single principle. From a computational perspective it may be intuitive at first to think of information processing as a clear sequence of transformations. Speech comprehension, for example, could be conceptualized as starting with a stage that analyzes the spectral features, followed by a stage that detects phonemes, and ending in a stage that identifies the semantics of the entire word (Bornkessel-Schlesewsky et al., 2015; Poeppel et al., 2012; Price, 2010). The planning of motor action is historically thought of as a sequence of computations that take a visual representation of the goal, into a body- and then hand-centred spatial representation. This is then followed by the computation of a desired displacement vector and ultimately the appropriate muscle commands (Flanders et al., 1992; Shadmehr and Wise, 2005).

Unfortunately, the brain does not usually oblige to arrange such computations anatomically in an orderly fashion, with each region corresponding to one distinct stage of information processing. Rather, most areas show a mixture of representations from multiple processing stages. For example, coding in the caudal premotor and primary motor cortex exhibit a mixture of muscle-like and extrinsic tuning in a range of different coordinate systems (Griffin et al., 2015; Kakei et al., 1999; Wu and Hatsopoulos, 2007). Similarly, in most sensory systems, regions show a mixture of selectivities, combining basic perceptual features with context information and attentional signals. Therefore, our default assumption should not be a representation can be explained by a single feature set, but rather that single regions usually exhibit a mixture of different representations.

This should be reflected in the way we perform analysis of brain data. Rather than trying to find sparse solutions and plot “winner take all” maps, we require techniques that allow us to rigorously test for a mixtures of representations. Different approaches have been developed for this purpose. For example, using
RSA, Khaligh-Razavi and Kriegeskorte (2014) combined the predictions derived from different layers of a deep-neural network with a categorical model to explain object representations in inferior temporal cortex.

Similarly, using an encoding approach, Heer et al. (2017) considered a mixture of spectral, articulatory, and semantic properties to explain representations in the auditory processing stream. Using PCM, Yokoi et al. characterize finger sequence representations in cortical motor areas through a mixture of single-finger, and 2nd-higher order finger-transition representations (Yokoi et al., 2018) (Fig 4).

![Finger transition models](image)

Figure 4: **Flexible representational models.** The second moment of the data is modelled as a combination of model components, where the mixing proportions are specified by a second-level parameter. Each model component represents a unique feature set. In this example (Yokoi et al., 2017), the activity patterns for 6 finger sequences (data) are modelled as a combination of a feature set that models the first finger in the sequence, a feature sets that encode all 2nd-order transitions between fingers, followed by even higher-order transitions. The highest level component has unique pattern for each sequence.

Methodologically, this is a model selection and inference problem that is relatively well studied in the statistical literature (Clyde, 1999). The first step is to be able to fit more complex models to the data. If we combine different models to explain the neural representations, what is the optimal weighting of each of these feature sets or components? In the context of PCM, the solution to this problem is simple: The second-moment matrix of the data can be modelled using a weighted linear combination of the predicted second-moment matrix for each component $G_i$. Linear combination assumes that features across different components are mutually independent. In this case, the overall second moment matrix of the combined model would be

$$G = \sum_i \theta_i G_i$$

where $\theta_i$ are the non-negative component weights. Because the analytical derivates of the marginal likelihood in respect to the weight can be easily derived, the optimal weights can be estimated very efficiently. Note that this approach is very similar to weighted RSA, with the advantage that PCM automatically deals with
the dependence of different elements of $G$.

The second step is to find the best feature sets to combine. This can be achieved by fitting all possible combination of the candidate model components (i.e., feature sets). Because each of the $K$ component can independently be either present or absent, this surmounts to fitting $2^K$ models. For small number of components, every combination usually be fitted easily. Once the number of candidate components becomes large, one needs to resort to model search strategies, such a step-wise approaches or approximate Bayesian approaches (Clyde, 1999). In evaluating each model, we need to take into account the increasing complexity of models with larger number of components. Even though PCM does not require cross-validation for fixed models for any number of features, as the parameters on the first level are integrated out in the marginal likelihood, this analytical approach cannot be applied easily to the parameters on the second level (Fig. 2). Practically, we can resort here to using AIC, BIC (Burnham and Anderson, 2004) or cross-validation within or across participants (Diedrichsen et al., 2017). Each of these measures gives us an estimate of the model evidence for each combination of features.

The last step is to make inferences about the underlying representation. Rather than inferences about which specific combination of features is best, we usually want to make inferences about the presence or absences of a specific component. For example, we would like to map how strong the evidence for a semantic representation is in secondary auditory areas in the context of the other competing explanations. Because representational models are models about probability distribution of activity profiles, the concept of “variance-partitioning” (Heer et al., 2017), which is useful in simple ANOVA-based approaches, is statistically not applicable here. Rather, we need a probabilistic approach. We can either evaluate the strength of evidence for each component in terms of a Bayes factor, or the posterior probability of each component. These calculation are based in the estimate of the model evidence (the likelihood of the data given the model), provided by PCM. We can then use the knock-in or knock-out change in evidence for each component, starting from the best model (Shen and Ma, 2017). If the component in question is not present in the best model, we can use the loss in likelihood when the component is added as a measure of evidence against the component. If the component is present in the best model, we can use the loss in the likelihood when the component is dropped as a measure of evidence in favour of the component. A more complete and robust solution is Bayesian model averaging. In this approach we would compare the total posterior probability of all models that contain the component to the posterior probability of all models that do not contain the component (Clyde, 1999; Shen and Ma, 2017). Inference of the group level can then be conducted using normal frequentist tests using log-Bayes factors, or by using Bayesian models of inter-subject variability (Stephan et al., 2009).

Inference on flexible representational models is an area of multivariate analysis that currently is developing quickly. We should stress that our discussion is not restricted to models, in which the second-moment matrix of different model components combine linearly (Eq. 4). Exactly the same principles apply to models in which the second moment matrix is a differentiable (but otherwise arbitrary) function of the second-
level parameters, $G = F(\theta)$. This allows the user to model receptive fields with different widths, arbitrary correlations between feature sets, and the relative re-weighting of individual features within correlated feature sets. In the context of RSA, such models have been described as mixed vs. weighted RSA (Khaligh-Razavi et al., 2016; Khaligh-Razavi and Kriegeskorte, 2014). In PCM, model estimation and inference for these two types is performed with exactly the same machinery (Diedrichsen et al., 2017). These techniques, as well as publically available software tools, enable researchers to make inferences on relatively complex models. It is therefore important to raise awareness of how to build and infer on complex representational models within each of the different approaches.

New developments

Overall, the structure and inference on standard representational models is starting to be well understood. There are, however, a number of novel directions of method development that are particularly promising and exciting.

Throughout the chapter we have assumed that the prior on the feature weights is Gaussian, prompting us to use the second moment as the sufficient statistic for model definition and comparison. However, we can also construct representational models that predict a clustering of activity profiles around specific feature directions, with arbitrary combinations of features being rare. This corresponds to the hypothesis of a sparse representation. This idea can be implemented in a PCM-like approach by using a multivariate Gamma distribution as a prior (Norman-Haignere et al., 2013). Similarly, using different dissimilarity measure and ways of construction representational space can RSA sensitive to aspects of the activity profile distribution that go beyond the second moment. However, evidence that such models provide a better description of fMRI data as compared to models with Gaussian priors is still missing.

Another frontier of method developments is to soften one of the assumptions that is at the core of RSA and PCM: namely that the actual spatial layout of the activity patterns does not matter. While the exact layout of the different activity profiles on the cortical sheet likely reflects a lot of random biological variation, there may be some aspects of the spatial arrangement that may matter. For example, some features are represented in the fine neuron-by-neuron variation in activity profiles, while other are encoded in more broad spatial gradients of functional sensitivities. While fMRI will be strongly biased to pick up the features that are represented at a larger spatial scale, there may be aspects of the spatial structure in the fMRI signal itself that could reveal meaningful variations in representations. More comprehensive spatiotemporal models combined with improvements in the effective spatial resolution of fMRI will be needed to effectively reveal such organization.

Finally, the generation of better models of brain representations is a highly active area of current research.
While traditional representational models are motivated by sets of hypothesized and hand-crafted features, new theories are increasingly informed by machine learning techniques. For example, numerous studies have systematically compared representational geometries emerging in the hidden layers of deep neuronal networks to the representational geometries found in areas of the brain that are expected to perform similar function (see Chapter X for overview). These developments are exciting and promise to elevate the study of brain representation to the next level — namely to build models of brain computations that would be real enough to actually perform the underlying tasks.

References


Diedrichsen, J., Zareamoghaddam, H., Provost, S., 2016. The distribution of crossvalidated mahalanobis


Kamitani, Y., Tong, F., 2006. Decoding seen and attended motion directions from activity in the human vi-


Shadmehr, R., Wise, S.P., 2005. The computational neurobiology of reaching and pointing. MIT Press,


2016.