Representational models and the feature fallacy

Jörn Diedrichsen

Brain and Mind Institute, Western University, Canada

Department of Statistical and Actuarial Sciences, Western University, Canada

Department of Computer Science, Western University, Canada

September 5, 2018

Abstract

In this chapter, I discuss models that specify how neuronal population activity relates to things that happen in the world - i.e., models of neuronal representations. Focussing on the application to functional magnetic resonance imaging (fMRI) data, I discuss current approaches to estimate and test such models. Encoding models conceptualize representations in terms of sets of underlying features. We show that these models ultimately test hypotheses about the distribution of activity profiles in the space of the experimental conditions, and that the exact choice of feature sets is to some degree arbitrary. Over-interpreting the significance of specific feature sets constitutes an intellectual dead-end, termed here "feature fallacy". The same representational models can also be tested using a Bayesian approach, Pattern component modelling (PCM), which abstract from the underlying features and allows direct model evaluation. It also provides a powerful means to test more flexible representational models, including models that consist of combinations of different feature sets.

1 Introduction

Discovering how the human brain gives rise to intelligent behaviour is one of the most daunting scientific challenges of our time. How do neurons represent and process information? In trying to answer this question, many researchers have taken an experimental approach, independently varying stimuli or task conditions, and repeatedly measuring the activity of neurons. The responses of each neuron across experimental conditions form a tuning function or activity profile. The goal of the approach is ultimately to discover the mapping between the stimulus characteristics and the neuronal response (Wu et al., 2006).

Of course, neurons do not work in isolation. Therefore, representational models need to be models of population codes, the activity of groups of neurons. Models of population codes usually specify a class of activity profiles. For example, the responses of neurons in M1 can be described as having cosine tuning to movement direction. While each neuron has a different preferred direction (fires maximally for a different movement), 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 I am concerned with such models of population activity, and will review current methods that can be used to specify and compare different models.

1.1 Measuring representations with fMRI

While many of the techniques discussed here are inspired by sensory neurophysiology, I will concentrate on their application to fMRI data. The concept of an activity profile can be generalized to the main measurement unit of fMRI, the voxel. A collection of neighbouring voxels can be analyzed as a "population code" in the hope that this will reveal something about the underlying neuronal representation. A number of different analysis methods have been proposed to achieve this goal. I refer to these methods collectively as *representational fMRI analysis* to distinguish them from the classical fMRI approach, which simply asks whether a region increases its activity in response to a specific task.

One may say that analyzing a "population code" across voxels in the hope to learn something about the underlying neuronal population code is simply one step too far. The activity measured in a single voxel combines activity of millions of neurons; any information encoded in the activity differences within a voxel will be lost (Kriegeskorte and Diedrichsen, 2016). Furthermore, the sluggish hemodynamic response removes nearly all useful temporal information from the signal. Finally, fMRI does not measure neural activity directly, but through the indirect lens of vascular responses, with many possible non-linearities and irregular spatial spread. Despite these severe limitations, representational fMRI analysis has been successful in uncovering many known characteristics of basic sensory and motor representations (Ejaz et al., 2015; Kay et al., 2008; Norman-Haignere et al., 2015). With adequate caution, it is therefore possible to use representational fMRI analysis to study higher-order regions of the human brain - already the approach has provided insights into the nature of spatial (Kim et al., 2017), sequential (Yokoi et al., 2018), categorical (Kriegeskorte, Mur, Ruff, et al., 2008) and semantic (Huth et al., 2016) representations. It is here that the power of representational fMRI analysis lies, as many of these concepts are not easily accessible in animal models.

1.1.1 Describing activity profiles using features

The goal of representational fMRI analysis is to build models of the activation profiles of groups of voxels across an ideally rich set of experimental conditions. An intuitive way to characterize activity profiles is through a flexible combination of features, an approach 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).

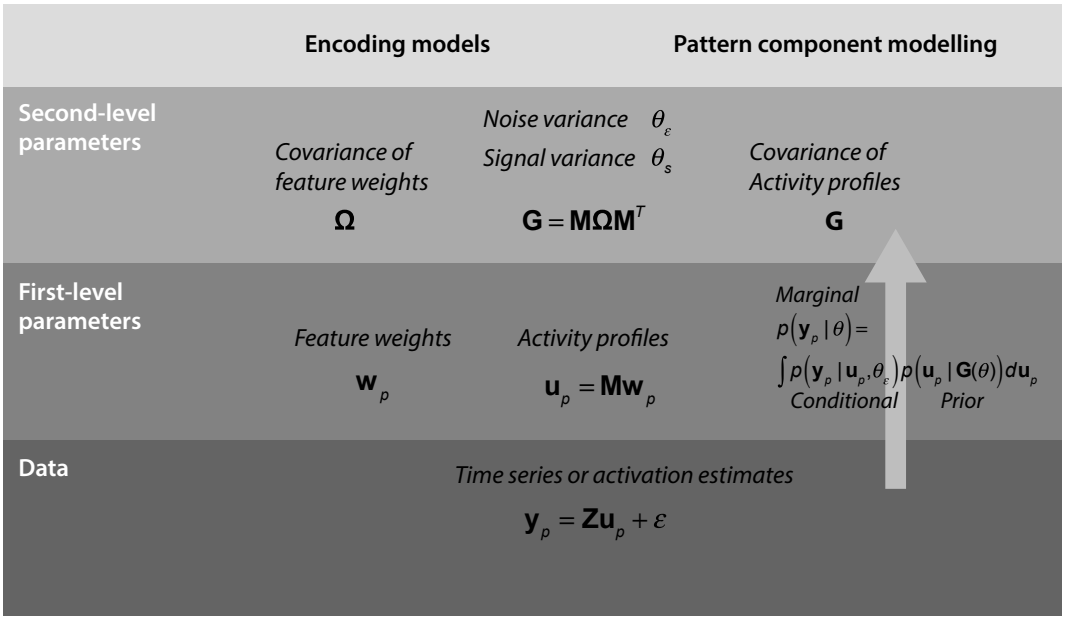

Figure 1: *Multiple levels of a representational model. The data ($\mathbf{y}_p$) are either fMRI time-series or activity estimates from a lower-level analysis for each voxel p. Encoding models (left column) model the data using a set of features* $\mathbf{M}$ *and feature weights* $\mathbf{w}_p$*. Second-level parameters determine the distribution of feature weights and noise across voxels. In PCM (right column) the distribution is directly specified on the activity profiles* $\mathbf{u}_p$*. The marginal likelihood of the data under a given model can be directly determined by integrating out of first-level parameters (arrow)*

To understand the relationship between different representational analysis methods, it is useful to take a multi-level modelling view (Fig. 1). On the level of the data, the fMRI data of the p^{th} voxel, $\mathbf{y}_p$, is expressed as a function of a design matrix $\mathbf{Z}$ that captures the temporally delayed and smoothed response to each event or condition, and an activity profile $\mathbf{u}_p$, which determines the size of the response to each of the experimental condition. In encoding models, the activity profile is then modelled by a linear combination of a set of features: Each column of the matrix $\mathbf{M}$ corresponds to a feature, and the feature weights, $\mathbf{w}_p$, determine to what degree voxel p responds to those features. Each feature set spans a specific subspace of the activation profiles that are "allowed" under the model. Any deviation from these permissible profiles would reduce the measure of fit. Early encoding models used multiple regression to estimate the feature weights (Mitchell et al., 2008), followed by either decoding or classical statistical approaches to determine the quality of the fit.

Using standard regression, however, has severe limitations. When the number of features approaches the number of distinct experimental conditions, all models fit the data equally well. Therefore, it is common practice to introduce a second modelling level that specifies a prior distribution on the voxels weights $p(\mathbf{w}_p)$, often assumed to be Gaussian with mean $\mathbf{0}$ and (co)variance $\mathbf{\Omega}\theta_s$. Matrix $\mathbf{\Omega}$ determines the shape of the distribution and θ_s simply scales the overall variance of the signal. Together with the feature matrix, this prior specifies how likely activity profiles are under a specific model. If we take the data to be the vector of estimated activity profiles $\hat{\mathbf{u}}_p$, obtained from a first level time-series model, and if we assume that these are estimated with error variance θ_ϵ,

then the best linear predictor of $\mathbf{w}_p$ is

$$\hat{\mathbf{w}}_p = \left(\mathbf{M}^T\mathbf{M} + \boldsymbol{\Omega}^{-1}\theta_\epsilon/\theta_s\right)^{-1}\mathbf{M}^T\hat{\mathbf{u}}_p. \quad (1)$$

The term $\boldsymbol{\Omega}^{-1}\theta_\epsilon/\theta_s$ shrinks the estimates towards the more likely regions of the prior distribution. In the case of $\boldsymbol{\Omega} = \mathbf{I}$, the above equation simplifies to Ridge regression.

The "fit" of such model is usually compared using crossvalidation. Leaving out a small subset of data (˜10%), the voxel weights are estimated from the remaining training set. The prediction of the left-out data is then assessed using crossvalidated R^2 or correlation between measured and predicted voxel activities. Crossvalidation automatically penalizes model complexity, making it possible to compare models with different number of features directly. In essence, crossvalidation assesses the likelihood of the activity profiles under the hypothesized distribution of the activity profiles, independent of the actual voxel-feature weights. This distribution is a Gaussian (implicitly assumed when using Eq. 1), with zero mean and a (co)variance matrix that depends on the features and their prior, $\mathbf{G} = \mathbf{M}\boldsymbol{\Omega}\mathbf{M}^T$. Matrix $\mathbf{G}$ therefore fully specifies the representational model.

1.2 The feature fallacy

This observation has an important consequence: Two representational models are identical if the predicted (co)variance matrix $\mathbf{G}$ is identical. When we find a well-fitting feature model, we need to consider that there are many other feature sets that predict the same activity profile distribution and therefore describe and predict the data equally well. Thus, features sets may be useful tools to describe distributions of activity profiles, but they do not carry a special significance in themselves. With the term *feature fallacy* I am referring to the common mistake of confusing the tools that we use to describe the data and the very thing that we seek to understand.

As an example, consider the representation of finger movements in M1 and S1 (Ejaz et al., 2015). In this experiment participants produced isometric finger presses with each finger 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 middle and ring finger activity is highly correlated across voxels, whereas thumb activity is relatively independent. The covariance matrix of the activity profiles is well-preserved across participants and coincides with the covariance matrix of finger movements in every-day life. 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 cross-validated fit will be especially good when the prior variance $\boldsymbol{\Omega}$ 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 the cortical surface (Fig. 2b), revealing a somatotopic map, an orderly representation of individual digits from thumb (ventral) to little finger (dorsal).

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".

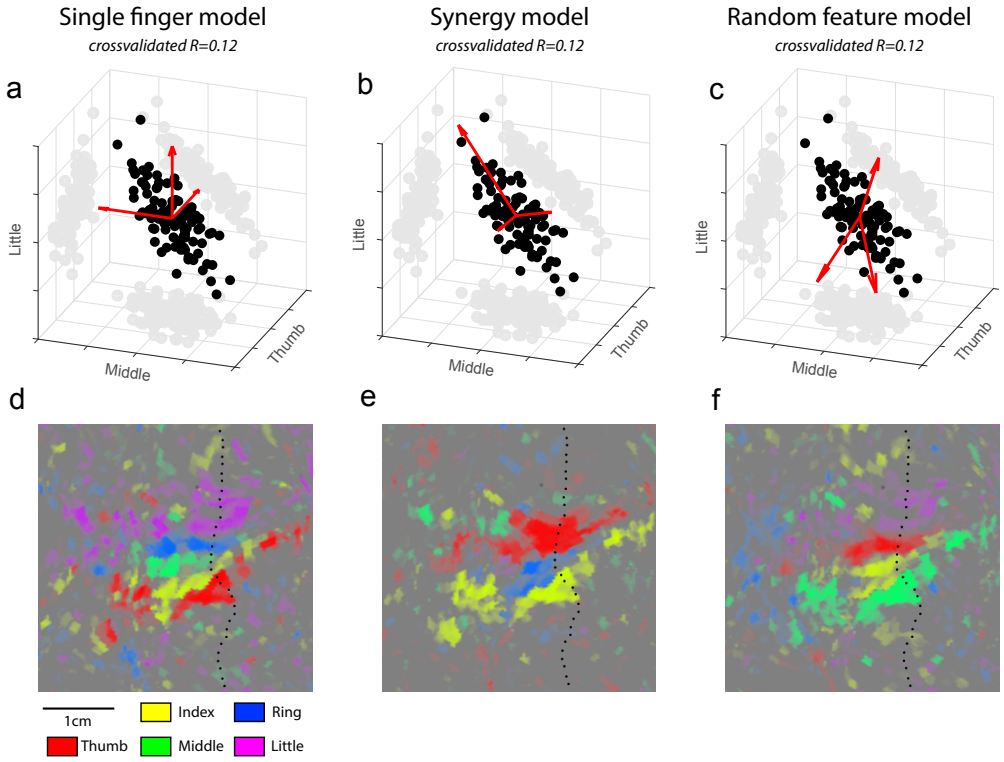

Figure 2: *Three feature models that describe the activity during finger movements in M1 and S1 equally well. (**a**) Distribution of voxels in the space of three of the experimental conditions. Each dot represents a voxel's activity profile across thumb, middle and ring finger. The distribution can 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. The colour saturation reflects the strength of the tuning, with grey areas showing no finger preference. Dotted line: fundus of the central sulcus. Length scale is approximate. (**c**) The distribution can also be described using the principal components of the natural statistics of finger movements (synergies). (**d**) Surface map of feature weights for the synergy model with voxels either scoring high (+) or low (-) on each PC. (**e**) A model with 5 random feature vectors explains the data equally well, but (**f**) lead to a different feature map.*

In this coordinate systems, the weights for the different features are approximately uncorrelated. When mapping the synergy preference of each voxel, an ordered "synergy representation" (Leo et al., 2016) emerges (Fig. 2d). Finally, we could equally describe the distribution with random features (Fig. 2e). As long as the prior (co)variance matrix (and hence the regularization) is adjusted accordingly, 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. Even when we constrain the prior (co)variance matrix to be diagonal, there are in infinite number of feature sets that would lead to the same prediction for left-out data. With equally strong conviction we can therefore conclude that primary sensorimotor cortex represents fingers, natural synergies, or random features. All three conclusions would be valid to some degree, but none of them would provide a deeper insight into underlying neuronal computations. Debates about which feature set is most appropriate therefore miss 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 the distribution are 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). They are useful tools in constructing representational models, especially in cases in which there no discrete experimental conditions – for example when a continuous movie is used as a stimulus. Taking the representation of features too literally, however, constitutes an intellectual dead-end. Neurophysiological research, for decades, has striven to determine whether the firing rate of M1 neurons is better described in terms of muscle activities, extrinsic movement direction, or synergies. The ultimate answer has been that none of these features 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, 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.

1.3 Pattern component modelling (PCM)

This insight motivated the development of PCM, which seeks to evaluate the likelihood of the observed activity profiles under the hypothesized distribution directly. In the multi-level view of representational models (Fig. 1), PCM is functionally equivalent to an encoding model with a Gaussian prior. Instead of defining features, 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.

$$p(\mathbf{y}_p|\theta) = \int p(\mathbf{y}_p|\mathbf{u}_p)p(\mathbf{u}_p|\mathbf{G}(\theta))d\mathbf{u}_p \quad (2)$$

The first term of the integral is the conditional probability of the data given the activity profiles, the second term 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 be Gaussian. The result is the marginal likelihood, the probability of the data under the assumed distribution, independent of the actual value of individual activity profiles. The marginalization achieves the same as the crossvalidation employed in encoding models: it corrects for the complexity (number of features) of the model.

To determine which model provides the most appropriate description of the data, we can therefore simply chose the model with the highest marginal likelihood. The ratio of the likelihoods is the Bayes factor, a measure of the evidence of one model over the other (Kass and Raftery, 1995). This approach is valid for models that predict a fixed representational structure, i.e. models based on a single feature set with only one signal variance and noise variance parameter on the second level (θ). Under these circumstances, the marginal likelihood can serve as an approximation of the model evidence - the probability of the data given the model. Using Bayes factors allows for more powerful model comparison than possible through encoding models or representational similarity analysis (RSA) (Diedrichsen and Kriegeskorte, 2017).

All second-level parameters can be efficiently optimized as analytical derivatives of the marginal likelihood with respect to these parameters are easily derived. A Matlab implementation of the corresponding algorithms is openly available (Diedrichsen et al., 2017). In sum, the core idea behind PCM is to abstract from the actual activity patterns and model features, to make direct and powerful inferences about representational models.

1.4 Representational similarity analysis (RSA) and representational spaces

The abstraction from activity patterns and features is shared with a number of other representational analysis techniques, first and foremost RSA (Kriegeskorte, Mur, and P. Bandettini, 2008). A central concept in this approach is the notion of a representational space (Guntupalli et al., 2016; 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). The relationship between the different activity patterns in this space defines the representation. In RSA, the relationship is quantified through a dissimilarity measure, with higher values indicating more distinct activity patterns. The representational geometry can be summarized succinctly by the matrix of all pairwise dissimilarities, the representational dissimilarity matrix (RDM).

Dissimilarity measures have the intuitive appeal of reflecting how strongly the distinction between two conditions 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 read out. An especially useful dissimilarity measure is the crossvalidated estimate of the Mahalanobis distance (Diedrichsen

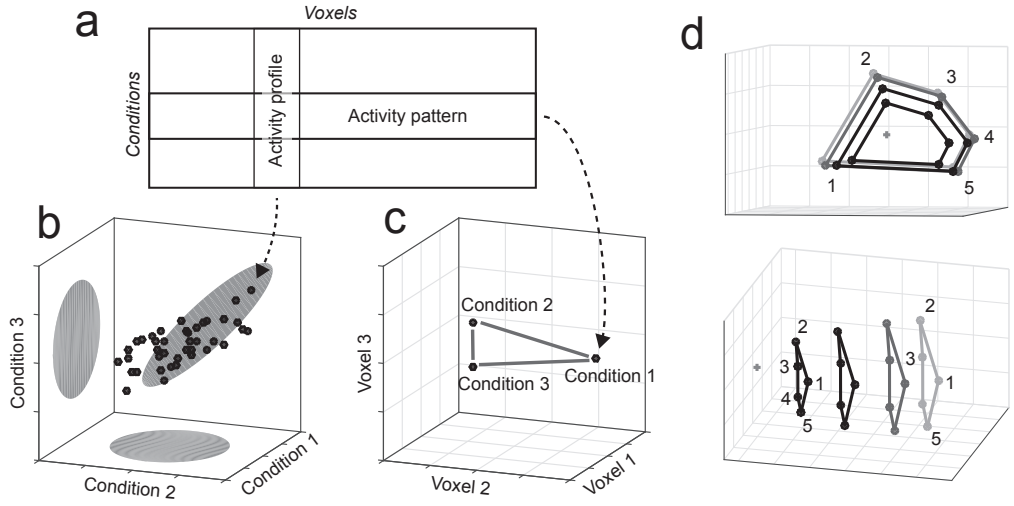

Figure 3: *Representational spaces.* ***(a)*** The data consist of repeated measures of the same set of voxels across a range of conditions. Each column of the matrix constitutes an activity profile, each row an activity pattern across voxels. (**b**) The activity profiles can be plotted in the space 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 (lines). (**d**) Two views of a low-dimensional projection of the representational geometry of individual finger movements in M1 (1:thumb - 5: little finger) at 4 different movement speed (black: slow - gray: fast).*

and Kriegeskorte, 2017). 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, it can be used to assess whether there is a true differences between two activity patterns.

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

$$\mathbf{G} = \sum_p \mathbf{u}_p \mathbf{u}_p^T / P. \tag{3}$$

Both Euclidean ($d_{1,2} = \sqrt{G_{1,1} - 2G_{2,1} + G_{2,2}}$) and correlation distances ($d_{1,2} = 1 - G_{2,1}/\sqrt{G_{1,1}G_{2,2}}$) are direct functions of this matrix. As the RDM, the second-moment matrix determines the representational geometry completely, with the only difference that it also specifies the distance of each condition from resting baseline.

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, are encoded in the overall activity. Subtracting the mean activity profile (column mean in Fig. 3a) would remove such difference and severely distort the representational geometry. Although sometimes performed, such subtraction are therefore not meaningful from a representational standpoint. By not performing this subtraction, PCM and encoding approaches assume the mean of the signal distribution to be zero. Any meaningful mean activity differences are therefore captured in the second moment matrix. In sum, all three approaches define models by the second moment of the predicted activity profile distribution. The only technical difference is how exactly the mismatch between empirical and predicted second moment matrix is measured (Diedrichsen and Kriegeskorte, 2017).

Using representational spaces, one can visualize the representational geometry of a population code without the need to define a model a-priori. To be able to generate graphs in 2 or 3 dimensions, we typically need to reduce the dimensionality of the space. A common approach here is to use the first 3 eigenvectors of $\mathbf{G}$, i.e., the 3 patterns that best differentiate between conditions. It is, however, also useful to explore other views, for example by picking dimensions that maximize specific experimental contrasts of interest (Diedrichsen et al., 2017; Kobak et al., 2016). While visualizations of representational spaces can be very powerful (Fig. 3d), they need to be treated with caution, as different views of the same data 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, representational spaces are an important concept for cognitive neuroscience. Representational spaces abstract from the underlying spatial activity pattern and from hypothetical features that are 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 content of a population code, but not the details of how it is laid out on the cortical sheet.

1.5 Flexible representational models

Comparison of representational models can be performed using encoding approaches, RSA or PCM. While PCM offers slightly more powerful inferences (Diedrichsen and Kriegeskorte, 2017), the small increase in power alone may not constitute an overwhelming argument for its use. The strength of PCM, 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 voxel-feature weights, encoding models with a single ridge-coefficient predict a fixed distribution of activity profiles.

It is, however, a rare stroke of luck if we can explain the neural activity patterns in a specific region fully using a single feature set. 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 (Poeppel et al., 2012).

Unfortunately, nature does not usually accommodate neuroscientists by arranging computations in an anatomically 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 extrinsic and muscle-like tuning (Wu and Hatsopoulos, 2007). Similarly, most sensory regions respond to a mixture of basic perceptual features, context information, and attentional signals. Therefore, our default assumption should not be that a representation can be explained by a single feature set, but rather that each region will exhibit a specific mixture of different representations.

This should be reflected in the way we analyze brain data. We therefore require techniques that allow us to 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. characterized finger sequence representations in cortical motor areas through a mixture of representations of single fingers and transition between finger presses (Yokoi et al., 2018).

Methodologically, this is an 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, which requires us to find the optimal weighting of each feature set. In the context of PCM, the second-moment matrix of the data can be modelled using a weighted linear combination of the predicted second-moment matrix of $i = 1...I$ components

$$\mathbf{G} = \sum_i \theta_i \mathbf{G}_i, \tag{4}$$

where θ_i are the non-negative component weights. Linear combination here assumes that the features across different components are mutually independent. Because the analytical derivates of the marginal likelihood in respect to each weight can be easily derived, the optimal weights can be estimated very efficiently.

The second step is to find the best feature sets to combine. This can be achieved by fitting all possible combinations of components (i.e., feature sets). Because each of the K component can independently be either present or absent, this amounts to fitting 2^K models, which is easily possible for small K. Once the number of candidate components becomes large, one needs to resort to model search strategies, such a step-wise or approximate Bayesian approaches (Clyde, 1999). In evaluating each model, we need to take into account the increasing complexity of models with more components. Even though PCM does not require crossvalidation for fixed models, as first-level parameters are integrated out in the marginal likelihood, this analytical approach cannot easily be applied to second-level parameters (Fig. 2). Practically, we can resort here to using AIC, BIC 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 presence or absence of a specific component. For example, we would like to map how much evidence there is for a semantic representation in secondary auditory areas *in the context* of the other competing explanations. It has been suggested to use the concept of "variance-partitioning" (Heer et al., 2017) to express what proportion of the variability can be explained uniquely by a feature set. This concept from unregularized linear models does, however, not readily generalize to the probabilistic setting. For example, the combination of two feature sets often provides a poorer description of the data than either feature set alone, even if the two feature sets account for different aspects of the data. Thus, a probabilistic approach is needed. One attractive approach is to evaluate the strength of evidence for each component in terms of a Bayes factor. We can apply Bayesian model averaging and divide the total posterior probability of all models containing the component by the posterior probability of all models not containing the component (Clyde, 1999; Shen and Ma, 2017). Inference on 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 is developing quickly. Our discussion is not restricted to models, in which the second-moment matrix of different components combine linearly (Eq. 4). Exactly the same principles apply to models in which $\mathbf{G}$ is a differentiable (but otherwise arbitrary) function of the second-level parameters, $\mathbf{G} = F(\theta)$. In PCM, these types of models can be handled with the existing machinery (Diedrichsen et al., 2017). This allows the user to model receptive fields with different widths, arbitrary correlations between feature sets, and relative weighting of individual features within correlated feature sets. These techniques enable researchers to make inferences on relatively complex representational models.

1.6 New developments

The structure and inference of standard representational models is starting to be well understood. Among the many novel directions of method development, I highlight three 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 comparison. However, we can also construct representational models that predict a clustering of activity profiles around specific feature directions. A classical example would be the tuning of V1 neurons that respond to specific locations in space (with each location being a feature), with neurons showing responses for a number of disparate locations being rare. The corresponding distribution of activity profiles would not be Gaussian anymore, but would have elongations along specific directions. Such a distribution can be modelled within a PCM-like approach by using a multivariate Gamma distribution as a prior (Norman-Haignere et al., 2013). Similarly, using a different dissimilarity measure and ways of constructing representational spaces could render 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 assumption that is at the core of RSA and PCM: namely that the spatial layout of activity patterns does not matter. While the exact layout on the cortical sheet likely reflects to a large degree random biological variation, there may be some aspects of the spatial arrangement that do matter. For example, some features are represented in the fine neuron-by-neuron variation in activity profiles, while others are encoded on a coarser level. While the overall information content of these two representations may be identical, the two architectures would differ in which components of the representation can interact through short-range intracortical connections, which in turn may have substantial consequences for the information processing in this region. Unfortunately, fMRI is strongly biased towards features that are represented at a larger spatial scale (Kriegeskorte and Diedrichsen, 2016). Thus, improvements in the spatial resolution of fMRI combined with more comprehensive spatiotemporal models will be needed to study this aspect of representations 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 neural networks to the representational geometries found in brain areas that are expected to perform similar function (see Chapters by Yamins and Storrs & Kriegeskorte in this volume). 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 actually be able to perform the underlying tasks.

1.7 Acknowledgements

The work was supported by the James S. McDonnell Foundation (Scholar award), a Natural Sciences and Engineering Research Council (Discovery Grant, RGPIN-2016-04890), and the Canada First Research Excellence Fund (BrainsCAN).

References

Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., Shenoy, K.V., 2012. Neural population dynamics during reaching. Nature 487, 51–6.

Clyde, M.A., 1999. Bayesian model averaging and model search strategies, in: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics. Oxford University Press, Oxford, UK, pp. 157–185.

De Angelis, V., De Martino, F., Moerel, M., Santoro, R., Hausfeld, L., Formisano, E., 2017. Cortical processing of pitch: Model-based encoding and decoding of auditory fMRI responses to real-life sounds. Neuroimage. https://doi.org/10.1016/j.neuroimage.2017.11.020

Diedrichsen, J., Kriegeskorte, N., 2017. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS Comput Biol 13, e1005508. https://doi.org/10.1371/journal.pcbi.1005508

Diedrichsen, J., Yokoi, A., Arbuckle, S.A., 2017. Pattern component modeling: A flexible approach for understanding the representational structure of brain activity patterns. Neuroimage. https://doi.org/10.1016/j.neuroimage.2017.08.051

Ejaz, N., Hamada, M., Diedrichsen, J., 2015. Hand use predicts the structure of representations in sensorimotor cortex. Nat Neurosci 18, 1034–40. https://doi.org/10.1038/nn.4038

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

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, 2919–34. https://doi.org/10.1093/cercor/bhw068

Heer, W.A. de, Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. J Neurosci 37, 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017

Huth, A.G., Heer, W.A. de, Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–8. https://doi.org/10.1038/nature17637

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795. https://doi.org/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, 352–5. https://doi.org/10.1038/nature06713

Khaligh-Razavi, S.M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain it cortical representation. PLoS Comput Biol 10, e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kim, M., Jeffery, K.J., Maguire, E.A., 2017. Multivoxel pattern analysis reveals 3D place information in the human hippocampus. J Neurosci 37, 4270–4279. https://doi.org/10.1523/JNEUROSCI.2703-16.2017

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C.E., Kepecs, A., Mainen, Z.F., Qi, X.L., Romo, R., Uchida, N., Machens, C.K., 2016. Demixed principal component analysis of neural population data. Elife 5. https://doi.org/10.7554/eLife.

10989

Kriegeskorte, N., Diedrichsen, J., 2016. Inferring brain-computational mechanisms with models of activity measurements. Proceedings of the Royal Society.

Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: Integrating cognition, computation, and the brain. Trends Cogn Sci 17, 401–12. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci 2, 4. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126–41. https://doi.org/10.1016/j.neuron.2008.10.043

Leo, A., Handjaras, G., Bianchi, M., Marino, H., Gabiccini, M., Guidi, A., Scilingo, E.P., Pietrini, P., Bicchi, A., Santello, M., Ricciardi, E., 2016. A synergy-based hand control is encoded in human motor cortical areas. Elife 5. https://doi.org/10.7554/eLife.13420

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., Just, M.A., 2008. Predicting human brain activity associated with the meanings of nouns. Science 320, 1191–5. https://doi.org/10.1126/science.1152876

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage 56, 400–10. https://doi.org/10.1016/j.neuroimage.2010.07.073

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, 19451–69. https://doi.org/10.1523/JNEUROSCI.2880-13.2013

Norman-Haignere, S., Kanwisher, N.G., McDermott, J.H., 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88, 1281–96. https://doi.org/10.1016/j.neuron.2015.11.035

Poeppel, D., Emmorey, K., Hickok, G., Pylkkanen, L., 2012. Towards a new neurobiology of language. J Neurosci 32, 14125–31. https://doi.org/10.1523/JNEUROSCI.3244-12.2012

Shen, S., Ma, W.J., 2017. Quantifying the importance of guessing, decision noise, and variable precision in explaining behavioral variability in visual perception. BioRxiv. https://doi.org/10.1101/153650

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46, 1004–17. https://doi.org/10.1016/j.neuroimage.2009.03.025

Wu, M.C., David, S.V., Gallant, J.L., 2006. Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci 29, 477–505. https://doi.org/10.1146/annurev.neuro.29.051605.113024

Wu, W., Hatsopoulos, N.G., 2007. Coordinate system representations of movement direction in the premotor cortex. Exp Brain Res 176, 652–7. https://doi.org/10.1007/s00221-006-0818-7

Yokoi, A., Arbuckle, S.A., Diedrichsen, J., 2018. The role of human primary motor cortex in the production of skilled finger sequences. J Neurosci. https://doi.org/10.

1523/JNEUROSCI.2798-17.2017