CHAPTER

15

# Working with high-dimensional feature spaces: the example of voxel-wise encoding models

*Mohammad Babakmehr, Ghislain St-Yves, Thomas Naselaris*

**Department of Neurosciences, Medical University of South Carolina, Charleston, SC, United States**

## 15.1 Introduction

When using machine learning to investigate brain disorders, one often depends on high-dimensional, noisy datasets. For example, in a typical neuroimaging experiment, brain activity is sampled in tens or hundreds of thousands of voxels which may or may not contain useful information. Making sense of such high-dimensional datasets can be extremely challenging. In this chapter, we discuss how to deal with high-dimensional data in the context of machine learning studies of brain disorders. In light of the fact that the vast majority of applications of machine learning to brain disorders have used neuroimaging data, we focus on *voxel-wise encoding models*—a powerful approach that aims to predict activity in single voxels, which is evoked by stimulus or task conditions. In particular, we provide the reader with the intuition and some of technical knowledge required to apply this approach to their own data.

We offer a summary of different classes of methods for analyzing neuroimaging data. A first class of methods relies on the *this-then-that* principle, that is, when *this* changes then *that* activates. Hypotheses are carefully built into the structure of the experiments and successive changes are performed whose effect is monitored. This effect may be behavioral, as in cognitive experiments, or a quantification of physiological responses like a neural spike count, a membrane potential, or a functional magnetic resonance imaging (fMRI) measure of blood oxygen

https://doi.org/10.1016/B978-0-12-815739-8.00015-8

level dependent (BOLD) activity. Many of the examples provided in this chapter focus on the latter. The preferred sensory input of a neuron or bit of brain can be discovered by looking at the sensory stimulus that produced the maximum amount of activity (Churchland, Sejnowski, & Poggio, 2016). This sort of procedure does not require a model; therefore, it generally leaves to be desired in terms of prediction and comparison. A second class of methods are based on the *similar-go-together* principle, that is, if we can distinguish two different conditions (say pathological and healthy) based on some measurement, then there must be a basis for that difference in the measurement. These methods are a clustering approach because they involve finding a way to separate two or more groups in some feature space (Parsons, Haque, & Liu, 2004). Multivariate pattern analysis (Detre et al., 2006) falls in this category. The main defect of this approach is that it fails to attribute proper importance to the causes of the success of the classification (Naselaris et al., 2015). Finally, the class of methods that we focus on here resolves the ambiguities of the previous two methods. This class of method is based on the principle that *prediction entails understanding* (to some degree at least). As we will see, these methods require that one builds an explicit set of assumptions to learn to make accurate predictions about a measure of interest. Furthermore, as prediction is quantitative in nature, these methods allow for a greater degree of cross-model comparison and hypothesis testing.

## 15.2 Voxel-wise encoding modeling

In principle, a voxel-wise encoding model is defined by the characterization of three components: (1) a *model class*, (2) a *noise distribution* or, more directly, a *loss function*, and (3) a *training procedure*, which include *regularization* which is a reflection of a *prior distribution over parameters* (Wu, Gao, Bienenstock, Donoghue, & Black, 2006). In practice, however, the model class and training procedure is heavily restricted by the type of data available. For encoding models of fMRI BOLD voxel responses, the typical number of stimulus—response pairs, with repeats, that can be acquired in a typical experiment is limited to an order of $10^3$—$10^4$ for a range of $10^4$—$10^6$ voxels. For example, a recent effort at acquiring a large dataset of image—response pairs managed to collect the responses to 5000 natural images (Chang, Pyles, Gupta, Tarr, & Aminoff, 2018) ("natural images" are pictures of things one may see during normal exploration of our visual world—like a cat sunning on a patio or a dog barking at the moon). Working in this relatively undersampled regime means that, in practice, the main learning procedure workhorse is close to *linear regression* and that the encoding model class takes the form

$$r = w^T \varphi(s) + noise \tag{15.1}$$

where $r$ is a vector whose entries describe the magnitude of the responses to each units. $\varphi(s)$ is some nonlinear mapping from the stimulus vector $s$ to some adequate feature space (more on that below) and $w$ is a matrix of feature weights which specifies the extent to which each feature contribute to each unit activation (recall that $w^T$ means the *transpose* of $w$, which is just $w$ with the rows and the columns swapped). Finally, *noise* is the component of activity that cannot be related to the stimulus. It is very common to assume Gaussian noise, which directly implies that the loss function, the scalar function that quantifies the model goodness of fit, ought to be of the form

$$L = \left| r - w^T \varphi(s) \right|^2 \tag{15.2}$$

where we discarded the potential contribution of the covariance matrix of the noise for simplicity (this is also often done in practice). This expression essentially measures the discrepancy between the measurement $r$ and the model prediction $w^T \varphi(s)$. Note that, in virtue of the linearity of the last operation of the model, which shares all the same causes $\varphi(s)$ relating to the stimulus, the encoding model can be separated into as many independent units (or voxel-wise models in the case of fMRI BOLD activity) as there are elements in $r$. This has extremely important consequences for our ability to model the large number of voxels in a typical fMRI experiment. Fig. 15.1 provides some intuition for the previous formal description. As shown in Fig. 15.1, we can describe activity in the brain as a point in an *activity space*. The axes correspond—in the case of fMRI—to individual voxels. If voxel 1 is strongly activated, and voxel 2 is only weakly activated, the point characterizing the brain's activity will be located far from the origin along the voxel 1 axis and close to the origin along the voxel 2 axis. We can also characterize the stimulus or behavior we are interested in as a point in a distinct *input space*. Here, the axes correspond to the sensory or behavioral states that vary during the experiment. In Fig. 15.1, the input space is a *pixel space*, where each axis corresponds to the brightness of one pixel, and any image can be represented as a point in this space.

The statement that the main workhorse of encoding model is linear regression might seem unintuitive because the brain is necessarily a highly nonlinear system. We know that the brain has to be a nonlinear system because the majority of interesting cognitive processes impose a *nonlinear* relationship between stimuli and behavior. For example, you are able to identify the face of someone you know at any location in your visual field, at any scale, in well- or poorly lit environments, with a wig and fake mustache, etc. Each of these transformations radically alters the retinal input to your visual system (the image of the familiar face) without affecting the behavior (recognition of the face). Thus, the parts of the brain that enable face recognition *must* perform some kind of nonlinear computation between retinal input and cognitive or behavioral output. How could then *linear* methods for analyzing brain data reveal anything about how or where face identification is happening?

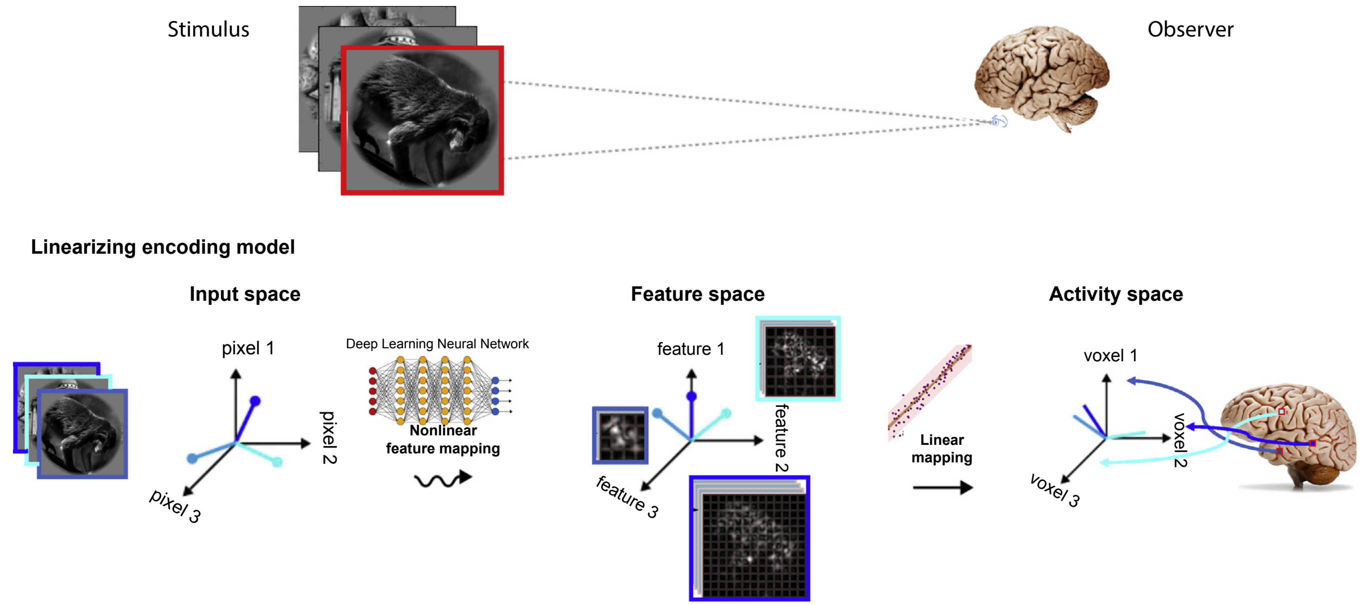

FIGURE 15.1 The feature space metaphor. The data analysis techniques discussed here can be viewed as methods for discovering *linear* relationships between *nonlinear* cognitive processes and brain activity. The figure shows that the data analysis techniques can all be thought of as different ways of mapping points in an abstract feature space to activity in the brain (or vice versa). Experimental stimuli used to indicate task conditions are represented by an input space (left). In experiments using visual stimuli, the axes of the input space are pixels and each point in the space represents a different image. Here, we show a 3D space constructed by considering just three of the pixels in the natural images shown. Each image will have a potentially unique position in the pixel space, as illustrated by the three vectors with different lengths and orientations. Brain activity measured in each voxel is represented by an activity space (right). The axes of the activity space correspond to different voxels and each point in the space represents a unique pattern of activity across voxels (different colors in the activity space). In between the input and activity spaces is a feature space (middle). The mapping between the input space and the feature space is *nonlinear.* For example, the nonlinear mapping might be implemented as a deep neural network (small diagram above nonlinear mapping *arrow*). The mapping between the feature space and activity space is *linear* and can be estimated by an appropriately regularized linear regression (*pink line [dark gary line in print version]* above linear mapping *arrow*). The nonlinear feature mapping serves as a model of the cognitive process of interest. Example feature maps corresponding to the "monkey" image are shown for each the three features axes. *Adapted from Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity.* Neuron, *63(6), 902–915.*

It would in fact be very difficult to gain an understanding of face identity by measuring activity in the retina because the relationship between the retinal input and the identify of faces is, as we just pointed out, nonlinear. A more direct approach is to identify one or more brain regions (defined as any collection of units from which activity can be measured) where the relationship between face identity and brain activity is *linear*. That is, if one can find a spot in the brain where the face identity is encoded, the activity in that region of the brain should actually be *linearly* related to the identity of faces. By linearly related we mean, for example, that the identification of a face could be "read-out" from the activity in this region simply by determining if the activity patterns evoked by images of any two faces can be cleanly separated by a plane. Another (more formal) way to state it is that if we let $r$ be the activity measured in this region in the brain and $\varphi$ be a simple binary code for identifying faces (i.e., "John" = (1,0,0), "Mary" = (0,1,0), "Joe" = (0, 0, 1)), then

$$r = w^T \varphi + noise \tag{15.3}$$

Note the subtle but important difference between Eqs. (15.1) and (15.3). In Eq. (15.1), we encoded the stimulus through some nonlinear mapping, $\varphi(s)$, which we could choose to perform face identification (assuming we can, as it is a difficult problem). On the other hand, in Eq. (15.3), we are providing explicitly the encoded face identity has an input space.

Therefore, whether one formulates the encoding model with Eq. (15.1) or (15.3), one has to ask the question: *what sort of feature would this or that little bit of brain be interested in*? This is the essential way in which encoding models allow for discrete hypothesis testing. The hypotheses take the form of the mapping $\varphi(s)$ from pixel space to feature space or that the representations $\varphi$ are represented *somewhere* in the activity of the brain. To find out if and where that *somewhere* is, we fit the model and test it for validity (more on that later).

### 15.2.1 Encoding models from the Gabor feature space to deep learning networks

The problem laid out by encoding models is always one of prediction. How well do encoding model generalize to new examples? How well can we decode novel activity patterns, i.e., can we predict seen images based on modeled activity of object categories not used to train the encoding model?

To answer these questions, Kay, David, Prenger, Hansen, and Gallant, 2008; Kay, Naselaris, Prenger, and Gallant, 2008 built an encoding model based on a Gabor wavelet feature space. Gabor wavelets have a very interesting computational interpretation because it can be shown that they form a maximally informative basis set of features for efficient image encoding (Lee, 1996) that emerge naturally out of learning a sparse representation of natural images (Olshausen & Field, 1996). The brain has a difficult job to do, namely, represent everything we need to know about the very complex world around using the vast but finite resources of our brains. Therefore, it was both surprising and not at all surprising that the brain would have had to evolve such a clever, efficient scheme for encoding all of this information. It has been observed that individual simple cell neurons respond to phase-encoded wavelets, but whole populations of neurons tend to lose this phase dependence and respond preferentially to phase-independent Gabors wavelet (David & Gallant, 2005) like complex cells. Phase independence is therefore a nonlinear operation.

Gabor wavelets are filters that extract specific low-level features from a local region of a visual scene. These low-level visual features include retinotopic location, edge orientation, and spatial frequency. By filtering an image through a bank of Gabor wavelets that vary in spatial location and characteristic orientation and frequency, we obtain a description of the scene in terms of these low-level or "structural" (Naselaris, Prenger, Kay, Oliver, & Gallant, 2009) features. If this particular low-level description of scenes is encoded in activity in the human early visual cortex, we should be able to use it to predict the activity that will be evoked in early visual cortex when a human subject looks at *anything*. Kay, David, et al. (2008); Kay, Naselaris, et al. (2008) presented an experiment to test this prediction. In this experiment, subjects passively fixated a stream of briefly (~4 s) flashed photographs (nearly 1500 over the course of the entire experiment). The photographs were randomly selected pictures of the world (known in the visual neuroscience literature as "natural scenes"). While subjects observed the photographs, BOLD activity was measured in voxels located near the posterior pole, including areas V1, V2, V3, V4, LO, and MT+. The runs (a sequence of back-to-back conditions) were split into training and testing sets. The training runs were then used to perform the voxel-wise modeling part of voxel-wise encoding modeling.

The encoding model in the Kay, David, et al. (2008); Kay, Naselaris, et al. (2008) study was built from a set of hundreds of Gabor wavelets. To generate a predicted response using the Gabor wavelet encoding model, the input stimulus was filtered by each wavelet in the set. Each wavelet came in a pair of orthogonal phase-shifted wavelet. The square root of the sum of the square of the responses to both wavelet produces the feature space $\varphi(s)$ used by the encoding model. As previously described, the predicted response is a weighted sum of outputs of this fixed nonlinearity. For each voxel, there is one parameter per Gabor wavelet. Thus, the model parameters capture something about the specific features that drive activity measured in that particular voxel. For example, if for one particular voxel, the model parameters assigned to Gabors with a vertical orientation are all large and positive, while the model parameters assigned to Gabors with all other orientations are small, then the activity in that particular voxel robustly encodes the presence of vertically oriented edges in a scene. Inferring the model parameters from the training data is a process called "model fitting."

This Gabor-based model was found to accurately predict early visual area and generalize to decoding unseen images, which indicates that these visual features are robustly encoded in the early visual area and are systematically activated for *any* natural image. Furthermore, because natural scenes were used in the series of studies beginning with Kay, David, et al. (2008); Kay, Naselaris, et al. (2008), it should be possible to test other competing models that were not based on Gabor wavelets, but on other feature spaces. Here is where the hypotheses testing enters the picture. Using the same linear machinery, we can build alternative encoding models that use very different nonlinear feature spaces. If an alternate feature space leads to better model predictions, it is a better model of representation in the brain area under study.

Currently in visual neuroscience, a very compelling family of feature spaces are those extracted from the hidden layers of a deep learning (DL) network that has been trained to categorize objects. DL is a new trend in the area of pattern recognition that has been widely used to address complicated visual processing tasks, such as human-level object recognition, image segmentation, and image captioning (LeCun, Bengio, & Hinton, 2015). In its most prevalent form, DL uses a deep convolutional architecture which produces many feature maps of different resolutions as an intermediary step toward fulfilling whatever computational task they are designed to perform. A feature map is a spatially organized set of features that share something in common (like an oriented edge at different position. In fact, the Gabor features of Kay, David, et al. (2008); Kay, Naselaris, et al. (2008) could be organized into such feature maps: one map for all Gabors sharing the same orientation, envelope, and frequency). Interestingly, DL networks trained to recognize objects learn many qualitatively different features across their processing hierarchy, from Gabor-like representations at the lowest level to semantic representations (meaning words and concepts) at the highest (see Chapter 10 for a detail explanation of convolutional neural networks). Several recent studies have compared the Gabor model to an encoding model based on the feature maps of a DL network (Güçlü & van Gerven, 2015; Kriegeskorte, 2015; St-Yves & Naselaris, 2018; Wen, Shi, Chen, & Liu, 2018a, 2018b; Wen et al., 2018a, 2018b). In early visual areas, it is difficult to beat the Gabor model. On the other hand, in high-level visual areas, the more abstract representations provided by DL networks provide a much better set of features. Some results from a recent paper (St-Yves & Naselaris, 2018) comparing the two models in terms of their prediction are shown in Fig. 15.2.

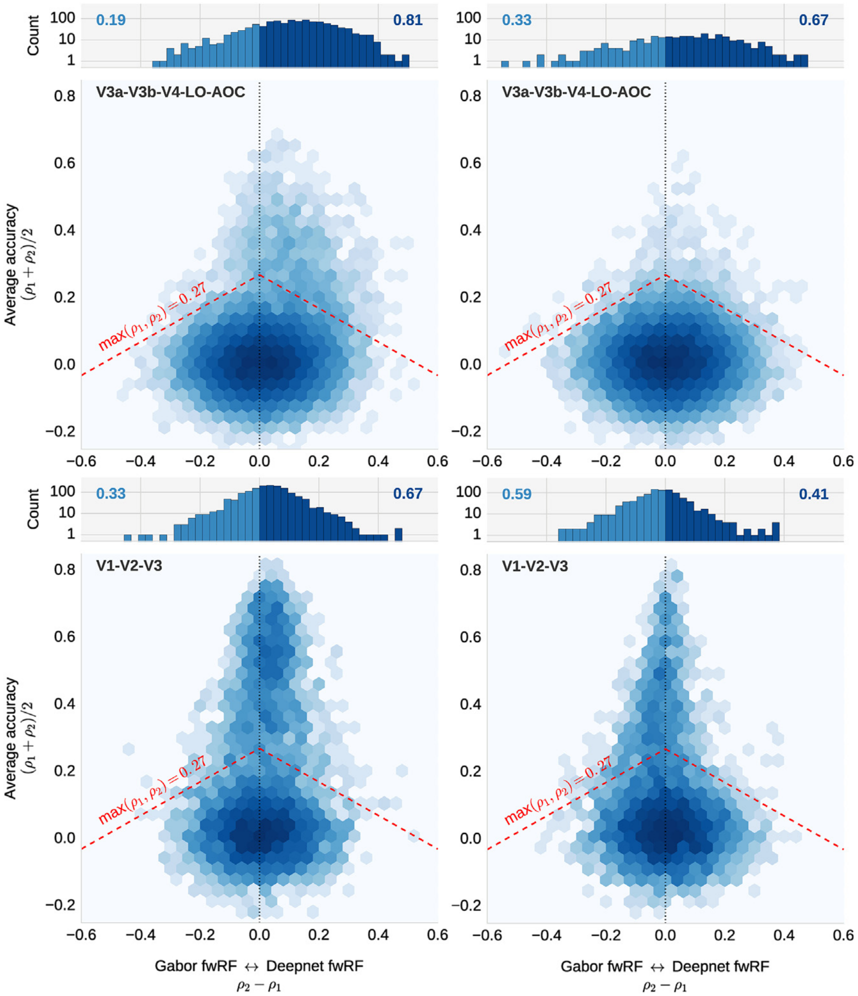

FIGURE 15.2 Comparison of the Gabor-fwRF and Deepnet-fwRF models. This plot provides an example of how encoding model comparisons can be used to test hypotheses about the representations underlying distinct brain areas. The encoding models considered here are the "Gabor-fwRF" encoding model that uses a Gabor wavelet transform as its feature space, and the "Deepnet-fwRF" encoding model that uses the feature maps of a deep neural network trained to classify objects in the foreground of natural scenes. These different features spaces imply very different hypotheses about the functioning of the visual system. These plots reveal advantages of one model over the other at explaining variance in a set of BOLD measurements. In this case, the BOLD measurements are from voxels in low-level visual areas (areas V1, V2, and V3, lower row) and high-level visual areas (areas V3a/b, the lateral occipital complex, and the anterior occipital cortex). The position along the *vertical axis* indicates the average prediction accuracy for the models under comparison; shifts to the right or left along the *horizontal axis* indicated a relative improvement in prediction accuracy for one of the models (model 1 is presented to the left of model 2). The color of each hexagonal bin indicates the number of voxels in a local region of the plot (log scaled). The histogram at the top of each plot represents the distribution of relative improvements for all voxels whose prediction accuracy is above 0.27

These results demonstrate the utility of encoding model as a hypothesis testing tool. In the case of the DL, a single model can be applied to analyze responses across the entire visual system and then easily compared with other alternatives. Importantly, this kind of system-wide analysis is being applied not just in vision but in other sensory and cognitive domains (Breedlove, St-Yves, Olman, & Naselaris, 2018, p. 462226; Kell, Yamins, Shook, Norman-Haignere, & McDermott, 2018).

The comparison of the models above relies on a measure of prediction accuracy. To measure prediction accuracy for a voxel-wise encoding model, some portion of the data is held out from the training set. After model fitting, the model for each voxel is used to generate predictions that can be correlated against measured activity (or estimated response amplitudes) in the held-out data. Note that the signals being predicted in this case are inherently noisy. In testing encoding models, it is therefore common to use data from dedicated validation runs in which individual stimuli or sequences of stimuli are repeated. This makes it possible to average out the noise in the responses to a specific stimulus. Of course, when we speak of voxel-wise encoding models, the words "applying" and "fitting" elide a considerable amount of work and technical knowledge. We now turn to a discussion of how these models are actually "applied" and "fitted."

## 15.2.2 Fitting encoding model: training regularization

Suppose we have measured a pattern of activity $r = [r_1, r_2, \ldots, r_M]$, in response to the presentation of a stream of photographs or a movie, $s = [s_1, s_2, \ldots, s_M]$, that we will call the *stimulus*. The data include voxels from the entire visual system—from the posterior occipital lobe up through anterior temporal cortex. We would like to infer a mathematical model from a subset of training pair $\{(s_i, r_i); i = 1 : M\}$, such that the model would be able to predict the associated activity $r_{test} = [r_{M+1}, r_{M+2}, \ldots, r_N]$, from another subset of stimulus, $s_{test} = [s_{M+1}, s_{M+2}, \ldots, s_N]$, named test

($P < .001$, randomization test) for at least one of the two models, which correspond graphically to all voxels above the red *dashed line*. The number on each side represents the fraction of voxels that are improved under that model. In the plots, a shift in the data toward the left indicates an advantage for Gabor-fwRF model while a shift of the data toward the *midline* indicates an advantage for the Deepnet-fwRF. This plot shows that, for intermediate brain areas, the Deepnet-fwRF outperforms the Gabor-fwRF models. For early visual areas, the Deepnet-fwRF weakly outperforms the Gabor-fwRF. The Deepnet-fwRF thus seems to have the strongest overall advantage for brain areas that require complex feature spaces. *Adapted from* St-Yves G. and Naselaris T., The feature-weighted receptive field: An interpretable encoding model for complex feature spaces, NeuroImage 180, 2018, 188–202.

dataset. A voxel-wise encoding model for $K$ generic features extracted from the image $x_i$ takes the form:

$$\widehat{r}_i = w^T \varphi(s_i) = \sum_k w_k \varphi^k(s_i) \tag{15.4}$$

where $w \in R^K$ is a weight vector where each element $w_k$ represents the contribution of each extracted feature in the overall activity at each sample of time. Most methods for fitting the model parameters are based on, or at least inspired by, the principle of likelihood maximization. The idea is that we want to find the model parameters that result in the highest likelihood for the actual observed signal. As stated above, in the case where the noise term is assumed to be Gaussian, maximum likelihood is equivalent to finding the $w$ that minimizes the square of the distance between the model's prediction $w^T\varphi(s_i)$ and the observed responses $r_i$.

The optimal solution for $w$ in this case is well known and can be found in any textbook covering multivariate linear regression (Bishop, 2006); however, our experience has been that for all but the most low-dimensional models the straightforward unregularized regression approach usually works very poorly and some regularization is necessary to find solutions that generalize even when the number of model parameters exceed the number of data samples. Ridge regression can be an effective regularization method but introduce an additional free parameter to the model. One particularly effective solution that has been employed in a number of publications (Naselaris et al., 2009, 2012) is called *coordinate descent with early stopping*. This is an iterative algorithm that gradually approximates the optimal (in the least-squares sense) value of $w$ by incrementally modifying one parameter at a time. Starting from initial values of zero, we increment the single model parameter, $w_k$, that results in the greatest reduction in cost (as measured using the training data). After each increment, we measure the cost on a held-out early stopping set. After an undetermined number of iterations, the loss on the held-out set will begin to rise and the algorithm halts, thus preventing overfitting on the training set. It can be shown that early stopping procedure equivalent result to an optimally tuned ridge regression without additional parameter (Raskutti, Wainwright, & Yu, 2014). We have found this to be an extremely robust algorithm for fitting high-dimensional models.

### 15.2.3 Fitting encoding models: structural regularization

In an ideal world, it would not be necessary to test hypotheses by fitting independent encoding models and comparing their predictions accuracies. Instead,[1] one would just fit a single encoding model that

[1] In a really ideal world, we could spend days of leisure instead of fitting encoding models.

contained all possible feature spaces and let the algorithm for selecting the weights $w$ sort out which feature space worked best as a predictor of neural response. To get close to this ideal requires fitting models with many parameters—far more, in fact, than the number of data samples, it is feasible to obtain in most neuroimaging experiments. If we want to make progress toward this ideal, then "vanilla" regression of the kind described above won't do.

Consider, for example, the DL-based model above. Fitting such a model for a single voxel requires regressing the feature maps of every layer in a DL network onto measured activity. The DL network will typically have $O(10)$ (on the order of 10) layers containing $O(100) - O(1K)$ feature maps containing $O(100)$ pixels. A model that assigned a single weight to each pixel of each feature map would have $O(100K)$ to $O(1M)$ regression parameters. How to fit such a high-dimensional regression model with $O(1K)$ data samples?

One modeling trick that can greatly reduce the number of parameters needed to fit a model is to factorize the spatial and feature dimension of the feature maps. This *space-feature separability* is possible because the cortex is highly structured in a very special way, which our models ought to take advantage of. In St-Yves & Naselaris (2018), the parameters that govern a voxel's receptive field (i.e., the region of visual space in which stimuli drive variance in the voxel's measured BOLD response) were separated from the parameters the govern its tuning to features. The parameters needed to specify the receptive field $\theta$ for each voxel have to be learned alongside the parameters that specify feature tuning, forming a set of parameter much smaller than the nonfactorized set. This encoding approach is termed the *feature-weighted receptive field* (fwRF) model. In the case of St-Yves & Naselaris (2018), there were three receptive field parameters $\theta = (\mu_x, \mu_y, \sigma)$ describing the position and size of the receptive field and one feature weight per feature map, meaning that the number of parameters was reduced by the number of pixels-per-feature map, i.e., an $O(1K)$ reduction. The predictive expression for this model was

$$\widehat{r}_i = \sum_k^K w_k \sum_i^I \sum_j^J g(x_i, y_j; \theta) \psi_{ij}^k(s_i) \tag{15.5}$$

where the function $g(x_i, y_j; \theta)$ is the receptive field, and the "parameters" are feature maps. The model thus assumes that activity measured in a single voxel will not encode distinct features at distinct locations, but rather a weighted combination of features at a *single* location. A graphical description of the model is shown in Fig. 15.3.

However, the receptive field function introduces an additional difficulty, as it is a nonlinear function of its parameters. To minimize the loss between the model prediction and the observed activity, the feature

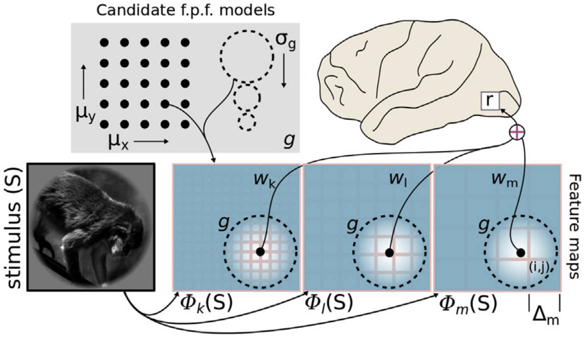

**FIGURE 15.3** The fwRF model. (A) A schematic illustration of a fwRF for a single voxel (*gray box* on brain, top right). The fwRF predicts the brain activity measured in the voxel, *r*, in response to any visual stimulus, *S* (bottom left). The stimulus is transformed into one or more feature maps (three feature maps, $\varphi_k$, $\varphi_l$, and $\varphi_m$, are shown in blue with pink [gray in print version] borders). Each feature map is filtered by a 2D Gaussian feature pooling field, *g*, that is sampled from a grid of candidate feature pooling fields (*gray box* at top left; candidate feature pooling field centers ($\mu_x$, $\mu_y$) are illustrated by the grid of *black points*, while candidate feature pooling field radii, $\sigma_g$, are illustrated by *dashed circles*. The output of the feature pooling field filtering operation (illustrated as *black dots* in the center of the dashed feature pooling fields on each feature map) for each feature map is then weighted by a feature weight (*black curves* labeled $w_k$, $w_l$, $w_m$). These weighted outputs are summed to produce a prediction of the activity *r*. *Adapted from* St-Yves G. and Naselaris T., The feature-weighted receptive field: An interpretable encoding model for complex feature spaces, NeuroImage 180, 2018, 188–202.

weights *w* are optimized through early stopping gradient descent over a range of possible receptive field parameters in parallel. This produces several estimates of loss and the minimal value is selected over the possible receptive field candidate. This is known as grid search, the simplest and most basic form of optimization (for details, see St-Yves & Naselaris, 2018).

## 15.3 Applications to brain disorders

A major challenge in applying fMRI to the understanding of brain disorder is the dearth of reliable single-subject modeling techniques. To study brain disorder using fMRI, researchers typically align MR images from multiple subjects who have been diagnosed with a common disorder (or designated as controls) and then averaging their functional responses. Although this approach is undoubtedly useful for identifying systematic difference between pathological cases and control cases, from the point of view of an individual, combining brains in this way compromises more signal than noise. A major hope in developing voxel-wise encoding is to design unique patient-dependent brain models for each subject. All of the results in the examples above are single subject: no aligning of brains is necessary. Performing such analyses requires a shift in experimental design that requires more sampling per subject. But the potential for payoff in terms of understanding individual brains seems quite high. However, it is not entirely clear how the study of *representations* translates into quantitative or qualitative assessment of *functions*.

Mental imagery would be another potential field for encoding models. Mental imagery has started to receive attention from a quantitative predictive modeling perspective (Breedlove et al., 2018; Cichy, Heinzle, & Haynes, 2011; Jug, Kolenik, Ofner, & Farkaš, 2018). The most recent work, Breedlove et al. (2018), demonstrates that the voxel-wise encoding models developed to model the cortical activity evoked by visual stimulus can be used to model activity evoked from merely imagining the stimulus. The resulting model shows idiosyncratic difference from the encoding model of a related vision experiment and offers a new window into our internal mental image distortions. These models can play a critical role in developing and testing hypotheses about the neuronal basis of brain disorders in individual subjects, while providing insights into possible treatment strategies (Pearson, Naselaris, Holmes, & Kosslyn, 2015).

Classification across experimental conditions and across subjects, or classes of subjects, could also in principle be performed using encoding model predictions. This method would have an advantage in that the basis for the difference would be well-grounded in a hypothesis-driven and quantitative framework, rather than on potentially unknown idiosyncratic brain differences. Thus, model-based approaches have an obvious advantage for their capacity to relate the effect (e.g., classification into one of several categories) with the underlying representational basis for that difference in the form of the hypotheses embedded in the feature space of the model.

## 15.4 Conclusion

We have detailed a framework for voxel-wise encoding modeling that provides quantitative generalizable predictions and can serve as a hypothesis testing tool for the representational basis that best describes neural activity in various brain states. An encoding model explains brain activity on the basis of various features of a well-controlled stimulus. This is why, in the existing literature, such models have been applied most successfully in the context of vision, audition, and mental imagery. In the context of brain disorders, activity in high-level brain areas tends to be resistant to modeling using features derived from efficient coding hypotheses (Gabor-like). However, in recent years, significant progress has been made in the development of voxel-wise encoding models with the advent of deep neural networks, allowing the extraction of new sets of task-driven features that are well-suited to explain the activity in high-level brain areas.

One remaining question concerns the difficulties that arise in testing different theories regarding mechanisms that lead to similar representations. For example, the absence or presence of a representation in a certain condition does not indicate the mechanism by which this representation may vanish or originate. The mechanism that the encoding model uses to produce the representation may be very different to the one that the brain uses. The Occam's razor principle—which states that, when presented with competing hypotheses, one should select the simplest solution—will be useful here. Finally, moving from static to dynamic representations is the current frontier of encoding models and is an effort that will require combining different measurement techniques with new way of building constrained encoding models.

## 15.5 Key points

- A voxel-wise encoding model is defined by the characterization: (1) a *model class*, (2) a *loss function*, and (3) a *training procedure.*
- In voxel-wise encoding modeling, the encoding model can be separated into a set of features, extracted from the external stimulus, and linearly weighted to estimate the neural activity recorded from a neural unit.
- Regularization, whether through the training procedure or structural regularization of the model class, is an important strategy for obtaining good quantitative generalizable predictions that are easily interpretable.
- Voxel-wise modeling provides a possible framework for quantitative hypothesis testing.
- Hypothesis testing and the generation of interpretable findings may help shed light on the neuronal basis of brain disorders and provide insights into possible treatment strategies.

## References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Breedlove, J. L., St-Yves, G., Olman, C. A., & Naselaris, T. (2018). *Human brain activity during mental imagery exhibits signatures of inference in a hierarchical generative model*. bioRxiv.

Chang, N., Pyles, J. A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2018). *BOLD5000: A public fMRI dataset of 5000 images*. arXiv preprint arXiv:1809.01281.

Churchland, P. S., Sejnowski, T. J., & Poggio, T. A. (2016). *The computational brain*. MIT press.

Cichy, R. M., Heinzle, J., & Haynes, J. D. (2011). Imagery and perception share cortical representations of content and location. *Cerebral Cortex, 22*(2), 372–380.

David, S. V., & Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems, 16*(2—3), 239—260.

Detre, G., Polyn, S. M., Moore, C. D., Natu, V. S., Singer, B., Cohen, J. D., et al. (June 2006). The multi-voxel pattern analysis (MVPA) toolbox. In *Poster presented at the annual Meeting of the organization for human brain mapping*.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience, 35*(27), 10005—10014.

Jug, J., Kolenik, T., Ofner, A., & Farkaš, I. (2018). Computational model of enactive visuospatial mental imagery using saccadic perceptual actions. *Cognitive Systems Research, 49*, 157—177.

Kay, K. N., David, S. V., Prenger, R. J., Hansen, K. A., & Gallant, J. L. (2008). Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. *Human Brain Mapping, 29*(2), 142—156.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452*(7185), 352.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron, 98*, 630—644.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417—446.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

Lee, M. S. (1996). *U.S. Patent No. 5,546,129*. Washington, DC: U.S. Patent and Trademark Office.

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage, 105*, 215—228.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron, 63*(6), 902—915.

Naselaris, T., Stansbury, D., & Gallant, J. L. (2012). Cortical representation of animate and inanimate objects in complex natural scenes. *Journal of Physiology Paris, 106*, 239—249.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607.

Parsons, L., Haque, E., & Liu, H. (2004). Subspace clustering for high dimensional data: A review. *Acm Sigkdd Explorations Newsletter, 6*(1), 90—105.

Pearson, J., Naselaris, T., Holmes, E. A., & Kosslyn, S. M. (2015). Mental imagery: Functional mechanisms and clinical applications. *Trends in Cognitive Sciences, 19*(10), 590—602.

Raskutti, G., Wainwright, M. J., & Yu, B. (2014). Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research, 15*(1), 335—366.

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage, 180*, 188—202.

Wen, H., Shi, J., Chen, W., & Liu, Z. (2018a). Deep residual network predicts cortical representation and organization of visual features for rapid categorization. *Scientific Reports, 8*(1), 3752.

Wen, H., Shi, J., Chen, W., & Liu, Z. (2018b). Transferring and generalizing deep-learning-based neural encoding models across subjects. *NeuroImage, 176*, 152—163.

Wu, W., Gao, Y., Bienenstock, E., Donoghue, J. P., & Black, M. J. (2006). Bayesian population decoding of motor cortical activity using a Kalman filter. *Neural Computation, 18*(1), 80—118.

## Further reading

Chen, M., Han, J., Hu, X., Jiang, X., Guo, L., & Liu, T. (2014). Survey of encoding and decoding of visual stimulus via FMRI: An image analysis perspective. *Brain Imaging and Behavior, 8*(1), 7−23.

Kay, K. N., & Gallant, J. L. (2009). I can see what you see. *Nature Neuroscience, 12*(3), 245.

Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2013). GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Frontiers in Neuroscience, 7*, 247.

Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron, 79*(5), 1025−1034.