Interpreting fMRI Decoding Weights: Additional Considerations

P.K. Douglas[*]
Modeling and Simulation Department,UCF
Department of Psychiatry and Biobehavioral Medicine, UCLA
Los Angeles, CA 90024
pdouglas@ist.ucf.edu

Ariana Anderson
Department of Psychiatry and Biobehavioral Medicine
UCLA, CA, USA

Abstract

Ideally, decoding models could be used for the dual purpose of prediction and neuroscientific knowledge gain. However, interpreting even linear decoding models beyond classification accuracies is difficult. Haufe and colleagues suggested projecting feature weights onto activation maps may enhance interpretability. Here we show that redundancy and noise, typically found in fMRI data, can further complicate interpretation – even after projection onto activation maps. In the presence of non-Gaussian noise seen with artifacts such as scanner drift, motion, and periodic biorhythms, projecting feature weights onto activation maps yields spurious patterns; for example, noise typically associated with scanner drift can cause the activation weights of noise sources to be greater than those of task-associated signals. These analyses suggest that certain decoding or backward models are inherently limited by their approaches to successfully model the structure of incoming noise, and that improper assumptions of this noise structure yields misleading interpretive maps.

1 Introduction

Encoding and decoding are the two most popular approaches used to study functional MRI (fMRI) neuroimaging data. Both methods have complementary goals, yet anti-correlated analysis pathways. Encoding models are referred to as forward models; these begin with the stimulus and attempt to understand how measured responses vary due to variation in the stimuli. Decoding models, conversely, are backward models, which start with the responses or voxel measurements and attempt to see how much can be learned about the stimulus from response patterns.

Encoding methods have a longer history of application; these classically adopt a mass univariate approach, where each voxel measurement is treated independently at the first level of analysis. Spatial dependencies are introduced later during inference through random field theory (Worsley et al. 1996). In typical encoding studies, an experimental variable or stimulus is manipulated, a general linear model is applied along with spatial smoothing, and the result is a map of functional activation in the brain.

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Encoding models suffered from two key issues related to interpretation and analysis. First, many experimentalists would infer that a particular cognitive process was engaged on the basis of activation in a particular brain region. However this type of reverse inference is only deductively valid if a particular brain region is activated exclusively during a particular cognitive process (Poldrack 2006), (see Figure 1, left column). Secondly, the underlying neural code was thought to be distributed and the relative modularity remained unclear. For example, the fusiform face area was previously thought to be a homogeneous module showing strong activation preference for faces (e.g. Kanwisher and McDermott 1997). However Cukur et al. (2014) later demonstrated that at least three different distributed subpopulations with unique tuning profiles exist across the FFA. Therefore analyzing each voxel independently coupled with random field theory would potentially miss distributed patterns of activity that were related to stimulus categories.

Decoding models provided some degree of a solution to these issues. Decoding models use machine learning techniques to find patterns of voxels that collectively discriminate between experimental conditions. The allure of decoding methods rested on the notion that greater insight into the patchy functional segregation in the brain may be achieved by analyzing voxel measurements in combination as opposed to individually (Friston 2009). Secondly, the multivariate nature of decoding approaches makes it possible to cancel out noise, thereby improving sensitivity (e.g., Blankertz et al. 2011). Decoding studies grew in popularity as they demonstrated the capability to discriminate stimulus categories that had previously eluded general linear model analysis (Kamitani Tong 2005; Haynes et al. 2007; etc.). Classification techniques ranging from decision trees to deep neural networks have been applied to neuroimaging data. However, by far the most widely adopted practice is to assume that the data from different categories can be shattered linearly with a hyperplane.

However, there are also issues with the interpretation of decoding analyses (see Figure 3 , right column). The process of training and testing a classifier furnishes a set of feature weights, typically called the *w* vector, whose elements collectively influence how the data are projected into this space, along with predictive accuracies. In numerous decoding studies, the feature weights were mapped onto brain data directly. However, the *w* vector is orthogonal to the decision boundary and from this geometric perspective, it is clear that feature weights should not be interpreted in a direct manner (e.g., Guyon Elisseeff 2002; Haufe et al. 2014). In this sense, presenting feature weights onto brain maps may have led to an implied interpretation that these images corresponded to meaningful representational patterns in the neural code. Furthermore, the presence of stimulus information in a brain region does not imply that this information serves the function of representing the stimulus in the context of the brain's overall operation (Kriegeskorte 2011). Therefore, the analogous issue of reverse inference also exists within the decoding model approach.

2 Activation Maps

Several methods have been developed for "reading out" which voxels are contributing the most to classifier performance (Norman et a. 2006). One issue with interpreting backward models or decoding results is that significant weight may be observed in spatial locations whose measurements are statistically independent from the brain process under study. Haufe et al. (2014) highlighted this issue as follows:

Suppose we have two measurement channels, $\tilde{x}_1(n)$ and $\tilde{x}_2(n)$, and we are trying to discriminate between two tasks. The first channel is informative, and the class means differ. In the second channel, the class means are equal, and therefore this measurement or voxel does not contain class-specific information. However, it is possible for the Bayes-optimal classification according to a linear discriminant analysis (LDA) to assign a considerably stronger weight to the noise channel (see Figure 2, top row middle column).

Their proposed remedy for the linear case is as follows. In the backward model case, the goal is to extract latent factors $\tilde{s}(n)$ as functions of the observed data, $\tilde{x}(n)$. In the linear case, the mapping from observations to factors can be summarized by an N x M transformation matrix, $W \in^{MxK}$, and the backward model can be described by:

$$W^T \tilde{x}(n) = \tilde{s}(n) \tag{1}$$

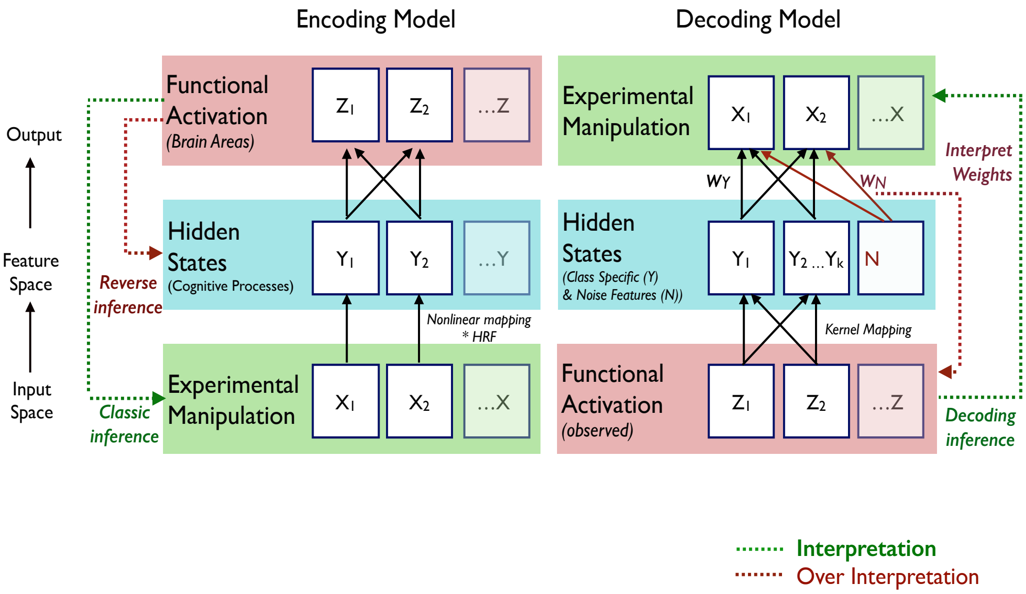

Figure 1: A graph representing the relationship between experimental manipulations, congitive processes, and observed variables. In the left column, the classic analysis flow of an encoding or forward model is shown. At the input level, an experimental manipulation is made which induces hidden cognitive processes to take place. Measurements are then made, which are typically analyzed using a general linear model. Infering that these activation patterns are related to the stimulus is a classic inference, however infering that activated regions implies that a certain cognitive process is taking place consistutes reverse inference. In the right column, a backward model or decoding model analysis is shown. At the input level, here, is the functional activations measurements, and the stimulus or predicted experimental manipulation is at the output level. Interpreting the feature weights directly and in isolation is incorrect, since this approach is inherently multivariate.

In order to read out the weights as meaningful, one must construct activation maps, *A(n)*, from the extraction filters. In the square case where K=M, we simply multiply by W^{-T}, as follows:

$$X(n) = W^{-T}\tilde{s}(n), \tag{2}$$

And our activation patterns become:

$$A(n) = W^{-T} \tag{3}$$

Although this approach of projecting feature weights or extraction filters onto activation patterns is useful (Haufe et al. 2014), read out from this process remains complex. Here we show that different types of noise to be expected in functional MRI measurements, can contaminate the ability to recover information even in an the extremely simple case: a linear discriminant analysis (LDA) classifier with two features. Redundancy is also thought to be common in fMRI measurements, since spatially contiguous voxels may contain similar information. This observation has motivated the use of both spatial smoothing and Gaussian random field application to forward model analysis in fMRI data.

However, we show here that this redundancy also further complicates the issue of interpretation in backward models.

3 Noise Effects Recovery of Activations

Here, we continue with this same illustrative example as before. We have two measurement channels, $\tilde{x}_1(n)$ and $\tilde{x}_2(n)$, and we are trying to discriminate between two tasks. The first channel is informative, and the class means differ. In the second channel, the class means are equal, and does not contain class-specific information. Here, we are interested in testing the extent to which back projecting the weights or extraction filters onto activations is effective, when the noise channel, $\tilde{x}_2(n)$, has varying noise profiles. In each case, the correlation of the data with the task $\tilde{y}$ for the informative voxel measurement remains the same $r = Corr(\tilde{y}, \tilde{x_1}) = 0.75$, and the correlation of the noise channel provides poor separability of the measurements, $r = Corr(\tilde{y}, \tilde{x_2}) = 0.09$. For classification, we use a simple LDA model, assuming Gaussian within-class distributions.

In Figure 2, four different time courses are simulated that represent noise commonly measured in fMRI. In the top row, Gaussian noise is shown. When LDA is applied in combination with this noisy measurement and the informative measurement, the feature weights or extractions assign a strong weight to the noise channel, and a weak weight to the informative channel. However, when these extractions are mapped onto activations using Eq 2., we effectively recover the true infomation contained in each channel.

However, in each of the following three cases, the information is not effectively recovered. Scanner drift is very common in the scanning environment, and temporally contiguous portions of drift typically remain in the data, even after high pass filtering. In the case of drift noise, the informative measurement is correctly assigned a higher weight than the noise measurement. However, after projecting onto activations, the reverse happens - the noise measurement is assigned a stronger weight than the channel with class specific information.

Noise due to head motion is also common in the scanning environment. Although motion correction algorithms exist, head motion that remains in the data has continued to cause problems with fMRI analysis (eg. Power 2014). In the case where the motion introduced causes a step change in the noise, the result is simlar to that of above in the drift motion case; the activations do not reflect the actual information within the data.

4 Redundancy Effects Recovery of Activations

In functional neuroimaging, features can be extracted from a diverse set of variables. Interestingly the extent to which features are statistically independent can also influence the ability to interpret decoding output. For example, in the rare circumstance where all features are completely independent and orthogonal, then their inner product is zero, making the weights directly interpretable. In this case, a naive Bayes classifier may be applied since each element is treated as conditionally independent (e.g., Douglas et al. 2014). However, in most cases, the features are not statistically independent (i.e., spatially adjacent neuroimaging voxels) and may contain redundant information.

In early decoding studies, feature selection was applied a separate step for computational and numerical considerations. In the simplest case the data were simply variance thresholded. Since then, t-test filtering, wrapper methods and a priori selection of an ROI have been used to reduce the number of inputs. However, many of these methods favor selection of features that maximize accuracy individually (Guyon et al. 2002). The searchlight method has also remained popular, whereby spatially contiguous and highly correlated voxels are pooled together. However, this method may still result in spatially distinct searchlight features that have highly correlated temporal information.

Here, we continue with our example with the simple LDA case. However, instead of using a noise channel, we use a redundant channel, $\tilde{x}_j(n)$. Figure 3 shows the timecourse of each of our channels, which are both themselves informative. Figure 3 (d) shows what happens to the activation weight on the initial informative channel $\tilde{x}_1(n)$ when additional redundant features are added to the LDA classifier.

Our results show that introducing redundant channels lowers the resulting activation weights for the informative channel. This similarly suggests that groups of highly-relevant features may show

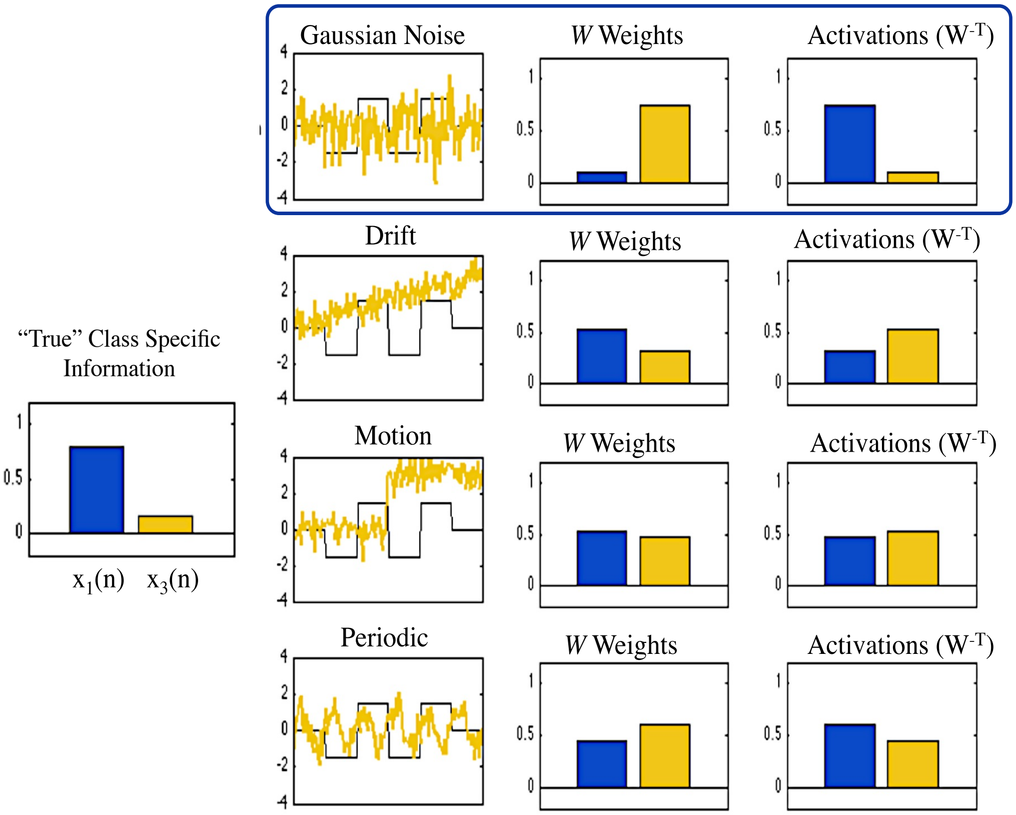

Figure 2: Noise and the simple linear discriminant analysis case with two data measurements. On the left hand side, the r value is shown for each measurment channel. The blue channel, x_1 is the informative channel and the gold channel x_2 is the noise channel. This value remains constant in each case shown. On the top row, the gold channel contains the timecourse data for the gaussian noise case, which is overlaid onto a black line which represents the time course of the task switching between experimental conditions as is the case in classic block designs in fMRI studies. The middle column shows the extraction filter weights. In the Guassian case on the top row, the noise channel is assigned a stronger weight than the informative channel. However, the correct weights are recovered as activations, after applying the Haufe et al. (2014) method (top right). Rows 2, 3, and 4, show results for scanner drift, head motion, and periodic noise. In each of these cases, the activaitons no longer reflect the channel information.

activaton weights that are similar to those of noise, because the individual activation weights are diminished when they are large in number, despite their relevance. A low activation weight is not specific for an irrelevant unit.

5 Conclusions

Even with linear classifiers, the feature weights or model parameters interact to predict the response pattern label. This multivariate interaction renders the interpretation of feature weights complex. Projecting extraction filters onto activations can improve the ability to interpret decoding or backward models. However, we observed that this method is ineffective when the noise characteristics deviate from the assumption that they are Gaussian distributed. Scanner drift, head motion, and sources of periodic noise are common in the fMRI environment, and may pose problems even after data scrubbing. In the case when these three types of noise were present in a measurement channel, the activation method was unable to recover the class specific information contained in either the informative channel or the noise channel - even in the simplest case of a binary discrimination task with two features using LDA.

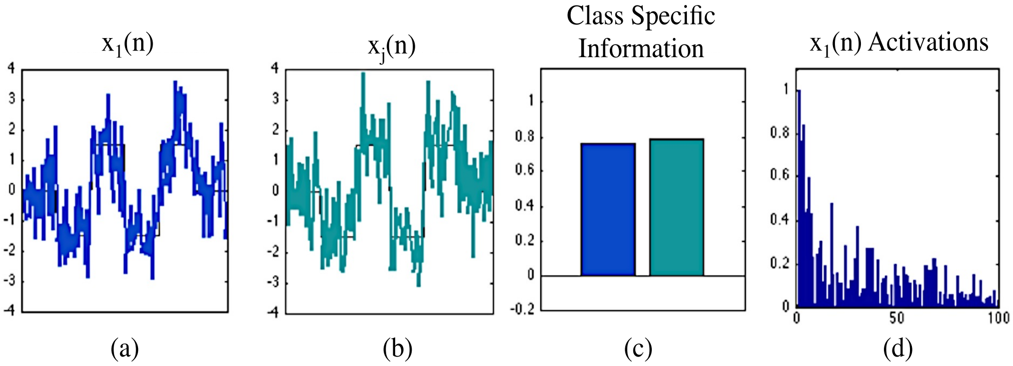

Figure 3: Redundancy and the simple linear discriminant analysis case with a binary classification task. (a) the timecourse of x_1 is shown, overlaid onto the black line which indicates the task switching (b) the timecourse of an example redundant feature is similarly shown (c) the r or class specific information in each channel is shown (d) the activation weight on the initial channel is show over increasing the number of redundant data measurements from 1 to 100.

It can also be shown that utilizing the activation recovery method is circular when the assumptions of LDA are met. In the case where the noise follows a Gaussian distribution and the covariance matrices are full rank, the resulting activation weights are proportional to the values that one would have obtained had they used a classic encoding forward model within a general linear model context for examining fMRI contrasts.

Redundancy is also problematic to interpreation. We observed a diminution in the activations information present in our informative channel that appeared to decline nearly exponentially. After only a few redundant voxels, the information dropped considerably, and dropped nearly to zero after one hundred. Initially, one hundred redundant features may appear large. However, fMRI data typically contain over 100,000 voxels. If voxels themeselves are being used as the features for decoding, it does not seem unreasonable to expect that many would be measuring activations time courses that are similar, and may even be spatially distributed due to the functional connectivity of the human brain. Both non-Gaussian noise and redundancy should be taken into consideration when using the activation method, which is now implemented in many neuroimaging analysis toolboxes. Feature selection methods that reduce redundancy, and or imposing additional sparsity constraints (e.g., L1 norm, elastic method, etc.) may reduce problems with redundancy (Xie et al. 2017).

Although most techniques for classification of neuroimaging data have been based on linear methods (Blankertz et al. 2008; Douglas et al. 2011; Douglas et al. 2013), deep neural networks (DNN) have recently started to be applied to these data (e.g., Khaligh-Razavi and Kriegeskorte 2014). A limiting factor in the applicability of DNNs to neuroimaging data was the idea that these black box methods were not interpretable. Layer-wise relevance propagation (LRP) has recently provided a principled mechanism for attributing the share that each input variable contributes to the classification decision. This method was recently applied to interpret DNNs used to classify EEG data, revealing how biologically plausible attributes at each layer of the network contributed to the classification decision (Strum et al. 2016). Moving forward, DNNs in combination with the LRP method may prove highly useful within the neuroimaging domain where the application of machine learning is dually motivated by classification performance and interpretation.

Acknowledgments

The authors are very thankful for funding from a NARSAD young investigator award, and an NIH K01 MH110645-01A1 award.

References

1. Worsley, K. J. Local Maxima and the Expected Euler Characteristic of Excursion Sets of 2 , F and t Fields. Adv. Appl. Probab. 26, (1994).
2. Poldrack, R. A. Can cognitive processes be inferred from neuroimaging data? Trends Cogn. Sci. 10, 59–63 (2006).
3. Kanwisher, N., McDermott, J. Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. Off. J. Soc. Neurosci. 17, 4302–4311 (1997).
4. Cukur, T., Huth, A. G., Nishimoto, S. Gallant, J. L. Functional Subdomains within Human FFA. J. Neurosci. 33, 16748–16766 (2013).
5. Friston, K. J. Modalities, Modes, and Models in Functional Neuroimaging. Science 326, 399–403 (2009). 6. Blankertz, B., Lemm, S., Treder, M., Haufe, S. Müller, K.-R. Single-trial analysis and classification of ERP components — A tutorial. NeuroImage 56, 814–825 (2011).
7. Kamitani, Y. Tong, F. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8, 679–685 (2005).
8. Haynes, J.-D. Rees, G. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523–534 (2006).
9. Guyon, I. Elisseeff, A. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 1157–1182 (2003).
10. Haufe, S. et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87, 96–110 (2014).
11. Kriegeskorte, N. Pattern-information analysis: From stimulus decoding to computational-model testing. NeuroImage 56, 411–421 (2011).
12. Norman, K. A., Polyn, S. M., Detre, G. J. Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10, 424–430 (2006).
13. Power, J. D. et al. Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage 84, 320–341 (2014).
14. Douglas, P. K. et al. Single trial decoding of belief decision making from EEG and fMRI data using independent components features. Front. Hum. Neurosci. 7, (2013).
15. Xie, J, Douglas, P. K, Wu, Y. N, Brody, A. L, Anderson, AE (2017). Decoding the Encoding of Functional Brain Networks: an fMRI Classification Comparison of Non-negative Matrix Factorization (NMF), Independent Component Analysis (ICA), and Sparse Coding Algorithms. Journal of Neuroscience Methods
16. Douglas, PK, Harris, S. Yuille,A, Cohen, MS. Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. Neuroimage. May 15;56(2):544-53 (2011)
17. Khaligh-Razavi, SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. Nov 6;10(11) (2014).
18. Sturm, I, Lapuschkin, S., Samek, W, Muller, K.-R Interpretable deep neural networks for single-trial EEG classification. Journal of Neuroscience Methods 274:141–145 (2016)