Component encoding: Interpretable and predictive models of neural computation

David Skrill (david_skrill@urmc.rochester.edu)
Department of Biostatistics and Computational Biology
University of Rochester Medical Center, Rochester, NY 14642

Sam V. Norman-Haignere (samuel_norman-haignere@urmc.rochester.edu)
Departments of Biostatistics and Computational Biology, Neuroscience, Brain and Cognitive Science, Biomedical Engineering
University of Rochester Medical Center, Rochester, NY 14642

Abstract

A central goal of sensory neuroscience is to build parsimonious computational models that can both predict neural responses to natural stimuli and reveal interpretable functional organization in the brain. Statistical "component" models can learn interpretable, low-dimensional structure across different brain regions and subjects, but lack an explicit "encoding model" that links these components to the stimuli that drive them, and thus cannot generate predictions for new stimuli or generalize across different experiments. The predictive power of standard encoding models has improved substantially with advances in deep neural network (DNN) modeling, but producing simple and generalizable insights from these models has been challenging. To overcome these limitations, we develop "component-encoding models" (CEMs) which approximate neural responses as a weighted sum of a small number of component response dimensions, each approximated by an encoding model. We show using simulations and fMRI data that our CEM framework can infer a small number of interpretable response dimensions across different experiments with non-overlapping stimuli and subjects (unlike standard components) while maintaining and even improving the prediction accuracy of standard encoding models.

Keywords: encoding model; natural stimuli; functional MRI; deep neural networks; latent variable model

Introduction

Understanding the neural computations that allow people to derive information from natural stimuli, such as speech and music, is a fundamental goal of sensory neuroscience. This task is challenging in part because natural stimuli are complex and high dimensional and in part because sensory systems encode natural stimuli using a highly nonlinear stimulus-response mapping.

Statistical component methods have proven useful in addressing the first challenge by revealing interpretable and low-dimensional structure. For example, human fMRI responses to natural sounds can be accurately approximated by a small number (5-10) component response patterns, each reflecting selectively for specific acoustic features (e.g., frequency, spectrotemporal modulation) and sound categories (e.g., speech, music, singing) within distinct sub-regions of the auditory cortex (Norman-Haignere et al. (2015), Boebinger et al. (2021)). Component models, however, cannot predict responses to new stimuli or generalize across different experiments testing distinct stimuli and subjects because they lack an encoding model that links neural responses to the stimuli.

Advances in deep neural network (DNN) modeling have substantially improved the predictive power of standard encoding models across many different sensory and cognitive systems (Yamins et al. (2014), Yamins & DiCarlo (2016), Kell et al. (2018), Schrimpf et al. (2021)). Encoding model are typically fit by mapping a high-dimensional feature set, learned by a pretrained DNN, onto a set of neural responses, using a separate mapping for each response. Despite their impressive predictive power, deriving generalizable scientific insights from these models has been challenging, in part because they learn a different high-dimensional mapping for each neuron, electrode, or voxel.

To combine the strengths of these two approaches, we develop "component-encoding models" (CEMs) that synthesize the respective benefits of these two approaches.

Model Definition

The input to a CEM is a collection of response timecourses, concatenated as a $n_{time} \times n_{channel}$ matrix, $\mathbf{D}$, where the channel dimension could reflect any neural response (e.g., voxel, electrode). Responses from multiple stimuli and subjects are concatenated across the time and channel axis, respectively.

In a standard component model, the data matrix is approximated as the product of a low-dimensional response matrix ($\mathbf{R}$: $n_{time} \times n_{components}$) and weight matrix ($\mathbf{W}$: $n_{components} \times n_{voxels}$), where the weights determine the contribution of each component to each channel (e.g., voxel):

$$\mathbf{D} \approx \mathbf{RW} \tag{1}$$

The solution to equation 1 is typically constrained by additional statistical criteria since matrix factorization is otherwise ill-posed (accomplished here by maximizing non-Gaussianity of the weights; Norman-Haignere et al. (2015)). Importantly, the factorization is performed using statistical properties of the data matrix alone without any information about the stimuli.

Encoding models are typically fit by approximating each channel as the weighted sum of a high-dimensional feature set, computed from the stimuli:

$$\mathbf{D} \approx \mathbf{NB} \tag{2}$$

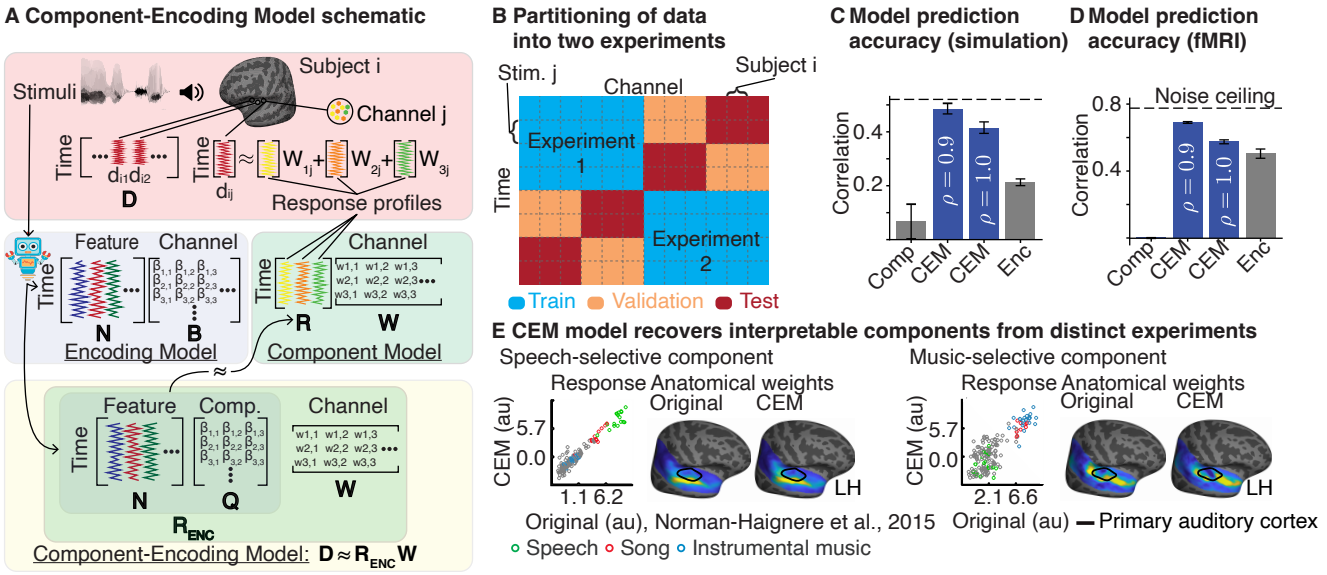

Figure 1: **A** Schematic of component-encoding model (CEM). See text for details. **B** Schematic showing the partitioning of a single experiment into sub-experiments (blue on-diagonal blocks) with non-overlapping time-points/stimuli and subjects; off-diagonal blocks are used for validation and test. **C, D** Prediction accuracy (correlation in test data) of simulated (C) and real (D) fMRI data from two CEMs (blue bars) either partially ($\rho = 0.9$) or fully ($\rho = 1$) constrained by an encoding model, along with standard component-only and encoding-only models (gray bars) (dotted line shows the noise ceiling). **E** CEMs recover interpretable speech- (left) and music-selective (right) components from non-overlapping sub-experiments, closely matching those inferred from component-only models with complete data.

where $\mathbf{N}$ is the feature matrix ($n_{time} \times n_{feature}$, e.g., unit responses from a pre-trained DNN) and $\mathbf{B}$ ($n_{feature} \times n_{channel}$) maps from the features to the response, separately for each channel.

We construct a CEM by approximating the low-dimensional component response matrix, instead of the neural data directly, using an encoding model:

$$\mathbf{R} \approx \mathbf{R}_{ENC} = \mathbf{NQ} \tag{3}$$

In our framework, we simultaneously encourage the encoding model to explain the data while enabling the CEM to learn low-dimensional structure that is not fully predictable by the encoding model ($\mathbf{R} \neq \mathbf{R}_{ENC}$) by minimizing the following loss with respect to the model parameters ($\hat{\mathbf{R}}_{ENC}, \hat{\mathbf{R}}, \hat{\mathbf{Q}}, \hat{\mathbf{W}}$):

$$\rho||(\mathbf{D} - \hat{\mathbf{R}}_{ENC}\hat{\mathbf{W}})||_F^2 + (1-\rho)||(\mathbf{D} - \hat{\mathbf{R}}\hat{\mathbf{W}})||_F^2 \tag{4}$$

The first term fully relies on the encoding model, while the second term is fully data-driven term, with ρ controlling their relative strength. Critically, the component weights, $\hat{\mathbf{W}}$, are shared between these two terms, which forces alignment between the constrained ($\hat{\mathbf{R}}_{ENC}$) and unconstrained ($\hat{\mathbf{R}}$) component responses.

Results

To demonstrate the utility of CEMs, we focus on an important application: inferring components across different experiments, testing distinct stimuli and subjects. Standard component models cannot accomplish this task (as demonstrated below) because there is no overlap between the time or channel dimension, and even if there is anatomical correspondence, functional responses vary substantially with respect to anatomy (Saxe et al. (2006)).

The data from two different experiments can be represented as a block-diagonal matrix (blue blocks, **Fig.1B**) with off-diagonal blocks corresponding to the "missing" stimuli from each experiment. A successful CEM should make it possible to predict the missing data. To test this possibility, we partitioned data from a single experiment into two "sub-experiments" as illustrated in Fig.**1B**. We trained the model on these sub-experiments (blue diagonals), and used the remaining off-diagonal blocks for validation (green diagonals) and test (red diagonals).

We tested our approach using simulations and real data from a prior fMRI experiment (Norman-Haignere et al. (2015)) (11,065 voxels, 10 participants, 165 natural sounds). Simulated data was designed to be similar to the fMRI data in SNR, dimensionality, and encoding model prediction accuracy. Features were extracted from the audio embedding of a DNN pre-trained on a large set of natural sounds (CLAP; top 40 principal components; Wu* et al. (2023), Chen et al. (2022)). We fit the CEM using 6 components, as in the prior study, and compared its performance with a component-only model (ρ=0) and encoding-only model (ρ=0 and $\mathbf{Q}$ equal to the identity matrix; results were similar when encoding models were fit using ridge regression).

For both simulated and real data, we found that CEMs sub-

stantially outperformed our component-only model with prediction accuracies approaching the noise ceiling (the ceiling is high because voxel responses were averaged across time and repetitions) (**Fig.1C,D**). Moreover, CEMs that allowed for encoding-model error ($\rho = 0.9$) outperformed CEMs fully constrained by the encoding model ($\rho = 1$), which in turn outperformed an encoding-only model that learned a separate mapping for each voxel. This finding shows that CEMs can out-perform standard encoding models both because CEMs can discard unreliable, voxel-specific response variation and because they can model low-dimensional structure that is not fully predictable by an encoding model. The learned components from our fMRI data exhibited clearly interpretable functional and anatomical structure that closely matched those from the original study using complete data (**Fig.1E**). For example, we observed distinct music- and speech-selective components that clustered in different regions of non-primary auditory cortex.

These findings demonstrate that CEMs can synthesize the strengths of component and encoding models, while overcoming their weaknesses. Unlike standard encoding models, CEMs can explain responses across many different regions, subjects, and experiments using a small number of interpretable response dimensions, even when those response dimensions are not fully explainable from an existing encoding model. Unlike component models, CEMs can make predictions for new stimuli, which enables successful generalization across multiple experiments. Our framework is not specific to sounds or fMRI and thus is likely to be broadly useful in modeling neural responses to natural stimuli across many different sensory and cognitive systems.

References

Boebinger, D., Norman-Haignere, S. V., McDermott, J. H., & Kanwisher, N. (2021). Music-selective neural populations arise without musical training [Journal Article]. *J Neurophysiol*, *125*(6), 2237-2263. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/33596723 doi: 10.1152/jn.00588.2020

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., & Dubnov, S. (2022). Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *Ieee international conference on acoustics, speech and signal processing, icassp.*

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 [Journal Article]. *Neuron*, *98*(3), 630-644 e16. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29681533 doi: 10.1016/j.neuron.2018.03.044

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition [Journal Article]. *Neuron*, *88*(6), 1281-1296. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26687225 doi: 10.1016/j.neuron.2015.11.035

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: a defense of functional localizers [Journal Article]. *Neuroimage*, *30*(4), 1088-96; discussion 1097-9. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16635578 doi: 10.1016/j.neuroimage.2005.12.062

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing [Journal Article]. *Proc Natl Acad Sci U S A*, *118*(45). Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/34737231 doi: 10.1073/pnas.2105646118

Wu*, Y., Chen*, K., Zhang*, T., Hui*, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Ieee international conference on acoustics, speech and signal processing, icassp.*

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex [Journal Article]. *Nat Neurosci*, *19*(3), 356-65. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26906502 doi: 10.1038/nn.4244

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex [Journal Article]. *Proc Natl Acad Sci U S A*, *111*(23), 8619-24. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/24812127 doi: 10.1073/pnas.1403112111